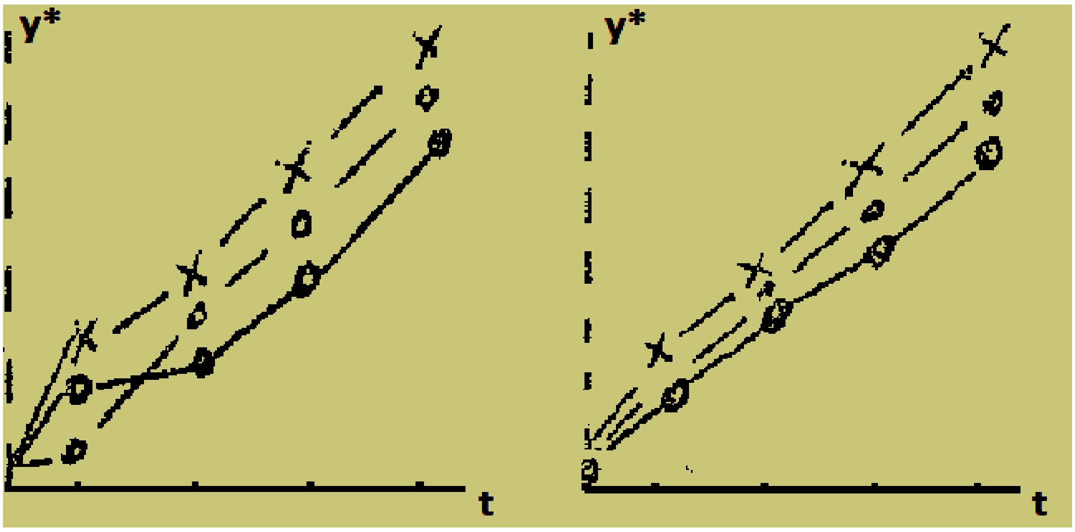
Mihail Motzev

BUSINESS FORECASTING

(A CONTEMPORARY DECISION MAKING APPROACH)



(*Second edition – printed version*)

2021

Business Forecasting: A Contemporary Decision Making Approach
© 1st edition, 2016, Mihail Motzev & Eudaimonia Productions Ltd.
ISBN 978-619-7209-17-4
© 2nd edition, 2021, Mihail Motzev, ISBN 978-1-7370258-1-8

Brief **CONTENTS**

Table of CONTENTS

PREFACE

Business forecasts importance has been recognized world-wide. Just recently, in about one year period, a couple of new books about forecasting and related problems, written by two very influential people[1], have been published.

Understanding of forecasting techniques is crucial not only for managers, but also for any decision maker, notwithstanding the number of their responsibilities in contemporary business organizations, must be alert to the improper use of forecasting techniques because inaccurate forecasts can lead to poor decisions. Unfortunately, not all managers/decision makers realize this. One important reason for this is because many college graduates, including those with degrees in business, do not ever study forecasting, except as a sidelight in courses like Statistics and Operations Management which have other primary objectives.

Another reason is that most of the existing books in Business Forecasting present only the basic techniques and many contemporary approaches, like business intelligence and predictive analytics, knowledge discovery from data and data mining (including artificial neural networks, genetic algorithms, self-organization and other intelligent tools), which are very useful today in forecasts development, are just mentioned.

The focus of this book is in incorporating the latest findings from both theory and practical research. The book not only presents general principles and fundamentals that underlie forecasting practice, but also introduces both standard and advanced approaches of forecasting, with main emphasis on data mining and predictive analytics. The purpose of the book is to help decision makers and forecasters carry out their job and to enable students to prepare for a managerial and analytical career.

[1] Nate Silver, who was named one of the world's 100 Most Influential People by the Time Magazine – see The Signal and the Noise (Why So Many Predictions Fail –_But Some Don't), The Penguin Press, New York, 2012, and Alan Greenspan, the former Chairman of the Board of Governors of the Federal Reserve System, who served for the longest period of time so far (approximately 19 years), for four Presidents of the USA – see The Map and the Territory (Risk, Human Nature, and the Future of Forecasting), The Penguin Press, New York, 2013.

BOOK ORGANIZATION OF THE SECOND EDITION

This textbook has been designed to provide students with enough understanding of the fundamentals and the major details of all important techniques, which are necessary to develop an appropriate forecast, to select a good one among many others and to apply it when solving real-life business problems.

The content of the book is organized into a few major groups:

- The fundamentals of business forecasting are discussed in Chapters 1 and 2 – Chapter 1 (which combines now the former Chapters 1 and 2) presents a brief historical view and some basic insights about forecasting and also discusses the nature and the need for Business Forecasting, and the common features to all forecasts as well as the characteristics of good forecasts. Chapter 2 presents some important fundamentals from other areas and their relations to Business Forecasting, such as Decision Making and the General System Theory. It also discusses the main steps in Business Forecasting, the general forecasting model and the major model building approaches in forecasting.
- Forecast error and accuracy are discussed in Chapter 3 (former Chapter 4) along with the forecasting techniques evaluation and selection of the most appropriate forecasting model. One very important point here is *Cross-Validation principle* (i.e. using an external supplement) as *regularization method* in *Best-model selection procedure* as an analog of *Gödel's incompleteness theorems.* Another original input here is the introduction of *Self-Organizing Modeling* and its applications in variables and model selection.
- The next group presents and discusses intuitive (subjective) techniques (Chapter 4 - former 5) and basic quantitative techniques used in business forecasting (Chapter 5 – former 6), such as Naïve Forecasts and Graphics, Moving Averages, Exponential Smoothing and other simple models and smoothing techniques.
- Chapters 6, 7 and 8 (former Chapters 7 and 8) cover the most common techniques, considered by many professionals the "classics" in business forecasting, regression and time series analyses. The former one presents regression models and their application in forecasting, discussing in detail all related topics, such as model building, estimation methods, aptness of the model, residual analysis, multicollinearity, etc. The latter (now in two parts) covers Time Series Analysis and

Autoregressive models. Time-Series decomposition, Seasonally-Adjusted forecasts, Regression with Time Series data and ARIMA methodology are presented. A special point of interest and original input here is *Time Series Forecasting using Data Mining techniques.*

- Complex Forecasting methods and techniques are discussed in Chapter 9, which presents advanced approaches like causal econometric modeling (both Single Equation Models and Simultaneous Equations Models) and emphasizes on complex model building and forecasting using *Self-Organizing Data Mining*. *Group Method of Data Handling (GMDH)* and GMDH based algorithms are also discussed and a special attention is dedicated to their implementations in developing *Artificial Neural Networks* for predictive modeling and in particular to the *Multi-Layered Net of Active Neurons (MLNAN)* technique, which is a multilayer *GMDH algorithm* for multi-input to multi-output models identification.
- The last chapters present new techniques and methods in Business Forecasting, which in general are referred to as *Business Intelligence (BI)* and *Business Analytics (BA)*. Chapter 10 introduces *BI&BA* and their relations with Decision Support Systems, Competitive Intelligence, Data Warehousing and so on, as well as *Data Mining* as an important component of BA. The last topics in this chapter present *BI* solutions and platforms, and *Data Mining* integrated with *BI* applications. Chapter 11 discusses in detail the major *Data Mining* techniques, such as *clustering*, *decision trees* and *Artificial Neural Networks (ANNs)*. *Self-Organizing Data Mining*, which incorporates some of the most advanced techniques, such as *Genetic Algorithms*, *Multi Stage Selection Procedures*, *Cross-Validation* and so on, is presented in Chapter 12. This chapter discusses in particular *Group Method of Data Handling (GMDH)* algorithms and their applications in business forecasting, emphasizing on specific software platforms like KnowledgeMiner and techniques for predictive modeling, such as *Multi-Layered Net of Active Neurons (MLNAN).*

NEW AND UNIQUE EMPHASIS

The following original contributions from this book to the Business Forecasting theory and practice can be pointed out and summarized as follows:

∗ Chapter 2, Sections:

– 2.1. *Business Forecasting and Decision Making* and 2.2. *General System Theory and Business Forecasting* present some important fundamentals from other functional and theoretical areas such as *Decision Making* and the *General System Theory* and also discuss their relations with *Business Forecasting*. This way a new perspective is put on business forecasting and the decision-making approach of the book is emphasized.

– 2.4. *General Forecasting Model and Model Building Approaches;* it presents both the ***theory-driven*** one and the ***data-driven*** or experimental systems analysis approaches, discussing pros and cons for each of them and emphasizing the need for new hybrid approach.

∗ Chapter 3:

– Sections 3.2. *Measures of Forecast Accuracy* and 3.3. *Forecasting Techniques Evaluation and Model Selection*, discuss ***Cross-Validation*** *principle* (i.e. using an *external supplement* or a *regularization method,* according to *Gödel's incompleteness theorems*) as a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. This is asymptotically equivalent to *Akaike's Criterion ICA* in measuring the Forecast Accuracy and the Forecasting Error, and for evaluating a forecasting technique, but it is more objective and it helps to address the *overfitting* problem as well.

– Section 3.4. *Self-Organizing Modeling* introduces one little-known, but very effective approach in complex systems modeling and forecasting. The theory of ***self-organizing modeling*** has widened the capabilities of system identification, forecasting, pattern recognition and multi-criteria decision making. In fact, it provides a new, third view (also known as *"hybrid approach"*) to the *theory-driven* (or theoretical systems analysis) and *data-driven* (or experimental systems analysis) approaches in model building problem: *"what variables, which method, which model.* This section also introduces one real-life application of this approach – the ***Group Method of Data Handling (GMDH)***, which is a heuristic, self-organizing modeling method developed by A.G. Ivakhnenko (1968). It contains a family of inductive algorithms (discussed in the following chapters) for computer-based mathematical modeling of multi-parametric datasets that feature fully automatic structural and parametric identification of models.

✱ Chapter 6, Section 6.3. *Multiple Regression and Model Building*:

▬ GMDH is suggested as a tool to address multicollinearity. Particular GMDH algorithms that solve the issue of multicollinearity and many other problems in regression model building and forecasting are presented in Chapters 8, 9 and 12.

▬ Since the results of a stepwise regression are often used incorrectly without adjusting them for the occurrence of model selection, the model assessment and evaluation of its performance should be done on a set of data not used for training. The usage of *Cross-validation* procedure, as a part of *Self-organizing data mining* algorithm, was suggested. In *cross-validation*, a regression model is usually given a dataset of known data on which training is run (training dataset), and a dataset of unknown data (or first seen data) against which the model is tested (testing dataset) and validated. If the validation error increases while the training error steadily decreases, then usually a situation of overfitting have occurred. The best predictive and fitted model would be when the validation error has its global minimum.

▬ When selecting an appropriate functional form for the model, *Self-Organizing Data mining* methods, such as *GMDH*, not only provide better platform for model building for expert forecasters, but also give a great support to nonqualified users, who cannot comprehend the rules on how to build and select the right model specification. To prove this statement, special *GMDH* algorithms that perform the modeling process as highly automated procedure, according to data patterns or particular properties of the regression model, are discussed in detail later on in Chapters 8, 9 and 12.

✱ Chapter 8:

▬ Section 8.3. *Time Series Forecasting Using Data Mining Techniques*. Outside the traditional statistical modeling (stochastic methods) for time series modeling and forecasting, an enormous amount of forecasting is done using *Data Mining* techniques. *Artificial Neural Networks (ANNs)* are suggested to address existing problems. This section:

- Introduces *GMDH*-type *ANN* as a multilayered inductive procedure, which is equivalent to the *ANNs* with polynomial activation function of neurons.
- Discusses ***Multi-Stage Selection*** algorithms as specific *GMDH*-type *ANNs* that use the idea of ***Genetic Algorithms (GA)*** and multilayer organization.
- Presents a unique, developed with author participation, working prototype of a ***Multi-Layered Net of Active Neurons (MLNAN)*** as a very useful tool in business forecasting for building multi-input to single-output models (different type of regression models) as

well as econometric models of SE (i.e. multi-input to multi-output models – presented in Chapter 9).

- Comments some examples of using *ANNs* for Time Series Analysis and Forecasting and the results obtained.

∗ Chapter 9. *Complex Forecasting Techniques*:

▬ Section 9.2. *Simultaneous Equations Models*. A variable could be dependent in one equation and a regressor in others in statistical models of linear S*imultaneous Equations* (*SE*), i.e. multi-input to multi-output models, which are one of the most advanced model's types. These models are not usually presented in forecasting books, while in sections 9.2 and 9.3 *SE* are introduced in detail and existing problems and potential solutions are presented and discussed.

▬ Section 9.3. *Complex Model Building and Forecasting Using Self-Organizing Data Mining*. In this section:

- *GMDH* algorithms for complex models specification are presented. This Self-organizing modeling technique is based on statistical learning networks, which are networks of mathematical functions that capture complex (both linear and non-linear) relationships in a compact and rapidly executable form. Such networks subdivide a problem into manageable pieces or nodes and then automatically apply advanced regression techniques to solve each of these much simpler problems. These tools are very useful for addressing model-building problems already discussed in the book.
- Special attention is given to the *GMDH* Algorithms for ***Self-organization of Active-Neuron Neural Networks***. The ***Multi-Layered Net of Active Neurons (MLNAN)*** algorithm, described in Chapter 8, is presented and applied for developing a small macroeconomic forecasting model in the form of *Simultaneous Equations*. The model and its properties are analyzed and a comparative evaluation of *ex-ante* and *ex-post* forecasts are discussed in detail.
- *MLNAN* is used for building *SE* models with both *stationary* and *dynamic coefficients*. A comparative evaluation of the predictions with the stationary and the dynamic models is presented and discussed.
- *MLNAN* applications in *Distributed lag models, AR, ARMAX* and *VAR models* are presented and discussed. All these examples provide evidences that the *MLNAN*, as other *SODM* techniques, is a cost-effective tool for building such models with high accuracy and reliability.

- Applications of ***Self-Organizing Data Mining (SODM)*** techniques and the *MLNAN* algorithms in macroeconomic modeling for complex systems, such as the US Economy, Germany, Bulgaria and other countries are presented and discussed. Results confirm that these tools are very useful for addressing model-building problems, for example, overfitting is eliminated by the use of external criterion, i.e. the cross-validation method. The small number of independent measurements (or short time-series) does not cause problems too, because the inverted matrix size is always 2x2 (pair-wise combinations). This feature helps in dealing with the problem of multicollinearity as well. The autocorrelation is eliminated by adding automatically (when needed) lagged variables and so forth. Last, but not least, these techniques are totally automated procedures with strong user-friendly interface which provides opportunities for a forecaster (or decision maker), who at the crucial points of the process has options to apply additional insights, knowledge or hypotheses.

∗ Chapter 10. *FORECASTING, BUSINESS INTELLIGENCE AND BUSINESS ANALYTICS:*

‒ Introduces *Business Intelligence* and *Business Analytics* (*BI&BA*) and their relations with *Decision Support Systems, Competitive Intelligence, Data Warehousing* and so on, as well as *Data Mining* as an important component of *BA*.

‒ *BI* solutions and platforms, and *Data Mining* integrated with *BI* applications, as tools for model building and forecasting, are presented and discussed.

∗ Chapter 11. *BUSINESS FORECASTING AND DATA MINING*:

‒ Introduces *Knowledge Discovery from Data (KDD)* and presents *Data Mining* techniques as a powerful tool in analyzing the massive amounts of data and turning the information located in the data into successful decisions. *KDD* process has a wide range of applications and business forecasting is just one of them.

‒ The most comprehensive *data mining* techniques such as *decision trees, clustering and artificial neural networks* (*ANNs*) are discussed in detail revealing their advantages and weaknesses. In addition, the need of a new approach to them (*ANN*s in particular) to address existing problems is pointed out and *Statistical Learning Networks*, like *GMDH* based *ANNs* are suggested as a potential solution.

∗ Chapter 12. *SELF-ORGANIZING DATA MINING AND FORECASTING.*

‒ Section 12.1. *GMDH based Self-Organizing Data Mining Algorithms;* it presents these algorithms, their advantages, details and potential applications. In GMDH algorithms, models are generated adaptively from data in the form of networks of active neurons. In this procedure, a repetitive generation of populations of competing models of growing complexity, corresponding validation, and model selection are done until an optimal complex model, neither too simple nor too complex, has been identified. This modeling approach grows a tree-like network out of data of input and output variables (seed information) in a pair-wise combination and competitive selection from a simple single individual (neuron) to a desired final solution that does not have an overspecialized behavior (model). In this approach, neither the number of neurons and the number of layers in the network, nor the actual behavior of each created neuron is predefined. The modeling is self-organizing because the number of neurons, the number of layers, and the actual behavior of each created neuron are adjusting during the process of self-organization.

- Section 12.4. *Self-Organizing Data Mining Platforms*. This section discusses possible platforms for automated model building and forecasting. In this way, modern data mining tools are no longer restricted to specialists. As more organizations adopt predictive analytics into decision-making processes and integrate it into their operations, they are creating a shift in the market toward the business users as the primary consumers of the information. Business users want tools they can use on their own and vendors are responding by creating software that removes the mathematical complexity, provides user-friendly graphic interfaces, and/or builds in shortcuts that can, for example, recognize the kind of data available and suggest an appropriate predictive model. Predictive analytics has become sophisticated enough to adequately present and dissect data problems, so that any data-savvy information worker can utilize these methods to analyze data and retrieve meaningful, useful results. Modern tools like KnowledgeMiner software present results using simple charts, graphs, and scores that indicate the likelihood and/or the level of possible outcomes.

Using the unique software *KnowledgeMiner (yX) for Excel*, an original *Multi-Stage Selection Procedure* based on *GMDH MLNAN* algorithm is designed and applied for building complex forecasting models and conducting different simulation experiments with real-life business systems.

- Section 12.5. *Forecasting Applications of Self-Organizing Data Mining.* Many applications of *GMDH based SODM* have been summarized and some important examples are presented and discussed. The results obtained so far prove that *SODM* and the *GMDH based ANNs* in particular provide opportunities to shorten the design time and reduce the cost and the efforts in model building and forecasting. *MLNAN* and similar techniques are able to develop reliably even complex models with lower overall error rates than other methods.

ACKNOWLEDGMENTS

The author is indebted to many colleagues around the world who helped in the preparation and the publication of this book. Their comments and insights have helped to complete this unique text in Business Forecasting. Particular thanks first of all to Alexey Grigoryevich Ivakhnenko (Head of the Combined Control Systems Division of the Glushkov Institute of Cybernetics, Kiev – RIP) and to Angel Marchev Sr. (retired at the University of National and World Economy, Sofia), Johann Mueller (retired at the University of Applied Sciences, Dresden) and Frank Lemke (Owner and Managing Director at KnowledgeMiner Software, Berlin).

Many thanks to the colleagues at the Walla Walla University School of Business and all other university staff who helped in some ways for this dream to come true.

Particular gratitude also is in order for the publishers at Eudaimonia Productions Ltd. and all reviewers, especially my students Michaela Honner and Andrew Rapp, and Ophelia Pamukchieva from the University of National and World Economy, Sofia, for their assistance throughout the production and finalization of this book.

The author also wishes to acknowledge the support from his family and to thank all of them for the understanding and support during the long days of this project.

Finally, the author is thankful for all the inspirations, wisdom and other blessings provided by our Creator, the Source of eternal knowledge and wisdom!

***

CHAPTER 1. FUNDAMENTALS OF BUSINESS FORECASTING

1.1. Introduction to Forecasting

*“**To predict the future**”* has been a human aspiration since ancient times. One can find evidence in many legends and manuscripts from antiquity. For instance, God revealed information about the future through prophets like Jeremiah and Daniel. Greek army leaders visited the oracle in Delphi to hear the opinion of their gods before attacking the city of Troy.

Predictions have often been made, from antiquity until the present, by resorting to paranormal or supernatural means, such as prophecy or by observing omens. Disciplines including water divining, astrology, numerology, and fortune telling, along with many other forms of divination, have been used for centuries and even millennia to predict or attempt to predict the future.

In literature, vision and prophecy are literary devices used to present a possible timeline of future events. For example, Charles Dickens' “A Christmas Carol” - after Scrooge confronts the visions given to him by the Ghosts of Christmas Past, Present, and Yet to Come, he asks whether the future he has seen can be changed, i.e. he wants to know whether he can change the outcome of the ghosts' prophecies[1].

One very interesting example could be found in the Bible. In the Epistle of James the author is giving a warning specifically about business forecasting[2]. Somewhat unusually, he focuses first on the principle of trusting God. He opens with sobering words: “Go to now, ye that say, ‘To-day or to-morrow we will go into such a city, and continue there a year, and buy and sell, and get gain:’ whereas ye know not what [shall be] on the morrow. For what [is] your life? It is even a vapour, that appeareth for a little time, and then vanisheth away.” From a glance, it might seem that James is condemning even short-term business planning. The process of planning ahead, however, is not his concern. Imagining that we are in control of what happens is the true problem.

The last verse helps us see James’s real point: “For that ye [ought] to say, ‘If the Lord will, we shall live, and do this, or that.’” The problem is not planning in general; it is planning as if the future lies in our hands. We are responsible to use wisely the resources, abilities, connections, and time that God gives us. But we are not in control of the outcomes. Most

[1] http://en.wikipedia.org/wiki/Prediction#Supernatural_.28prophecy.29

[2] See Epistle of James (KJV) Ch. 4, vs 13–15.

businesses are well aware how unpredictable outcomes are, despite the best planning and execution that money can buy. The annual report of any publicly traded corporation will feature a detailed section on risks the company faces, often running many pages. Statements such as “Our stock price may fluctuate based on factors beyond our control” make it clear that secular corporations are highly attuned to the unpredictability James is talking about[3].

In April 1968, *The Club of Rome* was founded by Aurelio Peccei, (an Italian industrialist, VP of FIAT), and Alexander King (a Scottish scientist). It was formed when a small international group of people from the fields of academia, civil society, diplomacy, and industry, met at a villa in Rome, Italy, and defined the original prospectus of the Club of Rome titled "The Predicament of Mankind." This prospectus was founded on a humanistic architecture and the participation of stakeholders in democratic dialogue. Later on, the Club of Rome Executive Committee in the summer of 1970 opted for a mechanistic and elitist methodology for an extrapolated future.

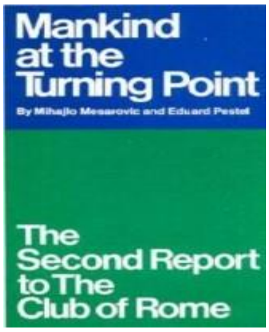
The club states that its essential mission is “to act as a global catalyst for change through the identification and analysis of the crucial problems facing humanity and the communication of such problems to the most important public and private decision makers as well as to the general public."[4]

The Club of Rome raised considerable public attention with its report “Limits to Growth”, which has sold 12 million copies in more than 30 translations, making it the best-selling environmental book in world history (Turner, 2008, p. 52). Published in 1972 and presented for the first time at the International Students' Committee (ISC) annual Management Symposium in St. Gallen, Switzerland, it predicted that economic growth could not continue indefinitely because of the limited availability of natural resources, particularly oil. The 1973 oil crisis increased public concern about this problem. However, even before Limits to Growth was published, Eduard Pestel and Mihajlo Mesarovic of Case Western Reserve University had begun work on a far more elaborate model (it distinguished ten world regions and involved 200,000 equations compared with 1000 in the Meadows model). The research had the full support of the Club and the final publication, *Mankind at the Turning Point* was accepted as the official Second Report to the Club of Rome in 1974. In addition to providing a more refined regional breakdown, Pestel and

[3] https://www.theologyofwork.org/new-testament/general-epistles/james-faith-works/business-forecasting-james-413-17/

[4] http://www.clubofrome.org/eng/about/3/

Mesarovic had succeeded in integrating social as well as technical data. The Second Report revised the predictions of the original Limits to Growth and gave a more optimistic prognosis for the future of the environment, noting that many of the factors were within human control and therefore that environmental and economic catastrophe were preventable or avoidable, hence the title.



Mankind at the Turning Point

By Mihajlo Mesarovic and Eduard Pestel

The Second Report to The Club of Rome

Another important point in the Club of Rome research is its third report published under the title “Reshaping the International Order” (RIO)[5]. It was formulated by a group of about twenty experts from both developing and developed countries. The initiative to undertake this study of the international order was taken by the Club of Rome Board, especially by its chairman Aurelio Peccei and the study was financed by the Netherlands Ministry of Foreign Affairs at the initiative of the Minister for Development Cooperation, Jan Pronk. The report was presented to the Club of Rome in a meeting at Algiers, hosted by the Algerian Government in 1976.

The project coordinator Jan Tinbergen, a Dutch economist, and Ragnar Frisch shared the first Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel in 1969 for having developed and applied dynamic models for the analysis of economic processes. Jan Tinbergen was also a consultant to the *League of Nations*. From 1945 till 1955 he served as the first director of the *Netherlands Bureau for Economic Policy Analysis*. He was a member of the *Royal Netherlands Academy of Arts and Science* and the *International Academy of Science*. In 1956 he founded the *Econometric Institute* at the *Erasmus Universiteit Rotterdam* together with *Henri Theil*, who was also his successor in Rotterdam. The *Tinbergen Institute* was named in his honor.

In 2001 the Club of Rome created *tt30* as a spin-off, an anticipatory thinking (systems thinking) youth think tank[6] for people around the age of 30. Today, the Club has national associations in many countries, including a number of European and Asian countries, USA, Canada, Mexico, Brazil, Puerto Rico and Venezuela. These associations analyze national problems in terms of the same factors and give advice to the countries’ decision-makers.

Global Forecasts importance increases every year, and The Global Forecast System (GFS) – a global numerical weather prediction system containing a global computer model and

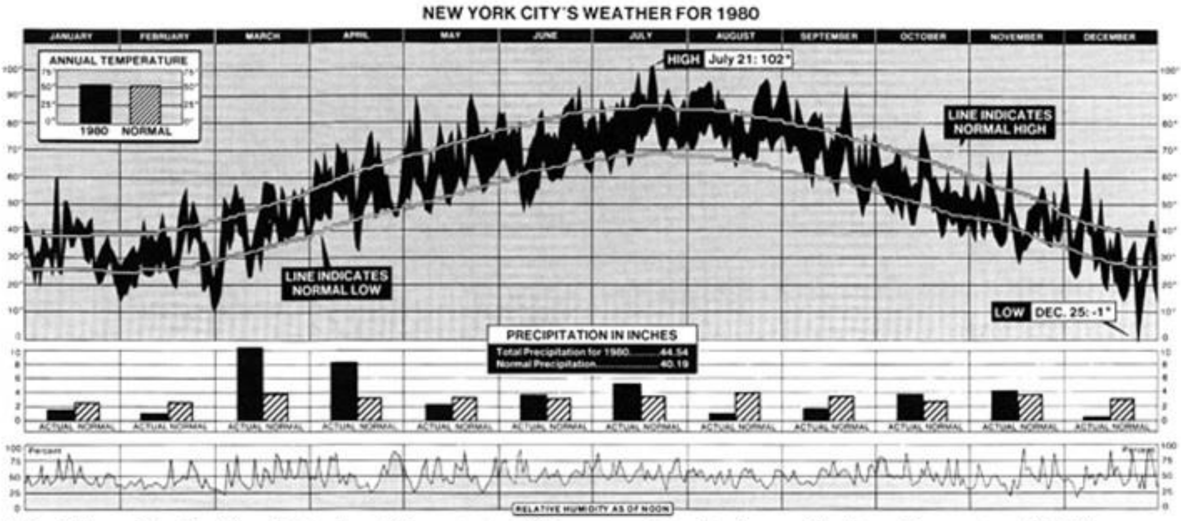
[5] Published as: *Reshaping the International Order, A Report to the Club of Rome*, Jan Tinbergen (coordinator), E.P. Dutton, New York, 1976.

[6] A think tank (also called a policy institute) is an organization, institute, corporation, group, or individual that conducts research and engages in advocacy in areas such as social policy, political strategy, economy, science or technology issues, industrial or business policies, or military advice.

variational analysis, run by the US National Weather Service (NWS) can be given as an example. The GFS ensemble[7] is combined with Canada's Global Environmental Multiscale Model ensemble to form the North American Ensemble Forecast System (NAEFS).

1.2. The Nature of Business Forecasting

According to *Webster's Online Dictionary*, the term **"*forecast*"** was first used in popular English literature sometime before 1321 and its definitions are: "*A prediction about how something (as the weather) will develop*" (noun) and *"Predict in advance"* (verb)[8].



Tufte, Edward R., *The Visual Display of Quantitative Information*. Graphic Press, Cheshire, Connecticut (2001)

Encyclopædia Britannica defines ***economic forecasting*** as the prediction of any of the elements of economic activity. Such forecasts may be made in great detail or may be very general. In any case, they describe the expected future behavior of all or part of the economy and help form the basis of planning[9].

Searching for the meaning of "**What is forecast?**", in most dictionaries we will find similar explanations, for example Microsoft Encarta Dictionary in its last, 2009 edition returns:

- **suggest what will happen:** to predict or work out something that is likely to happen such as the weather conditions for the days ahead;
- **be an early sign of something:** to be an advance indication of something that is likely or certain to happen;
- **prediction of future developments:** an estimation or calculation of what is likely to happen in the future, especially in business or finance.

[7] Ensemble forecasting is a numerical prediction method that is used to attempt to generate a representative sample of the possible future states of a dynamical system.

[8] http://www.websters-online-dictionary.org/definition/forecast

[9] http://www.britannica.com/EBchecked/topic/178385/economic-forecasting

Another, online dictionary (Merriam-webster.com)[10]:

- **to calculate or predict** some future event or condition usually as a result of study and analysis of available pertinent data.

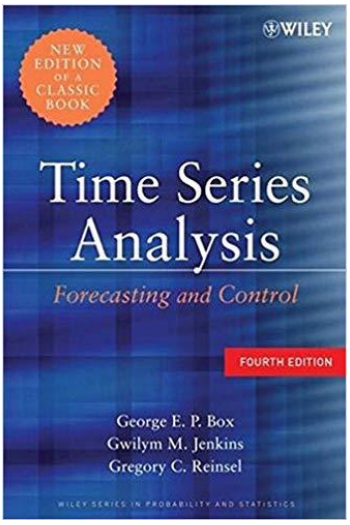
Textbooks also give similar definitions:

- **predicting** important variables for an individual company or perhaps for one component of a company (Hanke & Wichern, 2005, p. 4);
- **a statement** about the **future value** of a variable such as demand (Stevenson, 2009, p. 71) and so on.

We can summarize all of them and make the following definition:

Business forecasting *is the science of developing information about the future through different methods in order to assist in making more effective business decisions*.

Economic forecasting is probably as old as the first organized economic activity, but modern forecasting received its impetus from the Great Depression of the 1930s. The effort to understand and correct the worldwide economic disaster led to the development of a much greater supply of statistics and also of the techniques needed to analyze them. Many of the forecasting methods were developed in the nineteenth and the beginning of the twentieth century, for example the ***regression analysis***, one of the most frequently used forecasting technique.



After World War II, many governments committed themselves to maintaining a high level of employment. Business organizations manifested more concern anticipating the future. Many trade associations now provide forecasts of future trends for their members, and a number of highly successful consulting firms have been formed to provide additional forecasting help for governments and businesses. More forecasting procedures, like ***Box-Jenkins ARIMA***, ***Time-series decomposition*** and others, were developed and with the advent of electronic computers, especially the proliferation of the personal computer (PC) and associated software, forecasting has received more and more attention.

And new techniques for business forecasting continue to be developed (and/or adopted from other scientific areas) as management concern with the forecasting process continues to grow.

[10] http://www.merriam-webster.com/dictionary/forecast

Managers now have the ability to utilize very sophisticated data analysis techniques for forecasting purposes, like ***data mining, machine learning*** and other ***intelligent tools***.

Nowadays there is a wide variety of forecasting techniques that are in use. In many respects, they are quite different from each other, as we shall discover in the next chapters. Nonetheless, certain features are common to all methods, and it is important to know and recognize them. These features can be summarized in the following seven groups:

A. Forecasting techniques generally assume that ***the same underlying causal system that existed in the past will persist into the future*** – a manager cannot simply delegate forecasting to models or computers and then forget about it, because unplanned occurrences can wreak havoc with forecasts. For instance, weather-related events, tax increases or decreases, and changes in prices of competing products or services can have a major impact on demand. Consequently, a manager must be alert to such occurrences and be ready to override forecasts, which assume a stable causal system.

This assumption may be a very strong restriction, especially in long-term forecasting with time series data sets when most elements and relationships between them are dynamic and change, both within the system and in its environment. In Chapters 9 and 12 we will learn how to improve the forecast in such a case using "***dynamic coefficients***".

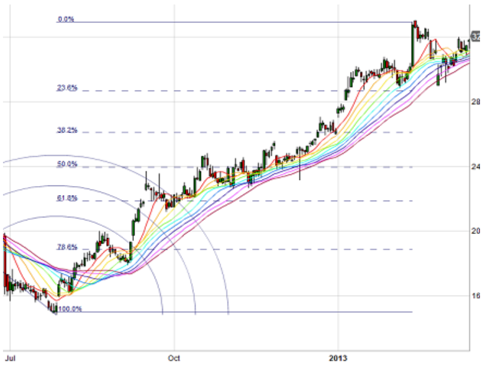
B. Business forecasting holds a very ***strong relationship with the planning function*** in business organizations. Demand planning, for example, also referred to as supply chain forecasting embraces both statistical forecasting and a consensus process. This is one aspect of forecasting that is often ignored. However, if leaders cannot differentiate between a plan and a forecast (as presented in Example 2 below), and the organization doesn't consider forecasting function as a driving force for the highest possible performance, they should not be surprised when one or more of the following occurs:

- They have excess inventory and don't know why;
- Production and logistics are constantly engaged in fire drills to deliver product and thus forced to use emergency/expedited shipping;
- Item-level business planning is nonexistent, inaccurate, or considered an exercise in futility.

Additional problems may also arise like lack of serious effort to improve forecasts; no incentives, metrics, or performance reviews regarding forecasting; forecasts are created, reviewed, and acted upon only at low organizational levels; forecasts are continually changing, adjusted to account for earlier misses; forecasting tools are not either properly understood or not available and so on.

In summary, we should point out that ***forecasting can be described as predicting what the future will look like, whereas planning predicts what the future should look like***.

C. Each forecast must consider ***the Global Optimum of the business organization.*** Forecasts are needed throughout the entire business organization. For example, an inventory system struggles with uncertain demand. The inventory parameters require estimates of the demand and forecast error distributions. The two stages of these systems, forecasting and inventory control, must not be examined independently. Demand forecasting is not an end in itself, and stock control models should not be considered as if there were no preceding stages of computation. It is important to understand the interaction between different departments and functions in business organizations, like demand forecasting and inventory control, and their influence on the performance of the inventory system.



In most similar uses of forecasts, decisions in one area have consequences in other areas. Therefore, it is vital for all affected areas to agree on a ***common forecast***. However, this may not be easy to accomplish. Different departments often have very different perspectives on a forecast, making a ***consensus forecast***[11] difficult to achieve. For example, salespeople, by their very nature, may be overly optimistic with their forecasts and may want to "reserve" capacity for their customers. This can result in excess costs for operations and inventory storage. Conversely, if demand exceeds forecasts, operations and the supply chain may not be able to meet demand, which would mean lost business and dissatisfied customers.

The negative consequences of this problem may consist of the following:

- **Multiple forecasts exist** - forecasts of marketing, logistics, etc. Each one is driving its functional area with its own forecasts, which are not in sync with anyone else;
- In spite of a single department positive results, the overall company performance is unsatisfactory and the organization may have a **big financial loss in the end of the period**;
- Similar efforts are made for forecasting "D" type items (i.e. that make <1% of company sales), as well as for "A" items, which drive more than 80% of the business, and so on.

[11] Consensus forecasts are predictions of the future that are created by combining together several separate forecasts which have often been created using different methodologies and/or forecasting teams.

Some best practices and recommendations of how to deal with problems in topics **B**. and **C**. are given and will be discussed in Chapter 13.

D. Forecasts are made with ***reference to a specific time horizon*** – the time horizon may be fairly short (e.g., an hour, day, week, or month), or somewhat longer (e.g., the next six months, the next year, the next five years, or the life of a product or service). **Short-term forecasts** pertain to ongoing operations. **Long-range forecasts** can be an important strategic planning tool. Long-term forecasts pertain to new products or services, new equipment, new facilities, or something else that will require a somewhat long lead time to develop, construct, or otherwise implement.

Forecast accuracy decreases as the time period covered by the forecast (i.e. the ***time horizon***) increases. Generally speaking, short-range forecasts must contend with fewer uncertainties than longer-range forecasts, so they tend to be more accurate. An important consequence of this point is that flexible business organizations – those that can respond quickly to changes in demand – require a shorter forecasting horizon and, hence, benefit from more accurate short-range forecasts than competitors who are less flexible and who must, therefore, use longer forecast horizons (Stevenson, 2009, p. 73).

How to select an appropriate forecasting method and a model for a given time horizon is discussed in Chapters 2, 3 and 4.

E. Forecasts are not perfect and actual results differ from predicted values – the presence of ***randomness*** precludes a perfect forecast. A particular focus of this is on the errors that are an inherent part of any forecast. Predictions of outcomes are rarely precise and the forecaster can only endeavor to make the inevitable errors as small as possible.

Risk and uncertainty are central to forecasting and prediction, and it is considered a good practice to indicate the degree of uncertainty attached to forecasts. How to measure forecast accuracy and to select the most reliable model is discussed in detail in Chapter 4.

One important moment in this regard is how to reduce the forecasting error. For example, forecasts for groups of items tend to be more accurate than forecasts for individual items because forecasting errors among items in a group usually have a **canceling effect** (i.e. nullification and neutralization). Opportunities for grouping may arise if parts or raw materials are used for multiple products or if a product or service is demanded by a number of independent sources. Another example is applying different transformations (reciprocal, logarithmic and so on) to the initial data set. These and some other techniques used to reduce the forecasting error are presented in Chapter 13.

F. Another important feature is that ***forecasting is* a *never ending process***. Forecasts are needed continually, and as time moves on, the impact of the forecasts on actual performance is measured, original forecasts are updated, decisions are revised, and modified if necessary, and so on (see Fig.1-1). A good example (Lucey, 1991, p. 3) is presented in the Forrester Research paper[12] under Best practice No.1 of Demand Management topic, part "Develop forecasts of weekly demand": "Retailers and their suppliers should ensure that their forecasts are sufficiently frequent to offer maximum intervention opportunities … The switch from monthly to weekly forecasting also provides more opportunities to take corrective action when sales are lagging".

G. Reducing time, efforts and cost of forecasting and transforming it from a complex, difficult to understand process to a smooth, easy to use and understand tool, are the goals of *intelligent techniques* discussed in Chapters 10, 11 and 12.

H. Forecasts are developed ***using a model*** which can be very simple or a very complex one. Usually, a model is defined as a simplified abstract view of the complex reality. A scientific model represents different systems and/or processes in a logical way.

Models form the basis for any decision. They support and assist decision makers in many different ways. Models make it possible to recognize (or identify) the structure and the functions of the systems. This leads to a deeper and better understanding of the problem. In general, models can be analyzed more easily, faster and cheaper than the original problem. They help to find appropriate means for cause-and-effect influence on an object and to predict what the system has to expect in the future. Eventually, models make it possible to run experiments with the system of interest and apply "what-if" analysis.

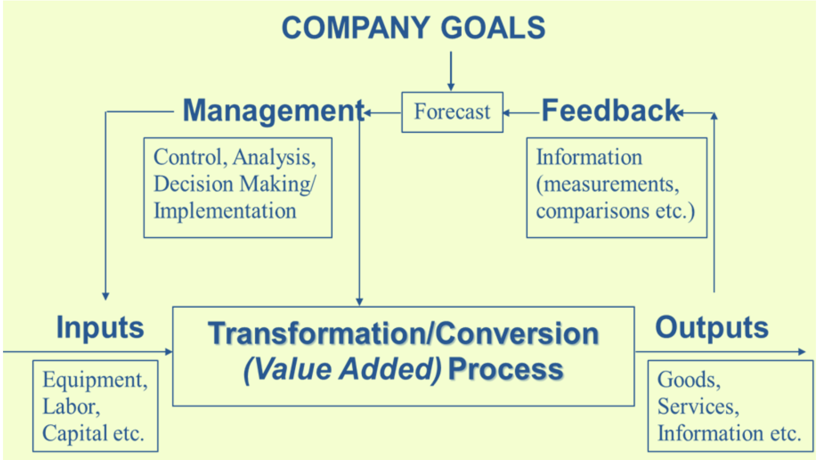

COMPANY GOALS
Management
Forecast
Feedback
Control, Analysis, Decision Making/ Implementation
Information (measurements, comparisons etc.)
Inputs
Transformation/Conversion (Value Added) Process
Outputs
Equipment, Labor, Capital etc.
Goods, Services, Information etc.

Fig.1-1 Forecast as an Element of the Business Process Cycle

[12] Forrester interviewed 20 vendor and user companies including: ConAgra Foods, JDA, Just Group, Oracle, SAP, and Shaw Industries

Of course, the use of models does not guarantee good decisions. Often, nonqualified users cannot comprehend the rules on how to use the model, or may incorrectly apply it and misinterpret the results. How to develop a reliable model, how to select a good one among many similar models, how to apply this model and develop a useful business forecast – all these questions are very important and will be discussed in detail in this textbook.

It is important to clarify two very important aspects of business forecast:

a) It is a statement about the ***future value***, the ***expected level*** of a variable such as demand – the level of demand may be a function of some ***structural variation*** such as long-term trend or a ***causal association*** such as between demand and supply, i.e. it is a statement about what will happen under specific conditions. Defining these conditions is our number one goal in any forecasting task (called ***structural identification*** of the forecasting model as explained in Chapter 2);

b) It represents the real-life business variable with some ***accuracy,*** related to the potential size of forecast error (discussed in Chapter 3). Predictions are rarely perfect and goal number two (known as ***parametric identification*** of the forecasting model) for any forecaster is to make the inevitable errors as small as possible.

There are also other important elements of business forecasting like data analysis, managing the forecasting process, improving the forecast, combining forecasts with other management tools (for example Management Information Systems and Decision Support Systems) and so on. All of them are discussed in this textbook and enough information is given to help students understand their nature, how to use them and when to apply them.

1.3. The Need for Business Forecasting

Organizations operate in an atmosphere of uncertainty and decisions that must be made now affect the future of an organization. Educated guesses about the future are more valuable to organization managers than uneducated guesses as many authors (Hanke et al., 2005), (Stevenson, 2009), (Markidakis, 1986) and others pointed out.

In the past, before the advent of modern forecasting techniques and the power of the electronic computers, the manager’s judgment, based on experience and very often just intuition, was the only tool available in decision making. This situation totally changed in the second part of the last century. Both practitioners and scientists realized that decisions generated using only judgment are not as accurate as those involving the judicious application of quantitative techniques (Markidakis, 1986, p. 17):

"*Humans possess unique knowledge and inside information not available to quantitative methods. Surprisingly, however, empirical studies and laboratory experiments have shown that their forecasts are not more accurate than those of quantitative methods. Humans tend to be optimistic and underestimate future uncertainty. In addition, the cost of forecasting with judgmental methods is often considerably higher than when quantitative methods are used.*"

"That's okay, I don't know what the chart means either."

Nowadays new technologies and new disciplines have sprung up overnight; government activities at all levels have intensified; competition in many areas has become more keen; international trade and multinational companies have stepped up in almost all industries; social help and service agencies have been created and have grown; and the Internet has become one of the most important sources of data and decision-making information. All these factors have combined to create an organizational climate that is more complex, fast-paced, and competitive than ever before. And as the world in which organizations operate has been changing constantly, forecasts have always been necessary. Organizations that cannot react quickly to changing conditions and cannot foresee the future with any significant degree of accuracy are doomed to extinction. The modern tools of forecasting, along with the capabilities of the computer, have become indispensable for organizations operating in the modern world (Hanke et al., 2005, p. 2).

Who and why needs business forecasts?

- *Governments* forecast unemployment, interest rates, and expected revenues from income taxes for policy purposes;
- *Marketing executives* forecast demand, sales, and consumer preferences for strategic planning;
- *College administrators* forecast enrollments to plan for facilities and for faculty recruitment;
- *Wholesalers* forecast demand to control inventory levels, hire employees and so on.

Every organization (large or small, private or public, business or nonprofit) needs and uses forecasting either explicitly or implicitly because it must plan to meet the conditions of the future for which it has imperfect knowledge. Questions like "If we increase our advertising budget by 5%, how will sales be affected?", "What is a year-by-year loan balance of our bank over the next 5 years?”, or “What factors can we identify that will help explain the variability in our monthly unit sales?”, etc. have been asked and will be asked, and all of them require the use of an appropriate forecasting procedure.

Bernstein (1996, pp. 21-22) effectively summarizes the role of forecasting in organizations:

You do not plan to ship goods across the ocean, or to assemble merchandise for sale, or to borrow money without first trying to determine what the future may hold in store. Ensuring that the materials you order are delivered on time, seeing to it that the items you plan to sell are produced on schedule, and getting your sales facilities in place all must be planned before that moment when the customers show up and lay their money on the counter. The successful business executive is a forecaster first; purchasing, producing, marketing, pricing, and organizing all follow.

According to Stevenson (1998), both researchers and business managers agree that forecasts are very important today. Al Enns, Director of Supply Chain Strategies at Motts North America (Stamford, Connecticut) also emphasizes (as cited in Hill, 1998, pp. 70-80), the importance of forecasting – “*I believe that forecasting or demand management may have the potential to add more value to a business than any single activity within the supply chain. I say this because if you can get the forecast right, you have the potential to get everything else in the supply chain right. But if you can’t get the forecast right, then everything else you do essentially will be reactive, as opposed to proactive planning*”.

The need for personnel with forecasting expertise is also growing as pointed out in (Chaman, 1999, p. 2). In a survey, conducted by the Institute of Business Forecasting, at the end of the last century, it was found that there were substantial increases in the staffing of forecasters in full-time positions within American companies.

Understanding of forecasting techniques is essential and it is crucial not only for managers – any decision maker with more or fewer responsibilities in contemporary business organization must be alert to the improper use of forecasting techniques, because inaccurate forecasts can lead to poor decisions. Unfortunately, not everybody understands this even nowadays. Dilgard (2009, p. 4), after more than 15 years of experience in logistics and supply chain as a consultant, project manager, and logistics leader in Fortune 500 companies, found very negative forecasting practices in corporate America:

> *“I have worked in dozens of medium and large companies as a consultant or logistics manager, and not one of them had forecasting/demand planning processes that could be*

considered effective. They did not believe they needed to pay much attention to improving forecasting, though they needed some forecasting capability, which might not be fully leveraged or properly understood. Furthermore, they did not feel that they need a complete overhaul requiring major changes in existing processes. In my experience, many companies have no forecasting capability at all because they feel it's not necessary at their companies. Some of these companies are profitable, well heeled, and stable, while others are in obvious decline, struggling for a way to survive. It does not seem to matter how the company is performing, how old it is, whether it is large or small, technologically savvy or not - none possessed a satisfactory, value-adding forecasting capability and process for continuous improvement. This does not mean, of course, that there are not any companies serious about forecasting, but my observation is that such firms are in the minority. Forecasting problems are myriad, but one consistent theme abounded: executives do not know the difference between a Forecast and a Plan. Hence, organizations are confused as well."

Forecasting is important, but apparently, in spite of this, there are many issues and wrong applications in this area. Some of the most important negative consequences are summarized by Dilgard (2009, pp. 6-7) in the following examples of the worst forecasting practices:

Example 1: Warehouse planner drives forecasts for the entire organization: I was brought into a large multinational instruments company to develop and improve the logistics and procurement capability for its $400 million North American sales and service organization. There were many challenges to be faced, as always, but one that prohibited improving customer service, delivery, warehouse operations, and marketing information was lack of a collaboratively built and widely understood forecast. Per-item demand history and expected future sales were owned and managed by a warehouse planning manager rather than the product managers! Marketing leaders, product managers, and I (the logistics leader for the marketing organization) had no input into demand planning, which drove procurement, shipping, and warehouse storage of expensive instruments and accompanying spare parts. Also, the warehouse planner has a different set of incentives than our marketing organization - his central purpose was to keep inventory down. But with that kind of approach, you can imagine marketing/sales leader complaints: We don't have enough stock to satisfy our customers! Our customers wait too long! We don't have input into expected demand! We can't incorporate market data into the forecast! The warehouse has no incentive to meet customer demand, etc.

Furthermore, marketers did not understand how the warehouse manager calculated future demand. By pulling historical demand from the ERP system and plugging it into a homemade Access database, he estimated future demand by simply extrapolating historical demand with moving averages and made adjustments where necessary using known manufacturing capacity and his own intuition. The organization had not provided better tools or a collaborative process to add market data into the forecast. The warehouse manager wasn't apprised of sales, discounts, large upcoming customer contracts, and so on.

The company faced many other familiar forecasting challenges - islands of analysis, low organizational level of forecasting; no math, statistics, or tools deployed; no marketing input, no concentration of efforts on "A" items; no collaboration; and no metrics or performance review to drive improvement.

Due to budget constraints and lack of top management buy-in to invest in forecasting, my team built a minimal, short-term solution. The objective was, at the very least, to provide a vehicle for integrating marketing input into the forecast. We built an Excel-based tool with minimal mathematical capability (moving averages, seasonal adjustment calculations, and the like) and devoted efforts to items that mattered most - the top 80% of total sales.

Every month, we pushed the tool, populated with demand history and a rough forecast (through the end of the year plus one year), to product managers and asked them to make adjustments based on their market information. We requested that they document market information for future reference and start tracking forecast error. The forecast was then pushed to factories in China; in this way, not only was their manufacturing planning greatly improved, but also rancor associated with lack of forecast collaboration was reduced.

Interestingly, once product managers were involved in the process, our team had to prod them to provide input. They were concerned that their input would actually be used to alter their warehouse buying plan or the plant production plan. They did not want to see their efforts go to waste. Even after proving the actual benefit of their input, obtaining product manager input was still sometimes difficult because providing a good forecast was not part of their job description or incentive plan. They also did not understand how a better forecast could translate into more available inventory for their customers. Both were huge obstacles to forecast improvement, a problem I witnessed at many organizations.

Example 2: Plan vs. Forecast: Parts requirement from Bill of Materials (BOM) blowout considered a "**forecast**". As the divisional logistics executive at a Fortune 500 diversified manufacturer, I was charged with improving inventory and materials management at five distinct and separate companies. Each had different systems, cultures of people, products, processes, and locations. But one thing was consistent: There was a complete lack of forecasting at virtually all levels.

One company in my division was engaged in the "fire drill" method of pleasing the customer. Rather than analyzing historical data and deploying forecasting techniques for new products or large purchase customers, it decided that its products could not be forecasted. Leaning heavily on suppliers, managers, and employees, it scrambled to meet customer demand as it arrived, and lived every day in an emergency mode.

When I began inquiring about planning and forecasting processes, I quickly learned why there was no forecasting capability. I met with hostile resistance from the division's CFO immediately and often. The CFO, a competent accountant also charged with managing IT, had been helpful in other areas. I explained how improved forecasting could help meet customer demand and reduce excess and obsolete inventory, which was a significant issue. However, the CFO responded in this way:

How can you say we don't forecast?! We already have a great forecasting capability in our ERP (Enterprise Resource Planning) system! We simply plug in the product we need to build, and it blows out each item we need to buy or build, and even lead times.

When the VP of Procurement pointed out that the CFO had identified a plan, not a forecast, there was a sense of sinking feeling in the room - this was an educational challenge of great proportion. The CFO and most other members of the leadership team did not understand the difference between a forecast and a plan.

The presidents of three other companies in the division stated that their company's products simply could not be forecasted or planned for and that trying to forecast or plan was an exercise in futility. The companies, in their mind, were limited to only responding to customer orders. To be sure, there were many unpredictable customer orders, but to completely ignore analyzing historical demand was startling and self-defeating.

Example 3: "Forecast" is a sales target, developed by top executives and pushed down to managers to make work. At one company, a very famous and large fashion retailer, forecasting had two major difficulties:

- The "forecast" was actually an aggregate-level sales revenue target, created by management to drive sales growth across the company.

- The actual forecasting function was done at a micro level by home office analysts, simply for replenishing store inventory at a store level. No forecast existed except a simple moving average with minor qualitative intervention.

The aggregate sales target sounded like this: Our sales target for 2009 is $2 billion. Every department and all sales divisions need to figure out how they can do their part to achieve this figure.

Sound familiar to your budgeting process? All organizational functions were doing their own forecasting to build a plan to meet the top management's sales target. The problem with that was middle managers lacked item-level historical demand or a collaborative forecast to build a case on how they could meet their "forecast." A budgetary sales target without serious forecasting credentials for support generally results in missed and poorly understood results and mysteriously built inventory. Companies are then stunned that sales targets are not met, targets are not realistic, or inventory is high. Often, forecasts are continually "adjusted" throughout the year, and there is little or no information available as to what has been changed.

Most companies have only anecdotes, stories, and assumptions to rely upon. Wins and misses are found too late, and adjustments are made too slowly to seriously redeploy or reduce inventory investment. Heads sometimes will roll, but mostly, companies just live with mistakes, pain (inventory), or missed opportunities. A robust forecasting function will quickly calculate hits and misses and have a ready-made plan to address them.

We can resume the above experiences simply by saying that there is a deep, ingrained misunderstanding at even the most modern U.S. companies about what a forecast is and how it should be used in business planning and management. Great difficulty lies in companies already feeling they forecast very well, or their products cannot be forecasted or planned, and forecasting is not a function where money should be thrown. Forecasting is deemed impossible without significant knowledge of forecasting, which many leaders believe nobody has.

One very important reason for all this to happen is because many college graduates, including those with degrees in business, do not ever study forecasting, except as a sidelight in a course that has other primary objectives, such as Business Statistics and Operations Management. Even more, most of the existing books in Business Forecasting, like (Hanke & Wichern, 2005), (Wilson & Keating, 2002) and others, present only the basic techniques and many contemporary approaches, like ***business intelligence*** and ***predictive analytics***, ***knowledge discovery from data*** and ***data mining*** (incl. *artificial neural networks, genetic algorithms, self-organization* and other intelligent tools), which are very useful today for preparing business forecasts, are only briefly mentioned.

1.4. Characteristics of Good Forecasts

In the beginning of this chapter, we defined ***Business forecasting*** as the science of developing ***information*** about the future through different methods to assist in making more effective business decisions. In fact, what managers do need is good information, i.e. information which has been used and created value. Consequently, the forecasting information is valuable to a business only when it leads to actions which create value or market behavior that gives a competitive advantage.

This general statement needs clarification in terms of some qualities, which can be used to determine if particular forecasting information is good or not. There are some classifications in the theory (Stevenson, 2009), (Lucey, 1991) and others, which list numerous characteristics.

First of all, managers need ***relevant information*** to assist them in planning, controlling, and decision making. According to Lucey (1991, p. 12) relevant information is information which:

a) Increases knowledge;

b) Reduces uncertainty;

c) Is usable for the intended purpose.

In effect, this is the overriding quality. Information must be relevant to the problem being considered (i.e. ***to be meaningful)***. Too often reports, messages and projections contain irrelevant parts which make understanding more difficult and cause frustration to the user. Relevance is affected by many other qualities, as described below and in Chapters 11 and 12 we are going to discuss different techniques, used to extract knowledge from data.

Reducing uncertainty is very closely related to the next important characteristic of good information & forecast – to be sufficiently ***accurate*** for its purpose, i.e. to be relied upon by the manager and for the purpose for which it is intended.

There is no such thing as absolute accuracy and raising the level of accuracy increases cost but does not necessarily increase the value of information. It is important that the ***degree of accuracy*** should be clearly stated. This will enable users to plan for possible errors and will provide a basis for comparing alternative forecasts. The most useful criteria and measures of forecast accuracy are discussed in detail in Chapter 3.

The ***level of accuracy*** must be related to the decision level involved and expressed implicitly by means of significant figures. At operational levels, information may need to be accurate to the nearest penny, $, kilogram or minute. A sales invoice, for example, will be accurate to the penny. On the other hand, a Sales manager at the tactical level will probably be best suited to information rounded to the nearest $100, whilst at the strategic level rounding to the nearest ten thousand dollars or higher is common.

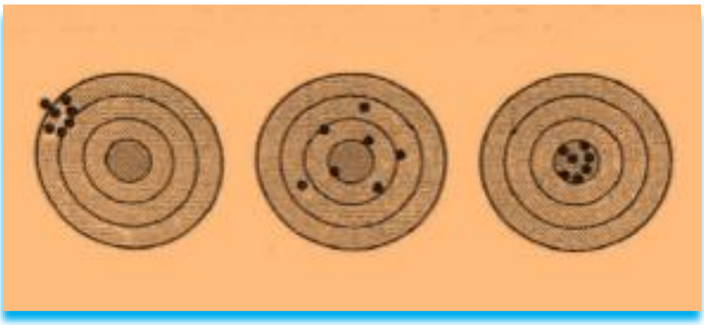


Fig.1-2 Distinction between Accuracy and Precision (according to Lucey (1991, p. 20)

It means also, that the forecast should be expressed in ***meaningful units***. Financial planners should know how many dollars will be needed, production planners should know how many units will be needed, and schedulers should know what machines and skills will be required. Eventually, the choice of units depends on user needs.

One important point is that accuracy should not be confused with ***precision***. Information may be inaccurate but precise or vice versa (see Fig. 1-2). The analogy used here to explain the difference between accuracy and precision is the target comparison. In this analogy, repeated measurements are compared to arrows that are shot at a target. Accuracy describes the closeness of arrows to the bull’s-eye at the target center. Arrows that strike closer to the bullseye are considered more accurate. The closer a system's measurements to the accepted value, the more accurate the system is considered to be.

To continue the analogy, if a large number of arrows are shot, precision would be the size of the arrow cluster. (When only one arrow is shot, precision is the size of the cluster one would expect if this were repeated many times under the same conditions.) When all arrows are grouped tightly together, the cluster is considered precise since they all struck close to the same spot, even if not necessarily near the bullseye. The measurements are precise, though not necessarily accurate.

However, it is *not* possible to reliably achieve accuracy in individual measurements without precision—if the arrows are not grouped close to one another, they cannot all be close to the bull’s-eye. Their *average* position might be an accurate estimation of the bull’s-eye, but the individual arrows are inaccurate.

In the fields of engineering and statistics (Dodge, 2003) the ***accuracy*** of a measurement system is the degree of closeness of measurements of a quantity to its actual (true) value. The ***precision*** of a measurement system, also called reproducibility or repeatability, is the degree to which repeated measurements under unchanged conditions show the same results. How to measure the ***Forecast Accuracy*** will be discussed, as we mentioned above, in Chapter 3.

The next characteristic of good information & forecast is to be from a source in which the user has confidence. To use the results from a forecast decision makers must have confidence in its source. ***Reliable confidence*** usually is available when:

a) the source has been reliable in the past;

b) there is good communication between the forecast producer and the user.

For example, confidence exists when the decision maker has been consulted over the content, format, and timing of the forecast and there is frank discussion over possible uncertainties and inaccuracies. Especially at strategic levels, management will cross check information from various sources to increase confidence in the forecast provided.

Another, related to previous one, characteristic is that the forecast should be ***reliable***, i.e. it should work consistently. A technique that sometimes provides a good forecast and sometimes a poor one will leave users with the uneasy feeling that they may get burned every time a new forecast is issued (Stevenson, 2009, p. 74).

Good forecast should also be ***complete enough*** for the problem and should contain the ***right level of detail***. Ideally, all the information required for a decision should be available. In the real world, of course, this never happens. What is required is that the information is complete in respect of the key elements of the problem. This means that there must be close liaison between information providers (forecasters) and users to ensure that the key factors are identified. For example, a supermarket chain in making a strategic decision whether or not to place a new superstore on the outskirts of a town would identify such things as population density, road access, presence of competitors, and so on as key factors in the decision and would not try to include every detail about the town in their initial analysis.

At the same time, the forecast information should contain the least amount of detail consistent with effective decision making. Every superfluous character means extra storage, more processing, extra assimilation and possibly poorer decisions. The level of detail should vary with the level in the organization – the higher the level the greater the degree of compression and summarization, although the information, particularly at lower levels, often has to be very detailed too, but the general rule of "**as little as possible**" consistent with effective use, must always apply.

Timing is the next very important characteristic. Good forecast information is that which is available in time to be used. It means, it should be communicated in time for its purpose. To an extent, the need for speed can conflict with the need for accuracy although modern forecasting methods (discussed in Chapters 10, 11 & 12) can produce accurate information very rapidly. Delays in data gathering, preprocessing, model development or communication (delivery) of the results can transform potentially vital information into worthless pieces of paper.

Forecasts should be produced at a frequency which is related to the type of decision or activity involved. Very often reports are produced routinely at quite arbitrary intervals - daily, weekly, monthly and so on - without regard to the time cycle of the activity involved. At operational levels, this may mean a requirement for information to be available virtually continuously (say on a PC screen), but at other levels much longer intervals are likely to be appropriate which should not be determined merely by the conventions of the calendar.

Timely forecast means also that forecasting process length should be determined precisely. Usually, a certain amount of time is needed to respond to the information contained in a forecast. For example, capacity cannot be expanded overnight, nor can inventory levels be changed immediately. Hence, the forecasting horizon must also cover the time necessary to implement possible changes.

A good forecast is also that, which is ***understandable*** by the user. Understandability is what transforms data into information. If the information is not understood it cannot be used and thus cannot add value. Many factors affect understandability including:

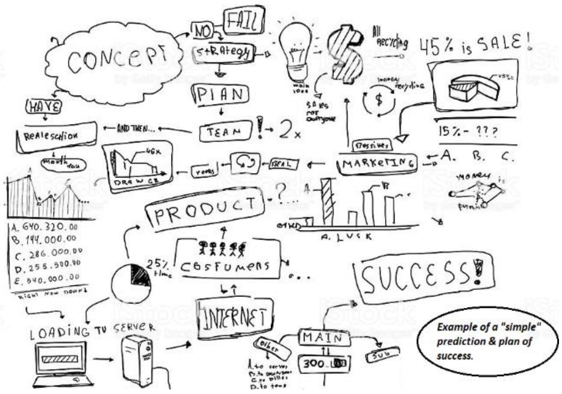
a) *Preferences of the user* – some people prefer information in the form of pictures and graphs, others prefer narrative. Some are happy with statistical and numeric presentations whilst others do not understand them. Experiments show that some people absorb concrete facts in detail whilst others evaluate situations as a whole with little regard for factual detail. This variability means that the same forecast will inevitably receive many interpretations.

b) *Remembered knowledge* – although the working of memory is not well understood there is no doubt that the extent of remembered knowledge, including technical (IT) knowledge, influences understanding. Understanding is thus a result of the association of memory and the received information.

c) *Environmental factors* – as well as the individual characteristics mentioned above a number of environmental factors influence understanding. These include group pressures, time available, trust in the information system (software) used and so on.

d) *Language* – information is conveyed by means of signals or messages. These may be in a code (for example, a mathematical equation) or in a natural language, such as English or Spanish. Natural languages are very rich in the range of information they can accommodate but are inherently ambiguous. Mathematical notations or programming languages can be very precise but lack the capacity to cope with a wide range of concepts. (Language and perception are of great importance to information specialists and this point is developed further below).

e) Last, but not least factor is that the forecasting technique should be *simple* in terms to understand and use. Users often lack confidence in forecasts based on sophisticated techniques; they do not understand either the circumstances in which the techniques are appropriate or the limitations of the techniques. Misuse of techniques is an obvious consequence. Not surprisingly, fairly simple forecasting techniques enjoy widespread popularity because users are more comfortable working with them.



Example of a "simple" prediction & plan of success.

In light of the relative complexity of some inclusive but sophisticated forecasting techniques, it could be recommended that users go through an evolutionary progression in adopting new forecast techniques. That is to say, a simple forecast method well understood is better implemented than one with all-inclusive features but unclear in certain facets.

One group of requirements is related to the process of *communications*. A good forecast is one which is communicated by an appropriate ***channel of communication*** and to the ***right person***. To be usable by the manager, information must be transmitted by means of a communication process. Communication involves the interchange of facts, thoughts, value judgments and opinions and the communication process may take many forms; face-to-face conversations, telephone calls, informal and formal meetings, conferences, memoranda, letters, reports, tabulations, wireless transmissions and so on. Whatever the process, good communication occurs when the sender and receiver are in accord with the meaning of a particular message.

The channel of communication should be selected with regard to such things as the nature and purpose of the information, the speed required and, above all, the requirements of the user. The typical output of formal forecasting is a printed report. It has its uses, of course, but there is research evidence that many managers, especially at senior levels, obtain most of their information orally. They use written reports merely to confirm or reinforce information they already have. In this connection, wireless communications could improve a lot of the process. More about channels of communication could be found in (Lucey, 1991), (Laudon & Laudon, 2010) and others.

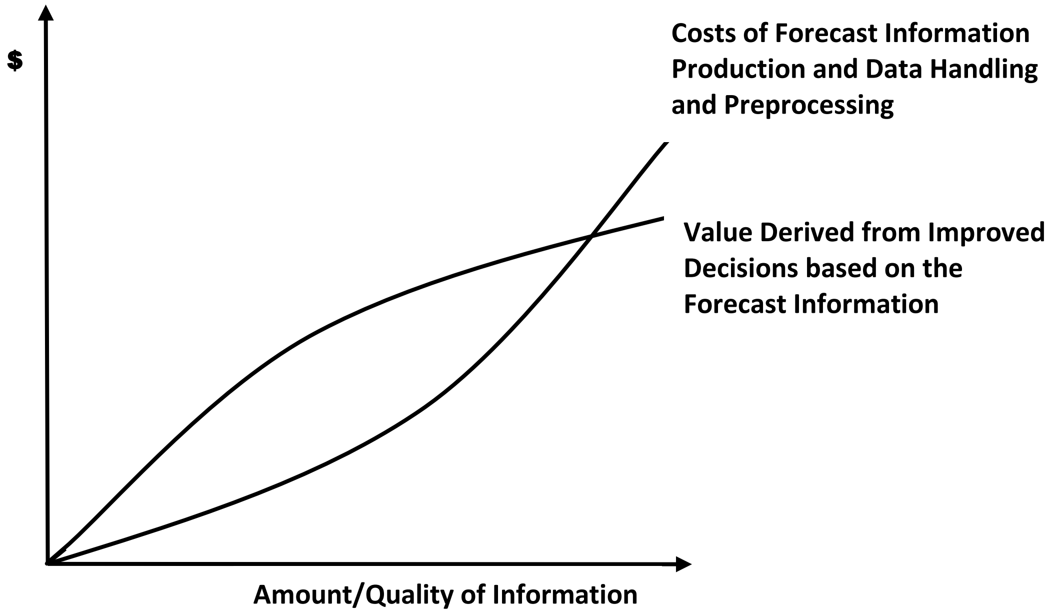


Fig.1-3 Forecast Information – Cost and Value (adopted from Lucey 1991, p. 18)

The second point here is that to be used effectively the forecast should be delivered to the ***right person***. Each manager has a defined sphere of activity and responsibility and should receive forecast information to help him carry out his designated tasks. It sounds curious, but in practice, this is not always as easy as it sounds. It is quite common for information to be supplied to the wrong level in the organization. A superior may not pass it on to the person who needs it whilst a subordinate may hold on to information in an attempt to make himself seem indispensable. Information systems designers need to analyze the key decision points in an organization in order to direct forecast information exactly where it is required.

The last group of requirements may contradict sometimes with the rest, discussed above. From a business standpoint the good forecast should be ***cost-effective***, i.e. the benefits should outweigh the costs (see Fig. 1-3). As we mentioned already, forecasting information is valuable to a business only when it leads to actions which create *value* or market behavior that gives a *competitive advantage*. Nevertheless, we need a little more clarifications about a few related terms.

Effectiveness means the capability of producing an effect. In management, effectiveness relates to *getting the right things done*. The term *effective* is sometimes used in a quantitative way, "being very or not much effective". However, it does not inform on the direction (positive or negative) and the comparison to a standard of the given effect. ***Efficacy***, on the other hand, is the ability to produce a desired amount of the desired effect, or success in achieving a given goal. Contrary to efficiency, the focus of efficacy is the achievement as such, not the resources spent in achieving the desired effect. Therefore, what is effective is not necessarily efficacious, and what is efficacious is not necessarily efficient. An ordinary way to distinguish among effectiveness, efficacy, and efficiency:

- ***Efficacy*** is getting things done, i.e. meeting targets;
- ***Effectiveness*** is doing "right" things, i.e. setting right targets to achieve an overall goal (the *effect*);
- ***Efficiency*** is doing things in the most economical way (good input to output ratio).

Economic efficiency is used to refer to a number of related concepts. It is the using of resources (inputs) in such a way as to maximize the production of goods and services (outputs) (Sullivan & Sheffrin, 2003, p. 15). One economic system is more efficient than another (in relative terms) if it can provide more goods and services for society without using more resources. In absolute terms, a system (incl. forecasting system) can be called ***economically efficient*** if[13]:

- Nothing can be made better off without making something else worse off.
- More output cannot be obtained without increasing the number of inputs.
- Production proceeds at the lowest possible per-unit cost.

It will be seen from the above that many, many things need to be right before particular forecast information can be considered as good. As we can see many of the factors relate to social and behavioral characteristics, proving again that that forecasting is both a science and an art. In Chapter 2 we'll discuss the forecasting process along with the effective model building procedure.

1.5. Understanding and Managing Business Forecasting

There are many approaches at present, which can be used to implement "**To predict the future**" concept. In general, they can be summarized in three major groups:

- **ANTICIPATION** – a logical model of the future with an uncertain level of reliability, like: *"If it rains outside, take the umbrella. Otherwise, leave the umbrella home"*. Nowadays, anticipation is an important part of the so called artificial intelligence (AI), where it is the concept of an agent making decisions based on predictions, expectations, or beliefs about the future[14];
- **PREDICTION** – nonprobabilistic approach (it is expected to happen for sure) for example: *"The sun will rise again tomorrow"*. It is a statement or claim that a particular event will occur in the future;

[13] These definitions of absolute efficiency are not equivalent, but they are all encompassed by the idea that nothing more can be achieved given the resources available.

[14] These methods are discussed in Chapter 12.

- **FORECAST** – information based on the theory of probability. Forecasting is the process of making statements about events whose actual outcomes (typically) have not yet been observed. A commonplace example might be an estimation of the expected value for some variable of interest (customer demand, or resource availability) at some specified future date.

Both ***prediction*** and ***forecast*** typically refer to formal statistical methods employing time series, cross-sectional or longitudinal data, or alternatively to less formal judgmental methods. The usage can differ between areas of application: for example in hydrology, the terms "forecast" and "forecasting" are sometimes reserved for estimates of values at certain specific future times, while the term "prediction" is used for more general estimates, such as the number of times floods will occur over a long period.

In a scientific context, a prediction is a rigorous, (usually quantitative), statement forecasting what will happen under specific conditions. The scientific method is built on testing assertions that are logical consequences of scientific theories. This is done through repeatable experiments or observational studies.

Since the terms ***prediction*** and ***forecast*** are often used as synonyms in real life business (as far as they are used to provide future information for decisions making), the same view would be applied in this textbook, and **forecasting** would be considered as the particular approach which implements the "***To predict the future***" concept in contemporary business organizations.

Current trends emphasize the increasing need for management to deal with complex issues and, in particular, the need to develop sophisticated methods for dealing with future uncertainties. They emphasize the growing importance of combining good judgment and sophisticated data manipulation methods into sound business forecasting. As these trends and the increasingly dynamic business environment continue to unfold, the ability of business leaders to react quickly and profitably to changing events is brought into sharper focus. As mentioned by (Hanke and Wichern, 2005, p. 513) "The basic business question "What will happen next?" will assume even greater importance".

There are many factors that affect the whole forecasting process and several key questions should always be raised if the forecasting process is to be properly managed:

- Why is a forecast needed?
- Who will use the forecast, and what are their specific requirements?
- What level of detail or aggregation is required, and what is the proper time horizon?
- What data are available, and will the data be sufficient to generate the needed forecast?

- How much will the forecast cost?
- How accurate can we accept the forecast to be?
- Will the forecast be made in time to help the decision-making process?
- Does the forecaster clearly understand how the forecast will be used in the organization?
- Is a feedback available to evaluate the forecast after it is made and adjust the forecasting process accordingly?

Understanding ***Business Forecasting*** is essential today and it is crucial for any decision maker, with more or fewer responsibilities in contemporary business organization, who must be alert to the improper use of forecasting techniques, because inaccurate forecasts can lead to poor decisions. A good forecast requires both, the knowledge about the forecasting techniques available and professional competences in the area of interest. The former is a prerequisite for forecast development and the latter is necessary for its appropriate use and implementation.

The effectiveness of ***Business forecasting*** depends also on management attitude and understanding. Long time ago, (Makridakis, 1986, p. 33) noted:

"The usefulness and utility of forecasting can be improved if management adopts a more realistic attitude. Forecasting should not be viewed as a substitute for prophecy but rather as the best way of identifying and extrapolating established patterns or relationships in order to forecast. If such an attitude is accepted, forecasting errors must be considered inevitable and the circumstances that cause them investigated."

An important point that should be mentioned here is that the management information systems·(MIS) of modern firms have increased in sophistication and usefulness in recent years. Their first benefit to the forecasting process involves their enormous capability to collect and record data throughout the organization.

The second benefit is that the availability of inexpensive personal computers and forecasting software has tended to move the forecasting function downward in the organization. It is now possible for managers to have access to sophisticated forecasting tools at a fraction of the cost of such capability just a few years ago. However, the knowledge required to properly use this capability does not come with the hardware or software package; the need to understand the proper use of forecasting techniques has increased as the computing capability has moved out of the hands of the "experts" into those of the users in an organization.

This text has been designed to provide students with enough understanding of fundamentals and the major details of all important techniques, which are necessary to develop an appropriate forecasting model, to select a “good” model among many others and to apply it when solving real-life business problems. The focus of the book is in incorporating the latest findings from both theory and practical research. It not only presents general principles and fundamentals that underlie forecasting practice, but also introduces both standard and advanced approaches to forecasting with a main emphasis on data mining and predictive analytics.

It is worth noting, that today many people have no doubts about the needs for forecasting and probably they would agree that forecasting is both a science and an art. It is an art because one can never be sure what the future holds. At the same time, it is also a science because one can extrapolate from historical data, so it's not a total guess. In the following chapters we will discuss the full range of forecasting techniques in business, from simple ones (like naïve methods, smoothing and extrapolation – ch.ch. 4 and 5) to the most sophisticated and complicated methods (incl. artificial neural networks, self-organizing data mining and others – ch.ch. 8, 9, 10, 11 and 12).

***

Summary and Conclusions

- Forecasting is used from ancient times in different ways, first in local and after that in global perspective. Nowdays, the Club of Rome reports and other similar projects act as a global catalyst for change through the identification and analysis of the crucial problems facing humanity. The importance of such ***Global Forecasts*** increases every year.
- In First Chapter we introduced ***"To predict the future"*** concept and the main approaches at present, which can be used to implement it – *anticipations, predictions,* and *forecasts*.
- As far as both ***prediction*** and ***forecast*** are used to provide future information for decisions making we are going to use these terms as synonyms in this textbook.
- We also accepted the most common definition that ***business forecasting is the science of developing information about the future through different methods to assist in making more effective business decisions***.
- The ***forecasting science*** is a particular approach, which implements "**To predict the future**" concept in contemporary business organizations. Its importance has been recognized within a wide range of individual to global, world-wide scale companies.
- ***Economic and Business forecasting*** are the predictions of any of the elements of economic activity. Such forecasts may be made in great detail or may be very general. They describe the expected future behavior of all or part of the economy and help form the basis of planning.
- ***Business forecasting*** is necessary because all organizations operate in an atmosphere of uncertainty and decisions, which affect the future of the organization, must be made today. Educated guesses about the future are more valuable to organization managers than uneducated guesses and today every organization uses forecasting either explicitly or implicitly because it must plan to meet the conditions of the future, for which it has imperfect knowledge.
- Large part of the chapter discusses in detail two important points, the features common to all forecasts and the characteristics of a "Good Forecast". Features like ***randomness, time horizon, the global optimum, model base*** and others were described and analyzed.
- Together with these features, the most important characteristics, like ***relevance, accuracy, timeliness, confidence, understandability,*** and ***cost-effectiveness,*** which particular forecast information must have, in order to be considered as good one, were also described.

- References and links to other chapters in the textbook, where these points or their elements will be discussed, were pointed out as well.
- ***Understanding of forecasting techniques is essential and it is crucial*** not only for managers – any decision maker with more or fewer responsibilities in contemporary business organization must be alert to the improper use of forecasting techniques because inaccurate forecasts can lead to poor decisions.
- Not all managers/decision makers understand the above, as the negative examples reveal this in the chapter.
- One important reason for this problem is that many college graduates, including those with degrees in business, do not ever study forecasting.
- Most of the existing books in Business Forecasting present only the basic techniques and many contemporary approaches are not discussed at all, or just briefly introduced.
- ***The primary goal of this textbook*** is to provide students with enough understanding of fundamentals and the major details of all important techniques which are necessary to develop an appropriate forecasting model, to select a "good" model among many others and to apply it when solving real-life business problems.

KEY TERMS

Accuracy	***17***	***Accuracy vs. Precision***	***18***
Anticipation	***23***	***Box-Jenkins ARIMA***	***5***
Business forecasting	***5, 25***	***Business intelligence***	***16***
Canceling effect	***8***	***Causal association***	***10***
Consensus (common) forecast	***7***	***Cost-effective***	***22***
Data mining	***6, 16***	***Economic efficiency***	***23***
Economic Forecasting	***4***	***Effectiveness***	***22, 23***
Efficacy	***22, 23***	***Expected level***	***10***
Forecast	***4, 24***	***Forecasting***	***23***
Information (characteristics)	***17***	***Intelligent tools***	***6***
Long-range forecasts	***8***	***Machine learning***	***6***
Model	***9***	***Parametric identification***	***10***
Plan vs. Forecast	***14***	***Prediction***	***23***
Predictive analytics	***16***	***Randomness***	***8***
Regression analysis	***5***	***Short-term forecasts***	***8***
Structural identification	***10***	***Structural variation***	***10***
Time horizon	***8***	***Time-series decomposition***	***5***

CHAPTER EXERCISES

Conceptual Questions:

1. Why is forecasting important for businesses? Discuss.
2. What is the definition of business forecasting? Discuss and illustrate with examples.
3. What is the difference between anticipation, prediction, and forecast? Discuss and illustrate with examples.
4. What contribution did the Club of Rome make in global forecasts? Discuss.
5. How has forecasting developed since the Great Depression? Discuss and illustrate with examples.
6. What are the two main aspects of a business forecast? Discuss.
7. What are the seven features common to all forecasts? Discuss and illustrate with examples.
8. What does economic efficiency refer to? Discuss.
9. What are the characteristics of a good forecast? Discuss and illustrate with examples.

Business Applications:

Problem 1. Amazon.com has become one of the most successful online merchants. Sales are one of the variables that measure its success. An article entitled „*Amazon CEO takes long view*" (USA Today, Byron Acohido, July 6, 2005) presented the following sales figures (in $ milion) for the period 1995 to 2004:

Year	1995	1996	1997	1998	1999
Sales	0.5	15.7	147.7	609.8	1,639.8
Year	2000	2001	2002	2003	2004
Sales	2,761.9	3,122.9	3,932.9	5,263.7	6,921.1

a) Graph these data and indicate whether they appear to have a trend. Discuss.

b) Produce a time-series plot for these data. Identify the specific pattern and the potential equation that should be used to obtain the following years' forecasts.

c) Compute the main descriptive statistics about Amazon.com sales. Discuss the findings.

Problem 2. How can a real estate firm determine the selling price for a house? This is what the managers want to be able to predict, i.e. the sales price is the dependent variable in their study. The list of potential independent (explanatory) variables, or factors that may cause changes in the selling price, usually consist of:

- House size in square feet.
- Age of the house in years.
- Number of bedrooms.
- Number of bathrooms.
- Number of other rooms.
- House location.
- Garage size (number of cars).
- Condition of the house and other less important variables.

File Houses.xmls contains data for a sample of 240 residential properties:

- Produce a scatter plot for each pair of those factors and the dependent variable. Specify the pattern of the relationship (if any) for each pair. Discuss.
- Develop the correlation matrix for this set of data. Select and exclude the independent variables whose correlation magnitude with the dependent variable is the weakest, i.e. <0.4.
- Determine if the association with the dependent variable is significant (use a significance level of 0.05) for each factor showing moderate correlation ($0.4< r <0.7$). Discuss the findings.

INTEGRATIVE CASE

HEALTHY FOOD SUPPLY CHAIN & STORES

Part 1: An Introduction – First Steps in Sales Forecasting

Healthy Food Stores is a fast-growing retail food provider with 12 stores in a northwestern state. To achieve faster growth the company has been engaged in various kinds of advertising. Sales and Marketing Department decides to study the effect the company advertising dollars have on sales and some monthly data has been collected for the past 4 years (file Sales.xlsx).

Company historical records contain the sales volume for each month along with the advertising dollars for the traditional and online ads alike. It's important to note that the company information system (IS) provides opportunities for deriving any secondary data from these records. Since the top executives believe that sales might depend on advertising expenditures in previous months rather than in the month the ads appeared both sales and advertising values have lagged one and/or two months.

The executive board members at *Healthy Food Stores* decide to evaluate their advertising efforts along with some other factors that may cause an increase in the sales for each month. Their goal is to examine the collected data to possibly reveal important relationships, which will help determine future advertising expenditures to achieve a fast growing and well-balanced sales. It is decided to accept a forecast only if its error is less than 5%.

Case Questions

1. What are the first steps in a company data analysis to prepare useful sales forecasts? Discuss.
2. Create a list of the potential factors which could be derived from data already collected.
3. Open file Sales.xlsx in MS Excel and perform the following:
    - Graph all sales data and indicate whether they appear to have a trend.
    - Produce a time-series plot for these data. Is there any specific pattern?
    - Compute the main descriptive statistics about company sales. Discuss the findings.
4. Write a short report (about two pages not counting charts and tables) on the questions above, discussing all important findings and draw relevant conclusions about this part of the Integrative Case.

References

Bernstein, P. L. (1996). *Against the Gods: The Remarkable Story of Risk.* New York, NY: John Wiley & Sons.

Chaman, L. J. (1999). Explosion in the Forecasting Function in Corporate America. *Journal of Business Forecasting*, Summer, 2.

Dilgard, L. (2009). Worst Forecasting Practices in Corporate America and Their Solutions. *Journal of Business Forecasting,* Flushing: Summer, *28*(2), 4-11.

Dodge, Y. (2003). *The Oxford Dictionary of Statistical Terms*, OUP.

Hanke, J., & Wichern, D. (2008). *Business Forecasting.* PEARSON & Prentice Hall.

Hill, Jr., S. (1998). A Whole New Outlook. *Manufacturing Systems*, *16*(9), 70-80.

Laudon, K., & Laudon, J. (2017). *Management Information Systems: Managing the Digital Firm.* Pearson Prentice Hall.

Lucey, T. (1991). *Management Information Systems.* DP Publications Lim.

Makridakis. S. (1986). The Art and Science of Forecasting. *International Journal of Forecasting*, *2*, 15-39.

Stevenson, H. (Ed.) (1998). DO LUNCH OR BE LUNCH. *Boston: Harvard Business School Press*.

Stevenson, W. J. (2017). *Operations Management*. McGraw-Hill/Irwin.

Sullivan, A., & Sheffrin, S. (2003). *Economics: Principles in action.* Pearson Prentice Hall.

Turner, G. (2008). A Comparison of The Limits to Growth with Thirty Years of Reality. Socio-Economics and the Environment in Discussion (SEED). *CSIRO Working Paper Series,* ISSN 1834-5638.

Wilson, J., & Keating, B. (2008). *Business Forecasting.* McGraw-Hill.

CHAPTER 2. THE FORECASTING PROCESS

2.1. Business Forecasting and Decision Making

Identifying business needs and determining solutions to business problems is a never-ending cycle in real life business. There is a common logical sequence in any ***problem solving*** and according to Pólya (1945, p. 5) there are four phases of the work:

- ***Understanding the problem*** - we must see clearly what is required: What is the unknown? What are the data? What is the condition?
- Devising a plan - find the connection between the data and the unknown: Have you seen the problem before? Do you know a related problem? Look at the unknown and try to think of a familiar problem having the same or a similar unknown! Here is a problem related to yours and solved before, could you see it? Eventually, decision makers should obtain a plan of the solution.
- **Carrying out the plan** - Carry out your plan of the solution, check each step. Can you see clearly that the step is correct? Can you prove that it is correct?
- **Looking back -** Examine the solution obtained: Can you check the result? Can you derive the result differently? Can you use the result (or method) for some other problem?

Problem solving and ***decision making*** are very close – as a matter of fact decision making is a part of every particular problem-solving approach (Anderson et. al., 1995). It might be regarded as a problem-solving activity which is terminated when a satisfactory solution is found (see Fig.2-1).

Human performance in decision-making terms has been the subject of active research from several perspectives. From a *psychological* perspective, it is necessary to examine individual decisions in the context of a set of needs, preferences an individual has and values they seek. From a *cognitive* perspective, the decision-making process must be regarded as a continuous process integrated into the interaction with the environment. From a *normative* perspective, the analysis of individual decisions is concerned with the logic of decision making and rationality and the invariant choice it leads to (Kahneman & Tversky, 2000).

Most of the ***decision theory*** is normative (or prescriptive), i.e., it is concerned with identifying the best decision to take, *assuming an ideal decision maker who is fully informed, able to compute with perfect accuracy, and fully rational.* The practical application of this prescriptive approach (how people *actually* make decisions) is referred to as ***decision analysis*** (Howard, 1966), and aimed at finding tools, methodologies and software to help people make better decisions.

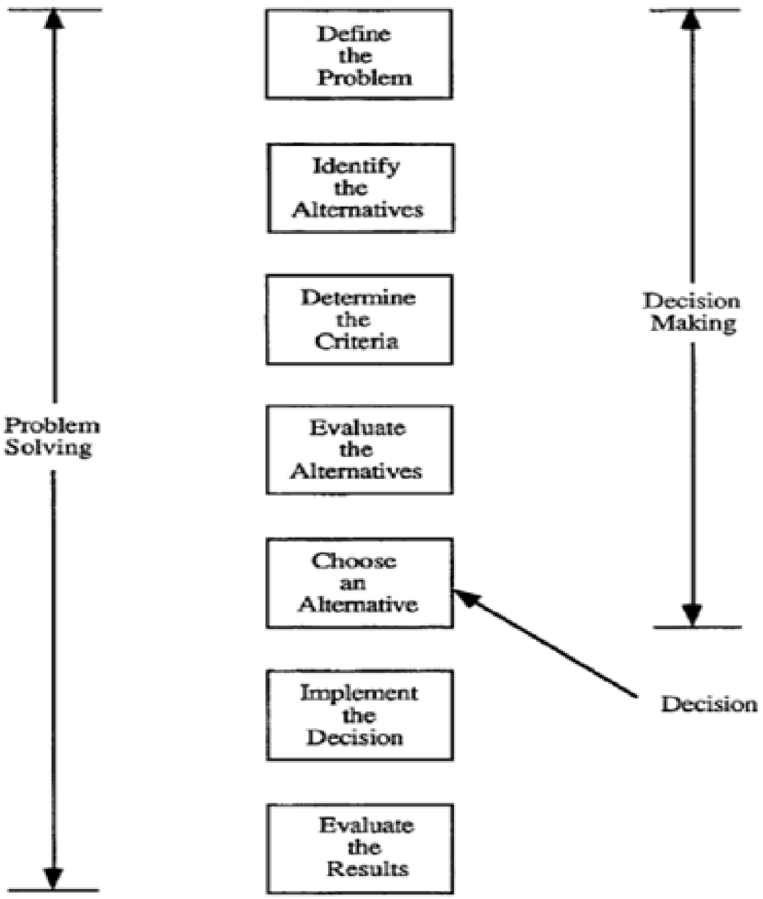


Fig.2-1 Problem Solving and Decision Making
Source: (Anderson et al., 1995, p. 4)

It should be noted that there is concern that these tools do not lead to real improvement in decision making. According to Klein (2003) and other authors, people do not make decisions using tools and the intuitive style of decision making needs to replace the disaggregated approaches commonly used by most decision analysts. Decision analysts point out that their approach is prescriptive, providing a prescription of what actions to take based on sound logic, rather than a descriptive approach, describing the flaws in the way people do make decisions.

Overall *a good decision maker should understand both approaches*, understanding how people go wrong in making decisions and providing a sound basis for them to make better decisions. Furthermore, past research studies like (Dawes & Corrigan, 1974), (Fischhoff et al., 1982), and others conclusively show how even the simplest decision analysis methods are superior to "unaided intuition" and there are several areas within decision analysis, which deal with normative results that are provably optimal for specific quantifiable decisions, and for which human intuition alone will almost never be correct or even close to correct. For example, the optimal order scheduling in a manufacturing facility or optimal hedging strategies are purely mathematical and their results are necessarily provable.

Logical decision making is an important part of all science-based professions, where specialists apply their knowledge in a given area to making informed decisions. The connection with business forecasting is that we may not think that we are forecasting, but our choices will be directed by our anticipation of results of our actions or inactions.

As a matter of fact, nowadays *forecasts are an important element in making informed decisions* and they affect decisions and activities throughout business organizations, for example in Accounting and Finance (cost/profit estimates, cash flow and funding), Human resources (hiring, recruiting and training), Marketing (pricing, promotion, strategy), Operations (schedules, MRP, workloads), Product/service design (new products and services) and so on.

It is not excessive to say that all business decisions are based on forecasts. Every decision becomes operational at some point in the future, so it should be based on forecasts of future conditions. Managers should use forecasting models and forecast information to assist them in decision-making process for better decisions. Researchers and scientists should use them to develop better products/services or improve current ones. Stevenson (1998) writes: "Prediction is at least two things: important and hard. Important, because we have to act, and hard because we have to realize the future we want, and what is the best way to get there.”

This book discusses various ways of making forecasts that rely on logical methods of manipulating the data that have been generated by historical events (see Chapters 5, 6, 7 and 8). Their purpose is to help managers and administrators do a better job of anticipating, and hence a better job of managing uncertainty, by using effective forecasting and other predictive techniques.

2.2. General System Theory and Business Forecasting

Systems theory is an interdisciplinary theory about the nature of complex systems in nature, society, and science, and is a framework by which one can investigate and/or describe any group of objects that work together to produce some result. This could be a single organism, any organization or society, or any electro-mechanical or informational artifact. As a technical and general academic area of study, it predominantly refers to the science of systems that resulted from Von Bertalanffy's **G**eneral **S**ystem **T**heory - **GST** (Von Bertalanffy, 1968), among others, in initiating what became a project of systems research and practice.

Many early systems theorists aimed at finding a GST that could explain all systems in all fields of science. Von Bertalanffy's objective was to bring together, under one heading, the organismic science that he had observed in his work as a biologist. His desire was to use the word "***system***" to describe those principles which are common to systems in general. In GST, he writes:

> *...there exist models, principles, and laws that apply to generalized systems or their subclasses, irrespective of their particular kind, the nature of their component elements, and the relationships or "forces" between them. It seems legitimate to ask for a theory, not of systems of a more or less special kind, but of universal principles applying to systems in general* (Von Bertalanffy, 1968, p. 32).

Systems theory is an area of study specifically developed following the World Wars from the work of Ludwig von Bertalanffy, Anatol Rapoport, Kenneth E. Boulding, William Ross Ashby, Margaret Mead, Gregory Bateson, C. West Churchman, and others in the 1950s. It was catalyzed by the cooperation in the Society for General Systems Research (SGSR), a predecessor of the current International Society for the Systems Sciences[1] (ISSS), known to be one the first interdisciplinary and international co-operations in the field of systems theory and systems science. This organization was initiated in 1954 as "Society for the Advancement of General Systems Theory", got formally underway as "Society for General Systems Research" and was eventually renamed in 1988.

Cognizant of advances in science that questioned classical assumptions in the organizational sciences, Bertalanffy's idea to develop a theory of systems began as early as the interwar period, publishing "An Outline for General Systems Theory" in the *British Journal for the Philosophy of Science*[2]. Where assumptions in Western science from Greek thought with Plato and Aristotle to Newton's *Principia*[3] have historically influenced all areas from the hard to social sciences, the original theorists explored the implications of twentieth-century advances in terms of systems.

Subjects like *complexity, self-organization, connectionism,* and *adaptive systems* had already been studied in the 1940s and 1950s. In cybernetics, researchers like Norbert Wiener, William Ross Ashby, John von Neumann, and Heinz von Foerster examined complex systems using mathematics. Von Neumann discovered cellular automata[4] and self-reproducing systems, Aleksandr Lyapunov and Jules Henri Poincaré worked on the foundations of chaos theory. During the period 1929 to 1951 Robert Maynard Hutchins at the University of Chicago had

[1] http://isss.org/world/

[2] See Vol 1, No. 2, 1950.

[3] See http://en.wikipedia.org/wiki/Philosophiae_Naturalis_Principia_Mathematica

[4] A cellular automaton (pl. *cellular automata*) is a discrete model studied in computability theory, mathematics, physics, complexity science, theoretical biology, and microstructure modeling.

undertaken efforts to encourage innovation and interdisciplinary research in the social sciences, aided by the Ford Foundation with the interdisciplinary Division of the Social Sciences established in 1931. Numerous scholars had been actively engaged in ideas before, but in 1937 von Bertalanffy presented the general theory of systems at a conference at the University of Chicago.

The Cold War affected the research project for systems theory in ways that sorely disappointed many of the seminal theorists. Some began to recognize theories defined in association with systems theory had deviated from the initial GST view. The economist Kenneth Boulding, an early researcher in systems theory, had concerns over the manipulation of systems concepts. Boulding concluded from the effects of the Cold War that abuses of power always prove consequential and that systems theory might address such issues. Since the end of the Cold War, there has been a renewed interest in systems theory with efforts to strengthen an ethical view.

The systems view was based on *several fundamental ideas*. *First, all phenomena can be viewed as a web of relationships among elements, or a system*. *Second, all systems,* whether electrical, biological, or social, *have common patterns, behaviors, and properties* that can be understood and used to develop greater insight into the behavior of complex phenomena and to move closer toward a unity of science. System philosophy, methodology, and application are complementary to GST.

Science systems thinkers consider that:

- a ***system*** is a dynamic and complex whole, interacting as a structured functional unit;
- energy, material and information flow among the different elements that compose the system;
- a system is a community situated within an environment;
- energy, material and information flow from and to the surrounding environment via semi-permeable membranes or boundaries;
- systems are often composed of entities seeking equilibrium but can exhibit oscillating, chaotic, or exponential behavior.

Consequently, a *holistic system is any set (group) of interdependent or temporally interacting parts.* ***Parts*** are generally systems themselves and are composed of other parts, just as systems are generally parts or ***holons***[5] of other systems (see Fig.2-2).

[5] A *Holon* (Greek) is something that is simultaneously a whole and a part.

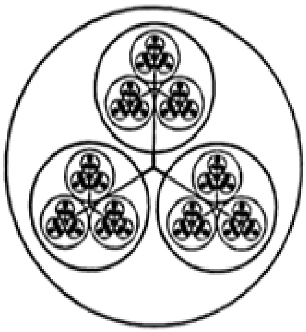


Fig.2-2 Systems, Relationships and Holons, represented by circles
Source: (Skyttner, 2006, p. 63)

Following these explanations, we will accept the following general definition of a system:

A System is a set of interrelated parts (elements) that must work together to achieve a common goal.

The systems thinking approach incorporates several main principles **(*tenets*)** as presented by Skyttner (2006):

- *Interdependence* of objects and their attributes - independent elements can never constitute a system;
- *Holism* - emergent properties not possible to be detected by analysis should be possible to be defined by a holistic approach;
- *Goal seeking* - systemic interaction must result in some goal or final state;
- *Inputs and Outputs* - in a closed system inputs are determined once and constant; in an open system additional inputs are admitted from the environment;
- *Transformation* of inputs into outputs - this is the process by which the goals are obtained;
- *Entropy* - the amount of disorder or randomness present in any system;
- *Regulation* - a method of feedback is necessary for the system to operate predictably;
- *Hierarchy* - complex wholes are made up of smaller subsystems;
- *Differentiation* - specialized units perform specialized functions;
- *Equifinality* - alternative ways of attaining the same objectives (convergence);
- *Multifinality* - attaining alternative objectives from the same inputs (divergence).

For example, using the principle of "Multifinality", a supermarket could be considered to be:

- a "profit making system" from the perspective of management and owners
- a "distribution system" from the perspective of the suppliers
- an "employment system" from the perspective of employees

- a "materials supply system" from the perspective of customers
- an "entertainment system" from the perspective of loiterers
- a "social system" from the perspective of local residents
- a "dating system" from the perspective of single customers

As a result of such thinking, new insights may be gained into how the supermarket works, why it has problems, how it can be improved or how changes made to one component of the system may impact the other components.

Fig.2-3 summarizes some of these principles (Hierarchy, Interdependence, Holism, etc.) and the systems model in Fig.2-4 represents others (Inputs/Outputs, Transformation, Regulation etc.) in a different view.

Science systems and the application of science systems thinking has been grouped into three categories (in fact two groups and a combination of them) based on the techniques used to tackle a system:

- ***Hard systems*** — involving simulations, often using computers and the techniques of operations research. Useful for problems that can justifiably be quantified. However, it cannot easily take into account unquantifiable variables (opinions, culture, politics, etc.), and may treat people as being passive, rather than having complex motivations.
- ***Soft systems*** — for systems that cannot easily be quantified, especially those involving people holding multiple and conflicting frames of reference. Useful for understanding motivations, viewpoints, and interactions and addressing qualitative as well as quantitative dimensions of problem situations.

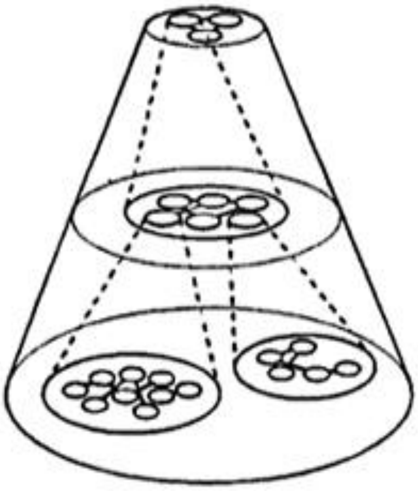


Fig.2-3 A Multilevel systems Hierarchy
Source: (Skyttner, 2006, p. 61)

While soft system thinking treats all problems as ill-defined or not easily quantified, hard systems approaches (the so-called structured methods) assume that:

- the problems associated with such systems are well-defined;
- they have a single, optimum solution;
- a scientific approach to problem-solving will work well;
- technical factors will tend to predominate.

The two approaches, *Hard systems* and *Soft systems*, form the basic platforms in model development, which is discussed in detail in section 2.4. ***Cybernetics*** is another interdisciplinary approach closely related to GST, which also forms very important fundamentals in modeling.

As we mentioned above cybernetics was defined by Norbert Wiener (1948), as the study of control and communication in the animal and the machine. Stafford Beer called it the science of effective organization and Gordon Pask extended it to include information flows "in all media" from stars to brains. It includes the study of feedback, black boxes, and derived concepts such as communication and control in living organisms, machines, and organizations, including ***self-organization***[6].

Cybernetics is a broad field of study, but the essential goal of cybernetics is to understand and define the functions and processes of systems that have goals and that participate in circular, causal chains that move from action to sensing to comparison with the desired goal, and again to action (see Fig.2-4). Studies in cybernetics provide a means for examining the design and function of any system, including social systems such as business management and organizational learning, for the purpose of making them more efficient and effective.

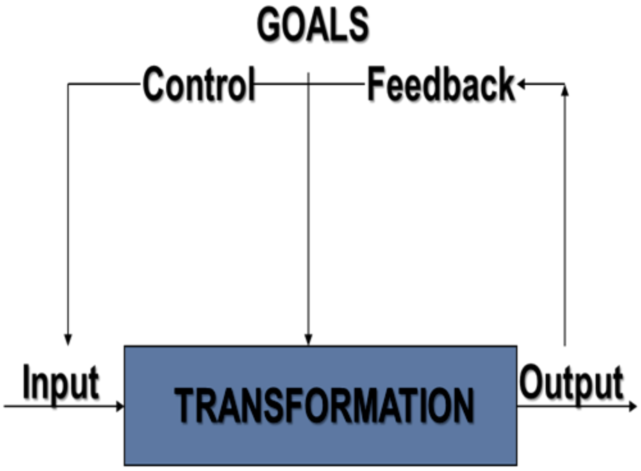


Fig.2-4 General Systems Model

[6] The idea of *self-organization* is explored in detail in Chapters 4 and 12.

As a starting point for the comprehension of the basic terms of cybernetics, a system may be represented by three boxes: the black, the grey and the white (see Fig.2-5). The purposeful action performed by the box is its ***function***. Inside each box, there are ***structural components***, the static parts, ***operating components*** which perform the processing, and ***flow components***, the matter/energy or information being processed.

Each box contains processes of input, transformation, and output. (Note that output can be of two kinds: products useful for the super-system and/or waste. Also, note that the input to one system may be the output of its subsystem.) Taken together these processes are called *throughput*, to avoid focus on individual parts of internal processes (Skyttner, 2006, p. 72).

The box colors denote different degrees of user interest in the knowledge (or understanding) of the internal working process of a system. A **black box** is a primitive element that behaves in a certain way without giving any clue to the observer how exactly the result is obtained. A ***grey box*** offers partial knowledge of selected internal processes and the ***white box*** represents a wholly transparent view, giving full information about internal processes.

The systems framework is also fundamental to organizational theory as organizations are complex dynamic goal-oriented processes. A systemic view on organizations is trans-disciplinary and integrative. In other words, it transcends the perspectives of individual disciplines, integrating them based on a common "code", or more exactly, on the basis of the formal apparatus provided by systems theory. The systems approach gives primacy to the interrelationships, not to the elements of the system. It is from these dynamic interrelationships that new properties of the system emerge.

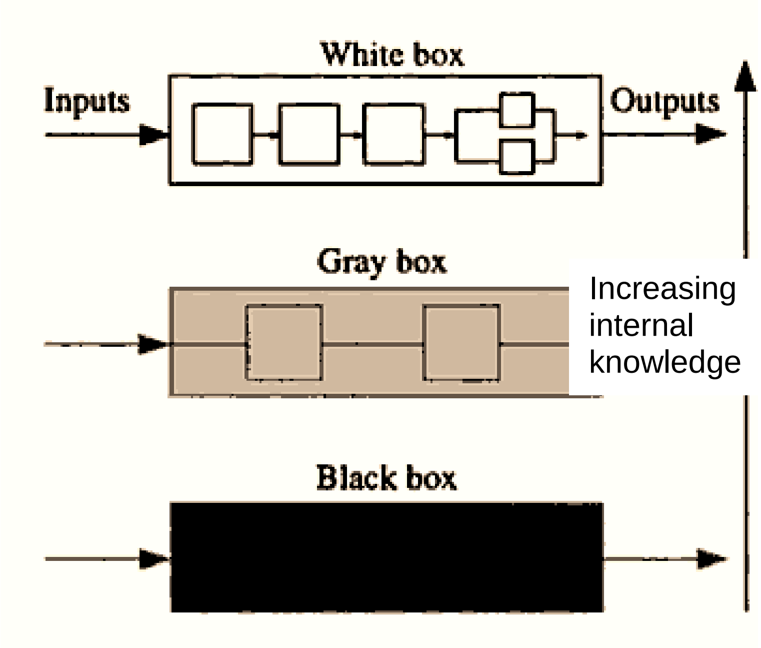


Fig.2-5 Degrees of Internal Knowledge

Ilya Prigogine[7] has studied emergent properties, suggesting that they offer analogs for living systems. An *emergent behavior* or *emergent property* can appear when a number of simple entities (or *elements*) operate in an environment, forming more complex behaviors as a collective, and *emergence* is the way complex systems and patterns arise out of a multiplicity of relatively simple interactions, i.e. "***The whole is greater than the sum of the parts***".

The quote above also involves *synergy,* which is where different entities cooperate advantageously for a final outcome. If used in a business application, it means that teamwork (cooperation of people with different complementary skills) will produce an overall better result than if each person was working toward the same goal individually.

The stock market (or any market for that matter) is an example of emergence on a grand scale. As a whole, it precisely regulates the relative security prices of companies across the world, yet it has no leader; there is no one entity which controls the workings of the entire market. Agents, or investors, have knowledge of only a limited number of companies within their portfolio and must follow the regulatory rules of the market and analyze the transactions individually or in large groupings. Trends and patterns emerge, which are studied intensively by technical analysts.

The World Wide Web (www) is another popular example of a decentralized system exhibiting emergent properties. There is no central organization rationing the number of links, yet the number of links pointing to each page follows a power law in which a few pages are linked to many times and most pages are seldom linked to. A related property of the network of links in the www is that almost any pair of pages can be connected to each other through a relatively short chain of links.

During the second half of the 20th century, the science systems thinking had increasingly been used to tackle a wide variety of subjects in fields such as computing, engineering, epidemiology, information science, health, manufacture, management, and the environment. Many examples are related to business and forecasting: Organizational architecture, Linear and Complex Process Design, Supply Chain Design, Business continuity planning, Delphi method, Futures studies, Leadership development, Quality function deployment, Quality management, Program management, Project management and others. For example, Fig.2-6 uses the general systems model shown in Fig.2-4 to represent Business Organization as a System.

[7] **Ilya, Viscount Prigogine** (25 January 1917 – 28 May 2003) was a Russian-born naturalized Belgian physical chemist and Nobel Laureate noted for his work on dissipative structures, complex systems, and irreversibility – see "Self-Organization in Non-Equilibrium Systems", 1977, Wiley.

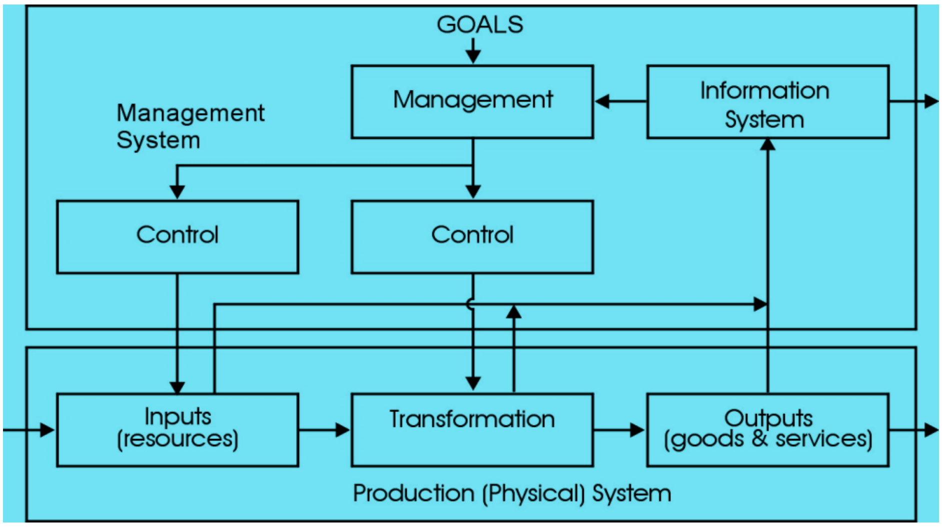


Fig.2-6 Business Organization as a System

There are two direct uses for forecasts when considering the business organization as a system. One is to help managers ***plan the system*** (i.e. the business organization), and the other is to help them ***plan the use of the system***. ***Planning the system*** generally involves long-range plans and forecasts about the types of products and services to offer, what facilities and equipment to have where to locate, and so on. ***Planning the use of the system*** refers to short-range and intermediate-range forecasting, which involve tasks such as planning inventory and workforce levels, planning purchasing and production, budgeting, and scheduling.

2.3. Main steps in Business Forecasting Process

In both theory and practice, one can find two slightly different approaches to the business forecasting process. First one is based on the assumption that we can select a forecasting technique before data analysis. According to Stevenson (2017, p. 74) there are six basic steps in the forecasting process (see Fig.2-7):

1. *Determine the purpose of the forecast.* How will it be used and when will it be needed? This step will provide an indication of the level of detail required in the forecast, the number of resources (personnel, computer time, dollars, etc.) that can be justified, and the level of accuracy necessary.

2. *Establish a time horizon.* The forecast must indicate a time interval, keeping in mind that usually accuracy decreases as the time horizon increases, and to also provide the time necessary to implement possible changes.

3. *Select a forecasting technique* – the variety of forecasting techniques and criteria, used to select the most appropriate one, are discussed in Chapter 5.

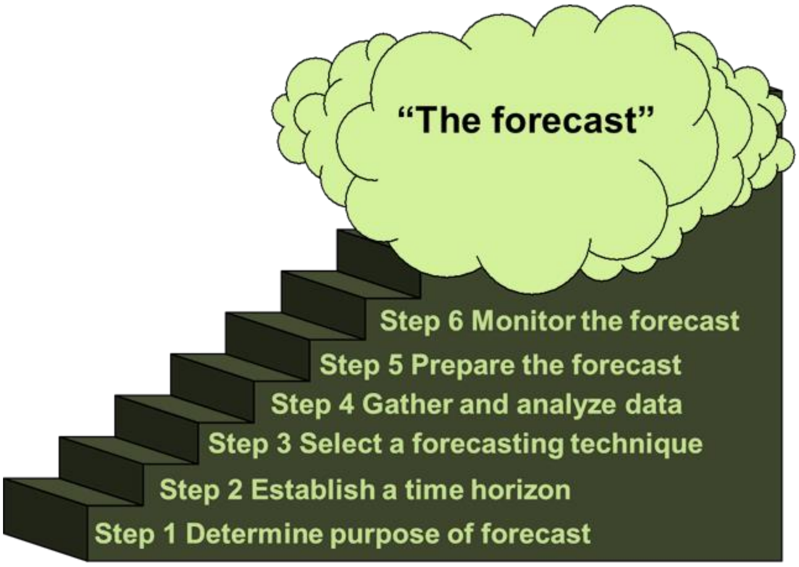

"The forecast"
Step 6 Monitor the forecast
Step 5 Prepare the forecast
Step 4 Gather and analyze data
Step 3 Select a forecasting technique
Step 2 Establish a time horizon
Step 1 Determine purpose of forecast

Fig.2-7 Steps in the Forecasting Process (a)

4. *Collect, clean, and analyze appropriate data.* Obtaining the data can involve significant effort. Once available, the data may need to be "*cleaned*", i.e. to get rid of outliers and obviously incorrect data before analysis.

5. *Make the forecast* – this is the actual model forecasts that are generated from the appropriate data.

6. *Monitor the forecast.* A forecast must be monitored to determine whether it is performing in a satisfactory manner. If it is not, reexamine the method, assumptions, data validity, and so on, make modifications as needed, and prepare a revised forecast.

Sometimes, an additional action may be necessary. For example, if demand was much less than the forecast, an action such as a price reduction or a promotion may be needed. Conversely, if demand was much more than predicted, increased output may be advantageous. That may involve working overtime, outsourcing, or taking other measures.

Most formal forecasting procedures involve extending the experiences of the past into the future. Thus, they involve the assumption that the conditions that generated past data are indistinguishable from the conditions of the future except for those variables explicitly recognized by the forecasting model.

For example, a human resource department is hiring employees, in part, based on a company entrance examination score because, in the past, examination score seemed to be an important predictor of job performance rating. As far as this relationship continues to hold, forecasts of future job performance, hence hiring decisions, can be improved by using examination scores. If, for some reason, the association between examination score and job performance changes, then forecasting job performance ratings from examination scores using the historical model will yield inaccurate forecasts and potentially poor hiring decisions. This is what makes forecasting difficult. The future is not always like the past. To the extent that it is true, quantitative forecasting methods work well. To the extent it is not true, inaccurate forecasts can result. However, it is generally better to have some reasonably constructed forecast than no forecast.

The recognition that forecasting techniques operate on the data generated by historical events leads to the second approach in forecasting process identification. (Hanke et al., 2008, p. 5) define five steps in the forecasting process (see Fig.2-8):

Step 1. *Problem formulation and data collection* are treated as a single step because they are intimately related. The problem determines the appropriate data. If a quantitative forecasting methodology is being considered, the relevant data must be available and correct. Often accessing and assembling appropriate data is a challenging and time-consuming task. If appropriate data are not available, the problem may have to be redefined or a non-quantitative forecasting methodology employed. Collection and quality control problems frequently arise whenever it becomes necessary to obtain pertinent data for a business forecasting effort.

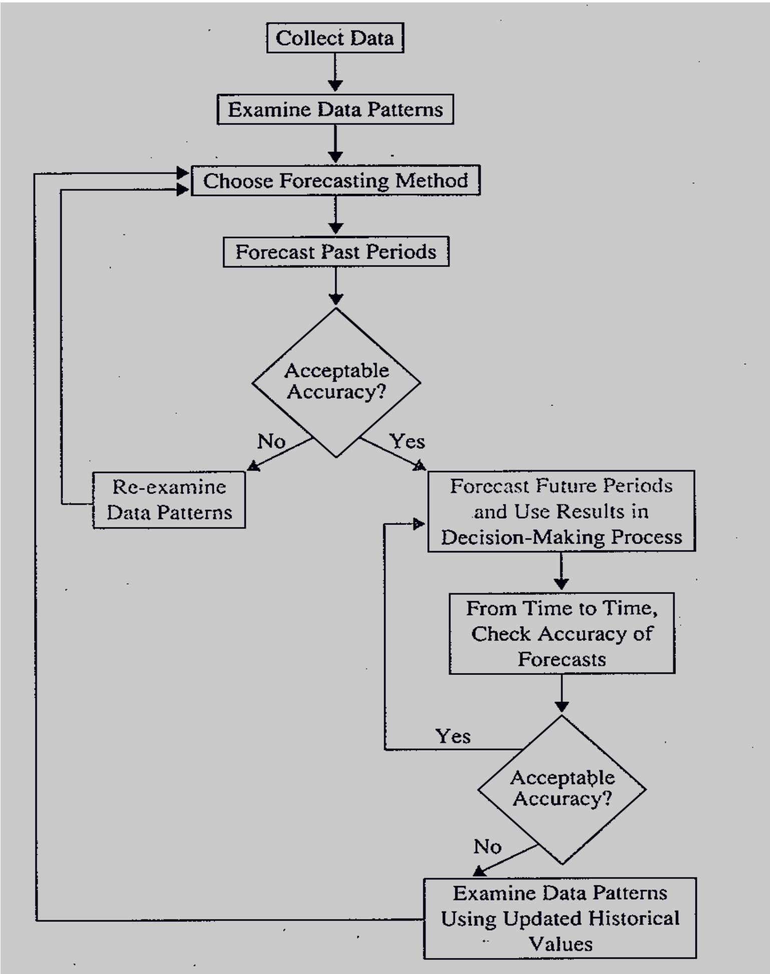

Collect Data
Examine Data Patterns
Choose Forecasting Method
Forecast Past Periods
Acceptable Accuracy?
No
Yes
Re-examine Data Patterns
Forecast Future Periods and Use Results in Decision-Making Process
From Time to Time, Check Accuracy of Forecasts
Yes
Acceptable Accuracy?
No
Examine Data Patterns Using Updated Historical Values

Fig.2-8 Steps in the Forecasting Process (b)

Step 2. D*ata preprocessing (manipulation and cleansing)*, is often necessary because is possible to have too much data as well as too little in the forecasting process. Some data may not be relevant to the problem and/or other data may have missing values that must be estimated. Also, data may have to be re-expressed in units other than the original units and/or may have to be preprocessed (for example, accumulated from several sources and summed). Some data may be appropriate but only in certain historical periods (for example, in forecasting the sales of small cars one may wish to use only car sales data after the oil embargo of the 1970s rather than data over the past 50 years). Ordinarily, some effort is required to get data into a form that is required for using certain forecasting procedures.

Step 3. *Model building and evaluation* involve fitting the collected data into a forecasting model that is appropriate in terms of minimizing forecasting error. The simpler the model, the better it is in terms of gaining acceptance of the forecasting process by managers who must make the firm's decisions. Often a balance must be struck between a sophisticated forecasting approach that offers slightly more accuracy and a simple approach that is easily understood and gains the support of, and is actively used by, the company's decision-makers, i.e. judgment is involved in this selection process. It is our hope that the reader's ability to exercise good judgment in the choice and use of appropriate forecasting models will increase after studying this book, which discusses numerous forecasting models and their applicability. Despite user experience, sometimes the choice of the model is too subjective. How to make this process more objective will be discussed in Chapter 3 and Chapter 12 will discuss model building techniques, which limit the user involvement in the process to the inclusion of well-known a priori knowledge.

Step 4. *Model implementation* (the actual forecast) consists of the actual model forecasts, which are generated once the appropriate data have been collected (and possibly reduced) and an appropriate forecasting model has been chosen. Forecasting for recent periods, in which the actual historical values are known, is often used to evaluate the accuracy of the process. Next, the forecasting errors are analyzed and summarized, which is a part of Step 5.

Step 5. *Forecast evaluation* involves comparing forecast values with actual historical values. In this process, a few of the most recent data values are often held back from the data set being analyzed. After the forecasting model is completed, forecasts are made for these periods and compared with the known historical values. Some forecasting procedures sum the absolute values of the errors and may report this sum or divide it by the number of forecast attempts to produce the average forecast error. Other procedures produce the sum of squared errors, which is then compared with similar figures from alternative forecasting methods. Some procedures also track and report the magnitude of the error terms over the forecasting period. Examination of error patterns often leads the analyst to a modification of the forecasting procedure. Specific methods of measuring forecasting errors are discussed in Chapter 3.

Another list of steps might be added to this one – Feedback after the forecasting process is underway to determine if sufficient accuracy has been obtained and if management is finding the forecast useful and cost-effective in the decision-making process.

Armstrong (2001) used 139 standards and principles to summarize knowledge about forecasting. These principles, organized into 16 categories, cover formulating problems, obtaining information, implementing methods, evaluating methods, and using forecasts. The main steps, as we can see, are very close to the second approach described above. His goal, as he pointed out, was that we could not generalize only "Five Principles Used by Successful Forecasters," because they would never be appropriate for all the different situations that can arise. Of course, we do not need all 139 of them in any situation. Nearly all of the principles are conditional on the characteristics of the situation. They include the major principles, but ignore some that are specific only to a certain forecasting method.

Users can examine the forecasting processes by systematically judging them against these 139 forecasting principles presented. When managers receive forecasts, they often cannot judge their quality. Instead of focusing on the forecasts, however, they can decide whether the *forecasting process* was reasonable for the situation. By examining the forecasting processes and improving them, managers may increase accuracy and reduce costs.

Despite the fact that some of the principles could be argued, it is a good idea to use the checklist of principles provided in this article to assist in auditing the forecasting process, which can help any forecast's user to find ways to improve the forecasting process and/or to avoid legal liability for poor forecasting. In fact, most principles (and in particular the main groups) are either based on expert opinion or widely accepted in terms of common sense, that means it is difficult to imagine that things could be otherwise. In this textbook we are using many common principles that are part of the 139-item list, and at the same time there are some unique ones representing the newest developments and achievements in the forecasting such as Self-Organizing Data Mining, Group Method of Data Handling, Statistical Learning Networks, and Multi-Stage Selection Algorithms, which are discussed in Chapters 3, 8, 9, 11 and 12.

2.4. General Forecasting Model and Model Building Approaches

All forecasts are developed using a model (see Fig.2-9[8]), which can be very simple or very complex, as mentioned earlier. Sterman (1991, p. 209) noted:" As computers have become faster, cheaper, and more widely available, computer models have become commonplace in forecasting and public policy analysis, especially in economics, energy and resources, demographics, and other crucial areas. As computers continue to proliferate, more and more policy debates, both in government and the private sector, will involve the results of models. Though not all of us are going to be model builders, we all are becoming model consumers, regardless of whether we know it (or like it)".

It is known from GST that connection between input and output variables can be expressed by Volterra functional series (Madala & Ivakhnenko, 1994, p. 19), discrete analogue of which is Kolmogorov-Gabor polynomial (2-1):

$$y = a_0 + \sum_{i=1}^{M} a_i x_i + \sum_{i=1}^{M}\sum_{j=1}^{M} a_{ij} x_i x_j + \sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k=1}^{M} a_{ijk} x_i x_j x_k , \qquad (2\text{-}1)$$

Where $X(x_1, x_2, ..., x_M)$ is input variables vector.

$A(a_1, a_2, ..., a_M)$ - vector of coefficients or weights.

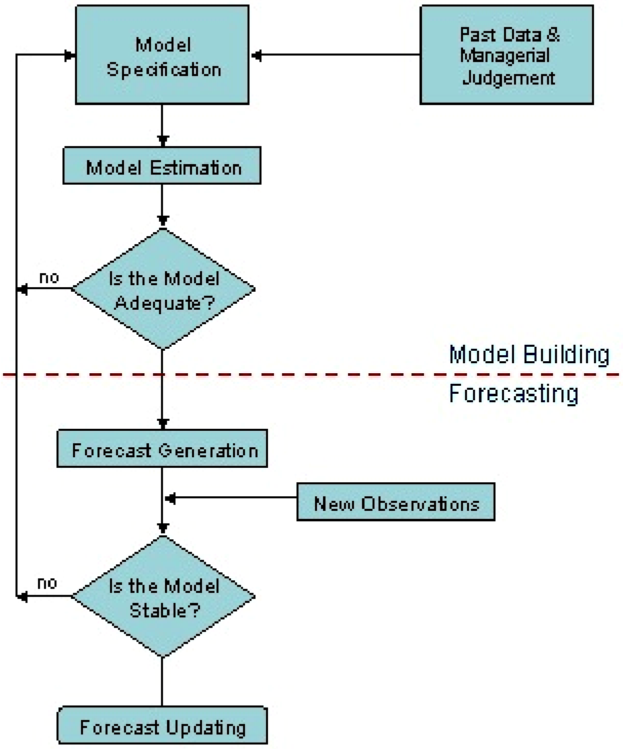


Fig.2-9 Forecasting System – The Model-Building and The Forecasting Phases

[8] See http://home.ubalt.edu/ntsbarsh/stat-data/Forecast.htm#rrstatthink

Components of the input vector *X* can be ***independent (exogenous) variables***, functional forms, or finite difference terms. Other non-linear reference functions, such as difference, logistic and/or harmonic can also be used for model constructions. In forecasting, this general systems model can be presented with the following mathematical expression:

$$Y = F(X, \varepsilon) \tag{2-2}$$

Where **F** can be any mathematical function describing the forecasted variable **Y** ***(the output***) as a function of ***input variables*** **X** and the ***stochastic component*** **ε** (***model error***).

In many areas (economy, ecology, sociology etc.) the objects of interest are ill-defined systems that can be characterized by inadequate a priori information about the system, a large number of immeasurable variables, noisy and/or small data samples (short time-series data), and fuzzy objects with attributive variables.

For such ill-defined objects, the hard systems thinking based on the assumption that the world could be understood objectively and that knowledge about the world can be validated by empirical means need to be substituted by a soft systems paradigm. This approach is based on the reflection that humans have an incomplete and fuzzy understanding of the world. In this sense, models do not reflect the world, but they capture the logic of the world. They are useful constructions that may help to discuss and consider the world (for more details see (Keys, 1991).

Problems of complex objects modeling like systems identification, pattern recognition, approximation and extrapolation, and forecasting can be solved either by ***deductive logical-mathematical modeling*** or by ***inductive sorting-out methods***. In the first approach (known as ***theory-driven approach*** or ***theoretical systems analysis***), models can be developed based on existing theory. Deductive methods have advantages in cases of rather simple modeling problems. Here, the theory of the object being modeled is well known and valid, and thus it is possible to develop a model from physically based principles employing user's knowledge of the problem. The elements of this model are represented by different variables described by numerical values. Then, the cause-and-effect relationships are formulated as mathematical equations or their equivalents.

A key assumption of this approach is that it is possible to construct and manipulate a model of the problem under study. The analytic process involves breaking-down parts of the world into constituent parts that will be simpler than their aggregate. These smaller parts are then more manageable, and an understanding of the original focus of interest is gained by bringing these separate pieces of knowledge together. In such a way the model gives an explanation of the behavior of the actual processes of the real system.

In summary (see Fig.2-10), the ***theory-driven*** approach starts with a well-known theory to select the appropriate model and then uses observed data to compute the model coefficients. One typical example (Mueller & Lemke, 2003, p. 13) from marketing is discussed below:

A company wants to model and predict its product demand. It has recorded several characteristics for past periods (month or weeks) which it expects/or knows due to market theory to have an impact on product demand. These variables might be: dollars spent for advertising, number of products in the pipeline, number of products in stock, number of new consumers, consumer confidence, consumer income, and inflation as input variables or environmental influences and number of products sold, profit margin, and gross profit as output variables of the system. The theory-driven modeling approach forces the company's research staff to formulate qualitatively the interdependence structure among all three output and all seven input variables and to define the relevant dynamics (time lags) for each variable.

Researchers and/or managers must decide whether the number of products sold is influenced by the dollars spent on advertising or by the number of products in the pipeline or by both or by any other possible constellation. Then, the staff must agree also, for instance, whether consumer confidence of one month, two months or six months ago has an influence on next month's product sales. Most important, however, is that the problem is composed of several aspects: global, macro and micro economical, psychological and political. So, eventually there are two more questions:

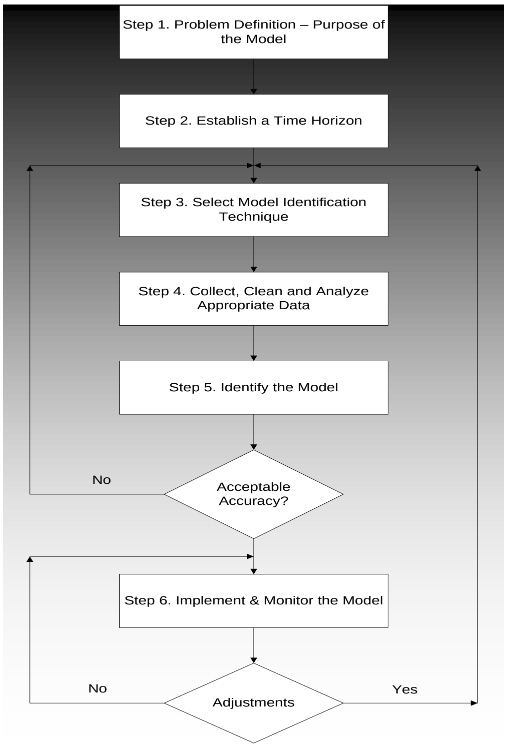

Step 1. Problem Definition – Purpose of the Model
Step 2. Establish a Time Horizon
Step 3. Select Model Identification Technique
Step 4. Collect, Clean and Analyze Appropriate Data
Step 5. Identify the Model
No
Acceptable Accuracy?
Step 6. Implement & Monitor the Model
No
Adjustments
Yes

Fig.2-10 Steps in Model Building Process (Theory-Driven Approach)

- What theory do they have to focus on at a given stage and how these theories should be combined?
- What are the rules for connecting different theories?

Here, it is marketing theory that may be somewhat helpful. Usually, however, these rules are not known completely. Since theory-driven modeling relies on them anyway, apparently the research staff members must make wild guesses several times in their work. Each time they change their assumptions, they may get totally different results. This, of course, is reasonable. However, what or who will decide what the true assumptions are, this is the problem.

In complex, ill-defined systems, such as most business processes, a researcher a priori has only insufficient knowledge about the relevant theory of the system under study. Thus, theory-driven modeling is influenced considerably by the fact that the researcher is a priori uncertain regarding the selection of the model structure due to insufficient knowledge about its variables and relationships.

This insufficient a priori information concerns necessary knowledge about the object such as those relating to:

- the main factors of influence and classification of variables into endogenous (dependent) and exogenous (independent);
- the functional form (equation) of the relation between the variables including the dynamic specification of the model, and
- the description of errors such as their correlation structure.

Consequently, the comprehensive application of theory-driven approach in practice and theoretical systems research is hampered by several essential problems linked to modeling, for example:

a. Model building requires much qualified scientific work. For example, the elaboration of the German economic development model was done by a group of two to three scientists and some staff members working full time from 1962 up to 1968 (Krelle et al., 1969).
b. The different scope of knowledge about the subject of study forces the researcher to an arbitrary approach to include the theoretical facts into a mathematical description (equation) and therefore it influences the uncertainty of the results. The tacit hope that an actual process will continue to go as presupposed in the assumptions is, seriously taken, not at all justified. The effect of model simplification is among the basic problems of classical econometrics and optimization and it implies to make various assumptions, which are often resulting in a considerable shortage regarding models' validity in reality.

c. The parameter identification (Step 5 in Fig.2-10) is linked to considerable difficulties and causes the researcher to study the conditions for that identification. As Sterman noted: ”The regression procedures used to estimate parameters yield unbiased estimates only under certain conditions. These conditions are known as *maintained hypotheses* because they are assumptions that must be made in order to use the statistical technique. The maintained hypotheses can never be verified, even in principle, but must be taken as a matter of faith” (Sterman, 1991, p. 223). Even sophisticated techniques which do not impose too many restrictive assumptions always involve other a priori hypotheses that cannot be validated.

d. For any required model property, a specific object can be described commonly by various mathematical models with a similar degree of accuracy and it is up to the researcher to select one or the other model. However, what will be the criterion for that choice?

e. Traditional mathematical modeling is characterized by the stigma that modeling means simplification, isomorphic relation, and one-sided reflection and that, on the other hand, the mathematical modeling needs exact description.

f. Finally, there is the Zadehs principle of incompatibility, described in (Kosko, 1994): “As the complexity of a system increases, our ability to make precise and significant statements about its behavior diminishes until a threshold is reached beyond which precision and significance (or relevance) become almost mutually exclusive characteristics. A corollary principle may be stated succinctly as the closer one look at a real-world problem the fuzzier becomes its solution.”

To solve the problems mentioned above, a considerable amount of specialized-scientific systems analysis, modeling, mathematical-statistical and computational work is necessary. This causes the need to extend the arsenal of well-proven principles of modeling by new and appropriate procedures. Further development of modeling has also contributed to improving the epistemological function of the models. This function is limited due to its links to the subjective ideas of the modeler, the uncertainty of a priori information and the two-stage model building (structure identification and parameters estimation).

According to Sterman (1991, p. 222) “The chief weak spots in econometric models stem from the assumptions of the underlying economic theory on which they rest: assumptions about the rationality of human behavior, about the availability of information that real decision makers do not have, and about equilibrium. Many economists acknowledge the idealization and abstraction of these assumptions, but at the same time point to the powerful results that have been derived from them. However, a growing number of prominent economists now argue that these assumptions are not just abstract, they are false.”

These flaws in theory-driven modeling have generated serious criticism from within the economics profession. Lester Thurow notes that econometrics has failed as a method for testing theories and is now used primarily as a "showcase for exhibiting theories." Yet as a device for advocacy, econometrics imposes few constraints on the prejudices of the modeler. He concludes: ”By simple random search, the analyst looks for the set of variables and functional forms that give the best equations. In this context, the best equation is going to depend heavily upon the prior beliefs of the analyst. If the analyst believes that interest rates do not affect the velocity of money, he finds the 'best' equation that validates his particular prior belief. If the analyst believes that interest rates do affect the velocity of money, he finds the 'best' equation that validates this prior belief” (Thurow, 1983, p. 108).

But the harshest assessment of all comes from Nobel laureate Wassily Leontief: ”Year after year economic theorists continue to produce scores of mathematical models and to explore in great detail their formal properties; and the econometricians fit algebraic functions of all possible shapes to essentially the same sets of data without being able to advance, in any perceptible way, a systematic understanding of the structure and the operations of a real economic system” (Leontief, 1982, p. 107).

Additionally, it is necessary to reduce the high share of time modeling that it has at the solution of tasks by means of a computerized draft of mathematical models. A variety of computer-based ideas has already been developed in the past for improving the computing instruments in this application domain, supporting both the modeling and the forecasting in this way. A possible trend of further development in this scientific area is based on theoretical systems analysis and leads to computer simulation. The methodology of simulation (algorithmic description) is developed successfully on the level of mathematical modeling and on the level of computerized programming. As a result, special simulation languages have been developed to support the application of models on a computer as executable programs (Mueller & Lemke, 2003).

In the ***data-driven*** approach (or ***experimental systems analysis***), using inductive sorting-out methods, models are derived from real-life data (see Fig.2-11). In many cases, when it is impossible to create a model by theoretical systems analysis, sophisticated applications of inductive sorting-out methods may be able to reveal hidden patterns of relationships and the corresponding model. The data-driven approach generates a description of the system behavior from observations of real systems evaluating how it behaves (what are the outputs) under different conditions (inputs). This is like statistical modeling and its goal is to infer general laws from specific cases. However, usually, the mathematical relationship that assigns an input to an output (and in this way imitates the behavior of a real-world system using these relationships) has nothing to do with the real processes running in the system. The system is not described in its details and functions and is treated as a black box (see Fig.2-5).

The task of experimental systems analysis is to select mathematical models from data of N observations and of M system variables X(it) (where **i** varies from l to M, and t=1 to N), to select the structure of the mathematical model (***structure identification***) and to estimate the unknown parameters (***parameter identification***). Statistically based principles of model building (Regression Analysis – see Chapter 6), which require a priori information about the structure of the mathematical model available are usually used. A good deal of work goes into identifying, gathering, cleansing and labeling data; specifying the questions to be asked about data and finding the right way to discover useful patterns from data. Unfortunately, such processing can take up a big part of the whole project in terms of efforts, time and funds.

The results obtained so far show that methods of experimental systems analysis cannot be used to analyze the causes of events for fuzzy objects. There are a few important facts that should be underlined:

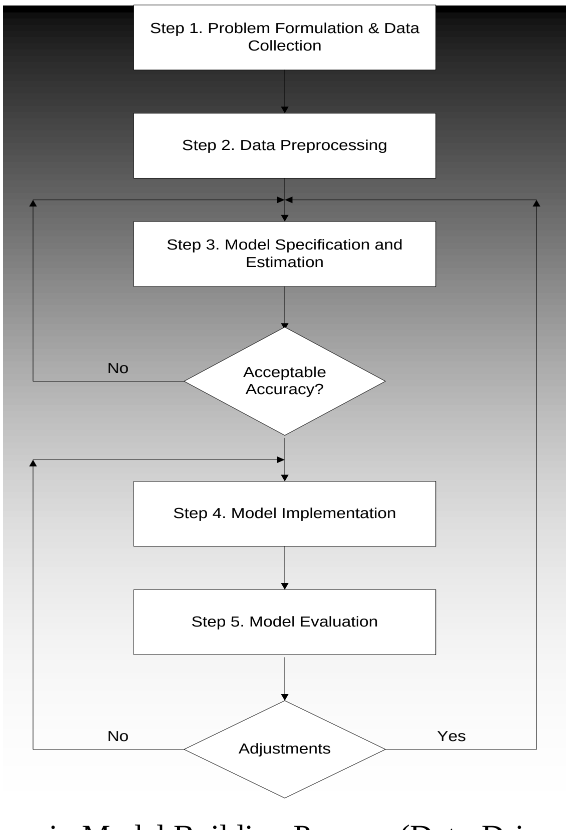

Step 1. Problem Formulation & Data Collection
Step 2. Data Preprocessing
Step 3. Model Specification and Estimation
No
Acceptable Accuracy?
Step 4. Model Implementation
Step 5. Model Evaluation
No
Adjustments
Yes

Fig.2-11 Steps in Model Building Process (Data-Driven Approach)

A. The goal of data-driven modeling is to estimate the unknown relationship between the output (y) and the input (x) from a set of past observations. A model, built in this way, is only able to represent a relation between inputs and outputs, which values are within the observed sample of data.

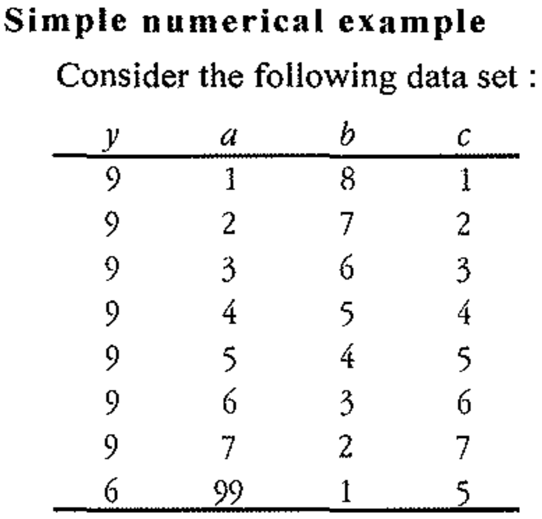
B. Many other factors that are not observed or controlled may influence the system's output. Therefore, knowledge of observed input values does not uniquely specify the output. This uncertainty of the output is caused by the lack of knowledge of unobserved factors. The result is a statistical dependency between the observed inputs and outputs.

C. There is a difference between statistical dependency and causality – the task of learning/estimation of statistical dependency between observed inputs and outputs can occur in the following situations or any of their combinations:

- output causally depends on the (observed) inputs;
- inputs causally depend on the output(s);
- input-output dependency is caused by other (unobserved) factors;
- input-output correlation is non-causal (i.e. there is a spurious correlation).

Statistical techniques fail to distinguish between correlations and causal relationships. Statistical techniques used to estimate parameters don't prove whether a relationship is causal. They only reveal the degree of past correlation between the variables, and these correlations may change or shift as the system evolves (Sterman, 1991).

It follows that causality cannot be inferred from data analysis alone. Instead, each of the four possibilities (or their combinations) is specified, and, therefore, causality must be assumed or demonstrated by arguments outside the data. The following example from Mueller and Lemke (2003 p. 18) shows how properties of input data can affect the system's model (output).



Simple numerical example

Consider the following data set :

y	*a*	*b*	*c*
9	1	8	1
9	2	7	2
9	3	6	3
9	4	5	4
9	5	4	5
9	6	3	6
9	7	2	7
6	99	1	5

If we use the general model equation (2-2), then:

$$\mathbf{Y = F\ (a,b,c,\varepsilon)} \tag{2-3}$$

where F can be any mathematical function describing the unknown (dependent) variable Y as a function of input variables a, b & c, and the stochastic component **ε** (model error).

If we apply a linear model and the ***Least Squares*** method[9], the following solutions are:

1. $y = 9.3 - 0.033a - 0.033b$
2. $y = 0.00001 + b + c$
3. $y = 9 - 0.0319a + 0.0319c$

All three models are very precise with model error very close to zero, but apparently not very useful for the system's analysis. With insufficient a priori information about the system, there are *several methodological problems* one must focus on before applying data-driven methodologies. Besides those mentioned in theory-driven modeling, the incomplete - since finite - database we use leads to an indeterminacy of the model and the computational data derived from it. Also, the effectiveness of different conclusions drawn from this data by means of mathematical statistics is limited. This incompleteness of the theoretical knowledge and the insufficiency of data cause the following problems:

a. The model-adequate estimation of the unknown coefficients of parametric models is commonly based on traditional statistical approaches and assumptions (see Chapter 7). Statistical analysis includes some useful hypothesis tests, but they can only be verified within the observed samples of data (Sterman, 1991).

b. According to ***the set principle of modeling***, many models can exist with a sufficient same degree of adequateness for a given sample of input and output observations. Therefore, the task of selecting a model from an existing data sample of endogenous and exogenous variables is an ill-defined task. It is not possible, generally, to select an optimal model from the number of possible models without some additional, external information (Madala & Ivakhnenko, 1994, pp. 11-17). For example, the regression learning problem is ill-defined in case of absence of any assumption about the nature of the continuous approximating functions. For limited training data, the solution that minimizes the empirical risk is not unique. There is an infinite number of functions from the class of continuous functions that can interpolate that data points yielding the minimum solution with respect to a given loss function.

[9] ***Least squares (LS)*** means that the overall solution minimizes the sum of the squares of the model errors **ε** (see Chapter 7).

c. Considering time series, most of them contain a trend component. With modeling based on these non-stationary time series, the danger emerges that relations between different variables with a similar growth trend will be established, although these relations do not exist in reality. Collinearity between the predictor variables confuses the interpretation of the associated parameters but can also be harmful to predictions.

d. Generally, dynamic systems are characterized by various growth processes. Therefore, differential equations obtained from observed growth processes are, as a rule, not stable. Utilization of unstable models or models that include unstable partial models is very dangerous because a small deviation in the initial conditions will cause much larger deviations in the model results.

e. Traditional models based on statistical techniques are unable to provide a guide to performance under conditions that have not been experienced previously. Researchers assume that the correlations indicated by the historical data will remain valid in the future. In reality, those data usually span a limited range and provide no guidance outside historical experience. As a result, models are often less than robust: faced with new policies or conditions, the models break down and lead to inconsistent results (Sterman, 1991).

f. Modeling assumes that functional relations are of relative constancy over the evaluated period. To satisfy this requirement, short time series must be applied. This means the modeler must meet contradictory requirements when establishing mathematical models. Due to the limited quantity of data the uncertainty of estimation increases with many model parameters. On the other hand, however, the reality is reflected more convincingly with growing model complexity since reality is complex.

g. To judge the quality of models merely by formal criteria like the closeness of fit of the model and true system is doubtful. Instead, it is necessary to have a purposeful judgment of the quality of model adaptation based on the suitability of the model to solve a predefined task (see Chapter 3). Transforming the requirements on a model to an adequately formalized criterion usually involves considerable difficulties.

h. A rule for parametric models is that the number of unknown model parameters must be smaller than the number of observations[10]. However, complex systems require many systems variables to be measured, since the necessary dimension of the state

[10] i.e. to have enough ***degrees of freedom*** (n-k), where **n** is the sample size of observations and **k** is the number of unknown parameters (independent variable coefficients) to be estimated.

space in which the system trajectory will be completely described in without redundancy is commonly unknown. On the other hand, the number of observations cannot be extended infinitely, because many economic and ecological systems, for example, are characterized by a strongly restricted set of available observations. If the number of predictors is large, one problem is that traditional modeling methods quickly become ill-behaved and computational unmanageable due to many parameters that have to be estimated simultaneously. If the number of systems variables in a data sample is larger than the number of observations, the modeling task is called an undetermined task[11]. Such under-determined tasks can be solved by means of inductive selection procedures that will be discussed in Chapter 12.

To solve this collection of possible problems it is necessary to develop appropriate data-driven tools for automatic modeling, because most users' primary interest is limited only in their field and they may not have time for learning advanced mathematical, cybernetic and statistical techniques and/or for using dialog-driven modeling tools (Mueller & Lemke, 2003, p. 19).

In summary, we can conclude that problems exist in both groups and a possible solution is in the unification of these methodologies. ***Knowledge discovery from data*** and in particular ***data mining*** techniques can help researchers analyze the massive amounts of data and turn information located in the data into successful decisions. These techniques are discussed in Chapters 10, 11, and 12.

***

[11] It is also referred to as ***overfitting*** which occurs when a model is excessively complex, such as having too many parameters relative to the number of observations

SUMMARY AND CONCLUSIONS

- In this chapter we discussed the Forecasting Process and its relations with other scientific areas like Decision Making, General System Theory and Model Building.
- There is a common logical sequence in any ***problem solving*** and there are four general phases – ***Understanding the problem, Devising a plan, Carrying out the plan*** and ***Looking back.***
- ***Problem solving*** and ***decision making*** are very close –decision making is a part of every particular problem-solving approach and it might be regarded as a problem-solving activity which is terminated when a satisfactory solution is found Decision Making.
- A good decision maker should understand both approaches ***(prescriptive*** and ***descriptive***), understanding how people go wrong in making decisions and providing a sound basis for them to make better decisions.
- ***Forecasts are an important element in making informed decisions*** and they affect decisions and activities throughout business organizations.
- ***Systems theory*** is an interdisciplinary theory about the nature of complex systems in nature, society, and science, and is a framework by which one can investigate and/or describe any group of objects that work together to produce some result.
- The systems view is based on *several fundamental ideas – First, all phenomena can be viewed as a web of relationships among elements, or a system*. *Second, all systems*, whether electrical, biological, or social, *have common patterns, behaviors, and properties* that can be understood and used to develop greater insight into the behavior of complex phenomena and to move closer toward a unity of science. System philosophy, methodology and application are complementary to GST.
- ***A System is a set of interrelated parts (elements) that must work together to achieve a common goal.***
- ***Hard systems*** involve simulations, often using computers and the techniques of operations research. They are useful for problems that can justifiably be quantified. However, it cannot easily take into account unquantifiable variables (opinions, culture, politics, etc.), and may treat people as being passive, rather than having complex motivations.
- ***Soft systems*** are used for systems that cannot easily be quantified, especially those involving people holding multiple and conflicting frames of reference. They are

useful for understanding motivations, viewpoints, and interactions and addressing qualitative as well as quantitative dimensions of problem situations.

- *Box colors denote different degrees of user interest in the knowledge (or understanding) of the internal working process of a system:* a ***black box*** is a primitive element that behaves in a certain way without giving any clue to the observer how exactly the result is obtained; a ***grey box*** offers partial knowledge of selected internal processes and the ***white box*** represents a wholly transparent view, giving full information about internal processes.
- There are two different approaches of the business forecasting process. The first one assumes that we can select a forecasting technique before data analysis. The second approach is based on the assumption that forecasting techniques operate on data generated by historical events and the most appropriate technique should be selected by fitting the collected data into a forecasting model that is appropriate in terms of minimizing forecasting error.
- All forecasts are developed using a model which can be very simple or very complex one. There are two groups of methods – ***deductive logical-mathematical modeling*** (known as ***theory-driven approach*** or ***theoretical systems analysis***), assuming that models can be developed on the basis of existing theory and ***inductive sorting-out methods*** (known as ***data-driven*** approach or ***experimental systems analysis***), in which models are derived from real-life data.
- Problems exist in both groups and a possible solution is in unification of these methodologies. ***Knowledge discovery from data*** and in particular ***data mining techniques*** can help researchers analyzing the massive amounts of data and turning information located in the data into successful decisions. These techniques are discussed in Chapters 10, 11 and 12.

KEY TERMS

Black box	***43***	***Data-driven approach***	
Decision analysis	***35***	(***experimental systems analysis***)	***55***
Decision making	***35***	***Degrees of freedom***	***59***
General System Theory	***37***	***Grey box***	***43***
Hard System	***41***	***Holons***	***39***
Independent variables		***Knowledge discovery from data,***	
exogenous (inputs)	***51***	***Data mining***	***60***
Least Squares method	***58***	***Output variable***	***51***
Overfitting	***60***	***Parameter identification***	***56***

Problem solving	***35***	***Self-organization***	***38, 42***
Soft System	***41***	***Stochastic component*** (ε) (***model error***)	***51***
Structure identification	***56***	***System***	***39, 40***
Theory-driven approach (***theoretical systems analysis)***	***52***	***White box***	***43***

Chapter Exercises

Conceptual Questions:

1. List all four phases of common logical sequence in any problem.
2. What are the steps in decision making process? Discuss.
3. How does the white box differ from the black box? Discuss.
4. List the main steps in the business forecasting process? What are the differences in the two main approaches?
5. What are the two approaches in model building? Discuss.

Business Applications:

A young graduate decided to return home and help his parents in their farm business. In the beginning he conducted analysis of the olive harvests. He noted that olives ripen in September. Each March his parents try to determine if the upcoming harvest would be bountiful. If his analysis and forecast indicate it would, his parents would enter into agreements with the owners of all the olive oil presses in the region. In exchange for a small deposit months ahead of the harvest, they would obtain the right to lease the presses at market prices during the harvest. If his prediction about the harvest was correct and demand for oil presses boomed, his parents could make a great deal of money.

1. Identify the following forecasting elements in the context of this scenario:
    - Input (independent or exogenous) variables.
    - Output (dependent or endogenous) variable.
    - Forecasting horizon (i.e. the number of periods between when the forecast is made and time period to which it applies).
    - Forecasting period (i.e. the unit of time for which forecasts are to be made – week, month, quarter, year, etc.).
    - Forecasting interval (usually, it is the frequency with which new forecasts are prepared).
2. What forecasting approach will fit better the above case. Discuss.

INTEGRATIVE CASE

HEALTHY FOOD SUPPLY CHAIN & STORES

Part 2: Organization as a system – developing a system model

In Chapter 1 we introduced *Healthy Food Stores* – a fast-growing retail food provider with 12 stores in a northwestern state. The company executives decided to study the effect that company advertising dollars have on sales. They hoped that examining collected historical data would reveal relationships that would help determine future advertising expenditures and predict monthly sales volumes for the upcoming quarter.

Company historical records revealed that sales experienced a seasonal effect (as shown in the sales-data chart in Part 1 of the study) and for this reason, a categorical (qualitative) variable was used to indicate each month. Another important finding, which was also revealed from the initial data analysis and time-series plots in Part 1, was that the sales experienced some positive trend.

Finally, the company executives wanted to find out if the *Healthy Food Stores* advertising would have any effect on its major competitors' advertising budgets the following month. An important, categorical variable was used for this purpose, which was coded as 1 (representing a little amount in the following month competitors' advertising), 2 (a moderate amount) and 3 (a significant amount).

After clarifying the above details, the research team in the company created a data file (Data.xslx) containing the following variables:

- Sales volume (in \$ thousand) – the dependent variable of interest $\mathbf{Y_t}$.
- Traditional advertising (in \$ thousand) – independent variable $\mathbf{X_{1t}}$.
- Online advertising (in \$ thousand) – independent variable $\mathbf{X_{2t}}$.
- Time period (a series from 1 to 48) – independent variable **t** used to indicate sales trend.
- A series of 12 numbers – a categorical (dummy) variable to indicate each month, where January code is 1, February – 2, through December – 12 $\mathbf{X_{3t}}$.
- Another categorical variable (coded 1, 2 or 3) to indicate competitors' advertising efforts the following month $\mathbf{X_{4t}}$.
- Along with these current data, company IS (as mentioned in the previous part) can derive time lags for any variable (dependent $\mathbf{Y_{t-s}}$ or independent $\mathbf{X_{t-s}}$). Initially, it was decided to use sales and advertising values lagged one and two months (s=1, 2).

After identifying the basic parameters and input (independent) and output (dependent) variables of the forecasting scenario, the research team discussed the main elements of the forecasting process. They agreed that the forecasting horizon should be up to twelve months (i.e. one year) and that they should update the forecast every quarter, keeping in mind that accuracy decreases as time horizon increases, and sufficient time is needed to implement possible changes.

Case Questions

1. What do you think about research team decisions? Discuss.
2. Produce an Input/Output system model for this specific case, identifying all important elements (refer to Fig. 2-4 to 2-6).
3. What model building approach would you recommend to the research team? Discuss.
4. Write a short report (about two pages not counting charts and tables) on the questions above, discussing all important findings and draw relevant conclusions about this part of the Integrative Case.

References

Anderson, D., Sweeney, D. & Williams T. (1995). *Quantitative Methods for Business*. West Publishing Company.

Armstrong, S. (2001). Standards and Practices for Forecasting. *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Kluwer Academic Publishers.

Dawes, R. M. & Corrigan, B. (1974). Linear Models in Decision Making. *Psychological Bulletin 81*(2), 93–106.

Fischhoff, B., Phillips, L. D. & S. Lichtenstein. (1982). Calibration of Probabilities: The State of the Art to 1980. Kahneman, D. & Tversky, A. (Eds), *Judgement under Uncertainty: Heuristics and Biases*. Cambridge University Press.

Hanke, J. & Wichern, D. (2008). *Business Forecasting.* PEARSON & Prentice Hall.

Howard, R. (1966). Decision Analysis: Applied Decision Theory. *Proceedings of the 4th International Conference on Operational Research* (pp. 55-77).

Kahneman, D. & Tversky, A. (2000). *Choice, Values, Frames*. The Cambridge University Press.

Keys, P. (1991). *Operational Research and Systems*. Plenum Press. New York and London.

Klein, G. (2003). *The Power of Intuition*, Doubleday. New York.

Kosko, B. (1994). *Fuzzy Thinking.* An Imprint of Harper Collins Publ. Flamingo.

Krelle, W., Beckerhoff, H., Langer, H. & FuB, H. (1969). *Ein Prognosesystem fiir die wirtschaftliche Entwicklung der BRD*. Verlag Anton Hain. Meisenheim am Glan.

Leontief, W. (1982). Academic Economics. *Science 217*, 104-107.

Madala, H. R., & Ivakhnenko, A. G. (1994). *Inductive Learning Algorithms for Complex Systems Modelling*. Boca Raton, FL: CRC Press Inc.

Mueller J. A., & Lemke, F. (2003). *Self-Organizing Data Mining: An Intelligent Approach To Extract Knowledge From Data.* Victoria, BC: Trafford Publishing.

Pólya, G. (1945). *How to Solve It.* Princeton University Press.

Skyttner, L. (2006). *General Systems Theory: Problems, Perspective, Practice*. World Scientific Publishing Company.

Sterman, J. D. (1991). A Skeptic's Guide to Computer Models. In Barney, G. O. et al. (Eds.), *Managing a Nation: The Microcomputer Software Catalog* (pp. 209-229). Boulder, CO: Westview Press.

Stevenson, H. (ed.) (1998). DO LUNCH OR BE LUNCH. *Boston: Harvard Business School Press.*

Stevenson, W. J. (2017). *Operations Management*. McGraw-Hill/Irwin.

Thurow, L. (1983). *Dangerous Currents.* New York, NY: Random House..

von Bertalanffy, L. (1968). *General System Theory: Foundations, Development, Application.* New York: George Braziller.

Wiener, N. (1948). *Cybernetics or Control and Communication in the Animal and the Machine*. New York, NY: John Wiley & Sons.

CHAPTER 3. FORECAST ACCURACY AND MODEL SELECTION

3.1. Forecast Error

Risk and uncertainty are central to forecasting and prediction and it is generally considered a good practice to indicate the degree of uncertainty attached to each forecast. This Chapter discusses how to measure forecast accuracy and to select the most appropriate forecasting model.

Reducing uncertainty requires that a forecast should be sufficiently ***accurate*** for its purpose and to be relied upon by the decision maker who will use it. In the past, most commonly accuracy was used as a description of systematic errors, a measure of statistical bias, i.e. ***Accuracy*** was considered as the proximity of measurement results to the true value and ***Precision*** as the degree to which repeated measurements, under unchanged conditions, show the same results. Often, ***Precision*** is measured with respect to detail and ***Accuracy*** is measured with respect to reality (Acken, 1997, pp. 281-306).

A shift in the meaning of these terms appeared with the publication of the International Organization for Standardization (**ISO**) 5725 series of standards in 1994, which was last reviewed and confirmed in 2018. The purpose of ISO 5725-1 "is to outline the general principles to be understood when assessing ***accuracy*** (trueness and precision) of measurement methods and results, and in applications, and to establish practical estimations of the various measures by experiment" (ISO 5725-1, 1994, p.1). This standard uses two terms "***trueness***" and "***precision***" to describe the accuracy of a measurement method. ***Trueness*** refers to the closeness of agreement between the arithmetic mean of a large number of test results and the true or accepted reference value. ***Precision*** refers to the closeness of agreement between test results.

In this regard, the chart presented in Fig.1-2 should be replaced by Fig.3-1. According to the ISO 5725, ***accuracy*** is a general term which consists of both ***trueness*** and ***precision***.

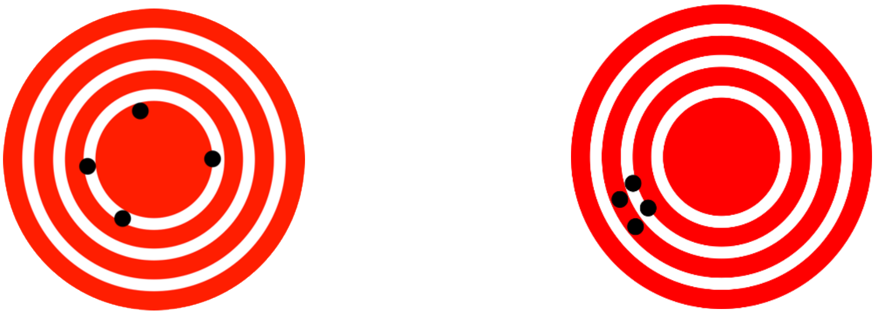


Fig.3-1. A) Low accuracy due to poor precision. B) Low accuracy due to poor trueness (Source:https://commons.wikimedia.org/wiki/File:Accuracy_and_precision-highaccuracylowprecision.png)

There is no such thing as ***absolute accuracy*** and raising the level of accuracy increases cost but does not necessarily increase the value of information. It is important that the ***degree of accuracy*** should be clearly stated in the beginning of each particular forecasting process. This will enable users to plan for possible errors and will provide a basis for comparing alternative forecasts.

As mentioned in Chapter 1, forecasts are not perfect, and actual results usually differ from predicted values. Predictions of outcomes are rarely precise, and the forecaster can only endeavor to make the inevitable errors as small as possible. The difference between the actual value and the predicted value for the corresponding period is the ***forecast error.*** By default, this error is defined using the value of the outcome minus the value of the forecast (3-1):

$$\mathrm{e}_t = y_t - F_t \tag{3-1}$$

where $\mathbf{e_t}$ is the forecast error at period $\mathbf{t}$ (***t={1, 2, 3...N}***);

- ***N*** is the forecasting interval (or the size of the dataset);
- $\mathbf{y_t}$ is the actual value at period $\mathbf{t}$ and
- $\mathbf{F_t}$ is the forecast for period $\mathbf{t}$.

Forecast error can be a calendar forecast error or a cross-sectional forecast error, when we want to summarize the forecast error over a group of units. If we observe the average forecast error for a time-series of forecasts for the same product or phenomenon, then we call this a ***calendar-forecast error*** or ***time-series forecast error***. If we observe this for multiple products for the same period, then this is a ***cross-sectional performance error***.

Many forecasting techniques have been developed as an attempt to reduce the forecast error. Unfortunately, real-life experience shows that no single technique works in every situation. Many types of forecasting techniques will be discussed in the following chapters. In the next sections of this Chapter we will learn how to measure and evaluate forecast accuracy and how to select the best forecasting model out of many potential good ones.

3.2. Measures of Forecast Accuracy

Each forecast represents the real-life business variable with some accuracy, related to the particular size of the forecast error. It is good to know some general facts, for example, the fact that forecast accuracy decreases as ***time horizon*** increases, i.e. short-range forecasts usually contend with fewer uncertainties than long-range forecasts, and thus they tend to be more accurate.

However, it is more important to know the degree of each particular forecast accuracy. There are different techniques to quantify the degree of accuracy. In most cases, the forecast is compared with an outcome at a single time-point and a summary of forecast errors is constructed over a collection of such time-points.

In some cases, a forecast may consist of predicted values over a number of lead-times. In this case an assessment of forecast error may need to consider more general ways of assessing the match between the time-profiles of the forecast and the outcome. If a main application of the forecast is to predict when certain thresholds will be crossed, one possible way of assessing the forecast is to use the timing-error, i.e. the difference in time between the value at which the outcome crosses the threshold and when the forecast does so. When there is interest in the maximum value being reached, assessment of forecasts can be done using any of:

- the difference of times of the peaks;
- the difference in the peak values of the outcome and the forecast;
- the difference between the peak value of the outcome and the forecast value at that time point.

The forecast error should always be calculated using actual data as a base. Traditionally, measures of fit are used to evaluate how well the forecasts match the actual values. If given the computed forecasts (**F**$_t$) and their errors (**e**$_t$) at period **t**, for a dataset ***t={1, 2, 3...N}***, then:

– **Mean Forecast Error (MFE)** is the average value of the forecast errors (3-2):

$$\mathrm{MFE} = \frac{1}{\mathrm{N}}\sum_{t=1}^{N} e_t \tag{3-2}$$

The main properties of this measure of accuracy are:

- It is a measure of the average deviation of forecasted values from actual ones.
- It shows the direction of the errors and, thus, it measures the ***Forecast Bias***[1].
- In MFE, the effects of positive and negative errors cancel out and there is no way to know their exact amount.
- A zero value of MFE does not mean that forecasts are perfect (i.e. contain no error) – rather, it only indicates that the forecasts are on proper target.
- It does not penalize extreme errors.
- It depends on the scale of measurement and it is also affected by data transformations.

[1] A ***forecast bias*** occurs when there are consistent differences between actual outcomes and previously generated forecasts of those quantities, i.e. forecasts may have a general tendency to be too high or too low.

- For a good forecast, i.e. to have a minimum bias, it is desirable that the MFE is as close to zero as possible.

– **Mean Absolute Deviation (MAD)** or **Mean Absolute Error (MAE)** is the average absolute value of the differences between the actual value at period **t** and the forecast for period **t**. It is easy to compute and is most useful to measure the forecast error in the same units as the original series (3-3):

$$\mathrm{MAE} = \frac{1}{\mathrm{N}}\sum_{t=1}^{N}\left|e_t\right| \qquad (3\text{-}3)$$

MAE (MAD) properties are:

- It measures the average absolute deviation of forecasted values from original ones.
- It shows the magnitude of overall error, occurred due to forecasting.
- In MAE, the effects of positive and negative errors do not cancel out.
- Unlike MFE, MAE does not provide any idea about the direction of errors.
- Like MFE, MAE also depends on the scale of measurement and data transformations.
- It weights errors linearly, but it is not sensitive to extreme values.
- Extreme forecast errors are not penalized by MAE.
- For a good forecast, the obtained MAE should be as small as possible.

– **Mean Absolute Percentage Error (MAPE)** puts errors in perspective. It is useful when the size of the forecast variable is important in evaluating. It provides an indication of how large the forecast errors are in comparison to the actual values. It is also useful to compare the accuracy of different techniques on same or different data. The computing formula is (3-4):

$$\mathrm{MAPE}\,(\%) = \frac{1}{\mathrm{N}}\sum_{t=1}^{N}\left(\left|e_t\right|/\,y_t\right)\mathrm{x}100 \qquad (3\text{-}4)$$

MAPE important features are:

- This measure represents the percentage of average absolute error that occurred.
- It is independent of the scale of measurement but affected by data transformations.
- Unlike MFE, MAPE does not show the direction of error.
- MAPE does not penalize extreme deviations.
- In this measure, opposite signed errors do not offset each other (i.e. the effects of positive and negative errors do not cancel out).
- For a good forecast, the obtained MAPE should be as small as possible.

– **Mean Percentage Error (MPE)** is useful when it is necessary to determine whether a forecasting method is biased. If the forecast is unbiased MPE will produce a value that is close to zero. Large negative values mean overestimating and large positive values indicate that the method is consistently underestimating. A disadvantage of this measure is that it is undefined whenever a single actual value is zero. In terms of its computation, it is an average percentage error (3-5):

$$\text{MPE}\,(\%) = \frac{1}{\text{N}}\sum_{t=1}^{N}(e_t\,/\,y_t)\text{x}100 \qquad (3\text{-}5)$$

The properties of MPE are:

- MPE represents the percentage of average error occurred, while forecasting.
- It is independent of the scale of measurement but affected by data transformations.
- Like MFE, MPE shows the direction of errors occurred.
- MPE does not penalize extreme deviations.
- Opposite signed errors affect each other and cancel out, i.e. as MFE, by obtaining a value of MPE close to zero, we cannot conclude that the model is perfect.
- It is desirable that for a good forecast with a minimum bias, the obtained MPE should be as close to zero as possible.

– **Mean squared error (MSE)** or **Mean Squared Prediction Error (MSPE)** is the average of the squares of the errors, i.e. the differences between the actual value and the forecast at period **t**:

$$\text{MSE} = \sum\left(e_t\right)^2/(n-1) \qquad (3\text{-}6)$$

The MSE is the second moment (about the origin) of the error, and thus incorporates both the variance of the estimator and its bias. For an unbiased estimator, the MSE is the variance of the estimator. Like the variance, MSE has the same units of measurement as the square of the quantity being estimated.

MSE (MSPE) properties are:

- It is athe measure of average squared deviation of forecasted values.
- Since the opposite signed errors do not offset one another, MSE gives an overall idea of the error.
- It penalizes extreme errors (it squares each) occurred while forecasting.
- MSE emphasizes the fact that the total forecast error is in fact greatly affected by large individual errors, i.e. large errors are more expensive than small errors.

- MSE does not provide any idea about the direction of overall error.
- It is sensitive to the change of scale and data transformations.
- Although MSE is a good measure of overall forecast error it is not as intuitive and easily interpretable as the other measures discussed above.

– **Root Mean Squared Error (RMSE)** is the square root of calculated MSE. In an analogy to standard deviation, taking the square root of MSE yields the ***root-mean-squared error (RMSE)***, which has the same units as the quantity being estimated (3-7):

$$RMSE = \sqrt{\text{MSE}} \qquad (3\text{-}7)$$

The properties of **RMSE** are:

- Mathematically, RMSE is nothing but the square root of calculated MSE.
- All the properties of MSE, but the last one, hold for RMSE as well.
- Unlike MSE, RMSE measures the forecast error in the same units as the original series and it has an easy and clear business interpretation.

The RMSE serves to aggregate the magnitudes of the errors in predictions for various times into a single measure of predictive power. Thus, it is a good measure of accuracy, but only to compare forecasting errors of different models for a particular variable and not between variables, as it is scale dependent.

– **Normalized Mean Squared Error (NMSE)** is the MSE divided by the variance of observed values of a variable being predicted (3-8):

$$\text{NMSE} = \text{MSE}/\sigma^2 \qquad (3\text{-}8)$$

Its features are:

- NMSE normalizes the obtained MSE after dividing it by the test variance.
- It is a balanced error measure and is very effective in judging the forecast accuracy of a model.
- The value is often expressed as a percentage, where lower values indicate less residual variance.
- The smaller the NMSE value, the better forecast.
- Other properties of NMSE are the same as those of MSE.

A similar measure is the **Normalized Root Mean Squared Error (NRMSE)**, which is the RMSE divided by the range of observed values of a variable being predicted (3-9):

$$\text{NRMSE} = \text{RMSE}/(y_{max}\text{-}y_{min}) \qquad (3\text{-}9)$$

The advantage of NRMSE is that unlike NMSE it measures the forecast error in the same units as the original series and it has an easy and clear business interpretation.

– **The Coefficient of variation of the RMSE, CV(RMSE)** is defined as the RMSE normalized to the mean of the observed values (3-10):

$$\mathrm{CV(RMSE)} = \mathrm{RMSE}/\bar{\mathrm{y}} \qquad (3\text{-}10)$$

It is the same concept as the ***coefficient of variation*** (***CV***) except that RMSE replaces the standard deviation. The CV is useful because the standard deviation of data must always be understood in the context of the mean of the data. In contrast, the actual value of the CV is independent of the unit in which the measurement has been taken, so it is a dimensionless number. For comparison between datasets with different units or widely different means, we should use the CV instead of the standard deviation. The smaller the CV(RMSE) value, the better the forecast. In general, its properties are same as those of RMSE and NRMSE. In addition:

- When the mean value is close to zero, the CV(RMSE) will approach infinity and is therefore sensitive to small changes in the mean. This is often the case if the values do not originate from a ratio scale.
- Unlike the RMSE, it cannot be used directly to construct prediction intervals.

– **Theil's U-statistics** (3-11) is defined as a normalized measure of total forecast error, which varies between zero and one. By the rule of the thumb, for a good forecast accuracy, the **U**-statistic should be close to zero:

$$\mathrm{U} = \frac{\sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}}{\sqrt{\frac{1}{n}\sum_{t=1}^{n} f_t^2}\,\sqrt{\frac{1}{n}\sum_{t=1}^{n} y_t^2}} \qquad (3\text{-}11)$$

Its properties are:

- It is a normalized measure of the total forecast error.
- $0 \le \mathrm{U} \le 1; \mathrm{U} = 0$ means a perfect fit.
- This measure is affected by the change of scale and data transformations.
- For assessing good forecast accuracy, it is desirable that the U-statistic is close to zero.

– **The forecast skill** or **skill score (SS)** is a scaled representation of forecast error that relates the forecast accuracy of a particular forecast model to some reference model (3-12). A perfect forecast results in a forecast skill of one, a forecast with a similar skill to the reference forecast would have a skill of zero, and a forecast which is less skillful than the reference forecast would have negative skill values.

$$SS = 1 - (MSE_{forecast} / MSE_{reference}) \qquad (3\text{-}12)$$

– **Average of Errors** (also referred to as **E-bar 3-13**) is sometimes useful when combining **k** forecasts (or **k** models) have been used (usually as an attempt to reduce the forecast error):

$$\overline{E} = \sum_{i=1}^{k} e_i \qquad (3\text{-}13)$$

where $\mathbf{e_i}$ is the forecast error for a particular technique (or model) **i** (i=1, 2, 3...k). Sometimes, weights could be appropriate if the combining forecast value is not a simple average.

– **The Tracking signal (TS)** is a simple indicator that a forecast bias is present in the forecast model (3-14). When forecasts are being produced on a repetitive basis, the performance of the forecasting system may be monitored using a tracking signal, which provides an automatically maintained summary of the forecasts produced up to any given time period.

It is most often used when the validity of the forecasting model might be in doubt. It monitors any forecasts that have been made in comparison with the actual values and warns when there are unexpected departures of the outcomes from the forecasts. One common form[2] of tracking signal is the ratio of the cumulative sum of forecast errors to the mean absolute deviation **(MAD)**.

$$TS = \sum e_t / MAD \qquad (3\text{-}14)$$

As we can see, there is a variety of measures of the forecast accuracy with different properties. They could be summarised in the following major types of forecast-error metrics:

- absolute errors (MAD and MSE);
- percentage-error metrics (MPE and MAPE);
- relative-error metrics, like NMSE, SS or U coefficient.

It is always a good idea to track the ***forecast bias*** since a normal property of a good forecast is that it is not biased. A typical measure of bias of forecasting procedure is the arithmetic mean (or expected value) of the forecast errors ***MFE*** or ***MPE***, but other measures of bias are also possible. For example, a median-unbiased forecast would be one where half of the forecasts are too low and half too high.

[2] APICS Dictionary, see http://www.apics.org/dictionary/dictionary-information?ID=4390.0

We have already discussed more than a dozen important measures for judging forecast accuracy of a fitted model. Each of these measures has some unique properties, different from others. In experiments, it is better to consider more than one performance criteria. This will help to obtain a reasonable knowledge about the amount, magnitude and direction of overall forecast error. For this reason, experienced forecasters usually use more than one measure for judgment.

Which metrics should be used depends on the particular case and its specific goals. For example, accurate demand forecasts are imperative in order to maintain an optimized inventory and effective supply chain. While forecasts are never perfect, they are necessary to prepare for actual demand. Understanding and predicting customer demand is vital to manufacturers and distributors to avoid stock-outs and maintain adequate inventory levels.

As already mentioned, it is better to consider more than one performance criteria. This will help to obtain a reasonable knowledge about the amount, magnitude and direction of the overall forecast error. For this reason, experienced researchers usually use more than one measure for judgment so MPE, MAPE, RMSE and CV(RMSE) are a very good choice. The main benefit of this group is that it provides good information about the bias and the precision of the forecast. In addition, since CV(RMSE) penalizes extreme errors and MAPE does not, a researcher's goal should be to obtain close enough values for both criteria

Forecast accuracy in the supply chain is typically measured using the MAPE. Many practitioners, however, define and use the MAPE as the Mean Absolute Deviation divided by Average Sales. This is in effect a ***volume weighted MAPE*** and it is also referred to as the ***MAD/Mean ratio***.

A simpler and more elegant method to calculate MAPE across all the products forecasted is to divide the sum of the absolute deviations (e_t) by the total sales of all products. This calculation (Sum $\mathbf{e_t}$)/(Sum $\mathbf{y_t}$) is also known as ***WAPE*** (Weighted Absolute Percent Error). Another interesting option is the ***weighted MAPE***. The advantage of this meassure is that it could weight errors, so we can define how to weight for our relevant business, ex-gross profit or ABC. The problem is that for seasonal products it returns indetermined results when sales equal zero and that is not symetrical, i.e. we can be much more innacurate if sales are higher than if they are lower than the forecast. To correct the above issue, some authors (Makridakis et al., 1998) proposed ***symmetric Mean Absolute Percentage Error (sMAPE)***.

Hyndman (2006) suggests another measure, which he referred to as *scale-free error metric*. He proposed a ***Mean Absolute Scaled Error (MASE)***, which expresses each error as a ratio to an average error from a baseline method:

$$MASE = \sum_{t=1}^{n} |q_t| / n \qquad (4\text{-}15)$$

where $\mathbf{q_t}$ is a scalled error defined as:

$$q_t = \frac{e_t}{\sum_{i=2}^{n} |y_i - y_{i-1}| / (n-1)} \qquad (4\text{-}16)$$

where the numerator $\mathbf{e_t}$ is the forecast error for a given period **t**, and the denominator is the average forecast error of the one-step ***naïve forecast*** method (discussed in Chapter 5), which uses the actual value from the prior period as the forecast, i.e. $\mathbf{F_t = y_{t-1}}$.

According to this study, the scale-free error metric "can be used to compare forecast methods on a single series and also to compare forecast accuracy between series. This metric is well suited to intermittent-demand series, because it never gives infinite or undefined values except in the irrelevant case where all historical data are equal" (Hyndman, 2006, p.46).

The example above shows that there are many attempts to improve the traditional forecast-error metrics, including some adjustments, modifications and even new formulas and criteria. However, in both studies (Makridakis et al., 1998; Hyndman, 2006) and in many others (for example Madala & Ivakhnenko (1994), Mueller & Lemke (2003) and others), there is another extremely important point, which must be emphasized here – the ***cross-validation*** approach.

Cross Validation

To judge the quality of models merely by formal criteria like the closeness of fit of a model and a true system based on one data set is doubtful. Instead, it is necessary to have a purposeful judgment of the quality of the model adaptation based on the suitability of the model to solve a predefined task. In spite of user experience, sometimes the choice of a model is too subjective – to make it more objective is a primary goal in many contemporary studies.

It is important to evaluate forecast accuracy using genuine forecasts. It is not correct to look at how well a model fits the historical data – the accuracy of forecasts can only be determined by considering how well a model performs on new data that were not used when fitting the model. When choosing models, it is common to use a portion of the available data for fitting and use the rest of the data for the model validation. These testing data should be used to measure how well the model is likely to forecast on new data.

In addition to the reasons mentioned above, a few more points should be considered:

- A model which fits the data well does not necessarily forecast well.
- A "*perfect*" fit with zero prediction error can always be obtained by using a model with large enough number of parameters (as shown in Fig.3-3).

- ***Overfitting*** a model to data is as bad as failing to identify the systematic pattern in these data (see Section 3.4).

In most cases, it is not possible to select an optimal model from many possible models without some extra information. In (Makridakis et al., 1998) and (Hyndman, 2006) this process is referred to as ***out-of-sample forecasting***. In Madala & Ivakhnenko (1994) and Mueller & Lemke (2003) the term ***external information*** is used, referring to Gödel's ***external complement*** (Gödel, 1931). In general, it is known as ***cross-validation*** or ***rotation estimation*** (Stone, 1974).

Cross-validation is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent dataset. It is mainly used in studies where the goal is prediction, and we want to estimate how accurately a predictive model will perform in practice. In a prediction problem, a model is usually given a dataset of known data on which training is run (***training dataset***), and a dataset of unknown data (or first seen data) against which the model is tested (***testing or validating dataset***).

The size of the testing set is typically about 20%-30% of the total sample, although this value depends on how long the sample is and how far ahead, we want to forecast. It should ideally be at least as large as the maximum forecast horizon required.

There is another, more sophisticated version of training/testing sets in cross-validation. For cross-sectional data it works as follows:

1. Select (it could be random) observation ***i*** for the testing set and use the remaining observations in the training set. Compute the error on the test observation.
2. Repeat the above step for ***i = 1, 2, … N-1,*** where ***N*** is the total number of observations.
3. Compute the forecast accuracy measures based on all errors obtained.

This is a much more efficient use of the available data, as we only omit one observation at each step. However, it can be very time consuming to implement.

For time-series data, the procedure is very similar and will be explained in Chapter 8. Cross-validation applications in model building and model selection are discussed in the next Sections.

4.3. Forecasting Techniques Evaluation and Model Selection

A. Forecast evaluation

To determine the value of a forecast we need to measure it against some baseline, or a minimally accurate reference forecast. There are many types of forecast that, while producing impressive-looking skill scores, are nonetheless naïve. For example, a *persistence*[3] forecast can still rival even those of the most sophisticated models.

[3] ***Persistence forecast*** in a nutshell: Q.What are the sales going to be like today? A.Same as they were yesterday.

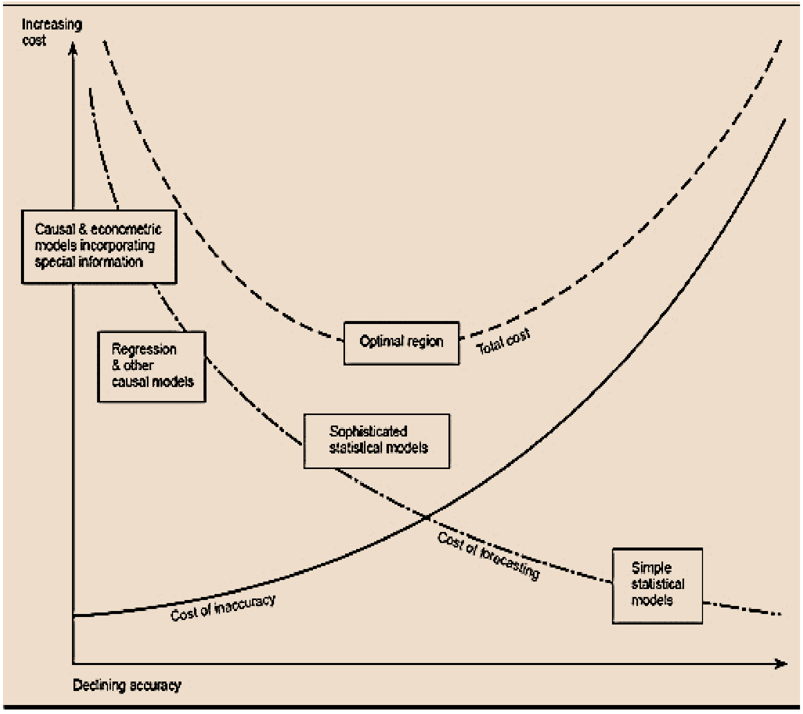


Fig.3-1. Cost of Forecasting Versus Cost of Inaccuracy for a Medium-Range Forecast, Given Data Availability

(Source: John C. Chambers, Satinder K. Mullick, Donald D. Smith, How to Choose the Right Forecasting Technique, Harvard Business Review, July 1971)

Two of the most important factors used to evaluate a forecast in the real life are the *Forecast Cost* and the *Forecast Accuracy*. Additional factors include the availability of historical data, computational power and relevant software, time to gather and analyze the data, forecast horizon and others. The example in Fig. 3-1 shows how cost and accuracy increase with sophistication and charts this against the corresponding cost of forecasting errors, given some general assumptions. The most sophisticated technique that can be economically justified is one that falls in the region where the sum of the two costs is minimal.

The variety of measures of forecast accuracy introduced in the previous section have different properties and could be used for different purposes. Though each case is particular and has its specific goals, there are some general rules of using the measures of forecast accuracy:

- To measure a forecast's usefulness or reliability, most frequently the researchers use MAD/MAE, MSE/RMSE, MPE, and TS.
- To compare the accuracy of two different techniques, the most common measures are MAPE, NMSE, the RMSE normalized value CV(RMSE), and SS.
- There are some important points that need clarification as well, in terms of the specific properties of the forecasting technique used, like the validity of the technique assumptions and the significance of the model parameter estimations.
- Lastly, it is also important if the forecasting technique is simple to use and easy to understand for decision makers.

It should be noted that despite the fact that forecasts are based on models, the use of models does not guarantee a good forecast. For example, nonqualified users cannot comprehend the rules for using the model, or may incorrectly apply it and misinterpret the results. Also, as mentioned above, unfortunately, no single technique/model works in every situation and selecting the most appropriate (cost-effective) one among many similar models is a never-ending task in contemporary forecasting.

B. Model selection

Generally, ***model selection*** is defined as the task of selecting a good model from a set of candidate models, given data. In simple cases, a pre-existing set of data is considered. However, the task can also involve the design of experiments so that the data collected is well-suited to the problem of model selection. Usually, among the candidate models of similar predictive and/or explanatory power, the simplest model is most likely to be the best choice.

Model selection is very important and in its most basic formis one of the fundamental tasks of scientific inquiry. A standard example of model selection is that of curve fitting, whereas a set of observations and other background knowledge is given , we must select a curve that describes the function that generated these values. Determining the principle that explains a series of observations is often linked directly to a model predicting those observations. One classical example is when Galileo performed his inclined plane experiments, in fact, he demonstrated that the motion of the balls fitted the parabola predicted by his model.

Model selection process typically begins defining the set of candidate models. Often simple models such as polynomials are used, at least initially. Once the set of candidate models has been chosen, there are different techniques, mostly from mathematical analysis, that allow to select the best of these models. What is meant by best is controversial – a good model selection technique will balance ***goodness of fit***[4] with simplicity. Sometimes, more complex models will be better able to adapt their shape to fit the data (for example, a fifth-order polynomial can exactly fit six data values), but the additional parameters may not represent anything useful. It is possible that those six observations are just randomly distributed about a straight line.

There are two major groups of methods for model selection:

A. The ***scientific method*** is a body of techniques for investigating phenomena, acquiring new knowledge, or correcting and integrating previous knowledge. To be termed scientific, a method of inquiry must be based on empirical and measurable evidence subject to specific

[4] The ***goodness of fit*** describes how well the model fits a set of observations. Measures of goodness of fit typically summarize the discrepancy between observed values and the values expected under the model in question.

principles of reasoning. The Oxford Dictionary defines the scientific method as "a method or procedure that has characterized natural science since the 17th century, consisting in systematic observation, measurement, and experiment, and the formulation, testing, and modification of hypotheses."[5]

The main characteristic which distinguishes the scientific method from other methods of acquiring knowledge is that scientists seek to let reality speak for itself, supporting a theory when a theory's predictions are confirmed and challenging a theory when its predictions prove false. The essential elements of scientific method are:

- Operation – Some action done to the system being investigated;
- Observation – What happens when the operation is done to the system;
- Model – An abstraction or the phenomenon itself at a certain moment;
- Utility Function for evaluating models – A measure of the usefulness of the model to explain, predict, and control, and of the cost of use of it. Some specific elements of any scientific utility function are the refutability[6] of the model, its simplicity and so on.

B. ***The Statistical hypothesis test*** is a method of statistical inference using different data sets from a scientific study or real-life observations, usually results from an experiment or a probabilistic sample. A result is called statistically significant if it has been predicted as unlikely to have occurred by chance alone, according to a pre-determined threshold probability, known as the ***significance level (α).*** Statistical hypothesis testing is sometimes called ***confirmatory data analysis***, in contrast to ***exploratory data analysis***, which may not have pre-specified hypotheses. ***Exploratory Data Analysis (EDA)*** is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily ***EDA*** is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. ***EDA*** is also different from the ***InitialData Analysis (IDA)*** (see Chapter 10) – ***EDA*** encompasses ***IDA***, which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed.

In both groups of methods, briefly discussed above, the model selection task involves a hypothesis test about the model significance. The most common measures used as criteria for decision are ***MAPE*** and ***SS***. There are some other specific criteria for model selection like

[5] See http://www.oxforddictionaries.com/definition/english/scientific-method?q=scientific+method

[6] Refute - to prove something to be false or somebody to be in error, either through logical argument or by providing evidence to the contrary

Akaike information criterion, *Bayesian information criterion*, *Deviance information criterion* and *Mallows's C_p*.[7]

Akaike information criterion (AIC) is a measure of the relative quality of a statistical model for a given set of data. *It deals with the trade-off between the goodness of fit of the model and the complexity of the model.* Founded on information theory, it offers a relative estimate of the information lost when a given model is used to represent the process that generates the data. The AIC penalizes the number of parameters less strongly than the ***Bayesian information criterion (BIC)*** does . Yang (2005) makes a comparison of ***AIC*** and ***BIC*** in the context of regression modeling. In particular, ***AIC*** is asymptotically optimal in selecting the model with the least mean squared error, under the assumption that the exact "*true*" model is not in the candidate set[8] and ***BIC*** is not. Yang further shows that the rate at which ***AIC*** converges to the optimum is, in a certain sense, the best possible.

It should be noted, however, that ***AIC*** does not provide a test of a model in the sense of testing a hypothesis; i.e. ***AIC*** directly can tell nothing about the quality of the model in an absolute sense. If all the candidate models fit poorly, ***AIC*** will not give any warning of that.

In a research done by Stone (1977) a logarithmic assessment of the performance of a predicting density is found to lead to an asymptotic equivalence of choice of model by ***cross-validation*** and ***AIC*** when maximum likelihood estimation is used within each model. It also shows, that maximizing ***AIC*** is equivalent to minimizing ***Mallows' C_p*** in the case of normal multiple linear regression.

Eventually, the importance and the qualities of ***cross-validation*** as a model selection tool could be proved by the latest sophisticated techniques such as machine learning and, later on, data mining where ***cross-validation*** is used to compare the performances of different predictive modeling procedures, for example in optical character recognition, support vector machines (SVM), k-nearest neighbors (KNN) and others, discussed in Chapter 11.

Model selection is a long procedure involving a lot of computations, especially in cases where there is a large number of potential explanatory variables and no underlying theory on which to base the model selection. To address this problem, different automatic procedures were elaborated since the middle of the 20th century.

[7] See http://en.wikipedia.org/wiki/Model_selection

[8] This makes both AIC and BIC sensitive to the type of asymptotic analysis adopted (Stone, M. 1979. Comments on Model Selection Criteria of Akaike and Schwarz. *Journal of the Royal Statistical Society*. Ser. B, Vol. 41, pp. 276-278).

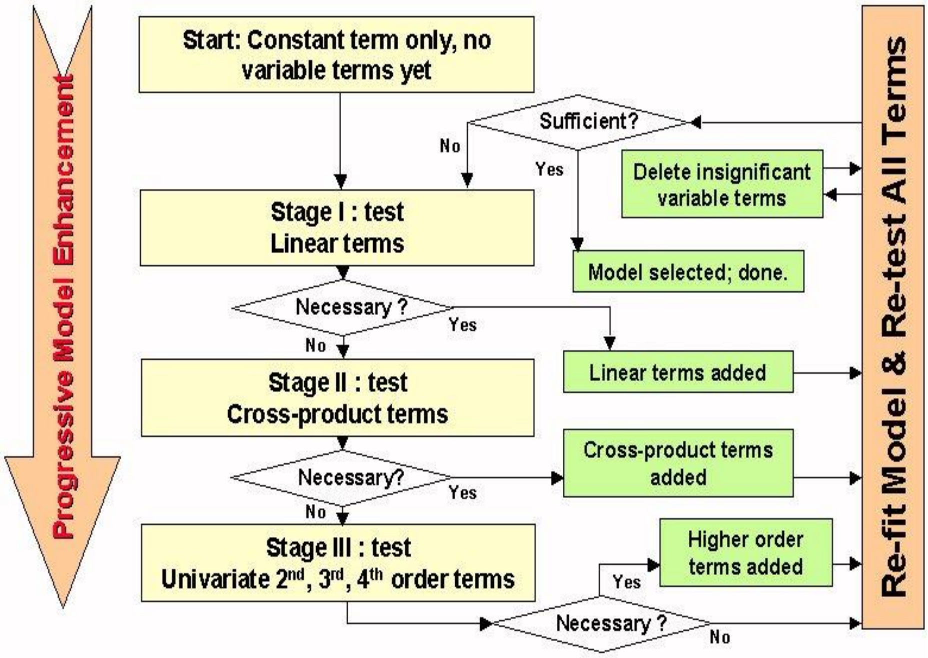

Progressive Model Enhancement
Start: Constant term only, no variable terms yet
Sufficient?
No
Yes
Delete insignificant variable terms
Stage I : test Linear terms
Model selected; done.
Necessary ?
Yes
No
Linear terms added
Stage II : test Cross-product terms
Necessary?
Yes
No
Cross-product terms added
Stage III : test Univariate 2nd, 3rd, 4th order terms
Higher order terms added
Yes
Necessary ?
No
Re-fit Model & Re-test All Terms

Fig.3-2 Example of stepwise regression

Stepwise regression is one of the first and the most popular techniques for regression model building in which the choice of predictive variables is carried out by an automatic procedure (Draper & Smith, 1981). Usually, this takes the form of a sequence of F-tests or t-tests, but other criteria could be used as well, such as ***adjusted R-square***[9], ***AIC***, ***BIC***, and others.

In the example shown in Fig.3-2 necessity and sufficiency are determined by F-tests. For additional consideration, when planning an experiment, computer simulation, or scientific survey to collect data for the model, we must keep in mind the number of parameters (**p**) to estimate and adjust the sample size accordingly. In the past, an efficient design of experiments, which explores the relationships between several explanatory variables (**k**) and one or more dependent variables existed only for **k<17**. Nowadays, there are more efficient designs, in particular in Data Mining, discussed in Chapter 11, requiring fewer observations, even for **k>16**.

Another main issue with the stepwise regression is that it searches a large space of possible models. Hence it is prone to overfitting the data, i.e. stepwise regression will often fit much better in-sample than it does on new out-of-sample data. It means that the tests used would be biased since they are based on the same data. Wilkinson and Dallal (1981) computed percentage points of the multiple correlation coefficients by simulation and showed that a final model, obtained by forward selection, said by the F-procedure to be significant at 0.1%, was in fact only significant at 5%.

[9] The coefficient of determination, denoted $\boldsymbol{R^2}$ or $\boldsymbol{r^2}$ and pronounced ***R-squared***, is a number that indicates how well data fit a statistical model. ***The adjusted*** $\boldsymbol{R^2}$ is almost the same as $\boldsymbol{R^2}$, but it penalizes the statistic as extra variables are included into the model (see Chapter 7).

Other points of criticism have also been made: Hurvich and Tsai (1990) found that when estimating the degrees of freedom, the number of the candidate independent variables from the best fit selected is smaller than the total number of final model variables, causing the fit to appear better than it is when adjusting the $\boldsymbol{r^2}$ value for the number of degrees of freedom (i.e. it is important to consider how many degrees of freedom have been used in the entire model, not just count the number of independent variables in the resulting fit). Roecker (1991) mentioned that models which are created may be over-simplifications of the real models of the data.

Mark and Goldberg (2001) proposed a technique to test for errors in models created by step-wise regression. It is not relying on the model's F-statistic, its significance, or multiple R. Instead the technique assesses the model against a set of data that was not used to build the model (like in cross-validation). This is often done by developing a model based on a sample of the dataset available (e.g., 70%) and using the remaining 30% of the dataset to assess the accuracy of the model. Accuracy is then measured as the actual MSE, MAPE, or mean error between the predicted value and the actual value in the hold-out sample. This method is particularly valuable when data is collected in different settings (e.g., different times, social vs. solitary situations) or when models are assumed to be generalizable.

This would address some issues with the step-wise regression, but unfortunately others still remain unsorted, like ***multicollinearity***, when two or more predictor variables in the model are highly correlated; ***autocorrelation***, when there is correlation between values of the process at different time points as a function of the time points or of the time lags, and so on. In the next section, a new, better approach in modeling will be discussed.

3.4. Self-Organizing Modeling

The concept of self-organization is central to the description of biological systems, from the subcellular to the ecosystem level. There are also cited examples of "self-organizing" behavior found both in the natural sciences and the social sciences such as economics or anthropology. The term "***self-organizing***" was introduced to contemporary science in 1947 by the psychiatrist and engineer William Ross Ashby (1947). It was taken up by other world recognized scientists like Heinz von Foerster, Gordon Pask, Stafford Beer and Norbert Wiener (1948) himself. In cybernetics, Norbert Wiener saw the first step in Self-organization as being able to copy the output behavior of a black box.

Self-organization as a word and concept was used by scientists associated with general systems theory in the 1960s, but did not become commonplace in scientific literature until its adoption by physicists and researchers in the field of complex systems in the 1970s and 1980s.

After 1977's Ilya Prigogine (1977) Nobel Prize, the *thermodynamic concept of self-organization* received some attention from the public.

According to (Camazine et al., 2003, p. 8), ”*In biological systems self-organization is a process in which pattern at the global level of a system emerges solely from numerous interactions among the lower-level components of the system. Moreover, the rules specifying interactions among the system's components are executed using only local information, without reference to the global pattern.”*

Self-organizing systems typically display emergent properties. In general, rules of behavior of a *self-organizing system* are determined internally but modified by environmental inputs (Madala & Ivakhnenko, 1994), which is the basic idea in the ***cross-validation*** approach as well.

Many scientists (Madala & Ivakhnenko, 1994; Mueller & Lemke 2003; Holland et al. 1986) consider the global view of *problem-solving* as a process of search through a state space – a problem is defined by an initial state. There are one or more goal states to be reached, a set of operators that can transform one state into another, and a number of constraints, that an acceptable solution should meet. Problem-solving techniques are used for selecting an appropriate sequence of operators that will succeed in transforming the initial state into a goal state through a *series of steps*, known as ***multi-stage (or multi-layered) selection procedure*** (Fig.3-3). A *selection approach* is taken on classifying the systems. This is based on an attempt to impose rules of "*survival of the fittest*" on an ensemble of simple productions.

This group is further enhanced by criterion rules which implement processes of *genetic cross-over* and *mutation* on the productions in the population. Thus, productions that survive a process of selection are not only applied but also used as "*parents*" in the synthesis of new productions. Here an "***external agent***" is required to play a role in laying out the basic architecture of those productions upon which both selective and genetic operations are performed.

These classification systems do not require any *a priori* knowledge of the categories to be identified – the knowledge is very much implicit in the structure of the productions, i.e. it is assumed as the *a priori* categorical knowledge, embedded in the classifying systems. The concepts of "*natural selection*" and "*genetic evolutions*" are viewed as a possible approach to normal levels of implementation of rules and representations in information processing models.

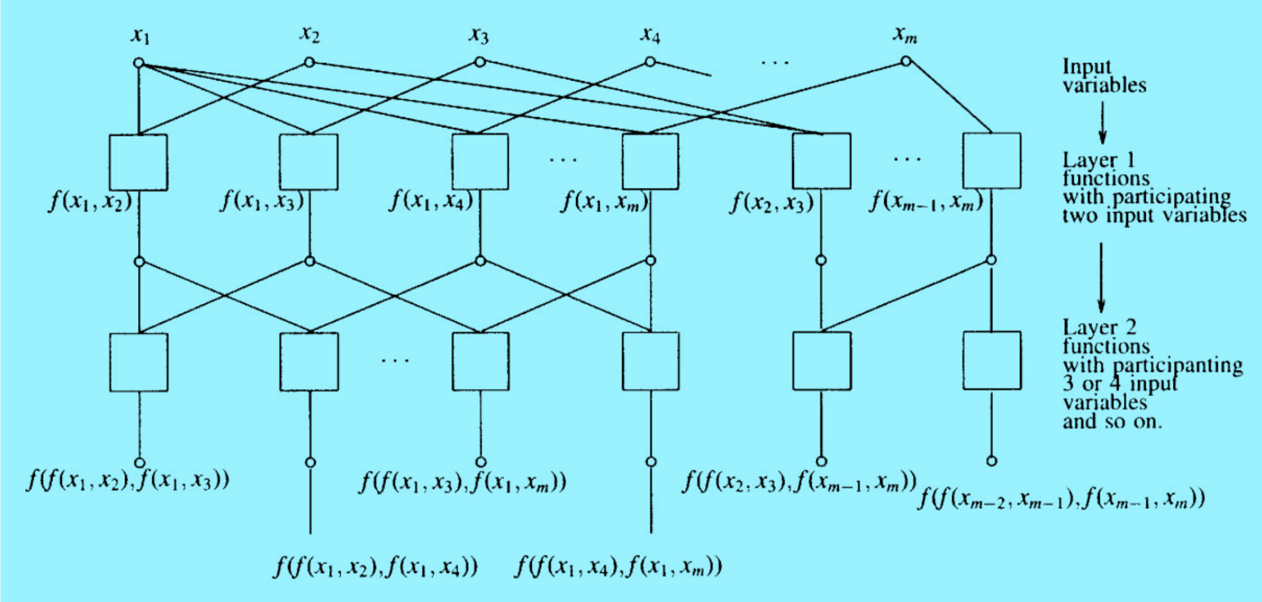

x_1
x_2
x_3
x_4
. . .
x_m
Input variables
Layer 1 functions with participating two input variables
$f(x_1, x_2)$
$f(x_1, x_3)$
$f(x_1, x_4)$
$f(x_1, x_m)$
$f(x_2, x_3)$
$f(x_{m-1}, x_m)$
Layer 2 functions with particippanting 3 or 4 input variables and so on.
$f(f(x_1, x_2), f(x_1, x_3))$
$f(f(x_1, x_3), f(x_1, x_m))$
$f(f(x_2, x_3), f(x_{m-1}, x_m))$
$f(f(x_{m-2}, x_{m-1}), f(x_{m-1}, x_m))$
$f(f(x_1, x_2), f(x_1, x_4))$
$f(f(x_1, x_4), f(x_1, x_m))$

Fig.3-3 Example of a Self-Organizing Multilayered Procedure
(Source: Madala & Ivakhnenko, 1994, p. 8)

Simplification of self-organization is regarded as its fundamental problem from the very beginning of its development. The modeling methods created for the last few decades based on the concepts of neural and inductive computing ensure the solution of comprehensive problems of complex systems modeling as applied to cybernetic systems (Madala & Ivakhnenko, 1994; Mueller & Lemke, 2003). They constitute an arsenal of means by which—either on the basis of notions concerning system structures and the processes occurring in them or on the basis of observations of the parameters of these systems—we can construct system models that are accessible for direct analysis and are intended for practical use.

One of the fundamentals of self-organizing modeling is the ***external agent***, which is equivalent to the ***regularization method***[10]. Gödel (1931) published his works on mathematical logic showing that the axiomatic method itself had inherent limitations and that the principal shortcoming was the so-called inappropriate choice of "***external supplement***" (or ***external complement***):

- The first incompleteness theorem states that no consistent system of axioms whose theorems can be listed by an *effective procedure* (essentially, a computer program) is capable of proving all facts about the natural numbers. For any such system, there will always be statements about the natural numbers that are true, but which are improvable within the system.

[10] ***Regularization***, in mathematics and statistics and particularly in the fields of machine learning, refers to a process of introducing additional information in order to solve an ill-posed problem or to prevent overfitting.

- The second incompleteness theorem shows that if such a system is also capable of proving certain basic facts about the natural numbers, then one particular arithmetic truth the system cannot prove is the consistency of the system itself.

According to the incompleteness theorem, as discussed by (von Neumann, 1966), it is in principle impossible to find a unique model of an object on the basis of empirical data without using an ***external supplement***, as defined by Beer (1959, p. 280). Hence ***external supplement***, ***external complement, external agent, regularization,*** and ***cross-validation*** are synonyms expressing the same concept.

For example, as mentioned in the previous section, in regression analysis (see Chapter 7) the errors ε_i (***i=1:N***) computed on the basis of all (***N***) observations monotonically decrease when the model complexity gradually increases (Fig.3-4). It equals zero when the number of coefficients ***m*** of the model becomes equal to the number of observations ***N*** or the degrees of freedom equal to zero (d.f.=0). It means that every equation with ***m*** coefficients can be regarded as a perfect model. It is impossible, in principle, to find a unique model in such a situation. Usually experienced modelers use trial and error techniques to find a unique model without claiming (or sometimes even understanding) that they consciously or unconsciously use an ***external supplement*** to build a unique model. The risk involved in the trial and error methods is that the researcher could not appropriately select the ***external agent***.

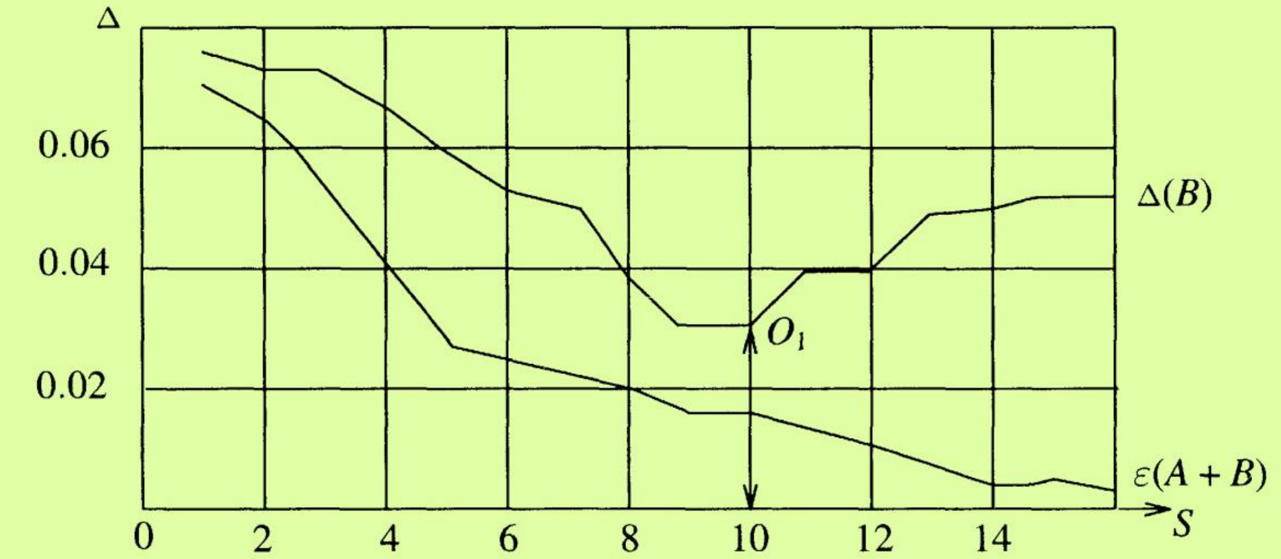
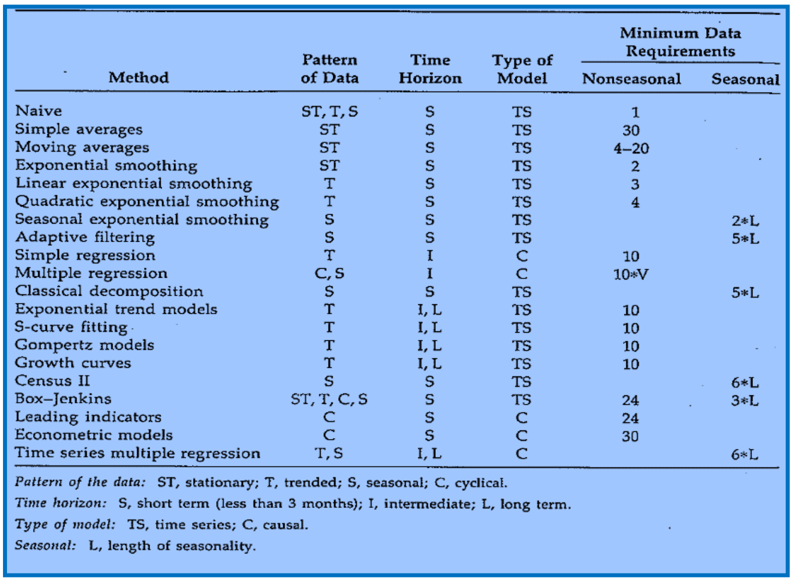


Fig.3-4 Variation in least square error $\varepsilon(A + B)$ and error measure of an "external complement" $\Delta(B)$ for a regression equation of increasing complexity S; O_1 is the model of optimal complexity

(Source: Madala & Ivakhnenko, 1994, p.11])



Method	Pattern of Data	Time Horizon	Type of Model	Minimum Data Requirements	
				Nonseasonal	Seasonal
Naive	ST, T, S	S	TS	1	
Simple averages	ST	S	TS	30	
Moving averages	ST	S	TS	4–20	
Exponential smoothing	ST	S	TS	2	
Linear exponential smoothing	T	S	TS	3	
Quadratic exponential smoothing	T	S	TS	4	
Seasonal exponential smoothing	S	S	TS		2*L
Adaptive filtering	S	S	TS		5*L
Simple regression	T	I	C	10	
Multiple regression	C, S	I	C	10*V	
Classical decomposition	S	S	TS		5*L
Exponential trend models	T	I, L	TS	10	
S-curve fitting	T	I, L	TS	10	
Gompertz models	T	I, L	TS	10	
Growth curves	T	I, L	TS	10	
Census II	S	S	TS		6*L
Box–Jenkins	ST, T, C, S	S	TS	24	3*L
Leading indicators	C	S	C	24	
Econometric models	C	S	C	30	
Time series multiple regression	T, S	I, L	C		6*L

Pattern of the data: ST, stationary; T, trended; S, seasonal; C, cyclical.

Time horizon: S, short term (less than 3 months); I, intermediate; L, long term.

Type of model: TS, time series; C, causal.

Seasonal: L, length of seasonality.

Fig.3-5 Choosing a Forecasting Technique and a Model

To support nonqualified users, who cannot comprehend the rules on how to build and select the right model, experienced researchers design special charts and/or tables with instructions for the most appropriate modeling technique according to data patterns or properties of the system of interest, such as the guidelines (Wilson & Keating, 2008) presented in Fig.3-5.

The principle of self-organization can be formulated as follows: *When the model complexity gradually increases, certain criteria, which are called* ***selection criteria*** *or objective functions, and which have the property of* ***external agent*** *pass through a minimum. Achievement of a* ***global minimum*** *indicates the existence of a* ***model of optimum complexity*** (Fig.3-4).

The statement that there exists a unique model of optimum complexity, determinable by the self-organization principle, forms the basis of the so-called inductive approach in modeling (Madala & Ivakhnenko, 1994). The optimum complexity of the mathematical model of a complex system is found by the minimum of a chosen objective function (or criteria) which has the properties of an ***external supplement***.

The theory of self-organization has widened the capabilities of system identification, forecasting, pattern recognition and multi-criteria decision making. In fact, self-organizing modeling provides a new, third view (also known as "***hybrid approach***") to the ***theory-driven*** (or ***theoretical systems analysis)*** and ***data-driven*** (or ***experimental systems analysis***) ***approaches*** in the model building problem "what variables, which method, which model." This idea is also an important part of self-organizing data mining (Mueller & Lemke, 2003) discussed in Ch.12. It is outlined below, and particular details are explained and discussed in Chapters 7, 8, 9, 10, 11 and 12:

A. *Variables selection*

Choice of appropriate variables is one part of data pre-processing (see Chapter 10) and is strongly connected to the problem identification. It is important here to find out the most relevant variables for a specific modeling task from a large set of available variables, which in turn forms the data source for modeling. Along with some solid theoretical knowledge (if available) on which variables are useful for the particular case, we suggest applying self-organizing algorithms (see (Madala & Ivakhnenko, 1994; Mueller & Lemke, 2003) and Chapters 8, 9 and 12), or at least correlation analysis with cross-validation, at this stage of modeling to select all significant variables or classes of variables more objectively.

B. *Method Selection*

The answer to this question is also task dependent. However, it is object dependent as well. It means that one and the same modeling task (prediction for example) may require different modeling methods for different systems. Successful modeling requires adequateness between object and model concerning their fuzziness, descriptive and predictive power. In other words, a modeling method needs to reflect an object's uncertainty appropriately. For example, regression is a very good tool to predict almost certain processes like many technical systems. However, one can hardly expect that regression models will do well on fuzzy processes that are sometimes even contradictory with themselves. Common examples are financial markets (Mueller & Lemke 2003).

There are several automated inductive modeling methods (Madala & Ivakhnenko, 1994), based on cybernetic principles of self-organization, which have been developed in recent decades. On their basis many particular algorithms appropriate for specific modeling tasks and work at different levels of objects' uncertainty were designed. Some of them, directly related to forecasting, will be discussed in Chapters 8, 9 and 12 when analyzing the specific problems.

C. *Model Selection*

The key fact here is that any model created by any method can only reflect a method-specific and therefore also freedom-of-choice dependent subset of the true behavior of the system. This is because any model can be created only for a section of the real-world's infinity, and it is the main reason why different models predict better at different times.

One solution to increase robustness and reliability of prediction is creating several different models, first, using different modeling methods. Since it is impossible, however, to know in advance which model will produce the best forecast for actual conditions, we suggest a synthesis of all models by a hybrid solution. It can be expected that this collective solution reflects reality more thorough than any single model can do over a period of time. Self-organization modeling, fast and powerful computer hardware platforms, and advanced software technologies make this approach realistic today. ***Group Method of Data Handling (GMDH)*** (Madala & Ivakhnenko, 1994) is its first real-life application.

Based on multiple application of self-organizing modeling (Madala & Ivakhnenko, 1994; Mueller & Lemke, 2003; Motzev, 2011), it is possible and also reasonable to extract reliable knowledge from data more readily, more universal, and more robust that makes forecasting something different from fortune-telling. Systems theory and practice have been showing that predictive controlled systems are less disturbance-sensitive, more stable and have more adaptive behavior than systems which work on historical and actual information only. In this way, many extreme and often very costly situations can be avoided.

It should be noted that some researchers show a few limitations of the ***Cross-validation*** approach. The most important is that it only yields meaningful results if the validation set and training set are drawn from the same population. In many applications of predictive modeling, the structure of the system being studied evolves over time. This can introduce systematic differences between the training and validation sets. One example pointed out is that if a model for predicting stock values is trained on data for a certain five-year period, it is unrealistic to treat the subsequent five-year period as a draw from the same population.

This is true sometimes and was mentioned already in Chapter 1 as a feature common to most forecasts. Forecasting techniques generally assume that ***the same underlying causal system that existed in the past will persist into the future,*** which could be a very strong restriction especially in long-term forecasting with time series data sets, when most elements and relationships between them are dynamic and change. It is not a real problem if managers stay alert to such occurrences and are ready to override forecasts, which assume a stable causal system. In Chapters 9 and 12 we will also discuss how to improve the forecast in such a case using ***dynamic coefficients***.

It is more dangerous when cross-validation is misused. If it is misused and a true validation study is subsequently performed, the prediction errors in the true validation are likely to be much worse than would be expected based on the results of cross-validation. Here are some examples of how cross-validation can be misused:

- By performing an initial analysis to identify the most informative features using the entire data set – if feature selection or model tuning is required by the modeling procedure, this must be repeated on every training set. If cross-validation is used to decide which features to use, an inner cross-validation to carry out the feature selection on every training set must be performed.

- By allowing some of the training data to be also included in the test set – this can happen due to "twinning" in the data set, whereby some exactly identical or nearly identical samples are present in the dataset.

If carried out properly, and if both the validation and the training sets are from the same population, cross-validation is nearly unbiased. In combination with the other fundamentals of self-organizing modeling, similar procedures (like for example ***GMDH***) have become one of the most powerful tools in model building and forecasting. As mentioned above, important applications will be discussed in Chapters 8, 9 and 12.

***

Summary and Conclusions

Chapter 4 discusses some of the most important topics in forecasting – ***forecast error and accuracy, as well as forecasting techniques evaluation and model selection.***

- It is important that the ***degree of accuracy*** should be clearly stated in the beginning of each particular forecasting process. This will enable users to plan for possible errors and will provide a basis for comparing alternative forecasts.
- By default, the difference between the actual value and the forecast value for the corresponding period is referred to as the ***forecast error.***
- ***Mean Absolute Deviation (MAD)*** or ***Mean Absolute Error***, is the average absolute value of the differences between the actual value at period **t** and the forecast for period **t**. It is most useful to measure the forecast error in the same units as the original series, but it is not sensitive to extreme values – it weights errors linearly.
- ***Mean squared error (MSE)*** or Mean Squared Prediction Error (MSPE) provides a penalty for large forecasting errors (it squares each). It is the average squared value of the differences between the actual value and the forecast at period **t.**
- ***Mean Absolute Percentage Error (MAPE)*** puts errors in perspective. It is useful when the size of the forecast variable is important in evaluating. It provides an indication of how large the forecast errors are in comparison to the actual values.
- ***Mean Percentage Error (MPE)*** is useful when it is necessary to determine whether a forecasting method is biased. If the forecast is unbiased MPE will produce a value that is close to zero. Large negative values mean overestimating and large positive values indicate that the method is consistently underestimating. A disadvantage of this measure is that it is undefined whenever a single actual value is zero.
- A ***forecast bias*** occurs when there are consistent differences between actual outcomes and previously generated forecasts of those quantities, i.e. forecasts may have a general tendency to be too high or too low.
- ***The forecast skill*** or ***skill score – SS*** is a scaled representation of forecast error that relates the forecast accuracy of a particular forecast model to some reference model. A perfect forecast results in a forecast skill of one, a forecast with similar skill to the reference forecast would have a skill of zero, and a forecast which is less skillful than the reference forecast would have negative skill values.
- The ***tracking signal (TS)*** is a simple indicator that forecast bias is present in the forecast model. It monitors any forecasts that have been made in comparison with the

actual values, and warns when there are unexpected departures of the outcomes from the forecasts. One common form of tracking signal is the ratio of the cumulative sum of forecast errors to the mean absolute deviation **(MAD)**.

- The variety of measures of the forecast accuracy have different properties. They could be summarised in the following major types of forecast-error metrics: absolute errors (MAD and MSE); percentage-errors (MPE and MAPE) and relative-error metrics, like SS.
- ***Cross-validation*** is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. A model is usually a given dataset of known data on which training is run (***training dataset***), and a dataset of unknown data (or first seen data) against which the model is tested (***testing dataset***).
- To determine the value of a forecast we need to measure it against some baseline, or minimally accurate reference forecast. Two of the most important factors used to evaluate a forecast in the real life are the forecast cost and the forecast accuracy. Additional factors include the availability of historical data, computers and relevant software, time needed to gather and analyze the data, forecast horizon and others.
- General rules in measuring ***forecast accuracy***: *to measure a forecast's usefulness or reliability* use *MAD/MAE, MSE, MPE and TS; to compare the accuracy of two different techniques – MAPE and SS*; important points that need clarification as well, in terms of the *specific properties* of the forecasting technique used are the validity of the technique assumptions, significance of the model parameter estimations and so on; it is also important if the forecasting technique is *simple to use and easy to understand* for decision makers.
- ***Model selection*** is the task of selecting a good model from a list of candidate models, generated from a given dataset.
- ***Goodness of fit*** describes how well the model fits a set of observations. Measures of goodness of fit typically summarize the discrepancy between observed values and the values expected under the model in question.
- ***Stepwise regression*** is one of the first and the most popular techniques for regression models building in which the choice of predictive variables is carried out by an automatic procedure. Usually, this takes the form of a sequence of F-tests or t-tests, but other criteria could be used as well, such as adjusted R-square, ***AIC***, ***BIC*** and others.

- ***Cross-validation*** could address some issues with the stepwise regression, but others still remain, like ***multicollinearity***, when two or more predictor variables in the model are highly correlated; ***autocorrelation***, when there is correlation between values of the process at different time points as a function of the time points or of the time lags, and so on.
- The principle of self-organization states: *When the model complexity gradually increases, certain criteria, which are called* ***selection criteria*** *or objective functions and which have the property of* ***external agent*** *pass through a minimum. Achievement of a* ***global minimum*** *indicates the existence of a* ***model of optimum complexity.***
- ***External supplement***, ***external complement, external agent, regularization*** and ***cross-validation*** are synonyms expressing the same concept.
- If carried out properly, and if the validation set and training set are from the same population, cross-validation is nearly unbiased. Combined with the other fundamentals of self-organization such procedures are among the most powerful tools in model building and forecasting. Very important applications will be discussed in Ch. 8, 9 and 12 accordingly.

KEY TERMS

Accuracy	***67***	***Adjusted R-square***	***82***
Akaike information criterion (AIC)	***81***	***Average of Errors***	***74***
Bayesian information criterion (BIC)	***81***	***Coefficient of Variation of the RMSE, CV(RMSE)***	***73***
Cross-validation (rotation estimation)	***76, 77, 81, 84, 86, 89***	***Dynamic coefficients***	***89***
Exploratory Data Analysis (EDA)	***80***	***External supplement (complement)***, ***External agent, Regularization***	***84, 85, 86***
Forecast bias	***69, 74***	***Forecast cost & accuracy***	***77***
Forecast error	***68***	***Forecast Skill (Skill Score – SS)***	***73***
Goodness of fit	***79***	***Group Method of Data Handling (GMDH)***	***89***
Hybrid approach in model building	***87***	***InitialData Analysis (IDA)***	***80***
Mean Absolute Deviation (MAD)		***Mean Absolute Percentage Error (MAPE)***	***70***
Mean Absolute Error (MAE)	***70***		
Mean Absolute Scaled Error (MASE)	***75, 76***	***Mean Forecast Error (MFE)***	***69***
Mean Percentage Error (MPE)	***71***	***Mean Squared Error (MSE)***	***71***
Model selection	***79***	***Multicollinearity, autocorrelation***	***83***
Multi-stage (Multi-layered) selection procedure	***84***	***Naïve forecast***	***76***

CHAPTER EXERCISES

Conceptual Questions:

1. What is the definition of forecast error? Explain.
2. Why do researchers split the dataset into training and testing datasets? What is the name of this technique? Discuss.
3. How should one evaluate a forecast? What are the most important factors? Discuss and illustrate with examples.
4. What are the pros and cons of stepwise regression? Discuss.
5. What are the fundamentals of self-organizing modeling? List and discuss at least three of them.

Business Applications:

The manager of *Sweet Onion Inc.* ordered weather forecast for the next calendar year. The report he received (file SweetOnion.xlsx) contains the results from three different forecasting techniques. For each of them, the report provides computed values for MAD, MSE, MAPE, and MPE, calculated on a testing dataset. The manager wants to evaluate forecast's usefulness and reliability and compare the accuracy of all three forecasts in order to select the "best" of them:

- Is there any biased forecast? What measure will identify it?
- How to determine if there is a forecasting technique, which produced to many extreme values? (Hint: compare measures that weight errors in a different way).
- Compare the accuracy of all different techniques. What measure should be used?
- Summarize all findings and select the "best" forecast. Explain your decision.

Write a short report (up to two pages) discussing your answers.

INTEGRATIVE CASE

HEALTHY FOOD SYPPLY CHAIN & STORES

Part 3: Measuring Forecast Accuracy

In Chapter 1 we introduced *Healthy Food Stores* – a fast-growing retail food provider with 12 stores in a northwestern state. The company executives decided to study the effect that company advertising dollars have on sales. They hoped that examining collected historical data would reveal relationships that would help determine future advertising expenditures and predict monthly sales volumes for the upcoming quarter.

After identifying basic parameters, along with input (independent) and output (dependent) variables of the forecasting scenario in Part 2, the research team determined the main elements of the forecasting process:

- Forecasting horizon of up to twelve months;
- Quarterly forecast updates, since accuracy decreases as time horizon increases, and sufficient time is needed to implement possible changes;
- Development of different models based on data patterns, if any, and evaluation of their accuracy in order to select the most appropriate one;
- Selection of the best forecasting model with no more than 5% forecasting error.

Since historical data had already been collected, the research team decided to develop different models and to evaluate their accuracy in order to select the most appropriate and accurate one to be used in computing future sales volumes.

The research team knew that in order to determine the value of each potential forecast they needed to measure each one against a baseline (i.e. minimally accurate) reference forecast. They decided to use as a baseline the ***one-step naïve forecast*** (which uses the actual value from the prior period as the forecast, i.e. $\mathbf{F_t = y_{t-1}}$) and to compute as a basis the most common measures of accuracy MFE, MAD, MAPE, MPE, MSE and CV(RMSE).

Case Questions

1. Open the file Data.xlsx from Part 1 and develop formulas to produce a one-step naïve forecast for the whole possible period. Considering the use of cross-validation technique what are the two data sets, i.e. how would you advise to split the existing data sample – which one should be the training dataset and which one the testing dataset? Explain.
2. Prepare formulas to compute MFE, MAD, MAPE, MPE, MSE and CV(RMSE), calculated:
   a) For the whole dataset;

b) For the testing dataset only.

3. Analyze the results:
   - Is the forecast biased? What measure will identify it?
   - Are there too many extreme values? What measures should be used to determine this? (Hint: compare measures that weight errors in a different way).
   - How ”good” is the relative accuracy of the naïve forecast? What measure will identify this?
4. What overall recommendations would you make to the research team? Explain why.
5. Write a short report (about two pages not counting charts and tables) on the questions above, discussing all important findings and draw relevant conclusions about this part of the Integrative Case.

References

Acken, J. M. (1997). "none". *Encyclopedia of Computer Science and Technology, 36, 281–306.*

Ashby, W. R. (1947). Principles of the Self-Organizing Dynamic System. *Journal of General Psychology, 37*, 125-128.

Beer, S. (1959). *Cybernetics and Management*. London: English University Press.

Camazine, S., Deneubourg, J-L., Franks, N. R., Sneyd, J., Theraulaz, G., & Bonabeau, E. (2003). *Self-Organization in Biological Systems*. Princeton University Press.

Draper, N., & Smith, H. (1981). *Applied Regression Analysis.* New York, NY: John Wiley & Sons.

Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme, I. *Monatshefte für Mathematik und Physik, 38*, 173–98.

Holland, J. H., Holyoak, K J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of Inference, Learning, and Discovery*. Cambridge, MA: MIT Press.

Hurvich, C., & Tsai, C. (1990). The impact of model selection on inference in linear regression. *American Statistician. 44*, 214–217.

Hyndman, R. (2006, June). Another Look at Forecast-Accuracy Metrics for Intermittent Demand. *Foresight, 4*, 43-46.

ISO 5725-1, (1994). Accuracy (trueness and precision) of measurement methods and results - Part 1: General principles and definitions p. 1. Retrieved from https://www.iso.org/obp/ui/#iso: std:iso:5725:-1:ed-1:v1:en

Madala, H. R., & Ivakhnenko, A. G. (1994). *Inductive Learning Algorithms for Complex Systems Modelling*. Boca Raton, FL: CRC Press Inc.

Makridakis, S. G., Wheelwright, S. C., & Hyndman, R. J. (1998). *Forecasting: Methods and Applications.* New York, NY: John Wiley & Sons.

Mark, J., & Goldberg, M. (2001, January). Multiple regression analysis and mass assessment: A review of the issues. *The Appraisal Journal,* 89–109.

Motzev, M. (2011). Intelligent Techniques In Business Games And Simulations – A Hybrid Approach. In M. Beran (Ed.), *Changing the world through meaningful play* (pp. 81-86), Spokane, WA: Eastern Washington University.

Mueller J. A., & Lemke, F. (2003). *Self-Organizing Data Mining: An Intelligent Approach To Extract Knowledge From Data.* Victoria, BC: Trafford Publishing.

Prigogine, I. (1977). *Self-Organization in Non-Equilibrium Systems*, Wiley.

Roecker, E. B. (1991). Prediction error and its estimation for subset-selected models. *Technometrics, 33*, 459-468.

Stone, M. (1974). Cross-Validatory Choice and Assesment of Statistical Predictions. *Journal of the Royal Statistical Society, Ser. B, 36*, 111-133.

Stone, M. (1977). An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion. *Journal of the Royal Statistical Society, Ser. B, 39*, 44-47.

von Neumann, J., & Burks, A. W. (1966). *Theory of Self Reproducing Automata*. Urbana, IL: University of Illinois Press.

Wiener, N. (1948). *Cybernetics or Control and Communication in the Animal and the Machine*.

New York, NY: John Wiley & Sons.

Wilkinson, L., & Dallal, G.E. (1981). Tests of significance in forward selection regression with an F-to enter stopping rule. *Technometrics*, *23*, 377-380.

Wilson, J., & Keating, B. (2008). *Business Forecasting.* McGraw-Hill.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? *Biometrika*. *92*, 937-950

CHAPTER 4. FORECASTING METHODS AND APPROACHES TO FORECASTS

4.1. Summary of Forecasting Methods

Theory provides many sources of information about a variety of forecasting methods. One attempt to summarize all of them and prepare a methodology tree for forecasting is presented in Fig.4-1. Though it contains the most important methods, there are several missing links representing some of the most recent developments of intelligent techniques, like ***Deep Statistical Learning***, ***Self-Organizing Data Mining***, and ***Predictive Analytics*** (discussed in Chapters 10, 11 and 12). There are also other questionable points like whether ***Artificial Neural Networks*** are only used in Univariate techniques, etc.

As mentioned earlier in Chapter 3, *hybrid techniques* based on *Self-Organization* principles have been developed during the last few decades. One successful and world-wide recognized method (Ivakhnenko, 1966) is the ***Group Method of Data Handling (GMDH)***. It provides many types of algorithms (see Table 4.1), which solve most typical problems in the area of prediction modeling (see Farlow, 1984; Madala & Ivakhnenko, 1994; Mueller & Lemke 2003; Motzev, 2018). Important applications of GMDH in forecasting are discussed in Chapters 7, 8, 9 and 12.

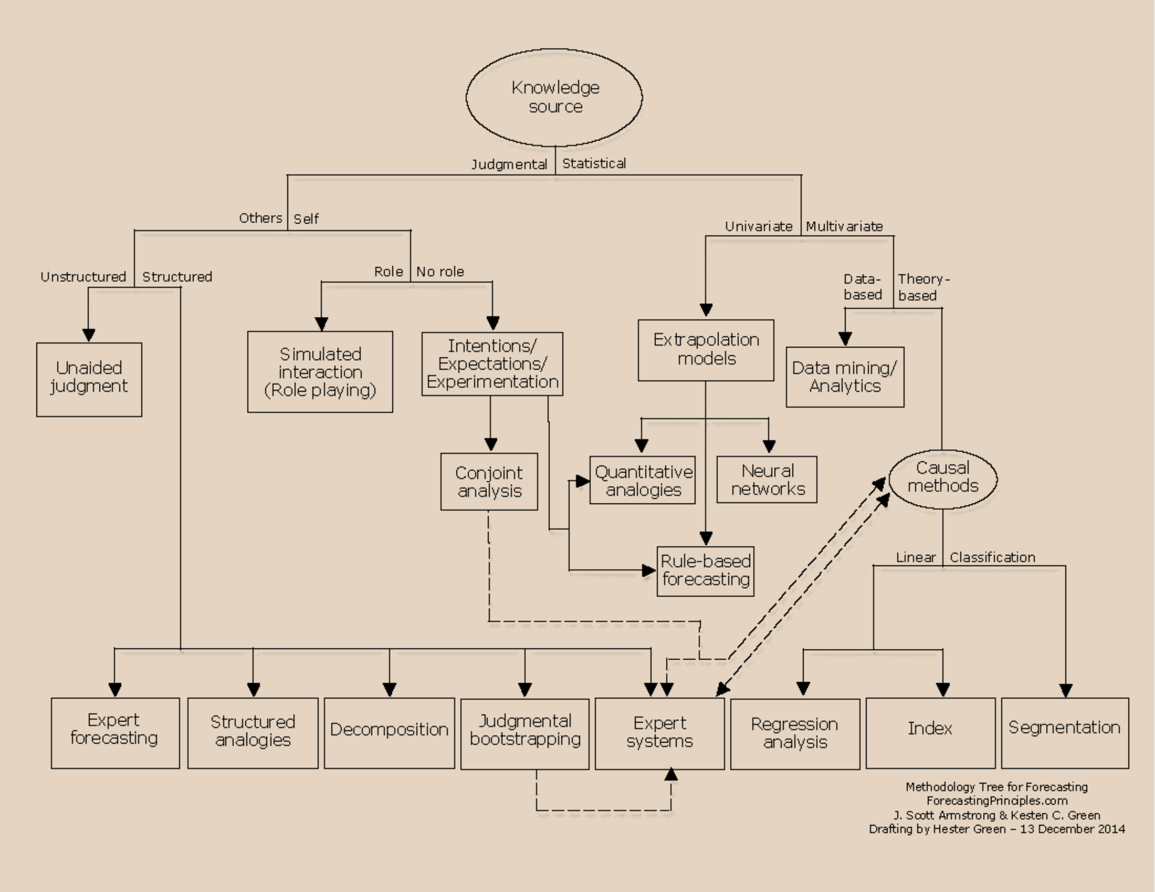


Fig.4-1 Methodology Tree for Forecasting
(Source: Armstrong, 2014)

There are two general approaches to forecasting: qualitative and quantitative. ***Qualitative*** (sometimes referrred to as ***Intuitive*** or ***Subjective***) ***methods*** consist mainly of subjective inputs (usually opinion and judgment of experts), which sometimes defy precise numerical description. They are appropriate when historical data are not available and are usually applied to intermediate- or long-range decisions. Examples of qualitative forecasting methods like the Delphi method, Executive and Sales-force Opinions, Forecast by analogy and Reference class forecasting, Scenario Writing, and others are discussed in the next Section of the book.

Quantitative methods involve either the projection of historical data (***extrapolation***) or the development of ***associative models*** that attempt to utilize causal (explanatory) variables to make a forecast. These methods are usually applied to short- or intermediate-range decisions. Examples of quantitative forecasting methods like Times Series Analysis, Autoregressive Models, and Regression Models will be discussed in Chapters 5, 6, 7, 8 and 9.

Qualitative techniques permit inclusion of ***soft information*** (e.g., human factors, personal opinions, hunches) in the forecasting process. These factors are sometimes omitted or downplayed when quantitative techniques are used because they are difficult or impossible to quantify. Quantitative techniques consist mainly of analyzing ***objective (hard) data***. They usually avoid personal biases that sometimes contaminate qualitative methods. In practice, either or both approaches might be used to develop a forecast.

Recently developed ***hybrid techniques*** provide solutions in both directions – qualitative (non-parametric) and quantitative (parametric). The following Table 4.1 summarized the most important ***GMDH*** algorithms developed so far. Some important applications, as mentioned above, will be discussed in Chapters 8, 9 and 12 in parallel with the traditional techniques.

Table 4.1 Basic GMDH algorithms[1]

	GMDH algorithms	
Variables	**Parametric**	**Non-parametric**
Continuous	- Combinatorial (COMBI) - Multilayered Iterative (MIA) - Objective System Analysis (OSA) - Harmonical - Two-level (ARIMAD) - Multiplicative-Additive (MAA)	- Objective Computer Clusterization (OCC); - "Pointing Finger" (PF) clusterization algorithm; - Analogues Complexing (AC)
Discrete or binary	- Harmonical Rediscretization	- Algorithm on the base of Multilayered Theory of Statistical Decisions (MTSD)

[1] Source: http://www.gmdh.net/GMDH_alg.htm

There is a rich history of forecasting based on subjective and judgmental methods, some of which remain useful even today. These methods are probably most appropriately used when the forecaster is faced with a severe shortage of historical data and/or when quantitative expertise is not available. Sometimes, even nowadays, a judgmental method may even be preferred to a quantitative one. Very long-range forecasting or new product development are typical examples of such a situation. Another reason is that the complicated nature of some quantitative techniques requires specific knowledge, which unfortunately many managers do not have. Last but not least, very often the so called analysts, or forecasters, prefer (for some private reasons) to present the sophisticated methods as impossible to understand, rather than explain them and teach users how and when to apply each one.

In fact, as the research (Makridakis, 1986) has shown, when historical data is available, the judgmental modification of the forecasts produced by analytical methods reduces the accuracy of the forecasts. This finding may be attributed to some bias on the part of the forecaster, possibly because of a tendency to be overly optimistic or to underestimate future uncertainty. It has also been shown that using a judgment component in the forecasting process increases the forecasting cost. In spite of this, executives frequently consider their own judgment superior to other methods of predicting the future. Makridakis (1986, p. 63) states, "People prefer making forecasts judgmentally. They believe that their knowledge of the product, market, and customers as well as their insight and inside information gives them a unique ability to forecast judgmentally”.

As we have already mentioned, it is our understanding that users should go through an evolutionary progression in adopting new forecasting techniques. Experience shows that a simple forecasting method, which is well understood, will be better implemented than one with all-inclusive features but unclear in certain facets. For this reason, the textbook is discussing a large variety of techniques, including some naïve and very simple ones, which every user should know and understand. Next section describes several subjective forecasting methods and in Chapter 6 a few more simple methods are reviewed.

4.2. Subjective (Intuitive) Methods

There are some situations when forecasters should rely solely on judgment and opinion to develop a forecast. If management must have it quickly, there may not be enough time to collect and analyze quantitative data. At other times, especially when political and/or economic conditions are changing, available data may be obsolete and more up-to-date information might not yet be available. Similarly, the introduction of new products and the redesign of existing products or packaging, suffer from the absence of historical data that would be useful in forecasting. In such instances, forecasts are based on executive opinions, consumer surveys, opinions of the sales staff, and opinions of experts. Sometimes, these sources provide insights that are not otherwise available. It is worth noting, however, that with the development of new intelligent techniques, and in particular the ***knowledge discovery from data*** (discussed in Chapter 11), the number of such cases declines.

A. Delphi method

Sometimes, a manager may solicit opinions from a number of other managers and staff. Occasionally, outside experts are needed to help with a forecast. Advice may be needed on political or economic conditions in the United States or a foreign country, or some other aspect of importance with which an organization is not familiar.

The ***Delphi method*** is a systematic, interactive forecasting technique which relies on a panel of experts. The experts answer questionnaires in two or more rounds. After each round, a facilitator provides an anonymous summary of the experts' forecasts from the previous round as well as the reasons they provided for their judgments. Thus, experts are encouraged to revise their earlier answers in light of the replies of other members of their panel. It is believed that during this process the range of the answers will decrease, and the group will converge towards the "correct" answer. Finally, the process is stopped after a pre-defined stop criterion (e.g. a number of rounds, achievement of consensus, or stability of results) and the mean or median scores of the final rounds determine the results.

Delphi is based on the principle that forecasts from a structured group of experts are more accurate than those from unstructured groups or individuals, i.e. its main assumption is that group judgments are more valid than individual judgments. The name "Delphi" derives from the Oracle of Delphi. The authors of the method were not happy with this name, because it implies "something oracular, something smacking a little of the occult". It was developed at the beginning of the Cold War to forecast the impact of technology on warfare. In 1944, General Henry H. Arnold ordered the creation of the report for the U.S. Army Air Corps on the future technological capabilities that might be used by the military.

Different approaches were tried, but the shortcomings of traditional forecasting methods, such as theoretical approach, quantitative models, or trend extrapolation, in areas where precise scientific laws have not been established yet, quickly became apparent. To combat these shortcomings, the Delphi method was developed by Project RAND during the 1950-1960s by Helmer, Dalkey and Rescher (1998). Experts were asked to give their opinion on the probability, frequency, and intensity of possible enemy attacks. Other experts could anonymously give feedback. This process was repeated several times until a consensus emerged (see Fig.4-2).

The following key characteristics of the Delphi method help the participants to focus on the issues at hand and separate Delphi from other methodologies:

Anonymity of the participants – usually all participants maintain anonymity. Their identity is not revealed even after the completion of the final report. This prevents the authority, personality, or reputation of some participants from dominating others in the process. Arguably, it also frees participants (to some extent) from their personal biases, minimizes the "bandwagon effect" or "halo effect", allows free expression of opinions, encourages open critique, and facilitates admission of errors when revising earlier judgments.

Structuring of information flow – the initial contributions from the experts are collected in the form of answers to questionnaires and their comments to these answers. The panel director controls the interactions among the participants by processing the information and filtering out irrelevant content. This avoids the negative effects of face-to-face panel discussions and solves the usual problems of group dynamics.

Regular feedback – participants comment on their own forecasts, the responses of others and on the progress of the panel as a whole. At any moment they can revise their earlier statements. While in regular group meetings participants tend to stick to previously stated opinions and often conform too much to the group leader, the Delphi method prevents this.

Role of the facilitator – the person coordinating the Delphi method is usually known as a facilitator or leader, and facilitates the responses of their panel of experts, who are selected

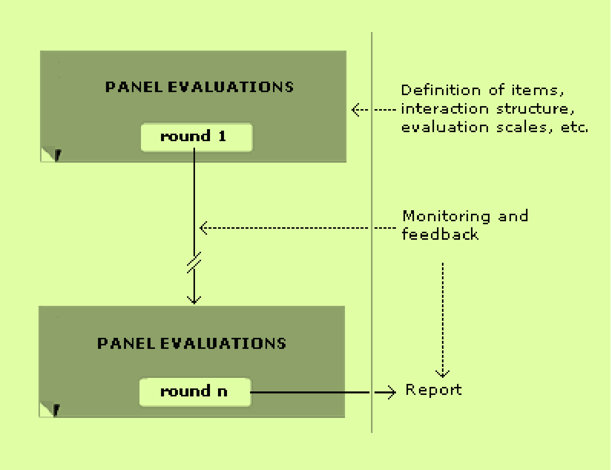


Fig.4-2 The Delphi Method communication structure[2]

[2] Source: http://en.wikipedia.org/wiki/Delphi_method

for a reason, usually that they hold knowledge on an opinion or view. The facilitator sends out questionnaires, surveys, etc. and if the panel of experts accepts, they follow instructions and present their views. Responses are collected and analyzed, then common and conflicting viewpoints are identified. If consensus is not reached, the process continues through thesis and antithesis, to gradually work towards synthesis, and building consensus.

The Delphi method can be summarized by the following six steps:

1. Participating panel members are selected.
2. Questionnaires asking for opinions about the variables to be forecasted are distributed to panel members.
3. Results from panel members are collected, tabulated, and summarized.
4. Summary results are distributed to panel members for their review and consideration.
5. Panel members revise their individual estimates, taking account of the information received from the other, unknown panel members.
6. Steps 3 through 5 are repeated until no significant changes result.

Through this process, there is usually movement toward centrality, but there is no pressure on panel members to alter their original projections. Members, who have strong reason to believe that their original response is correct, no matter how widely it differs from others, can freely stay with it. Thus, in the end, there may not be a consensus.

A number of Delphi forecasts are conducted using websites that allow the process to be conducted in real-time (Fig.4-3). Some innovations came from the use of computer-based and web-based Delphi conferences. According to Turoff & Hiltz (1996) in computer-based Delphis:

- the iteration structure used in the paper Delphis, which is divided into three or more discrete rounds, can be replaced by a process of “round-less” continuous interaction, enabling panelists to change their evaluations at any time;
- the statistical group response can be updated in real-time, and shown whenever a panelist provides a new evaluation.

Since the 1970s, the use of the Delphi technique in public policymaking has introduced a number of methodological innovations. In particular:

- the need to examine several types of items (not only forecasting items but, typically, issue items, goal items, and option items) leads to introducing different evaluation scales which are not used in the standard Delphi. These often include desirability, feasibility (technical and political) and probability, which the analysts can use to outline different scenarios: the desired scenario (from desirability), the potential scenario (from feasibility) and the expected scenario (from probability);

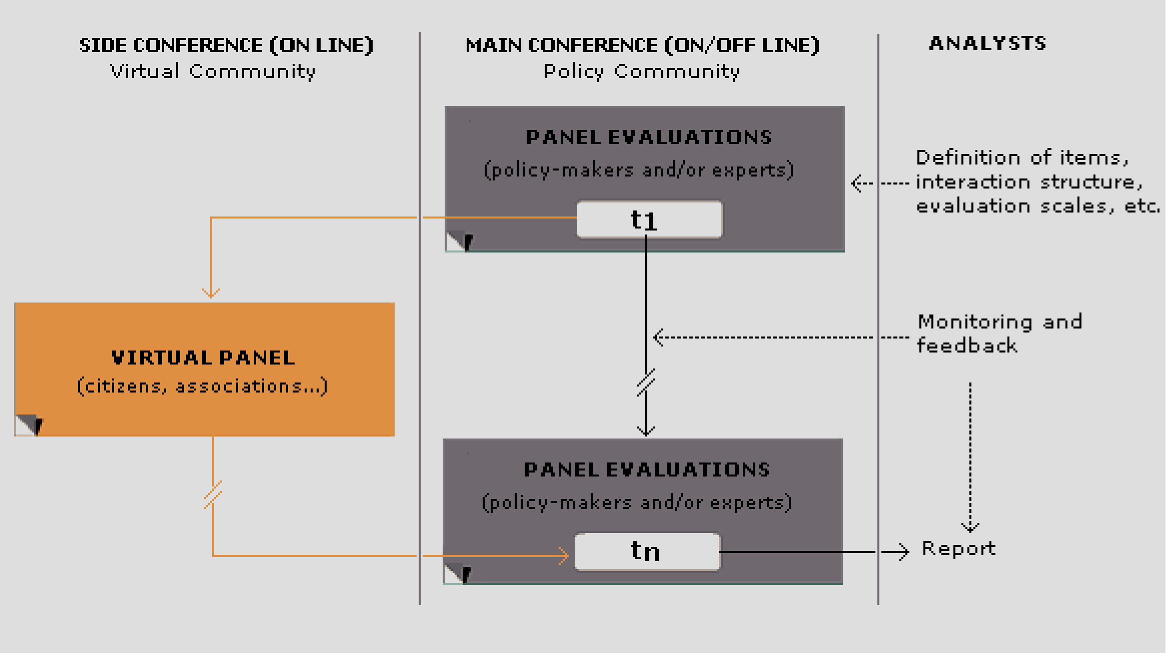

SIDE CONFERENCE (ON LINE)
Virtual Community
MAIN CONFERENCE (ON/OFF LINE)
Policy Community
ANALYSTS
PANEL EVALUATIONS
(policy-makers and/or experts)
t1
Definition of items, interaction structure, evaluation scales, etc.
VIRTUAL PANEL
(citizens, associations...)
Monitoring and feedback
PANEL EVALUATIONS
(policy-makers and/or experts)
tn
Report

Fig.4-3 A web-based communication structure (Hyperdelphi)

- the complexity of the issues posed in public policy-making makes participants give more importance to the arguments supporting the evaluations of the panelists, so they are often invited to list arguments for and against each option item, and sometimes they are given the possibility to suggest new items to be submitted to the panel;
- for the same reason, the scaling methods, which are used to measure panel evaluations, often include more sophisticated approaches such as multi-dimensional scaling.

The first applications of the Delphi method were in the field of science and technology forecasting. Later the method was applied in other areas, especially those related to public policy issues, such as economic trends, health, and education. It was also applied successfully and with high accuracy in business forecasting. For example, in one case reported by Basu and Schroeder[3], the Delphi method predicted the sales of a new product during the first two years with an inaccuracy of 3–4% compared with actual sales. Quantitative methods produced errors of 10–15%, and traditional unstructured forecast methods had errors of about 20%.

Overall, the track record of the Delphi method is mixed. There have been many cases when the method produced poor results. Still, some authors attribute this to poor application of the method and not to the weaknesses of the method itself. It must also be realized that in areas

[3] Basu, Sh. & Schroeder, R. (1977, May). Incorporating Judgments in Sales Forecasts: Application of the Delphi Method at American Hoist & Derrick, *Interfaces*, *7(3)*, 18–27.

such as science and technology forecasting the degree of uncertainty is so great that exact and correct predictions are impossible, and a high degree of error is to be expected.

Another weakness of the Delphi method is that future developments are not always predicted correctly by a consensus of experts. First, the issue of ignorance is important. If panelists are misinformed about a topic, the use of Delphi may only add confidence to their ignorance. Second, sometimes unconventional thinking of amateur outsiders may be superior to expert thinking.

One of the initial problems of the method was its inability to make complex forecasts with multiple factors. Potential future outcomes were usually considered as if they had no effect on each other. Later on, several extensions to the Delphi method were developed to address this problem, such as cross impact analysis, which takes into consideration the possibility that the occurrence of one event may change probabilities of other events covered in the survey. Still, the Delphi method can be used most successfully in forecasting single scalar indicators.

Despite these shortcomings, today the Delphi method is a widely accepted forecasting tool and has been used successfully for thousands of studies in areas varying from technology forecasting to drug abuse. The advantage of the Delphi method is that noted experts can be asked to carefully consider the subject of interest and to reply thoughtfully to the viewpoints of others without the interference of group dynamics. The result, if the process is handled carefully, may be a good consensus of the future along with several alternative scenarios.

B. Prediction Markets

The Delphi method is similar to another structured forecasting approach, the prediction markets. ***Prediction markets*** (also known as *predictive markets, information markets, decision markets, idea futures, event derivatives, or virtual markets*) are speculative markets created for the purpose of making predictions. The current market prices can then be interpreted as predictions of the probability of the event or the expected value of the parameter. For example, a prediction market security might reward a dollar if a particular candidate is elected, such that an individual who thinks the candidate had a 70% chance of being elected should be willing to pay up to 70 cents for such a security.

People who buy low and sell high are rewarded for improving the market prediction, while those who buy high and sell low are punished for degrading the market prediction. Evidence so far suggests that prediction markets are at least as accurate as other institutions predicting the same events with a similar pool of participants.

Delphi has characteristics similar to prediction markets as both are structured approaches that affect aggregate diverse opinions from groups. Yet, there are differences that may be decisive for their relative applicability for different problems (Green et al., 2007).

Some advantages of prediction markets derive from the possibility of providing incentives for participation:

- They can motivate people to participate over a long period of time and to reveal their true beliefs.
- They aggregate information automatically and instantly incorporate new information into the forecast.
- Participants do not have to be selected and recruited manually by a facilitator. They themselves decide whether to participate if they think their private information is not yet incorporated in the forecast.

Delphi seems to have the following advantages over prediction markets:

- Participants reveal their reasoning.
- It is easier to maintain confidentiality.
- Potentially quicker forecasts if experts are readily available.

C. Executive Opinions

The judgments of experts in any area are a valuable resource. Based on years of experience, such judgments can be useful in the forecasting process. Using the method known as the ***jury of executive opinion***, a forecast is developed by combining the subjective opinions of the managers and executives who are most likely to have the best insights about the firm's business. To provide a breadth of opinions, it is useful to select these people from different functional areas. For example, personnel from finance, marketing, and production might be included.

The person responsible for making the forecast may collect opinions in individual interviews or in a meeting where the participants have an opportunity to discuss various points of view. The latter has some obvious advantages such as stimulating deeper insights, but it has some important disadvantages as well. For example, if one or more strong personalities dominate the group, their opinions will become disproportionately weighted in the final consensus that is reached.

A small group of upper-level managers (e.g., in marketing, operations, and finance) may meet and collectively develop a forecast. This approach is often used as a part of long-range planning and new product development. It has the advantage of bringing together the considerable knowledge and talents of various managers. However, there is a risk that the view of one person will prevail, and the possibility that diffusing responsibility for the forecast over the entire group may result in less pressure to produce a good forecast.

It is worth noting that the Delphi method is similar to the jury of executive opinion in taking advantage of the wisdom and insight of people who have considerable expertise in the area to be forecast. It has the additional advantage, however, of anonymity among the participants. The experts, perhaps five to seven in number, never meet to discuss their views and what is more, none of them even knows who else is on the panel.

The Delphi method may be superior to the jury of executive opinion since strong personalities or peer pressures have no influence on the outcome. As mentioned above, the processes of sending out questionnaires, getting them back, tabulating, and summarizing can be expedited by using advanced computer capabilities, including networking and e-mails.[4]

D. Sales-force Opinion (composite)

The sales force can be a rich source of information about future trends and changes in buyer behavior. These people have daily contact with buyers and are the closest contact most firms have with their customers. If the information available from the sales force is organized and collected in an objective manner, considerable insight into future sales can be obtained.

Members of the sales force are asked to estimate sales for each product they handle. These estimates are usually based on each individual's subjective "feel" for the level of sales that would be reasonable in the forecast period. Often a range of forecasts will be requested, including a most optimistic, a most pessimistic, and a most likely forecast. Typically, these individual projections are aggregated by the sales manager for a given product line and/or geographic area. Ultimately the person responsible for the firm's total sales forecast combines the product-line and/or geographic forecasts to arrive at projections that become the basis for a given planning horizon.

While this process takes advantage of information from sources very close to actual buyers, a major problem with the resulting forecast may arise if members of the sales force tend to underestimate sales for their product lines and/or territories. This behavior is particularly likely when the salespeople are assigned quotas on the basis of their forecasts and when bonuses are based on performance relative to those quotas. Such a downward bias can be very harmful to the firm. Scheduled production runs are shorter than they should be, raw-material inventories are too small, labor requirements are underestimated, and in the end, customer will is generated

[4] See, for example, Husbands, B. S. (1982, Summer). Electronic Mail System Enhances Delphi Method. *Journal of Business Forecasting I*, *4*, 24-27.

by product shortages. The sales manager with ultimate forecasting responsibility can offset this downward bias, but only by making judgments that could, in turn, incorporate other bias into the forecast Robin Peterson has developed a way of improving sales-force composite forecasts by using a prescribed set of learned routines as a guide for salespeople as they develop their forecasts.[5]

Members of the sales (or the customer service) staff are often good sources of information because of their direct contact with consumers. They are often aware of customers' plans for the future. There are, however, several drawbacks to using sales force opinions. One is that staff members may be unable to distinguish between what customers would *like* to do and what they actually *will* do. Another is that these people are sometimes overly influenced by recent experiences. Thus, after several periods of low sales, their estimates may tend to become pessimistic, or after periods of good sales, they may tend to be too optimistic. In addition, if forecasts are used to establish sales quotas, there will be a conflict of interest because it is to the salesperson's advantage to provide low sales estimates.

E. Surveys of Customers and the General Population

In some situations, it may be practical to survey customers for more information about their buying intentions. This practice presumes that buyers plan their purchases and follow through with their plans. Such an assumption is probably more realistic for industrial sales than for sales to households and individuals. It is also more realistic for big-ticket items such as cars or personal computers than for convenience goods like toothpaste or tennis balls.

Survey data concerning how people feel about the economy are sometimes used by forecasters to help predict certain buying behaviors. One of the commonly used measures of how people feel about the economy comes from a monthly survey conducted by the University of Michigan Survey Research Center. The Center produces an Index of Consumer Sentiment (ICS) based on a survey of 500 individuals, 40 percent of whom are respondents who participated in the survey six months earlier and the remaining 60 percent are new respondents selected on a random basis. This index has its base period in 1966 when the index was 100. High values of the ICS indicate more positive feelings about the economy than do lower values. Thus, if the ICS goes up, one might expect that people are more likely to make certain types of purchases.

[5] See: Peterson, R. T. (1993). Improving Sales Force Composite: Forecasting by Using Scripts. *Journal of Business Forecasting, Fall,* 10-14

Because it is the consumers who ultimately determine demand, it seems natural to solicit input from them. In some instances, every customer or potential customer can be contacted. However, usually there are too many customers or there is no way to identify all potential customers. Therefore, organizations seeking consumer input usually resort to consumer surveys, which enable them to *sample* consumer opinions. The obvious advantage of consumer surveys is that they can tap information that might not be available elsewhere. On the other hand, a considerable amount of knowledge and skill is required to construct a survey, administer it, and correctly interpret the results for valid information.

Surveys can be expensive and time-consuming. In addition, even under the best conditions, surveys of the general public must contend with the possibility of irrational behavior patterns. For example, much of the consumer's thoughtful information gathering before purchasing a new car is often undermined by the glitter of a new car showroom or a high-pressure sales pitch. Along the same lines, low response rates to a mail survey could make the results suspect. If these and similar pitfalls can be avoided, surveys can produce useful information.

F. Scenario Writing

Scenario writing involves defining the particulars of an uncertain future by writing a "*script*" for the environment of an organization over many years in the future. New technology, population shifts, and changing consumer demands are among the factors that are considered and woven into this speculation to provoke the thinking of top management.

A most likely scenario is usually written along with one or more less likely, but possible, scenarios. By considering the posture of the company for each of these possible future environments, top management is in a better position to react to actual business environment changes as they occur and to recognize the long-range implications of subtle changes that might otherwise go unnoticed. In this way, the organization is in a better position to maintain its long-term profitability rather than concentrate on short-term profits and ignore the changing technological environment in which it operates.

The scenario-writing process is often followed by a discussion phase, sometimes by a group other than the one that developed the scenarios. Discussion among the groups can then be used to defend and modify viewpoints so that a solid consensus and alternative scenarios are developed. For example, scenarios might be developed by a company's planning staff and then discussed by the top management team. Even if none of the scenarios are subsequently proven to be totally true, this process encourages the long-range thinking of the top management team and better prepares it to recognize and react to important environmental changes.

Example: A company that manufactures industrial telephone and television cables decides to conduct a scenario-writing exercise prior to its annual weekend retreat. Each member of the retreat group is asked to write three scenarios that might face the company 5 years from now: a worst-case, a most likely, and a best-case scenario. After these writing assignments are completed, and just before the weekend retreat, the president, and his senior vice president summarized the contributions into the following three scenarios on which they intend to focus the group's discussion during the two-day retreat:

1. Internet usage continues to grow rapidly but slowly moves away from cable in favor of satellite access. Even telephone service relies increasingly on noncable means, as does home television reception. The company sees its sales and profits on a non-ending decline and will soon be out of business.

2. Internet and home television service continues to grow rapidly and is provided by several sources. Satellite service is widely used, but buried cables continue to be an integral part of high-tech service, especially in large cities, both in private housing and industrial applications. The company's leadership in cable development and deployment results in increased sales and profits.

3. Due to technical problems and security issues, satellite usage for the Internet and television service declines until it is used primarily in rural areas. Cable service grows rapidly in both residential and industrial applications, and the company prospers as its well-positioned products and industry leadership result in industry dominance.

The company president and senior vice president intend to have extensive discussions on each of these three scenarios. They want to have long-range strategies developed that will accommodate all future possibilities and believe that focusing on these three cases will energize both themselves and their management team.

G. Forecast by Analogy and Reference Class Forecasting

Forecast by analogy is a forecasting method that assumes that two different kinds of phenomena share the same model of behaviour. For example, one way to predict the sales of a new product is to choose an existing product which "looks like" the new product in terms of the expected demand pattern for sales of the product. "Used with care, an analogy is a form of scientific model that can be used to analyze and explain the behavior of other phenomena."[6]

Reference class forecasting, or ***comparison class forecasting***, is another method of predicting the future, through looking at similar past situations and their outcomes.

[6] Morlidge, S., & Player, S. (2010). *Future Ready: How to Master Business Forecasting*. Wiley&Sons, p. 287.

Reference class forecasting predicts the outcome of a planned action based on actual outcomes in a reference class of similar actions to that being forecasted.

Kahneman and Tversky (1979) found that human judgment is generally optimistic due to overconfidence and insufficient consideration of distributional information about outcomes. Therefore, people tend to underestimate the costs, completion times, and risks of planned actions, whereas they tend to overestimate the benefits of those same actions. Such error is caused by actors taking an "inside view," where the focus is on the constituents of the specific planned action instead of on the actual outcomes of similar ventures that have already been completed.

Kahneman and Tversky (1979) concluded that disregard of distributional information is perhaps the major source of error in forecasting. On that basis, they recommended that forecasters "should, therefore, make every effort to frame the forecasting problem so as to facilitate the utilization of all the distributional information that is available". (p. 316).

Using distributional information from previous ventures similar to the one being forecast is called taking an "outside view". Reference class forecasting is a method for taking an outside view on planned actions.

Reference class forecasting for a specific project involves the following three steps:

a) Identify a reference class of past, similar projects.
b) Establish a probability distribution for the selected reference class for the parameter that is being forecasted.
c) Compare the specific project with the reference class distribution, in order to establish the most likely outcome for the specific project.

The first instance of reference class forecasting in practice was a forecast carried out in 2004 by the UK government of the projected capital costs for an extension of Edinburgh Trams.[7]

Advantages and Disadvantages of Subjective Methods

Subjective (i.e., qualitative or judgmental) forecasting methods are sometimes considered desirable because they do not require any particular mathematical background of the individuals involved. As future business professionals become better trained in quantitative forms of analysis, this advantage will become less important.

Historically, another advantage of subjective methods has been their wide acceptance by users. However, our experience suggests that users are increasingly concerned with how the forecast was developed, and with the most subjective methods it is difficult to be specific in this

[7] See "Council to borrow £231m for Edinburgh trams project". BBC News (BBC), 19 August 2011.

regard. The underlying models are, by definition, subjective. This subjectivity is nonetheless the most important advantage of this class of methods. There are often forces at work that cannot be captured by quantitative methods. They can, however, be sensed by experienced business professionals and can make an important contribution to improved forecasts. Wilson and Allison-Koerber have shown this dramatically in the context of forecasting sales for a large piece of food-service equipment produced by the Delfield Company[8]. Quantitative methods reduced errors to about 60 percent of those that resulted from the subjective method that had been in use. When the less accurate subjective method was combined with the quantitative methods, errors were further reduced to about 40 percent of the level when the subjective method was used alone. It is clear from this result, and others, that there is often important information content in subjective methods.

The disadvantages of subjective methods were nicely summarized by Charles W. Chase, Jr., when he was with Johnson & Johnson Consumer Products, Inc. He stated that "the disadvantages of qualitative methods are: (1) they are almost always biased; (2) they are not consistently accurate over time; (3) it takes years of experience for someone to learn how to convert intuitive judgment into good forecasts."[9]

4.3. Quantitative Methods

Quantitative (also known as ***mathematical, statistical***) ***techniques*** using the power of the computer have come to dominate the forecasting landscape. There are three major groups, which will be discussed in detail in the following chapters of the textbook.

A. Time Series methods

A ***time series*** is a sequence of data points, measured typically at successive points in time spaced at uniform time intervals. Example of time series are the daily closing values of the Dow Jones Industrial Average. ***Time series analysis*** comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. ***Time series forecasting*** is the use of a model to predict future values based on previously observed data. The most common methods are: Naïve and Visual forecasts; Moving average & Weighted moving average; Exponential smoothing; Trend estimation; Growth curve and others. Basic techniques are discussed in Chapter 5 and more advanced methods in Chapter 7.

[8] See: Wilson, J. & Keating, B. (1998). *Business Forecasting.* Irwin McGraw-Hill, p. 405.

[9] See: Chase, C. W. Jr. (1991, Spring). Forecasting Consumer Products. *Journal of Business Forecasting,* p. 4.

B. Causal/Econometric methods

Some forecasting methods try to identify the underlying factors that might influence the variable that is being forecast. For example, including information about climate patterns might improve the ability of a model to predict umbrella sales. These methods often take account of regular seasonal variations as well. In addition to climate, such variations can also be due to holidays and customs: for example, one might predict that sales of college football apparel will be higher during the football season than during the offseason.

All of the above mentioned methods use the assumption that it is possible to identify the underlying factors that might influence the variable that is being forecasted. If the causes are understood, projections of the influencing variables can be made and used in the forecast.

In general, the causal methods include the following groups:

- Regression analysis includes a large group of methods for predicting future values of a variable using information about other variables. These methods include both parametric (linear or non-linear) and non-parametric techniques and are discussed in Chapters 6 and 9.
- Autoregressive models, autoregressive integrated moving average (ARIMA), also known as Box-Jenkins, and Autoregressive moving average with exogenous inputs (ARMAX). These techniques are discussed in Chapter 8.
- Econometrics models, incl. single functions and systems of equations are discussed in Chapter 9.

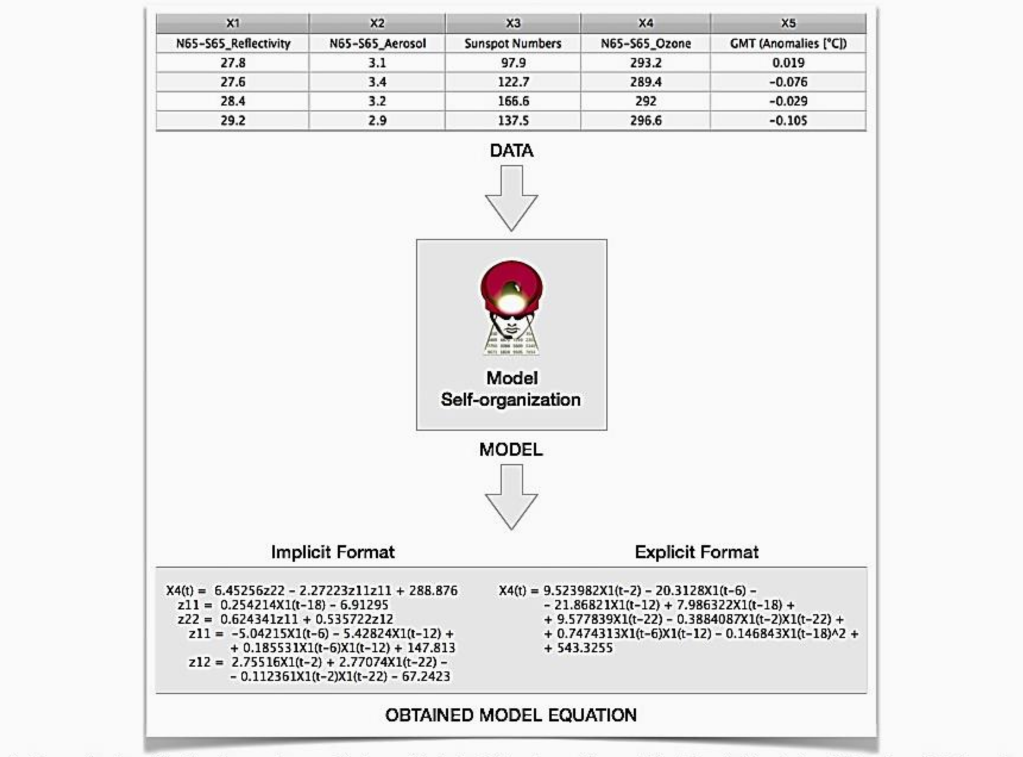
C. Artificial Intelligence methods and Data Mining

Artificial Intelligence (AI) is the “intelligence” exhibited by machines or software. The central problems (or goals) of AI research include reasoning, knowledge, planning, learning, natural language processing (communication), perception and the ability to move and manipulate objects. Currently, popular approaches include statistical methods, computational intelligence and traditional symbolic AI. There are a large number of tools used in AI, including versions of search and mathematical optimization, logic, methods based on probability and economics, and many others. The AI field is interdisciplinary, in which a number of sciences and professions converge, including computer science, psychology, linguistics, philosophy, and neuroscience.

Machine learning is a subfield of computer science and artificial intelligence that deals with the construction and study of systems that can learn from data, rather than follow only explicitly programmed instructions.

Data mining (the analysis step in Knowledge Discovery in Databases process, or KDD (Fayyad et al., 1996), is also an interdisciplinary subfield of computer science and artificial intelligence. Nowadays, it is a common understanding that ***Machine Learning***, ***Data Mining***, and ***Pattern Recognition***[10] are conflated. In fact, ***Data Mining*** is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the ***Data Mining*** process is to extract information from a data set and transform it into an understandable structure (model) for further use (see Fig.4-4).

The major groups of techniques that are found in most comprehensive ***Data mining*** tools are *decision trees, neural networks* and *clustering.* They are discussed in detail in Chapters 10 and 11. Special attention will be given to the hybrid approach ***Group Method of Data Handling (GMDH)***, which comprises ***genetic algorithms***, ***multi-stage (or multi-layered) selection procedures*** and other principles of ***self-organization.*** Its real-life applications are discussed in Chapters 8, 9 and 12 accordingly.



X1	X2	X3	X4	X5
N65-S65_Reflectivity	N65-S65_Aerosol	Sunspot Numbers	N65-S65_Ozone	GMT (Anomalies [°C])
27.8	3.1	97.9	293.2	0.019
27.6	3.4	122.7	289.4	-0.076
28.4	3.2	166.6	292	-0.029
29.2	2.9	137.5	296.6	-0.105

Implicit Format

```
X4(t) =  6.45256z22 - 2.27223z11z11 + 288.876
  z11 =  0.254214X1(t-18) - 6.91295
  z22 =  0.624341z11 + 0.535722z12
   z11 =  -5.04215X1(t-6) - 5.42824X1(t-12) +
          + 0.185531X1(t-6)X1(t-12) + 147.813
   z12 =  2.75516X1(t-2) + 2.77074X1(t-22) -
          - 0.112361X1(t-2)X1(t-22) - 67.2423
```

Explicit Format

```
X4(t) = 9.523982X1(t-2) - 20.3128X1(t-6) -
        - 21.86821X1(t-12) + 7.986322X1(t-18) +
        + 9.577839X1(t-22) - 0.3884087X1(t-2)X1(t-22) +
        + 0.7474313X1(t-6)X1(t-12) - 0.146843X1(t-18)^2 +
        + 543.3255
```

OBTAINED MODEL EQUATION

Self-organization of optimal complex models from data in *Insights* along with analytical descriptions in implicit and explicit format.

Fig.4-4 Example of ***Data mining platform*** for model building

(Source: http://knowledgeminer.eu/about.html)

***

[10] ***Pattern recognition*** is nearly synonymous with ***machine learning*** (Bishop, 2006, p. vii). This branch of AI focuses on the recognition of patterns and regularities in data.

SUMMARY AND CONCLUSIONS

Chapter 4 discusses the types of forecasting. Today, there is a large variety of methods used in Business Forecasting which we can summarize in a few major categories:

Judgmental methods

As discussed above, judgmental forecasting methods incorporate intuitive judgments, opinions and subjective probability estimates.

- Surveys
- Delphi method
- Scenario building
- Executive and Sales-force opinions
- Forecast by analogy etc.

Time Series methods

Time series methods use historical data as the basis of estimating future outcomes.

- Naïve and Visual forecasts
- Moving average
- weighted moving average
- Exponential smoothing
- Extrapolation
- Linear prediction
- Trend estimation
- Growth curve etc.

Causal/Econometric methods

Some forecasting methods use the assumption that it is possible to identify the underlying factors that might influence the variable that is being forecast. For example, sales of umbrellas might be associated with weather conditions. If the causes are understood, projections of the influencing variables can be made and used in the forecast.

- Regression analysis using linear regression or non-linear regression
- Autoregressive moving average (ARMA)
- Autoregressive integrated moving average (ARIMA) e.g. Box-Jenkins
- Econometrics models (systems of equations).

Data Mining methods

- Artificial neural networks
- Support vector machines
- Genetic Algorithms

- Group method of data handling (GMDH)

This textbook presents the traditional, state-of-the-art methods, but the real emphasis is on the new, advanced methods, such as Data Mining techniques, which will be discussed in detail.

KEY TERMS

Artificial intelligence (AI)	***112***
Associative models	***98***
Causal / Econometric methods	***112***
Delphi method	***100***
Executive Opinions	***105***
Extrapolation	***98***
Forecast by analogy, Reference class forecasting	***109***
Hybrid techniques, Group Method of Data Handling (GMDH)	***98, 111***
Machine learning**, **Data Mining**, **Pattern Recognition	***113***
Prediction markets (information markets, decision markets, etc).	***104***
Qualitative (***Intuitive*** or ***Subjective***) ***methods***	***98, 99***
Quantitative methods	***98***
Sales-force Opinions	***106***
Scenario Writing	***108***
Time series, Time series analysis**, **Time series forecasting	***111***

CHAPTER EXERCISES

Conceptual Questions:

1. What is the Delphi method and how is it applied? Discuss.
2. List all 6 steps in the Delphi method. Discuss and illustrate with examples.
3. What are the major pros and cons of prediction market vs Delphi method? Explain.
4. What are the fundamentals of self-organizing modeling? List and discuss at least three of them.
5. What are the major groups ***Quantitative techniques?*** Discuss and illustrate with examples.

Business Applications:

The M&M company plans to launch a brand new product on the market. Marketing manager Daniel Steel knows that the Judgmental forecasting is usually the only available method for new product forecasting as historical data are unavailable. The approaches we have already outlined (***Delphi, forecasting by analogy, executive opinion scenario forecasting*** and so on) are all applicable when forecasting the demand for a new product.

Write a short essay (up to two pages) discussing your opinion about using the following methods in this particular case:

- Executive opinion.
- Sales-force composite
- Survey of customer intentions
- Your overall recommendation to Daniel Steel.

INTEGRATIVE CASE

HEALTHY FOOD SUPPLY CHAIN & STORES

Part 4: The Art of Forecasting – Getting started

In Chapter 1 we introduced *Healthy Food Stores* – a fast-growing retail food provider with 12 stores in a northwestern state. The company executives decided to study the effect that company advertising dollars have on sales. They hoped that examining collected historical data would reveal relationships that would help determine future advertising expenditures and predict monthly sales volumes for the upcoming quarter.

After identifying basic parameters, along with input (independent) and output (dependent) variables of the forecasting scenario in Part 2, the research team determined the main elements of the forecasting process:

- Forecasting horizon of up to twelve months;
- Quarterly forecast updates, since accuracy decreases as time horizon increases, and sufficient time is needed to implement possible changes;
- Development of different models based on data patterns, if any, and evaluation of their accuracy in order to select the most appropriate one;
- Selection of the best forecasting model with no more than 5% forecasting error.

The benchmark forecast computed in Part 3 (using the baseline of ***one-step naïve forecast*** as a reference forecast) provided basic values for the most common measures of accuracy MFE, MAD, MAPE, MPE, MSE and CV(RMSE).

It was time to learn more about the opinion of some important people from the Healthy *Food Stores* Company concerning this specific case. The research team planned to use three different techniques in order to reach the three groups of most important people in the company.

Case Questions

1. What are the three groups of most important people in the company regarding this particular case? Explain why.
2. What are the most suitable Subjective (judgmental) forecasting methods to study the effect company advertising dollars have on sales? Discuss and illustrate with appropriate examples each of them.
3. What overall recommendations would you make to the research team? Explain why.
4. Write a short report (about two pages not counting charts and tables) on the questions above, discussing all important points and draw relevant conclusions about this part of the Integrative Case.

References

Armstrong, S. (2014). Methodology Tree for Forecasting, Retrieved March 30, 2020, from http://www.forecastingprinciples.com/index.php?option=com_content&task=view&id=16&Itemid=16

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer.

Farlow, S. (Ed.). (1984). *Self-Organizing Methods in Modeling*. Marcel Dekker Inc.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996, Fall). From Data Mining to Knowledge Discovery in Databases. *American Association for Artificial Intelligence Magazine*, 37-54

Green, K. C., Armstrong, J. S., & Graefe, A. (2007, Fall). Methods to Elicit Forecasts from Groups: Delphi and Prediction Markets Compared. Forthcoming in Foresight. *The International Journal of Applied Forecasting, 8*, 17-20.

Ivakhnenko, A. G. (1966). Group Method of Data Handling –A Rival of the Method of Stochastic Approximation. *Soviet Automatic Control*, *13*, 43-71.

Kahneman, D., & Tversky, A. (1979). Intuitive Prediction: Biases and Corrective Procedures. In S. Makridakis & S. C. Wheelwright (Eds.), *Studies in the Management Sciences: Forecasting*, *12*. Amsterdam: North Holland.

Madala, H. R., & Ivakhnenko, A. G. (1994). *Inductive Learning Algorithms for Complex Systems Modelling*. Boca Raton, FL: CRC Press Inc.

Makridakis, S. (1986). The Art and Science of Forecasting. *International Journal of Forecasting*, *2*, 15-39.

Motzev, M. (2018). A Framework for Developing Multi-Layered Networks of Active Neurons for Simulation Experiments and Model-Based Business Games Using Self-Organizing Data Mining with the Group Method of Data Handling. In: H. Lukosch, G. Bekebrede, & R. Kortmann (Eds.), *Simulation Gaming. Applications for Sustainable Cities and Smart Infrastructures* (pp. 191-201). ISAGA 2017. Lecture Notes in Computer Science, vol 10825. Springer, Cham. https://doi.org/10.1007/978-3-319-91902-7_19

Mueller J. A., & Lemke, F. (2003). *Self-Organizing Data Mining: An Intelligent Approach To Extract Knowledge From Data*. Victoria, BC: Trafford Publishing.

Rescher, N. (1998). *Predicting the Future*. Albany, NY: State University of New York Press.

Turoff, M., & Starr, R. H. (1996). Computer-based Delphi processes. In M. Adler, & E. Ziglio (Eds.), *Gazing Into the Oracle: The Delphi Method and Its Application to Social Policy and Public Health* (pp. 56-89). London: Kingsley Publishers.

CHAPTER 5. BASIC QUANTITATIVE TECHNIQUES

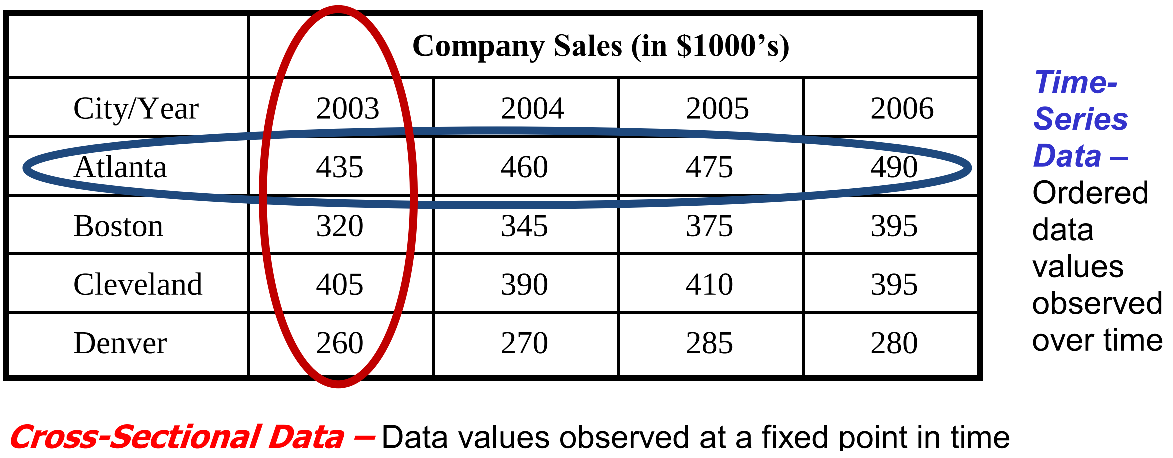
5.1. Naïve Forecasts and Graphical Techniques

Forecasting, or making Predictions, as discussed in previous chapters, is the process of making statements about events whose actual outcomes have not yet been observed. Both might refer to formal statistical methods employing time-series, cross-sectional, or multi-dimensional data, or alternatively to less formal judgmental methods.

Cross-sectional data or a cross-section of a study population is a type of data collected by observing many subjects (such as individuals, firms or regions) at the same point of time, or without regard to differences in time. Analysis of cross-sectional data usually consists of comparing the differences among the subjects.

For example, if we want to measure the current demand level for a new product in a population, we could randomly draw a sample of 1,000 people from that population (also known as a cross-section of that population), record their preferences and calculate what percentage of that sample is willing to buy the new product. This cross-sectional sample provides us with a snapshot of that population, at that point in time. Note that we do not know, based on one cross-sectional sample, if demand is increasing or decreasing, we can only describe the current proportion.

Cross-sectional data differs from ***time-series data***, in which the same small-scale or aggregate entity is observed at various points in time, for example, longitudinal data which follow one subject's changes over the course of time. Another variant, panel data (or ***Time-Series-Cross-Sectional (TSCS) data***), combines both and looks at multiple subjects and how they change over the course of time (see Fig.5-1). Panel analysis uses panel data to examine changes in variables over time and differences in variables between subjects.



	Company Sales (in $1000's)			
City/Year	2003	2004	2005	2006
Atlanta	435	460	475	490
Boston	320	345	375	395
Cleveland	405	390	410	395
Denver	260	270	285	280

Time-Series Data – Ordered data values observed over time

Cross-Sectional Data – Data values observed at a fixed point in time

Fig.5-1 Examples of different data types

A ***time series*** is a sequence of data values, measured typically at successive points in time spaced at uniform time intervals. A typical example of time series is the daily closing values of the Dow Jones Industrial Average. Time series are used virtually everywhere – in statistics, signal processing, econometrics, pattern recognition, mathematical finance, forecasting, earthquake prediction, control engineering, astronomy, communications engineering, and largely in any domain of applied science which involves time-based measurements.

Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. ***Time series forecasting*** is the use of a model to predict future values based on previously observed values. A commonplace example might be an estimation of some variable of interest at some specified future date. While regression analysis is often employed in order to test theories that the current values of one or more independent time series affect the current value of another time series, this type of analysis of time series is not called "time series analysis" – it is usually referred to as "***Regression with Time Series Data***" (see Ch. 7). ***Time series analysis*** focuses on comparing values of a single time series or multiple dependent time series at different points in time.

A number of different notations are in use for time-series analysis. A common notation specifying a time series ***X*** that is indexed by the natural numbers is written

$$\boldsymbol{X = \{X_1, X_2, ...\}}$$

Another common notation is

$$\boldsymbol{Y = \{Y_t : t \in T\},}$$

where *T* is the ***index set***[1].

Mathematically, if we denote time by the variable ***t***, and sales by ***X***, then the function denoted ***X(t)*** indicates that ***X*** (the dependent variable sales) is a function of ***t***.

A. Charts and Other Visual Techniques

The clearest way to examine a regular time series manually is with a line chart such as the one shown in Fig.5-2, made with a spreadsheet program. The number of cases (the number of customer complaints) was standardized to a rate per 100,000 and the percent change per year in this rate was calculated. The nearly steadily dropping line shows that the customer complaints were decreasing in most years, but the percent change in this rate varied by as much as +/- 10%, with some flows in 1975 and around the early 1990s. The use of both vertical axes allows the comparison of two time series in one graphic.

[1] In mathematics, an ***index set*** is a set whose members label (or index) members of another set. For instance, if the elements of a set ***A*** may be *indexed* or *labeled* by means of a set ***T***, then ***T*** is an **index set**.

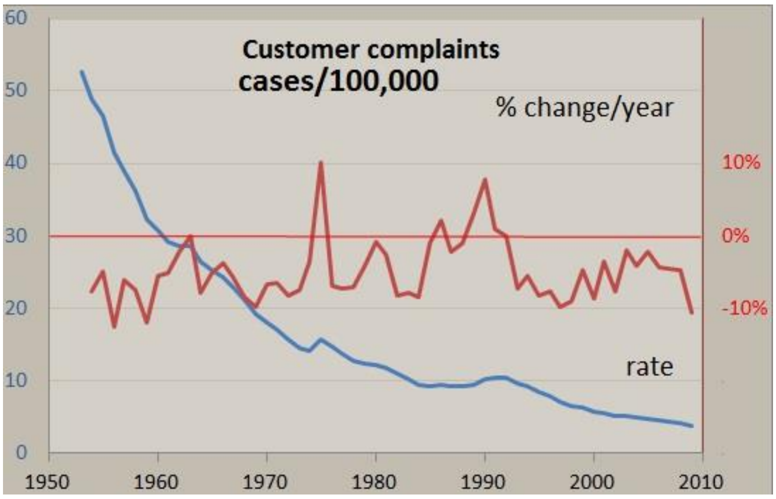


Fig.5-2 Example of Time series combo chart

A ***line chart*** or ***line graph*** is a type of chart which displays information as a series of data points (known as 'markers') connected by straight line segments. It is a basic type of chart common in many fields. ***Line charts*** show how a particular data changes at equal intervals of time. When a line chart is used to visualize a trend in data over intervals of time (i.e. time series) the line is drawn chronologically.

Data collected from experiments or by observations are often visualized by tables. For example, if one were to collect data, let us say, on the speed of a product line at certain points in time, one could visualize the data by a data table such as Table 5.1.

The table "visualization" is a good way of displaying exact values but can be a poor way to understand the underlying patterns that those values represent. Sometimes, the table display is incorrectly conflated with the data itself, though it is just another visualization of the data. Understanding the process described by the data in Table 5.1 should be aided by producing a line chart of Speed versus Time, as shown in Fig.5-3.

Table 5.1 Speed Vs Time

Elapsed Time (s)	Speed (m/s^{-1})
0	0
1	3
2	7
3	12
4	20
5	30
6	45

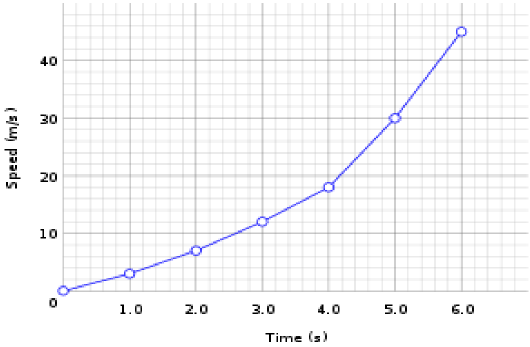


Fig.5-3 Graph of Speed Vs Time

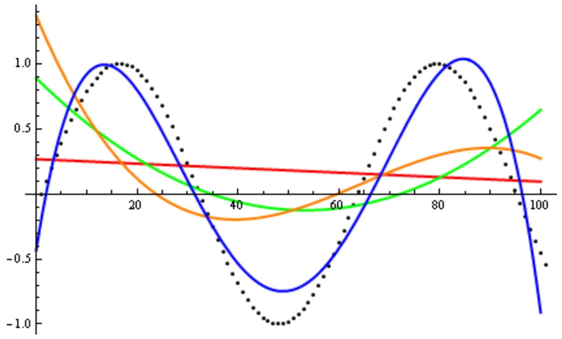


Fig.5-4 Example of Best-fit layers with Polynomial curves – the red line is a first-degree polynomial, the green line is the second degree, the orange line is the third degree and the blue is a fourth-degree polynomial

Time-Series charts often include an overlaid mathematical function depicting the best-fit trend of the scattered data. This layer is referred to as a ***best-fit*** layer (see Fig.5-4). It is simple to construct a "best-fit" layer consisting of a set of line segments connecting adjacent data points, however, such a "best-fit" is usually not an ideal representation of the trend of the underlying scatter data for the following reasons:

- It is highly improbable that the discontinuities in the slope of the best-fit would correspond exactly with the positions of the measurement values.
- It is highly unlikely that the experimental error in the data is negligible, yet the curve falls exactly through each of the data points.

In either case, the best-fit layer can reveal trends in the data. Furthermore, measurements such as the gradient or the area under the curve can be made visually, leading to more conclusions or results from the data. A true best-fit layer should depict a continuous mathematical function whose parameters are determined by using a suitable error-minimization technique (like ***Least Squares (LS)*** method – see Chapter 6), which appropriately weights the error in the data values. Such curve fitting functionality is often found in graphing software or spreadsheets. Best-fit curves may vary from simple linear equations to more complex quadratic, polynomial, exponential, and periodic curves.

Sometimes, a ***fan chart*** could be used in time series analysis. A fan chart is a chart that joins a simple line chart for observed past data, by showing ranges for possible values of future data together with a line showing a central estimate or most likely value for the future outcomes (see Fig.5-5). As predictions become increasingly uncertain the further into the future one goes, the more these forecast ranges spread out, creating distinctive wedge or "fan" shapes, hence the term. Alternative forms of the chart can also include uncertainty for past data, such as preliminary data that is subject to revision.

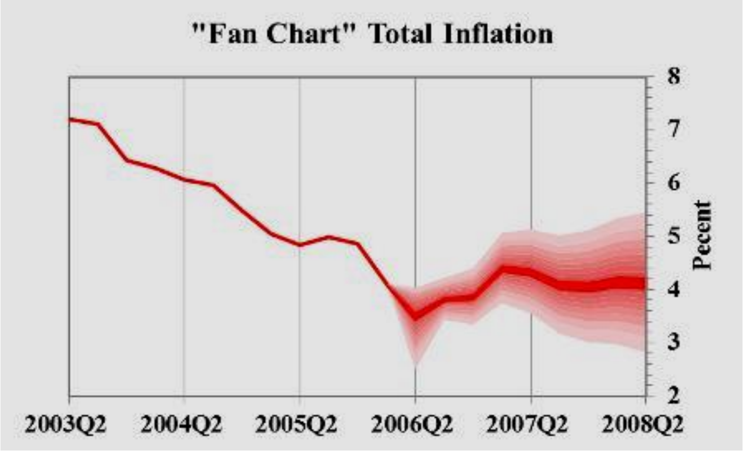


Fig.5-5 Hypothetical Fan Chart of the Inflation Rate

The term "***fan chart***" was coined by the Bank of England, which has been using these charts and this term since 1997 in its "Inflation Report"[2] to describe its best prevision of future inflation to the general public. Fan charts have been used extensively in finance and monetary policy, for instance, to represent forecasts of inflation.

Predicted future values can be diagrammed in various ways and the most common is by a single predicted value, and an upper and lower range around it. Another one is by various future intervals, depicted by varying degrees of shading (darkest near the center of the range, fainter near the ends of the range – see Fig.5-5).

There are different ways to represent the forecast density depending on the shape of the forecasting distribution:

- If the forecast density is symmetric, the fan centers at the mean (which coincides with the mode and median) forecast, and the ranges expand like ***confidence intervals*** by adding and subtracting multiples of the forecasting standard error to the mean forecast (because of this they are referred to as ***equal tail ranges***). We can add and subtract one, two and three forecasting standard errors for approximate coverage of 68%, 95% and 99.7%. These charts can easily be built through standard Excel functions.
- If the forecast density is non-symmetric, centering the fan at the mean and using equal tail ranges might not be appropriate as it would overstate the forecast uncertainty. In this case, it is better to center the fan at the more likely forecast (the mode) and use Highest Probability Density (HPD) ranges. HPDs are by definition the shortest ranges covering a given probability, say 50%, and are centered at the mode. In this case, it is usual to include increasing probability ranges of 10%, 20%, …, 90%, for instance.

[2] Source: Britton, E., Fisher, P., & Whitley, J. (1998, February). *The Inflation Report Projections: Understanding the Fan Chart*. Bank of England Quarterly Bulletin.

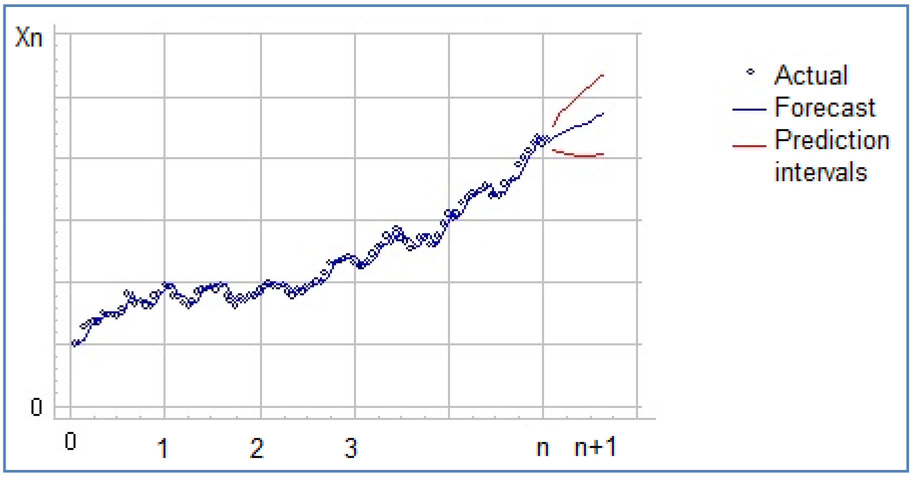


Fig.5-6 Example of Forecast chart and Prediction intervals

In both cases, it is always a good idea to add prediction intervals to the main line graph. A ***prediction interval*** is an estimate of the interval in which future observations will fall, with a certain probability, given what has already been observed.

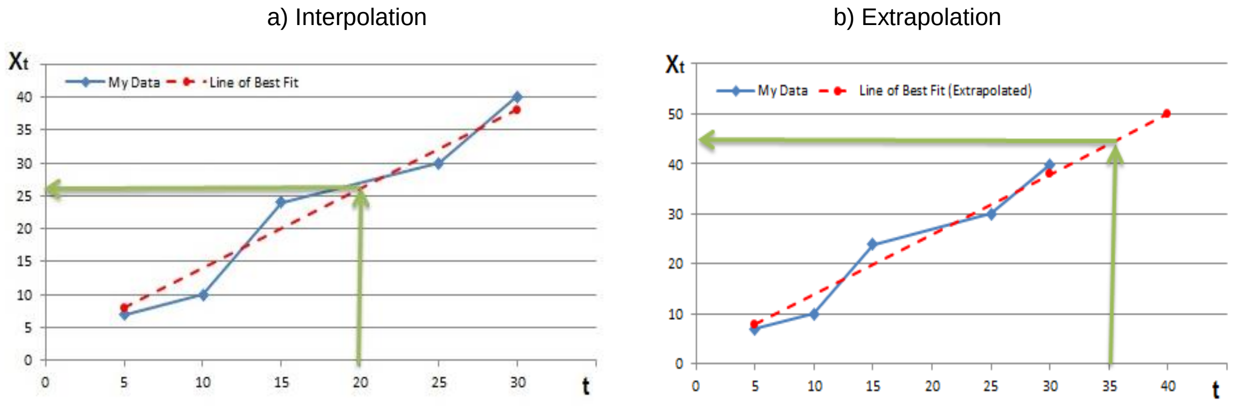
Prediction intervals in forecasting predict the distribution of individual future points, whereas confidence intervals in statistics predict the distribution of estimates of the true population mean or other measures of interest that cannot be observed. For example, if we make the parametric assumption that the underlying distribution is a normal distribution, and has a sample set $\{X_1, ..., X_n\}$, then confidence intervals may be used to estimate the population mean μ and population standard deviation σ of the underlying population, while prediction intervals may be used to estimate the value of the next sample variable[3], X_{n+1} (see Fig.5-6).

Another simple technique is the manual estimation of the best-fit line, i.e. using ***the rule of thumb***[4]. Manually, we can make an estimate of a line of best fit and draw it on the chart by adding a new data series consisting of two points (see Fig.5-7). If we want to know what the measurement would be for a location where no measurement was taken (known as ***Interpolation*** task), we can use the chart and two quick lines to show that for 20th period (i.e. when **t**=20) we would expect a measurement of about 26 units, i.e. $\mathbf{X_t}$=26 (Fig.5-7 a).

Alternatively, ***extrapolating*** the line of Best-Fit beyond the observed data can be used (the same technique can be applied to estimate what some future value may be). In the chart in Fig.5-7 b) (an example for the 35th period, i.e. when **t**=35), we can see that the forecasted value would be of about 45 units for period 35, i.e. $\mathbf{X_t}$=45.

[3] Prediction interval construction will be discussed in Chapter 6.

[4] A rule of thumb is a principle with broad application that is not intended to be strictly accurate or reliable for every situation.


a) Interpolation

b) Extrapolation

Fig.5-7 Manual estimation of the best-fit line

Benefits of the Manual estimation (rule of thumb):

- Applicable to simple models;
- Can be used without a computer or a calculator in the field;
- Gives the user a better feel for the data.

Problems of the Manual estimation (rule of thumb):

- Only applicable to simple models;
- Reliant on the accuracy of our estimate of the trend;
- No measure of how accurately our estimate fits the data.

B. Naïve Approach

Naïve forecasts are some of the most cost-effective forecasting models that provide a benchmark against which more sophisticated models can be compared. They are simple and sometimes surprisingly effective. For cross-sectional and stationary time series data, this approach states that the forecast for any period equals the historical average. For time series data that are stationary in terms of their first differences, the naïve forecast equals the previous period's actual value and so on. The most common naïve techniques are discussed below:

Average Method

For purposes of statistical forecasting, the simplest case is that of a variable whose values are independently and identically randomly distributed, like the example presented in Fig.5-8. The values of this time series appear to have been independently drawn from a common probability distribution, suggesting that future observations will be drawn from the same distribution. The natural forecast to use for all future values is, therefore, the sample mean of the past data because by definition the mean is an unbiased estimator and also it minimizes the mean squared forecasting error.

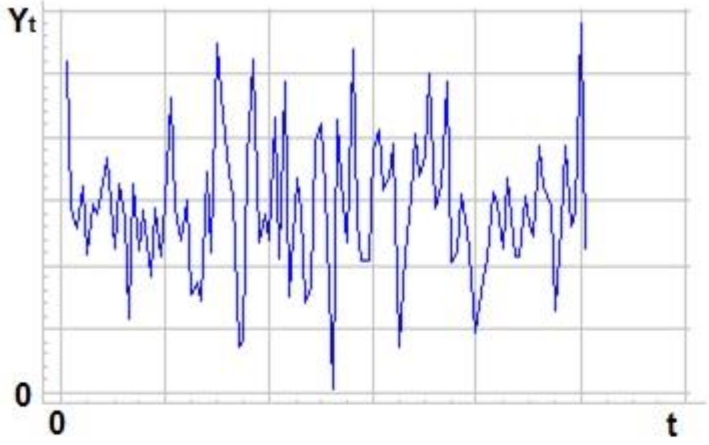


Fig.5-8 Example of stationary variable $\boldsymbol{Y_t}$

Hence, the forecasts of all future values can be computed as the mean of the observed historical data. If we let the historical data be denoted by $Y = \{Y_t\text{: } t \in T\}$, where T is the index set, then the forecast equation is simply:

$$\boldsymbol{Y^*_{(T+1)} = \overline{Y} = (Y_1 + \cdots + Y_T)/T} \qquad (5\text{-}1)$$

(The notation $\boldsymbol{Y^*_{(T+1)}}$ is a short-hand for the estimate of $\boldsymbol{Y_{(T+1)}}$ based on the data Y_1 ... Y_T)

The equation (5-1) may seem too simple and obvious to be of much importance, but it is actually the building block for a number of more sophisticated models, which will be discussed in the next chapters of the book.

This technique can also be used for cross-sectional data when we are predicting a value not included in the dataset. Then the prediction for this value is the average of the values that have been observed. The remaining techniques are only applicable to time series data.

Naïve Forecast (Random Walk)

In the simple naïve approach, the forecast is set to be the value of the last observation. That is, the forecasts of all future values $\boldsymbol{Y^*_{(T+1)}}$ are set to be $\boldsymbol{Y_T}$, where $\boldsymbol{Y_T}$ is the last observed value:

$$\boldsymbol{Y^*_{(T+1)} = Y_T} \qquad (5\text{-}2)$$

Sometimes, this technique works remarkably well for many economic and financial time series. In fact, most naturally-occurring time series in business and economics are not at all stationary (at least when plotted in their original units). Usually, they exhibit various kinds of trends, cycles, and seasonal patterns. Fig.5-9 represents a time series $Y = \{Y_t\text{: } t \in T\}$, which exhibits steady, if somewhat irregular, linear growth. The average technique described above would obviously be inappropriate here.

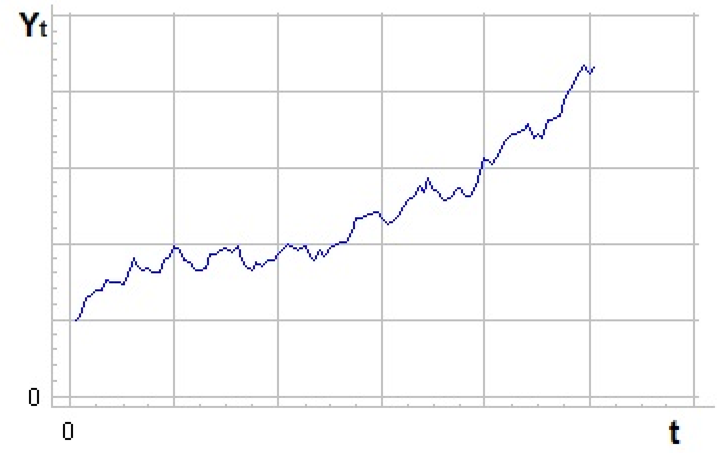


Fig.5-9 Example of non-stationary variable $\boldsymbol{Y_t}$

When faced with similar time series data that show irregular growth, the best strategy may not be to try to directly predict the level of the series at each period $\boldsymbol{Y_t}$. Instead, it may be better to try to predict the change that occurs from one period to the next ($\boldsymbol{Y_t - Y_{t-1}}$).

Very often it is helpful to look at the first difference of the series, to see if a predictable pattern can be discerned there. For practical purposes, it is just as good to predict the next change as to predict the next level of the series, since the predicted change can always be added to the current level to yield a predicted level. Fig. 5-10 displays the first difference of the irregular growth series presented above (see Fig. 5-9).

The new variable $\boldsymbol{d_t}$ looks stationary and quite random like the stationary variable $\boldsymbol{Y_t}$ from Fig.5-8, i.e. it shows a pattern that we previously fitted with the average model. Hence, we can apply the equation (5-1), which now will look as follows:

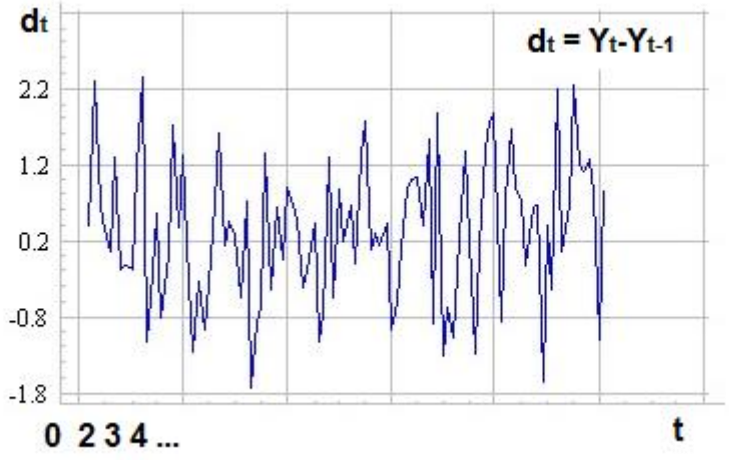


Fig.5-10 First differences of the non-stationary variable $\boldsymbol{Y_t}$

(Note: there is no $\boldsymbol{d_t}$ for $\boldsymbol{t=1}$)

$$(\boldsymbol{Y_t}\text{-}\boldsymbol{Y_{t-1}}) = \boldsymbol{\bar{d}} = (\mathbf{d_2}+\cdots+\mathbf{d_t})/(\boldsymbol{T}\text{-}\mathbf{1}) \tag{5-3}$$

Where $\boldsymbol{\bar{d}}$ (the constant term) is the mean of the first differences, i.e., the average change from one period to the next. Solving (5-3) for $\boldsymbol{Y_t}$ will return new equation:

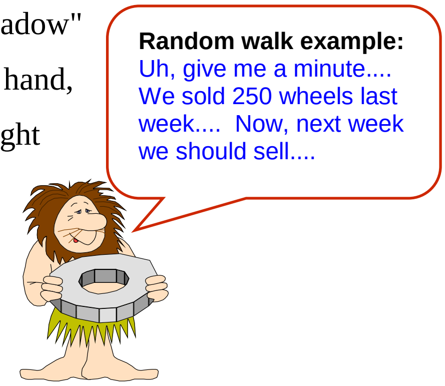
$$\boldsymbol{Y_t} = \boldsymbol{Y_{t-1}} + \boldsymbol{\bar{d}} \tag{5-4}$$

Consequently, the forecasting model is:

$$\boldsymbol{Y^*_{(T+1)}} = \boldsymbol{Y_T} + \boldsymbol{\bar{d}} \tag{5-5}$$

In other words, we predict that this period's value will equal the last period's value plus a constant representing the average change between periods. This naïve technique is often referred to as "***random walk***"[5] model. It assumes that, from one period to the next, the original time series merely takes a random "step" away from its last recorded position. If the constant term in the random walk model is zero, then equation (5-5) returns the same forecast as the simple naïve model (5-2). This technique is known as ***"random walk without drift"***.

Notice that the one-step forecasts within the sample merely "shadow" the observed data, lagging exactly one period behind. On the other hand, the long-term forecasts outside the sample follow a horizontal straight line anchored on the last observed value, i.e. the horizontal appearance of the long-term forecasts is rather unsatisfactory if we believe that the upward trend observed in the past is genuine.


Random walk example:
Uh, give me a minute.... We sold 250 wheels last week.... Now, next week we should sell....

Drift Method (Random Walk with Drift)

A variation on the naïve method is to allow the forecasts to increase or decrease over time, where the amount of change over time (known as the ***drift***) is set to be the average change seen in the historical data, i.e. equation (5-5). Then, the general forecast for period $\boldsymbol{T+l}$ is:

$$\boldsymbol{Y^*_{(T+l)}} = \boldsymbol{Y_T} + (\boldsymbol{l} \times \boldsymbol{\bar{d}}) \tag{5-6}$$

This is equivalent to drawing a line between the first and last observation, and extrapolating it into the future. It is useful when the time series being fitted by a random walk model has an average upward (or downward) trend that is expected to continue in the future. Then we should include a non-zero constant term in equation (5-5), assuming that the random walk undergoes a "drift". Hence, this technique is also known as a ***random-walk-with-drift.***

[5] The traces of an inebriated person (who steps randomly to the left or right at the same time as he steps forward) are an example of a ”random walk”.

Seasonal Naïve Technique

A similar method is useful for highly seasonal data. In this case, we set each forecast to be equal to the last observed value from the same season of the year (the same month of the previous year for instance). Formally, the forecast for period ***T+l*** is:

$$Y^*_{(T+l)} = Y_{(T+l-km)} \qquad (5\text{-}7)$$

where ***m*** is the seasonal period;

$k = [(l-1)/m] + 1$ and **[…]** denotes the integer part of the algebraic expression within [].

It looks more complicated than it really is. For example, with monthly data, the forecast for all future January values is equal to the last observed January value. With quarterly data, the forecast of all future Q3 values is equal to the last observed Q3 value (where Q3 means the third quarter). Similar rules apply for other months and quarters, and for other seasonal periods.

Fig.5-11 shows the three naïve methods without drift applied to the Australian quarterly beer production data[6]. Sometimes one of these simple techniques will be the best forecasting method available. But in most cases, these methods will serve as benchmarks rather than the method of choice. Their errors should be used as a reference model accuracy in equation (3-6) when computing the forecast skill, or skill score, ***SS***. There, a perfect forecast results in an ***SS*** of one, a forecast with similar skill to the reference forecast would have a skill of zero, and a forecast which is less skillful than the reference forecast would have negative skill values. In other words, whatever forecasting models we develop, they should be compared to these simple models to ensure that the new one is better than these simple alternatives. If not, then the new forecasting model is not worth considering.

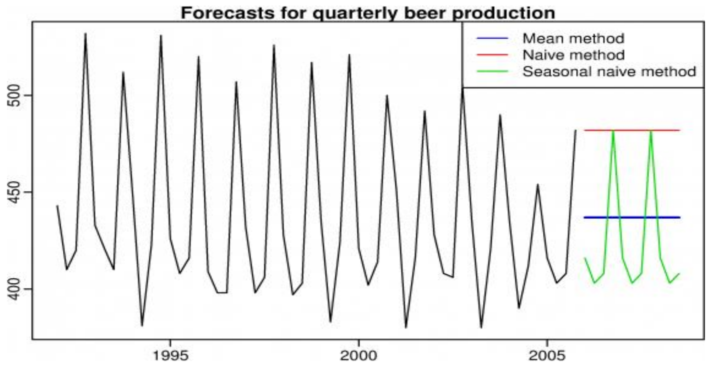


Fig.5-11 Comparison between three naïve methods (Source: https://www.otexts.org/fpp/2/3)

[6] Source: https://www.otexts.org/fpp/2/3

6.2. Moving Averages

A Moving average is a calculation technique to analyze data points by creating a series of averages of different subsets of the full data set. It is also known as a moving mean or rolling mean, rolling average or running average. Given a series of numbers and a fixed subset size, the first element of the moving average is obtained by taking the average of the initial fixed subset of the number series (see eq. 5-8). Then the subset is modified by "shifting forward"; that is, excluding the first number of the series and including the next number following the original subset in the series. This creates a new subset of numbers, which is averaged. This process is repeated over the entire data series. The plotline connecting all the (fixed) averages is the moving average. Viewed simplistically it can be regarded as smoothing the data.

The basic assumption behind averaging and smoothing models is that the time series is locally stationary with a slowly varying mean. First, we take a moving (local) average to estimate the current value of the mean and then use that as the forecast for the near future. This can be considered as a compromise between the average model and the random-walk-without-drift-model. The same strategy can be used to estimate and extrapolate a local trend. A moving average is often called a "smoothed" version of the original series because short-term averaging has the effect of smoothing out the bumps in the original series. By adjusting the degree of smoothing (the width of the moving average), we can hope to strike some kind of optimal balance between the performance of the mean and random walk models.

A moving average is commonly used with time series data to smooth out short-term fluctuations and highlight longer-term trends or cycles. The threshold between short-term and long-term depends on the application, and the parameters of the moving average will be set accordingly. For example, it is often used in technical analysis of financial data, like stock prices, returns or trading volumes. It is also used in economics to examine the gross domestic product, employment or other macroeconomic time series.

A moving average may also use unequal weights for each value in the subset to emphasize some particular values. Variations include simple, cumulative, or weighted forms.

<u>Simple (Equally-Weighted) Moving Average</u>

The simplest way to smooth a time series is to calculate a simple, or equally-weighted, moving average (***MA***). The forecast (or the smoothed value) for ***Y*** ($Y = \{Y_t\colon t \in T\}$) at time ***t+1*** that is made at time ***t*** equals the simple average of the most recent ***m*** observations:

$$\boxed{\hat{Y}_{t+1} = \frac{Y_t + Y_{t-1} + \dots + Y_{t-m+1}}{m}} \tag{5-8}$$

where "***Y-hat***" or $\hat{Y}$ stands for a forecast of the time series ***Y*** ($Y = \{Y_t: t \in T\}$) made at the earliest possible prior date by a given model.

A moving average of order ***m*** can be written as

$$\hat{Y}_t = \frac{1}{m}\sum_{j=-k}^{k} Y_{t+j} \qquad (5\text{-}9)$$

where ***m=2k+1***, that is the smoothed value at time ***t***, is obtained by averaging values of the time series within ***k*** periods of ***t***.

Observations that are nearby in time are also likely to be close in value, and the average eliminates some of the ***randomness*** (also known as ***noise***) in the data, leaving a smooth trend-cycle component. It is known also as ***m-MA***, i.e. a moving average of order ***m***.

The forecast (5-8) is an average, centered at period (***t-(m+1)/2***), which implies that the estimate of the local mean will tend to lag behind the true value of the local mean by about *(m+1)/2* periods. Thus, we say the average age of the data in the simple moving average is *(m+1)/2* relative to the period for which the forecast is computed. This is the amount of time by which forecasts will tend to lag behind turning points in the data, i.e. if we are averaging the last 5 values, the forecasts will be about 3 periods late in responding to turning points.

This introduces a phase shift into the data of half the time window (i.e. the time period used) length. For example, if the observations were all the same except for one high data point, the peak in the smoothed data would appear half a window length later than when it actually occurred. Where the phase of the result is important, this can be simply corrected by shifting the resulting series back by half the time window length.

For a number of applications, it is advantageous to avoid the shifting induced by using only past data. Hence a central moving average can be computed, using data equally spaced on either side of the point in the series where the mean is calculated. This requires using an odd number of observations in the MA time window.

When ***m=1***, the ***simple MA model*** is equivalent to the ***random walk model without growth***. If ***m*** is very large (comparable to the length of the estimation period), the simple MA model is equivalent to the ***average model***. As with any parameter of a forecasting model, it is customary to adjust the value of ***k*** in order to obtain the best "fit" to the data, i.e., the smallest forecast errors on average.

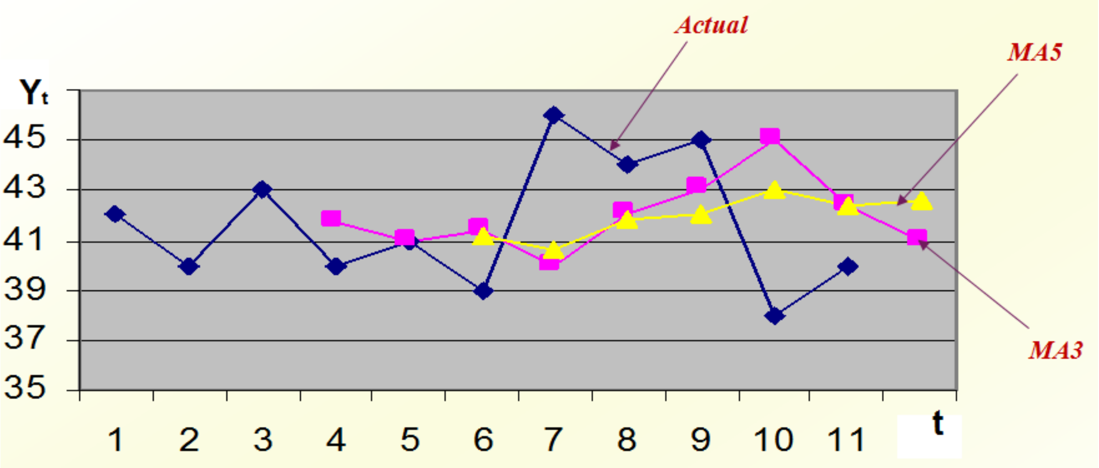


Fig.5-12 Visual comparison between actual data, three and five period MA forecasts

The choice of an integer ***k > 1*** is arbitrary. A small value of ***k*** will have less of a smoothing effect and be more responsive to recent changes in the data, while a larger ***k*** will have a greater smoothing effect, and produce a more pronounced lag in the smoothed series (see Fig.5-12). In choosing the value of ***k***, in fact, we are making a tradeoff between two effects – ***filtering (smoothing) out*** more ***noise (randomness)*** vs. being too slow to respond to trends and turning points.

One disadvantage of this technique is that it cannot be used on the first ***k−1*** terms of the time series without the addition of values created by some other means. This means effectively extrapolating outside the existing data, and the validity of this section would, therefore, be questionable and not a direct representation of the data.

If the data used are not centered around the mean, a simple MA lags behind the latest observation by half of the time window length. A simple MA can also be disproportionately influenced by old values dropping out or new data coming in. Another characteristic of the simple MA is that if the data have a periodic fluctuation, then applying a simple MA of that period will eliminate that variation because the average always contains one complete cycle. Fortunately, a perfectly regular cycle is rarely encountered in real-life business.

A major drawback of the simple MA is that it lets through a significant amount of the "***signal***"[7] shorter than the time window length. Worse, it actually *inverts it*. This can lead to unexpected events, such as peaks in the smoothed data appearing where there were troughs in the observations. It also leads to the result being less smooth than expected since some of the higher frequencies are not properly removed.

[7] A ***signal*** as referred to in communication systems "is a ***function that conveys information about the behavior or attributes of some phenomenon***". The information in a signal is usually accompanied by ***noise*** - an error or undesired random disturbance of a useful information signal.

The problem can be overcome by iterating the process three times, with the window being shortened by a factor of 1.4303 at each step[8]. This removes the negation effects and provides a better-behaved model. This solution is often used in real-time audio filtering since it is computationally quicker than other comparable filters.

The simple MA model can be easily customized in several ways to fine-tune its performance. If there is a consistent trend in the data, then the forecasts of any of the simple MA models will be biased, because they do not contain any trend component. The presence of a trend will tend to give an edge to models with lower values of ***m***, regardless of the amount of noise that needs to be smoothed out. We can fix this problem by simply adding a constant to the simple MA forecasting equation, analogous to the drift term in the random-walk-with-drift model. The new, ***Simple Moving Average with Trend*** is:

$$\hat{Y}_{t+1} = \frac{Y_t + Y_{t-1} + \ldots + Y_{t-m+1}}{m} + d \qquad (5\text{-}10)$$

Another way to fine-tune the simple MA model is to use a ***Tapered Moving Average*** rather than an equally weighted MA. For example, in the 5-order MA model, we could choose to put only half as much weight on the newest and oldest values, like this:

$$\hat{Y}_{t+1} = \frac{\frac{1}{2}Y_t + Y_{t-1} + Y_{t-2} + Y_{t-3} + \frac{1}{2}Y_{t-4}}{4} \qquad (5\text{-}11)$$

This average is centered on three periods in the past, like the 5-period simple MA model, but when an unusually large or small value is observed, it doesn't have as big an impact when it first arrives or when it is finally dropped out of the calculation, because its weight is ramped up or down over two periods. In other words, the ***tapered moving average*** is more robust to outliers in the historical data.

Cumulative Moving Average

In a ***cumulative moving average*** (***CMA***), the data arrive in an ordered data series, and the user would like to get the average of all of the data up until the current observation. For example, an investor may want the average price of all of the stock transactions for a particular stock up until the current time. As each new transaction occurs, the average price at the time of the transaction can be calculated for all of the transactions up to that point using the cumulative

[8] For more details see http://climategrog.wordpress.com/2013/05/19/triple-running-mean-filters/

average, typically an equally weighted average of the sequence of ***m*** values ***Y*** ($Y = \{Y_t: t \in T\}$) up to the current time:

$$CMA_m = \frac{Y_1 + Y_2 + \cdots + Y_m}{m} = \frac{1}{m}\sum_{i=1}^{m} Y_i \tag{5-12}$$

The brute-force method to calculate this would be to store all of the data and calculate the sum and divide by the number of data every time a new observation arrived. However, it is possible to simply update the cumulative average as a new value $\boldsymbol{Y_{m+1}}$, becomes available, using the formula:

$$CMA_{m+1} = \frac{Y_{m+1} + m.CMA_m}{(m+1)} \tag{5-13}$$

where $\boldsymbol{CMA_0}$ can be taken to be equal to zero.

Thus, the current cumulative average for a new observation is equal to the previous cumulative average, times ***m***, plus the latest observation $\boldsymbol{Y_{m+1}}$, all divided by the number of data received so far, ***m+1***. When all of the observations arrive (***m = T***), then the cumulative average will equal the final average.

Weighted Moving Average

A weighted average is any average that has multiplying factors to give different weights to data at different positions in the sample. In general, a weighted ***m-MA*** can be written as:

$$\hat{Y}_t = \sum_{j=-k}^{k} \mathrm{w_j Y_{t+j}} \tag{5-14}$$

where ***k=(m−1)/2*** and the weights are given by ***w*** = {$\boldsymbol{w_{-k},\ldots,w_k}$}. It is important that all the weights sum to one and that they are symmetric so that $\boldsymbol{w_j=w_{-j}}$.

The ***Random Walk*** model is the special case in which ***m=1***. The simple ***m-MA*** is a special case where all the weights are equal to ***1/m***. It is important to note that the simple MA model has the following properties:

- as ***m*** gets larger each individual observation in the recent past receives less weight, because each of the past ***m*** observations has a weight of ***1/m*** in the averaging formula (5-14). This implies that larger values of ***m*** will filter out more of the period-to-period randomness and yield smoother-looking series of forecasts.
- the first term in the average is “1 period old” relative to the point in time for which the forecast is being calculated; the second term is two periods old, and so on up to the $\boldsymbol{m^{th}}$ term. Hence, the “***average age***” of the data in the forecast is ***(m+1)/2***. This is the amount

by which the forecasts will tend to lag behind in trying to follow trends or respond to turning points. For example, with m=5, the average age is 3, so that is the number of periods by which forecasts will tend to lag behind what is happening now.

Notice that the long-term forecasts from the simple MA model are a horizontal straight line, just as in the random walk model. Thus, the simple MA model assumes that there is no trend in the data. However, whereas the forecast from the random walk model is simply equal to the last observed value, the forecasts from the simple MA model are equal to a weighted average of recent values.

A slightly more intricate method for smoothing a raw time series $\boldsymbol{X = \{X_1, X_2, ..., X_T\}}$ is to calculate a weighted MA by first choosing a set of weighting factors $\boldsymbol{w} = \{\boldsymbol{w_1, w_2, \ldots, w_k}\}$ such that:

$$\sum_{n=1}^{k} w_n = 1$$

and then using these weights to calculate the smoothed statistics $\boldsymbol{s_t} = \{\boldsymbol{s_1, s_2, ..., s_k}\}$

$$s_t = \sum_{n=1}^{k} w_n x_{t+1-n} = w_1 x_t + w_2 x_{t-1} + \cdots + w_k x_{t-k+1}. \qquad (5\text{-}15)$$

A major advantage of weighted moving averages is that they yield a smoother estimate of the trend-cycle. Instead of observations entering and leaving the calculation at full weight, their weights are slowly increased and then slowly decreased resulting in a smoother curve. In practice, the weighting factors ***w*** are often chosen to give more weight to the most recent terms in the time series and less weight to older data.

It is worth noting that this technique has the same disadvantage as the simple MA technique, i.e. it cannot be used until at least ***k*** observations have been made and that it entails a more complicated calculation at each step of the smoothing procedure. In addition to this, if the data from each stage of the averaging is not available for analysis, it may be difficult if not impossible to reconstruct a changing signal accurately (because older samples may be given less weight). If the number of stages missed is known, however, the weighting of values in the average can be adjusted to give equal weight to all missed samples to avoid this issue.

Other weighting systems are used occasionally – for example, in share trading, a volume weighting will weight each time period in proportion to its trading volume. Another weighting, used by actuaries, is a 15-ordered MA (a central moving average). The symmetric weight coefficients are −3, −6, −5, 3, 21, 46, 67, 74, 67, 46, 21, 3, −5, −6, −3.

Outside the world of finance, weighted MA has many forms and applications. Each weighting function has its own characteristics. A mean does not just "smooth" the data. A mean is a form of low-pass filter. The effects of the particular filter used should be understood in order to make an appropriate choice.

Moving Median

From a statistical point of view, the MA, when used to smooth data or to estimate the underlying trend in a time series, is susceptible to rare events such as rapid shocks or other anomalies (i.e. extreme values), since in computing formula (5-9) it takes into account all values from the particular data set. A more robust estimate of the trend is the ***simple moving median (MM)*** of the time series ***Y*** ($Y = \{Y_t \colon t \in T\}$) over ***n*** time points:

$$\boldsymbol{MM = Median\ (Y_t, Y_{t-1}, \dots, Y_{t-n+1})} \tag{5-16}$$

where the ***Median*** is the middle value in the sorted time series.

Statistically, the MA is optimal for recovering the underlying trend of the time series when the fluctuations about the trend are normally distributed. However, the normal distribution does not place a high probability on very large deviations from the trend which explains why such deviations will have a disproportionately large effect on the trend estimate. If the fluctuations are instead assumed to be Laplace distributed, then the ***MM*** is statistically optimal. For a given variance, the Laplace distribution places a higher probability on rare events than does the normal, which explains why the ***MM*** tolerates shocks better than the ***MA***.

Moving Average Regression Model

In a ***moving average regression model***, a variable of interest $\boldsymbol{X = \{X_1, X_2, ..., X_T\}}$ is assumed to be a weighted moving average of an unobserved error term and the weights in the moving average are parameters to be estimated:

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \tag{5-17}$$

where μ is the mean of the series, the θ_1, ..., θ_q are the parameters of the model and the ε_t, ε_{t-1}... are ***white noise***[9] error terms. The value of q is called the order of the MA model.

Thus, a moving-average model is conceptually a ***linear regression*** of the current value of the series against current and previous (unobserved) random variations. The random variations at each point are assumed to be mutually independent and to come from the same, typically a

[9] White noise refers to a ***statistical model for signals*** and signal sources, rather than to any specific signal.

normal distribution. This and other similar forecasting models will be discussed in Chapters 6, 7 and 8.

5.3. Exponential Smoothing

Exponential smoothing was proposed in the late 1950s by Robert Brown (1956) and then expanded by Charles Holt (1957) as double exponential smoothing. The term triple exponential smoothing was first suggested by Holt's student, Peter Winters (1960). Exponential smoothing is a technique that can be applied to time series data, either to produce smoothed data for presentation, or to make forecasts. The observed phenomenon may be an essentially random process or it may be an orderly, but noisy, process.

Using the naïve method (5-2), all forecasts for the future are equal to the last observed value of the series, i.e. it assumes that the most current observation is the only important one and all previous observations provide no information for the future. This can be thought of as a weighted average (5-15) where all the weight is given to the last observation.

Using the average method (5-1), all future forecasts are equal to a simple average of the observed data, i.e. it assumes that all observations are of equal importance and they are given equal weight when generating forecasts. We often want something between these two extremes. For example, it may be sensible to attach larger weights to more recent observations than to observations from the distant past.

The simple MA has the undesirable property that it treats the last ***k*** observations equally and completely ignores all preceding observations. Intuitively, past data should be discounted in a more gradual fashion--for example, the most recent observation should get a little more weight than 2nd most recent, and the 2nd most recent should get a little more weight than the 3rd most recent, and so on. In other words, the more recent the observation, the higher the associated weight. Whereas in the simple MA the past observations are weighted equally, exponential smoothing assigns exponentially decreasing weights over time.

Brown's Simple Exponential Smoothing (exponentially weighted moving average)

The simplest of the exponentially smoothing methods is known as ***simple exponential smoothing (SES), exponential moving average (EMA)***, as well as ***exponentially weighted moving average (EWMA)***. This method is suitable for forecasting data without any systematic trend or seasonal pattern. The formulation below, which is most commonly used, is attributed to Brown (1956) and is known as "Brown's simple exponential smoothing":

The raw data sequence is represented by $x_t = \{x_1, x_2, ..., x_T\}$ and the output of the exponential smoothing algorithm is written as $s_t = \{s_1, s_2, ..., s_T\}$, which may be regarded as the best estimate of what the next value of x_t will be. When the sequence of observations begins at time t = 0, the simplest form of exponential smoothing is given by the formula:

$$s_t = \alpha \cdot x_{t-1} + (1-\alpha) \cdot s_{t-1} \qquad (5\text{-}18)$$

where α is the ***smoothing factor***, and $0 < \alpha < 1$, i.e. the smoothed statistic s_t is a simple weighted average of the previous observation x_{t-1} and the previous smoothed statistic s_{t-1}.

In other words, as time passes the smoothed statistic s_t becomes the weighted average of a greater and greater number of the past observations x_{t-n}, and the weights assigned to previous observations are in general proportional to the terms of the geometric progression $\{1, (1-\alpha), (1-\alpha)^2, (1-\alpha)^3, \ldots\}$. A geometric progression is the discrete version of an exponential function, so this is where the name for this smoothing method originated.

The term ***smoothing factor*** applied to α here is an unsuitable name, as larger values of α actually reduce the level of smoothing, and when $\alpha = 1$ the output series $s_t = \{s_1, s_2, \ldots, s_T\}$ is just the same as the original series $x_t = \{x_1, x_2, \ldots, x_T\}$ with lag of one time unit. In fact, when $\alpha=1$, then the SES model is equivalent to a random walk model without growth. If $\alpha=0$, the SES model is equivalent to the average model, assuming that the first smoothed value s_1 is set equal to the mean.

Values of α close to one have less of a smoothing effect and give greater weight to recent changes in the data, while values of α closer to zero have a greater smoothing effect and are less responsive to recent changes (see Fig.5-13). There is no formally correct procedure for choosing α. Sometimes the researcher's judgment is used to choose an appropriate factor, or simulations with different values for α could be performed to select the most suitable one. Alternatively, a statistical technique like ***Least Squares (LS)*** method mentioned above (see Chapter 6) may be used to optimize the value of α, i.e. to determine the value of α for which the sum of the squared differences $(s_{n-1}-x_{n-1})^2$ is minimized.

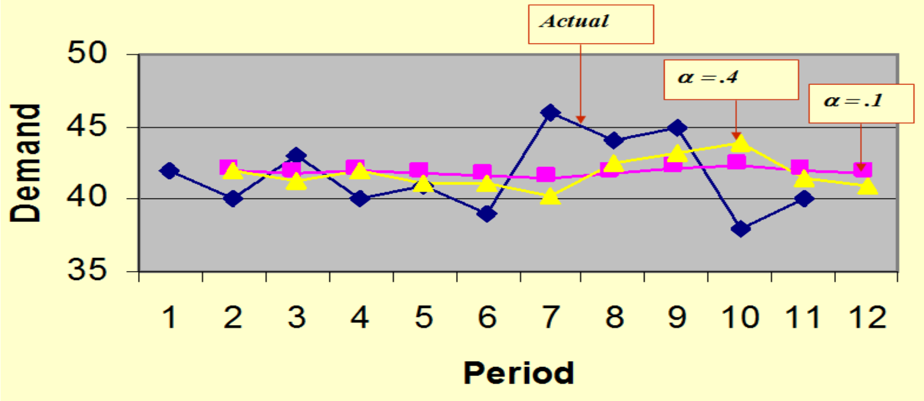


Fig.5-13 Visual comparison between actual data and different ***SES*** forecasts

Unlike some other smoothing methods, this technique does not require any minimum number of observations to be made before it begins to produce results. Simple exponential smoothing is easily applied, and it produces a smoothed statistic as soon as two observations are available.

In practice, however, a "good average" will not be achieved until several samples have been averaged together. For example, a constant signal will take approximately $3/\alpha$ stages to reach 95% of the actual value. To accurately reconstruct the original signal without information loss all stages of the exponential moving average must also be available because of older samples corrupt in weight exponentially. This is in contrast to a simple moving average, in which some samples can be skipped without much loss of information due to the constant weighting of samples within the average. If a known number of samples will be missed, one can adjust a weighted average for this as well, by giving equal weight to the new sample and all those to be skipped.

Choosing the initial smoothed value is another issue since s_1 is undefined. S_1 may be initialized in a number of different ways, most commonly by setting s_1 to x_1, though other techniques exist, such as setting s_1 to an average of the first four or five observations. The importance of the s_1 initializations effect on the resultant moving average depends on α – smaller α values make the choice of s_1 relatively more important than larger α values, since a higher α discounts older observations faster.

Whatever is done for s_1 it assumes something about values prior to the available data and if necessary, in errors. In view of this the early results should be regarded as unreliable until the iterations have had time to converge (in other words, to approach a given number). This is sometimes called a 'spin-up' interval. One way to assess when it can be regarded as reliable is to consider the required accuracy of the result. For example, if 3% accuracy is required, initializing with x_1 and taking data after five-time constants (defined above with eq. 5-18) will ensure that the calculation has converged to within 3% (only <3% of x_1 will remain in the result). Sometimes with a very small alpha, this can mean little of the result is useful. This is analogous to the problem of using a weighted average with a very long time window.

Exponential smoothing and MA have similar defects of introducing a lag relative to the input data. While this can be corrected by shifting the result by half the window length for a symmetrical kernel, such as a moving average (5-14), it is unclear how appropriate this would be for exponential smoothing. They also both have roughly the same distribution of forecast error when $\alpha = 2/(k+1)$. They differ in that exponential smoothing takes into account all past data, whereas MA only takes into account k past observations. From a computational point of view, they also differ in that MA requires that the past ***k*** data points be kept, whereas exponential smoothing only needs the most recent forecast value to be kept.

The basic step for an exponential forecast is equation (5-18). Here, for a given time series ***Y*** ($Y = \{Y_t: t \in T\}$) we can express the next forecast ***Y-hat*** directly in terms of previous forecasts and previous observations, in any of the following equivalent equations as pointed out in Nau (2014). They are all mathematically equivalent and any one of them can be obtained by rearrangement of any of the others.

In the first version, the forecast is computed by interpolating between the last observed value and the forecast that had been made for it:

$$\hat{Y}_{t+1} = \alpha Y_t + (1-\alpha)\hat{Y}_t \tag{5-19}$$

Written in this way, it is clear that the random walk model is an SES model with ***α=1***, and the constant-forecast model (of which the average model is a special case) is an SES model with ***α=0***. Hence the SES model is an interpolation between the average model and the random walk model with respect to the way it responds to new data. In general, this model performs better than either of them in situations where the random walk model over-responds and the average model under-responds.

In the second version, the forecast is a function of the previous forecast plus a fraction ***α*** of the previous error:

$$\hat{Y}_{t+1} = \hat{Y}_t + \alpha e_t \tag{5-20}$$

where $e_t = Y_t - \hat{Y}_t$ is the error made at time t.

Thus, the last forecast is adjusted in the direction of the error it made. If the error was positive, i.e., if the previous forecast was too low, then the next forecast is adjusted upward by a fraction ***α*** of that error. This version provides a nice interpretation of the meaning of "alpha," namely that ***α*** is the fraction of the forecast error that is believed to be due to an unexpected change in the level of the series rather than an unexpected one-time event. Models with larger values of ***α*** assume that what they are seeing are significant changes in the fundamental level of the series from one period to the next. In the limit as ***α***→1 (which is the random walk model) all of the variation from one period to the next is believed to be due to a change in the fundamental level rather than just a temporary deviation. In the limit as ***α***→0 (which is the constant model), the fundamental level of the series is assumed to never change, and all of the period-to-period variation is attributed to temporary deviations from it.

The forecast can also be written as an exponentially weighted (i.e. discounted) moving average of all past values with the discount factor 1-α:

$$\hat{Y}_{t+1} = \alpha[Y_t + (1\text{-}\alpha)Y_{t-1} + (1\text{-}\alpha)^2Y_{t-2} + (1\text{-}\alpha)^3Y_{t-3} +. ..] \tag{5-21}$$

The exponentially-weighted-moving-average form of SES model highlights the difference between it and the simple MA model. The SES forecast uses all past values but discounts their weights by a factor of ***1-α*** per period, while the simple ***m-MA*** model uses only the last ***m*** values and gives them equal weights of ***1/m***.

One important weakness, as mentioned above, concerns the trend. The SES model, like the simple ***m-MA*** model, assumes that there are no trends in the data, either short-term or long-term. Based on the ***random-walk-with-drift*** idea (5-6), the simplest solution for SES would be to modify equation (5-20) like the simple MA model (5-10), in order to incorporate a long-term linear trend by merely adding a drift term to the forecasting equation:

$$\hat{Y}_{t+1} = Y_t - (1\text{-}\alpha)e_t + d \tag{5-22}$$

where ***d*** is the average long-term growth for the whole time series. The basic component in equation (5-22) comes from (5-20) with the substitution that $e_t = Y_t - \hat{Y}_t$, i.e.

$$\hat{Y}_{t+1} = \alpha Y_t + (1-\alpha)\widehat{Y_t} = aY_t + \widehat{Y_t} - \alpha\widehat{Y_t} = \alpha(Y_t - \widehat{Y_t}) + \widehat{Y_t} = \alpha e_t + Y_t - e_t$$

$$\hat{Y}_{t+1} = Y_t - (1-\alpha)e_t$$

Holt’s Linear (Double) Exponential Smoothing

The simple ***m-MA*** models and the ***SES*** models assume that there is no trend of any kind in the data, which is usually OK or at least not-too-bad for one-step-ahead forecasts (Nau, 2014, p. 16). Both models can be modified to incorporate a constant linear trend (i.e. the average long-term growth per period) as shown above in (5-22), but the problem with short-term trends still exists. If a series displays a varying rate of growth or a cyclical pattern and if there is a need to forecast more than one period ahead, then estimation of a local trend might also be an issue.

The simple exponential smoothing model can be generalized to obtain a ***linear exponential smoothing (LES)*** model that computes local estimates of both level and trend. It means adding a second equation with a second constant, β, which must be chosen in conjunction with α. For this case, several methods were developed under the name "double exponential smoothing", "second-order exponential smoothing" or ***Holt’s Linear Exponential Smoothing***. The ***LES*** model introduces a term to take into account the possibility of a time series exhibiting some form of trend. This slope component is itself updated via exponential smoothing.

The basic logic is the same as with ***SES***, but we now have two smoothing constants, one for smoothing the level and one for smoothing the trend. The raw data sequence is represented by time series *{x_t}* beginning at time *t*=0. Series *{s_t}* represents the smoothed value for time *t*, and {b_t} is the best estimate of the trend at time *t*. Then, double exponential smoothing is given by the formulas:

$$s_1 = x_1$$
$$b_1 = x_1 - x_0$$

And for **t** > **1** by:

$$s_t = \alpha x_t + (1-\alpha)(s_{t-1} + b_{t-1})$$
$$b_t = \beta(s_t - s_{t-1}) + (1-\beta)b_{t-1} \tag{5-23}$$

where ***α*** is the *data smoothing factor* (0<α<1)

and ***β*** is the *trend smoothing factor* (0<β<1).

Note that the current value of the series s_t is used to calculate its smoothed value replacement in double exponential smoothing. The first smoothing equation adjusts s_t directly for the trend of the previous period, b_{t-1}, by adding it to the last smoothed value s_{t-1}. This helps to eliminate the lag and brings s_t to the appropriate base of the current value. The second equation then updates the trend, which is expressed as the difference between the last two values. The equation is similar to the basic form of single smoothing, but here applied to the updating of the trend.

The output of the algorithm (the forecast beyond x_t) is given as F_{t+m} and it is an estimate of the value of *x* at time ***t+m*** (***m>0***) based on the raw data up to time ***t***:

$$F_{t+m} = s_t + mb_t \tag{5-24}$$

The logic in the above algorithm could be better understood by following step-by-step explanations:

- At any time ***t*,** the model has an estimate s_t of the local level (the smoothed value) and an estimate b_t of the local trend. These are computed recursively from the value of ***X*** (or ***Y***) observed at time ***t*** and the previous estimates of the level and trend by two equations (5-23). If the estimated level and trend at time ***t-1*** are s_{t-1} and b_{t-1}, respectively, then the forecast for Y_t that would have been made at time ***t-1*** is equal to $s_{t-1}+b_{t-1}$. When the actual value is observed, the updated estimate of the level s_t is computed recursively by interpolating between ***Yt*** and its forecast, $s_{t-1}+b_{t-1}$, using weights of α and (1-α):

$$s_t = \alpha Y_t + (1-\alpha)(s_{t-1} + b_{t-1}) \tag{5-25}$$

- The change in the estimated level, namely (s_t–s_{t-1}), can be interpreted as a noisy measurement of the trend at time ***t***. The updated estimate of the trend is then computed recursively by interpolating between (s_t–s_{t-1}) and the previous estimate of the trend b_{t-1}, using weights of β and (1-β):

$$bt = \beta(s_t - s_{t-1}) + (1 - \beta)b_{t-1} \tag{5-26}$$

- Finally, the forecasts for the near future (for example next period t+1) that are made from time ***t*** are obtained by extrapolation of the updated level and trend:

$$\hat{Y}t{+}1 = s_t + mb_t \tag{5-27}$$

The interpretation of the trend-smoothing constant ***β*** is analogous to that of the level-smoothing constant ***α*** in SES. Models with small values of ***β*** assume that the trend changes only very slowly over time, while models with larger ***β*** assume that it is changing more rapidly. LES with a large ***β*** believes that the distant future is very uncertain because errors in trend-estimation become quite important when forecasting more than one period ahead.

Note that F_0 (or $\hat{Y}_0$) in (5-27) is undefined since there is no estimation for time ***t=0***. According to the definition $F_1=s_0+b_0$ is well defined and thus further values can be evaluated.

As in the case of single smoothing, there are a variety of schemes to set initial values for s_t and b_t in double smoothing. Setting the initial value b_0 is a matter of preference. S_1 is in general set to the first value of the time series y_1. Here are three suggestions for b_1[10].

$$b_1 = y_2 - y_1$$

$$b_1 = \frac{1}{3}\left[(y_2 - y_1) + (y_3 - y_2) + (y_4 - y_3)\right]$$

$$b_1 = \frac{y_n - y_1}{n - 1}$$

Alternatively, the smoothing constants ***α*** and ***β*** can be estimated using the ***Least Squares (LS) method***, by minimizing the mean squared error of the one-step-ahead forecasts.

[10] See http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc433.htm

Holt-Winters (Triple) Exponential Smoothing

Triple exponential smoothing was first suggested by Holt's student, Peter Winters (1960) but often it is referred to as Holt-Winters seasonal method. Triple exponential smoothing takes into account seasonal changes as well as trends. It comprises the forecast equation and three smoothing equations — one for the level s_t, one for trend b_t, and one for the seasonal component (or seasonal index) denoted by c_t, with smoothing parameters α, β, and γ.

Seasonality is defined to be the tendency of time-series data to exhibit behavior that repeats itself every ***m*** periods, i.e. ***m*** denotes the period of the seasonality, for example, the number of seasons in a year – for quarterly data m=4, for monthly data m=12 and so on.

$$\text{Observed series} = \text{Trend} + \text{Seasonal} + \text{Irregular} \qquad (5\text{-}28)$$

There are two variations to this method that differ in the nature of the seasonal component. The ***additive method*** (5-28) is preferred when the seasonal variations are roughly constant through the series, while the ***multiplicative method*** (5-29) is preferred when the seasonal variations are changing proportionally to the level of the series. For example, if every month of May we sell 2,000 more cars that we do in April the seasonality is ***additive*** in nature, however, if we sell 5% more cars in the summer months than we do in the spring months the seasonality is ***multiplicative*** in nature.

$$\text{Observed series} = \text{Trend} \times \text{Seasonal} \times \text{Irregular} \qquad (5\text{-}29)$$

With the additive method, the seasonal component is expressed in absolute terms in the scale of the observed series, and in the level equation for s_t (5-30) the series is seasonally adjusted by subtracting the seasonal component. Within each year the seasonal component will add up to approximately zero. With the multiplicative method, the seasonal component is expressed in relative terms (percentages) and the series is seasonally adjusted by dividing the seasonal component. Within each year, the seasonal component will sum up to approximately ***m***. The multiplicative model cannot be used when the original time series contains very small or zero values because it is not possible to divide a number by zero. In these cases, a pseudo additive model combining the elements of both the additive and multiplicative models is used.

The basic logic is similar to SES, but we now have three smoothing constants. The raw data sequence is represented by time series $\{x_t\}$ beginning at time $t=0$ with a cycle of seasonal change of length L. $\{s_t\}$ represents the smoothed value of the constant part at time t. $\{b_t\}$ represents the sequence of best estimates of the linear trend that are superimposed on the seasonal changes. $\{c_t\}$ is the sequence of seasonal correction factors and c_t is the expected proportion of the predicted trend at any time t in the cycle that the observations take on.

The method calculates a trend line for the data as well as seasonal indexes that weight the values in the trend line based on where that time point falls in the cycle of length ***L***. Triple exponential smoothing is given by the formulas (5-30):

$$\text{(5-30)}\qquad \begin{aligned} s_0 &= x_0 \\ s_t &= \alpha\frac{x_t}{c_{t-L}} + (1-\alpha)(s_{t-1}+b_{t-1}) && \text{Overall smoothing} \\ b_t &= \beta(s_t - s_{t-1}) + (1-\beta)b_{t-1} && \text{Trend smoothing} \\ c_t &= \gamma\frac{x_t}{s_t} + (1-\gamma)c_{t-L} && \text{Seasonal smoothing} \end{aligned}$$

where $\boldsymbol{\alpha}$ is the *data smoothing factor*, $0<\alpha<1$,
$\boldsymbol{\beta}$ is the *trend smoothing factor*, $0<\beta<1$,
$\boldsymbol{\gamma}$ is the *seasonal change smoothing factor*, $0<\gamma<1$.

The output of the algorithm (the forecast) $\boldsymbol{F_{t+m}}$, is an estimate of the value of ***x*** at time ***t+m*** ***(m>0)*** based on the raw data up to time ***t***.

$$F_{t+m} = (s_t + mb_t)c_{t-L+1+(m-1)} \qquad \text{(5-31)}$$

The general formula for the initial trend estimates $\boldsymbol{b_0}$ is given by:

$$b_0 = \frac{1}{L}\left(\frac{x_{L+1}-x_1}{L} + \frac{x_{L+2}-x_2}{L} + \ldots + \frac{x_{L+L}-x_L}{L}\right)$$

Setting the initial estimates for the seasonal indices $\boldsymbol{c_i}$ (i=1,2,...,L) is more complicated. As a rule of thumb, a minimum of two full seasons (i.e. ***2L*** periods) of historical data is needed to initialize a set of seasonal factors. If ***N*** is the number of complete cycles present in our data, then:

$$c_i = \frac{1}{N}\sum_{j=1}^{N}\frac{x_{L(j-1)+i}}{A_j} \qquad \forall i \quad = 1,2,\ldots,L$$

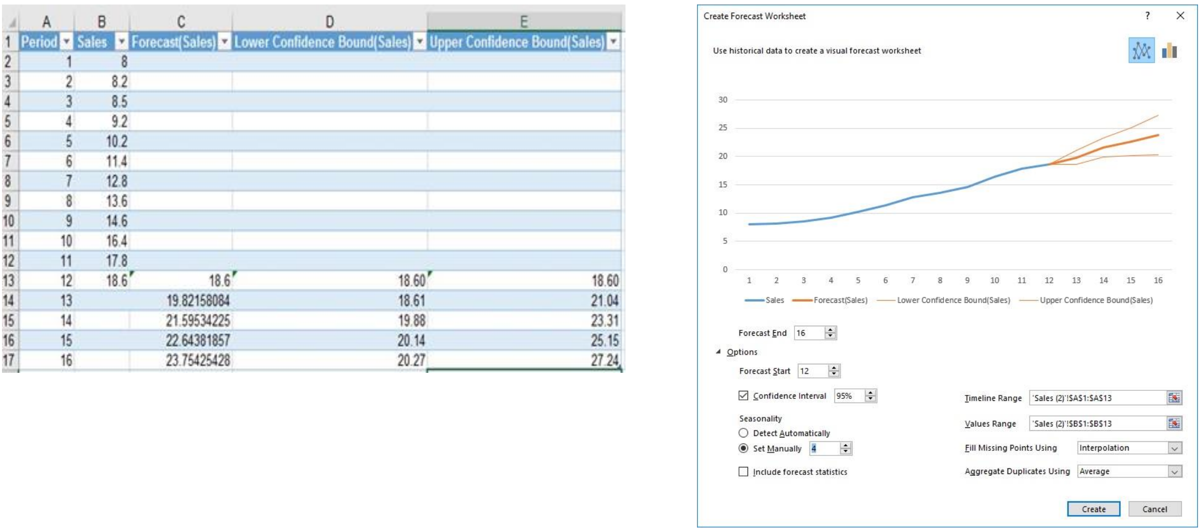
where:

$$A_j = \frac{\sum_{i=1}^{L} x_{L(j-1)+i}}{L} \qquad \forall j \quad = 1,2,\ldots,N$$

Note that $\boldsymbol{A_j}$ is the average value of *x* in the $\boldsymbol{j^{th}}$ cycle of our data.

Actually, we can use an easy algorithm with three steps to calculate the seasonal indices: ***step 1*** – compute the averages of $\boldsymbol{x_t}$ for each year; ***step 2*** – divide the observations $\boldsymbol{x_t}$ by the appropriate yearly mean; ***step 3*** – the seasonal indices are formed by computing the average of each season.[11]

[11] For details see http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc435.htm



	A	B	C	D	E
1	Period	Sales	Forecast(Sales)	Lower Confidence Bound(Sales)	Upper Confidence Bound(Sales)
2	1	8			
3	2	8.2			
4	3	8.5			
5	4	9.2			
6	5	10.2			
7	6	11.4			
8	7	12.8			
9	8	13.6			
10	9	14.6			
11	10	16.4			
12	11	17.8			
13	12	18.6	18.6	18.60	18.60
14	13		19.82158084	18.61	21.04
15	14		21.59534225	19.88	23.31
16	15		22.64381857	20.14	25.15
17	16		23.75425428	20.27	27.24

Create Forecast Worksheet

Use historical data to create a visual forecast worksheet

Forecast End 16

Options

Forecast Start 12

☑ Confidence Interval 95%

Seasonality

○ Detect Automatically

● Set Manually 4

□ Include forecast statistics

Timeline Range 'Sales (2)'!$A$1:$A$13

Values Range 'Sales (2)'!$B$1:$B$13

Fill Missing Points Using Interpolation

Aggregate Duplicates Using Average

Fig.5-14 Example of ***TES*** with MS Excel

The smoothing parameters are often selected between 0.02 and 0.2. It is again possible to estimate them by ***Least Squares (LS) method,*** minimizing the sum of the squared one-step-ahead errors, but there is no exclusive combination of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ which will minimize the squared errors for all $\boldsymbol{t}$.

There are many modifications of the Exponential smoothing. Pegels (1969), later extended by Gardner (1985), proposed to include methods with an additive damped trend, and Taylor (2003) added methods with a multiplicative damped trend[12]. Most of the Exponential smoothing methods involve complex calculations and using statistical software or at least MS Excel (see Fig.5-14).

In summary, there are many different models to forecast a set of cross-sectional or time-series data (Box et al., 2016). It should be noted that all models are based on specific assumptions how the particular real-life case works. We need to understand what these assumptions are. Moreover, we should be certain that the assumptions of our chosen model are true and be able to explain and defend them.

***

[12] For more details see https://www.otexts.org/fpp/7/6

SUMMARY AND CONCLUSIONS

- Chapter 5 discusses the basic types of forecasting techniques and models for ***cross-sectional*** and ***time series data***. These simple predictions provide ***benchmarks*** for all other forecasting techniques accuracy.
- ***Cross-sectional data*** is a set of data values collected by observing many subjects at the same point of time, whereas ***time series data*** are ordered data values observed at various points in time.
- ***Time series analysis*** comprises methods for analyzing time series data in order to extract meaningful information and ***Time series forecasting*** is the use of a model to predict future values based on previously observed data.
- ***Line chart*** or ***line graph*** is a type of chart which displays information as a series of data points connected by straight line segments. When a line chart is used to visualize a trend in time series data the line is drawn chronologically.
- ***Fan chart*** joins a simple line chart for observed past data, by showing ranges for possible values of future data together with a line showing a central estimate or most likely value for the future outcomes.
- ***Prediction interval*** is an estimate of the interval in which future observations will fall, with a certain probability, given what has already been observed.
- ***Naïve forecast*** for cross-sectional and stationary time series data states that the forecast for any period equals the historical average. For time series data that are stationary in terms of their first differences, the ***naïve forecast*** equals the previous period's actual value and so on.
- ***Naïve technique*** is often referred to as "***random walk***" model. It assumes that, from one period to the next, the original time series merely takes a random "step" away from its last recorded position.
- If the constant term in the ***random walk model*** is zero, then it returns the same forecast as the simple naïve model, known as ***”random walk without drift”***. If we assume that the random walk undergoes a drift, it becomes a ***random-walk-with-drift.***
- In ***Seasonal Naïve Technique***, we set each forecast to be equal to the last observed value from the same season of the year (the same month of the previous year for instance).
- ***Moving average*** is a calculation technique to analyze data by creating a series of averages of different subsets of the full data set. It is commonly used with time series data to smooth out short-term fluctuations and highlight long-term trends or cycles.

- The simplest way to smooth a time series is to calculate a ***simple, or equally-weighted, moving average (MA)***.
- ***Simple Moving Average with Trend*** means adding a constant to the simple MA forecasting equation, analogous to the drift term in the random-walk-with-drift model.
- In a ***cumulative moving average*** **(*CMA*)**, the data arrive in an ordered data series, and the user would like to get the average of all of the data up until the current observation.
- ***Weighted average*** is any average that has multiplying factors to give different weights to data at different positions in the sample.
- Exponential smoothing is a technique based on the ***Weighted average*** approach, that can be applied to time series data, either to produce smoothed data for presentation, or to make forecasts.
- ***Simple exponential smoothing (SES), exponential moving average (EMA)***, or ***exponentially weighted moving average (EWMA)*** is suitable for forecasting data without any systematic trend or seasonal pattern.
- ***Smoothing factor*** $\boldsymbol{\alpha}$ ($0<\boldsymbol{\alpha}<1$) is a simple weighted average of the previous observation $\boldsymbol{x_{t-1}}$ and the previous smoothed statistic $\boldsymbol{s_{t-1}}$. There is no formally correct procedure for choosing $\boldsymbol{\alpha}$. Values of $\boldsymbol{\alpha}$ close to one have less of a smoothing effect and give greater weight to recent changes in the data, while values of $\boldsymbol{\alpha}$ closer to zero have a greater smoothing effect and are less responsive to recent changes.
- The simple exponential smoothing model can be generalized to obtain a ***linear exponential smoothing (LES)*** model that computes local estimates of both level and trend. It is known as ***Holt's Linear Exponential Smoothing*** and it introduce a second equation with a second constant, β, which must be chosen in conjunction with α.
- The interpretation of the trend-smoothing constant $\boldsymbol{\beta}$ is analogous to that of the level-smoothing constant $\boldsymbol{\alpha}$ in SES. Models with small values of $\boldsymbol{\beta}$ assume that the trend changes only very slowly over time, while models with larger $\boldsymbol{\beta}$ assume that it is changing more rapidly.
- ***Holt-Winters (Triple) Exponential Smoothing*** takes into account seasonal changes as well as trends. It comprises the forecast equation and three smoothing equations — one for the level $\boldsymbol{s_t}$, one for trend $\boldsymbol{b_t}$, and one for the seasonal component (or seasonal index) denoted by $\boldsymbol{c_t}$, with smoothing parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$.
- The ***additive method*** is preferred when the seasonal variations are roughly constant through the series, while the ***multiplicative method*** is preferred when the seasonal variations are changing proportional to the level of the series.

- *There are many different models to forecast a set of cross-sectional or time-series data. It should be noted that all models are based on specific assumptions how the particular real-life case works. We need to understand what these assumptions are. Moreover, we should be certain that the assumptions of our chosen model are true and be able to explain and defend them.*

***

KEY TERMS

CHAPTER EXERCISES

Conceptual Questions:

1. What is the difference between ***cross-sectional*** and ***time series data***? Explain and illustrate with examples.
2. How can data visualization provide useful prediction information? Discuss.
3. List and briefly discuss the major pros and cons of the best-fit line manual estimation.
4. What are the similarities and the differences between ***Naïve techniques*** (average model, random walk and so on) and the ***simple m-MA model***? List and discuss at least three of them.
5. What are the major types of ***Exponential Smoothing?*** Discuss and illustrate with examples.

Business Applications:

The M&M company Sales Department recorded (see Table 5.2 Sales Data) their monthly sales (in $ thousands) of product ”A” for four years (2015 through 2019 – file SmallSales.xlsx):

- Plot the time series of sales of product A. Can you identify seasonal fluctuations and/or a trend pattern?
- Use MS Excel option ”Add Trendline” and append a 12 ordered MA trend to the chart. Does it make the trend pattern visualization better? Explain.
- Use MS Excel option Data/Data Analysis/Exponential smoothing to develop SES model and compute smoothed values for the Sales time series (Check Chart option). Explain your findings.

Write a short report (up to two pages) discussing your answers.

Table 5.2 Sales Data

Mont\Year	2015	2016	2017	2018	2019
Jan	742	741	896	951	1030
Feb	697	700	793	861	1032
Mar	776	774	885	938	1126
Apr	898	932	1055	1109	1285
May	1030	1099	1204	1274	1468
Jun	1107	1223	1326	1422	1637
Jul	1165	1290	1303	1486	1611
Aug	1216	1349	1436	1555	1608
Sep	1208	1341	1473	1604	1528
Oct	1131	1296	1453	1600	1420
Nov	971	1066	1170	1403	1119
Dec	783	901	1023	1209	1013

INTEGRATIVE CASE

HEALTHY FOOD SYPPLY CHAIN & STORES

Part 5: Simple Forecasting Models – First Steps in Numerical Predictions

In Chapter 1 we introduced *Healthy Food Stores* – a fast-growing retail food provider with 12 stores in a northwestern state. The company executives decided to study the effect that company advertising dollars have on sales. They hoped that examining collected historical data would reveal relationships that would help determine future advertising expenditures and predict monthly sales volumes for the upcoming quarter.

After identifying basic parameters, along with input (independent) and output (dependent) variables of the forecasting scenario in Part 2, the research team determined the main elements of the forecasting process:

- Forecasting horizon of up to twelve months;
- Quarterly forecast updates, since accuracy decreases as time horizon increases, and sufficient time is needed to implement possible changes;
- Development of different models based on data patterns, if any, and evaluation of their accuracy in order to select the most appropriate one;
- Selection of the best forecasting model with no more than 5% forecasting error.

The benchmark forecast computed in Part 3 (using the baseline of ***one-step naive forecast*** as a reference forecast) provided basic values for the most common measures of accuracy MFE, MAD, MAPE, MPE, MSE and CV(RMSE).

In Part 4, information about the opinion of some important people from the Healthy Food Stores Company, concerning this specific case, was collected. To study the effect company advertising dollars have on sales, the research team met three groups of most important people in the company – the company top executives, the sales managers from all 12 stores and the most experienced professionals from Advertising Department. To address each group, the research team applied the following methods accordingly – Delphi method to top executives' group, Sales-force composite to the sales managers and Scenario writing to the experienced professionals from Advertising Department.

After collecting such valuable information from different sources, the research team was ready to make its first steps in Numerical Predictions. Having enough skills in MS Excel and basic knowledge in Business Statistics they planned to develop different basic forecasting models, which could be used to expand the baseline of the ***one-step naïve forecast*** as reference forecasts.

Case Questions

1. Open the updated file Data.xlsx from Part 3 and create new spreadsheets for each group of techniques discussed in this Chapter as follows:
   a) NT spreadsheet for ***Naïve techniques***, i.e. Average model, Random Walk with Drift and Seasonal Naïve Technique (recall that Random Walk Without Drift returns the same predictions as the one-step naïve forecast);
   b) MA spreadsheet for ***Moving Average*** techniques – only for Simple (equally-weighted) 12 ordered Moving Average and Simple Moving Average with Trend (Hint: use average from the first differences of the series as a trend component);
   c) SES spreadsheet for ***Simple Exponential Smoothing*** – use MS Excel option Data/Data Analysis/Exponential smoothing to develop SES model and compute the smoothed values. Note: The damping factor is the value (1- α), i.e. the smaller alpha (larger the damping factor), the more the peaks and valleys are smoothed out. The larger alpha (smaller the damping factor), the closer the smoothed values are to the actual data points (see Fig.5-13). Values of 0.1 to 0.3 are reasonable smoothing constants. These values indicate that the current forecast should be adjusted 10 percent to 30 percent for error in the prior forecast. Larger constants yield a faster response but can produce erratic projections. Smaller constants can result in long lags for forecast values
   d) TES spreadsheet for ***Holt-Winters (Triple) Exponential Smoothing (TES).*** (Hint: use function FORECAST in MS Excel). Compute 12 periods forecast, i.e. set up parameter "Stop Forecast" at 48.
2. Use the formulas designed in Part 3 to compute MFE, MAD, MAPE, MPE, MSE and CV(RMSE) for each new model for the testing dataset of the last 12 months only.
3. Analyze the results:
   - How good is the accuracy of the new models relative to the ***one-step naïve forecast*** from Part 4? Discuss each model.
   - What model provides the best (so far) accuracy? Are there any initial assumptions/reasons leading to this conclusion?
4. Are there connections between the results from this Part and the findings in previous Parts of the Case? Explain and give details.
5. What overall recommendations would you make to the research team? Explain why.
6. Write a report on the questions above, discussing all important findings and draw relevant conclusions about this part of the Integrative Case.

References

Box, G., Jenkins, G., Reinsel, G., & Ljung, G. (2016). *Time Series Analysis: Forecasting and Control*. Wiley.

Brown, R. G. (1956). *Exponential Smoothing for Predicting Demand*. Cambridge, MA: Arthur D. Little Inc.

Gardner, E. S. (1985). Exponential smoothing: The state of the art. *Journal of Forecasting*, *4*(1), 1–28. https://doi.org/10.1002/for.3980040103

Holt, C. C. (1957). Forecasting Trends and Seasonal by Exponentially Weighted Averages. *Office of Naval Research Memorandum 52*, reprinted in Holt, C. C. (2004, January–March). Forecasting Trends and Seasonal by Exponentially Weighted Averages. *International Journal of Forecasting*, *20(1)*, 5–10. https://doi.org/10.1016/j.ijforecast.2003.09.015

Nau, R. (2014, August). *Forecasting with moving averages*. Fuqua School of Business, Duke University, Durham, North Carolina, United States. Retrieved from: http://people.duke.edu/~rnau/Notes_on_forecasting_with_moving_averages--Robert_Nau.pdf

Pegels, C. C. (1969). Exponential forecasting: Some new variations. *Management Science*, *15*(5), 311–315. https://doi.org/10.1287/mnsc.15.5.311

Taylor, J. W. (2003). Exponential smoothing with a damped multiplicative trend. *International Journal of Forecasting*, *19*, 715–725. https://doi.org/10.1016/S0169-2070(03)00003-7

Winters, P. R. (1960, April). Forecasting Sales by Exponentially Weighted Moving Averages. *Management Science*, *6(3)*, 324–342. https://doi.org/10.1287/mnsc.6.3.324

***

CHAPTER 6. REGRESSION MODELS

6.1. Association, Correlation and Dependence

Today everybody knows that there is a relationship between consumption and income. Although Keynes postulated that there is a positive relationship between consumption and income, he did not specify the precise form of the functional relationship between the two. A mathematical economist might suggest the following form of the Keynesian consumption function:

$$Y = \beta_0 + \beta_1 X \ (0 < \beta_1 < 1) \qquad (6\text{-}1)$$

where Y = consumption expenditure and X = income, and β_0 and β_1, known as parameters of the model, are, respectively, the ***intercept*** and ***slope*** coefficients of this ***linear equation***.

Since in business and economics most variables are random, a statistician would rather use the following model:

$$\overset{\text{Dependent Variable}}{y} = \underbrace{\overset{\text{Population y intercept}}{\beta_0} + \overset{\text{Population Slope Coefficient}}{\beta_1}\overset{\text{Independent Variable}}{x}}_{\text{Linear component}} + \underbrace{\overset{\text{Random Error term, or residual}}{\varepsilon}}_{\text{Random Error component}} \qquad (6\text{-}2)$$

The ***dependent variable y*** in (6-2) is also known[1] as a "response variable", "endogenous variable", "forecast variable", "regressand", "explained variable", "output variable", etc. The ***independent variable***[2] ***x*** is also known as a "regressor", "exogenous variable", "predictor variable", "factor", "explanatory variable", "input variable", etc.

The basic way to study an association between two variables is by chart – finding the overall pattern (if any) and analyzing the deviations from it. Scatter plots like Fig.6-1 are the most effective graphical technique to study a relationship between two quantitative variables.

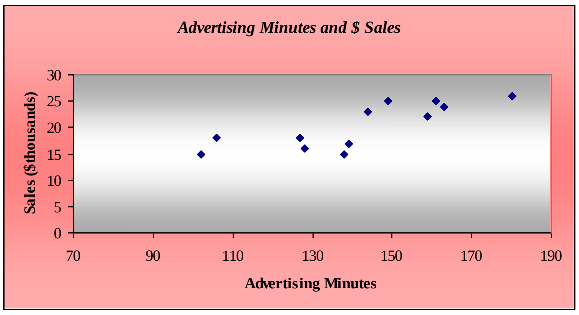


Fig.6-1 Example of Scatter plot representing a relationship between Sales and Advertising based on specific given data

[1] See http://en.wikipedia.org/wiki/Dependent_and_independent_variables

[2] Some authors prefer not to use the term "independent variable", because the quantities treated as "independent variable" may, in fact, not be statistically independent.

The ***scatterplot*** (or ***scatter diagram***) is one of the best tools for studying the association of two variables graphically. A ***Scatter Diagram*** is a chart that portrays the relationship between two variables. ***Scatterplots*** are especially helpful when the number of data is large studying a list is then virtually hopeless. A scatter diagram (see Fig.6-1) plots two measured variables against each other (for each individual). That is, the x (horizontal) coordinate of a single point in a scatterplot is the value of one measurement (X) of an individual, and the y (vertical) coordinate of that point is the other measurement (Y) of the same individual. Such a plot is known as *"a scatterplot of Y versus X"* or *"a scatterplot of Y against X"*.

An association is any relationship between two measured quantities that renders them statistically dependent. The term "***association***" is closely related to the term "***correlation***." Both terms imply that two or more variables vary according to some pattern. However, correlation is more rigidly defined by some correlation coefficient which measures the degree to which the association of the variables tends to a certain pattern.

Dependence, in general, is any statistical relationship between two random variables or two sets of data. ***Correlation*** refers to any of a broad class of statistical relationships involving dependence. Formally, dependence refers to any situation in which random variables do not satisfy a mathematical condition of ***probabilistic independence***[3]. In loose usage, correlation can refer to any departure of two or more random variables from independence, but technically it refers to any of the specialized types of ***relationship between mean values*** discussed in 6.2.

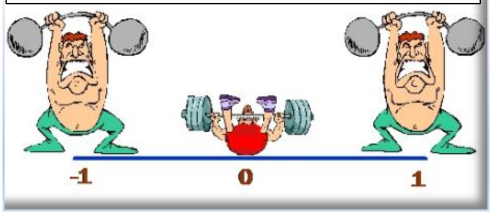
It is important to note that *neither association nor correlation establish causality*. This is necessary to state because studies which show correlation are sometimes misinterpreted or misconstrued to the effect that association by itself proves something useful. Association by itself does not prove or disprove anything and can only at best show that two variables are mathematically related, whether or not they are causally related. Likewise, it is quite common (and yet erroneous) for people or groups to state that "studies show..." some given conclusion which is actually based only on statistical association rather than the implied causality suggested by the person citing the studies. This is not to say that the studies themselves are invalid, but rather that a study which looks only for correlation can only establish that there is a correlation, not proof of why there is a correlation.

"*Correlation does not imply causation*" is a phrase in science that emphasizes that a correlation between two variables does not necessarily imply that one causes the other. Many statistical tests calculate correlation between variables. A few go further and calculate the

[3] Two random variables are independent if the realization of one does not affect the probability distribution of the other (see https://en.wikipedia.org/wiki/Independence_%28probability_theory%29).

likelihood of a true causal relationship. The assumption, that correlation proves causation, is considered a questionable cause logical fallacy in that two events occurring together are taken to have a cause-and-effect relationship.

Examples of negative, weak and positive correlation.



As with any logical fallacy, identifying that the reasoning behind an argument is flawed does not imply that the resulting conclusion is false. However, in casual use, the word "imply" loosely means "suggests" rather than "requires". The idea that correlation and causation are connected is certainly true – where there is causation, there is a likely correlation. Indeed, correlation is used when inferring causation.

Edward Tufte (2003, p.4), in a criticism of the brevity of "*correlation does not imply causation*", deprecates the use of "is" to relate correlation and causation (as in "*Correlation is not causation*"), citing its inaccuracy as incomplete. While it is not the case that correlation is causation, simply stating their nonequivalence omits information about their relationship. Tufte suggests that the shortest true statement that can be made about causality and correlation is one of the following:

- "Empirically observed covariation[4] is a necessary but not sufficient condition for causality."
- "Correlation is not causation but it sure is a hint."

The conventional dictum that "*correlation does not imply causation*" means that correlation cannot be used to infer a causal relationship between the variables. This dictum should not be taken to mean that correlations cannot indicate the potential existence of causal relations. However, the causes underlying the correlation, if any, may be indirect and unknown, and high correlations also overlap with identity relations (tautologies), where no causal process exists. Consequently, establishing a correlation between two variables is not a sufficient condition to establish a causal relationship (in either direction).

In other words, there can be no conclusion made regarding the existence or the direction of a cause-and-effect relationship only from the fact that Y and X are correlated. *Determining whether there is an actual cause-and-effect relationship requires further investigation, even when the relationship between Y and X is statistically significant, a large effect size is observed, or a large part of the variance is explained.*

[4] ***Covariance*** is a measure of how much two random variables change together – its normalized version is the ***correlation coefficient***.

6.2. Simple Linear Regression and Correlation

Sometimes the pattern of association is a simple linear relationship as in the case of the popular ***Pearson linear product moment correlation coefficient***, although other forms of correlation are better suited to non-linear associations. There are many statistical measures of association that can be used to infer the presence or absence of a relationship in a sample of data. The most important techniques, from a forecasting point of view, are the ***Correlation*** and ***Regression analysis***.

A. Correlation Analysis

Correlation Analysis is a group of statistical techniques to measure the association between two variables. It is only concerned with the strength of the relationship and no causal effect is implied.

There are several correlation coefficients, denoted ***ρ*** or ***r***, measuring the degree of association. The most common of these is the ***Pearson correlation coefficient*** which is sensitive only to a linear relationship between two variables (which may exist even if one is a nonlinear function of the other). Since then, other correlation techniques have been developed to be more robust than the Pearson correlation coefficient and to measure other types of relationships, for example nonlinear (Yule & Kendall, 1950, pp. 258-270; Kendall 1955).

The ***Pearson product-moment correlation coefficient*** was developed by Karl Pearson (1895) from a related idea introduced by (Sir) Francis Galton (1886). It is a measure of the linear relationship between two variables X and Y and ranges from +1 to −1 inclusive. A value of 1 implies that a linear equation (Fig.6-2) describes the relationship between X and Y perfectly, with all data points lying on a line for which Y increases as X increases. A value of −1 implies that all data points lie on a line for which Y decreases as X increases. A value of 0 implies that there is no linear correlation between the variables. The correlation coefficient is positive (and indicate a direct relationship) if and only if X and Y lie on the same side of their respective means. Thus, it is positive if X and Y tend to be simultaneously greater than, or simultaneously less than, their respective means. The coefficient is negative (i.e. indicate an inverse relationship) if X and Y tend to lie on opposite sides of their respective means.

Pearson's correlation coefficient between two variables is defined as the covariance of the two variables divided by the product of their standard deviations (see equation 6-3). The form of the definition involves a "product moment", that is, the mean (or the first moment of the origin) of the product of the mean-adjusted random variables, hence the modifier ***product-moment*** in the name.

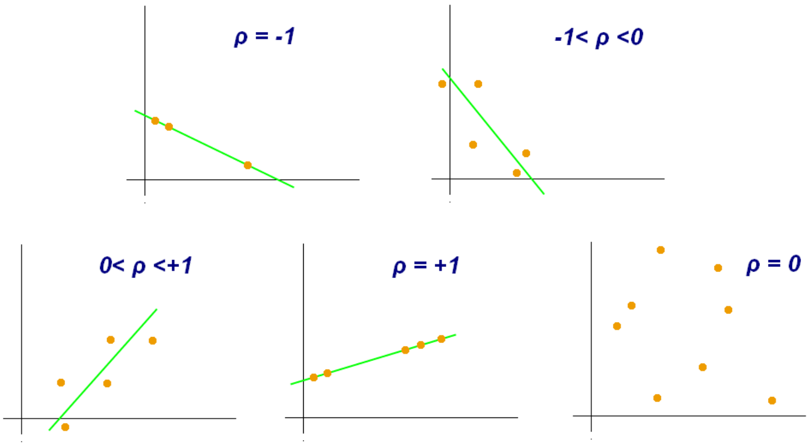


Fig.6-2 Examples of scatter diagrams with different values of correlation coefficient (ρ)

Calculations are very complicated and here we present only the general formula for a population:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \qquad (6\text{-}3)$$

where, **cov** is the covariance between X and Y, **σ_X** is the standard deviation of X, **μ_X** is the mean of X, and **E** is the expected value (the mean) of the first moment.

Some authors suggest guidelines for interpretation of the correlation coefficient (***ρ***). Of course, all such criteria are in some ways arbitrary and should not be observed too strictly. The interpretation of a correlation coefficient depends on the context and purposes. For example, a correlation of 0.8 may be very low if one is verifying a physical law using high-quality instruments, but it may be regarded as very high in the social sciences where there may be a greater contribution from complicating factors.

In business, in general, it is accepted that values of (***ρ***) close to 0 (0 up to +/- 0.4) indicate weak correlation. Values between (both positive and negative) 0.4 and 0.7 indicate moderate relationship, and between (both positive and negative) 0.7 and 1 – strong correlation. Values of -1.00 or +1.00 indicate perfect correlation, i.e. deterministic mathematical relationship[5].

If a population or data-set is characterized by more than two variables (see Section 6.3 in this Chapter), a ***partial correlation coefficient*** measures the strength of dependence between a pair of variables that is not accounted for by the way in which they both change in response to variations in a selected subset of the other variables.

[5] Mathematical models that are not ***deterministic***, because they involve ***randomness*** are known as ***stochastic***.

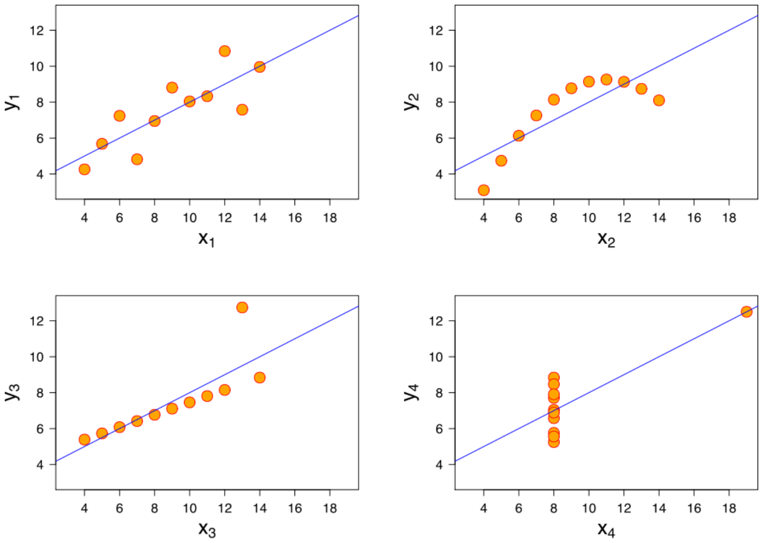


Fig.6-3 Four sets of data with the same correlation of 0.816

It should be noted that the Pearson correlation coefficient indicates the strength of a linear relationship between two variables, but its value generally does not completely characterize their relationship. For instance, if the conditional mean of Y given X, denoted E(Y|X), is not linear in X, the correlation coefficient will not fully determine the form of E(Y|X).

Fig.6-3 shows a well-known example of a set of four different pairs of variables, where all four ***y*** variables have the same mean (7.5), variance (4.12), correlation (0.816) and regression line (y = 3 + 0.5x). However, as can be seen on the plots, the distribution of the variables is very different. The first one (top left) seems to be distributed normally and corresponds to what one would expect when considering two variables correlated and following the assumption of normality. The second one (top right) is not distributed normally – while an obvious relationship between the two variables can be observed it is not linear. In this case the Pearson correlation coefficient does not indicate that there is an exact functional relationship, only the extent to which that relationship can be approximated by a linear relationship. In the third case (bottom left), the linear relationship is perfect, except for one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816. Finally, the fourth example (bottom right) shows another example when one outlier is enough to produce a high correlation coefficient, even though the relationship between the two variables is not linear.

These examples indicate that the correlation coefficient, as a summary statistic, cannot replace visual examination of the data. In addition, to characterize and analyze the relationship between two variables and use it to make predictions we need a further analysis, the ***Regression analysis***.

B. Regression Analysis

The concept of regression comes from genetics and was popularized by (Sir) Francis Galton (1886) during the late 19th century with the publication of Regression towards mediocrity in hereditary stature. Galton coined the term regression to describe an observable fact in the inheritance of multi-factorial quantitative genetic traits, namely that the offspring of parents who lie at the tails of the distribution will tend to lie closer to the center, the mean, of the distribution. He quantified this trend, and in doing so invented ***linear regression analysis***, thus laying the groundwork for much of modern statistical modeling.

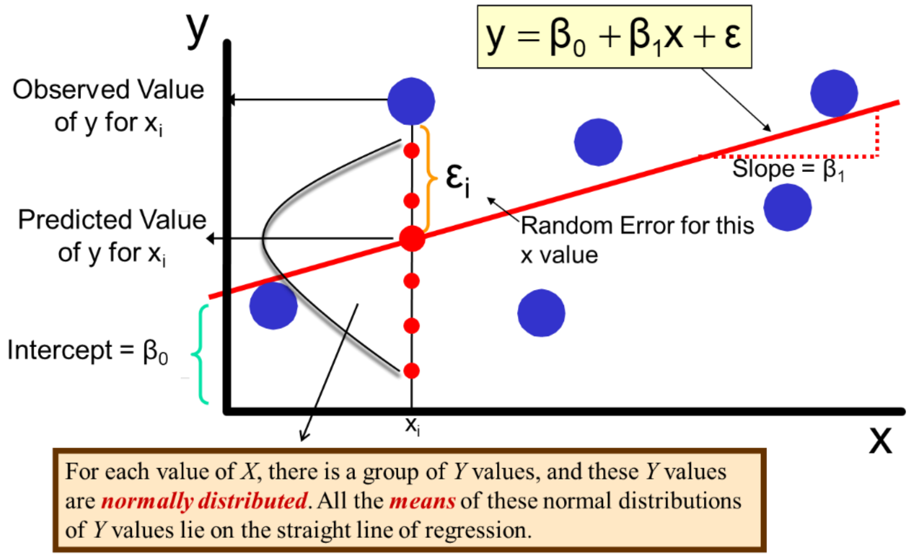
The intuitive explanation for the regression effect is simple – the variable we are trying to predict usually consists of a predictable component ("***signal***") and a statistically independent unpredictable component ("***noise***"). The best we can hope to do is to predict only that part of the variability which is due to the signal. Hence our forecasts will tend to exhibit less variability than the actual values, which implies a regression to the mean.

Regression analysis is widely used for predictions, where its use has substantial overlap with the field of machine learning and data mining (see Chapter 10). It is also used to find out which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. In general, the ***Regression analysis*** is used to:

- Predict the value of a dependent variable (Y) based on the value of at least one explanatory variable (X), whose measures are statistically independent. When there is more than one explanatory variable (see Section 6.3), all of them should be statistically independent.
- Explain the impact of changes in one (or more) explanatory variable (X) on the dependent variable (Y).

Many techniques for carrying out regression analysis have been developed. The Ordinary ***Least Squares (LS)***[6] and the ***Linear regression*** are the most familiar methods, developed long time ago, during the 19th century. The performance of regression analysis methods in practice depends on the form of the data generating process, and how it relates to the regression approach being used. Since the true form of the data-generating process is generally not known, regression analysis often depends to some extent on making assumptions about this process. These assumptions are testable if a sufficient quantity of data is available. In a case for making predictions, Regression models are sometimes useful even when the assumptions are moderately violated, although they may not perform optimally.

[6] The ***least-squares*** method is usually credited to Carl Friedrich Gauss (1809) but it was first published by Adrien-Marie Legendre (1805).


$y = \beta_0 + \beta_1 x + \varepsilon$
y
Observed Value of y for x_i
ε_i
Slope = β_1
Random Error for this x value
Predicted Value of y for x_i
Intercept = β_0
x_i
X

For each value of *X*, there is a group of *Y* values, and these *Y* values are ***normally distributed***. All the ***means*** of these normal distributions of *Y* values lie on the straight line of regression.

Fig.6-4 Example of data from a linear regression model.

Linear regression analysis is the most widely used of all statistical techniques – it is a study of *linear, additive relationships* between variables (Fig. 6-4), usually under the assumption of independently and identically normally distributed errors (Nau, 2014).

Justification for Regression Assumptions

Testing the assumptions of linear regression (also known as ***Aptness of the Model***) will be discussed in detail in next sections. Here, some general notes and comments about these classical assumptions for regression analysis are given:

A. Why should we assume that relationships between variables are *linear* (technically *linearity* means that the mean of the response variable ***Y*** is a linear combination of the regression coefficients ***β*** and the predictor variables ***X***). It is, because:

- linear relationships are the simplest non-trivial relationships that can be imagined (hence the easiest to work with);
- regression coefficients ***β*** have very clear business interpretation with both, cross-sectional and time-series data sets (see the examples in this and next sections of the book);
- the "true" relationships between the variables being studied are often at least approximately linear over the range of values that are of interest to us and
- even if they are not, we can often transform the variables (for example using their first differences as we did in section 5.1) in such a way as to linearize their relationships.

This is a strong assumption, and the first step in regression modeling should be to look at scatter diagrams (like Fig.6-1) of the variables (and in the case of time series data, plots of the variables vs. time), to make sure it is reasonable a priori. After fitting a model, plots (see Fig.6-5) of the ***errors*** (also known as ***residuals*** $\varepsilon_i = y_i - \hat{y}_i$, see equation (6-4) and the comments after that) should be studied to see if there are unexplained nonlinear patterns.

This is especially important when the goal is to make predictions for scenarios outside the range of the historical data, where departures from perfect linearity are likely to have the biggest effect. If we see evidence of nonlinear relationships, it is possible (though not guaranteed) that transformations of variables (similar to the first differences mentioned above) will straighten them out in a way that will yield useful inferences and predictions via a linear regression model.

Note that this assumption is much less restrictive than it may at first seem. Because the predictor variables ***X*** are treated as fixed values, rather than random variables (i.e. the predictor variables ***X*** are assumed to be error-free, that is, measured with no error), linearity is in fact only a restriction on the parameters (regression coefficients $\boldsymbol{\beta}$). The predictor variables ***X*** themselves can be arbitrarily transformed, which means that multiple copies of the same underlying predictor variable can be added, each one transformed differently. This trick is used, for example, in ***polynomial regression***, which uses linear regression to fit the dependent variable ***Y*** as an arbitrary polynomial function (up to a given rank) of a predictor variable. This makes linear regression an extremely powerful inference method.

B. Why should we assume that the effects of different explanatory variables ***X*** on the expected value of the dependent variable ***Y*** are *additive*?

This is another assumption, which is especially important in the case of multiple regression analysis, and stronger than most people realize. It implies that the marginal effect of one independent variable (i.e., its regression coefficient $\boldsymbol{\beta}$) does not depend on the current values of other independent variables. Nau (2014) asks "Why shouldn’t it"? It is plausible that one independent variable could amplify the effect of another.

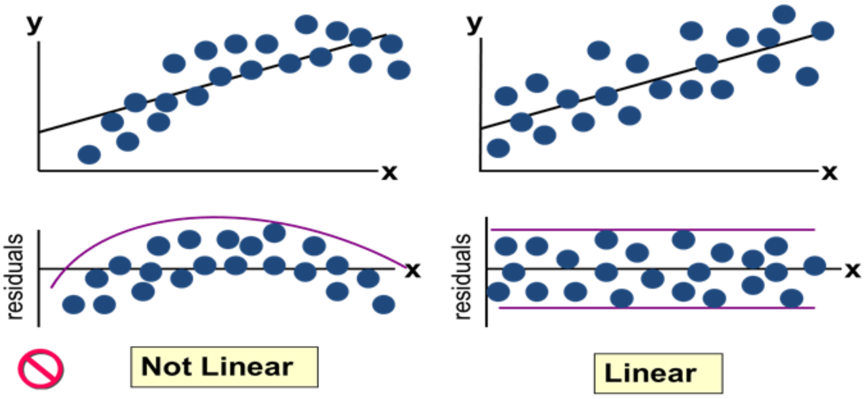


Fig.6-5 Residuals Test for Linearity

In a multiple regression model, the estimated coefficient of a given predictor variable ***X*** supposedly measures its effect while "controlling" for the presence of the other ***X’***s. However, the way in which this "controlling" is performed is extremely simplistic – multiples of other variables are just added or subtracted.

C. Why should we assume the errors of linear models are independent and normally distributed with constant variance?

- Independence of errors means that errors ($\boldsymbol{\varepsilon}$) in (6-2) are uncorrelated with each other. Actual statistical independence is a stronger condition than mere lack of correlation and is often not needed, although it can be exploited if it is known to hold.
- A further assumption made by linear regression is that the residuals or errors ($\boldsymbol{\varepsilon}$) are normally distributed. This is a consequence of the ***Central Limit Theorem***[7]. Much data in business and economics are obtained by adding or averaging numerical measurements performed on many different persons or products or locations or time intervals. As far as the activities that generate the measurements may occur somewhat randomly and somewhat independently, we might expect the variations in the totals or averages to be somewhat normally distributed.
- It is mathematically convenient, as it implies that the optimal coefficient estimates for a linear model are those that minimize the mean squared error (which are easily calculated). It also justifies the use of a host of statistical tests based on the normal family of distributions, like ***z***-, ***t***-distribution, and so on…
- Again, even if the "true" error process is not normal in terms of the original units of the data, it may be possible to transform their values so that our model's prediction errors are approximately normal.

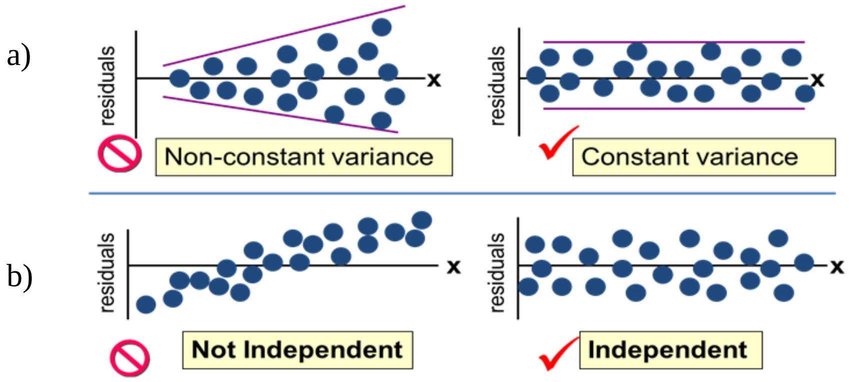


Fig.6-6 Graphical Test of residuals for independence and constant variance

[7] The central limit theorem (***CLT***) states that, *given certain conditions, the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined expected value and well-defined variance, will be approximately normally distributed, regardless of the underlying distribution*.

- ***Homoscedasticity***, which is also known as a homogeneity of variance, means that errors (ε) in (6-2) have the same variance, regardless of the values of the predictor variables. In practice, this assumption is invalid (i.e. the errors (ε) are ***heteroscedastic***) if the dependent variable ***Y*** can vary over a wide scale. In order to determine when a pattern of residuals violates model assumptions of ***homoscedasticity,*** it is wise to look for a systematic change in the absolute or squared residuals when plotted against the predicting outcome (Fig.6-6, a). The error will not be evenly distributed across the line in case of ***heteroscedasticity.*** In effect, residuals appear clustered and spread apart on their predicted plots for larger and smaller values of ***X***, and the mean squared error for the model will be wrong.

Note, that here, too, caution must be exercised. Even if the errors (ε) are approximately normally distributed, it is not guaranteed that they will also have the same variance (i.e. be identically normally distributed) for all values of the predictor variables ***X***. For example, if the dependent variable consists of daily or monthly total sales, there are probably significant day-of-week patterns or seasonal patterns. In such cases, the variance of the total will be larger on days or in seasons with greater business activity.

Some authors, (Hyndman & Athanasopoulos, 2013) consider that it “is useful to have the errors normally distributed with constant variance in order to produce prediction intervals and to perform statistical inference. While these additional conditions make the calculations simpler, they are not necessary for forecasting[8]."

Such statement could be arguable since simple linear regression estimation methods give less precise parameter estimates and misleading inferential quantities such as standard errors when substantial heteroscedasticity is present. However, various estimation techniques (like weighted least squares, heteroscedasticity-consistent standard errors and Bayesian linear regression techniques, see Box & Tiao 1992) can handle heteroscedasticity in a quite general way. It is also possible in some cases to fix this problem by applying different transformations to the dependent variable ***Y***, for example – use of seasonal adjustments to ***Y***; fit the logarithm of the response variable ***Y*** using a linear regression model, which implies that ***Y*** has a log-normal distribution rather than a normal distribution and others.

What is more important is that even if the unexplained variations in the dependent variable ***Y*** are approximately normally distributed, it is not guaranteed that these random variations or the errors (ε) will be statistically independent (Fig.6-6, b). This is an especially important

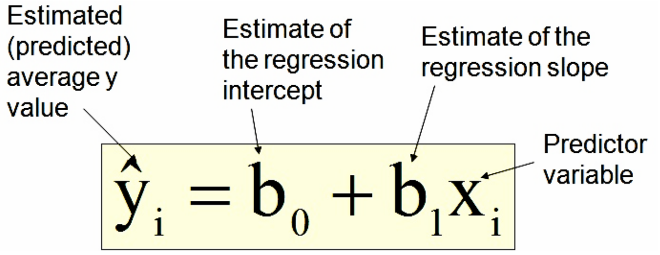
[8] See https://www.otexts.org/fpp/4/1

question when the data consists of time series. If the model is not correctly specified, it is possible that consecutive errors (*ε*) (or errors separated by some other number of periods) will have a systematic tendency to have the same, or the opposite signs (i.e. the errors (*ε*) will be related), a phenomenon known as "***autocorrelation***". This important question will be discussed in Chapter 7.5. Regression with Time Series Data.

Regression Coefficients Estimation Methods

A large number of procedures have been developed for parameters ***β*** estimation in (6-2) and inference in linear regression. These methods differ in the computational simplicity of algorithms, the presence of a closed-form solution, the robustness with respect to heavy-tailed distributions, and the theoretical assumptions needed to validate desirable statistical properties such as ***consistency***[9] and asymptotic ***efficiency***[10].

The sample regression line provides an estimate of the population regression line (6-2), given by equation (6-4):


Estimated (predicted) average y value
Estimate of the regression intercept
Estimate of the regression slope
Predictor variable
$\hat{y}_i = b_0 + b_1 x_i$

(6-4)

where:

– ***Y-hat*** (ŷ) is the estimated average value of dependent variable ***y*** (y_i = {y_1, y_2, … y_N});

– the intercept $\mathbf{b}_0$ is the estimated ***Y-hat*** value when regressor ***x***=0 (x_i = {x_1, x_2, … x_N});

– the slope of the line $\mathbf{b}_1$ is the average change in dependent variable ***y*** for each change of one unit in regressor ***x***;

– the individual random error (*ε*) terms $\mathbf{e}_i$ ($\mathbf{e}_i = y_i - \hat{y}_i$) have a mean of zero (ē = 0), i.e. the estimation is unbiased.

The ***Least Squares (LS)*** method is a standard approach to the approximate solution of overdetermined systems, i.e., sets of equations in which there are more equations than unknowns. "***Least squares***" means that the overall solution minimizes the sum of the squares of the errors made in the results of every single equation like (6-2).

[9] In statistics, ***consistency*** means that as the number of data points used increases indefinitely, the distributions of the estimates become more and more concentrated near the true value of the parameter being estimated.

[10] ***Efficiency*** is a term used in the comparison of various statistical procedures and, here in particular, it refers to a measure of the optimality of an estimator – essentially, a more efficient estimator needs fewer samples than a less efficient one to achieve a given performance and it attains the minimum variance for all parameters.

LS is the first technique used for the regression coefficient (***β***) estimation. In fact, it was developed long time before (Sir) Francis Galton publication of Regression towards mediocrity in hereditary stature in 1886. In 1809 Carl Friedrich Gauss published his method of calculating the orbits of celestial bodies. In that work, he claimed to have been in possession of the ***LS*** method since 1795. This naturally led to a priority dispute with Legendre (1805). However, to Gauss's credit, he went beyond Legendre and succeeded in connecting the ***LS*** method with the principles of probability and to the normal distribution.

In 1822, Gauss was able to state that the ***LS*** approach to regression analysis is optimal in the sense that in a linear model where the errors have a mean of zero, are uncorrelated, and have equal variances, the best linear unbiased estimator of the coefficients is the ***least-squares estimator***. This result is known as the ***Gauss–Markov theorem***[11].

The objective in ***LS*** consists of adjusting the parameters of a model function to best fit a data set. A simple dataset consists of ***n*** points, data pairs (x_i,y_i), i = {*1, 2, … n*}, where x_i is an explanatory variable and y_i is the dependent variable. The model function has the form $f(x,\beta_j)$, like equation (6-4). The goal is to find the parameter values β_j for the model which "best" fits the data. The ***LS*** method finds its optimum when the ***Sum of Squared Errors*** is a minimum:

$$SSE = \sum_{i=1}^{n} e_i^2 \qquad (6\text{-}5)$$

Least squares problems fall into two categories – linear or ***ordinary least squares*** and ***non-linear least squares***, depending on whether or not the residuals are linear in all unknowns. The linear LS problem occurs in statistical regression analysis and it has a closed-form solution[12] that is unique. In contrast, the non-linear LS problem has not closed-form solution and is usually solved by an iterative procedure, where at each iteration the non-linear system is approximated by a linear one, and thus the core calculation is similar in both cases.

Ordinary Least Squares (OLS)

OLS is an approach fitting a mathematical or statistical model to data in cases where the idealized value provided by the model for any data point (x_i,y_i) is expressed linearly in terms of the unknown parameters β_j of the model (6-2). Minimization of (6-5) results in a ***set of normal equations***, a set of simultaneous linear equations in the parameters, which are solved to yield the parameter estimators b_j. The resulting fitted model (6-4) can be used to summarize the data,

[11] See http://en.wikipedia.org/wiki/Gauss%E2%80%93Markov_theorem

[12] A closed-form solution (or closed-form expression) is any formula that can be evaluated in a finite number of standard operations, see http://en.wikipedia.org/wiki/Closed-form_expression

to predict unobserved values from the same system, and to understand the mechanisms that may underlie the system.

OLS is the simplest and thus most common estimator. It is conceptually clear and computationally straightforward. ***OLS*** estimates are used to analyze both experimental and observational data, which makes it very suitable for business forecasting where most historical data (both cross-sectional and time-series) are collected via observations.

The ***OLS*** method ***minimizes*** the ***Sum of Squared Errors (SSE)*** in (6-5) and leads to a closed-form expression like (6-4) for the estimated values $\boldsymbol{b_j}$ of the unknown parameters $\boldsymbol{\beta_j}$ in (6-2). The estimator is ***unbiased*** and ***consistent*** if the errors ($\boldsymbol{\varepsilon}$) have finite variance and are uncorrelated with the explanatory variable $\boldsymbol{x}$.

OLS is also ***efficient*** under the assumption that the errors ($\boldsymbol{\varepsilon}$) have finite and constant variance (i.e. ($\boldsymbol{\varepsilon}$) are homoscedastic). Unfortunately, as mentioned above, the condition of homoscedasticity can fail with either experimental or observational data.

Generalized Least Squares (GLS) and related techniques

Generalized least squares (GLS) is an extension of the ***OLS*** method, which allows efficient estimation of $\boldsymbol{\beta}$ when either heteroscedasticity, or correlations, or both are present among the error terms of the model, as long as the form of heteroscedasticity and correlation is known independently of the data. To handle heteroscedasticity when the error terms are uncorrelated with each other, ***GLS*** minimizes a weighted analogue to the sum of squared residuals (6-5) from ***OLS***, where the weight for the $\boldsymbol{i}^{\text{th}}$ case is inversely proportional to the variance of the $\boldsymbol{i}^{\text{th}}$ error (ε_i). This special case of ***GLS*** is called "***weighted least squares***".

GLS can be viewed as applying a linear transformation to the data so that the assumptions of ***OLS*** are met for the transformed data. For ***GLS*** to be applied, the covariance structure of the errors must be known up to a multiplicative constant.

Percentage LS (Tofallis, 2009) focuses on reducing percentage errors, which is useful in the field of forecasting or time series analysis. It is also useful in situations where the dependent variable has a wide range without constant variance, as here the larger residuals at the upper end of the range would dominate if ***OLS*** were used. When the percentage or relative error is normally distributed, ***least squares percentage regression*** provides ***maximum likelihood*** estimates (see below). Percentage regression is linked to a multiplicative error model, whereas ***OLS*** is linked to models containing an additive error term.

Iteratively reweighted least squares (IRLS) is used when heteroscedasticity, or correlations, or both are present among the error terms of the model, but where little is known about the covariance structure of the errors independently of the data (del Pino, 1989). In the first iteration, OLS, or GLS with a provisional covariance structure is carried out, and the errors are obtained from the fit. Based on these residuals, an improved estimate of the covariance structure of the errors can usually be obtained. A subsequent ***GLS*** iteration is then performed using this estimate of the error structure to define the weights. The process can be iterated to convergence, but in many cases only one iteration is sufficient to achieve an efficient estimate of $\boldsymbol{\beta}$ (Carroll, 1982).

Total least squares (TLS) (Nievergelt , 1994) is an approach to least squares estimation of the linear regression model (6-2) that treats the explanatory variable $\boldsymbol{x}$ and the dependent variable $\boldsymbol{y}$ in a more geometrically symmetric manner than ***OLS***. ***TLS***, also known as ***rigorous least squares*** and in special cases as ***orthogonal regression*** is a type of errors-in-variables regression, a LS data modeling technique in which observational errors on both dependent and independent variables are taken into account. It can be applied to both linear and non-linear models.

Maximum-likelihood estimation and related techniques

Maximum Likelihood Estimation (MLE) can be performed when the distribution of the error terms is known to belong to a certain parametric family of probability distributions. When it is a normal distribution with zero mean and finite variance, the resulting estimate is identical to the OLS estimate. GLS estimates are maximum likelihood estimates when error (ε) follows a multivariate normal distribution with a known covariance matrix.

In general, for a fixed set of data and underlying statistical model, the ***MLE*** selects the set of values of the model parameters that maximizes the likelihood function[13] (i.e. the probability to reach their unknown, true values). Intuitively, this maximizes the "agreement" of the selected model with the observed data, and for discrete random variables, it indeed maximizes the probability of the observed data under the resulting distribution. ***MLE*** gives a unified approach to estimation, which is well-defined in the case of the normal distribution and many other problems. However, in some complicated problems ***MLE*** are unsuitable or do not exist.

Ridge regression and other forms of penalized estimation such as ***Lasso regression***, deliberately introduce bias into the estimation of $\boldsymbol{\beta}$ in order to reduce the variability of the estimate. The resulting estimators generally have a lower mean squared error than the ***OLS***

[13] A likelihood function (often simply the likelihood) is a function of the parameters $\boldsymbol{\beta}$ of a statistical model. The likelihood of a set of parameter values, $\boldsymbol{\beta}$, given outcomes $\boldsymbol{x}$, is equal to the probability of those observed outcomes $\boldsymbol{x}$, given those parameter values (i.e. their conditional probability). In informal contexts, "likelihood" is often used as a synonym for "probability."

estimates, particularly when multicollinearity (see section 6.3) is present. They are generally used when the goal is to predict the value of the response variable ***y*** for values of the predictors ***x*** that have not yet been observed. These methods are not as commonly used when the goal is inference since it is difficult to account for the bias.

Least absolute deviation (LAD) regression is a robust estimation technique in that it is less sensitive to the presence of outliers than ***OLS*** but is less efficient than ***OLS*** when no outliers are present. It is equivalent to ***MLE*** under a Laplace distribution (also known as double exponential distribution because it can be thought of as two exponential distributions spliced together back-to-back) model for (ε).

Adaptive estimation – if we assume that error terms (ε_i) are independent of the regressors $\{x_i\}$, the optimal estimator is the 2-step ***MLE***, where the first step is used to non-parametrically estimate the distribution of the error term (Stone, 1975).

Bayesian linear regression applies the framework of Bayesian statistics to linear regression (Box & Tiao, 1992). In particular, the regression coefficients β are assumed to be random variables with a specified prior distribution. The prior distribution can bias the solutions for the regression coefficients, in a way similar to (but more general than) ridge regression or lasso regression. In addition, the Bayesian estimation process produces not a single point estimate for the "best" values of the regression coefficients but an entire posterior distribution, completely describing the uncertainty surrounding the quantity. This can be used to estimate the "best" coefficients using the mean, mode, median, or any other function of the posterior distribution.

All the multitude of techniques described above shows that there are many different methods for regression coefficients estimation. It is very important to understand and remember that all these methods are based on specific assumptions about error terms, explanatory variables, and parameters of the model. We have to understand what these assumptions are, be able to explain them and identify when they are true and apply the corresponding method accordingly.

C. Fitting the regression line

Suppose there is a dataset with N data points (x_i,y_i), $i = \{1, 2, \ldots N\}$, where x_i is an explanatory variable and y_i is the dependent variable. The function that describes x and y is given by equation (6-2). The goal is to find the equation of the straight line (6-4). Since there are many possible methods (as discussed above), and most of them involve pretty complicated calculations from calculus, the best way to find the parameters β_j estimators (i.e. unknown regression coefficients b_j) is by using computer software from a general type such as MS Excel, or special statistical software like Statgraphics, Gretl or similar ones.

This is how it works in the real-life business. In practice, we have a collection of observations for $\boldsymbol{x_i}$ and $\boldsymbol{y_i}$, and we do not know the values of $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$. These need to be estimated from the data available – a process known as "*fitting a line through the data*".

There are many possible choices for $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$, using different methods, each choice giving a different line. To illustrate "fitting the regression line" process and to discuss some important elements from its output, we are going to apply the ***OLS*** estimator. This is, as already mentioned, because: (a) linear relationships are the simplest and the easiest to work with; (b) linear regression coefficients $\boldsymbol{\beta_j}$ have very clear business interpretation; (c) the "true" relationship between the variables $\boldsymbol{x_i}$ and $\boldsymbol{y_i}$ is often at least approximately linear and (d) ***OLS*** is the basis (with some assumptions or data transformations) for the most estimation methods.

The OLS provides a way of choosing $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ effectively by minimizing the sum of the squared errors (6-5), which after some transformations will look like:

$$\sum_{i=1}^{N} \varepsilon_i^2 = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{6-6}$$

By using either calculus, the geometry of inner product spaces, or simply expanding to get a quadratic expression in $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$, it can be shown that the values of $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ that minimize the objective function (6-6) are:

$$\boldsymbol{b_1 = \hat{\beta}_1 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}} \tag{6-7}$$

and

$$\boldsymbol{b_0 = \hat{\beta}_0 = \bar{y} - b_1\bar{x}} \tag{6-8}$$

where $\bar{y}$ and $\bar{x}$ are the arithmetic means of observations for $\boldsymbol{y_i}$ and $\boldsymbol{x_i}$.

The estimated regression line is shown in Fig.6-7, where the "true line" is the unknown regression model for the population and the "estimated line" is the regression model for a random sample of size $\boldsymbol{N}$. The estimated regression line could be used for forecasting. For each value of $\boldsymbol{x_i}$, we can forecast the corresponding value of $\boldsymbol{y_i}$ using equation (6-4) with the estimated values for $\boldsymbol{b_0}$ and $\boldsymbol{b_1}$.

The estimated regression line has a few very important numerical properties:

- The line goes through the "center of mass" point ($\bar{x}$,$\bar{y}$).

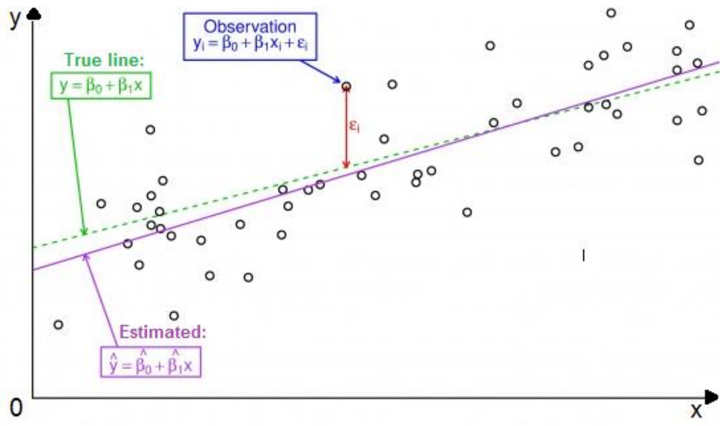

y
Observation
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
True line:
$y = \beta_0 + \beta_1 x$
ε_i
Estimated:
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
0
x

Fig.6-7 True and Estimated Regression lines for variables $\boldsymbol{y_i}$ and $\boldsymbol{x_i}$

- The sum of the errors (or residuals) is equal to zero (i.e. the estimators for $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ are unbiased):

$$\sum_{i=1}^{N} e_i = 0 \qquad (6\text{-}9)$$

- The linear combination of the residuals, in which the coefficients are the ***x***-values, is equal to zero too:

$$\sum_{i=1}^{N} x_i e_i = 0 \qquad (6\text{-}10)$$

As a result of these properties, it is clear that the arithmetic mean of the errors is zero, and the correlation between the errors and the observations for predictor variable $\boldsymbol{x_i}$ is also zero.

In a simple linear regression, where there is only one explanatory variable and a constant, the ***OLS*** coefficient estimates have a simple form that is closely related to the correlation coefficient (6-3). If we make some substitutions in (6-7) the new equation is:

$$b_1 = \hat{\beta}_1 = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} = r_{xy}\frac{S_y}{S_x} \qquad (6\text{-}11)$$

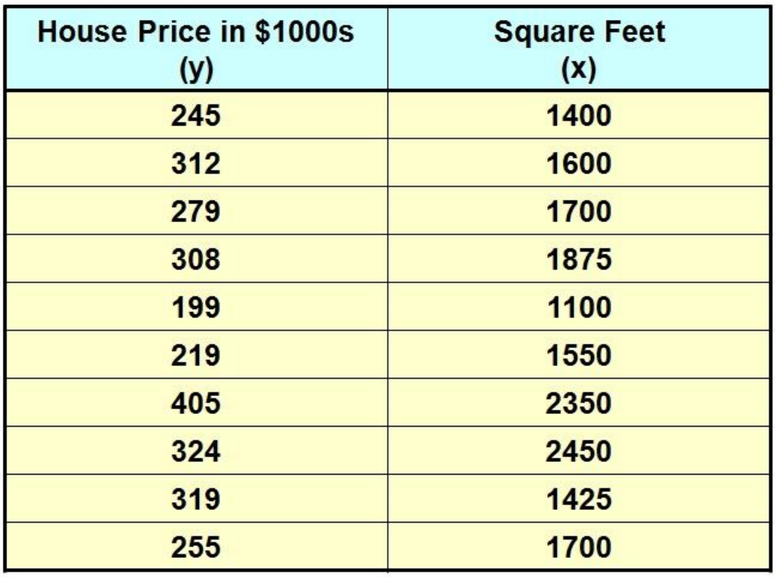
where $\boldsymbol{r_{xy}}$ is the sample correlation coefficient between variables $\boldsymbol{x_i}$ and $\boldsymbol{y_i}$;
$\boldsymbol{S_x}$ is the standard deviation of $\boldsymbol{x_i}$ and $\boldsymbol{S_y}$ is correspondingly the standard deviation of $\boldsymbol{y_i}$.

This shows the role which the correlation coefficient $\boldsymbol{r_{xy}}$ plays in the regression line of standardized data points. It is sometimes useful to calculate $\boldsymbol{r_{xy}}$ from the sample data independently, solving equation (6-11) for $\boldsymbol{r_{xy}}$. What is more important is that from (6-11) it is clear that in linear regression analysis the signs of correlation coefficient $\boldsymbol{r_{xy}}$ and estimator $\boldsymbol{b_1}$ are identical, i.e. the direction of the relationship between variables $\boldsymbol{x_i}$ and $\boldsymbol{y_i}$ according to both correlation coefficient $\boldsymbol{r_{xy}}$ and regression coefficient $\boldsymbol{b_1}$ must always be the same.

Interpreting the Results in Linear Regression Output

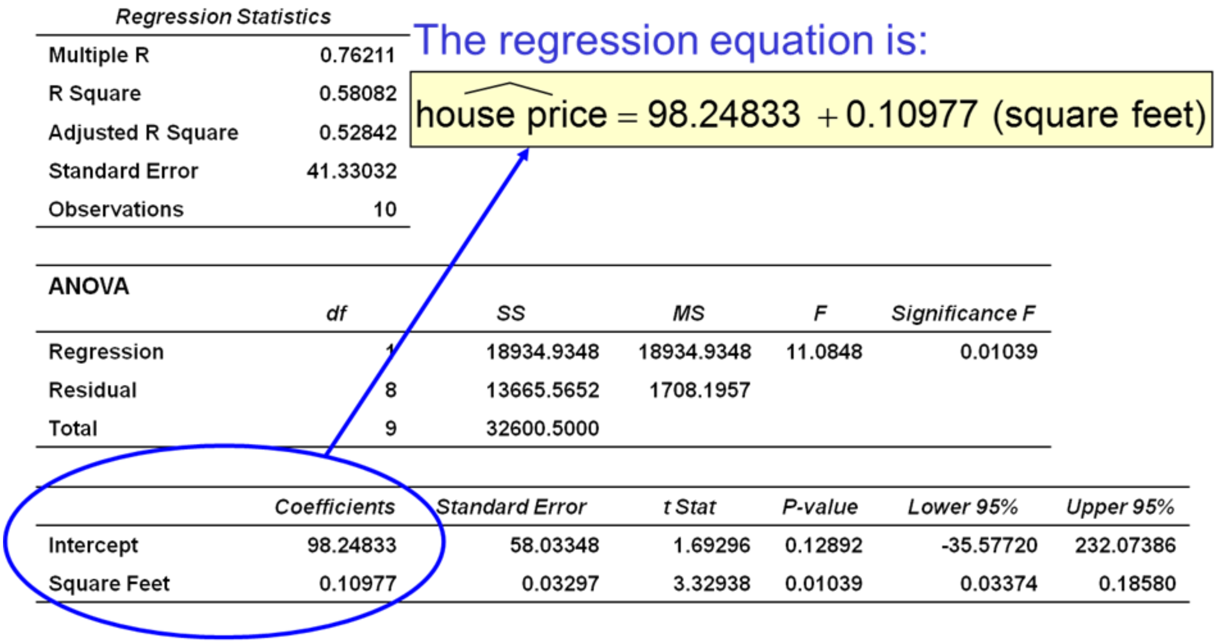
Example: Suppose, a real estate agent wishes to examine the relationship between the selling price of a home and its size, and he selected a random sample of 10 houses. Here, the variables (x_i,y_i), i = {*1, 2, … 10*}, are given as: y_i is the dependent variable "House price in $1,000s" and x_i is an explanatory variable "Square feet". The observations for variables y_i and x_i are presented in Table 6.1 below.

Table 6.1 Sample Data for *House Price* example



House Price in $1000s (y)	Square Feet (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Using MS Excel (Data/Data Analysis/Regression) with sample data from Table 6.1 will return the following output:



Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

The regression equation is:

$$\widehat{\text{house price}} = 98.24833 + 0.10977\ (\text{square feet})$$

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Fig.6-8 MS Excel Regression output for the *House Price* example

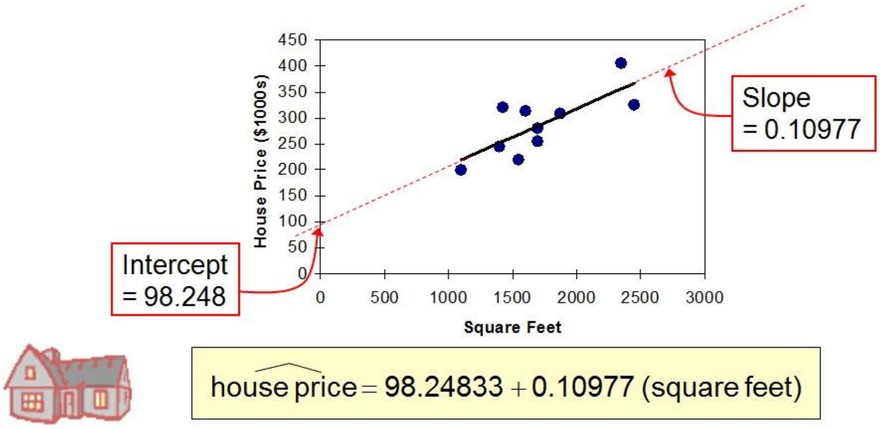


Fig.6-9 Scatter diagram and estimated regression line for the *House Price* example

The scatterplot and estimated regression line are presented in Fig.6-9. Note that there should be some variations in regression output, depending on the software used, but these differences mainly concern how the results are presented, not what is presented.

The first important element of the output is the estimator's part (i.e. the regression coefficients with their statistics). The first column always (no matter what software is used) presents the names of the explanatory variables (with or without intercept[14]), followed by their estimated regression coefficients. These details give us the information that we need to state the sample regression line. In our example, as we can see it is:

$$\widehat{\text{house price}} = 98.24833 + 0.10977x\ (\text{square feet}) \tag{6-12}$$

Interpreting the intercept $\mathbf{b_0}$ = 98.24833. As it often happens with the intercept, this is a case where the interpretation is nonsense as it is impossible for a house to have zero square feet. The interpretation of the intercept requires that $\boldsymbol{x_i} = 0$ is within the range of observed $\boldsymbol{x_i}$ values, i.e. the value of $\boldsymbol{x_i} = 0$ makes sense. Only then we can say that the intercept $\mathbf{b_0}$ is the predicted value ***Y-hat*** (**ŷ**) corresponding to $\boldsymbol{x_i} = 0$. In our case, $\mathbf{b_0 = 98.24833}$ just indicates that, for houses within the range of sizes observed, $98,248.33 is *the portion of the house price not explained by the square feet variable*.

It is important to note that the intercept in a regression model has only mathematical interpretation – the point where the regression line intersects the y-axis. It is rarely a number with any direct economic meaning, but even when $\boldsymbol{x_i} = 0$ case does not make sense at all, the intercept is an important part of the model. Despite the fact that sometimes, some researchers consider a linear regression model without the intercept term, in general, it always should be presented in the model. Without it, the slope coefficient $\mathbf{b_1}$ can be distorted unnecessarily.

[14] Sometimes we can see "***constant***" instead of "***intercept***", but they are just synonyms in this case.

Interpreting the slope b_1 = 0.10977. The slope measures the estimated change in the average value of y_i as a result of a one-unit change in x_i. Here, this coefficient tells us that for each additional one square foot of size the average value of a house will increase on average by 0.10977($1000) = $109.77.

The next important number in the output is the ***standard error of the regression*** S_ε, which measures how well the model has fitted the data. It is often known and presented in the regression output as "the standard deviation of the residuals", or just "the standard error", as it shown in Fig.6-8. It is $41.33032 in this example, but we should warn here that its evaluation can be highly subjective as it is scale dependent. This number is the estimated standard deviation of the "noise" in the dependent variable that is unexplainable by the independent variable(s), and it is a lower bound on the standard deviation of any of the model's forecast errors, under the assumption that the model is correct.

The point "explained vs. unexplained" variation needs some important clarifications. We assume that the total variation of dependent variable y_i is made up of two parts (6-13):

$$SST = SSE + SSR \qquad (6\text{-}13)$$

Total sum of Squares Sum of Squares Error Sum of Squares Regression

$$SST = \sum (y - \bar{y})^2 \qquad SSE = \sum (y - \hat{y})^2 \qquad SSR = \sum (\hat{y} - \bar{y})^2$$

where SST measures the variation of the y_i values around their mean $\bar{y}$;

– SSE is the variation attributable to factors other than the relationship between x_i and y_i;

– SSR is the explained variation attributable to the relationship between x_i and y_i.

Fig.6-10 explains graphically this point:

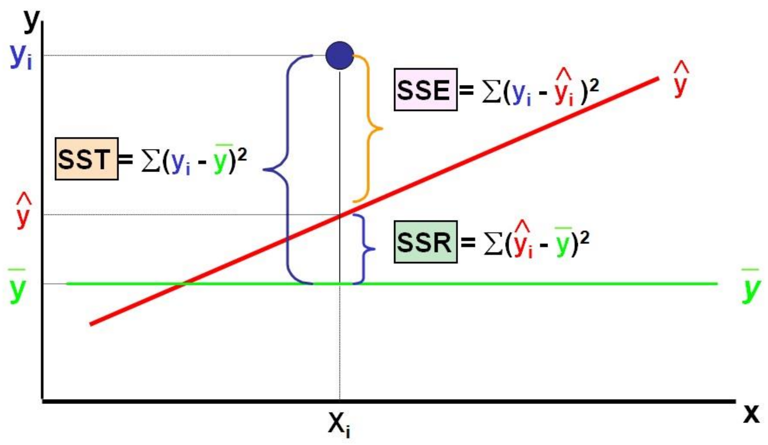

y
y_i
SST = $\Sigma(y_i - \bar{y})^2$
SSE = $\Sigma(y_i - \hat{y}_i)^2$
$\hat{y}$
$\hat{y}$
SSR = $\Sigma(\hat{y}_i - \bar{y})^2$
$\bar{y}$
$\bar{y}$
X_i
x

Fig.6-10 Explained and Unexplained Variation in regression models

The ***standard error of the regression*** $\boldsymbol{S_\varepsilon}$ or the variation (the scatter, or dispersion) of observations around the regression line is then estimated by:

$$\boxed{s_\varepsilon = \sqrt{\frac{SSE}{N-k-1}}} \qquad (6\text{-}14)$$

where ***SSE*** is the sum of squared errors, ***N*** is the sample size and ***k*** is the number of explanatory variables in the model. The denominator in the formula (***N-k-1***) is also known as ***degrees of freedom (d.f.)***, which is the number of values in the final calculation of a statistic that are free to vary (Walker, 1940).

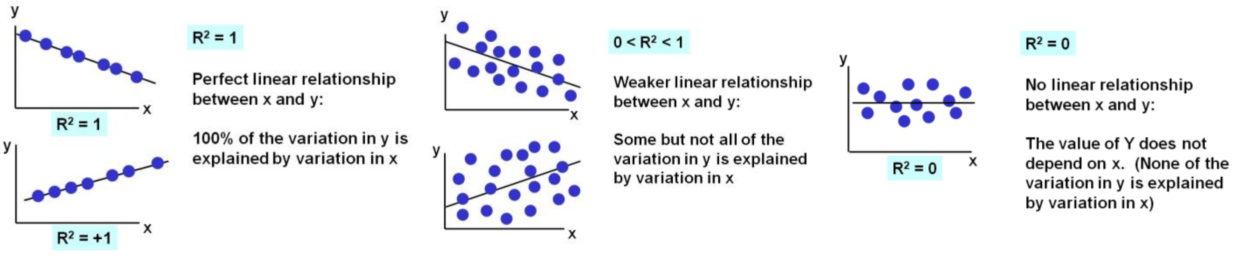
It is clear from formula (6-14) that $\boldsymbol{S_\varepsilon}$ is scale dependent and cannot be interpreted, explained or evaluated directly. The main reason we introduce it here is that it is required when constructing some important confidence intervals for regression coefficients.

A common way to summarize how well a linear regression model fits the data is via the coefficient of determination or $\boldsymbol{R^2}$ (***R-squared***), which is the square of the correlation between the observed values $\boldsymbol{y_i}$ and the predicted values ($\hat{y}$). Alternatively, it can also be calculated as:

$$\boxed{R^2 = \frac{SSR}{SST}} \quad \text{where} \quad \boxed{0 \le R^2 \le 1} \qquad (6\text{-}15)$$

This is the proportion of variation in the dependent variable $\boldsymbol{y_i}$ that is accounted for (or explained) by the regression model. In the case of the single independent variable, the coefficient of determination equals the square of the simple correlation coefficient $\boldsymbol{r_{xy}}$ (6-3), i.e. $\boldsymbol{R^2 = r_{xy}^2}$. Graphical illustration of ***R-squared*** is presented in Fig.6-11.

In our example $\boldsymbol{R^2 = 0.58082}$, which means that 58.08% of the variation in house prices, is explained by variation in square feet, i.e. by the regression model. There are no set rules of what a good $\boldsymbol{R^2}$ value is and in general, it is accepted that the larger $\boldsymbol{R^2}$ (or the closer to one), the better. Apparently, in our example 0.58082 is far away from this goal and we should look for some improvements in the regression model, mainly by adding more explanatory variables.



a) b) c)

Fig.6-11 Examples of ***R-squared***

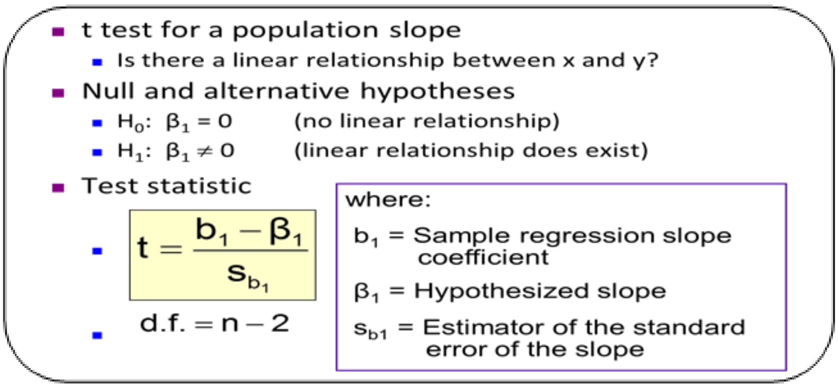
It is worth noting that the ***R-squared*** value is commonly used in forecasting too, but this is not correct. Validating a model's out-of-sample forecasting performance (i.e. cross-validation) is much better than measuring the in-sample ***R-squared*** value.

Let us return to estimator's part, i.e. the regression coefficients with their statistics, as shown in Fig.6-8. ***The standard error of a regression coefficient*** is the estimated standard deviation of the error in estimating this coefficient. The standard error of the slope (***S_{b1}***) is:

$$S_{b_1} = \frac{S_\varepsilon}{\sqrt{\sum (x-\bar{x})^2}} = \frac{S_\varepsilon}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}} \tag{6-16}$$

In our example ***S_{b1}* = 0.03297** and could be found next to the slope value in Fig.6-8. This number is the basis for testing a hypothesis about the slope significance. We use a hypothesis test (recall section 3.3 B) ***Statistical hypothesis test***) to formally examine whether there is enough evidence to show that variables ***x_i*** and ***y_i*** are related. If they are unrelated, then the slope parameter ***$\beta_1 = 0$***. So, we can construct a test to see if it is plausible that ***$\beta_1 = 0$*** given the observed data.

The logic of hypothesis tests is to assume the thing we want to disprove, and then to look for evidence that the assumption is wrong. In this case, we assume that there is no relationship between ***x_i*** and ***y_i***. This is known as the "null hypothesis" ***Ho***. Evidence against this hypothesis is provided by the value of ***b_1***, the slope estimated from the data (this is the "alternative hypothesis" ***H_1*** or ***H_A***). If ***b_1*** is significantly different from zero, we conclude that the ***Ho*** is incorrect and that the evidence suggests there is a relationship between ***x_i*** and ***y_i*** , i.e. ***H_A*** is the true one. Fig.6-12 summarizes the process of making statistical inference about the slope coefficient[15] in linear regression.



- t test for a population slope
  - Is there a linear relationship between x and y?
- Null and alternative hypotheses
  - H_0: $\beta_1 = 0$ (no linear relationship)
  - H_1: $\beta_1 \neq 0$ (linear relationship does exist)
- Test statistic
  - $t = \dfrac{b_1 - \beta_1}{s_{b_1}}$
  - $\text{d.f.} = n - 2$

where:

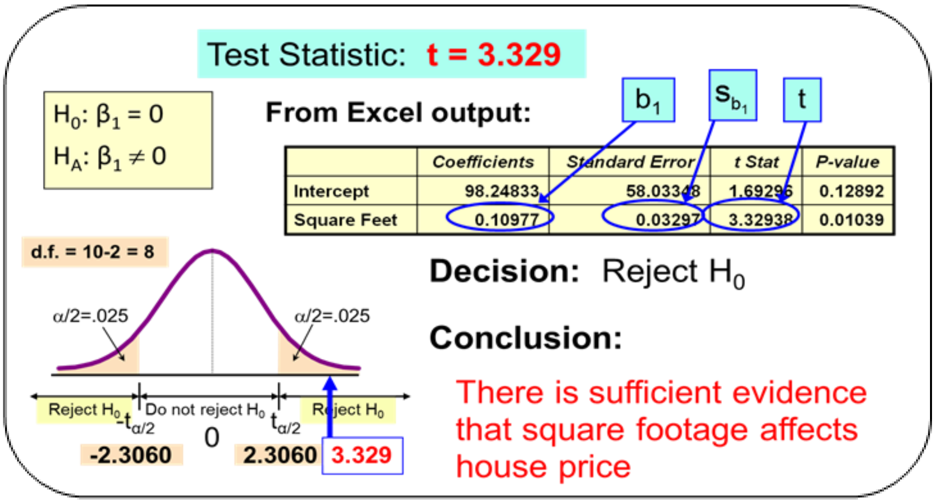
b_1 = Sample regression slope coefficient

β_1 = Hypothesized slope

s_{b1} = Estimator of the standard error of the slope

Fig.6-12 Inference about the Slope – ***t*** test

[15] In general, we do not care about the test of the intercept unless it is possible for all the independent variables to simultaneously go to zero and we are interested in whether the prediction should also be zero in such a situation.



Test Statistic: $t = 3.329$

H_0: $\beta_1 = 0$

H_A: $\beta_1 \neq 0$

From Excel output:

	Coefficients	Standard Error	t Stat	P-value
Intercept	98.24833	58.03348	1.69296	0.12892
Square Feet	0.10977	0.03297	3.32938	0.01039

Decision: Reject H_0

Conclusion:

There is sufficient evidence that square footage affects house price

Fig.6-13 Inferences about the Slope – *t* test example

In our example the question to ask would be "Does square footage of the house affect its sales price?" and the statistical test of significance is presented in Fig.6-13. Since the computed ***t*** value is within the region of ***Ho*** rejection, we make a decision to reject ***Ho*** and accept ***H_A***. The conclusion, based on the accepted level of significance ***α=0.05*** (i.e. 5% tolerable risk to reject ***Ho*** when it is true) and the available sample of observations, is that "*There is sufficient evidence that square footage affects house price*."

Note, that ***t*** test does not indicate precisely how strong our decision is, and how much ***x_i*** and ***y_i*** are related. Another statistic to help us with this question is the correlation coefficient ***r_{xy}***, which is available under the name "*Multiple R*" on the top side in the regression output in Fig.6-8. The value of 0.76211 indicates above the moderate strength of the relationship between the selling price of a home (in $1,000) and its size (measured in square feet). Again, like with the ***t*** test, we don't know if this is enough and even if we test this value of ***r_{xy}***, the result will be similar to the slope test, according to a pre-determined threshold level of significance ***α=0.05.***

To determine precisely how big must be the difference between ***β_1*** and ***b_1***, before we would reject the ***Ho***, we should calculate the probability of obtaining a value of ***b_1*** as large as we have calculated if the null hypothesis was true. This probability is known as the "***P-value***" and we compare it with the tolerable risk (level of significance ***α***) to reject ***Ho*** when it is true.

The details of the calculation need not concern us here, because the software (MS Excel in our case) will provide the P-values when we need them[16]. The ***P-value*** corresponding to the slope is in the column with the same name right after the ***t Stat*** column (Fig.6-8 or Fig.6-13) and in our example **P = 0.01039**. In other words, there is about 1% chance that observed data would have arisen if the null hypothesis were true.

[16] In manual process this is one additional step after computing the ***t*** test value and because of these extra calculations some researchers just omit it.

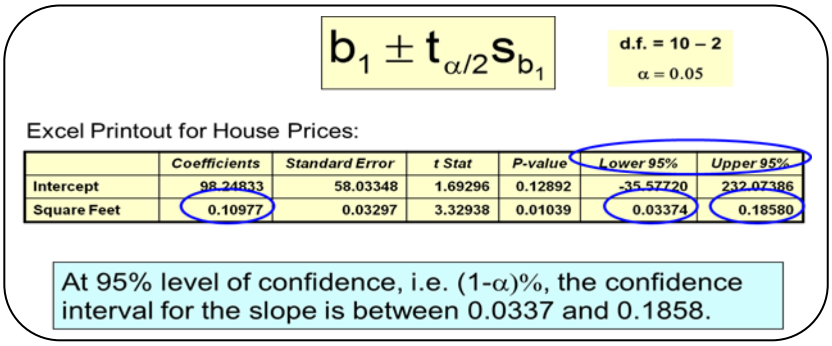
The ***P-value*** not only provides more precise information to make a correct decision in rejecting (or not) ***Ho***, it shows how big the risk of rejecting true ***Ho*** is as well. In our example, the ***P-value*** of 0.01039 indicates that the minimum threshold level of significance ***α*** to reject ***Ho*** should be just above this value, say 0.0104**.** In general, it is accepted that threshold levels of significance ***α*** between 0.01 and 0.05 indicate strong evidence to reject ***Ho***, and levels below 0.01 indicate very strong evidence in favor of alternative hypothesis $\boldsymbol{H_A}$, i.e. in rejecting ***Ho***. In our example, ***P-value*** of 0.01039 shows that it is much more likely that there is a relationship between square footage and house price.

The standard error of a regression slope ($\boldsymbol{S_{b1}}$) and the computed **t** statistics are used to calculate important limits for regression coefficient variation, known as "***Confidence Interval Estimate of the Slope***[17]". Fig.6-14 presents this process in summary.

By the rule of thumb, an approximate 95% confidence interval for a coefficient is the point estimate plus or minus two standard errors. The exact border values for a two-tailed 95% confidence interval are shown at the right-hand end of the estimator’s part table (see Fig.6-8 or 6-14), and in our example these values are [0.03374,0.18580]. The regression coefficient of Square Feet is 0.10977 with a standard error of 0.03297 and its confidence interval *does not include zero*, which means that $\boldsymbol{b_1=0}$ is not among the possible solutions at $\boldsymbol{\alpha=0.05}$ (i.e. tolerable risk of 5%), given the observed data.

The conclusion we can make is: "*Since the units of the house price variable are $1000s, we are 95% confident that the average impact on sales price is between $33.70 and $185.80 per square foot of house size*". This appears to be a big interval and not a reasonably precise estimate of the strength of the *price-size* relationship. Again, the recommendation is that we should look for some improvements in the regression model.

To avoid unnecessary overlaps in the comments, the other features of the regression output and corresponding statistical tests will be explained later in this chapter.



$$b_1 \pm t_{\alpha/2} s_{b_1}$$

d.f. = 10 – 2

$\alpha = 0.05$

Excel Printout for House Prices:

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

At 95% level of confidence, i.e. (1-α)%, the confidence interval for the slope is between 0.0337 and 0.1858.

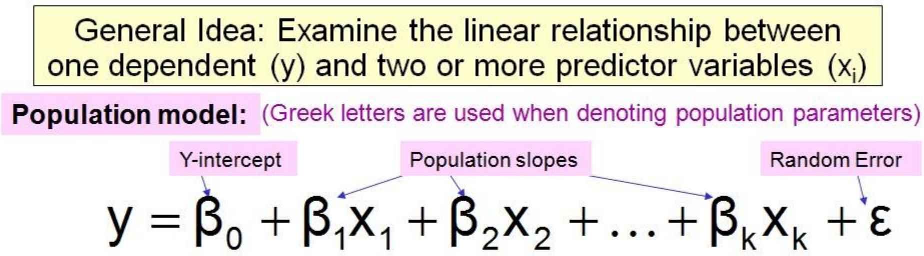
Fig.6-14 Confidence Interval Estimate of the Slope

[17] Confidence intervals were devised to give a plausible set of values the estimates might have if we repeated the experiment a very large number of times.

6.3. Multiple Regression and Model Building

The general idea in **Multiple Regression Analysis** is very similar to the one of a simple regression, however, there are several important points to understand and remember. First of all, **Multiple Regression Analysis** is a statistical process for estimating relationships among more than two variables. It includes many techniques for modeling and analyzing several variables when the focus is on the relationship between one dependent variable and two or more factors. The estimation target is a function of the explanatory variables and the general form of a multiple linear regression is similar to the simple regression form (6-2), but the number of explanatory variables **k** in the model is larger than one.

A. General Form of the Multiple Regression



General Idea: Examine the linear relationship between one dependent (y) and two or more predictor variables (x_i)

Population model: (Greek letters are used when denoting population parameters)

Y-intercept
Population slopes
Random Error

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon \qquad (6\text{-}17)$$

More specifically, regression analysis helps us understand how the typical value of the explained variable **y** changes when any one of the explanatory variables **x_j** (***j*** = {***1, 2, … k***} is varied, while the other regressors are held fixed. Thus, the corresponding regression coefficients **β_j** (***j*** = {***1, 2, … k***}) in (6-17) measure the effect of each regressor after taking account of the effect of all other regressors in the model, i.e. coefficients **β_j** measure the marginal effects of the explanatory variables. These coefficients are also known as ***partial regression coefficient*** or ***net regression coefficient***.

The general form (6-17) could be presented also as a model which relates **y** to a function of **X** and **β**:

$$\boldsymbol{y \approx f(X, \beta)} \qquad (6\text{-}18)$$

where **X** and **β** are vectors of the regressors and the unknown parameters.

Most commonly, regression analysis estimates the conditional expectation of the dependent variable given the explanatory variables – that is, the average value of the response variable when the regressors are fixed. The approximation is usually formalized as:

$$\boldsymbol{E(Y|X) = f(X, \beta)} \qquad (6\text{-}19)$$

To carry out regression analysis, the form of the function ***f*** must be specified. Here, it is very important to mention the differences between *linear* and *non-linear models,* between *linear* and *non-linear least squares*:

- The model function (***f***) in ***OLS*** (or ***Linear Least Squares - LLS***) is a linear combination of parameters of the form $f = X_{i1}\beta_1 + X_{i2}\beta_2 + \cdots$The model may represent a straight line, a parabola or any other linear combination of functions. In ***Non-Linear Least Squares (NLLS)*** the parameters appear as functions, such as $\boldsymbol{\beta^2}$**,** $\mathbf{e^{\beta x}}$ and so forth. If the derivatives $\partial f/\partial \beta_j$ are either constant or depend only on the values of the regressors, the model is linear in the parameters. Otherwise the model is non-linear.
- Algorithms for finding the solution to a NLLS problem require initial values for the parameters, but LLS does not.
- Like LLS, when solving the system of normal linear equations, solution algorithms for NLLS often require that the ***Jacobian determinant***[18] be calculated. Analytical expressions for the partial derivatives can be complicated. If analytical expressions are impossible to obtain either the partial derivatives must be calculated by numerical approximation or an estimate must be made of the ***Jacobian determinant***.
- In NLLS non-convergence (failure of the algorithm to find a minimum) is a common phenomenon, since data are fitted by a method of successive approximations, whereas the LLS is globally concave, so non-convergence is not an issue.
- NLLS is usually an iterative process, which has to be terminated when a convergence criterion is satisfied. LLS solutions can be computed using direct methods.
- In LLS the solution is unique, but in NLLS there may be multiple minima in the sum of squares.
- Under the condition that errors and regressors are uncorrelated, LLS yields unbiased estimates, but even under that condition NLLS estimates are generally biased.
- Last but not least, as already mentioned above and then explained in the *House Price* example, linear regression coefficients β_j have a very clear business interpretation, which most of the time non-linear coefficients do not.

These differences, advantages, and disadvantages must be considered whenever a selection of a non-linear model should be made or a solution to a non-linear problem is being sought. Things become much more complicated when studying and predicting complex object and processes in the form of systems of equations (see Chapter 9).

[18] See http://en.wikipedia.org/wiki/Jacobian_matrix_and_determinant

Another important point is that in practice we are ***fitting the regression line through the data,*** which means that the general ***Multiple Regression form*** (6-17) in fact is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \qquad (6\text{-}20)$$

In formula (6-20) we consider **N** observations of one dependent variable y_i ={y_1, y_2, … y_N} and **k** predictor variables x_{ij} = {x_{1j}, x_{2j}, … x_{Nj}}. Thus:

– y_i is the ***i^th*** observation of the dependent variable;

– x_{ij} is ***i^th*** observation of the ***j^th*** predictor variable ***(j = 1, 2, ..., k)***;

– βj represent parameters to be estimated, and

– ε_i is the ***i^th*** independent identically distributed normal error.

The sample regression line provides an estimate of the population regression line (6-20), given by equation (6-21), which is similar to equation (7-4) in simple linear regression.

$$\overset{\text{Estimated (or predicted) value of y}}{\hat{y}_i} = \overset{\text{Estimated intercept}}{b_0} + \overbrace{b_1 x_{i1} + b_2 x_{i2} + \ldots + b_k x_{ik}}^{\text{Estimated slope coefficients}} \qquad (6\text{-}21)$$

where:

– ***Y-hat*** (**ŷ**) is the estimated average value of dependent variable y_i (i = 1, 2, ..., N);

– the ***intercept*** **b₀** is the estimated **ŷᵢ** value when all regressors x_{ij} (j = 1, 2, ..., k) are zero;

– each particular ***slope*** (***partial, net*** or just ***regression coefficient***) **bⱼ** is the average change in dependent variable y_i for each change of one unit in regressor x_j holding all other predictor variables constant;

– the individual ***random error*** (ε) terms **eᵢ** (**eᵢ** = y_i - $\hat{y}_i$) have a mean of zero (**ē** = 0), i.e. the estimation is unbiased.

In case of two predictor variables, the multiple regression could be presented graphically:

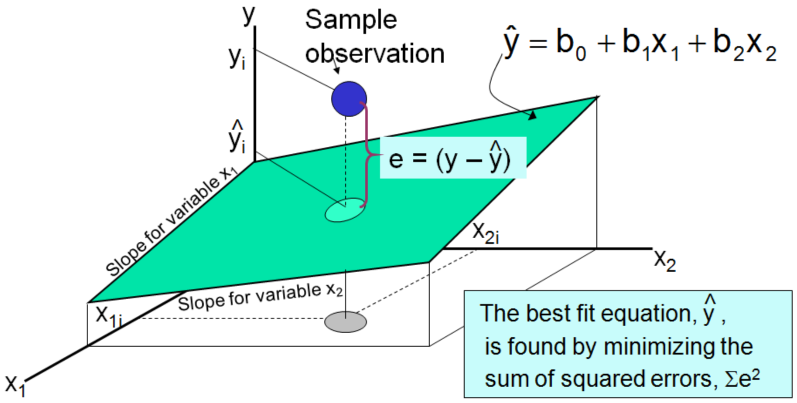


Fig.6-15 Graphical illustration of multiple regression with two predictor variables

B. Multiple Regression Assumptions

To perform a multiple regression analysis and fitting data by OLS, we must provide enough information about the dependent variable $\boldsymbol{y_i}$. If we assume that the vector of unknown parameters **β** is of length **k**, i.e. there are **k** explanatory variables $\boldsymbol{x_{ij}}$, then:

a) If N data points of the form (y,x) are observed, where $\boldsymbol{N < k}$, the classical approaches to regression analysis cannot be performed. Since the system of equations defining the regression model is underdetermined, there are not enough data to recover β.

b) If exactly $\boldsymbol{N=k}$ data points are observed and the function ***f*** is linear, the equation (6-18) can be solved exactly rather than approximately. This is the case of solving a set of **N** equations with **N** unknowns (the elements of **β**), which has a unique solution as long as the $\boldsymbol{x_{ij}}$ are linearly independent. If ***f*** is nonlinear, a solution may not exist, or many solutions may exist.

c) The most common situation is when $\boldsymbol{N > k}$ data points are observed. In this case, there is enough information in our data to estimate a unique value for **β** that best fits the data in some sense, and the regression model when applied to the data can be viewed as an overdetermined system[19] in **β**.

Only in the last case c), the regression analysis provides the tools for:

- Finding a solution for unknown parameters **β** that will minimize the distance between the measured and predicted values of the dependent variable $\boldsymbol{y_i}$.
- Under certain statistical assumptions, the regression analysis uses the surplus of information to provide statistical information about the unknown parameters **β** and predicted values of the dependent variable $\boldsymbol{y_i}$.

Statistical assumptions – when the number of observations (**N**), is larger than the number of unknown parameters (**k**), and the errors $\boldsymbol{\varepsilon_i}$ are normally distributed, then the excess of information contained in (**N−k−1**) measurements is used to make statistical predictions about the unknown model parameters. This excess of information, as mentioned above is the degrees of freedom (d.f.) of the regression. In general, the larger the d.f., the better.

Classical assumptions for multiple regression analysis and ***OLS*** include all assumptions discussed in section 6.2.B, such as the sample is representative of the population for the inference prediction; the error (**ε**) is a random variable with a mean of zero; the error terms ($\boldsymbol{\varepsilon_i}$) are uncorrelated (i.e. no autocorrelation) and have constant variance (homoscedasticity); prediction variables are measured with no error and so forth…

[19] In mathematics, a system of linear equations is considered overdetermined if there are more equations than unknowns. For details see http://en.wikipedia.org/wiki/Overdetermined_system

These are sufficient conditions for the LS estimator to possess desirable properties. In particular, these assumptions imply that the parameter estimates will be *unbiased, consistent*, and *efficient* in the class of linear unbiased estimators. It is important to note that actual data rarely satisfies the assumptions. That is, the method is used even though the assumptions are not true. Variation from the assumptions can sometimes be used as a measure of how far the model is from being useful. Regression outputs usually include analyses of tests on the sample data and methodology for the fit and usefulness of the model (see the example below).

There is one very important additional assumption, in order to perform the multiple regression, which concerns the explanatory variables. The predictors should be linearly independent, i.e. it is not possible to express any predictor as a linear combination of the others. When they are related a ***multicollinearity*** exists. If the goal is either inference or predictive modeling, the performance of OLS estimates can be poor if ***multicollinearity*** is present. Also, when high correlation exists between two explanatory variables, these two variables contribute redundant information to the multiple regression model.

Including two highly correlated predictors can adversely affect the regression results:

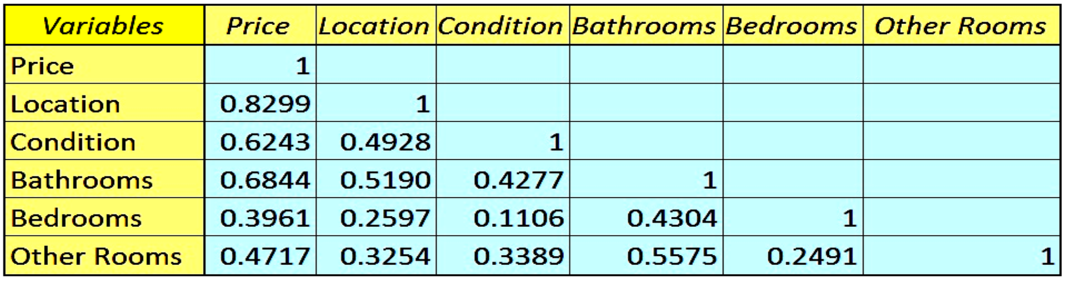
- No new information is provided.
- Can lead to unstable coefficients (large standard error and low t-statistics).
- Coefficient signs may not match prior expectations and others.

There are some indications of severe ***multicollinearity,*** which can help to detect it:

- Incorrect signs on the coefficients;
- Large change in the value of a previous coefficient when a new predictor variable is added to the model;
- A previously significant explanatory variable becomes insignificant when a new predictor is added;
- The estimate of the standard deviation of the model increases when a new predictor is added to the model.

The correlation matrix, like one presented in Table 6.2, can give us first evidence about potential ***multicollinearity*** in the regression model, if there are large (for instance $\boldsymbol{r_{x_ix_j}} > \mathbf{0.7}$) partial correlation coefficients between two predictors $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$. Since the single relationships in multiple regression have cumulative influence over the dependent variable (and that is why the multiple correlation coefficient in the model is always larger than any partial one – compare for example the Table 6.2 values and the regression output in Fig.6-16), often, even smaller than 0.7 partial correlation coefficients may cause significant multicollinearity.

Table 6.2 Correlation matrix between variables in expanded *House Price* example



Variables	*Price*	*Location*	*Condition*	*Bathrooms*	*Bedrooms*	*Other Rooms*
Price	1					
Location	0.8299	1				
Condition	0.6243	0.4928	1			
Bathrooms	0.6844	0.5190	0.4277	1		
Bedrooms	0.3961	0.2597	0.1106	0.4304	1	
Other Rooms	0.4717	0.3254	0.3389	0.5575	0.2491	1

In statistics, there are some tests to detect collinearity between regressors, such as the ***Variance Inflationary Factor (VIF)***:

$$\boxed{VIF_j = \frac{1}{1 - R_j^2}} \qquad (6\text{-}22)$$

where R^2_j is the coefficient of determination when the j^{th} explanatory variable is regressed against the remaining k – 1 predictor variables; if $\boldsymbol{VIF_j > 5}$, then $\mathbf{x_j}$ is highly correlated with the other explanatory variables.

According to some authors, (Chatterjee et al., 2000), if the underlying specification is correct, multicollinearity does not actually bias results – it just produces large standard errors in the related predictor variables. More importantly, the usual use of regression is to take coefficients from the model and then apply them to other data sets. However, if the pattern of multicollinearity in the new data differs from that in the data that was fitted, such extrapolation may introduce large errors in the predictions.

The simplest way to deal with the ***multicollinearity*** is to exclude one of the related predictors from the regression model. An explanatory variable may be dropped to produce a model with significant coefficients. However, we lose information, because we have dropped a variable. The omission of a relevant variable results in biased coefficient estimates for the remaining explanatory variables that are correlated with the dropped variable. Also, a variable omitted from the model may have a relationship with both the dependent variable and one or more of the predictor variables, which will result in the ***omitted-variable bias***[20]. Finally, the exclusion may interfere the goal of our study if the omitted regressor is part of the scenario case under analysis.

One universal approach to address many issues in statistics is to obtain more data (i.e. increase sample size N), if possible. More data can produce more precise parameter estimates (with lower standard errors), as seen from the formula in VIF for the variance of the estimate of a regression coefficient in terms of the sample size and the degree of multicollinearity.

[20] ***Omitted-variable bias (OVB)*** occurs when a model which incorrectly leaves out one or more important causal factors is created. The "bias" is created when the model compensates for the missing factor by over- or underestimating the effect of one of the other factors.

If the correlated explanators are different lagged values of the same underlying explanator, then a ***distributed lags technique*** can be used, imposing a general structure on the relative values of the coefficients to be estimated (see Chapters 8 and 9). Sometimes, ***Ridge regression*** or ***Principal component regression*** can be used (Hoerl & Kennard, 1970).

One of the best options to deal with the ***multicollinearity*** is to apply cross-validation based techniques (Golub et al., 1979). The ***GMDH*** (Madala, & Ivakhnenko, 1994), already introduced in previous chapters, provides good solutions to address this problem and many others as well. One ***GMDH algorithm*** that solves the issue of ***multicollinearity*** and many other problems in regression model building and forecasting is presented in Chapter 9.

C. Multiple regression model specification

At the beginning of every particular forecasting model building, before performing regression or any other analysis, the researcher should clarify several important points. The first group of them, which is referred to as ***Model Specification***, is to be done at the first phase of the process. Sometimes, it is wrongly considered that the specification is the whole process of developing a regression model (i.e. that ***Model Specification*** is the same as ***Model Building)***.

As a first phase of regression analysis, we should specify the model. If an estimated model is misspecified, it will be biased and inconsistent. The specification consists of the following:

a) *Problem identification* – this is the very first step when the researcher should decide what is the goal of the analysis (i.e. what he wants to do) and select the dependent variable, which describes the problem being studied.

b) *Determine the potential explanatory variables for the regression model* – as discussed above, scatter diagrams (like Fig.6-16) and *coefficients of partial correlation* in the *correlation matrix* can help to identify the most important predictors (i.e. those with the strongest relationship with the dependent variable).

c) *Gather sample data (observations) for all variables* – at this step, a random sample of cross-sectional or time-series data should be selected.

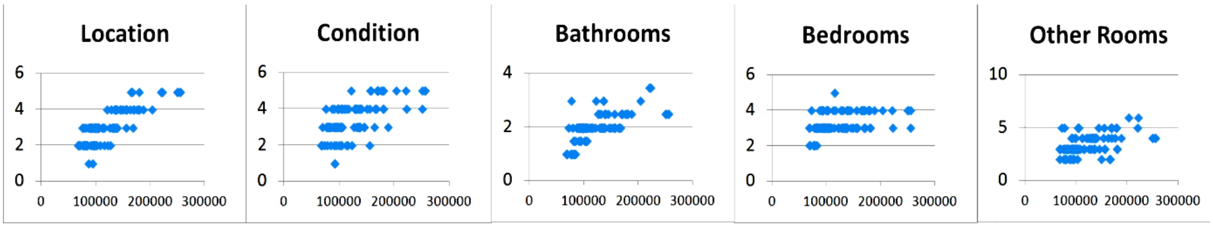


Fig.6-16 Scatterplots in the expanded *House Price* example

Sometimes, mostly in theory-driven approaches, selecting an appropriate functional form for the model is considered as a part of this early step of the model building process. Since we are using the general linear type of models (for reasons already discussed above) we can apply the stepwise regression (recall Chapter 3) later at the step of model estimation.

Here, at this phase of the model building process, one of the most important questions to clarify is the *necessary number of independent measurements* and the *number of potential explanatory variables* in the model. At first glance, this point looks like the so-called ***degrees of freedom*** (***d.f.***), which number (***N-k-1***) should always be larger than zero, but as we are going to see below it is different and has at least the same importance.

Consider a regression model which has two predictors ($\boldsymbol{x_1}$ and $\boldsymbol{x_2}$) and three unknown parameters ($\boldsymbol{\beta_0}$, $\boldsymbol{\beta_1}$, and $\boldsymbol{\beta_{02}}$). Suppose there are 10 observations for dependent variable $\boldsymbol{y_i}$ (i = 1, 2, ..., 10), but all of them at exactly the same value of explanatory variable vector **X**, i.e. there is only one level of measurements for the regressors $\boldsymbol{x_{ij}}$ ***(i=1; j=1,2)***. In this case, regression analysis fails to give a unique set of estimated values for the three unknown parameters, because there is not enough information – the number of normal equations in ***LS*** is smaller than the unknowns. Similarly, measuring at two different levels (values) of X would give enough data for a regression with two unknowns, but not for three or more unknowns. If we have observations for $\boldsymbol{y_i}$ at three different levels of the explanatory variable vector **X (**i.e. the number of normal equations in LS equals the number of unknown parameters to be estimated), then regression analysis would provide a unique set of estimates for the three unknown parameters ($\boldsymbol{\beta_0}$, $\boldsymbol{\beta_1}$, and $\boldsymbol{\beta_{02}}$).

This problem, known as ***overfitting*** (as introduced in Chapter 2) can also arise because of the ***multicollinearity.*** As mentioned above, when correlation exists between two explanatory variables, these two variables contribute redundant information to the multiple regression model. One principal danger of this data redundancy is that of ***overfitting*** in regression analysis. The best regression models are those in which the predictor variables each correlate highly with the dependent variable but correlate at most only minimally with each other. Such a model is often called "***low noise***" and will be statistically robust (that is, it will predict reliably across numerous samples of variable sets drawn from the same statistical population).

Overfitting generally occurs when a model is excessively complex, such as having too many parameters relative to their levels of observations. A model that has been overfitted will generally have poor predictive performance, as it can exaggerate minor fluctuations in the data. The possibility of overfitting exists because the criterion used for estimating the model is not the same as the criterion used to judge the efficacy of a model. In particular, a model is typically estimated by maximizing its performance (i.e. minimizing the sum of squared errors in (6-5)) on some set of training (in-sample) data. However, its efficacy is determined not by its performance on the training data, but by its ability to perform well on unseen, out-of-sample testing data.

As a simple example, consider a database of retail purchases that includes the item bought, the purchaser, and the date and time of purchase. It's easy to construct a model that will fit the training set perfectly by using the date and time of purchase to predict the other attributes, but this model will not generalize to new data at all, because those past times will never occur again.

In order to avoid ***overfitting***, it is necessary to use additional techniques (e.g. cross-validation or regularization), that can indicate when further training is not resulting in better generalization. The basis of those techniques is either (1) to explicitly penalize overly complex models, or (2) to test the model's ability to generalize by evaluating its performance on a set of data not used for training, which is assumed to approximate the typical unseen data that a model will encounter.

In cross-validation, a simple chart (see Fig.6-17) of the model error (**ε**) can detect the ***overfitting***. If the validation error increases (positive slope) while the training error steadily decreases (negative slope), then usually a situation of ***overfitting*** has occurred. The best predictive and fitted model would be where the validation error has its global minimum.

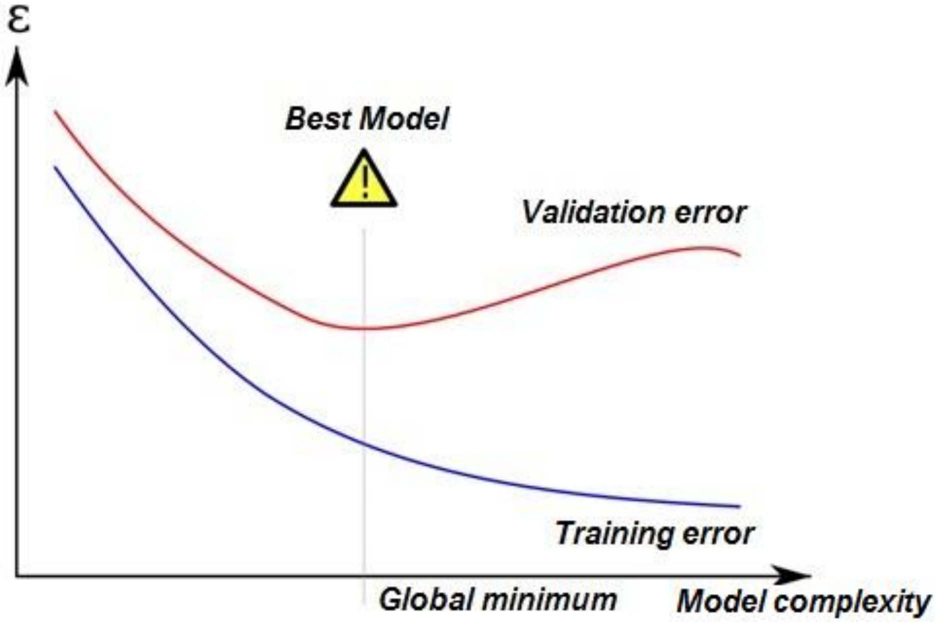


Fig.6-16 Example of a model error chart in cross-validation approach

D. Regression model estimation and interpretation

Recall the *House Price* example – in the end, a recommendation was made that we should look for some improvements in the regression model by means of adding more explanatory variables. To expand the model a few more regressors were considered, such as *Location* of the house, *Condition* of the house, number of *Bathrooms*, number of *Bedrooms* and number of *Other rooms*, i.e. the original predictor *Square feet* was transformed into three new variables. To describe all of them (including the dependent variable House *Price* ($1,000s) a sample of 100 observations was drawn.

The regressors *Location* and *Condition* of the house require some more comments. Both of them are not "pure" numerical variables, but rather qualitative ones. Such type of variables are known as ***Dummy Variables*** (Suits, 1957), or indicator variables, categorical variables, binary variables, qualitative variables, and so forth, and can assume two or more levels (yes or no, on or off, codes as 0 or 1, and others).

Dummy variables are "proxy" variables or numeric stand-ins for qualitative facts in a regression model. In regression analysis, the dependent variables may be influenced not only by quantitative variables (income, output, prices, etc.), but also by qualitative variables (gender, religion, geographic region, etc.). A dummy explanatory variable, which for some observation has a value of 0 will cause that variable's coefficient to have no role in influencing the dependent variable, while when the dummy takes on a value "1" or more its coefficient acts to alter the intercept. For example (see Fig.6-17), suppose Gender is one of the qualitative variables relevant to a regression. Then, female and male would be the categories included under the Gender variable. If a female is arbitrarily assigned the value of 1, then male would get the value 0. Thus the intercept is a constant term for males but would be the constant term plus (or minus) the coefficient of the gender dummy in the case of females.

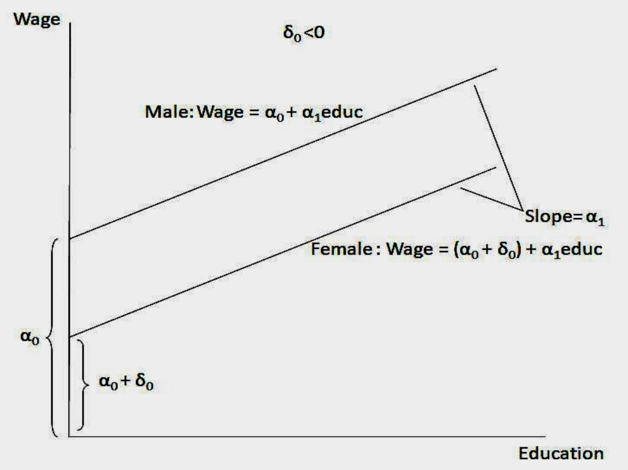


Fig.6-17 Example of a regression line for Wage with dummy variable Female

The interpretation of each of the coefficients associated with the dummy variables is that it is a measure of the effect of that category relative to the omitted category. In the above example, the coefficient associated with a Female will measure the effect of Female compared to Male on the forecast variable Wage.

If there is an ***outlier in the data***, rather than omit it, we can use a dummy variable to remove its effect. In this case, the dummy variable takes value one for that observation and zero everywhere else. Another useful application is for *Public holidays*. For daily data, the effect of public holidays can be accounted for by including a dummy variable predictor taking value one on public holidays and zero elsewhere.

Easter is different from most holidays because it is not held on the same date each year and the effect can last for several days. In this case, a dummy variable can be used with a value of one where any part of the holiday falls in the particular time period and zero otherwise. For example, with monthly data, when Easter falls in March then the dummy variable takes value 1 in March, when it falls in April, the dummy variable takes value 1 in April, and when it starts in March and finishes in April, the dummy variable takes value 1 for both months.

A precaution needs to be taken while using dummy variables for calculating the regression coefficients. For example, only six dummy variables are needed to code seven categories (for instance days of the week). That is because the seventh category (in this case Sunday) is specified when the dummy variables are all set to zero. If we try to add a seventh dummy variable for the seventh category, there will be too many parameters to estimate. The general rule is the number of dummy variables must be the number of categories minus one. So for quarterly data, we need three dummy variables, for monthly data – 11, and for daily data – six dummy variables.

Technically, it means that the constant terms in all the LS normal equations will obviously have a coefficient of 1 (since they are independent of all the variable terms). When the regression is expressed as a matrix equation (6-18), the columns of the coefficient matrix will be linearly dependent. In fact, if a vector-of-ones variable is present, this would result in perfect multicollinearity, so that the matrix inversion in the ***LS*** estimation algorithm would be impossible. This is referred to as the ***dummy variable trap*** (Suits, 1957). The solution is to drop one term from the equation for each set of dummy variables representing a categorical variable, because the column rank of the coefficient matrix is reduced by 1 for every categorical variable – don't forget that to apply the LS method the number of normal equations should at least be equal to the number of unknown parameters to be estimated.

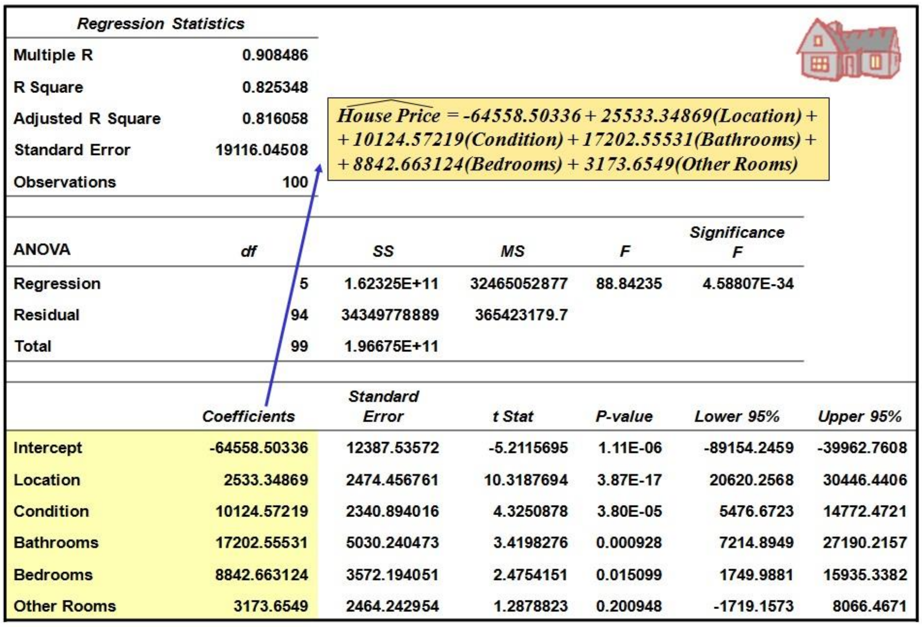
There are a few more special types of useful predictors, like a ***Piecewise Linear Trends*** and ***Splines***, ***Distributed lags***, ***Trading days*** and ***Decision*** and ***Intervention variables***, which are discussed later on in Chapters 8 and 9, because most of them concern time series data or complex forecasting models.

In our example, both regressors *Location* and *Condition* of the house are ***pseudo dummy variables***, since they can assume more than 0 or 1 values. Counting the available levels of measurement for all explanatory variables presented in our regression model, we will find that the column rank of the coefficient matrix (i.e. the number of normal equations in LS) is larger than the number of unknown parameters to be estimated. Technically, it means that the observed variables $(\boldsymbol{x_{ij}},\boldsymbol{y_i})$, $\boldsymbol{i} = \{\boldsymbol{1, 2, \dots\ 100}\}$, $\boldsymbol{j} = \{\boldsymbol{1, 2, \dots\ 5}\}$ provide enough information and we can perform multiple regression analysis, accepting that its assumptions are true.

Fitting the Regression Line

Computer software (such as MS Excel) is generally used to estimate the coefficients $\boldsymbol{\beta_j}$ and compute the measures of goodness of fit for multiple regression model. Fig.6-18 presents the regression output including all five explanatory variables.

A well skilled professional eye will spot from the very first glance that there is something wrong with this model. Without wasting time for further analysis, we can stop at this point since the *Other Rooms* regression coefficient has ***P-value*** of 0.200948, i.e. there is about 20% risk to reject the *Null hypothesis* that this coefficient is zero ($\boldsymbol{H_0}$: $\boldsymbol{\beta_5 = 0}$). This means that coefficient $\mathbf{b_5}$ is not significant (see the test of significance in Fig.6-12 and its example in Fig.6-13) and any further analysis is meaningless.



Regression Statistics	
Multiple R	0.908486
R Square	0.825348
Adjusted R Square	0.816058
Standard Error	19116.04508
Observations	100

$$\widehat{House\ Price} = -64558.50336 + 25533.34869(Location) + \\ + 10124.57219(Condition) + 17202.55531(Bathrooms) + \\ + 8842.663124(Bedrooms) + 3173.6549(Other\ Rooms)$$

ANOVA	df	SS	MS	F	Significance F
Regression	5	1.62325E+11	32465052877	88.84235	4.58807E-34
Residual	94	34349778889	365423179.7		
Total	99	1.96675E+11			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-64558.50336	12387.53572	-5.2115695	1.11E-06	-89154.2459	-39962.7608
Location	2533.34869	2474.456761	10.3187694	3.87E-17	20620.2568	30446.4406
Condition	10124.57219	2340.894016	4.3250878	3.80E-05	5476.6723	14772.4721
Bathrooms	17202.55531	5030.240473	3.4198276	0.000928	7214.8949	27190.2157
Bedrooms	8842.663124	3572.194051	2.4754151	0.015099	1749.9881	15935.3382
Other Rooms	3173.6549	2464.242954	1.2878823	0.200948	-1719.1573	8066.4671

Fig.6-18 MS Excel Regression output for the expanded *House Price* example

If we take into account the partial correlation coefficients from Table 6.2, we should expect similar results, because both regressors *Bedrooms* and *Other Rooms* have weak association with dependent variable *Price*. However, the relations between all variables in multiple regression models are formed in a much more complex way, than the simple correlation indicates (i.e. when the effect of other predictors has not been accounted for). After performing the regression analysis, sometimes we can find in the final output that a week partial correlation results in significant regression coefficient, like the case with the predictor variable *Bedrooms*.

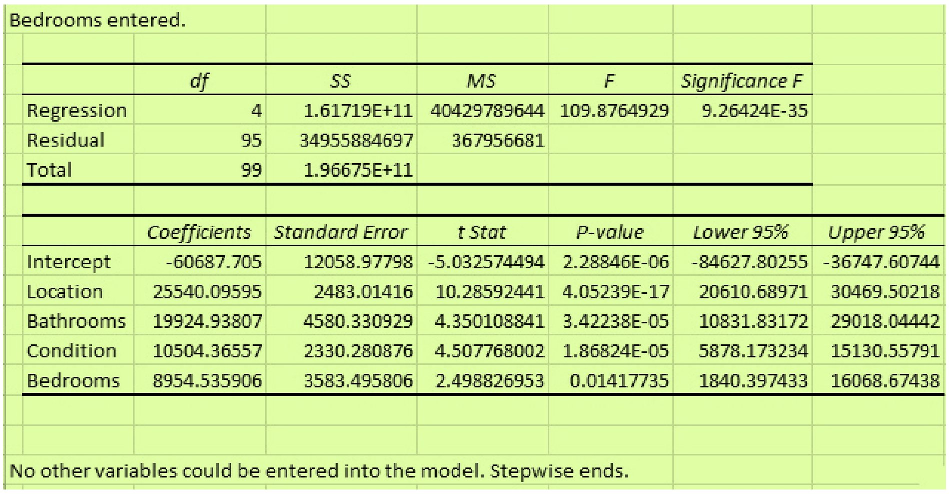
Though some authors (Hyndman & Athanasopoulos, 2013, Chapter 4.3 Selecting predictors) do not recommend them, the above two approaches (starting with the full list of explanatory variables and using the correlation matrix), are the most common techniques for two reasons. *First*, because these are the usual steps in multiple regression analysis (and every statistical software provide options for them), and *second*, because a good expert can identify many properties of the full-list predictors model (using multiple and multivariate tests, as in our example below) and modify the model specification accordingly. This, in most cases, will result in a better model, than recommended ***Stepwise Regression*** will produce, as we are going to show out here, in expanded *House Price* example.

To support nonqualified users, who cannot comprehend the rules on how to build and select the right model specification, experienced researchers designed special algorithms and computer software that perform the modeling process as a highly automated procedure, according to data patterns or particular properties of the regression model.

Where possible, all potential regression models can be fitted and the best one selected based on one of the measures discussed in Model selection section 4.3. This is known as "***best subsets***" regression or "***all possible subsets***" regression. Unfortunately, if there are a large number of predictors, it is not possible to fit all possible models. For example, 40 predictors leads to $\mathbf{2^{40}>1}$ trillion possible models! Consequently, as most authors agree[21], a strategy is required to limit the number of models to be explored.

The classical one, as mentioned in Chapter 3 is the ***Stepwise Regression***. The basic version of MS Excel does not support this option, but there are some Add-ins (most of them come for free), like PHStat, that have an expanded list of statistical techniques. These Add-ins provide advanced statistical procedures, including ***Multiple Sample Tests***, ***Best subsets regression***, ***Stepwise Regression*** and others. What is even more attractive is that these Add-ins do not even need installation, since they come as ready to use Excel applications (i.e. ***xla*** files).

[21] See section Criticism: https://en.wikipedia.org/wiki/Stepwise_regression



Bedrooms entered.

	df	SS	MS	F	Significance F
Regression	4	1.61719E+11	40429789644	109.8764929	9.26424E-35
Residual	95	34955884697	367956681		
Total	99	1.96675E+11			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-60687.705	12058.97798	-5.032574494	2.28846E-06	-84627.80255	-36747.60744
Location	25540.09595	2483.01416	10.28592441	4.05239E-17	20610.68971	30469.50218
Bathrooms	19924.93807	4580.330929	4.350108841	3.42238E-05	10831.83172	29018.04442
Condition	10504.36557	2330.280876	4.507768002	1.86824E-05	5878.173234	15130.55791
Bedrooms	8954.535906	3583.495806	2.498826953	0.01417735	1840.397433	16068.67438

No other variables could be entered into the model. Stepwise ends.

Fig.6-19 ***Stepwise Regression*** output for the expanded *House Price* example

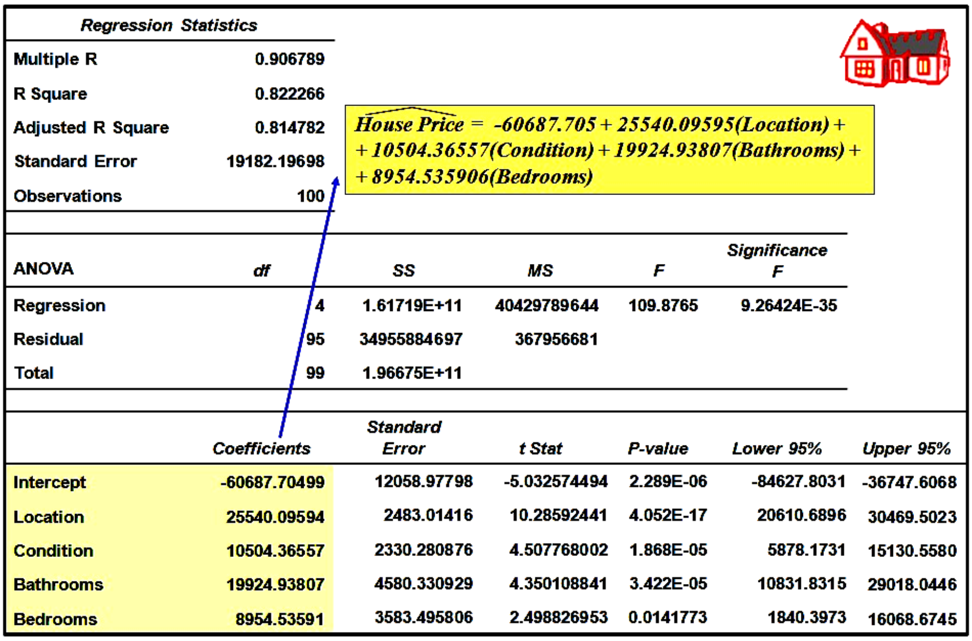
Applying the ***Stepwise Regression*** from MS Excel Add-ins PHStat, we can perform Stepwise Analysis of the regression model, based on selected observations for dependent variable Price and all regressors. In our example, the automated procedure enters *Location* variable at the first step, followed by *Bathrooms, Condition,* and *Bedrooms*. Notice, that the procedure does not follow the sequence in the explanatory variable list, but enters the regressors according to series of multivariate tests based on their P-values and partial R-square. Fig.6-19 presents the end of the ***Stepwise Regression*** in our expanded *House Price* example.

As we can see, the specified/generated by this procedure model omitted the regressor *Other rooms*, as we suggested earlier, and this is only one small remark regarding the ***Stepwise Regression*** qualities and much more important will follow. If we accept the so-specified model as the "best" model, we can perform further regression analysis, which returns the output presented in Fig.6-20.

Before using the model for any purpose (inference or prediction) we should evaluate the measures of its goodnes of fit, presented in the regression output (Fig.6-20) and compare some of them with the first model, containing all five explanatory variables (Fig.6-18). The Multiple coefficient of correlation R has very close values in both models, +.9085 in the full model and +.9068 in the four regressors version.

The coefficient of determination $\boldsymbol{R^2}$ (***R-squared***), which determine how much of the variation in dependent variable *House prices* is explained by the regression model, also changes insignificantly – less than one percent, from .8253 in the full model to .8223 in the new model (i.e. 0.3%). Here, it is worth noting a few very important properties of ***R-squared.***

R-squared never decreases when a new explanatory variable is added to the model. The intuitive reason that using an additional regressor cannot lower the $\boldsymbol{R^2}$ is as follows:



Regression Statistics	
Multiple R	0.906789
R Square	0.822266
Adjusted R Square	0.814782
Standard Error	19182.19698
Observations	100

$$\widehat{House\ Price} = -60687.705 + 25540.09595(Location) + 10504.36557(Condition) + 19924.93807(Bathrooms) + 8954.535906(Bedrooms)$$

ANOVA	df	SS	MS	F	Significance F
Regression	4	1.61719E+11	40429789644	109.8765	9.26424E-35
Residual	95	34955884697	367956681		
Total	99	1.96675E+11			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-60687.70499	12058.97798	-5.032574494	2.289E-06	-84627.8031	-36747.6068
Location	25540.09594	2483.01416	10.28592441	4.052E-17	20610.6896	30469.5023
Condition	10504.36557	2330.280876	4.507768002	1.868E-05	5878.1731	15130.5580
Bathrooms	19924.93807	4580.330929	4.350108841	3.422E-05	10831.8315	29018.0446
Bedrooms	8954.53591	3583.495806	2.498826953	0.0141773	1840.3973	16068.6745

Fig.6-20 MS Excel Multiple Regression output for the expanded *House Price* example excluding variable *Other Rooms*

Minimizing ***Sum of Squared Errors*** (6-5) is equivalent to maximizing $\boldsymbol{R^2}$ (6-15)[22]. When the extra explanatory variable is included, the data always have the option of giving it an estimated coefficient of zero, leaving the predicted values and the $\boldsymbol{R^2}$ unchanged. The only possible output, when the optimization (i.e. SSE minimization) problem will give a non-zero regression coefficient, is if doing so improves the $\boldsymbol{R^2}$.

We can summarize it, that $\boldsymbol{R^2}$ does not indicate whether:

- the explanatory variables are a cause of the changes in the dependent variable;
- omitted-variable bias exists;
- the correct regression was used;
- the most appropriate set of independent variables has been chosen;
- there is collinearity present in the data on the explanatory variables;
- the model can be improved by using transformed versions of the existing set of regressors;
- there are enough data points to make a solid conclusion.

When comparing models explaining the same dependent variable with a different number of explanatory variables, the main question to ask is "What is the net effect of adding a new regressor?" To answer this question we need to take into account that we lose a degree of freedom when a new regressor is added, which leads to another question "Did the new predictor

[22] Explain this using only the logic without calculations – hint: use equation (6-13).

variable add enough explanatory power to offset the loss of one degree of freedom?" Consequently, ***R^2*** is not good enough and we need a measure that shows the proportion of variation in dependent variable explained by all predictor variables and adjusted for the number of predictors used.

Adjusted R^2 (often written as $\bar{R}^2$ and pronounced "***R bar squared***") is an attempt to take account of the phenomenon of the ***R^2*** automatically and spuriously increasing when extra explanatory variables are added to the model. It is a modification due to Henri Theil (1961) of ***R^2*** that adjusts for the number of explanatory variables in a model relative to the number of data points and it is defined as:

$$\boxed{R_A^2 = 1-(1-R^2)\left(\frac{N-1}{N-k-1}\right)} \tag{6-23}$$

where N = sample size and k = number of explanatory variables.

The ***adjusted R^2*** can be negative and its value will always be less than or equal to that of ***R^2***. Unlike ***R^2***, the ***adjusted R^2*** increases when a new regressor is included only if the new regressor improves the ***R^2*** more than it would be expected by chance. If a set of explanatory variables with a predetermined hierarchy of importance are introduced into a regression one at a time, with the ***adjusted R^2*** computed each time, the level at which ***adjusted R^2*** reaches a maximum, and decreases afterward, would be the regression with the ideal combination of having the best fit without excess/unnecessary regressors.

The adjusted R^2 does not have the same interpretation as ***R^2*** – while ***R^2*** is a measure of fit, the ***adjusted R^2*** is instead a comparative measure of the suitability of alternative nested sets of regressors. ***Adjusted R^2*** penalizes excessive use of unimportant explanatory variables and is particularly useful in the model specification (variable selection) stage of model building.

In our example, the ***adjusted R^2*** changes from 81.48% in four regressors model to 81.61% in all regressors model. Apparently, adding the Other rooms predictor does not contribute too much and, in addition, its estimated parameter **b**$_5$ fails the test for significance (as explained earlier).

Other useful comparissons are the ***F****-Test for Overall Significance of the Model* and the *Tests of Individual Regressor's Significance.* The *Overall Significance test* shows if there is a linear relationship between all of the explanatory variables ***x_{ij}*** (***i***={***1, 2, …*** *N*}, ***j***={***1, 2, … k***}) considered together, and dependent variable ***y_i*.** It uses F[23] test statistic:

[23] The name was coined by George W. Snedecor, in honor of Sir Ronald A. Fisher. See http://en.wikipedia.org/wiki/F-test

$$\boxed{F = \frac{Explained.Variance}{UnExplained.Variance}}$$

or

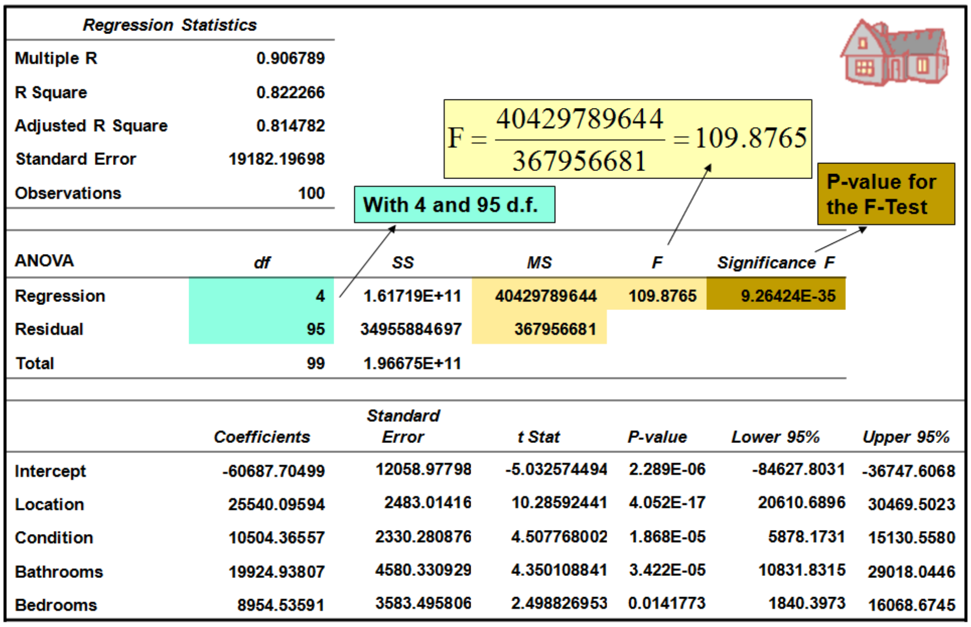
$$\boxed{F = \frac{\frac{SSR}{k}}{\frac{SSE}{N-k-1}}} \tag{6-24}$$

And the hypotheses tested are:

- H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ (no linear relationship)
- H_A: at least one $\beta_i \neq 0$ (at least one regressor x_{ij} affects dependent variable y_i)

Fig.6-21 shows the F-test results from the ANOVA Table in the MS Excel Multiple Regression output for the expanded *House Price* example excluding variable *Other Rooms*. The same test for the full model is available in Fig.6-18 and the corresponding values are F = 88.84235 and P-value of 4.58807E-34. In both cases, the Decision is to reject $\boldsymbol{H_0}$ in favor of $\boldsymbol{H_A}$, which means that the regression model does explain a significant portion of the variation in House Price variable (Fig.6-22).

When *Overall Significance of the Model test* is positive, i.e. at least one regressor $\boldsymbol{x_{ij}}$ affects the dependent variable $\boldsymbol{y_i}$, a *Tests of Individual Regressor's Significance* (such as the $\boldsymbol{t}$ test for Inference about the Slope in Fig.6-12) should follow.



Regression Statistics	
Multiple R	0.906789
R Square	0.822266
Adjusted R Square	0.814782
Standard Error	19182.19698
Observations	100

$$F = \frac{40429789644}{367956681} = 109.8765$$

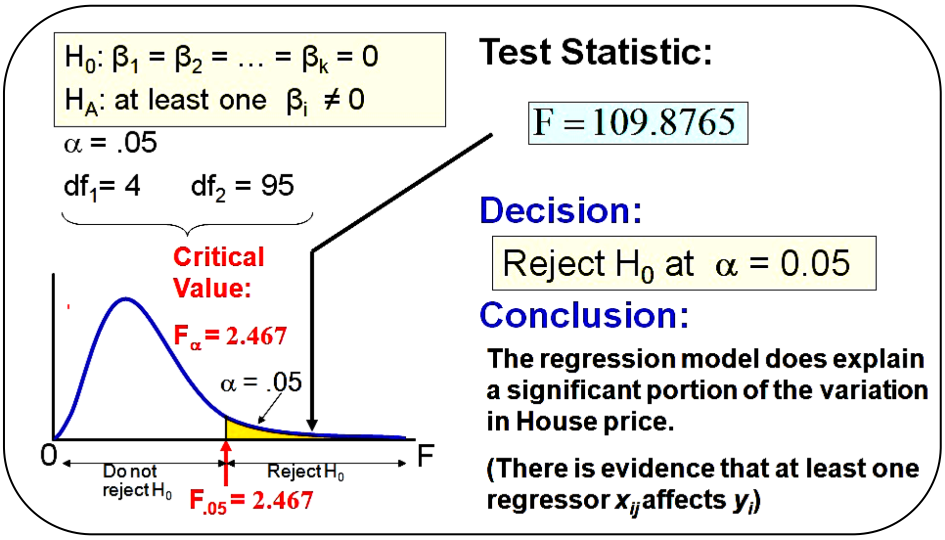
With 4 and 95 d.f.

P-value for the F-Test

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	4	1.61719E+11	40429789644	109.8765	9.26424E-35
Residual	95	34955884697	367956681		
Total	99	1.96675E+11			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-60687.70499	12058.97798	-5.032574494	2.289E-06	-84627.8031	-36747.6068
Location	25540.09594	2483.01416	10.28592441	4.052E-17	20610.6896	30469.5023
Condition	10504.36557	2330.280876	4.507768002	1.868E-05	5878.1731	15130.5580
Bathrooms	19924.93807	4580.330929	4.350108841	3.422E-05	10831.8315	29018.0446
Bedrooms	8954.53591	3583.495806	2.498826953	0.0141773	1840.3973	16068.6745

Fig.6-21 MS Excel Multiple Regression output emphasizing the ***F***-Test for the expanded *House Price* example excluding predictor *Other Rooms*



H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$

H_A: at least one $\beta_i \neq 0$

$\alpha = .05$

$df_1 = 4$ $df_2 = 95$

Critical Value:

$F_\alpha = 2.467$

Test Statistic:

$F = 109.8765$

Decision:

Reject H_0 at $\alpha = 0.05$

Conclusion:

The regression model does explain a significant portion of the variation in House price.

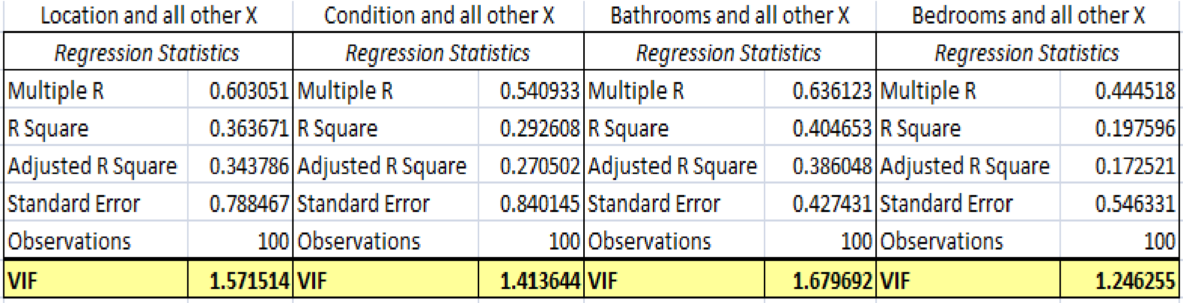
(There is evidence that at least one regressor x_{ij} affects y_i)

Fig.6-22 Inferences about the *Overall Significance of the Model* – ***F***-Test example

The test of individual regressors is used to determine which explanatory variables have nonzero regression coefficients (***b**$_j$*). It is a ***t*** test of individual regressor slopes, which shows if there is a linear relationship between the particular predictor ***x**$_{ij}$* and dependent variable ***y**$_i$*. Each ***t*** test and its ***P-value*** can be found in Fig.6-18 and Fig.6-20. As already mentioned above, all regressors but *Other Rooms* pass the *Test of Individual Regressor's Significance.*

To test if there is collinearity between regressors, the ***Variance Inflationary Factor (VIF)*** (6-22) was used. All ***VIF**$_j$* < **5**, as presented in Fig.6-23, which means that neither one of regressors **x**$_j$ is highly correlated with the other explanatory variables. Such result should be expected since the largest ***partial correlation coefficient*** within the predictor's section in the correlation matrix equals .5575 (as shown in Table 6.2). This is a moderate relationship and is between regressors *Bathrooms* and *Other Rooms*, which is omitted from the current model.

It looks like the four regressors model is the "best" one for the expanded *House Price* example. Unfortunately, neither test, used so far in our example, provides the optimal value of the test criterion. In addition, regression analysis performed on the observations available for the expanded *House Price* example is based on the claim that all OLS assumptions are true. As any other characteristics of the model, these assumptions should be tested as well.



Location and all other X		Condition and all other X		Bathrooms and all other X		Bedrooms and all other X	
Regression Statistics		*Regression Statistics*		*Regression Statistics*		*Regression Statistics*	
Multiple R	0.603051	Multiple R	0.540933	Multiple R	0.636123	Multiple R	0.444518
R Square	0.363671	R Square	0.292608	R Square	0.404653	R Square	0.197596
Adjusted R Square	0.343786	Adjusted R Square	0.270502	Adjusted R Square	0.386048	Adjusted R Square	0.172521
Standard Error	0.788467	Standard Error	0.840145	Standard Error	0.427431	Standard Error	0.546331
Observations	100	Observations	100	Observations	100	Observations	100
VIF	**1.571514**	**VIF**	**1.413644**	**VIF**	**1.679692**	**VIF**	**1.246255**

Fig.6-23 Tests about ***multicollinearity*** in four regressors model – ***VIF*** example

Aptness of the Regression Model Diagnostic (Residual Analysis)

Regression diagnostic could be one of a set of procedures that seek to assess the validity of the regression model in any of a number of different ways. This assessment may be an exploration of the model's underlying statistical assumptions, an examination of the structure of the model by considering formulations that have fewer, more or different explanatory variables, or a study of subgroups of observations, looking for those that are either poorly represented by the model (outliers) or that have a relatively large effect on the regression model's predictions.

A regression diagnostic may take the form of a graphical result, informal quantitative results or a formal statistical hypothesis test, each of which provides guidance for further stages of a regression analysis or additional tests.

A basic, though not quantitatively precise, way to check for problems that render a model inadequate is to conduct a visual examination of the residuals (i.e. model errors ε) to look for obvious deviations from randomness. If a visual examination suggests, for example, the possible presence of heteroskedasticity, then additional statistical tests can be performed to confirm or reject this hunch. If it is confirmed, the model needs modifications accordingly.

Different types of plots of the residuals from a fitted model provide information on the adequacy of different aspects of the model:

- sufficiency of the functional part of the model (i.e. examine for linearity assumption) – scatter plots of residuals versus predictors (see Fig.6-5);
- non-constant variation (i.e. examine for heteroskedasticity) – scatter plots of residuals versus predictors (see Fig.6-6 a); for data collected over time, also plots of residuals against time;
- independence of errors (i.e. examine relationship between the error terms and the regressors) – scatter plots of residuals versus predictors (see Fig.6-6 b) or in case of time-series data autocorrelation plot (see Chapter 8);
- normality of errors – histogram and normal probability plot, or box-plot.

Graphical methods have an advantage over numerical methods for model validation because they readily illustrate a broad range of complex aspects of the relationship between the model and the data. Fig.6-24 presents scatter plots of residuals versus predictors in the four regressors model and Fig.6-25 shows the symmetry (normality) of the residuals distribution for the expanded *House Price* example using Box-plot and a Histogram. Unfortunately, these charts are not perfect and we can see some discrepancies in both of them.

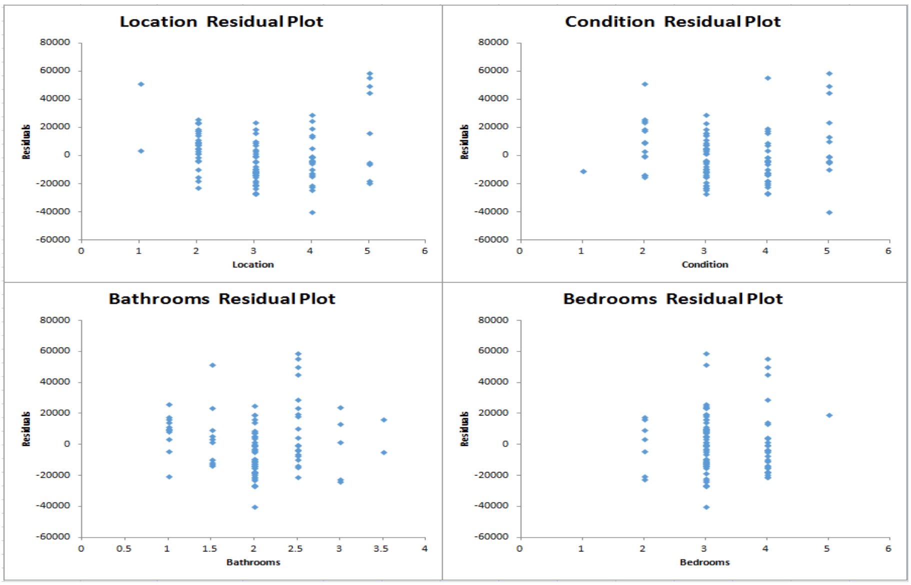


Fig.6-24 Scatter plots of residuals versus predictors in the four regressors model

The first warning is in Fig.6-24 and it comes from the plot of residuals versus predictor *Location*. It is clear that there is a nonlinear patern, which looks like second degree polynomial. This finding certainly needs more attention.

The second warning is in the charts in Fig 6-25, which represent the symmetry (normality) of the residuals' distribution for the expanded *House Price* example. Both of them, the Box-plot and the Histogram, show a positive skewness in distribution, maybe due to some extreme positive values – outliers in residuals. Although some authors (Hyndman & Athanasopoulos, 2013, Chapter 5.4 Residual diagnostics) claim that "this is not essential for forecasting", the lack of normality indicates possible improvements in the forecasting model, which apparently is not the "best" one. In combination with the nonlinear pattern in the plot of residuals versus predictor *Location*, there is only one meaningful conclusion – to modify and improve the regression model.

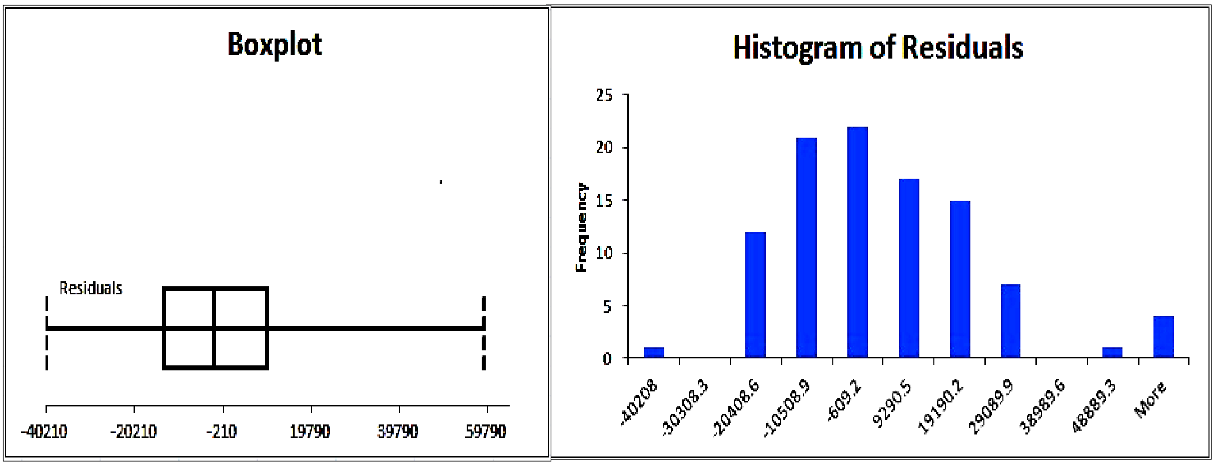
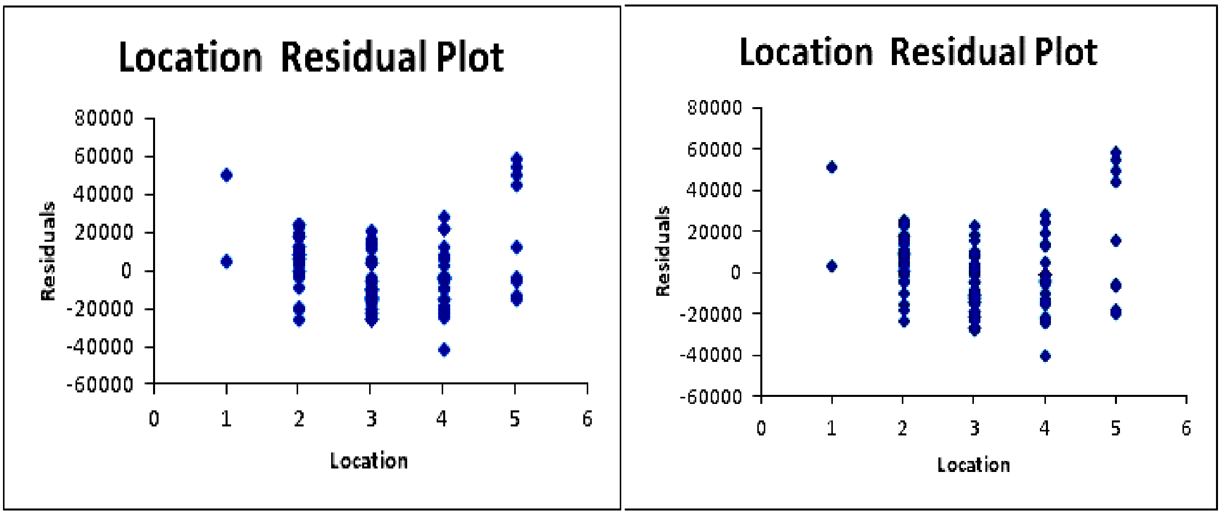


Fig.6-25 Box-plot and Histogram of the residuals in the four regressors model



a) Full-list regressors model b) Four regressors model

Fig.6-26 Comparisson between scatter plots of residuals versus *Location*

If we return to the the full-list predictors model, we will find the same nonlinear patern in the plot of residuals versus predictor *Location* (see Fig.6-26). Despite the fact that coefficient $\mathbf{b_5}$ is not significant and any further analysis is meaningless, an expert (or just a sceptic who wants to test everything)[24] would rather check all details, before leave out the regression output. As we have already mentioned, an expert can identify many properties of the full-list predictors model (using multiple and multivariate tests) and modify the model specification accordingly.

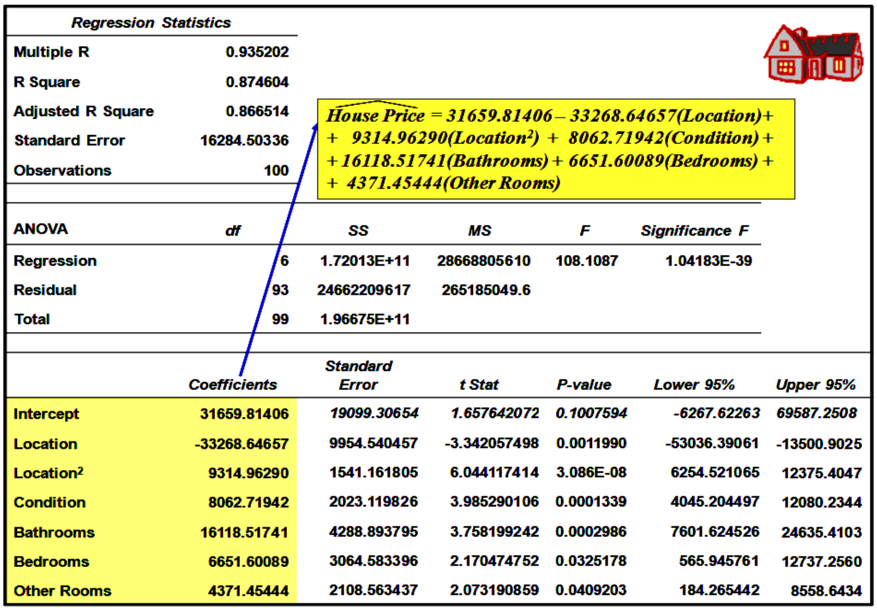
The easiest and fastest way to modify the regression model in this case is to transform the regressor *Location* from linear factor to second degree polynomial, that is to change the component $\boldsymbol{\beta_1 x_{i1}}$ in equation (6-20) to $\boldsymbol{\beta_1 x_{i1}}{+}\ \boldsymbol{\gamma_1 x^2{}_{i1}}$ and after that to update (if necessary) all following components of the equation accordingly:

$$\boldsymbol{y_i = \beta_0 + \beta_1 x_{i1} + \gamma_1 x^2{}_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i} \qquad (6\text{-}25)$$

Performing Multiple Regression Analysis on equation (6-25) with the new, updated list of explanatory variables and the same set of observations, including a new data column for *Location*^{2}, returns the output presented in Fig.6-27.

At the very first glance it is clear that this model is much better, than both models discussed so far. All measures of its goodness of fit show improvements – for instance, the Multiple coefficient of correlation ***R*** increased from +.9085 in the full model and +.9068 in the four regressors version to 0.9352 in the new model.

[24] Even the Bible says "Prove all things; hold fast that which is good. " (1 Thessalonians 5:21, KJV)



Regression Statistics	
Multiple R	0.935202
R Square	0.874604
Adjusted R Square	0.866514
Standard Error	16284.50336
Observations	100

$$\widehat{House\ Price} = 31659.81406 - 33268.64657(Location) + 9314.96290(Location^2) + 8062.71942(Condition) + 16118.51741(Bathrooms) + 6651.60089(Bedrooms) + 4371.45444(Other\ Rooms)$$

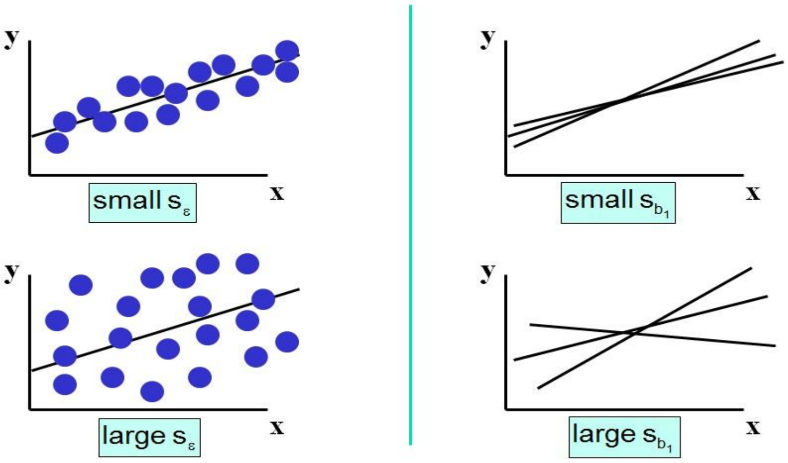
ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	6	1.72013E+11	28668805610	108.1087	1.04183E-39
Residual	93	24662209617	265185049.6		
Total	99	1.96675E+11			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	31659.81406	*19099.30654*	*1.657642072*	*0.1007594*	*-6267.62263*	*69587.2508*
Location	-33268.64657	9954.540457	-3.342057498	0.0011990	-53036.39061	-13500.9025
Location²	9314.96290	1541.161805	6.044117414	3.086E-08	6254.521065	12375.4047
Condition	8062.71942	2023.119826	3.985290106	0.0001339	4045.204497	12080.2344
Bathrooms	16118.51741	4288.893795	3.758199242	0.0002986	7601.624526	24635.4103
Bedrooms	6651.60089	3064.583396	2.170474752	0.0325178	565.945761	12737.2560
Other Rooms	4371.45444	2108.563437	2.073190859	0.0409203	184.265442	8558.6434

Fig.6-27 MS Excel Multiple Regression output for the expanded *House Price* example including second degree polynomial for regressor *Location*

The coefficient of determination $\boldsymbol{R^2}$ also changes significantly – about five percents, from .8253 in the full model and .8223 in the four regressors, to 0.8746 in the new model. With the knowledge about $\boldsymbol{R^2}$ properties, this was expected (since we increased the number of predictors), but the ***adjusted*** $\boldsymbol{R^2}$ also changes more than five percents – from 81.48% in four regressors model and 81.61% in all regressors to 86.65% in the new model (i.e. 5.17% and 5.04% more). Apparently, adding the *Location²* predictor does contribute significantly to the *House Price* model improvement.

The Standard Error of the estimate also shows a better value. In general, we cannot compare standard errors in different models, but it is possible when the models are estimated from the same data set and contain regressors from the same subset of variables, as shown in Fig.6-28.



a) Variation of observed y_i values from the regression line b) Variation in the slope $\mathbf{b_i}$

Fig.6-28 Comparing Standard Errors in Regression Analysis

In our example the ***standard error of the regression*** $\boldsymbol{S_\varepsilon}$ decreases by approximately $3,000 – from $19,182 in the four regressors model and $19,116 in all regressors one, to $16,285 in the new model.

The ***F*** *Test for Overall Significance of the Model* from ANOVA Table in MS Excel Multiple Regression output (Fig.6-27) again leads to the same Decision – to reject $\boldsymbol{H_0}$ in favor of $\boldsymbol{H_A}$, that is the regression model does explain a significant portion of the variation in the House Price variable. An important note should be made here – rejecting $\boldsymbol{H_0}$ in the *F-Test for Overall Significance of the Model* does not mean that the model is perfect. It means that the regression model explains a significant portion of the variation in dependent variable and at least one regressor $\boldsymbol{x_{ij}}$ affects $\boldsymbol{y_i}$**.** As we can see from our example, there is always a room for further improvements in the model.

What is much more important here are the findings[25] from the *Tests of Individual Regressor's Significance*. ***t*** tests and their P-values provide evidence that all partial regression coefficients $\mathbf{b_j}$ are significant (at $\boldsymbol{\alpha}$ = 0.041, which is just enough above the maximum observed P-value for $\mathbf{b_5}$) even for the regressor *Other Rooms*. At the same time, the intercept *Tests of Significance* failed (P-value = 0.1008). This raises another interesting point for discussion: "Do we need to include or not to include the CONSTANT in the regression equation?"

Most multiple regression models include a constant term (i.e., the intercept), since this ensures that the model will be unbiased. In a simple regression model, the constant represents the Y-intercept of the regression line, in unstandardized form. In a multiple regression model, the constant represents the value that would be predicted for the dependent variable if all the independent variables were simultaneously equal to zero. If we are not particularly interested in what would happen if all the independent variables were simultaneously zero, then we normally leave the constant in the model regardless of its statistical significance. In addition to ensuring that the in-sample errors are unbiased, the presence of the constant allows the regression line to "seek its own level" and provide the best fit to data which may only be locally linear.

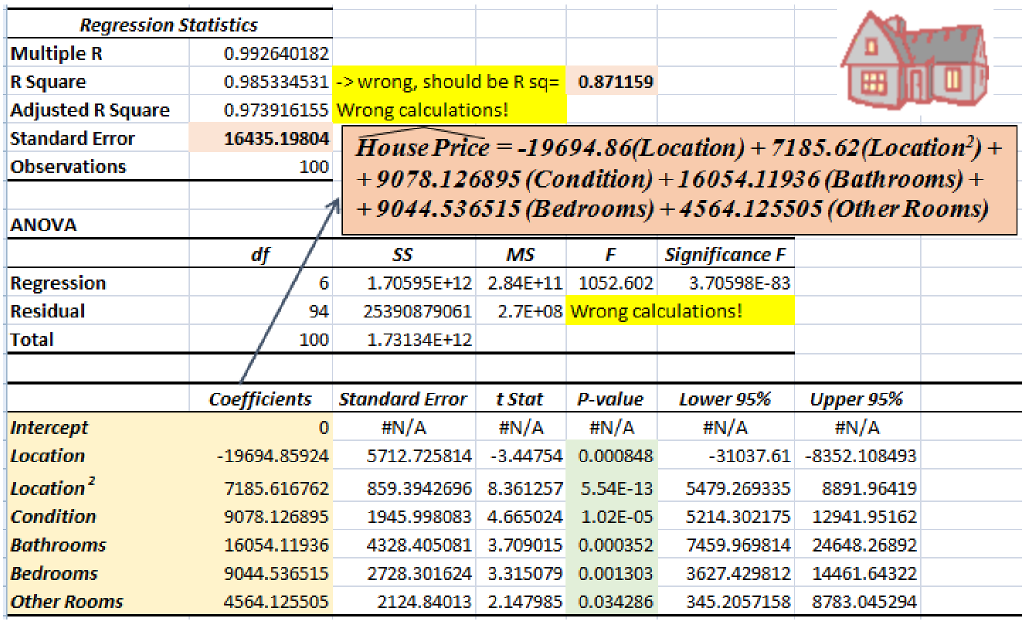
This case has been discussed for many years and it is known as ***Regression Through the Origin (RTO)***, which refers to linear regressions obtained by least-squares methods without a constant term. Joseph Eisenhauer (2003) summarizes the most important points and make very interesting conclusion:

[25] I found the file with this Example long time ago in the shared folders on our campus server and just a few years ago, when I used it in Business Statistics class, I realized how rich and full of opportunities it is. Thank you, my unknown assistant!

Regression through the origin is an important and useful tool in applied statistics, but it remains a subject of pedagogical neglect, controversy, and confusion. ... However, in the light of the unresolved debate, perhaps the strongest conclusion to be drawn from this review is that the practice of statistics remains as much an art as it is a science, and the development of statistical judgment is therefore as important as computational skill. (p. 80)

What is important to know and remember from this study is that ***R-squared*** and the ***F statistic*** do not have the same meaning in an RTO model as they do in an ordinary regression model, and they are not calculated in the same way. Fortunately, changing the value of the constant in the model changes the mean of the errors but does not affect the variance. Also, there is a measure analogous to $\boldsymbol{R^2}$ for the no-intercept model, which is the square of the sample correlation between observed and predicted values. It, therefore, gives an interpretable measure of the quality of an RTO model but does not help in comparing RTO with OLS model. For that purpose, the best measures appear to be the p-value of the OLS constant and the standard errors of the OLS and RTO regressions. Using these measures, together with the business interpretation of the case (i.e. if we are particularly interested in what would happen if all the independent variables were simultaneously zero), we should make the final decision whether the constant should be retained in the model or not.

In our example, after comparing outputs in Fig.6-27 and Fig.6-29 (taking into account the comments above), the conclusion is that there is no significant difference of including the intercept. To answer this question, we need out-of-sample data and cross-validation.



Regression Statistics			
Multiple R	0.992640182		
R Square	0.985334531	-> wrong, should be R sq=	0.871159
Adjusted R Square	0.973916155	Wrong calculations!	
Standard Error	16435.19804		
Observations	100		

$$\widehat{House\ Price} = -19694.86(Location) + 7185.62(Location^2) + + 9078.126895\,(Condition) + 16054.11936\,(Bathrooms) + + 9044.536515\,(Bedrooms) + 4564.125505\,(Other\,Rooms)$$

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	6	1.70595E+12	2.84E+11	1052.602	3.70598E-83
Residual	94	25390879061	2.7E+08	Wrong calculations!	
Total	100	1.73134E+12			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A
Location	-19694.85924	5712.725814	-3.44754	0.000848	-31037.61	-8352.108493
Location ^{2}	7185.616762	859.3942696	8.361257	5.54E-13	5479.269335	8891.96419
Condition	9078.126895	1945.998083	4.665024	1.02E-05	5214.302175	12941.95162
Bathrooms	16054.11936	4328.405081	3.709015	0.000352	7459.969814	24648.26892
Bedrooms	9044.536515	2728.301624	3.315079	0.001303	3627.429812	14461.64322
Other Rooms	4564.125505	2124.84013	2.147985	0.034286	345.2057158	8783.045294

Fig.6-29 MS Excel Multiple Regression output including second degree polynomial for regressor *Location* without *intercept*

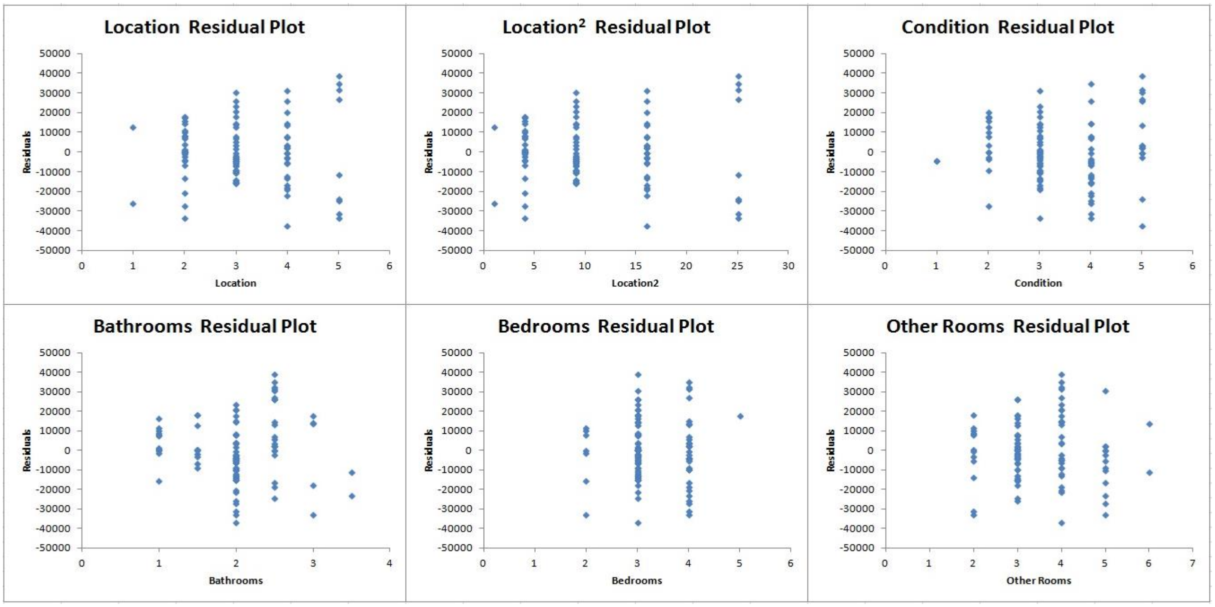


Fig.6-30 Scatter plots of residuals versus predictors in regression model including second degree polynomial for predictor *Location*

Residual Analysis of the regression model including second degree polynomial for predictor *Location* (presented in Fig.6-30 and Fig.6-31) does not reveal any violations of the ***OLS*** assumptions. On the contrary, there is no more non-linear pattern in the plot of residuals versus predictor *Location* (and anywhere else) and neither the Box-plot, nor the Histogram indicates any skewness in errors distribution.

To test if there is collinearity between regressors in the new model, including a second-degree polynomial for predictor *Location* (or simply *the six regressors model*), the ***VIF*** (6-22) was used. As expected, the ***VIF*** values for *Location* and *Location^2* are "> **5**", as presented in Fig.6-32, because these two regressors are highly correlated. In fact, *Location^2* is a linear combination of *Location* explanatory variable, but this is not a problem, as far as there are no errors in ***OLS*** computations and model tests and statistics do not show any discrepancies.

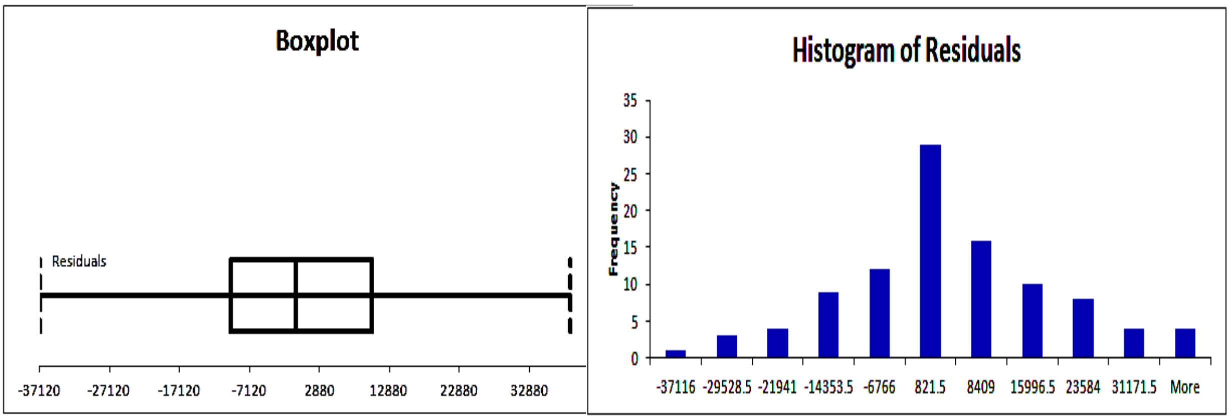
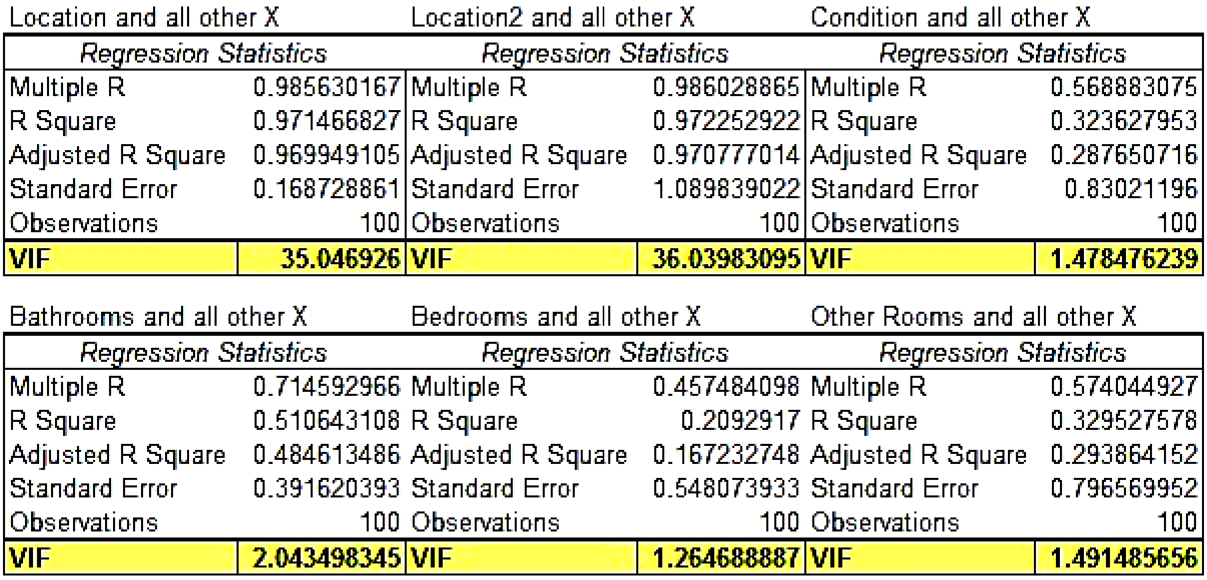


Fig.6-31 Box-plot and a Histogram of the residuals in regression model including second degree polynomial for predictor *Location*



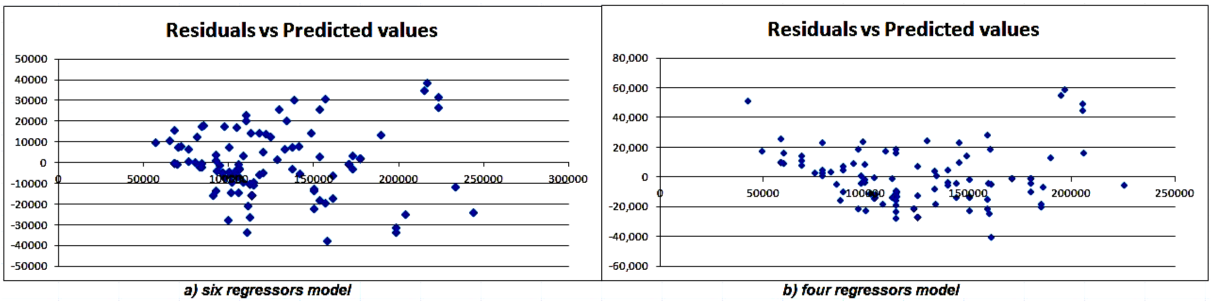
Location and all other X		Location2 and all other X		Condition and all other X	
Regression Statistics		*Regression Statistics*		*Regression Statistics*	
Multiple R	0.985630167	Multiple R	0.986028865	Multiple R	0.568883075
R Square	0.971466827	R Square	0.972252922	R Square	0.323627953
Adjusted R Square	0.969949105	Adjusted R Square	0.970777014	Adjusted R Square	0.287650716
Standard Error	0.168728861	Standard Error	1.089839022	Standard Error	0.83021196
Observations	100	Observations	100	Observations	100
VIF	**35.046926**	**VIF**	**36.03983095**	**VIF**	**1.478476239**

Bathrooms and all other X		Bedrooms and all other X		Other Rooms and all other X	
Regression Statistics		*Regression Statistics*		*Regression Statistics*	
Multiple R	0.714592966	Multiple R	0.457484098	Multiple R	0.574044927
R Square	0.510643108	R Square	0.2092917	R Square	0.329527578
Adjusted R Square	0.484613486	Adjusted R Square	0.167232748	Adjusted R Square	0.293864152
Standard Error	0.391620393	Standard Error	0.548073933	Standard Error	0.796569952
Observations	100	Observations	100	Observations	100
VIF	**2.043498345**	**VIF**	**1.264688887**	**VIF**	**1.491485656**

Fig.6-32 ***VIF*** tests about ***multicollinearity*** in the *six regressors model*

A pattern in the plot of the residuals against the fitted values indicates that there may be ***heteroscedasticity*** in the errors, i.e. the variance of the residuals may not be constant. As we can see in Fig.6-33 the plot from *the six regressors model* looks random enough and show no pattern. The *four regressors model* (selected by the ***Stepwise Regression***) is definitely not random and indicates possible ***heteroscedasticity.*** To overcome this problem, a transformation of the forecast variable (such as a logarithm or square root) or changing the model specification may be required – as it is in our expanded *House Price* example.

Finally, to test the assumption for independence of errors, a ***correlogram*** (in time series also known as an ***autocorrelation plot***) or a statistical test is required. In our Example, a ***Durbin-Watson*** test was performed using MS Excel Add-ins PHStat[26]. Despite the fact, that usually autocorrelation is presented mainly in time-series data, if the test shows non-random correlation in residuals, it may indicate real problems in the regression model.

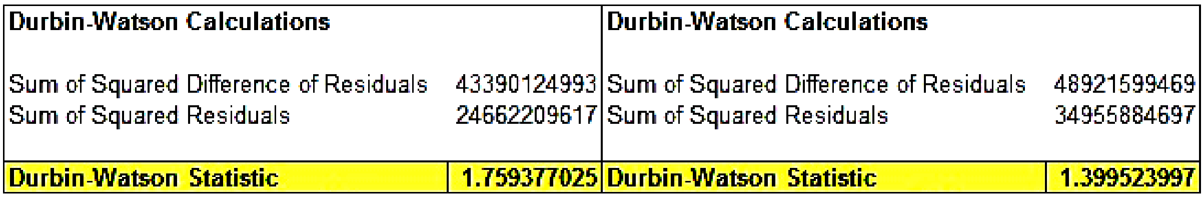


a) six regressors model

b) four regressors model

Fig.6-33 Scatter plots of residuals versus predicted values – test against ***heteroscedasticity***

[26] For advanced users we would recommend a professional statistical software, like Gretl, which is free, open-source software, and was developed "from statisticians for statisticians" – http://gretl.sourceforge.net/index.html



Durbin-Watson Calculations		Durbin-Watson Calculations	
Sum of Squared Difference of Residuals	43390124993	Sum of Squared Difference of Residuals	48921599469
Sum of Squared Residuals	24662209617	Sum of Squared Residuals	34955884697
Durbin-Watson Statistic	1.759377025	Durbin-Watson Statistic	1.399523997

a) six regressors model *b) four regressors model*

Fig.6-33 ***Durbin-Watson*** tests for ***autocorrelation*** in residuals

The value of the ***Durbin–Watson*** statistic ***d*** always lies between 0 and 4. If ***d*** is substantially less than 2, there is evidence of positive serial correlation. As a rough rule of thumb, if ***d*** is less than 1.0, there may be cause for alarm. Small values of ***d*** indicate that successive error terms are, on average, close in value to one another, or positively correlated[27]. If ***d* > 2**, on average, successive error terms are much different in value from one another, i.e., negatively correlated. In regressions, this can imply an underestimation of the level of statistical significance.

To test for positive autocorrelation at a given significance level (***α***), the test statistic ***d*** is compared to lower and upper critical values (*dL* and *dU*):

- If d < dL (given α), there is statistical evidence that the error terms are positively autocorrelated.
- If d > dU (given α), there is no statistical evidence that the error terms are positively autocorrelated.
- If dL < d < dU (given α), the test is inconclusive.

The critical values, *dL* and *dU*, vary by the level of significance (***α***), the number of observations (***N***), and the number of predictors (***k***) in the regression equation. Their derivation is complex and therefore users typically obtain them from Tables in appendices of statistical texts or online[28].

In our example, the critical values for ***Durbin–Watson*** test at 1% ***α*** (significance level), given sample size N=100 and number of regressors k +1 (for the intercept), are as follow:

N	k	dL	dU
100.	*4.*	*1.48241*	*1.60370*
100.	*5.*	*1.46203*	*1.62527*
100.	*6.*	*1.44142*	*1.64735*
100.	*7.*	*1.42061*	*1.66994*

[27] Positive serial correlation is a correlation in which a positive error for one observation increases the chances of a positive error for another observation.

[28] See http://web.stanford.edu/~clint/bench/dw01a.htm

Comparing these values with the computed ***Durbin–Watson*** statistics (***d***), as presented in Fig.6-33, the following conclusions should be done:

- For the *four regressors model* d=1.39952 < dL=1.46203 (at α=1%), i.e. *there is statistical evidence that the error terms are positively correlated*!
- For the *six regressors model* If d=1.75938 > dU=1.66994 (at α=1%), i.e. *there is no statistical evidence that the error terms are positively (or negatively) correlated.*

In summary, according to all computed statistics and tests of significance and goodness of fit, it is clear that the *six regressors model* is more reliable and better than the *four regressors model* selected by the ***Stepwise Regression.*** In addition, the *six regressors model* satisfies all statistical assumptions, while the *four regressors model* does not.

It should be mentioned that there is a lot of criticisms[29] of the ***Stepwise Regression*** procedure. It is pointed out that the results of stepwise regression are often used incorrectly without adjusting them for the occurrence of model selection, especially, the practice of fitting the final selected model as if no model selection had taken place and reporting of estimates and confidence intervals as if the least-squares theory were valid for them. Widespread incorrect usage and the availability of alternatives such as *Ensemble learning* (where "The most common approach used for model-selection is cross-validation selection[30]"), leaving all variables in the model or using expert judgment to identify relevant variables have led to calls to totally avoid stepwise model selection (Flom & Cassell, 2007).

Such criticisms based upon limitations of the relationship between a model and procedure, and data set used to fit it, are usually addressed by verifying the model on an independent data set, as in the cross-validation selection. Data mining methods, such as ***GMDH,*** not only provide a better platform for model building for expert forecasters but also give a great support to nonqualified users, who cannot comprehend the rules on how to build and select the right model specification. Special ***GMDH*** algorithms that perform the modeling process as highly automated procedure, according to data patterns or particular properties of the regression model, will be discussed in Chapters 8, 9 and 12.

In this way, modern data mining tools are no longer restricted to specialists. As more organizations adopt predictive analytics into decision-making processes and integrate it into their operations, they are creating a shift in the market toward business users as the primary consumers of the information. Business users want tools they can use on their own and vendors

[29] See http://en.wikipedia.org/wiki/Stepwise_regression

[30] See http://en.wikipedia.org/wiki/Ensemble_learning

are responding by creating software that removes the mathematical complexity, provides user-friendly graphic interfaces, and/or builds in shortcuts that can, for example, recognize the kind of data available and suggest an appropriate predictive model. Predictive analytics tools have become sophisticated enough to adequately present and dissect data problems, so that any data-savvy information worker can utilize them to analyze data and retrieve meaningful, useful results. For example, modern tools like KnowledgeMiner software (Mueller & Lemke, 2003) present results using simple charts, graphs, and scores that indicate the likelihood and/or the level of possible outcomes (Fig.6-33).

Interpreting Partial Regression Coefficients and Using the Model

Interpreting the intercept b_0 = 31659.81. In mathematics, this is the estimated average value of $\boldsymbol{y_i}$ when all $\boldsymbol{x_{ij}}$ = 0. To have a business interpretation, we should assume that all $\boldsymbol{x_{ij}}$=0 are within the range of observed $\boldsymbol{x_j}$ values. In general, it is reasonable for a house to have a zero price when all regressors (location, condition, number of bedrooms and so forth.) have zero values. Though plausible it is not of a great interest to look for a house with zero bedrooms for instance. In our expanded *House Price example* the only useful interpretation of the intercept **b_0 = 31659.81** is that it indicates, for houses within the range of regressors observed, that $**31659.81** is *the portion of the house price not explained by predictor variables in the model*.

Interpreting the slope b_5 = 6651.60. This slope tells us that the House price will increase, on average, by $**6651.60** for each additional *Bedroom*, net of the effects of changes due to all other regressors.

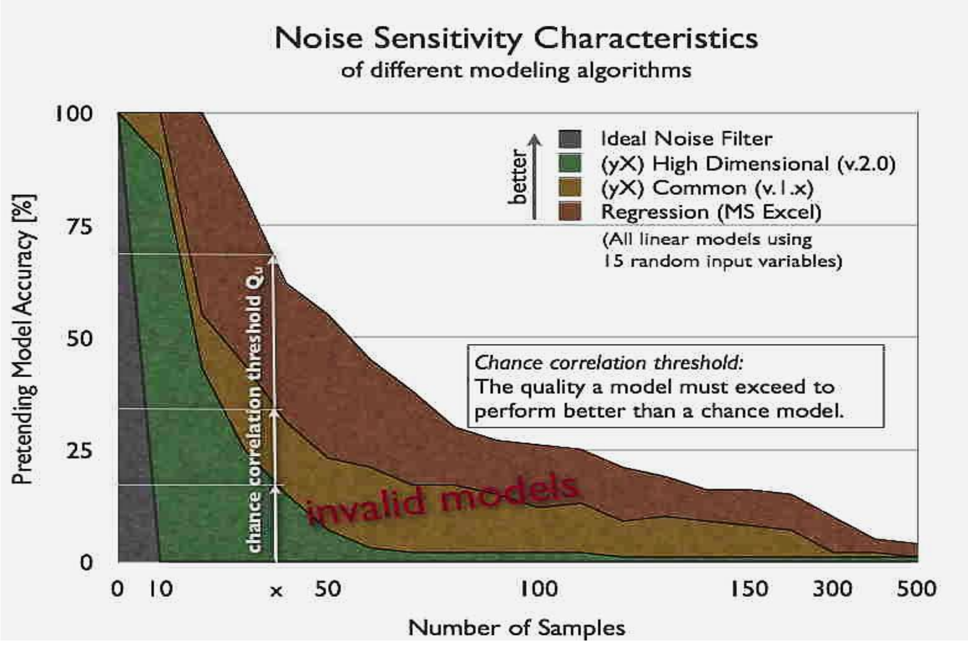
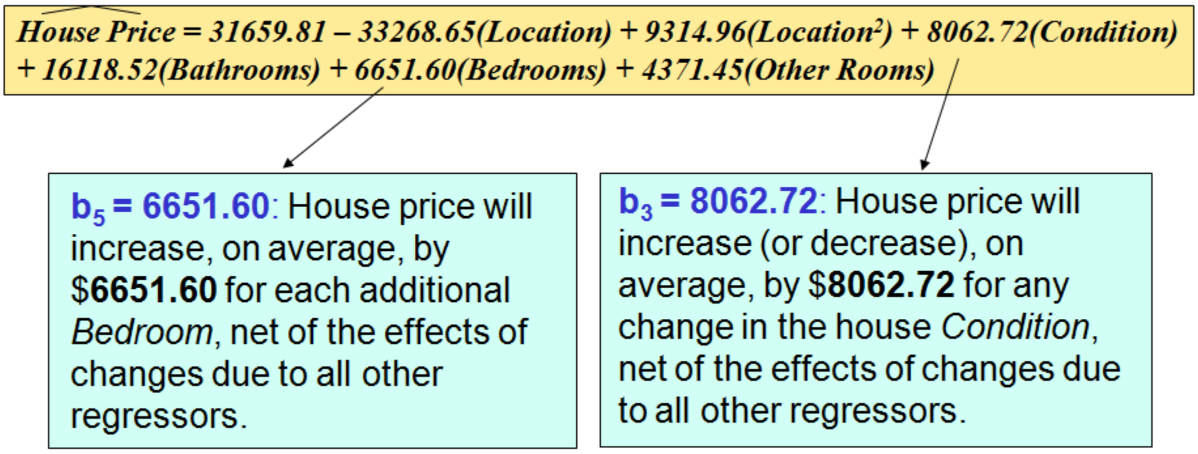


Fig.6-33 Example of *KnowledgeMiner* software model specification output
(Source: http://knowledgeminer.eu/about.html)



House Price = 31659.81 – 33268.65(Location) + 9314.96(Location^2) + 8062.72(Condition) + 16118.52(Bathrooms) + 6651.60(Bedrooms) + 4371.45(Other Rooms)

b_5 = 6651.60: House price will increase, on average, by $**6651.60** for each additional *Bedroom*, net of the effects of changes due to all other regressors.

b_3 = 8062.72: House price will increase (or decrease), on average, by $**8062.72** for any change in the house *Condition*, net of the effects of changes due to all other regressors.

Fig.6-34 Business interpretation of some estimated regression coefficients in the expanded *House Price example* (all data rounded to the nearest cent)

Interpreting the slope b_3 = 8062.72. House price will increase (or decrease), on average, by $8062.72 for any unit change in the house *Condition*, net of the effects of changes due to all other regressors. In a similar way we can interpret all estimated regression coefficients in the expanded *House Price example*, but for predictor *Location*. Because *Location* has two coefficients and its relationship with *House Price* is presented like second-degree polynomial, it is difficult to make clear business interpretation.

The explanation requires to consider polynomial component in equation (6-25) as one element:

$$\beta_1 x_{i1} + \gamma_1 x^2{}_{i1} \qquad (6\text{-}26)$$

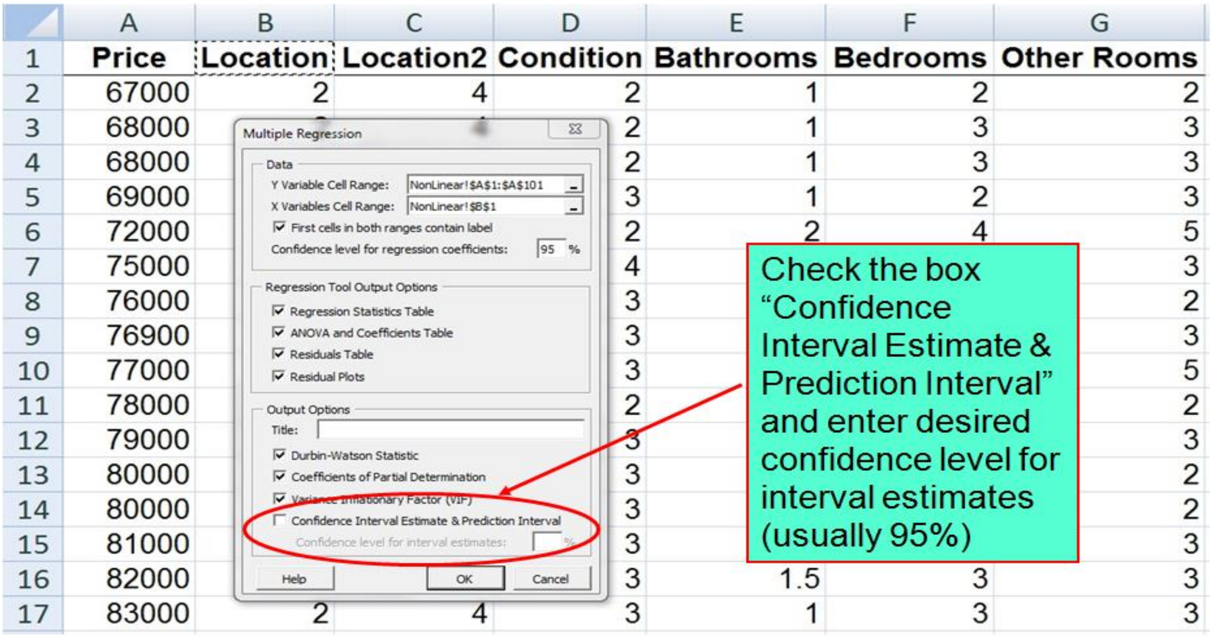
which is exactly a second-degree polynomial with linear and quadratic terms.

Basically, the sign in linear term is positive when the line represents an increase and negative means a decline. The sign in quadratic term is positive when the model is convex and negative when the curve is concave. Thus the quadratic term indicates which way the curve is bending but what the linear term means in the polynomial does not seem to make sense.

If we differentiate (6-26) with respect to x we get:

$$y' = b_1 + 2g_1 x \qquad (6\text{-}27)$$

What this shows is that $\mathbf{b_1}$ gives the rate of change when x is equal to zero. The coefficient $\mathbf{g_1}$ tells both the direction and steepness of the curvature (a positive value indicates the curvature is upwards while a negative value indicates the curvature is downwards). So, if **x=0** is not within the range of our observed values, the trick is to place the zero value within the range of our data. We can do this by centering the x, i.e. subtracting the mean of x from each value. Despite this clarifications, the business interpretation of polynomial terms remains unclear.



	A	B	C	D	E	F	G
1	Price	Location	Location2	Condition	Bathrooms	Bedrooms	Other Rooms
2	67000	2	4	2	1	2	2
3	68000			2	1	3	3
4	68000			2	1	3	3
5	69000			3	1	2	3
6	72000			2	2	4	5
7	75000			4			3
8	76000			3			2
9	76900			3			3
10	77000			3			5
11	78000			2			2
12	79000			3			3
13	80000			3			2
14	80000			3			2
15	81000			3			3
16	82000			3	1.5	3	3
17	83000	2	4	3	1	3	3

Multiple Regression

Data

Y Variable Cell Range: NonLinear!$A$1:$A$101

X Variables Cell Range: NonLinear!$B$1

☑ First cells in both ranges contain label

Confidence level for regression coefficients: 95 %

Regression Tool Output Options

☑ Regression Statistics Table

☑ ANOVA and Coefficients Table

☑ Residuals Table

☑ Residual Plots

Output Options

Title:

☑ Durbin-Watson Statistic

☑ Coefficients of Partial Determination

☑ Variance Inflationary Factor (VIF)

☐ Confidence Interval Estimate & Prediction Interval

Confidence level for interval estimates: %

Check the box "Confidence Interval Estimate & Prediction Interval" and enter desired confidence level for interval estimates (usually 95%)

Fig.6-35 Making predictions in MS Excel with Add-ins PHStat

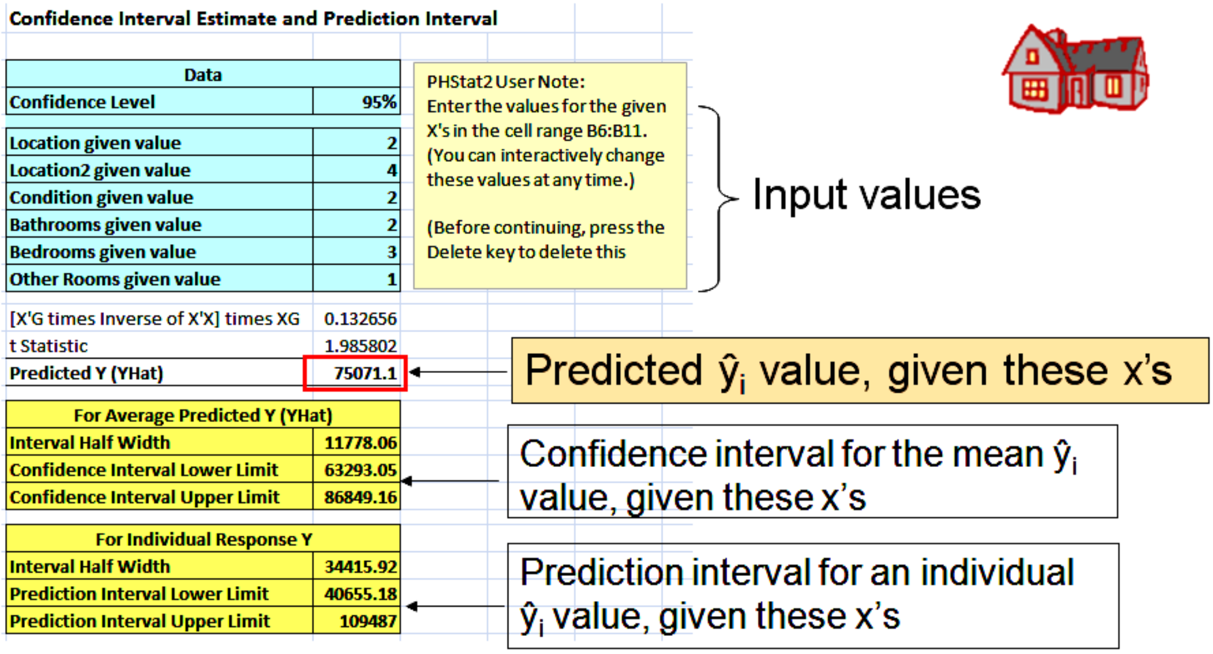
Using the regression model to make predictions simply means to solve the equation (6-25) for dependent variable y_i, given particular values for predictors x_j, after substituting the unknown parameters with their estimates, as presented in Fig.6-34, i.e.:

$$\widehat{\textit{House Price}} = 31659.81 - 33268.65^*(\textit{Location}) + 9314.96^*(\textit{Location}^2) + 8062.72^*(\textit{Condition}) + 16118.52^*(\textit{Bathrooms}) + 6651.60^*(\textit{Bedrooms}) + 4371.45^*(\textit{Other Rooms})$$

Suppose, we want to predict what the expected Price for a house is given *Location*=2, *Condition*=2, *Bathrooms*=2, *Bedrooms*=3 and *Other Rooms*=1. The forecasting equation is:

$$\widehat{\textit{House Price}} = 31659.81 - 33268.65^*(2) + 9314.96^*(4) + 8062.72^*(2) + 16118.52^*(2) + 6651.60^*(3) + 4371.45^*(1) = \$75071.1$$

It easy to program this equation as a formula in MS Excel, but it is better to use Add-Ins to automate this process. For example, in PHStat we just need to check a box in the data entry form for Regression analysis as presented in Fig.6-35. The program returns an output where we can interactively change predictor values at any time and obtain the forecast and its predicted intervals. Notice that there are two intervals, one for the mean House Price value (i.e. for the whole population) and another one for the individual sample, given predictor values (Fig.6-36).



Confidence Interval Estimate and Prediction Interval

Data	
Confidence Level	95%

Location given value	2
Location2 given value	4
Condition given value	2
Bathrooms given value	2
Bedrooms given value	3
Other Rooms given value	1

PHStat2 User Note:
Enter the values for the given X's in the cell range B6:B11.
(You can interactively change these values at any time.)

(Before continuing, press the Delete key to delete this

Input values

[X'G times Inverse of X'X] times XG	0.132656
t Statistic	1.985802
Predicted Y (YHat)	75071.1

Predicted $\hat{y}_i$ value, given these x's

For Average Predicted Y (YHat)	
Interval Half Width	11778.06
Confidence Interval Lower Limit	63293.05
Confidence Interval Upper Limit	86849.16

Confidence interval for the mean $\hat{y}_i$ value, given these x's

For Individual Response Y	
Interval Half Width	34415.92
Prediction Interval Lower Limit	40655.18
Prediction Interval Upper Limit	109487

Prediction interval for an individual $\hat{y}_i$ value, given these x's

Fig.6-36 Interactive output for Predictions in MS Excel with Add-ins PHStat2

It means, that for the whole population of houses in this area the predicted Price for a house with *Location*=2, *Condition*=2, *Bathrooms*=2, *Bedrooms*=3 and *Other Rooms*=1 is $75071.1 and it is expected to vary within the range $63,293.05 – $86,849.16. The forecast based on one particular (individual) sample data is the same ($75071.1), but the interval where it is expected to vary is much larger, from $40,655.18 to $109,487.

Does this appear to be a big interval or not? Is it a reasonably precise estimate of the *House Price* or not? We do not know. We can only make recommendations about the model and its characteristics and eventually, based on the model properties, to recommend if it needs any improvement. All the findings should be provided to an expert in the specific business (real estate in our example) and the final conclusions and decisions are up to his/her professional opinion.

As mentioned earlier in Chapter 6, a commonplace example might be estimation of some variable of interest at some specified future date. When regression analysis is employed to test theories that the current values of one or more independent time series affect the current value of another time series, this type of analysis of time series is usually referred to as “***Regression with Time Series Data***”. Making predictions in such case involve some complications, due to unknown values of expected regressor values and the presence of time-series autocorrelation. This topics will be discussed in Chapters 8 and 9.

***

SUMMARY AND CONCLUSIONS

Regression and ***Correlation analyses*** are widely used for predictions. Chapter 6 discusses techniques that apply the concepts of regression and correlation in business forecasting:

A) There are similarities and differences between ***Association, Correlation*** and ***Dependence,*** but *neither association, nor correlation establishes causality:*

- ***Correlation*** and ***causation*** are connected – where there is causation, there is a likely correlation, and correlation is used when inferring causation, but *correlation does not imply causation*, i.e. correlation cannot be used to infer a causal relationship between the variables.
- ***Correlation Analysis*** is a group of statistical techniques to measure the association between two or more variables. It is only concerned with the strength of the relationship and no causal effect is implied.
- The ***Pearson product-moment correlation coefficient*** is a measure of the linear relationship between two variables and ranges from "+1" (meaning perfect positive relationship) through "0" (which means no association) to "−1" inclusive (that is perfect inverse relationship).
- The ***scatterplot*** (***scatter diagram***) is a chart that portrays the relationship between two variables. It is especially helpful when the number of data is large. The correlation coefficient, as a summary statistic, cannot replace visual examination of the data.

B) Regression methods use assumptions, which could be summarized in the following:

- ***Classical assumptions*** for multiple regression analysis and ***OLS:*** the sample is representative of the population for the inference prediction; the error (**ε**) is a random variable with a mean of zero; the error terms (**ε_i**) are uncorrelated (i.e. no ***auto-correlation***) and have constant variance (***homoscedasticity***); prediction variables are measured with no error and so forth…
- ***Statistical assumptions*** – when the number of observations (**N**), is larger than the number of unknown parameters (**k**), and the errors **ε_i** are normally distributed, then the excess of information contained in (**N−k−1**) measurements is used to make statistical predictions about the unknown model parameters; in such case, parameter estimates of the regression model will be *unbiased, consistent*, and *efficient*.
- Additional assumption, in order to perform the multiple regression, is that the predictors should be linearly independent, i.e. not correlated. When they are related a ***multicollinearity*** exists, and the performance of OLS estimates can be very poor.

C) A large number of procedures have been developed for regression parameters **β** estimation, which differ in computational simplicity of algorithms, presence of a closed-form solution, robustness with respect to heavy-tailed distributions, and theoretical assumptions needed to validate desirable statistical properties such as consistency and efficiency:

- The ***Least Squares (LS)*** method is a standard approach to the approximate solution of overdetermined systems, i.e., sets of equations in which there are more equations than unknowns. "***Least squares***" means that the overall solution minimizes the sum of the squares of the errors ($\boldsymbol{\varepsilon}_i$).
- There is a big variety of techniques, used to *fit the regression line*, which could be summarized in a few major categories of estimators:

  - ***Ordinary Least Squares (OLS)*** is an approach fitting a mathematical or statistical model to data in cases where the idealized value provided by the model for any data point ($\boldsymbol{x_i,y_i}$) is expressed linearly in terms of the unknown parameters $\boldsymbol{\beta_j}$ of the model;

  - ***Generalized least squares (GLS)*** and related techniques are extension of the ***OLS*** method, which allows efficient estimation of $\boldsymbol{\beta}$ when either heteroscedasticity, or correlations, or both are present among the error terms of the model, as long as the form of heteroscedasticity and correlation is known independently of the data;

  - ***Maximum Likelihood Estimation (MLE)*** and related techniques can be performed when the distribution of the error terms is known to belong to a certain parametric family of probability distributions. When it is a normal distribution with zero mean and finite variance, the resulting estimate is identical to the ***OLS*** estimate. ***GLS*** estimates are maximum likelihood estimates when error ($\boldsymbol{\varepsilon}$) follows a multivariate normal distribution with a known covariance matrix. In general, for a fixed set of data and underlying statistical model, the ***MLE*** selects the set of values of the model parameters that maximizes the likelihood function, i.e. the probability to reach their unknown, true values;

  - ***Bayesian linear regression*** applies the framework of Bayesian statistics to linear regression – regression coefficients $\boldsymbol{\beta}$ are assumed to be random variables with a specified prior distribution, which can bias the solutions for the regression coefficients, in a way similar to (but more general than) *ridge regression* or *lasso regression.* In addition, the Bayesian estimation process produces not a single point estimate for the "best" values of the regression coefficients, but an entire posterior distribution, completely describing the uncertainty surrounding the quantity.

- *There are many different methods for regression coefficients estimation. It is very important to understand and remember that all these methods are based on specific assumptions about error terms, explanatory variables and parameters of the model. We need to understand what these assumptions are, and we should be able to explain them and identify when they are true, and we must apply the corresponding method.*

D) There are certain differences between *linear* and *non-linear models,* between *linear* and *non-linear least squares,* which advantages and disadvantages must be considered whenever a selection of a non-linear model should be made or a solution to a non-linear problem is being sought:

- The model function (***f***) in ***Linear Least Squares*** (***LLS***) is a linear combination of the parameters. The model may represent a straight line, a parabola or any other linear combination of functions. In ***Non-Linear Least Squares (NLLS)*** the parameters appear as functions, such as $\boldsymbol{\beta^2}$**,** $\mathbf{e^{\beta x}}$ and so forth. If the derivatives $\partial f/\partial \beta_j$ are either constant or depend only on the values of the regressors, the model is linear in the parameters. Otherwise the model is non-linear.
- Algorithms for finding the solution to a NLLS problem require initial values for the parameters, LLS does not.
- Like LLS, when solving the system of normal linear equations, solution algorithms for NLLS often require that the **Jacobian determinant** be calculated. Analytical expressions for the partial derivatives can be complicated. If analytical expressions are impossible to obtain either the partial derivatives must be calculated by numerical approximation or an estimate must be made of the **Jacobian determinant**.
- In NLLS non-convergence (failure of the algorithm to find a minimum) is a common phenomenon, since data are fitted by a method of successive approximations, whereas the LLS is globally concave so non-convergence is not an issue.
- NLLS is usually an iterative process, which has to be terminated when a convergence criterion is satisfied. LLS solutions can be computed using direct methods.
- In LLS the solution is unique, but in NLLS there may be multiple minima in the sum of squares.
- Under the condition that errors and regressors are uncorrelated, LLS yields unbiased estimates, but even under that condition NLLS estimates are generally biased.
- Last but not least, linear regression coefficients $\boldsymbol{\beta_j}$ have very clear business interpretation – most the time non-linear coefficients have not.

E) To perform a multiple regression analysis, we must provide enough information about the dependent variable ***y_i***. If we assume that the vector of unknown parameters **β** is of length **k**, i.e. there are **k** explanatory variables ***x_{ij}***, then ***N* > *k*** provides enough information in our data to estimate a unique value for ***β*** that best fits the data, and the regression model, when applied to the data can be viewed as an ***over-determined system*** in ***β.***

F) As a first phase of regression analysis, we should specify the model. If an estimated model is misspecified, it will be biased and inconsistent. The ***model specification*** (which should not be confused with the whole process of developing a regression model, that is the ***Model Building)*** consists of the following:

a) *Problem identification* – the researcher should decide what is the goal of the analysis and select the dependent variable, which describes the problem being studied.

b) *Determine the potential explanatory variables for the regression model* – scatter diagrams and *coefficients of partial correlation* can help to identify the most important predictors (i.e. those with the strongest relationship with the dependent variable).

c) *Gather sample data (observations) for all variables* – at this step, a random sample of cross-sectional or time-series data should be selected.

d) Sometimes, mostly in theory-driven approaches, *selecting an appropriate functional form* for the model is considered as a part of this step of the model building process.

G) One of the most important questions to clarify in the model building process, is the *necessary number of independent measurements* and the *number of potential explanatory variables* in the model. If the number of predictors is larger than the number of observations, the modeling task is called an undetermined task, which is also referred to as ***overfitting:***

- ***Overfitting*** generally occurs when a model is excessively complex, such as having too many parameters relative to their levels of observations but can also arise because of the ***multicollinearity*** – when correlation exists between two explanatory variables, these two variables contribute redundant information to the multiple regression model.
- A model that has been ***overfitted*** will generally have poor predictive performance, as it can exaggerate minor fluctuations in the data. The possibility of overfitting exists because the criterion used for estimating the model is not the same as the criterion used to judge the efficacy of a model. A model is typically estimated by maximizing its performance (i.e. minimizing the sum of squared errors) on some set of training (in-sample) data. However, its efficacy is determined not by its performance on the training data, but by its ability to perform well on unseen, out-of-sample testing data.

- To avoid ***overfitting***, it is necessary to use additional techniques (e.g. ***cross-validation*** or ***regularization***), that can indicate when further training is not resulting in better generalization. The basis of those techniques is either (1) to explicitly penalize overly complex models, or (2) to test the model's ability to generalize by evaluating its performance on a set of data not used for training, which is assumed to approximate the typical unseen data that a model will encounter.

H) **Regression model estimation** – this is the stage of the model building process when selecting an appropriate functional form for the model is to be considered as well:

- Where possible, all potential regression models can be fitted and the best one selected based on one of the measures discussed in Model selection. This is known as "***best subsets***" regression or "***all possible subsets***" regression.
- If there is a large number of predictors, it is not possible to fit all possible models and because of this, the typical technique used is the ***Stepwise Regression.***
- Data mining methods, such as ***GMDH,*** not only provide better platform for model building for expert forecasters but also give a great support to nonqualified users, who cannot comprehend the rules on how to build and select the right model specification. Special ***GMDH*** algorithms that perform the *modeling process as highly automated procedure*, according to data patterns or particular properties of the regression model, will be discussed in Chapters 8, 9 and 12.

I) ***Aptness of the regression model (Residual Analysis) –*** regression diagnostic could be one of a set of procedures that seek to assess the validity of the regression model in any of a number of different ways:

- The assessment may be an exploration of the model's underlying statistical assumptions, an examination of the structure of the model by considering formulations that have fewer, more or different explanatory variables, or a study of subgroups of observations, looking for those that are either poorly represented by the model (outliers) or that have a relatively large effect on the regression model's predictions.
- A regression diagnostic may take the form of a graphical result, such as ***Residual*** and ***Autocorrelation plots;*** informal quantitative results or a formal statistical hypothesis test (*F-Test for Overall Significance of the Model* and the *Tests of Individual Regressor's Significance,* ***R-squared*** and the ***Standard error of the regression*** S_{ε}), each of which provides guidance for further stages of a regression analysis or additional tests.

- ***Graphical methods*** often have an advantage over numerical methods for model validation because they readily illustrate a broad range of complex aspects of the relationship between the model and the data.
- The most important point is that the model assessment and the evaluation of its performance should be done on a set of data not used for training. The model validation is for assessing how the results of a statistical analysis will generalize to an independent data set. In ***cross-validation,*** a model is usually given a dataset of known data on which training is run (***training dataset***), and a dataset of unknown data (or first seen data) against which the model is tested (***testing dataset***).

J) The final step in the regression model building is to make recommendations about the model and its characteristics and eventually, based on the model properties, to recommend if it needs any improvement. All the findings should be provided to an expert in the specific business and the final conclusions and decisions are up to his/her professional opinion.

KEY TERMS

CHAPTER EXERCISES

Conceptual Questions:

1. What are the similarities and the differences between ***Association, Correlation and Dependence***? Explain and illustrate with examples.
2. What is the purpose of the ***Scatterplot***? What does the ***Pearson correlation coefficient*** measure? Discuss.
3. What are the similarities and the differences between ***Correlation*** and ***Regression analysis?*** List and briefly discuss the main goals of these techniques.
4. How do we fit the regression line? What are the classical and the statistical assumptions in regression analysis? List and discuss at least three of them.
5. What is a ***multicollinearity?*** How does it affect the regression results? What are the major indications of severe ***multicollinearity?*** How to address this problem? Discuss and illustrate with examples.

Business Applications:

The "*Fresh Food*" Inc. marketing manager wants to predict *what the yearly amount the families of four or more spend on food is.* In his study, he defined the dependent variable ***y*** "***Food***" as cross-sectional dataset $y_i = \{y_1, y_2, \dots y_{12}\}$.

He believes that there are three explanatory variables related to yearly food expenditures: total family income (***Income*** in \$100s), family size (***Size***) and whether the family has a child in college (***College***). These variables are defined as x_{ij} ($x_j = \{x_1, x_2, \dots x_{12}\}$), (***j=1, 2, 3***).

Note that the variable college is a dummy variable. It can take only one of two possible outcomes, that is, a child is a college student (value of ***"1"***) or not (value of ***"0"***).

To prove his claim and compute a forecast, the manager decided to use correlation and regression analysis. A sample of 12 families was selected and observed data were recorded accordingly (file *Food.xlsx*).

Open MS Excel file *Food.xlsx* and:

- Use MS Excel menu Data/Data Analysis and perform Correlation and Regression analysis (all steps are given in detail within the file).
- Comment and explain the findings.

Write a short report (up to two pages) explaining your answers.

INTEGRATIVE CASE

HEALTHY FOOD SUPPLY CHAIN & STORES

Part 6: Building Multiple Regression Model – How to begin?

In Chapter 1 we introduced *Healthy Food Stores* – a fast-growing retail food provider with 12 stores in a northwestern state. The company executives decided to study the effect that company advertising dollars have on sales. They hoped that examining collected historical data would reveal relationships that would help determine future advertising expenditures and predict monthly sales volumes for the upcoming quarter.

After identifying basic parameters, along with input (independent) and output (dependent) variables of the forecasting scenario in Part 2, the research team determined the main elements of the forecasting process:

- Forecasting horizon of up to twelve months;
- Quarterly forecast updates, since accuracy decreases as time horizon increases, and sufficient time is needed to implement possible changes;
- Development of different models based on data patterns, if any, and evaluation of their accuracy in order to select the most appropriate one;
- Selection of the best forecasting model with no more than 5% forecasting error.

The benchmark forecast computed in Part 3 (using the baseline of ***one-step naive forecast*** as a reference forecast) provided basic values for the most common measures of accuracy MFE, MAD, MAPE, MPE, MSE and CV(RMSE).

In Part 4, information about the opinion of some important people from the Healthy Food Stores Company, concerning this specific case, was collected and the research team applied the Delphi method to top executives group, Sales-force composite to the sales managers from all 12 stores and Scenario writing to the most experienced professionals from Advertising Department. After collecting such valuable information from different sources, in Part 5 the research team made its first steps in Numerical Predictions by developing different basic forecasting models, which could be used to expand the base-line of one-step naïve forecast as reference forecasts. Using their computational skills in MS Excel and basic knowledge in Business Statistics, they created spreadsheets for automated calculations and model development with Naïve techniques (Average model, Random Walk With Drift and Seasonal Naïve Technique), simple Moving Average, Simple Exponential Smoothing (SES) and Triple (Holt-Winters) Exponential Smoothing (TES)

Accumulating good experience during the previous parts, the research team continued with more confidence in the forecasting process. Next step would be to analyze the relationships between dependent variable sales and the predictors in the Healthy Food Stores. The goal was to build a multiple regression model representing the real system close enough and to make some diagnostic tests and experiments to determine its aptness.

Case Questions

1. Open the file Data.xslx in MS Excel and run correlation and regression analysis:
   a) Compute the correlation matrix. Which single predictor variable has the strongest correlation with the dependent variable "Sales"? Is there any pair of regressors with strong correlation that could be a potential cause for multicollinearity?
   b) Use the results from the regression output to state the multiple regression equation. Interpret the meaning of the partial regression coefficients.
   c) What percent of the variation in variable "Sales" is explained by the model?
   d) Conduct a global test (***F*** test) of hypothesis (use the ANOVA Table) to determine whether any of the regression coefficients are not zero (use the 0.05 significance level). Is there significant evidence of a linear relationship?
   e) Conduct a ***t*** test of hypothesis on each of the predictor variables. Would you consider eliminating any regressor as insignificant (use the 0.05 significance level)?
   f) Rerun the multiple regression equation with these variables (if any) eliminated. How much has the explained variation (Adj. $\boldsymbol{R^2}$) changed from the previous model?
   g) Conduct again a global test of hypothesis and test of hypothesis on each of the independent variables (use the 0.05 significance level). Is there evidence of significance in these relationships?
   h) Perform residual analysis with the residual plots provided in the regression output. Do you see any violations of the regression assumptions?
   i) Set up a Box plot of the errors. Does the normality assumption appear reasonable?
   j) Set up a plot of the predicted values and the residuals and analyze it. Do you see any violations of the assumptions?
2. Comment and analyze the model accuracy:
   - How ”good” is the accuracy of the forecast (for the training dataset only)?
   - Can we apply the cross-validation approach in this case? If yes, explain how.
3. What overall recommendations would you make to the research team? Explain why.
4. Write a report on the questions above, discussing all important findings and draw relevant conclusions about this part of the Integrative Case.

References

Box, G., & Tiao, G. (1992). *Bayesian Inference in Statistical Analysis*. Wiley.

Carroll, R. J. (1982). Adapting for Heteroscedasticity in Linear Models. *The Annals of Statistics, 10*(4), 1224-1233.

Chatterjee, S., Hadi, A. S., & Price, B. (2000). *Regression Analysis by Example* (3rd ed.). John Wiley and Sons.

del Pino, G. (1989). The Unifying Role of Iterative Generalized Least Squares in Statistical Algorithms. *Statistical Science, 4*(4), 394-403.

Eisenhauer, J. G. (2003). Regression through the Origin. *Teaching Statistics, 25*(3), 76-80.

Flom, P. L., & Cassell, D. L. (2007). Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use. *NESUG 2007*.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain and Ireland*, *15*, 246–263.

Gauss, C. F. (1809). *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientum*.

Golub, G., Heath, M., & Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics, 21*, 215–223.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55-67.

Hyndman, R., & Athanasopoulos, G. (2013). *Forecasting: principles and practice*. OTexts.

Kendall, M. G. (1955). *Rank Correlation Methods*. Charles Griffin & Co.

Legendre, A. (1805). *Nouvelles méthodes pour la détermination des orbites des comètes*. [New Methods for the Determination of the Orbits of Comets] (in French), Paris: Firmin Didot ("Sur la Méthode des moindres quarrés" [The least-squares method] appears as an appendix).

Madala, H. R., & Ivakhnenko, A. G. (1994). *Inductive Learning Algorithms for Complex Systems Modelling*. Boca Raton, FL: CRC Press Inc.

Mueller J. A., & Lemke, F. (2003). *Self-Organizing Data Mining: An Intelligent Approach To Extract Knowledge From Data*. Victoria, BC: Trafford Publishing.

Nau, R. (2014). *Notes on linear regression analysis*. Duke University, Durham, North Carolina. Retrieved from: http://people.duke.edu/~rnau/regintro.htm#top

Nievergelt, Y. (1994). Total Least Squares: State-of-the-Art Regression in Numerical Analysis. *SIAM Review, 36* (2), 258-264.

Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, *58*, 240–242.

Stone, C. J. (1975). Adaptive maximum likelihood estimators of a location parameter. *The Annals of Statistics, 3*(2), 267-284.

Suits, D. B. (1957). Use of Dummy Variables in Regression Equations. *Journal of the American Statistical Association, 52*(280), 548-551.

Theil, H. (1961). *Economic forecasts and policy.* Amsterdam, North-Holland Pub. Co.

Tofallis, C. (2009). Least Squares Percentage Regression. *Journal of Modern Applied Statistical Methods*, *7*, 526–534.

Tufte, E. R. (2003). *The Cognitive Style of PowerPoint*. Cheshire, Connecticut: Graphics Press.

Walker, H. M. (1940). Degrees of Freedom. *Journal of Educational Psychology, 31*(4), 253-269.

Yule, G. U., & Kendall, M. G. (1950). *An Introduction to the Theory of Statistics*. 14th Edition (5th Impr. 1968, pp. 258–270.), Charles Griffin & Co.

***

CHAPTER 7. TIME SERIES ANALYSIS AND FORECASTING – PART 1

7.1. Time Series Decomposition

A ***Time series*** is a sequence of data values, measured typically at successive points in time spaced at uniform time intervals as introduced in Chapter 5. Typical examples of time series are the monthly sales of a company, or quarterly cost of production, i.e. the time intervals can be annual, quarterly, monthly, weekly, daily, and so forth. Time series values are usually presented in Tables like Table 7.1 and plotted via line charts (see Fig.7-1), and have a broad application – in fact, time series are virtually everywhere.

A few basic techniques of ***Time series analysis*** and ***Forecasting*** were introduced and explored in Chapter 5**.** Here, more advanced predictive methods will be presented and discussed, emphasizing on some important models and their application in contemporary business analysis and management.

Time series analysis (***TSA***) comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. There are many methods for ***TSA*** (Box et al., 2016) usually divided into two classes: *frequency-domain methods* (spectral analysis) and *time-domain methods* (auto-correlation analysis). Additionally, ***TSA*** techniques may be divided into *parametric* and *non-parametric*[1], *linear* and *non-linear* and so on.

Table 7.1 Current US Inflation Rates – 1999-2014[2]

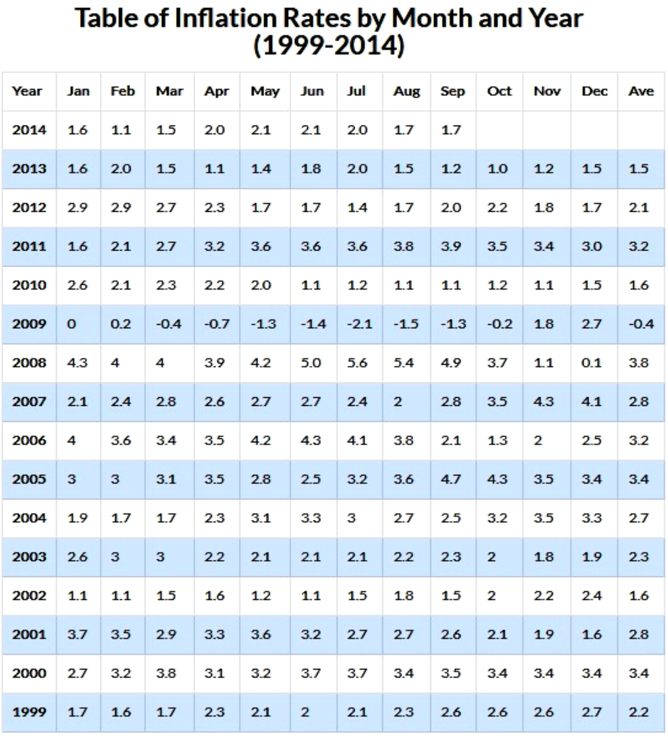
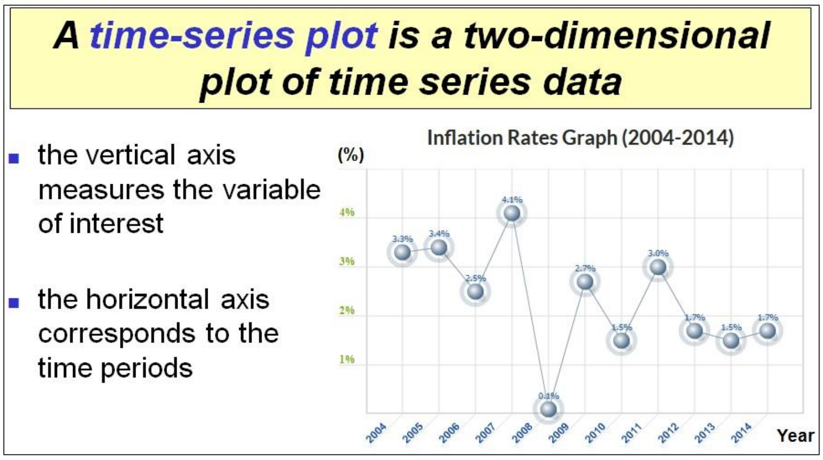


Table of Inflation Rates by Month and Year (1999-2014)

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ave
2014	1.6	1.1	1.5	2.0	2.1	2.1	2.0	1.7	1.7				
2013	1.6	2.0	1.5	1.1	1.4	1.8	2.0	1.5	1.2	1.0	1.2	1.5	1.5
2012	2.9	2.9	2.7	2.3	1.7	1.7	1.4	1.7	2.0	2.2	1.8	1.7	2.1
2011	1.6	2.1	2.7	3.2	3.6	3.6	3.6	3.8	3.9	3.5	3.4	3.0	3.2
2010	2.6	2.1	2.3	2.2	2.0	1.1	1.2	1.1	1.1	1.2	1.1	1.5	1.6
2009	0	0.2	-0.4	-0.7	-1.3	-1.4	-2.1	-1.5	-1.3	-0.2	1.8	2.7	-0.4
2008	4.3	4	4	3.9	4.2	5.0	5.6	5.4	4.9	3.7	1.1	0.1	3.8
2007	2.1	2.4	2.8	2.6	2.7	2.7	2.4	2	2.8	3.5	4.3	4.1	2.8
2006	4	3.6	3.4	3.5	4.2	4.3	4.1	3.8	2.1	1.3	2	2.5	3.2
2005	3	3	3.1	3.5	2.8	2.5	3.2	3.6	4.7	4.3	3.5	3.4	3.4
2004	1.9	1.7	1.7	2.3	3.1	3.3	3	2.7	2.5	3.2	3.5	3.3	2.7
2003	2.6	3	3	2.2	2.1	2.1	2.1	2.2	2.3	2	1.8	1.9	2.3
2002	1.1	1.1	1.5	1.6	1.2	1.1	1.5	1.8	1.5	2	2.2	2.4	1.6
2001	3.7	3.5	2.9	3.3	3.6	3.2	2.7	2.7	2.6	2.1	1.9	1.6	2.8
2000	2.7	3.2	3.8	3.1	3.2	3.7	3.7	3.4	3.5	3.4	3.4	3.4	3.4
1999	1.7	1.6	1.7	2.3	2.1	2	2.1	2.3	2.6	2.6	2.6	2.7	2.2

[1] The difference between *parametric model* and *non-parametric model* is that the former has a fixed number of parameters, while the latter grows their number with the amount of training data (Murphy, 2012, p.16).

[2] Source: http://www.usinflationcalculator.com/inflation/current-inflation-rates



A *time-series plot* is a two-dimensional plot of time series data

- the vertical axis measures the variable of interest
- the horizontal axis corresponds to the time periods

Fig.7-1 Example of Time Series Plot
(Source: http://www.usinflationcalculator.com/inflation/current-inflation-rates/)

In this Chapter, we are going to use the same notations for time series analysis as in Chapter 5. For example, the notation specifying a time series which is indexed by the natural numbers is written as

$$\boldsymbol{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots \boldsymbol{X}_T\} \text{ or } \boldsymbol{Y} = \{\boldsymbol{Y}_t\text{: } \boldsymbol{t} \in \boldsymbol{T}\}, \text{ where } \boldsymbol{T} \text{ is the index set.}$$

Time series models are used for predicting the future behavior of variables based on previously observed data. These models account for the fact that data points taken over time may have an internal structure (such as autocorrelation, trend or seasonal variation) that should be accounted for. As a result, standard regression techniques cannot be applied to time series data directly and a methodology has been developed to decompose the trend, seasonal and cyclical component of the series. Modeling the dynamic path of a variable can improve forecasts since the predictable component of the series can be projected into the future.

Time Series Decomposition (Kendall, 1976) based on rates of change is an important technique for all types of ***TSA***, especially for seasonal adjustment. It seeks to construct, from an observed time series, a number of component series (that could be used to reconstruct the original by additions or multiplications) where each of these has a certain characteristic or type of behavior. For example, time series are usually decomposed (as shown in Fig.7-2) into:

- ***Trend Component*** (**T$_t$**) is the long-run increase or decrease over time (overall upward or downward movement) that reflects the long-term progression of the series (secular variation[3]).

[3] The ***secular variation*** of a time series is its long-term non-periodic variation. Whether it is perceived as a secular or not depends on the timescale – a secular variation over a time scale of decades may be part of a periodic variation over a time scale of thousands of years. Natural quantities may have both ***periodic*** and ***secular variations***. Secular variation is also known as ***secular trend*** when the emphasis is on a ***linear long-term trend***.

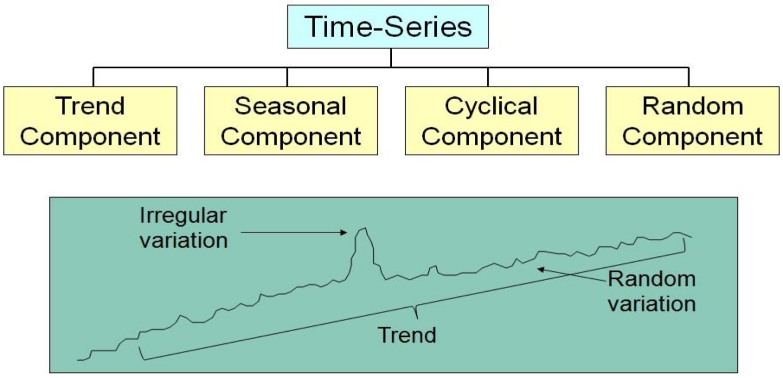


Fig.7-2 ***Time Series Decomposition*** and example of ***Time Series*** data with ***Trend*** and ***Residuals – Random*** and ***Irregular variation***

- ***Cyclical Component*** ($\mathbf{C}_t$) describes repeated but non-periodic fluctuations with duration of at least 2 years. In business, these fluctuations are wavelike variations of more than one year's duration due to changing economic conditions.
- ***Seasonal Component*** ($\mathbf{S}_t$) reflecting seasonality (seasonal variation) – short-term regular variations when a time series is influenced by seasonal factors (the quarter of the year, the month, or day of the week). The movement in data is classified quarterly, monthly or weekly (i.e. seasonality is always of a fixed and known period) and it is completed within the duration of a year and repeats itself year after year. Typical examples are weather variations (and related sales in winter and summer sports equipment), vacations (airline travel, greeting cards sales) and so forth.
- ***Random Component***, the ***Error***, or ***Irregular Component*** ($\mathbf{E}_t$) represents the residuals of the time series after all other components have been removed. These are unpredictable, random fluctuations due to random variations, which are sometimes referred to as the "***Noise***" in the time series that describes random, irregular influences. It is important to distinguish between the ***Random variations*** (caused by chance) and the ***Irregular variations*** (caused by unusual circumstances, natural disasters, accidents or unusual events) as presented in Fig.7-2.

Sometimes, another decomposition based on predictability is applied. The theory of ***TSA*** makes use of the idea of decomposing a times series into deterministic and non-deterministic components (or predictable and unpredictable components). The so-called Wold's theorem or Wold decomposition (Wold, 1954) in ***TSA***, implies that any stationary discrete-time stochastic process can be decomposed into a pair (or as a sum) of uncorrelated processes – a deterministic component, and a stochastic component which can itself be expressed as an infinite moving average. The usefulness of the Wold's theorem is that it allows the dynamic evolution of a time series variable $\boldsymbol{Y_t}$ to be approximated by a linear model (see section 8.1).

Inexperienced users often confuse cyclic behavior with seasonal behavior, but they are quite different. In general, the average length of cycles is longer than the length of a seasonal pattern, and the magnitude of cycles tends to be more variable than the magnitude of seasonal patterns. Also, the cyclic fluctuations are not of a fixed period. If the period is unchanging and associated with some aspect of the calendar, then the pattern is seasonal.

In some analysis, trend ($\mathbf{T_t}$) and cycle ($\mathbf{C_t}$) are considered as one, ***trend-cycle component*** ($\mathbf{T_t}$x$\mathbf{C_t}$) containing both trend and cycle. Sometimes, for example in Exponential smoothing equations (5-28) and (5-29), the ***trend-cycle component*** is simply called "***The trend***" component ($\mathbf{T_t}$), even though it may contain cyclic behavior as well.

If we consider the time series $\boldsymbol{Yt = \{Y_1, Y_2, \ldots Y_T\}}$ as comprising all four components we can present a time series in the same two variations that differ in the nature of the seasonal component:

a) as an additive model:

$$\boldsymbol{Y_t = T_t + S_t + C_t + E_t} \qquad (7\text{-}1)$$

b) or a multiplicative model:

$$\boldsymbol{Y_t = T_t \text{ x } S_t \text{ x } C_t \text{ x } E_t} \qquad (7\text{-}2)$$

To choose an appropriate decomposition model, we should examine a graph of the original series (see Fig.7-1) and try a range of models, selecting the one which yields the most stable seasonal component. As mentioned in Chapter 5, the additive model is most appropriate if the magnitude of the seasonal fluctuations or the variation around the trend-cycle does not vary with the level of the time series (i.e. it is independent of the current level of the series). When the variation in the seasonal pattern, or the variation around the trend-cycle, appears to be proportional to the level of the time series (i.e. it varies with changes in the level), then a multiplicative model is the most likely candidate. The multiplicative models are common with economic time series and used primarily for forecasting.

An important point, that needs some clarification here, is the change in the time series data when there is high inflation in the data range. Often, the change (or part of it) is due to the change in price because of inflation. Thus:

Total change = (real change in physical quantity +change in price because of inflation)

To measure the second component in this equation we must know what the ***Purchasing power*** of the currency is, i.e. how much of the change in dollar values is due to change in price because of inflation. ***Purchasing power,*** sometimes retroactively called ***adjusted for inflation***, is the number of goods or services that can be purchased with a unit of currency.

For example, if someone had taken one unit of currency to a store in the 1950s, it is probable that it would have been possible to buy a greater number of items than they would today, indicating that one unit would have had a greater ***purchasing power*** in the 1950s. Currency can be either commodity money or free-floating market-valued currency like US dollars, and so on.

If a monetary income stays the same, but the price level increases, the ***purchasing power*** of that income falls. Inflation does not always imply falling ***purchasing power*** of one's money income since it may rise faster than the price level. A higher ***real income*** means a higher ***purchasing power*** since ***real income*** refers to the income adjusted for inflation.

When high inflation is part of the process we want to predict we must adjust the process expressing terms in a series in ***constant dollars***. For this purpose, we can use a special type of accounting model (***Constant Dollar*** accounting) that converts[4] nonmonetary assets and equities from historical dollars to current dollars using a *general price index*. Monetary items are not adjusted, so they gain or lose ***purchasing power***. To understand better this idea we need to explain what an index number is and how to use it in business forecasting.

Index Numbers

An index is a statistical measure of changes in a representative sample of observations. These data may be derived from any number of sources, including company performance, prices, productivity, and employment. Some indices display market variations that cannot be captured in other ways. For example, the Economist provides a Big Mac Index that expresses the adjusted cost of a globally ubiquitous Big Mac as a percentage over or under the cost of a Big Mac in the U.S. in USD (estimated: $3.57)[5]. The least relatively expensive Big Mac price occurs in Hong Kong, at a 52% reduction from U.S. prices, or $1.71 U.S. Such indices can be used to help forecast currency values. From this example, it would be assumed that Hong Kong currency is undervalued, and provides a currency investment opportunity.

An index number is an economic data figure reflecting price or quantity compared with a standard or base value. The base period index by definition equals 100 and the index number is usually expressed as 100 times the ratio of this base value. For example, if a commodity costs twice as much in 2010 as it did in 2000, its index number would be 200 relatives to 2000. Index numbers allow relative comparisons over time and are used specially to compare business activity, the cost of living, and employment. They enable economists to reduce unwieldy business data into easily understood terms.

[4] This is similar to a currency conversion from old dollars to new dollars.

[5] Source: http://en.wikipedia.org/wiki/Index_%28economics%29

$$\boxed{I_t = \frac{y_t}{y_0}100} \qquad (7\text{-}3)$$

where $\boldsymbol{I_t}$ is the index number at time period $\boldsymbol{t}$ ($\boldsymbol{t} = \{\boldsymbol{1, 2, \dots T}\}$) where $\boldsymbol{T}$ is the index set***;***

$\boldsymbol{y_t}$ – observed value of the time series at time $\boldsymbol{t}$

$\boldsymbol{y_0}$ – observed value of the time series in the base period.

In economics, index numbers are generally time series summarizing movements in a group of related variables. In some cases, however, index numbers may compare geographic areas at a point in time. An example is a country's purchasing power parity. The best-known index number is the ***consumer price index (CPI)***, which measures changes in retail prices paid by consumers. The Cost-of-living index is a price index number that measures relative cost of living over time. Other common indices are the Producer Price Index, Stock Market Indexes, such as Dow Jones Industrial Average, S&P 500 Index, NASDAQ Index, and so on.

To answer the question "How much of the change in dollar values is due to change in price because of inflation?" i.e. what is the ***Purchasing Power*** (***PP***) we can use the formula:

Current PP of $1 = 100/Current CPI (7-4)

Then, to express dollar values in terms of ***constant dollars*** we use the formula:

Deflated dollar value (Constant dollars) = (Current dollar value) x (PP of $1) (7-5)

"***Constant dollars***" is an adjusted value of the currency used to compare dollar values from one time period to another. Due to inflation, the purchasing power of the dollar changes over time, so in order to compare dollar values from one year to another, they need to be converted from *nominal (current) dollar values* to constant dollar values, also known as real dollars, where all values are expressed in terms of a common reference year. The process of converting from nominal to real values is known as ***inflation adjustment***. Technically, it is like finding the common denominator when adding or comparing fractions[6].

The calculation for conversion of nominal dollars in year **A** to constant dollars in year **A** expressed in terms of year **B** prices can be based on any relevant price index, such as the ***CPI***. Any year can be used as a baseline for comparing **A** and **B** if it is consistent. For example, both values could be converted into year **C** dollars. When the base year is different from both, the current year **A** and the reference year **B**, the formula (7-5) is modified to:

Constant dollars $_{\textit{(year A)}}$ ***= Current dollars*** $_{\textit{(year A)}}$ **** CPI*** $_{\textit{(year B)}}$ ***/ CPI*** $_{\textit{(year A)}}$ (7-6)

[6] For example, if the price level is twice as high in year A as in the reference year B, then nominal dollars in year A are divided by 2 to obtain their equivalent in terms of year B prices.

In summary, we need to deflate a time-series (when high inflation is part of the data range), which means to adjust the observed values to a base year equivalent and to allow uniform comparisons over time. The general deflation formula is:

$$\boxed{y_{adj(t)} = \frac{y_t}{I_t}(100)} \qquad (7\text{-}7)$$

where $\boldsymbol{y_{adj(t)}}$ is adjusted time-series value at time ***t*** (***t***={***1, 2,…T***}), ***T*** is the index set;

$\boldsymbol{y_t}$ – observed value of the time series at time ***t***

$\boldsymbol{I_t}$ – index (such as CPI) at time period ***t.***

7.2. Trend Estimation and Forecasting

A ***trend line*** represents the long-term movement in time series data after other components have been accounted for. It tells whether a particular data set (say GDP, oil prices or stock prices) have increased or decreased over the period of time. ***Trend lines*** typically are straight lines, although some variations use higher degree polynomials depending on the degree of curvature desired in the line (see Fig.7-3).

Trend lines are used in business analytics to show changes in data over time. Having the advantage of being simple, ***linear trends*** do not require a control group, experimental design, or another sophisticated analysis technique. However, ***linear trend*** suffers from a lack of scientific validity in cases where other potential changes can affect the data.

When a series of measurements of a process are treated as a time series, ***trend estimation*** can be used to make and justify statements about tendencies in the data, by relating the measurements to the times at which they occurred. A ***trend line*** could simply be drawn by eye through a set of data points, but more properly their position and slope is calculated using estimation techniques like ***linear regression***.

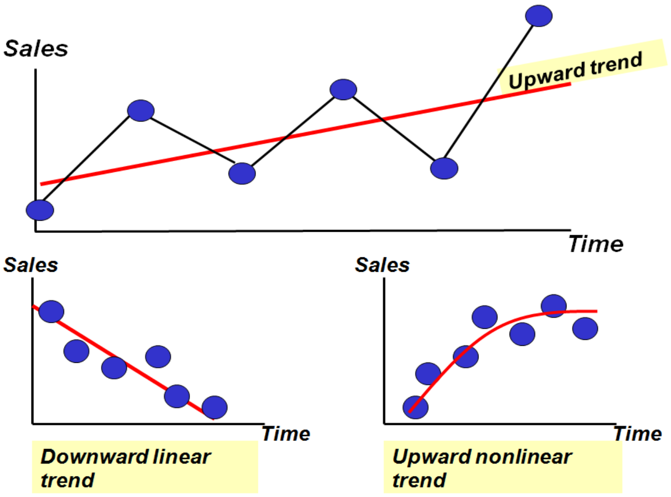


Fig.7-3 ***Trend*** examples

By using ***trend estimation,*** it is possible to construct a model which is independent of the theoretical knowledge about the nature of the process of an incompletely understood system (for example, physical, economic, or other system). This model can then be used to describe the behavior of the observed data and to make predictions. In particular, it may be useful to determine if measurements exhibit an increasing or decreasing trend, which is statistically distinguished from random behavior.

Trend line estimation is an example of ***simple linear regression*** application. If we present a trend line as a typical linear model (6-2), then we can apply the ***LS*** estimator as discussed in Chapter 6.2. In general, we can always present time series data as a trend plus noise:

Population y intercept

Population Slope Coefficient

Time as Explanatory Variable

Random Error term, or residual

Dependent Variable

$$y = \underbrace{\beta_0 + \beta_1 t}_{\text{Linear component}} + \underbrace{\varepsilon}_{\text{Random Error component}} \tag{7-8}$$

This is the population linear trend model where:

– ***y*** is the dependent variable $y_t = \{y_1, y_2, \ldots y_T\}$, ($t = \{1, 2, \ldots T\}$), ***T*** is the index set;

– ***t*** is the time considered as an explanatory variable in the regression model;

– ***ε*** is the residual, or the noise.

Applying ***OLS*** to equation (7-8) we can fit the trend line minimizing the sum of the squared errors (***ε***):

Estimated (predicted) average y value

Estimate of the regression intercept

Estimate of the regression slope

$$\hat{y}_t = b_0 + b_1 t \tag{7-9}$$

Predictor variable time

where:

– ***Y-hat*** (ŷ) is the estimated average value of dependent variable ***y*** ($y_t = \{y_1, y_2, \ldots y_T\}$);

– the intercept $\mathbf{b_0}$ is the estimated ŷ value when the regressor ***t***=0;

– the slope of the line $\mathbf{b_1}$ is the average change in dependent variable ***y*** for each change of one unit in regressor ***t***;

– the individual random error (***ε***) terms $\mathbf{e_t}$ ($\mathbf{e_t} = y_t - \hat{y}_t$) have a mean of zero ($\bar{e} = 0$), i.e. the estimation is unbiased.

The Linear trend model has another important advantage, giving a very clear business interpretation of the slope $\mathbf{b_1}$. In fact, this change in the dependent variable ***y*** for one unit in ***t*** (i.e. time period) represents the (economic) ***growth***, which is the increase in the market value of the goods and services produced by a company (or an economy) over time.

Year	Time Period (t)	Sales (y)
2009	1	20
2010	2	40
2011	3	30
2012	4	50
2013	5	70
2014	6	65

- Linear trend model is:

$$\hat{y} = 12.333 + 9.5714\ t$$

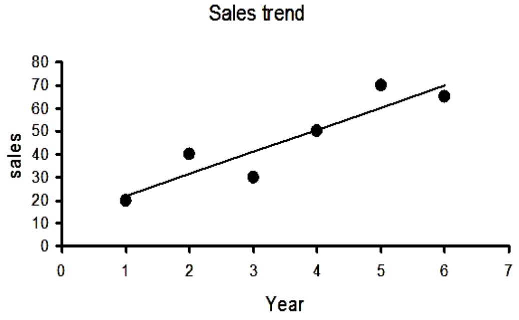


Fig.7-4 ***Linear Trend*** Example

As mentioned above, ***linear trend models*** are very simple in both fitting their lines and using the estimated equation to make predictions. Consider the Example presented in Fig.7-4. Applying ***OLS*** to these data is easy and could be done as to any other simple linear regression. Using MS Excel will provide all the calculations necessary in this case. Then, making predictions is like using the simple ***linear regression model:***

$$y^*_{(T+l)} = b_0 + b_1(t) \tag{7-10}$$

where:

– y^* is the forecast of dependent variable y ($y_t = \{y_1, y_2, \dots y_T\}$) for period ($T+l$), $l>0$;
– t is the time considered as a predictor, $t = \{T+1, T+2, \dots T+l\}$.

Technically, it means that we substitute t with the desired value for the forecasting period and solve equation (7-10) for y^*. The result, as presented in Fig.7-5 is rounded according to the original scale of units. In this case it is rounded down because rounding up will overestimate the expected Sales in the year 2015 (i.e. time period 7).

Year	Time Period (t)	Sales (y)
2009	1	20
2010	2	40
2011	3	30
2012	4	50
2013	5	70
2014	6	65
2015	7	??

- Forecast for 2015 - time period 7:

$$\hat{y} = 12.333 + 9.5714(7) = \\ = 79.33 \approx \boxed{79}$$

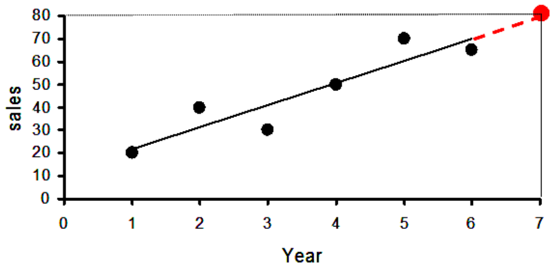


Fig.7-5 Example of Forecasting with a ***Linear Trend Model***

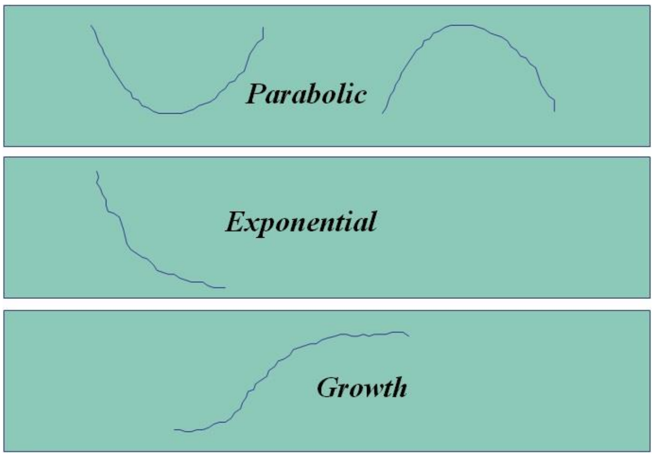


Fig.7-6 Examples of Common Nonlinear Trends

The use of a ***linear trend line*** has been the subject of criticism, leading to a search for alternative approaches to avoid its use in forecasting. For example, by default a linear trend in a model precludes the presence of fluctuations in the tendencies of the dependent variable over time. At the same time, many economic time series are in fact characterized by non-linear, for example, exponential growth. To address this problem, non-linear trend (see Fig.7-6) models were developed using different non-linear functions and models[7].

As already discussed in Chapter 6.3, there are both advantages and disadvantages, which must be considered whenever a selection of a non-linear model should be made, and unfortunately, in the case of trend estimation, there are not too many advantages. In fact, real-life business data need more advanced models than the single trend one (no matter linear or non-linear), especially when other potential changes can affect the data.

Typically, the quality of a particular method of trend estimation and extrapolation is limited by the assumptions about the function made by the method. In this section of the text, time series data have been assumed to consist of the trend plus noise, with the noise (i.e. the error) at each data point being independent and identically distributed random variables and to have a normal distribution. Real-life data may not fulfill these criteria. This is important, as it makes an enormous difference to the ease with which the statistics can be analyzed so as to extract maximum information from the time series. If there are other non-linear effects that have a correlation to the independent variable (such as cyclic influences), the use of least-squares estimation of the trend is not valid. Also, where the variations are significantly larger than the resulting straight-line trend, the choice of start and end points can significantly change the result.

[7] See for example http://www.hedengren.net/research/models.htm

Statistical inferences (tests for the presence of trend, confidence intervals for the trend, and so on) are invalid unless departures from the standard assumptions are properly accounted for. For example, in the case of residuals dependence, i.e. autocorrelated time series, it might be modeled using autoregressive moving average models (discussed in Chapter 8, Section 1).

One relatively simple and at the same time succesful technique is the Time Series decomposition, additive (7-1) or multiplicative (7-2). Technically, it means to estimate the time-series components and then to combine the predictions. One possible approach, known as ***Seasonally Adjusted Time Series Trend Forecasts*** is discussed in the next topic.

7.3. Seasonal Component Estimation and Forecasting

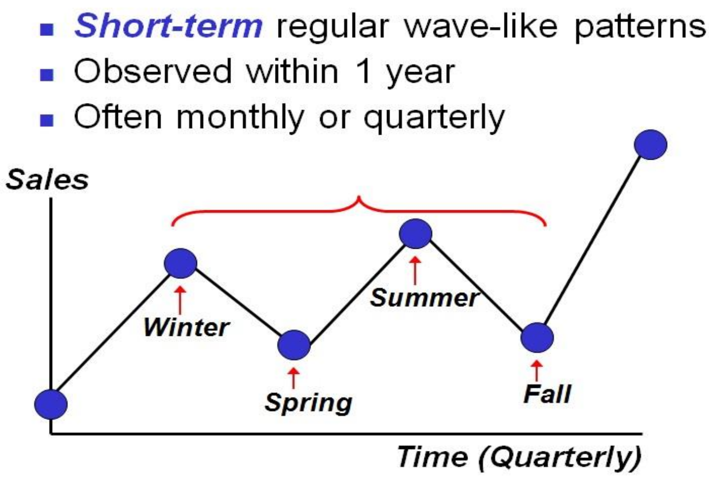
Many time series exhibit cyclic variation known as ***seasonality, seasonal variation, periodic variation,*** or ***periodic fluctuations. Seasonal variation*** is a component of a time series which is defined as the repetitive and predictable movement around the trend line in one year or less. It is detected by measuring the quantity of interest for small time intervals, such as days, weeks, months or quarters.

Organizations facing ***seasonal variations***, like the motor vehicle industry, are often interested in knowing their performance relative to the normal seasonal variation. The same applies to the ministry of employment which expects unemployment to increase in June because recent graduates are just arriving into the job market and schools have also been given a vacation for the summer.

Seasonality is quite common in economic time series. For example, retail sales tend to peak for the Christmas season and then decline after the holidays. So time series of retail sales will typically show increasing sales from September through December and declining sales in January and February.

Organizations affected by ***seasonal variation*** need to identify and measure this seasonality to help with planning for temporary increases or decreases in labor requirements, inventory, training, and so forth. Apart from these considerations, the organizations need to know if the variation they have experienced has been more or less than the expected, given the usual seasonal variations.

Unlike the trend and cyclical components, the ***seasonal component*** tends to happen with similar magnitude during the same time period. Often, the ***seasonal component*** causes the interpretation of a time series to be ambiguous. By removing it, we can focus on the other components. When the ***seasonal component*** is not removed from monthly data, year-on-year changes (i.e. annual data) are utilized in an attempt to avoid contamination with seasonality.



- *Short-term* regular wave-like patterns
- Observed within 1 year
- Often monthly or quarterly

Fig.7-7 General presentation of seasonality

There are several important reasons for studying seasonal variation:

- The description of the seasonal effect provides a better understanding of the impact this component has upon a particular time series.
- When establishing a seasonal pattern, we should eliminate it from the time-series to study the effect of other components such as cyclical and irregular variations. This elimination of the seasonal effect is referred to as ***deseasonalizing*** or ***seasonal adjustment of data.***
- Projecting the existing patterns into the future knowledge domain of the seasonal variations is a must for the time series prediction.

Seasonal Plots

The first step in studing the ***seasonal component*** is to detect ***seasonality***. There are a few simple, mainly graphical, techniques that can be used to identify a seasonal pattern. The ***run sequence plot***[8] (a chart like Fig.7-7) is a recommended first step for analyzing any time series. Although this plot can indicate a ***seasonal pattern***, ***seasonality*** is shown more clearly by few other charts.

One useful tool for detecting seasonality in a time series is the ***seasonal subseries plot***. For example, the ***seasonal subseries plot*** in Fig.7-8, which contains monthly data of Small Cap Fund Returns, reveals a strong seasonal pattern. The Fund Returns concentrations peak in May, steadily decrease through September, and then begin rising again until the May peak.

The ***seasonal subseries plot*** can provide answers to questions, like: "Do the data exhibit a seasonal pattern?"; "What is the nature of the seasonality?"; "Is there a within-group pattern (e.g., do May and August exhibit similar patterns)?" and so forth...

[8] A ***run chart***, also known as a ***run-sequence plot***, is a graph that displays observed data in a time sequence.

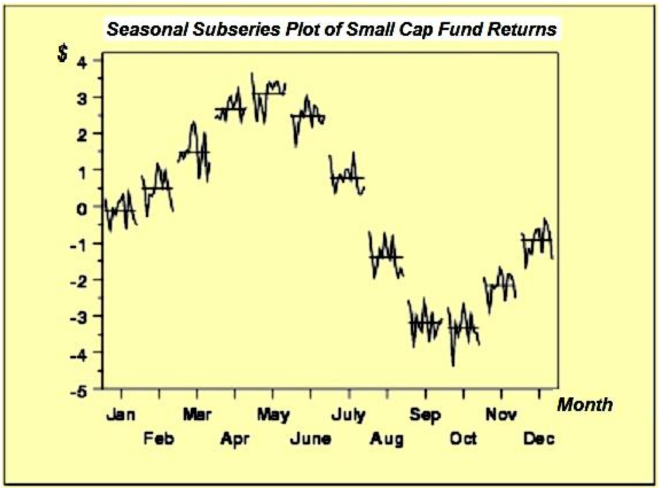


Fig.7-8 Example of ***seasonal subseries plot***

Multiple ***box plots*** can be used as an alternative to the ***seasonal subseries plot*** to detect ***seasonality***. ***Box plots*** are useful in any data series, particularly for detecting and illustrating location and variation changes between different groups of data. In general, ***box plots*** like Fig.7-9 show the seasonal difference (i.e. between-group patterns) quite well, but do not show within-group patterns. However, for large data sets, the ***box plot*** is usually easier to read than the ***seasonal subseries plot***.

Multiple ***box plots*** can be drawn together to compare multiple time series or to compare groups in a single time series data set. The box plot in Fig.7-9, represents the monthly sales revenue of four company branches. It shows that there is a significant difference in sales revenue with respect to both location and variation. Division 3 has the highest average sales revenue (about 72.5) and branch 4 has the least variable sales revenue with about 50% of its readings being within 1 sales revenue unit ($1,000).

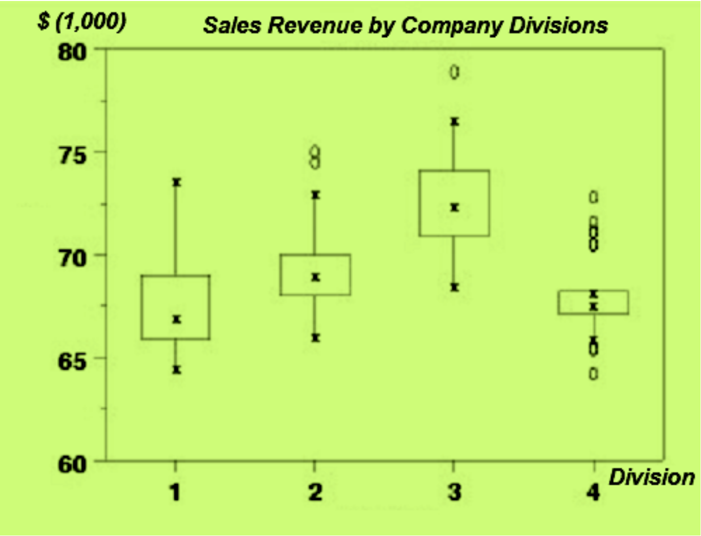


Fig.7-9 Example of multiple ***box plots***

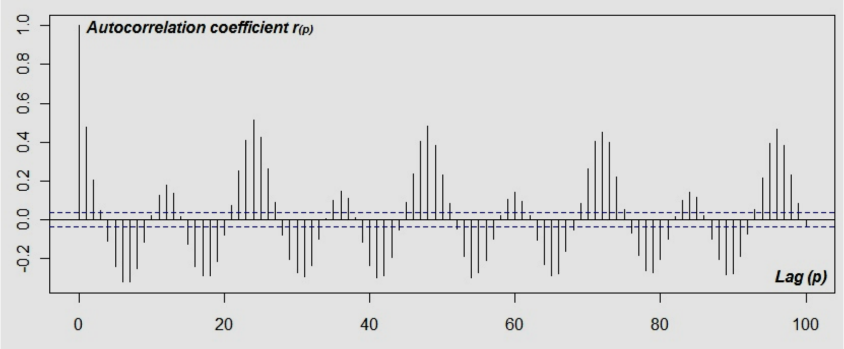


Fig.7-10 Example of ***autocorrelation plot***

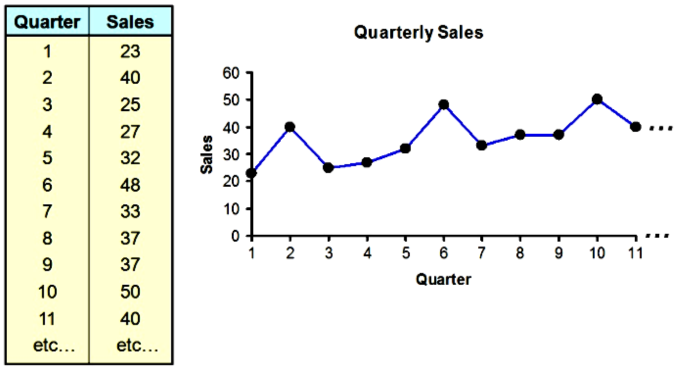
Both the ***seasonal subseries plot*** and the ***box plot*** assume that the seasonal periods are known. In most cases, the analyst will know this. For example, for monthly data, the period is 12 since there are 12 months in a year. However, if the period is not known, the ***autocorrelation plot*** (Box & Jenkins, 1976, pp. 28-32) can help. In time series analysis, an ***autocorrelation plot***, also known as a correlogram, is a plot (see Fig.7-10) of the time series autocorrelation coefficients $\boldsymbol{r_p}$, versus $\boldsymbol{p}$ (the time lags). If there is significant seasonality, the ***autocorrelation plot*** should show spikes at lags equal to the period. For example, for monthly data, if there is a seasonality effect, we would expect to see significant peaks at lag 12, 24, 36, and so on.

Seasonal Index

Seasonal variation is measured in terms of an index, known as a ***seasonal index***. It is an average that can be used to compare an actual observation relative to what it would be if there were no seasonal variations, i.e. the ***seasonal index*** measures how much the average for a particular period tends to be above (or below) the expected value. An index value is attached to each period of the time series within a year. This implies that if monthly data are considered, there are twelve separate seasonal indices, one for each month, or there are four index values for quarterly data. The following methods use seasonal indices to measure seasonal variations of a time-series data:

- Method of simple averages
- Ratio to trend method
- ***Ratio-to-moving average*** method

The Ratio-to-moving average is a very common technique used in ***Time Series Analysis*** and ***Forecasting.*** The measurement of seasonal variation by using the ***ratio-to-moving average*** method provides an index to measure the degree of the seasonal variation in a time series. The index is based on a mean of 100 (see equation (7-3) and the corresponding comments and explanations), with the degree of seasonality measured by variations away from the base. For better understanding, we will use the Sales Example, presented in Fig.7-11.



Quarter	Sales
1	23
2	40
3	25
4	27
5	32
6	48
7	33
8	37
9	37
10	50
11	40
etc…	etc…

Fig.7-11 Sales example Time Series Data and their ***Run-sequence Plot***

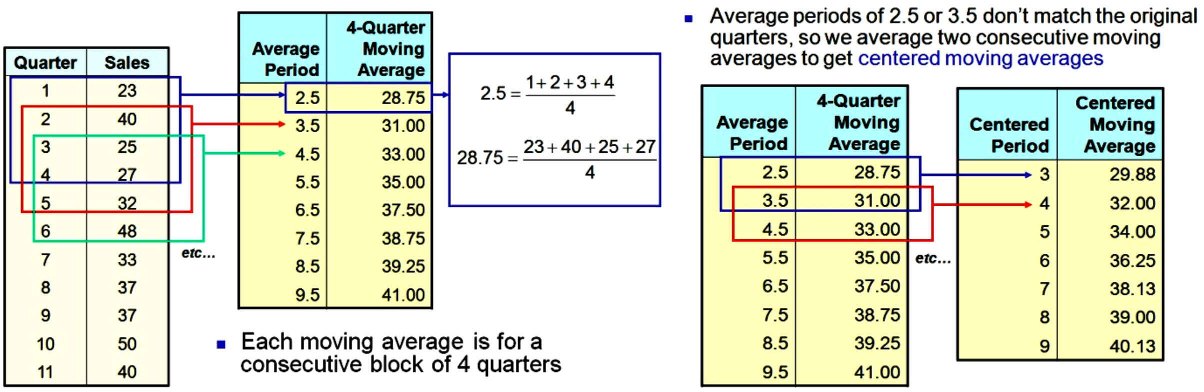
The ***Run-sequence Plot*** in Fig.7-11 reveals very clear seasonal variation by quarters. Assuming the multiplicative model (7-5), the seasonal component can be expressed in terms of ratio and percentage as Seasonal effect:

$$\mathbf{S_t = (T_t \times S_t \times C_t \times E_t)/(T_t \times C_t \times E_t)}*\mathbf{\mathit{100}} = \mathbf{Y_t/(T_t \times C_t \times E_t)}*\mathbf{\mathit{100}} \qquad (7\text{-}11)$$

i.e. the seasonality is expressed in terms of the amount that actual values $\mathbf{Y_t}$ deviate from the average values (or the trend) of a series. For example, an index of 1.20 for summer would indicate that the value in this quarter is 20% above the quarterly average.

The Ratio-to-moving average technique is a procedure with the following steps:

1. Find the centered four quarterly moving averages of the observed data values in the time-series (see Fig.7-12). Since these averages (see Fig.7-12, a) do not match the original quarters we have to center them as shown in Fig.7-12, b). Calculations are easy and even if there is no statistical software available, we can code and enter each of them as a formula into an MS Excel spreadsheet.



Quarter	Sales
1	23
2	40
3	25
4	27
5	32
6	48
7	33
8	37
9	37
10	50
11	40

Average Period	4-Quarter Moving Average
2.5	28.75
3.5	31.00
4.5	33.00
5.5	35.00
6.5	37.50
7.5	38.75
8.5	39.25
9.5	41.00

$$2.5 = \frac{1+2+3+4}{4}$$

$$28.75 = \frac{23+40+25+27}{4}$$

- Each moving average is for a consecutive block of 4 quarters

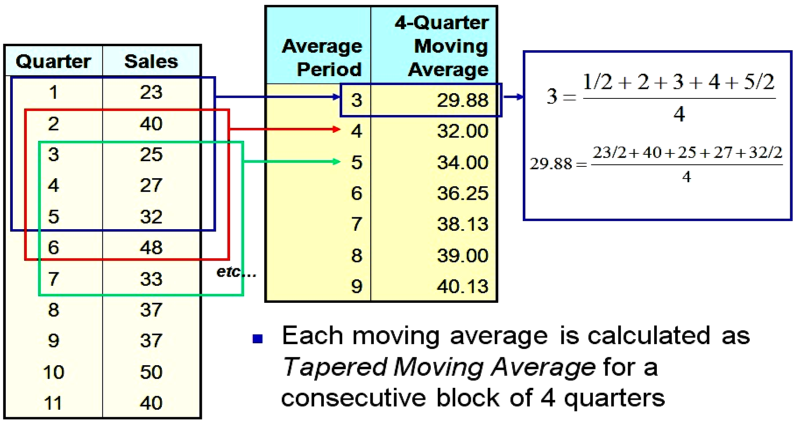
- Average periods of 2.5 or 3.5 don't match the original quarters, so we average two consecutive moving averages to get centered moving averages

Average Period	4-Quarter Moving Average
2.5	28.75
3.5	31.00
4.5	33.00
5.5	35.00
6.5	37.50
7.5	38.75
8.5	39.25
9.5	41.00

Centered Period	Centered Moving Average
3	29.88
4	32.00
5	34.00
6	36.25
7	38.13
8	39.00
9	40.13

a) b)

Fig.7-12 Calculating 4-Quarter ***Centered Moving Average*** in the Sales Example



Quarter	Sales
1	23
2	40
3	25
4	27
5	32
6	48
7	33
8	37
9	37
10	50
11	40

Average Period	4-Quarter Moving Average
3	29.88
4	32.00
5	34.00
6	36.25
7	38.13
8	39.00
9	40.13

$$3 = \frac{1/2 + 2 + 3 + 4 + 5/2}{4}$$

$$29.88 = \frac{23/2 + 40 + 25 + 27 + 32/2}{4}$$

- Each moving average is calculated as *Tapered Moving Average* for a consecutive block of 4 quarters

Fig.7-13 Calculating 4-Quarter ***Tapered Moving Average*** in the Sales Example

It should be noted that these computations could be done in one step if we use the ***Tapered Moving Average*** formula (5-11), which in the Sales Example is looking as follows:

$$\hat{y}_t = \frac{(1/2y_{t-2}+y_{t-1}+y_t+y_{t+1}+1/2y_{t+2})}{4} \qquad (7\text{-}12)$$

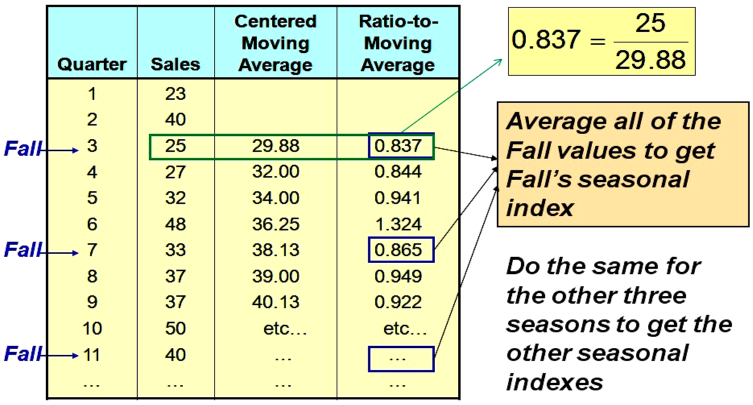
The calculations in the Sales Example are presented in Fig.7-13.

2. Estimate the seasonal component – it means to express each original data value of the time-series as a percentage of the corresponding centered moving average value obtained in step (1). In a multiplicative time-series model, we get (Original data values)/(Trend values). This implies that the ***ratio–to-moving average*** represents the seasonal and irregular components (residuals) as one **S$_t$xE$_t$** element, assuming that trend (**T$_t$**) and cycle (**C$_t$**) are also one, ***trend-cycle component*** (**T$_t$**x**C$_t$**) containing both trend and cycle as computed at step (1).

To estimate the **S$_t$xE$_t$** value we use the ***Ratio-to-Moving Average*** formula:

$$\boxed{S_t \times E_t = \frac{y_t}{T_t \times C_t}} \qquad (7\text{-}13)$$

Computations for the Fall quarter in Fig.7-14 give an example of how to do this.



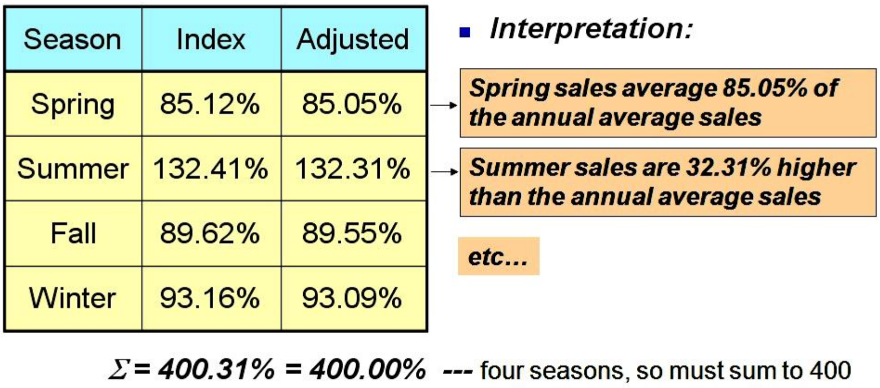
Quarter	Sales	Centered Moving Average	Ratio-to-Moving Average
1	23		
2	40		
3	25	29.88	0.837
4	27	32.00	0.844
5	32	34.00	0.941
6	48	36.25	1.324
7	33	38.13	0.865
8	37	39.00	0.949
9	37	40.13	0.922
10	50	etc…	etc…
11	40	…	…
…	…	…	…

$$0.837 = \frac{25}{29.88}$$

Average all of the Fall values to get Fall's seasonal index

Do the same for the other three seasons to get the other seasonal indexes

Fig.7-14 ***Ratio-to-Moving Average*** Steps (2) and (3) in the Sales Example



Season	Index	Adjusted
Spring	85.12%	85.05%
Summer	132.41%	132.31%
Fall	89.62%	89.55%
Winter	93.16%	93.09%

- ***Interpretation:***

Spring sales average 85.05% of the annual average sales

Summer sales are 32.31% higher than the annual average sales

etc...

Σ ***= 400.31% = 400.00%*** **---** four seasons, so must sum to 400

Fig.7-15 ***Seasonal Index*** values and interpretation in the Sales Example

3. Arrange these percentages according to months or quarter of given years. Find the averages over all months or quarters of the given years. Thus, the Spring quarter seasonal index in our Sales Example is:

$$\textbf{\textit{Spring}}_{(Si)} = (0.837+0.865)/2*100 = 85.12\% \qquad (7\text{-}14)$$

4. If the sum of these indices is not four for quarterly figures, we have to adjust them – i.e. multiply by a correction factor = 400/(sum of quarterly indices). Otherwise, the 4 quarterly averages will be considered as the seasonal indices.

In our Sales Example, since the sum is 400.31%, we compute the correction factor, which is 400/400.31 = 99.92 and then multiply each index by this value. The sum of the adjusted indices now is 400.00%.

Seasonal Adjustments

The main applications of ***Ratio-to-moving average*** are in ***Time Series Analysis*** and ***Forecasting.*** In an analysis**,** to get a clear picture of the nonseasonal components in time-series, we must deseasonalize original data values by removing the seasonal component from them. If the seasonal component is removed from the original observations, the resulting values are referred to as ***seasonally adjusted data***. For an additive model (7-1), the ***seasonally adjusted data*** are given by $\mathbf{y_t\text{-}S_t}$, and for the multiplicative (7-2), the seasonally adjusted values are obtained using $\mathbf{y_t/S_t}$.

If the variation due to seasonality is not of primary interest, the seasonally adjusted time-series can be useful. For example, monthly unemployment data are usually seasonally adjusted to highlight variation which is due to the underlying state of the economy rather than the seasonal variation. An increase in unemployment due to school leavers seeking work is seasonal variation while an increase in unemployment due to large employers laying off workers is non-seasonal. Most people who study unemployment are more interested in the non-seasonal variation and, usually, employment time-series are seasonally adjusted.

Technically, ***Seasonal Adjustment*** means to calculate the seasonal indices and use them to ***deseasonalize*** the original data. In the Sales Example, the original time series data is ***deseasonalized*** by dividing the observed value by its seasonal index, i.e.:

$$\boxed{T_t \times C_t \times E_t = \frac{y_t}{S_t}} \tag{7-15}$$

Fig.7-16 presents the ***Seasonal Adjustment*** of the Sales time-series as well as the plot of ***deseasonalized*** versus the ***seasonalized*** data. For example, the Spring's sales are typically only 85.05% of the normal quarterly value based on historical data. Then each Spring's sales should be seasonally adjusted by dividing by 0.8505. Similarly, since the Summer's sales are typically 132.31% of normal, then each Summer's sales should be seasonally adjusted by dividing by 1.3231. Thus, Summer's value would be adjusted downward while Spring's would be adjusted upward, correcting for the anticipated seasonal effect.

In ***Time Series Forecasting*** the ***seasonal adjustment*** means to incorporate seasonality into the Forecast, i.e. to adjust the forecast to the seasonal changes. The standard procedure goes through the following steps:

1. Obtain trend estimates for a desired period – using the linear trend equation (7-9) for example. An important point here is "What time series data to be used in regression model estimation – ***seasonalized*** or ***deseasonalized***?"

From a forecasting point of view, the differences look insignificant, as the two regression outputs show in Fig.7-17, but there is a big difference if we compare the two trend models. Even though both models pass all goodness of fit tests, there are a few very important points that need comments.

Quarter	Sales	Seasonal Index	Deseasonalized Sales
1	23	85.05%	27
2	40	132.31%	30
3	25	89.55%	28
4	27	93.09%	29
5	32	85.05%	38
6	48	132.31%	36
7	33	89.55%	37
8	37	93.09%	40
9	37	85.05%	44
10	50	132.31%	38
11	40	89.55%	45
12	???	93.09%	N.A.

$$\boxed{27 = \frac{23}{0.8505}}$$ **etc…**

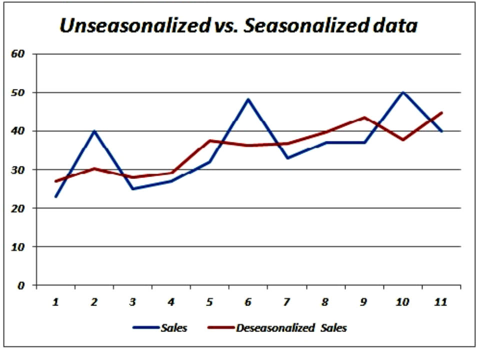
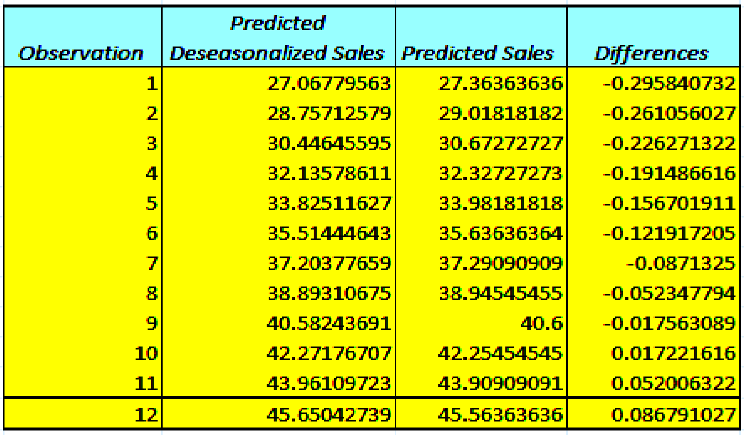


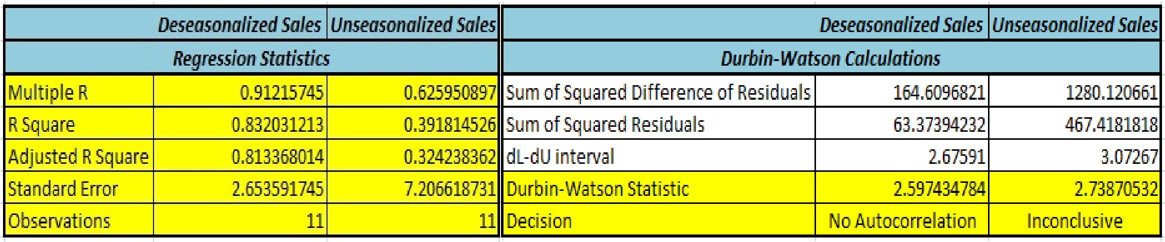
Fig.7-16 ***Deseasonalized*** sales table and plot versus ***seasonalized*** data in the Sales Example



Observation	Predicted Deseasonalized Sales	Predicted Sales	Differences
1	27.06779563	27.36363636	-0.295840732
2	28.75712579	29.01818182	-0.261056027
3	30.44645595	30.67272727	-0.226271322
4	32.13578611	32.32727273	-0.191486616
5	33.82511627	33.98181818	-0.156701911
6	35.51444643	35.63636364	-0.121917205
7	37.20377659	37.29090909	-0.0871325
8	38.89310675	38.94545455	-0.052347794
9	40.58243691	40.6	-0.017563089
10	42.27176707	42.25454545	0.017221616
11	43.96109723	43.90909091	0.052006322
12	45.65042739	45.56363636	0.086791027

Fig.7-17 Predicted sales, deseasonalized sales and their differences in the Sales Example

Fig.7-18 presents several of the regression model statistics for both seasonalized and deseasonalized sales linear trends in the Sales Example. Using deseasonalized time series data leads to a much better trend model, according to all statistics (Multiple R, i.e. Coefficient of correlation, Adjusted R square, Standard Error and D-W statistic). In addition, the residual plot (see Fig.7-19) of deseasonalized sales predictions looks quite random, while seasonalized errors have clear pattern of downward bias.



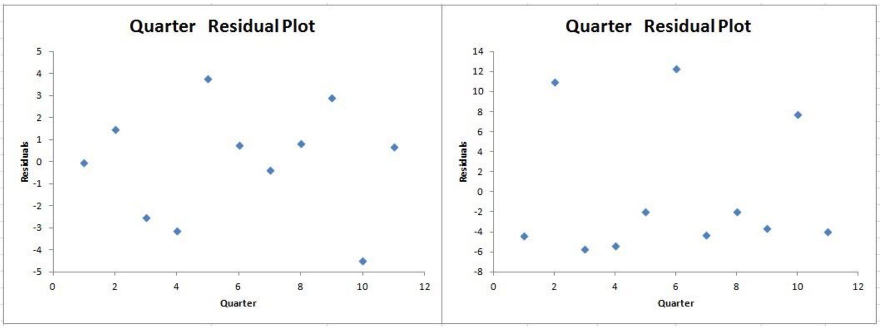
	Deseasonalized Sales	Unseasonalized Sales
Regression Statistics		
Multiple R	0.91215745	0.625950897
R Square	0.832031213	0.391814526
Adjusted R Square	0.813368014	0.324238362
Standard Error	2.653591745	7.206618731
Observations	11	11

	Deseasonalized Sales	Unseasonalized Sales
Durbin-Watson Calculations		
Sum of Squared Difference of Residuals	164.6096821	1280.120661
Sum of Squared Residuals	63.37394232	467.4181818
dL-dU interval	2.67591	3.07267
Durbin-Watson Statistic	2.597434784	2.73870532
Decision	No Autocorrelation	Inconclusive

a) General model statistics b) Durbin-Watson test for Autocorrelation

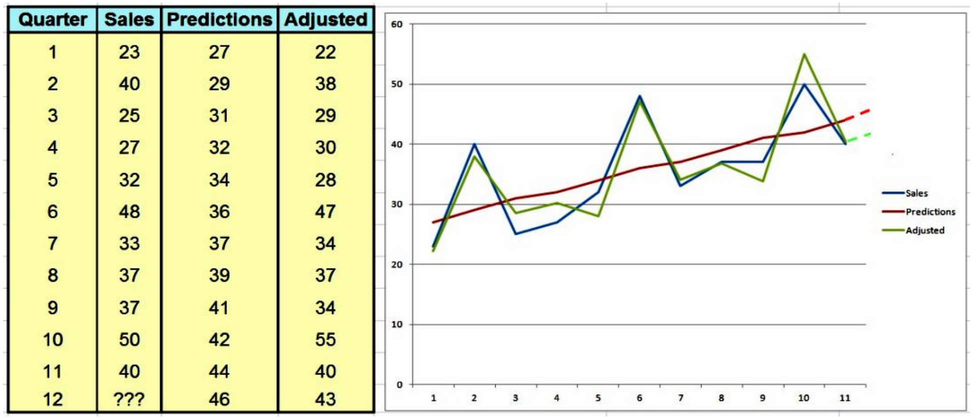
Fig.7-18 Regression model statistics for sales and deseasonalized sales in Sales Example

In summary, we can draw a conclusion, that it is better and should be recommended to use deseasonalized time series data (if available) when developing linear trend model.



a) Deseasonalized sales b) Seasonalized sales

Fig.7-19 Residual plots of deseasonalized and seasonalized predictions in the Sales Example



Quarter	Sales	Predictions	Adjusted
1	23	27	22
2	40	29	38
3	25	31	29
4	27	32	30
5	32	34	28
6	48	36	47
7	33	37	34
8	37	39	37
9	37	41	34
10	50	42	55
11	40	44	40
12	???	46	43

Fig.7-20 Sales, trend forecast and ***seasonally adjusted predictions*** in Sales Example

2. Add seasonality to the trend estimates – in the case of the multiplicative model (7-2), by multiplying the trend estimates by the corresponding seasonal index. For example, the Spring quarter seasonally adjusted forecasts are computed using the following equation:

$$\boldsymbol{Y^*}_{\textbf{(seasonally adjusted spring forecast)}} = \textbf{\textit{Trend forecast}}\ (\mathbf{T}_{\textbf{spring}}) \text{ x } \textbf{\textit{Seasonal index}}\ (\mathbf{S}_{\textbf{spring}}) \qquad (7\text{-}16)$$

The ***seasonal adjustment*** of the ***trend forecast*** improves the predictions and makes them much closer to the real-life observations, as presented in Fig.7-20. It combines the advantages of these two techniques and removes their disadvantages, increasing the reliability of the forecast and reducing the forecast error. ***Seasonally adjusted forecasts*** are more realistic than both the linear trends and the moving average predictions. Eventually, this type of forecast eliminates the main disadvantage of the MA, which in general tend to lag behind turning points in the time series data, for example, if we are averaging three consecutive values, the forecasts will be about one period late in responding to turning points as discussed in Chapter 5.2 and presented in Fig.7-16 in the Sales Example.

7.4. Cyclical Component Estimation and Forecasting

The ***Cyclical Component*** $\mathbf{C}_t$ describes repeated but non-periodic fluctuations in time-series, e.g. wavelike variations of more than one year's duration due to changing economic conditions as shown in Fig.7-21.

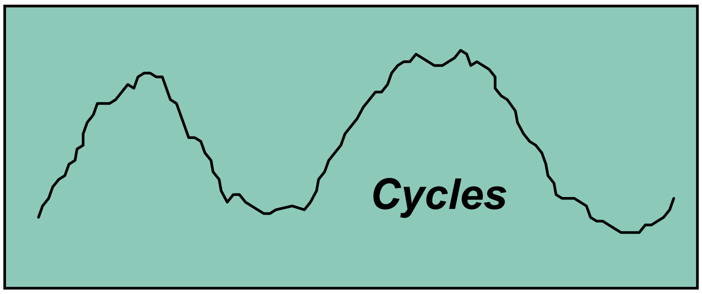


Fig.7-21 Example of time series data with cycles

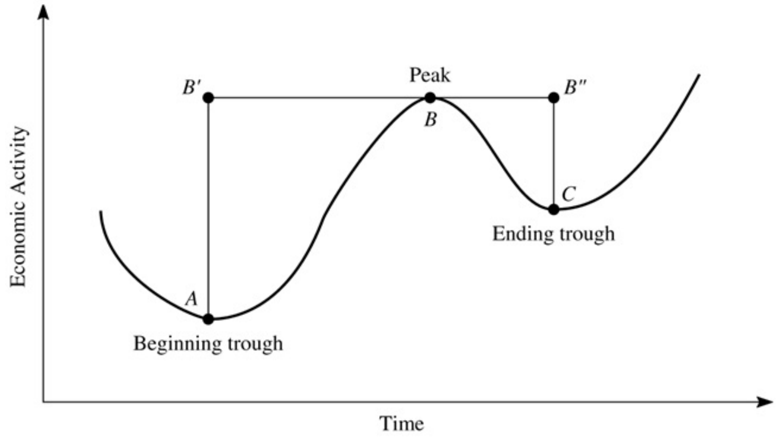


Fig.7-22 The General Business Cycle

There are different types of business cycles. The ***General Business Cycle*** goes through successive periods of expansion and contraction, expansion, contraction, and so on (see Fig.7-22). Almost all industries exhibit cyclicality to some extent but cyclical dynamics at the level of individual industries may present rather different patterns from those of the general business cycles. For example, while the fluctuations of many industries correlate with those in the aggregate economy, there were also many industries that are not sensitive to business cycles – such as the pharmaceutical, educational service, insurance carriers and public service industries. In fact, it was estimated that "in any one recession [during the 1980s and the 1990s] only 60% of all industrial sectors were actually in a downturn[9]."

The timing, duration, and amplitude of industry cycles can vary widely. Durable goods industries in the US are approximately three times more cyclical than nondurable-goods industries (Petersen & Strongin, 1996).

The timing of most individual economic time-series conforms only loosely, and in some series not at all, to that of the business cycle. In the long term, with yearly time-series the ***Seasonal Cycle*** usually tends to equal one. Thus, applying the multiplicative model (7-2) and assuming that annual data are seasonal variations free (as mentioned above), we can identify the ***Cyclical Component*** by eliminating the effects of the trend:

$$\mathbf{C_t = (T_t \times C_t \times E_t)/(T_t \times E_t)*\mathit{100} = Y_t/(T_t \times E_t)*\mathit{100}} \qquad (7\text{-}17)$$

To estimate the ***Cyclical Component,*** we use the ***Residual Method*** formula:

$$\boxed{C_t \times E_t = {}^{y_t}\!/_{T_t}} \qquad (7\text{-}18)$$

i.e. the residuals are treated as part of the cyclical component.

[9] ‘How was it for you?’ The Economist, 2001, Vol. 362: 4-6

For example, if the real value of $\boldsymbol{Y_t}$ in 2014 is 6.58 and the trend estimate for the same year is 7.69, then:

$$\mathbf{C_t = (6.58/7.69) \times 100 = 85.5} \qquad (7\text{-}19)$$

The ***cyclical index*** (7-19) shows the position of each $\mathbf{Y_t}$ value relative to the trend line. It is extremely difficult to estimate the ***cyclical index*** for more than a year or two into the future.

Economic indicators allow analysis of economic performance and predictions of future performance. One application of ***economic indicators*** is the study of ***business cycles***. ***Economic indicators*** include various indices, earnings reports, and economic summaries, such as unemployment rate, quits rate, housing starts, consumer price index (a measure for inflation), consumer leverage ratio, industrial production, bankruptcies, gross domestic product, retail sales, stock market prices, money supply changes and others (see Table 7.2).

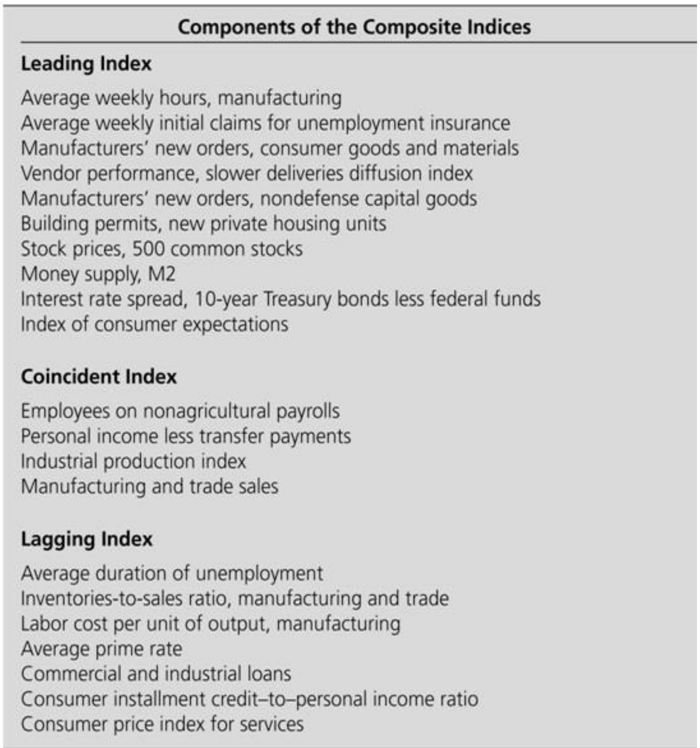
The leading business cycle dating committee in the United States of America is the National Bureau of Economic Research. The Bureau of Labor Statistics is the principal fact-finding agency for the U.S. government in the field of labor economics and statistics. Other producers of economic indicators include the United States Census Bureau and United States Bureau of Economic Analysis[10].

Economic indicators can be classified into three categories according to their usual timing in relation to the business cycle as presented in Table 7.2 and in Fig.7-23.

TABLE 7.2

U.S. Business Cycle Indicators

Source: The Conference Board (http://www.conference-board.org/economics/bci/component.cfm). Data in this table are from The Conference Board, which produces the U.S. Business Cycle Indicators.



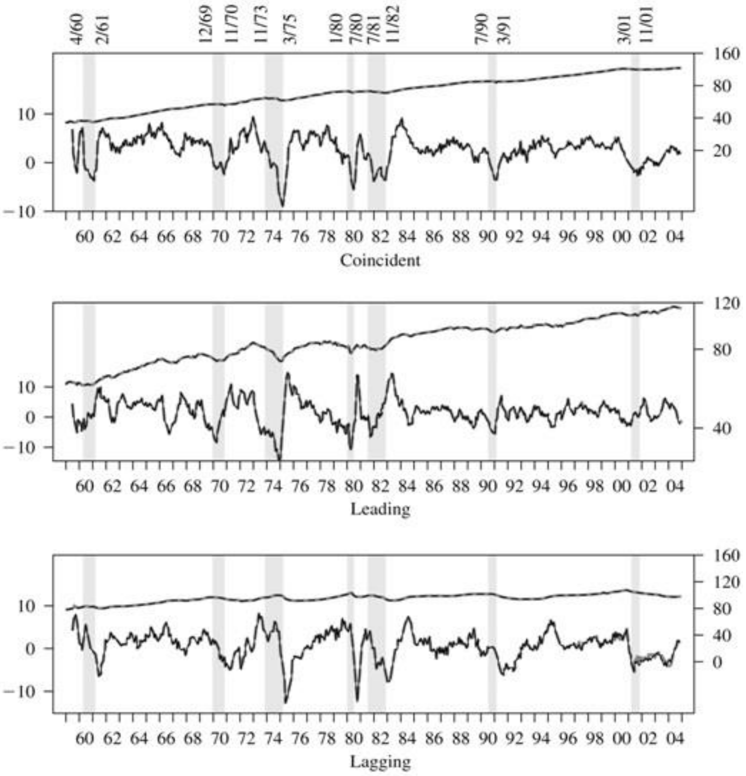
Components of the Composite Indices
Leading Index
Average weekly hours, manufacturing
Average weekly initial claims for unemployment insurance
Manufacturers' new orders, consumer goods and materials
Vendor performance, slower deliveries diffusion index
Manufacturers' new orders, nondefense capital goods
Building permits, new private housing units
Stock prices, 500 common stocks
Money supply, M2
Interest rate spread, 10-year Treasury bonds less federal funds
Index of consumer expectations
Coincident Index
Employees on nonagricultural payrolls
Personal income less transfer payments
Industrial production index
Manufacturing and trade sales
Lagging Index
Average duration of unemployment
Inventories-to-sales ratio, manufacturing and trade
Labor cost per unit of output, manufacturing
Average prime rate
Commercial and industrial loans
Consumer installment credit–to–personal income ratio
Consumer price index for services

[10] See http://en.wikipedia.org/wiki/Economic_indicator

Technically, economic indicators are business-related time series that are used to help access the general state of the economy, particularly with reference to the business cycle. ***Leading indicators*** are indicators that usually, but not always, change before the economy as a whole changes, i.e. they provide advance warning of probable changes in economic activity. These indicators move ahead of turns in the business cycle because decisions to expand take time to produce influences – hiring rates, construction contracts, etc. They are therefore useful as short-term predictors of the economy. Stock market returns are a leading indicator since the stock market usually begins to decline before the economy declines and usually begins to improve before the general economy begins to recover from a slump. Other leading indicators include the index of consumer expectations, building permits, and the money supply. The Conference Board publishes a composite Leading Economic Index consisting of ten indicators designed to predict activity in the U. S. economy six to nine months in future (see Table 7.2).

Coincident indicators change at approximately the same time as the whole economy, thereby providing information about the current state of the economy. There are four economic statistics comprising the ***Index of Coincident Economic Indicators*** as presented in Table 7.2, which reflect the current performance of the economy.

Official Business Cycles in the United States
For each graph the top line (right hand axis) represents the corresponding index, while the bottom line (left hand axis) is the six-month growth rate.



Source: *Business Cycle Indicators* 10, no. 1 (January 2005). The Conference Board (accessed from http://www.conference-board.org/economics/bci/methodology.cfm, May 21, 2005).

Fig.7-23 Official Business Cycles in USA

These indicators, such as Gross Domestic Product, industrial production, personal income, and retail sales, are comprehensive in coverage and tell us whether the economy is currently experiencing a recession or a slowdown, a boom or inflation. A coincident index may be used to identify, after the fact, the dates of peaks and troughs in the business cycle.

The ***Lagging Index*** tends to follow changes in the overall economy. ***Lagging indicators*** are indicators that usually change after the economy does. They confirm changes previously signaled and typically the lag is a few quarters of a year. The fluctuations of these series usually follow those of the coincident indicators – long-term unemployment, the yield on mortgage loans, etc. The unemployment rate is a lagging indicator since the employment tends to increase two or three quarters after an upturn in the general economy. In finance, Bollinger bands are one of the various lagging indicators in frequent use. In a performance measuring system, profit earned by a business is a lagging indicator as it reflects a historical performance. Similarly, improved customer satisfaction is the result of initiatives taken in the past. The ***Index of Lagging Indicators*** is published monthly by The Conference Board, a non-governmental organization, which determines the value of the index from seven components as listed in Table 7.2.

There are also three terms that describe an economic indicator's direction relative to the direction of the general economy:

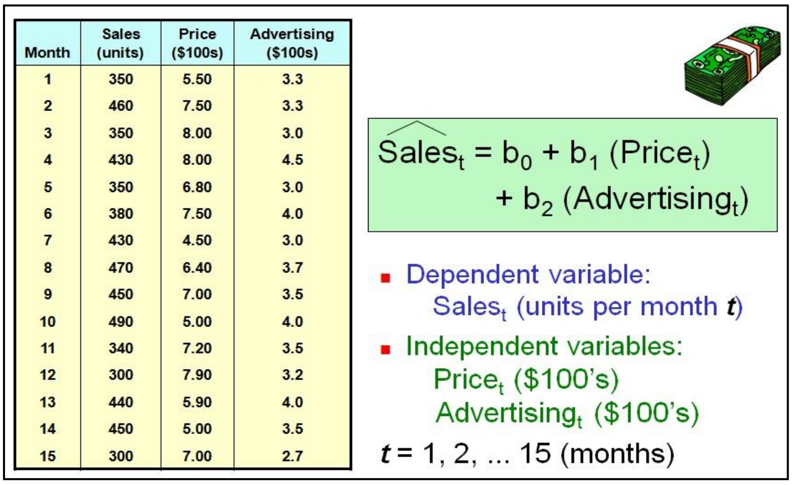
- ***Procyclical indicators*** move in the same direction as the general economy – they increase when the economy is doing well and decrease when it is doing badly. Gross domestic product (GDP) is an example of a procyclic indicator.

- ***Countercyclical indicators*** move in the opposite direction to the general economy. The unemployment rate is countercyclical since it rises when the economy is deteriorating.

- ***Acyclical indicators*** are those with little or no correlation to the business cycle – they may rise or fall when the general economy is doing well, and at the same time, they may rise or fall when the general economy is not doing well.

7.5. Regression with Time Series Data

The first techniques in ***Regression analysis*** were developed for cross-sectional data. The method of ***least squares*** grew out of the fields of astronomy and geodesy as scientists and mathematicians sought to provide solutions to the challenges of navigating the Earth's oceans during the Age of Exploration. The technique is described as an algebraic procedure for fitting linear equations to data and Legendre demonstrates in 1805 the new method by analyzing the same data as Laplace for the shape of the earth.



Month	Sales (units)	Price ($100s)	Advertising ($100s)
1	350	5.50	3.3
2	460	7.50	3.3
3	350	8.00	3.0
4	430	8.00	4.5
5	350	6.80	3.0
6	380	7.50	4.0
7	430	4.50	3.0
8	470	6.40	3.7
9	450	7.00	3.5
10	490	5.00	4.0
11	340	7.20	3.5
12	300	7.90	3.2
13	440	5.90	4.0
14	450	5.00	3.5
15	300	7.00	2.7

$$\widehat{Sales}_t = b_0 + b_1 (Price_t) + b_2 (Advertising_t)$$

- Dependent variable: $Sales_t$ (units per month t)
- Independent variables: $Price_t$ ($100's) $Advertising_t$ ($100's)

t = 1, 2, ... 15 (months)

Fig.7-24 Regression with time series data example

Today, ***Regression analysis*** techniques are very often used in ***Time series analysis*** and ***Forecasting.*** Making predictions in such case involve some complications, due to unknown values of the predictor variables and/or existing autocorrelation between time series data. There are a few issues that arise with time series data but not with cross-sectional data that we will consider in this section.

A. Unknown Predictors in Time Series Forecasting

Using regression techniques to model and forecast the trend in time series data (already discussed in this Chapter) by including the time $\boldsymbol{t} = \{\boldsymbol{1, 2, \dots T}\}$ as a predictor variable does not give rise to any problem, since $\boldsymbol{T}$ is the index set. Applying ***OLS*** to equation (7-8) we can fit the trend line minimizing the sum of the squared errors and then we can use the coefficient estimates in equation (7-10) to make predictions (see Fig.7-5).

In more complex regression models one or more real business variables are used as predictors (see Fig.7-24). Using a regression model to forecast time series data in such case poses a challenge in that future values of the predictor variable are needed to be input into the estimated model, but these are not known in advance.

One possible solution to this problem is to use "***scenario-based forecasting***", introduced and discussed in Chapter 4.2, Section F. Scenario Writing. Forecast intervals for scenario-based forecasts do not include the uncertainty associated with the future values of the predictor variables. They assume the value of the predictor is known in advance and a minimum (pesimistic) and a maximum (optimistic) predictions are computed instead.

An alternative approach is to use genuine forecasts for the predictor variable. For example, a pure time series-based approach (as discussed above) can be used to generate forecasts for the predictor variable or forecasts published by some other source such as a government agency can be used.

When using regression models with time series data, we need to distinguish between two different types of forecasts that can be produced, depending on what is assumed to be known when the forecasts are computed.

Ex-ante forecasts are made using only the information that is available in advance. For example, ex-ante forecasts of sales for the four quarters in 2019 should only use information that was available before 2019. These are the only genuine forecasts made in advance using whatever information is available at the time.

Ex-post forecasts are made using later information on the predictors. For example, ex-post forecasts of sales for each of the 2019 quarters may use the actual observations of demand for each of these quarters, once these have been observed. These are not genuine forecasts but are useful for studying the behavior of forecasting models.

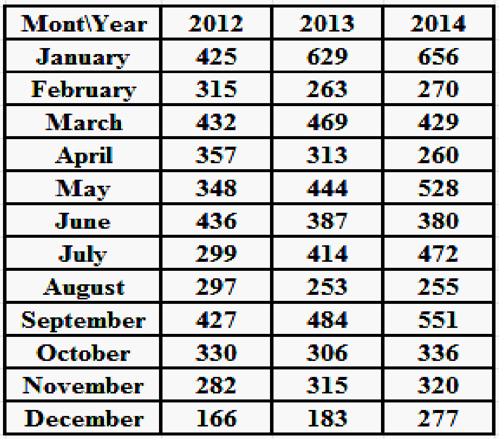
The model from which ex-post forecasts are produced should not be estimated using data from the forecast period. That is, ex-post forecasts can assume knowledge of the predictor variable (the ***x*** variable) but should not assume knowledge of the data that are to be forecast (the ***y*** variable).

A comparative evaluation of ***ex-ante*** forecasts and ***ex-post*** forecasts can help to separate out the sources of forecast uncertainty. This will show whether forecast errors have arisen due to poor forecasts of the predictors or due to a poor forecasting model.

B. Regression Analysis with Dummy Seasonal Variables

In regression analysis techniques such as ***OLS***, with a seasonally varying dependent variable being influenced by one or more explanatory variables, the seasonality can be accounted for and measured by including dummy variables, as discussed in Chapter 6.3 Section D. Technically, we need (***n-1***) ***dummy variables***, one for each of the seasons except for an arbitrarily chosen reference season, where ***n*** is the number of seasons (e.g., 4 in the case of meteorological seasons, 12 in the case of months, etc.).

Table 7.3 Monthly Sales for 2012-2014



Mont\Year	2012	2013	2014
January	425	629	656
February	315	263	270
March	432	469	429
April	357	313	260
May	348	444	528
June	436	387	380
July	299	414	472
August	297	253	255
September	427	484	551
October	330	306	336
November	282	315	320
December	166	183	277

Each ***dummy variable*** is set to "**+1**" if the data point is drawn from the dummy's specified season and "**0**" otherwise. Then the predicted value of the dependent variable for the reference season is computed from the rest of the regression, while for any other season it is computed using the rest of the regression and by inserting the value **"+1"** for the dummy variable for that season.

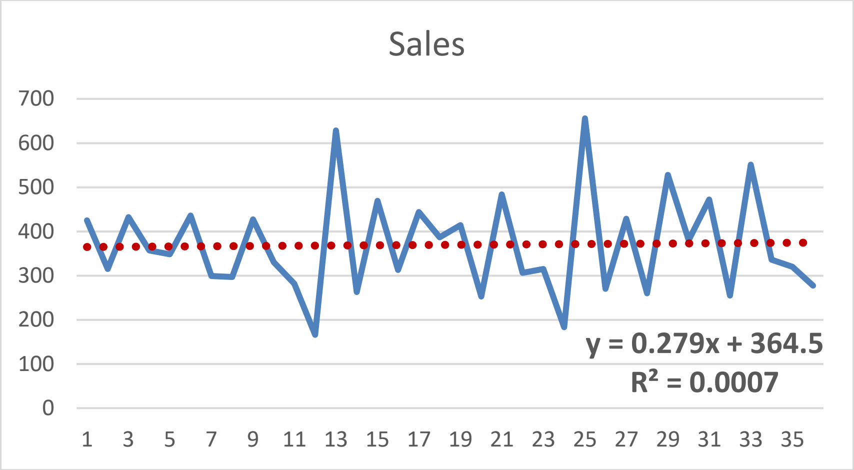
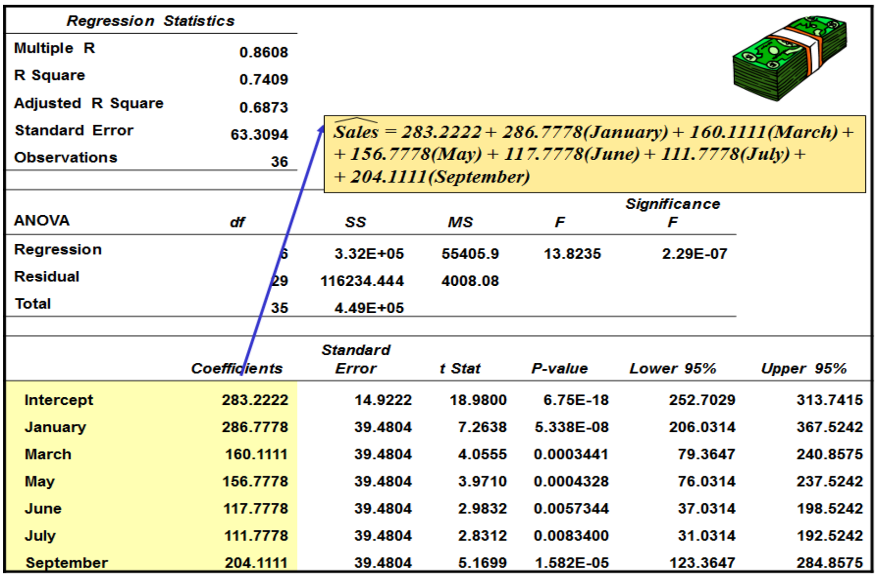


Fig.7-25 Time series plot of company sales and linear trend equation

Example: Suppose that a company manager wants to analyze the seasonal factor in company's monthly sales for the last three years, as presented in Table 7.3. He knows that the seasonality can be accounted for and measured by including dummy variables and he decided to perform regression analysis on dependent variable Sales with eleven explanatory variables – January, February and so on, up to November. Each regressor is coded as **"+1"** for the particular month and "**0**" otherwise.

Time series plot and the linear trend model in Fig.7-25 show that the process is ***stationary*** and there is no need to add a trend component into the model.

The output from ***OLS*** estimation, done by MS Excel regression analysis is presented in Fig.7-26. After conducting a series of test diagnostics, the manager identified a significant model (at 0.01 level of significance) with predictors January, March, May, June, July and September, which explained about 68% of the variation (mostly seasonal) in company Sales.



Regression Statistics	
Multiple R	0.8608
R Square	0.7409
Adjusted R Square	0.6873
Standard Error	63.3094
Observations	36

$$\widehat{Sales} = 283.2222 + 286.7778(January) + 160.1111(March) + 156.7778(May) + 117.7778(June) + 111.7778(July) + 204.1111(September)$$

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	6	3.32E+05	55405.9	13.8235	2.29E-07
Residual	29	116234.444	4008.08		
Total	35	4.49E+05			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	283.2222	14.9222	18.9800	6.75E-18	252.7029	313.7415
January	286.7778	39.4804	7.2638	5.338E-08	206.0314	367.5242
March	160.1111	39.4804	4.0555	0.0003441	79.3647	240.8575
May	156.7778	39.4804	3.9710	0.0004328	76.0314	237.5242
June	117.7778	39.4804	2.9832	0.0057344	37.0314	198.5242
July	111.7778	39.4804	2.8312	0.0083400	31.0314	192.5242
September	204.1111	39.4804	5.1699	1.582E-05	123.3647	284.8575

Fig.7-26 Regression output for Sales and the significant dummy regressors for each month

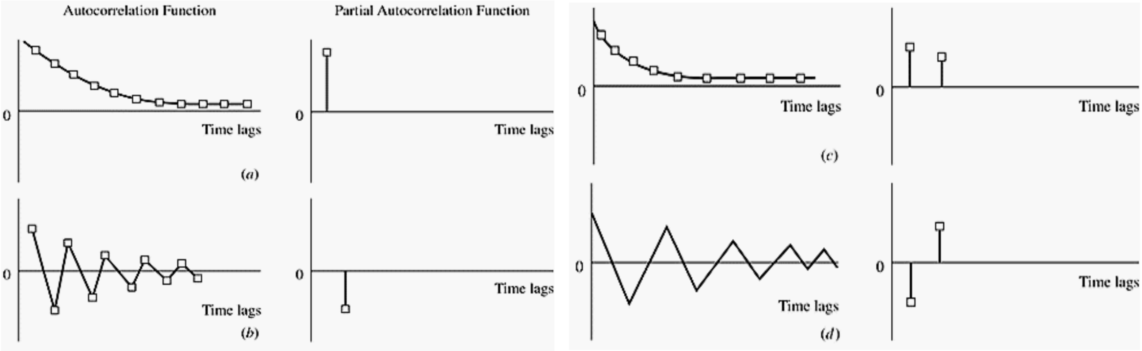


Fig.7-27 Examples of autocorrelation functions (***ACF***) with single and double-time lags

C. Residual Autocorrelation

With time series data it is highly likely that the value of a variable observed in the current time period will be influenced by its value in the previous period, or even the period before that, and so on (see Fig.7-27). Therefore, when fitting a regression model to time series data, it is very common to find ***autocorrelation in the residuals***. In such a case, the estimated model violates the assumption that errors are random and independent, and the forecasts may be inefficient – there is some information left over which should be utilized in order to obtain better forecasts.

When ***autocorrelation*** exists (i.e. successive observations over time are related to one another – see Fig.7-28) there are technical problems that arise, including:

- The ***Standard Error*** (***S_ε***) of the estimate can seriously underestimate the variability of the error terms.
- The usual inferences based on ***t*** and ***F*** statistics are no longer strictly applicable – both ***t*** and ***F***-scores are overestimated when the autocorrelation is positive at low lags.
- The ***standard errors of the regression coefficients*** (***S_{bj}***) underestimate the variability of the estimated regression coefficients.
- ***Spurious regression*** can result (see section D below).

The forecasts from a model with ***autocorrelated errors*** are still unbiased, and so are not “wrong”, but they will usually have larger prediction intervals than they must. Problematic autocorrelation of the errors, which themselves are unobserved, can generally be detected because it produces autocorrelation in the observable residuals. Since autocorrelation violates the ***OLS*** assumption that the error terms are uncorrelated, the Gauss Markov theorem does not apply, and the ***OLS*** estimators are no longer the ***Best Linear Unbiased Estimators (BLUE).***

- Here, residuals show a cyclic pattern, not random

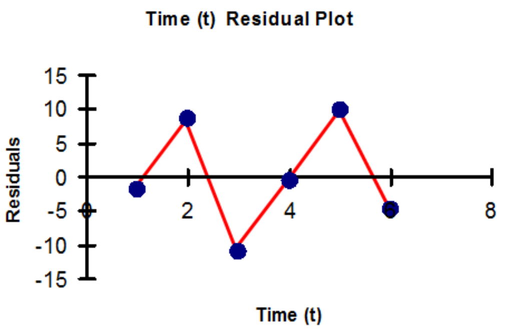


Fig.7-28 Correlation of the error terms (the residuals) over time

The traditional test for the presence of first-order autocorrelation is the Durbin–Watson statistic ***d*** (introduced and used already in Chapter 6). Durbin and Watson (1950, 1951) applied this statistic to the residuals from least squares regressions, and developed bounds tests for the null hypothesis that the errors are serially uncorrelated against the alternative that they follow a first-order autoregressive process:

H_0: $\rho = 0$ (residuals are not correlated)

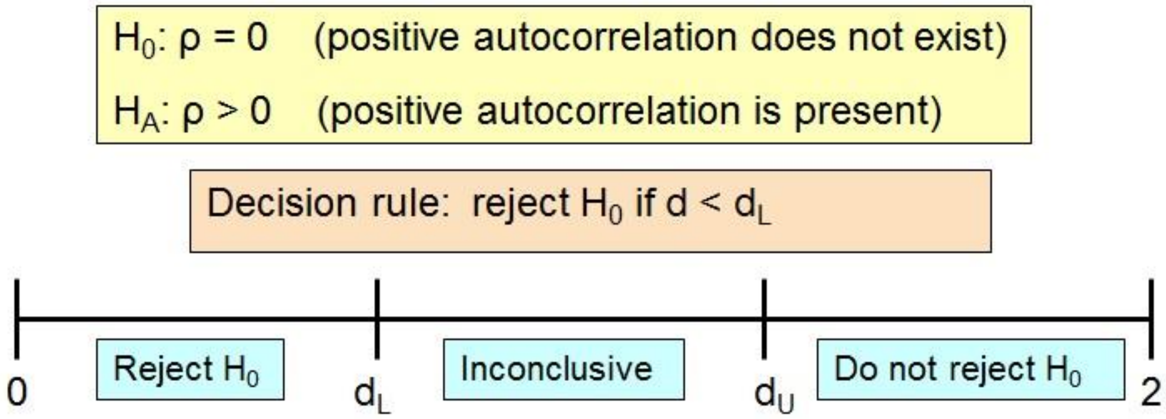
H_A: $\rho \neq 0$ (autocorrelation is present)

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2} \qquad (7\text{-}20)$$

where $\boldsymbol{e_t}$ are the residuals and ***t*** = {***1, 2, … T***}, where ***T*** is the index set, is the number of observations, and the value of **d** always lies between 0 and 4.

Interestingly, the ***Durbin-Watson statistic*** (**d**) and ***Pearson's correlation coefficient*** (**r**) look very different, in fact, if we have a lengthy sample, these statistics are closely related when applied to measure the first-order autocorrelation. We can easily and linearly, map back and forth between these two mathematical techniques for serial correlation – first we multiply equation (7-3) by -2 (which makes the (**r**) range from -2 to +2), then we add 2 and the new range for (**r**) is now from 0 to positive 4. Thus, we can find (see the calculations provided in this Chapter Exercises) that **d** approximately equals **2(1 − r)**.

Which technique one would employ to measure the correlation between the time-series data and their lags in practice depends slightly on how one's data is, how one wants to describe the autocorrelation, what statistical software is available, and what level of down-stream analysis may also need to be performed.


H_0: ρ = 0 (positive autocorrelation does not exist)

H_A: ρ > 0 (positive autocorrelation is present)

Decision rule: reject H_0 if d < d_L

Fig.7-29 Durbin–Watson test for positive first-order autocorrelation

Since the value of the Durbin–Watson statistic always range between 0 and 4, then **d=2** means **"no autocorrelation"**. If **d** is substantially less than 2, there is evidence of positive serial correlation, which means that a positive error for one observation increases the chances of a positive error for another observation. Small values of **d** indicate that successive error terms are, on average, close in value to one another, i.e. positively correlated. As a rough rule of thumb, if the Durbin–Watson statistic is less than 1.0, there may be cause for alarm.

The statistical test for autocorrelation can be done, depending on how the alternative hypothesis is formulated. To test for ***positive autocorrelation*** (see Fig. 7-29) we should conduct a one-tailed test against "H_A: Positive serial correlation exists (e_t is directly related to e_{t-1})".

The decision is made, at given significance level (***α***), comparing the test statistic ***d*** with the lower and upper critical values (**d_L** and **d_U**). These critical values vary by the level of significance (***α***), the number of observations (***T***), and the number of predictors (***k***) in the regression equation. Their derivation is complex and, as mentioned earlier in Chapter 6, users typically obtain them from Tables in appendices of statistical texts or online. Thus:

- If **d** < **d_L** (given α), then we reject the null hypothesis and accept the H_A, which means that there is statistical evidence that the error terms are positively autocorrelated.
- If **d** > **d_U** (given α), there is no statistical evidence that the error terms are positively autocorrelated and we cannot reject the null hypothesis "H_0: No serial correlation."
- If **d_L** < **d** < **d_U** (given α), the test is inconclusive, and we cannot make a decision.

When **d** > **2** successive error terms are, on average, much different in value from one another, i.e., negatively correlated. Negative serial correlation implies that a positive error for one observation increases the chance of a negative error for another observation and a negative error for one observation increases the chances of a positive error for another.

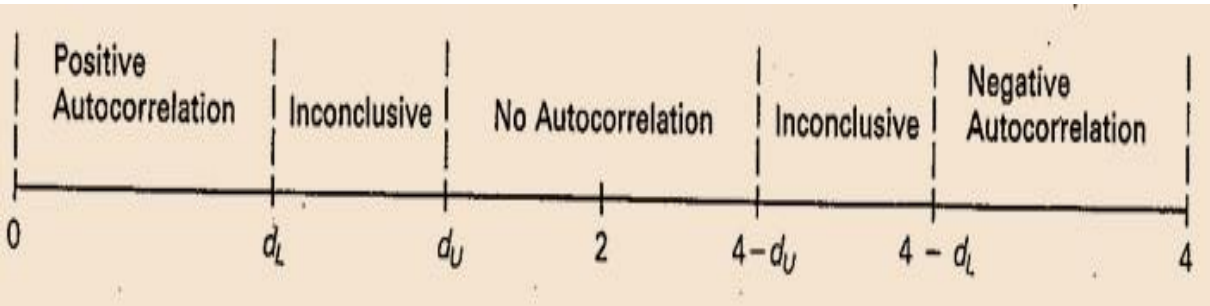


Fig.7-30 Critical values and regions for Durbin–Watson Test statistic

To test for ***negative autocorrelation*** (see Fig.7-30) at given significance level (**α**), we apply a procedure like the above one, using a test statistic (**4−*d***). Then:

- If **(4-d) < d$_L$** (given α), which means that **d** value is close to 4, there is statistical evidence that the error terms are negatively autocorrelated.
- If **(4-d) > d$_U$** (given α), which means that **d** value is larger than, but close to 2, there is **no** statistical evidence that the error terms are negatively autocorrelated.
- If **d$_L$ < (4-d) < d$_U$** (given α), the test is inconclusive, and we cannot make a decision.

Unfortunately, ***d*** is biased if the explanatory variables include a lagged dependent variable (i.e. for autoregressive moving average models, as explained later in section 8.1.) so that autocorrelation is underestimated. But for large samples, we can easily compute the unbiased normally distributed Durbin's ***h*** statistic using the ***d*** statistic and the estimated variance of the regression coefficient of the lagged dependent variable[11].

A more flexible test, covering autocorrelation of higher orders in the errors in a regression model, is the ***serial correlation test*** and its modifications, where the null hypothesis is that there is no serial correlation of any order. The test is more general than the Durbin–Watson statistics ***d*** (which is only valid for non-stochastic regressors) and ***h***, which is for testing the possibility of a serial correlation in residuals in first-order autoregressive models.

Although ***serial correlation*** does not affect the consistency of the estimated regression coefficients, it does affect our ability to conduct valid statistical tests. *First,* the **F**-statistic to test for overall significance of the regression may be inflated under positive serial correlation because the mean squared error will tend to underestimate the population error variance. *Second*, positive serial correlation typically causes the ***OLS*** standard errors for the regression coefficients to underestimate the true standard errors. Therefore, if a positive serial correlation is present in the regression, standard linear regression analysis will typically lead us to compute artificially small standard errors for the regression coefficient.

[11] See http://en.wikipedia.org/wiki/Durbin%E2%80%93Watson_statistic

These small standard errors will cause the estimated **t**-statistic to be inflated, suggesting significance where perhaps there is none. The inflated **t**-statistic, may in turn, lead us to incorrectly reject null hypotheses, about population values of the parameters of the regression model more often than we would if the standard errors were correctly estimated.

In regression analysis using time series data, ***autocorrelation*** in a variable of interest is typically modeled either with an ***autoregressive model (AR)***, a ***moving average model (MA)***, their combination as an ***autoregressive moving average model (ARMA)***, or an extension of the latter called an ***autoregressive integrated moving average model (ARIMA)***. Chapter 8 discusses all these models.

D. Spurious regression

Spurious correlation is a term coined by Karl Pearson (1897) to describe the correlation between ratios of absolute measurements that arises because of using ratios, rather than because of any actual correlations between the measurements. The phenomenon of spurious correlation is one of the main motives for the field of compositional data analysis[12], which deals with the analysis of variables that carry only relative information, such as proportions, percentages, and parts-per-million.

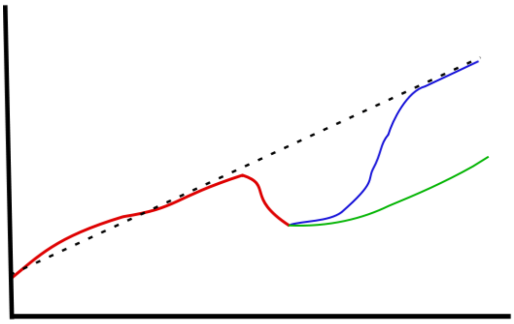
Most ***time-series*** have ***trends***, either deterministic or stochastic. The ***R-squared*** statistic used in assessing adequacy of regressions gives substantially misleading results for ***time-series*** with ***trends***. A simple verification for this is to pick any consumption series for any country and regress it against GNP for some other, dissimilar country (for example, Belgium and Argentina). There will be (unless we are unlucky) a strong correlation, and a regression with very high ***R-squared*** will result. This is known as ***spurious regression*** – even though there is no association between the two ***time-series***, the regression results suggest that there is a strong relationship.

Spurious regression refers to a regression that shows significant results due to the presence of a ***unit root***[13] in both (dependent and explanatory) variables. A linear stochastic process has a ***unit root*** if **"1"** is the root of the process's characteristic equation. Such a process is non-stationary. If the other roots of the characteristic equation have an absolute value less than one, then the first difference of the process will be stationary.

[12] In statistics, ***compositional data*** are quantitative descriptions of the parts of some whole, conveying exclusively relative information.

[13] A ***unit root*** is a feature of processes that evolve through time that can cause problems in statistical inference involving time series models.

Fig.7-31 Output with potential unit root



This diagram depicts an example of a potential unit root. The red line represents an observed drop in output. Green shows the path of recovery if the series has a unit root. Blue shows the recovery if there is no unit root and the series is trend stationary. The blue line returns to meet and follow the dashed trend line while the green line remains permanently below the trend. The unit root hypothesis also holds that a spike in output will lead to levels of output higher than the past trend[14].

Use of ***OLS*** to estimate the slope coefficients of the autoregressive model relies on the stochastic process being stationary. If the stochastic process is non-stationary, the use of ***OLS*** can produce invalid estimates. Granger and Newbold (1974) called such estimates spurious regression results high $\boldsymbol{R^2}$ values and high $\boldsymbol{t}$-ratios yielding results with no economic meaning.

To estimate the slope coefficients, we should first conduct a unit root test, whose null hypothesis is that a unit root is present. If that hypothesis is rejected, we can use ***OLS***. However, if the presence of a unit root is not rejected, then we should apply the difference operator to the series (see Fig.5-10). If another unit root test shows the ***differenced time-series*** to be stationary, ***OLS*** can then be applied to this series to estimate the slope coefficients.

For example, in first-order autoregression case $\Delta y_t = y_t - y_{t-1} = \varepsilon_t$ is stationary.

In the second-order autoregressive model:

$$\boxed{y_t = \beta_1 y_{t-1} + \beta_2 y_{t-2} + \varepsilon_t} \tag{7-21}$$

Equation (7-21) can be written as:

$$(1 - \lambda_1 L)(1 - \lambda_2 L)y_t = \varepsilon_t \tag{7-22}$$

where **L** is a ***lag operator*** that decreases the time index of a variable by one period.

In ***Time Series Analysis (TSA)***, the ***lag operator*** **(L)** or ***backshift operator*** **(B)** operates on an element of a time-series to produce the previous element. For example:

$$Ly_t = y_{t-1} \text{ for all } t > 1 \text{ or equivalently } y_t = Ly_{t+1} \text{ for all } t \geq 1$$

Note that the ***lag operator L*** can be raised to arbitrary integer powers so that:

$$L^{-1}y_t = y_{t+1} \text{ and } L^k y_t = y_{t-1}$$

[14] See http://en.wikipedia.org/wiki/Unit_root

If $\lambda_2=1$, the model (7-22) has a unit root and we can define $z_t=\Delta y_t$. Then $z_t = \lambda_1 z_{t-1} + \varepsilon_t$ is stationary if $|\lambda_1|<1$ and ***OLS*** can then be used to estimate the slope coefficient λ_1.

If the process has multiple unit roots, the difference operator (7-22) can be applied multiple times, as necessary.

Economists debate whether various economic statistics, especially output, have a unit root or are trend stationary. A unit root process with ***drift*** (see Chapter 5.1 B) is given in the first-order case by equation:

$$\boxed{y_t = y_{t-1} + b + \varepsilon_t} \tag{7-23}$$

where **b** is a constant term referred to as the ***drift*** term, and ε_t is white noise.

Any non-zero value of the noise term, occurring for only one period, will permanently affect the value of y_t as shown in Fig.7-30, so deviations from the line $y_t = a + b.t$ are non-stationary and there is no reversion to any trend line.

In contrast, a ***trend-stationary*** process is given by

$$\boxed{y_t = k.t + \upsilon_t} \tag{7-24}$$

where k is the slope of the trend and

υ_t is noise (white noise in the simplest case, or more generally, noise following its own stationary autoregressive process).

Here any transient noise will not alter the long-run tendency for y_t to be on the trend line (as also shown in Fig.7-31). This process is said to be ***trend-stationary*** because deviations from the trend line are stationary.

The issue of ***spurious regression*** is particularly popular in the literature on business cycles. Research on the subject began with Nelson and Plosser (1982) whose paper on GNP and other output aggregates failed to reject the unit root hypothesis for these series. While the literature on the unit root hypothesis may consist of arcane debate on statistical methods, the hypothesis carries significant practical implications for economic forecasts and policies. It should be noted, however, that cases of ***spurious regression*** might appear to give reasonable short-term forecasts, but they will generally not continue to work into the future.

E. Cross-validation with Time Series data

There are two possible versions – in the first one, we select a testing set of ***m*** observations (about 20%-30% of the total sample), based on how big the sample is and how far ahead we want to forecast. It should be at least as large as the maximum forecast horizon required, but no more than 30% of the total number of observations. The testing data set could be selected in a different way depending on the model purpose (Madala & Ivakhnenko, 1994), for example to use the last ***m*** observations, or by random selection and so forth.

In the second version, the ***cross-validation*** is similar to the procedure with cross-sectional data, described in Chapter 3.2, but for time series data the training set consists only of observations that occurred prior to the observation that forms the testing set (Mueller & Lemke, 2003). Thus, no future observations can be used in constructing the forecast. However, since it is not possible to get a reliable forecast based on a very small training set, the earliest observations are not considered as testing sets.

Suppose ***n*** observations are required to produce a reliable forecast. Then the ***cross-validation*** process works as follows:

- We select the observation at time (***n+i***) for the testing set and use the observations at times ***t= {1, 2, … (n+i-1)}*** to estimate the forecasting model. Then we compute the error on the forecast for the time (***n+i*).**
- The above step should be done for all ***i= {1, 2, … (T-n)}***, where ***T*** is the total number of observations and the forecast error should be measured on each (***n+i***) period accordingly.
- In the end, we compute the forecast accuracy measures based on all errors obtained.

This procedure is sometimes known as a "***rolling forecasting origin***" because the "origin" (***n+i-1***) at which the forecast is based, rolls forward in time.

With ***Time Series Forecasting***, one-step-ahead forecasts may not be as relevant as multi-step forecasts. In such case, the above ***cross-validation*** procedure based on a ***rolling forecasting origin,*** should be modified to allow multi-step errors to be used. Suppose we are interested in models that produce good ***l-step-ahead forecasts***. The procedure is as follows:

1. Select the observation at time (***n+l+i***) (where ***(n+l)<T, i= {0, 1, 2, … (T-n-l)}*** and ***T*** is the total number of observations) for the testing set, and use the observations at times ***t= {1, 2, … (n+i)}*** to estimate the forecasting model. At each step compute the ***l-step-error*** on the forecast for time (***n+l+i***).
2. In the end, compute the forecast accuracy measures based on all errors obtained.

Note that when ***l=1*** this procedure is identical with the one-step ***rolling forecasting origin***.

***

SUMMARY AND CONCLUSIONS

A ***time series*** is a sequence of data points, measured typically at successive points in time spaced at uniform time intervals. A few basic techniques of ***Time series analysis*** and ***Forecasting*** were introduced and explored in Chapter 5**.** Chapters 7 and 8 discuss more advanced predictive methods, emphasizing on some important models and their application in contempory business analysis and management.

- ***Time series analysis*** (***TSA***) comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. ***TSA*** methods are divided as: frequency-domain (spectral analysis) and time-domain methods (auto-correlation analysis); parametric and non-parametric; linear and non-linear methods.
- ***Time series models*** are used for predicting the future behavior of variables based on previously observed data. These models account for the fact that data points taken over time may have an internal structure (such as autocorrelation, trend or seasonal variation) that should be accounted for.

Time series forecasting is the use of a model to predict future values based on previously observed data. Different approaches and techniques are used:

A) ***Time Series Decomposition*** seeks to construct, from an observed time series, a number of component series (that could be used to reconstruct the original by additions or multiplications) where each of these has a certain characteristic or type of behaviour:

- ***Trend Component*** ($\mathbf{T_t}$) is the long-run increase or decrease over time (overall upward or downward movement) that reflects the long-term progression of the series.
- ***Cyclical Component*** ($\mathbf{C_t}$) describes repeated but non-periodic fluctuations with duration of at least 2 years. In business, these fluctuations are wavelike variations of more than one year’s duration due to changing economic conditions.
- ***Seasonal Component*** ($\mathbf{S_t}$) reflecting seasonality (seasonal variation) – short-term regular variations when a time series is influenced by seasonal factors (the quarter of the year, the month, or day of the week).
- ***Random Component***, the ***Error***, or ***Irregular Component*** ($\mathbf{E_t}$) represents the residuals of the time series after all other components have been removed. These are unpredictable, random fluctuations due to random variations (also known as the “***Noise***” in time series). It is important to distinguish between the ***Random variations*** (caused by chance) and the ***Irregular variations*** (caused by unusual circumstances).

B) Index Numbers – an index is a statistical measure of changes in a representative sample of observations. These data may be derived from any number of sources, including company performance, prices, productivity, and employment. An index number is an economic data figure reflecting price or quantity compared with a standard or base value. The best-known index numbers are the ***consumer price index (CPI),*** Producer Price Index, Stock Market Indexes, such as Dow Jones Industrial Average, S&P 500 Index, NASDAQ Index, and others.

C) Trend Estimation and Forecasting – ***Trend estimation*** can be used to make and justify statements about tendencies in time-series, by relating the measurements to the times at which they occurred:

- ***Linear trend*** has the advantage of being simple and does not require a control group, experimental design, or another sophisticated analysis technique. However, it suffers from a lack of scientific validity when other potential changes can affect the data.
- A ***trend line*** could simply be drawn by eye through a set of data points, but more properly their position and slope is calculated using estimation techniques like ***OLS***.

D) Seasonal Component Estimation and Forecasting – ***Seasonal variation*** is a component of a time series which is defined as the repetitive and predictable movement around the trend line in one year or less. It is detected by measuring the quantity of interest for small time intervals, such as days, weeks, months or quarters:

- The ***seasonal index*** measures how much the average for a particular period tends to be above (or below) the expected value.
- The measurement of seasonal variation by using the ***ratio-to-moving average*** method provides an index to measure the degree of the seasonal variation in a time series.
- If the seasonal component is removed from the original observations, the resulting values are ***seasonally adjusted***. Technically, ***Seasonal Adjustment*** means to calculate the seasonal indexes and use them to ***deseasonalize*** the original data.
- In ***Time Series Forecasting*** the ***seasonal adjustment*** means to incorporate seasonality into the Forecast, i.e. to adjust the forecast to the seasonal changes.

E) Cyclical Component Estimation and Forecasting – The ***Cyclical Component*** **C_t** describes repeated but non-periodic fluctuations in time-series, e.g. wavelike variations of more than one year's duration due to changing economic conditions.

- The ***General Business Cycle*** goes through successive periods of expansion, contraction, expansion, contraction, and so on.

- ***Economic indicators*** (such as ***Leading indicators, Coincident indicators, Lagging indicators***) allow analysis and predictions of future economic performance.

F) Regression with Time Series Data – using ***Regression analysis*** in ***Time series analysis*** and ***Forecasting*** involve some complications, due to unknown values of the predictor variables and/or existing autocorrelation between time series data:

- ***Unknown predictors in Time Series Forecasting*** **–** in complex regression models one or more real business variables are used as predictors and in such case to forecast time series data poses a challenge in that future values of the predictor variable are needed to be input into the estimated model, but these are not known in advance.
- ***Regression analysis with dummy seasonal variables*** **–** in regression analysis with a seasonally varying dependent variable, the seasonality can be accounted for and measured by including dummy variables.
- ***Residual autocorrelation*** **–** when fitting a regression model to time series data, it is very common to find ***autocorrelation in the residuals*** (i.e. the value of a variable observed in the current time period will be influenced by its value in the previous period, or even the period before that) and there are technical problems that arise.
- The traditional test for the presence of first-order autocorrelation is the ***Durbin–Watson*** statistic (***d***) and/or the ***serial correlation test*** and its modifications.
- ***Spurious regression and correlation*** – the correlation between ratios of absolute measurements that arises as a consequence of using ratios, rather than because of any actual correlations between the measurements. ***Spurious regression*** refers to a regression that shows significant results due to the presence of a ***unit root*** in both (dependent and explanatory) variables (i.e. such process is non-stationary).
- When ***differenced time-series (first-*** or ***higher order)*** are stationary, ***OLS*** can then be applied to this series to estimate the slope coefficients.
- ***Lag operators*** or ***backshift operators*** are used to describe non-stationary process.
- ***Cross-validation with Time Series data*** – the ***cross-validation*** is similar to the procedure with cross-sectional data, but here the training set consists only of observations that occurred prior to the observation that forms the testing set.
- One-step ahead forecasts use a ***rolling forecasting origin*** approach.
- When one-step ahead forecast is not relevant the ***cross-validation*** procedure based on a ***rolling forecasting origin*** should be modified to allow multi-step errors to be used.

KEY TERMS

CHAPTER EXERCISES

Conceptual Questions:

1. What is Time Series Decomposition and how does it work? Discuss.
2. List all steps in the Ratio-to-moving average technique and explain how seasonal adjustment of the trend forecast works.
3. What are the major issues in Regression analysis with Time Series Data? List and discuss at least three.
4. What are the steps in Cross-validation with Time Series data? List and discuss the differences with cross-section data.

Business Applications (continue from Chapter 5):

Open file Sales Data and continue product “A” Time series analysis”:

- Create formulas to compute Ratio-to-moving average to estimate the seasonal component.
- Remove the seasonal component from the original observations and provide the seasonally adjusted data for further analysis.
- Perform Regression analysis on smoothed (deseasonalized) data from previous step.
- Compute trend forecast for the next 12 months.
- Add seasonality to the trend estimates using the general equation (7-16).
- Compute the residuals for the basic period and the forecast errors for the 12-month period predictions.
- Design formulas, like the formulas in Part 4 of the Integrative case and compute MFE, MAD, MAPE, MPE, MSE and CV(RMSE) for the new model, for a validating dataset of the 12 new monthly forecasts.

Compare the accuracy achieved using the Ratio-to-moving average technique with the forecast errors of the models in Chapter 5 (Business Applications). Summarize and analyze all findings and write a short report (up to two pages) discussing your answers.

INTEGRATIVE CASE

HEALTHY FOOD SYPPLY CHAIN & STORES

Part 7: Time Series Analysis and Forecasting

In Chapter 1 we introduced *Healthy Food Stores* – a fast-growing retail food provider with 12 stores in a northwestern state. The company executives decided to study the effect that company advertising dollars have on sales. They hoped that examining collected historical data would reveal relationships that would help determine future advertising expenditures and predict monthly sales volumes for the upcoming quarter.

After identifying basic parameters, along with input (independent) and output (dependent) variables of the forecasting scenario in Part 2, the research team determined the main elements of the forecasting process:

- Forecasting horizon of up to twelve months;
- Quarterly forecast updates, since accuracy decreases as time horizon increases, and sufficient time is needed to implement possible changes;
- Development of different models based on data patterns, if any, and evaluation of their accuracy in order to select the most appropriate one;
- Selection of the best forecasting model with no more than 5% forecasting error.

The benchmark forecast computed in Part 3 (using the baseline of ***one-step naive forecast*** as a reference forecast) provided basic values for the most common measures of accuracy MFE, MAD, MAPE, MPE, MSE and CV(RMSE).

In Part 4, information about the opinion of some important people from the *Healthy Food Stores Company,* concerning this specific case, was collected and the research team applied the Delphi method to top executives group, Sales-force composite to the sales managers from all 12 stores and Scenario writing to the most experienced professionals from Advertising Department. After collecting such valuable information from different sources, in Part 5 the research team made its first steps in Numerical Predictions by developing different basic forecasting models. They created spreadsheets for Naïve techniques (Average model, Random Walk with Drift and Seasonal Naïve Technique), simple Moving Average, Simple Exponential Smoothing (SES) and Triple (Holt-Winters) Exponential Smoothing (TES), which were used to expand the base-line of one-step naïve forecast as reference forecasts.

In Part 6 the research team analyzed the relationships between Sales and all available predictors. After performing multiple correlation and regression analysis, researchers developed reliable forecasting model representing the real system with certain error.

In Part 7, the forecasting model will be expanded and probably improved by some additional Time series analyses and predictive techniques.

Case Questions

1. Open file Data.xslx. Following the steps, given in the Business Applications from the Chapter exercises above, apply the Ratio-to-moving average technique and compute the forecasts for the next 12 months period.
2. Open a new worksheet and copy/paste the time series data for dependent variable “Sales”. Create data for eleven Dummy seasonal variables and run Regression analysis using these variables as regressors:
   a) Conduct test of hypothesis to determine whether any of the regression coefficients are not zero (use the 0.05 significance level). Is there significant evidence of a linear relationship on each of the predictor variables? Would you consider eliminating any dummy variable as insignificant (use the 0.05 significance level)?
   b) Rerun the regression analysis with these variables eliminated (if any) and conduct again a global test of hypothesis and a test of hypothesis on each of the independent variables (use the 0.05 significance level). Is the new model significant? Is there evidence of significance of relationship on any of the remaining regressors? Is it necessary to make any further changes with the model? If yes, repeat step b) until the results obtained are satisfactory.
   c) Perform residual analysis with the residual plots provided in the regression output. Set up a Box-and-Whisker plot of the errors and a chart of the predicted values versus residuals and analyze all of them. Do you see any violations of the regression assumptions? If yes, adjust the model accordingly.
   d) Compute Sales forecasts for the next 12 months.
3. Use (copy/paste) the formulas created in Part 3 of the Integrative case and compute MFE, MAD, MAPE, MPE, MSE and CV(RMSE) for both of the new models. Use the testing dataset of 12 new monthly forecasts provided in spreadsheet Errors.
4. Comment and analyze the model accuracy:
   - How good is the accuracy of the new models compared to the ***one-step naïve forecast*** and all other models from the previous Parts? Explain.
   - Which model provides the best (so far) accuracy? Discuss.
5. What overall recommendations would you make to the research team and why?
6. Write a report on the questions above, discussing all important findings and draw relevant conclusions about this part of the Integrative Case.

References

Box, G., & Jenkins, G. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day.

Box, G., Jenkins, G., Reinsel, G., & Ljung, G. (2016). *Time Series Analysis: Forecasting and Control*. Wiley.

Durbin, J., & Watson, G. S. (1950). Testing for Serial Correlation in Least Squares Regression, I. *Biometrika, 37*(3–4), 409-428. doi:10.1093/biomet/37.3-4.409. JSTOR 2332391.

Durbin, J., & Watson, G. S. (1951). Testing for Serial Correlation in Least Squares Regression, II. *Biometrika, 38*(1–2), 159-179. doi:10.1093/biomet/38.1-2.159. JSTOR 2332325.

Granger, C., & Newbold, P. (1974). Spurious regressions in econometrics. *Journal of Econometrics, 2*(2), 111-120. doi:10.1016/0304-4076(74)90034-7

Kendall, M. (1976). *Time-Series* 2nd ed. Charles Griffin, (Fig. 5.1).

Madala, H. R., & Ivakhnenko, A. G. (1994). *Inductive Learning Algorithms for Complex Systems Modelling*. Boca Raton, FL: CRC Press Inc.

Mueller J. A., & Lemke, F. (2003). *Self-Organizing Data Mining: An Intelligent Approach To Extract Knowledge From Data.* Victoria, BC: Trafford Publishing.

Murphy, K. (2012). *Machine Learning: A Probabilistic Perspective*. MIT.

Nelson, C., & Plosser, C. (1982). Trends and Random Walks in Macroeconomic Time Series: Some Evidence and Implications. *Journal of Monetary Economics, 10*(2), 139-162. doi: 10.1016/0304-3932(82)90012-5

Pearson, K. (1897). Mathematical Contributions to the Theory of Evolution-On a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs. *Proceedings of the Royal Society of London, 60,* 489-498. doi:10.1098/ rspl.1896.0076

Petersen, B., & Strongin, S. (1996). Why are some industries more cyclical than others? *Journal of Business & Economic Statistics, 14*(2), 189-198.

Wold, H. (1954). *A Study in the Analysis of Stationary Time Series*. Second revised edition, with an Appendix on Recent Developments in Time Series Analysis by Peter Whittle. Almqvist and Wiksell Book Co., Uppsala.

Chapter 8. Time Series Analysis and Forecasting – Part 2

8.1. Autoregression and Moving Average Models

Time series models estimate difference equations containing stochastic components. Two commonly used forms of these models are ***autoregressive models (AR)*** and ***moving average (MA)*** models. The ***ARMA*** methodology (Box et al., 2016), developed by Box and Jenkins (1976, pp. 28-32), combines the ***AR*** and ***MA*** models to produce the ***autoregressive moving average (ARMA)*** model, which is the cornerstone of stationary time series analysis. ***ARIMA (autoregressive integrated moving average)*** models on the other hand are used to describe non-stationary time series (Box et al., 2016).

A. Autoregressive models (AR)

An ***autoregressive process*** operates under the premise that past values have an effect on current values. A process denoted as ***AR(1)*** is the ***first-order process***, meaning that the current value is based on the immediately preceding value. An ***AR(2)*** (i.e. ***second-order process***) has the current value based on the previous two values and so on.

<u>Estimation of AR parameters</u>

In ***autoregression,*** future values are estimated based on a weighted sum of past values. An ***autoregressive (AR) model*** is a representation of a type of random process and as such, it describes certain time-varying processes in nature, economics, etc. It is a special case of the more general ***ARMA*** model of ***time series*** and specifies that the output variable depends linearly on its own previous values:

$$Y(t) = \beta_0 + \sum_{i=1}^{p} \beta_i Y(t-i) + \varepsilon_t \tag{8-1}$$

where $\boldsymbol{y_t}$ is the dependent variable $\boldsymbol{y_t} = \{\boldsymbol{y_1, y_2, \dots y_T}\}$

- $\boldsymbol{\beta_1, \beta_2, \dots \beta_p}$ are the unknown parameters of AR model;
- $\boldsymbol{p}$ indicates the order of AR(p) model or the maximum lag value ($\boldsymbol{p} < \boldsymbol{T}$);
- $\boldsymbol{\beta_0}$ is a constant term (often omitted for simplicity);
- $\boldsymbol{t}$ is the current time period, $\boldsymbol{t} = \{\boldsymbol{1, 2, \dots T}\}$, where $\boldsymbol{T}$ is the index set;
- ε_t is the model error (residual or noise).

An ***AR*** model is simply a linear regression of the current value of the time-series against one or more prior values of the series. ***AR*** models can be analyzed and fitted with one of the various methods, including standard ***OLS***. Another advantage is that ***AR*** models have a straightforward business interpretation, like any other linear regression model.

Some parameter constraints are necessary for the model to remain wide-sense stationary. For example, processes in the ***AR(1)*** model with unit root ≥ 1 are not stationary. More generally, for an ***AR(p)*** model to be wide-sense stationary the roots of the polynomial[1] (8-2) must lie within the unit circle[2], i.e., each root $\boldsymbol{z_i}$ must satisfy $|\boldsymbol{z_i}|<\mathbf{1}$.

$$z^p - \sum_{i=1}^{p} \beta_i z^{p-i} \tag{8-2}$$

The partial ***AutoCorrelation Function*** (***ACF***) plays an important role in data analyses aimed at identifying the extent of the lag in an autoregressive model. The use of this function was introduced as part of the Box–Jenkins approach to time-series modelling. By plotting the partial ***ACF*** (see Fig.7-27) we could determine the appropriate lags **p** in an ***AR(p)*** model or in an extended ***ARIMA (p,d,q)*** model.

l-step-ahead forecasting

Once the parameters of the autoregression (8-2) have been estimated, the ***AR*** model can be used to forecast an arbitrary number of periods into the future:

$$\boxed{Y^*_{(T+l)} = b_0 + \sum_{i=1}^{p} b_i Y_{(T+l-i)}} \tag{8-3}$$

where $\boldsymbol{Y^*}$ is the forecast of dependent variable $\boldsymbol{Y}$ ($\boldsymbol{Y_t} = \{\boldsymbol{Y_1, Y_2, \ldots Y_T}\}$) for period ($\boldsymbol{T+l}$), $\boldsymbol{l} > \boldsymbol{0}$;

$\boldsymbol{p}$ is the order of the **AR** model ($\boldsymbol{p} < \boldsymbol{T}$)

$\boldsymbol{b_i}$ are the estimated **AR** coefficients ($\boldsymbol{i = 0, 1, \ldots p}$)

$\boldsymbol{l}$ is the time horizon, or the range of the forecast.

There is a sequence of steps we should follow to compute an ***l-step-ahead forecast***:

- First, we begin with ***l=1*** (i.e. ***T+1***) as the first period for which data is not yet available and substitute the known prior values $\boldsymbol{Y_{t-i}}$ (for ***t=T+l*** and ***i=1, 2, ...p***) into the **AR** model (8-3). The output of the autoregressive equation is the forecast for the first unobserved period.
- Next, we use ***l=2*** to refer to the next period for which data is not yet available and again the autoregressive equation (8-3) is used to make the forecast. However, there is one difference – since $\boldsymbol{Y_{(T+l-1)}}$, i.e. the value of one period prior to the one now being forecasted, is unknown, its expected value (the predicted value arising from the previous forecasting step) is used instead.

[1] A polynomial of lag operators is called a ***lag polynomial*** and it is a common notation for ***ARMA*** models, see http://en.wikipedia.org/wiki/Lag_operator

[2] In mathematics, a ***unit circle*** is a circle with a radius of one.

- For future periods we apply the same procedure, each time using one more forecasted value on the right side of the predictive equation (8-3) until, after ***p*** predictions, all ***p*** right-side values are predicted values from prior steps.

It should be noted that there are four sources of uncertainty, regarding predictions, obtained in this manner: (1) uncertainty as to whether the **AR** model (8-3) is the correct model; (2) uncertainty about the accuracy of the forecasted values ***Y**** that are used as lagged values in the right side of the autoregressive equation; (3) uncertainty about the true values of the autoregressive coefficients ($\boldsymbol{b_i}$); and (4) uncertainty about the value of the error terms ($\boldsymbol{\varepsilon_t}$), for the period being predicted. Each of the last three can be quantified and combined to give a confidence interval for the ***l-step-ahead predictions,*** which becomes wider as ***l*** increases because of the use of an increasing number of estimated values for the right-side of the predictive equation variables.

Evaluating the quality of AR model predictions

The predictive performance of the **AR** model can be assessed as soon as estimation has been done if ***cross-validation*** is used. As discussed above, in this approach some of the initially available data are used for parameter estimation purposes (***training dataset***), and some (usually from available observations later in the sample) are held back for out-of-sample testing (***testing dataset***). When some time has passed after the parameter estimation was conducted, more data will have become available and predictive performance can be evaluated then using the new data.

In either case, there are two aspects of predictive performance that should be evaluated – ***one-step-ahead*** and ***l-step-ahead*** performance. For one-step-ahead performance, the estimated parameters are used in the autoregressive equation along with observed values of dependent variable (***Y***) for all periods prior to the one being predicted, and the output of the equation is the one-step-ahead forecast. This procedure is used to obtain forecasts for each of the out-of-sample observations. To evaluate the quality of ***l-step-ahead*** forecasts, the forecasting procedure in the previous section is employed to obtain the predictions.

For a given set of predicted values and the corresponding set of actual values for (***Y***) for various time periods, the common evaluation techniques to use are: ***Mean Forecast Error (MFE), Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), Mean Percentage Error (MPE),*** **Coefficient of variation of the RMSE, CV(RMSE),** and other measures, discussed in Chapter 3.

The question of how to interpret the measured forecasting accuracy arises – for example, what is a "***high***" (***bad***) or a "***low***" (***good***) value for the error? There are two common approaches of comparison:

- *First*, the forecasting accuracy of a reference model (estimated under different modeling assumptions or different estimation techniques), can be used for comparison purposes.
- *Second*, the ***out-of-sample*** accuracy measure can be compared to the same measure computed for the ***in-sample*** data points (***training set***) for which enough prior data values are available (that is, dropping the first **p** data points, for which **p** prior data points are not available). Since the model was estimated specifically to fit the ***in-sample*** points as well as possible, it will usually be the case that the ***out-of-sample*** predictive performance will be poorer than the ***in-sample*** predictive performance. But if the predictive quality deteriorates ***out-of-sample*** by "*not very much*" (which is not precisely definable), then the forecaster may be satisfied with the performance.

It is a good practice, no matter which of the above approaches is used, always to define a threshold (as %) of the evaluation criteria (MAPE, NRMSE, CV(RMSE) and so forth). In general, in business forecasting a margin of 5%-7% error is considered as acceptable (occasionally even up to 10%-12%), however, sometimes (for example in Production and Operations Management) an error less than 1% is required.

B. Moving Average model

In time series analysis and forecasting, the ***moving average (MA)*** model is a common approach for modeling single time series data. The notation ***MA(q)*** refers to the moving average model of order **q**. Suppose we have a time-series $\boldsymbol{X = \{X_1, X_2, \ldots X_T\}}$. Then the MA model is given as follows:

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \qquad (8\text{-}4)$$

where **μ** is the mean of the series (often assumed to equal zero),

$\boldsymbol{\theta_1, \ldots, \theta_q}$ are the parameters of the model

$\boldsymbol{\varepsilon_t, \ldots, \varepsilon_{t-q}}$ are white noise error terms

q is the order of the MA model.

This can be equivalently written in terms of the backshift operator ***B*** as shown in (8-5).

$$X_t = \mu + (1 + \theta_1 B + \cdots + \theta_q B^q)\varepsilon_t. \qquad (8\text{-}5)$$

Thus, an MA model is conceptually a linear regression of the current value of the series against current and previous (unobserved) white noise error terms. The random errors at each point are assumed to be mutually independent and to come from the same distribution, typically a normal distribution, with allocation at zero and constant variance. Fitting the MA estimates is more complicated than with AR models because the lagged error terms are not observable. This means that iterative non-linear fitting procedures need to be used in place of linear least squares. MA models also have a less obvious interpretation than AR models.

The role of the random errors (**ε_t**) in the ***MA*** model differs from their role in the ***AR*** model in two ways:

- *First,* they are propagated to future values of the time series directly – for example, **ε_{t-1}** appears directly on the right side of the equation for ***X_t***. In contrast, in an ***AR*** model **ε_{t-1}** does not appear on the right side of the ***X_t*** equation (8-1), but it does appear on the right side of the ***X_t*** equation, and ***X_{t-1}*** appears on the right side of the ***X_t*** equation, giving only an indirect effect of **ε_{t-1}** on ***X_t***.
- *Second*, in the ***MA*** model an error **ε_t** affects ***X*** values only for the current period and ***q*** periods into the future – in contrast, in the ***AR*** model an error **ε_t** affects ***X*** values infinitely far into the future, because **ε_t** affects ***X_t***, which affects ***X_{t+1}***, which affects ***X_{t+2}***, and so on forever.

Sometimes the ***AutoCorrelation Function*** (***ACF***) and ***Partial AutoCorrelation Function*** (***PACF***) will suggest that an ***MA*** model would be a better model choice and sometimes both ***AR*** and ***MA*** terms should be used in the same ***ARMA*** model. Note that after each of those models is fit, the error terms should be independent and follow the standard assumptions for a univariate process, as explained in Chapter 6.2.

C. Autoregressive Moving Average models

In the statistical analysis of time series, ***autoregressive moving average (ARMA)*** models provide a parsimonious description of a (weakly) stationary stochastic process in terms of two polynomials, one for the auto-regression and the second for the moving average. The general ***ARMA*** model was described in the 1951 thesis in time series analysis hypothesis testing of Peter Whittle, who used mathematical analysis (Laurent series and Fourier analysis) and statistical inference (Whittle, 1983). ***ARMA*** has been promoted since 1971 by Box and Jenkins (1976) who expounded an iterative (Box–Jenkins) method for choosing models and estimating their parameters. This method was useful for low-order polynomials (of degree three or less).

Although both autoregressive and moving average approaches were already known (and were originally investigated by Yule (1927), the contribution of Box and Jenkins (1976) was in developing a systematic methodology for identifying and estimating models that could incorporate both approaches. This makes Box-Jenkins models a powerful class of models.

Given a time series of data ***X*** (***X* = {*X*₁, *X*₂, ... *X*ₜ}**), the ***ARMA*** model is a tool for understanding and predicting future values in this series. The model consists of two parts, an autoregressive ***(AR)*** part and a moving average ***(MA)*** part. The model is usually then referred to as the ***ARMA(p,q)*** model where **p** is the order of the autoregressive part and **q** is the order of the moving average part (as defined below).

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}. \qquad (8\text{-}6)$$

where $\boldsymbol{X_t}$ is the dependent variable $\boldsymbol{X = \{X_1, X_2, \ldots X_T\}}$,

- $\boldsymbol{\varphi_1, \varphi_2, \ldots \varphi_p}$ and $\boldsymbol{\theta_1, \ldots, \theta_q}$ are the unknown parameters of ***ARMA*** model,
- ***p*** indicates the order of AR(p) model or the maximum lag operator ($\boldsymbol{p} < \boldsymbol{T}$);
- ***q*** indicates the order of MA(q) model,
- ***c*** is a constant term,
- ***t*** is the current time period, $\boldsymbol{t} = \{\boldsymbol{1, 2, \ldots T}\}$, where ***T*** is the index set;
- $\boldsymbol{\varepsilon_t}$ are the error terms (or white noise).

The error terms $\boldsymbol{\varepsilon_t}$ in (8-6) are generally assumed to be independent identically distributed (i.i.d.) random variables, sampled from a normal distribution with zero mean, i.e. $\varepsilon_t \sim N(0,\sigma^2)$ where σ^2 is the variance. These assumptions may be weakened but doing so the properties of the model will change. In particular, a change to the i.i.d. assumption would make a rather fundamental difference.

Finding appropriate values of **p** and **q** in the ***ARMA(p,q)*** model can be facilitated by plotting the ***PACF*** (partial autocorrelation functions) for an estimate of **p**, and likewise using the ***ACF*** (autocorrelation functions) for an estimate of **q**. Further information can be extracted by considering the same functions for the residuals $\boldsymbol{\varepsilon_t}$ of a model fitted with an initial selection of **p** and **q**.

ARMA models in general can (after choosing **p** and **q**) be fitted by ***OLS*** to find the values of the parameters, which minimize the error term. It is generally considered good practice to find the smallest values of **p** and **q,** which provide an acceptable fit to the data.

Note that the ***ARMA*** model is a univariate model. Its extension, the ***autoregressive moving average model with exogenous inputs*** (***ARMAX***) is discussed in Chapter 9.

8.2. Autoregressive Integrated Moving Average Models and ARIMA Methodology

A. Autoregressive Integrated Moving Average models (ARIMA)

ARIMA models are used to describe non-stationary time series data. Box and Jenkins suggest (1976) differencing a non-stationary time series to obtain a stationary series to which an ***ARMA*** model can be applied. Non-stationary time series have a pronounced trend and do not have a constant long-run mean or variance.

An ***ARIMA*** model is a generalization of an ***ARMA*** model. These models are fitted to time series data either to better understand the data or to predict future points in the series. They are applied in some cases where data show evidence of non-stationarity, where an initial differencing step (corresponding to the "integrated" part of the model) can be applied to remove the non-stationarity.

The model is generally referred to as an ***ARIMA(p,d,q)*** model where parameters **p**, **d**, and **q** are non-negative integers that refer to the order of the autoregressive, integrated, and moving average parts of the model respectively.

Given a time series of data ***X*** (***X* = {*X*₁, *X*₂, ... *X*_T})**, the equation (8-6) for an ***ARMA(p′,q)*** model, after transformations similar to (7-22), can be rewritten as:

$$\left(1-\sum_{i=1}^{p'}\alpha_i L^i\right) X_t = \left(1+\sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_t \qquad (8\text{-}7)$$

where ***L*** is the lag operator,

α_i are the parameters of the autoregressive part of the model,

θ_i are the parameters of the moving average part and

ε_t are the error terms, which are assumed to be i.i.d. variables sampled from a normal distribution with zero mean.

If we assume that the polynomial $\left(1-\sum_{i=1}^{p'}\alpha_i L^i\right)$ has a unitary root of multiplicity ***d***, then an ***ARIMA(p,d,q)*** process expresses this polynomial factorization property with ***p=p′−d***, and can be generalized as follows:

$$\left(1-\sum_{i=1}^{p}\phi_i L^i\right)(1-L)^d X_t = \delta + \left(1+\sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_t \qquad (8\text{-}8)$$

This defines an ***ARIMA(p,d,q)*** process with a **drift = δ/(1−Σφᵢ)**. Some well-known special cases arise naturally. For example, an ARIMA(0,1,0) model is given by:

$$X_t = X_{t-1} + \varepsilon_t$$

i.e. it is equivalent to equation (5-5), which is simply a ***random walk***.

Theoretically, ***ARIMA*** models put it all together:

- Generalized random walk models that are fine-tuned to eliminate all residual autocorrelation;
- Generalized exponential smoothing models that can incorporate long-term trends and seasonality;
- Stationarized regression models that use lags of the dependent variables and/or lags of the forecast errors as regressors.
- The most general class of forecasting models for time series that can be stationarized by transformations such as differencing, logging, and/or deflating

ARIMA, actually, is not a forecasting model, but rather a procedure used to select from a group of forecasting models that best fit to the particular set of time series data. Box and Jenkins proposed a three-stage methodology which does not assume any particular pattern in the historical data of the series to be forecasted. Rather, it uses a three-step iterative approach of model identification, parameter estimation and validation. This three-step process is repeated several times until a satisfactory model is finally selected. Then this model can be used for forecasting future values of the time-series (see Fig.8-1).

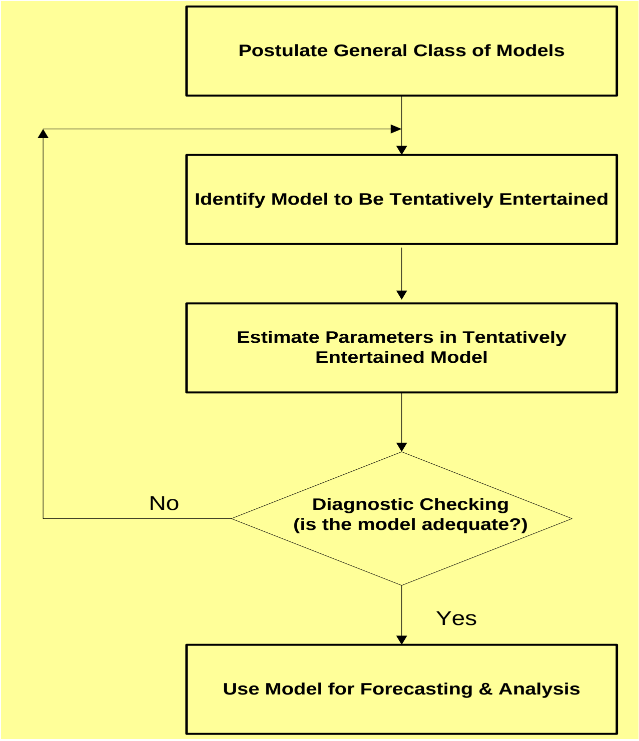

Postulate General Class of Models
Identify Model to Be Tentatively Entertained
Estimate Parameters in Tentatively Entertained Model
No
Diagnostic Checking (is the model adequate?)
Yes
Use Model for Forecasting & Analysis

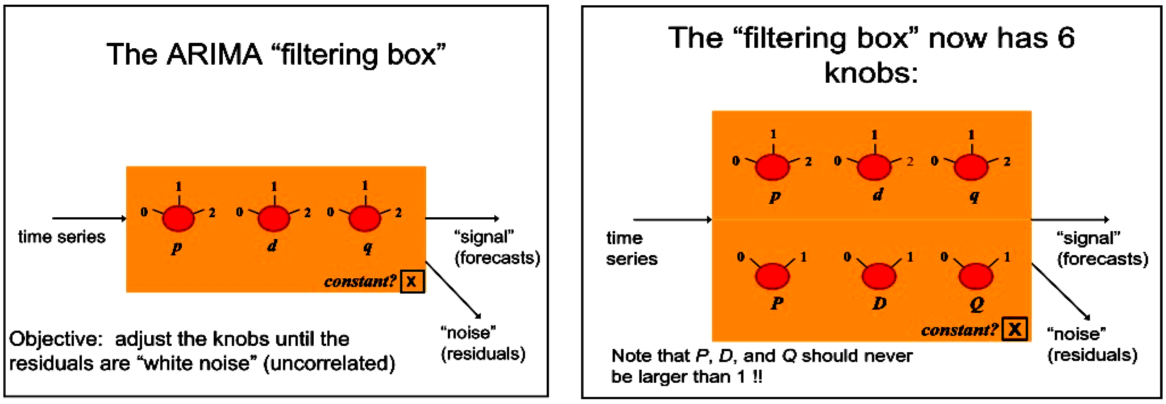
Fig.8-1 Box-Jenkins (1976) methodology

1) The identification of a tentative model stage involves identifying if the series is stationary or not and the presence of seasonality by examining plots of the series, autocorrelation and partial autocorrelation functions. ACF and PACF are also used to decide which (if any) AR or MA component should be used in the model.
2) Parameter estimation – using computation algorithms to arrive at coefficients that best fit the selected ARIMA model. The most common methods use maximum likelihood estimation or non-linear least-squares estimation.
3) Finally, the validation stage involves diagnostic checking such as plotting the residuals to detect outliers and evidence of model fit, and testing whether the estimated model conforms to the specifications of a stationary univariate process, i.e. the residuals should be independent of each other and constant in mean and variance over time. If the estimation is inadequate, we return to step one and attempt to build a better model.

This approach is more complicated than the other time-series models, but it is also capable of handling almost any type of time series data. The main features of Box-Jenkins models are:

- The Box-Jenkins model assumes that the time series is stationary. Box and Jenkins recommend differencing non-stationary series one or more times to achieve stationarity. Doing so produces an ARIMA model where "I" means "Integrated".
- Some formulations transform the series by subtracting the mean of the series from each data point. This yields a series with a mean of zero. Whether you need to do this or not depends on the software you use to estimate the model.
- Box-Jenkins models can be extended to include ***seasonal autoregressive and seasonal moving average terms*** (***SARIMA***). Although this complicates the notation and mathematics of the model, the underlying concepts for seasonal AR and seasonal MA terms are similar to the non-seasonal AR and MA terms.
- The most general Box-Jenkins model includes difference operators, autoregressive terms, moving average terms, seasonal difference operators, seasonal autoregressive terms, and seasonal moving average terms. As with modeling in general, however, only necessary terms should be included in the model.

Technically, ARIMA methodology could be presented as a "filtering box" (Nau, 2014), as shown in Fig.8-2, where the seasonal part of an ARIMA model is summarized by three additional numbers: P = # of seasonal autoregressive terms, D = # of seasonal differences and Q = # of seasonal moving-average terms:


The ARIMA “filtering box”
time series
0 1 2 p
0 1 2 d
0 1 2 q
constant? X
“signal” (forecasts)
“noise” (residuals)
Objective: adjust the knobs until the residuals are “white noise” (uncorrelated)
The “filtering box” now has 6 knobs:
time series
0 1 2 p
0 1 2 d
0 1 2 q
0 1 P
0 1 D
0 1 Q
constant? X
“signal” (forecasts)
“noise” (residuals)
Note that P, D, and Q should never be larger than 1 !!

a) Nonseasonal ARIMA b) seasonal ARIMA

Fig.8-2 ARIMA methodology as a "filtering box"

B. Forecasts using ARIMA models

The ARIMA model can be viewed as a "cascade" of two models. Given time series data and the corresponding variable ***Y*** (***Y****t* = {***Y****1*, ***Y****2*, ***…Y****T*}), the first model is non-stationary:

$$Y_t = (1 - L)^d X_t \tag{8-9}$$

while the second is wide-sense stationary:

$$\left(1 - \sum_{i=1}^{p} \phi_i L^i\right) Y_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right) \varepsilon_t \tag{8-10}$$

Now forecasts can be made for the variable Y_t*, using a generalization of the method of autoregressive forecasting (see equation (8-3) and its explanations).

C. <u>ARIMA discussions</u>

The following remarks regarding Box-Jenkins models should be noted.

a) Box-Jenkins models are quite flexible due to the inclusion of both autoregressive and moving average terms, but its complexity discourages many forecasters and managers from using it[3].

b) It is best suited to short-range (i.e. daily, weekly or monthly) forecasts[4].

c) Based on the Wold decomposition theorem, a stationary process can be approximated by an ARMA model – in practice, finding that approximation may not be easy[5].

d) Building good ARIMA models generally requires more experience than commonly used statistical methods such as regression[6].

[3] Source: http://help.sap.com/saphelp_45b/helpdata/en/35/8a524b52060634e10000009b38f9b9/content.htm
[4] Ibid.
[5] Source: http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc445.htm
[6] Ibid.

e) It is usually necessary to develop a new model whenever new data appear.

f) It requires a large amount of data – typically, effective fitting of Box-Jenkins models requires at least a moderately long series – Chatfield (1996) recommends at least 50 observations, but many others would recommend at least 100 observations[7] .

The strengths of the Box-Jenkins approach are its versatility (it can be used with most time series data) and its track record (its forecasting accuracy tends to exceed that of most time series models). A number of studies comparing forecasting models indicate that the Box-Jenkins approach provides some of the more accurate short-range forecasts (one to three periods out) of any time series models[8].

At the same time, authors such as Commandeur & Koopman (2007) argue that the Box-Jenkins approach is fundamentally problematic. The problem arises because in "the economic and social fields, real series are never stationary however much differencing is done". Thus, the investigator has to face the question: how close to stationary is close enough? As the authors note, "This is a hard question to answer". The authors further argue that rather than using Box-Jenkins, it is better to use state space methods, as stationarity of the time series is then not required.

A detailed study (Onwubolu, 2009) and investigation of the forecasting performance of two time-series forecasting techniques (***Elman neural network*** and ***self-organizing data mining***) against the autoregressive integrated moving average (***ARIMA***) model show that ***GMDH*** based techniques are able to develop even complex models reliably with better overall error rates than most of the current methods. This approach is discussed in the next section.

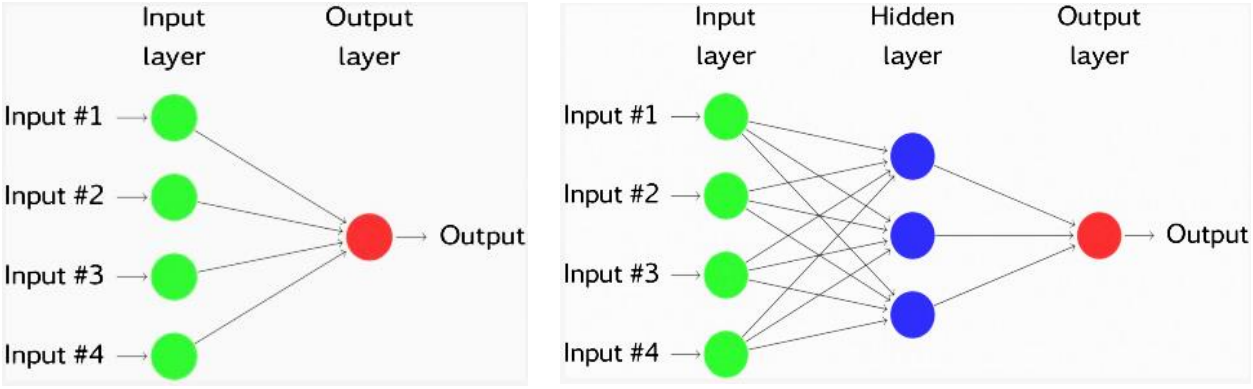
8.3. Time Series Forecasting Using Data Mining Techniques

In the previous sections, a few important stochastic methods for time series modeling and forecasting were discussed. Outside of traditional statistical modelling, an enormous amount of forecasting is done using ***Data Mining*** techniques[9]. Most of these techniques have no formal statistical model, sometimes prediction intervals are not computed, and there is limited model testing, but some of the data-mining tools have proven powerful predictors in specific contexts, especially when there is a vast quantity of available data. Most successfully among these techniques, a variety of ***Artificial Neural Networks (ANNs)*** have been used for predictions in many different areas.

[7] Ibid.

[8] Source: http://help.sap.com/saphelp_45b/helpdata/en/35/8a524b52060634e10000009b38f9b9/content.htm

[9] Data Mining and ANNs as a Data Mining technique are discussed in detail in Chapters 10, 11 and 12.



a) A simple Artificial Neural Network b) Multi-layer Feed-Forward Network

Fig.8-3 Examples of Artificial Neural Networks (ANNs)

Artificial Neural Networks (ANNs)

The basic objective of ANNs is to construct a model for mimicking the intelligence of the human brain into a machine (Berry & Linoff, 2000; Zhang et al., 1998). Similar to the work of a human brain, ANNs try to recognize regularities and patterns in the input data, "learn" from experience and then provide generalized results based on their known previous knowledge.

A neural network can be thought of as a network of "neurons" organized in layers as shown in Fig.8-3. The predictors form the first layer (the input layer), and the forecasts form the last layer (the output layer). There may be intermediate layers containing "hidden neurons". The very simplest networks contain no hidden layers and are equivalent to linear regression. Figure 8-3 a) shows the neural network version of a linear regression with four predictors. The coefficients attached to these predictors are referred to as "weights". The forecasts are obtained by a linear combination of the inputs. The weights are selected in the neural network framework using a "learning algorithm" that minimizes a "cost function" such as MSE.

If we add an intermediate layer with hidden neurons (Fig. 8-3 b), the neural network becomes non-linear. It is known in general as a ***multi-layered feed-forward network*** (***FNN***) where each layer of nodes receives inputs from the previous layers. The outputs of nodes in one layer are inputs to the next layer. The inputs to each node are combined using a weighted linear combination. The result is then modified by a nonlinear function before being output.

The parameters of the function and the weights are "learned" from the data. The weights most often take random values to begin with, which are then updated using the observed data. Consequently, there is an element of randomness in the predictions produced by a neural network. Therefore, the network is usually trained several times using different random starting points, and the results are averaged.

ANNs are generally presented as systems of interconnected neurons which can compute values from inputs and are capable of machine learning as well as pattern recognition and prediction thanks to their adaptive nature. Graphically, as shown in Fig.8-4, an ANN is presented as an interconnected group of nodes, akin to the vast network of "neurons" in a "brain". Each circular node represents an artificial neuron and an arrow represents a connection from the output of one neuron to the input of another.

There is no single formal definition of what an artificial neural network is. However, a class of statistical models may commonly be called "*Neural*" if they possess the following characteristics:

– consist of sets of adaptive weights, i.e. numerical parameters that are tuned by a learning algorithm, and

– are capable of approximating non-linear functions of their inputs.

The adaptive weights are conceptually connection strengths between neurons, which are activated during training and prediction.

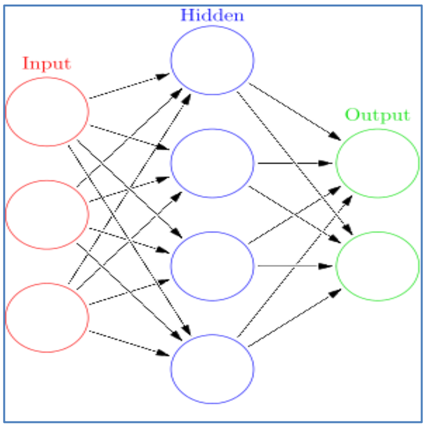


Fig.8-4 General model of ANNs

Although the development of ANNs was initially biologically motivated, afterward they have been applied in many different areas, especially for forecasting and classification purposes (Mueller & Lemke, 2003; Berry & Linoff, 2000). There are a few salient features of ANNs, which make them quite favorite for time series analysis and forecasting:

First, ANNs are data-driven and self-adaptive in nature (Berry & Linoff, 2000; Zhang et al., 1998). It means that there is no need to specify a particular model form or to make any a priori assumption about the statistical distribution of the data. The desired model is adaptively formed based on the features presented from the data. This approach is quite useful for many practical situations, where no theoretical guidance is available for an appropriate data generation process.

Second, ANNs are inherently non-linear, which makes them more practical and accurate in modeling complex data patterns, as opposed to various traditional linear approaches, such as ARIMA methods. Results from many studies (Mueller & Lemke, 2003; Onwubolu, 2009; Zhang et al., 1998) suggest and prove that ANNs made significantly better forecasting than various linear models.

Finally, as suggested by Hornik, Stinchcombe and White (1982), ANNs are universal functional approximators. They have shown that a network can approximate any continuous function to any desired accuracy. ANNs use parallel processing of the information from the data to approximate a large class of functions with a high degree of accuracy. Further, they can deal with situation, where the input data are erroneous, incomplete or fuzzy (Mueller & Lemke, 2003; Onwubolu, 2009).

ANNs applications in TSA and Forecasting

The most widely used ANNs in forecasting problems are the multi-layer ***FNN*** (Fig.8-4). In ***TSA***, within the ***FNN*** formulation, described above, the input nodes are the successive observations of the time series. Thus, given the dependent variable ***Y*** ($Y_t = \{Y_1, Y_2, \ldots Y_T\}$), the output (the forecast) ***Y**** is a function of the values Y_{t-i} ($i= 1, 2 \ldots p$}, where ***p*** is the number of input nodes. Another variation of ***FNN***, the ***Time Lagged*** (***TLNN***) architecture [13] is also widely used. In ***TLNN***, the input nodes are the time series values at some particular lags. For example, a typical ***TLNN*** for a time series, with seasonal period **s = 4** can contain the input nodes as the lagged values at time (t −1), (t – 2) and (t −4). The value at time **t** (Y_t) is to be forecasted using the values Y_{t-i} at lags **i** =1, 2 and 4.

Another proposed ***FNN*** is the ***seasonal ANN*** (***SANN***). The ***SANN*** model does not require any preprocessing of raw data. Also, ***SANN*** can learn the seasonal pattern in the series, without removing them, contrary to some other traditional approaches, such as the seasonal ***ARIMA***. When forecasting with ***SANN***, the number of input and output neurons should be taken as 12 for monthly and 4 for quarterly time series. The appropriate number of hidden nodes can be determined by performing suitable experiments on the training data.

There are some other types of neural models also proposed in literature, such as the *Probabilistic Neural Network* for classification problems, the *Generalized Regression Neural Network* for regression problems and others (Berry & Linoff, 2000).

If the model, cost function and learning algorithm are selected appropriately the resulting ANN can be extremely robust. Perhaps the greatest advantage of ANNs is their ability to be used as an arbitrary function approximation mechanism that 'learns' from observed data. However, using them is not so straightforward, and a relatively good understanding of the underlying theory is essential:

A. Model selection – this will depend on the data representation and the application. Overly complex models tend to lead to problems with learning. The number of hidden layers, and the number of nodes in each hidden layer, must be specified in advance. It is a trial-and-error process and could be optimized using cross-validation techniques.

After specifying a particular network structure, the next most important issue is the determination of the optimal network parameters. The selection of appropriate network parameters is crucial, while using ANNs for forecasting purposes. Also, a suitable transformation or rescaling of the training data is often necessary to obtain best results.

Another major problem is that an inadequate or large number of network parameters may lead to overfitting. This produces spuriously good within-sample fit, which does not generate better forecasts. To penalize the addition of extra parameters some model comparison criteria, such as AIC and BIC can be used (Faraway & Chatfield, 1998). *Network Pruning* and other regularization techniques (Berry & Linoff, 2000) are also quite popular in this regard.

A desired network model should produce reasonably small error not only on within sample (training) data but also on out of sample (test) data (Berry & Linoff, 2000). Due to this reason immense care is required while choosing the number of input and hidden neurons. However, it is a difficult task as there is no theoretical guidance available for the selection of these parameters and often experiments, such as cross-validation are conducted for this purpose (Mueller & Lemke, 2003).

B. Learning algorithm: There are numerous trade-offs between learning algorithms. Almost any algorithm will work well with the correct hyperparameters (i.e. parameters of a prior theoretical distribution, normal, beta, etc.) for training on a particular fixed dataset. However, selecting and tuning an algorithm for training on unseen data requires a significant amount of experimentation. ***Deep learning algorithms*** attempt to model high-level abstractions in data by using model architectures composed of multiple non-linear transformations. These algorithms are based on the learning of multiple levels of features or representations of the data. Higher level features are derived from lower level features to form a hierarchical representation.

Some of the most successful deep learning methods involve ANNs. A ***Deep Neural Network (DNN)*** is defined to be an artificial neural network with multiple hidden layers of units between the input and output layers. The extra layers enable composition of features from lower layers, giving the potential of modeling complex data with fewer units than a similarly performing shallow network.

Unfortunately, as with ANNs, many issues can arise with DNNs if they are naively trained. Two common issues are overfitting and computation time. DNNs are prone to overfitting because of the added layers of abstraction, which allow them to model rare dependencies in the training data. Regularization methods (typically cross-validation) can be applied during training to help combat overfitting.

Multi Layered Networks of Active Neurons (MLNAN)

The ***Group Method of Data Handling*** (***GMDH***), introduced in Chapter 3, is a family of inductive algorithms for computer-based mathematical modeling of multi-parametric datasets that features fully automatic structural and parametric optimization of models. In GMDH-type self-organizing algorithms, models are generated adaptively from input data in the form of an ANN of active neurons in a repetitive generation of populations of competing partial models of growing complexity. A limited number is selected from generation to generation by cross-validation, until an optimal complex model is finalized. (see Fig.3-2 for example).

There are many different ways to select the right order of a ***GMDH-type ANN***. The most popular one is known as multi-layer inductive procedure (see Fig.8-5). It is equivalent to the ANNs with polynomial activation function of neurons. Therefore, the algorithm with such an approach usually referred to as ***GMDH-type Neural Network (NN)*** or ***Polynomial NN***.

A GMDH model with multiple inputs (x_j) and one output (Y) is a subset of components of the base function (8-11):

$$Y(x_1,\ldots,x_n) = a_0 + \sum_{i=1}^{m} a_i f_i \qquad (8\text{-}11)$$

where f_i are elementary functions dependent on different sets of inputs ($i=1, 2 \ldots m$);

x_j ($j=1, 2 \ldots n$) are the inputs at the first layer (predictors)

a_0 is the constant term;

a_i are the unknown coefficients and

m is the number of the base function components.

In order to find the best solution, GMDH algorithms consider various component subsets of the base function (8-11) called partial models. Coefficients of these models are estimated by the LS method. GMDH algorithms gradually increase the number of partial model components as shown in Fig.8-6 and find a model structure with optimal complexity indicated by the minimum value of an external criterion.

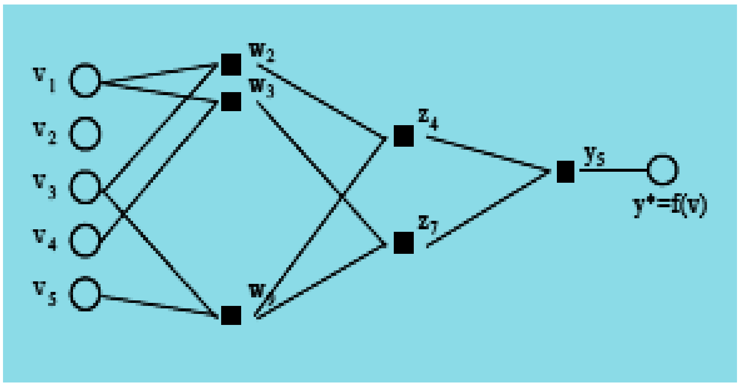


Fig.8-5 Optimal model **y*** selected by a neural network with three layers

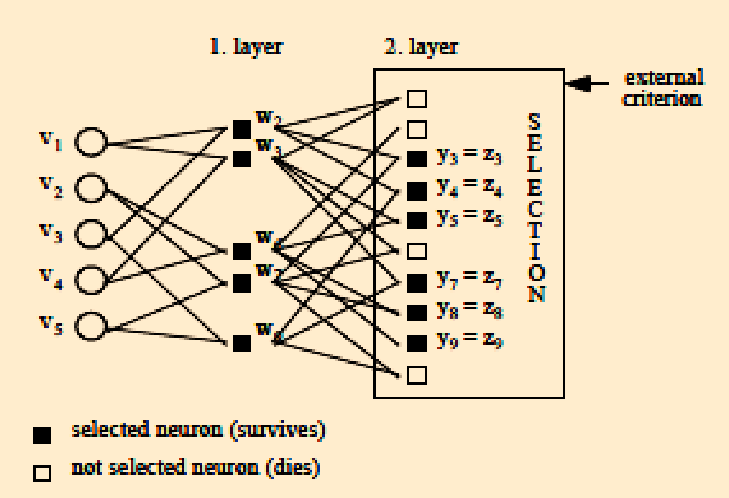


Fig.8-6 GMDH-type NN after selection of the best partial models at the second layer

This modeling approach grows a tree-like network out of data of input and output variables (seed information) in a pair-wise combination and competitive selection from a simple single neuron to a desired final output, a model without predefined characteristics. Here, neither the number of neurons and the number of layers in the network, nor the actual behavior of each created neuron is predefined. In this way, the model-building process is self-organizing because the number of neurons, the number of layers, and the actual behavior of each created neuron are identified or adjusting during the learning (training) process from layer to layer.

External criterion as a cross-validation technique is one of the key features of GMDH-type NN. Another important feature is the rule of layers complication, i.e. the partial models should be simple, without quadratic terms.

The requirements of each partial model are defined by selected external criteria (one or more), for example the traditional Criterion of Regularity (CR) requires the minimization of the sum of the squared residuals (ε), i.e. Least Squares. Following Dennis Gabor (1971) work, a "***freedom of choice***" at each layer is provided and the total number of possible partial models is limited by the threshold value of the external criterion for selection, similar to ***Genetic Algorithms*** (***GA***). Because of the relation with GA, sometimes procedures of this type are referred to as ***Multi-Stage Selection Algorithms*** (Motzev & Marchev, 1988).

A variety of GMDH-type NN algorithms were developed and applied successfully in many different areas (Madala & Ivakhnenko, 1994; Mueller & Lemke, 2003; Motzev & Marchev, 1988). Some of them, designed for developing of predictive models as a ***Multi-Layered Net of Active Neurons (MLNAN)*** are very useful in business forecasting for building both multi-input to single-output models (for example different type of regression models) and econometric models of simultaneous equations (i.e. multi-input to multi-output).

The basic idea (Motzev, 1985) in such MLNAN is that first the elements on a lower level are estimated and the corresponding intermediate outputs are computed, which are then used to estimate the parameters of the elements of the next levels. At the first layer, all possible pairs of the inputs are considered as potential factors, and only some of the best partial models (in the sense of the selection criterion – mainly coefficient of correlation) are used as inputs for the second layer. In the succeeding layers all possible pairs of the intermediate partial models from the preceding layer(s) are connected as inputs to the components of the next layer(s). This means that the output of a component at a processed level is or may become an input, depending on a local threshold value (the selection criterion here is the coefficient of determination of the partial model), to several other components at the next level. Finally, when additional layers provide no further improvement, the network synthesis stops.

For single time-series, equation (8-11) could be presented in a form similar to the general autoregression model (8-1) and then make predictions using form (8-3). Thus, in times series analysis and forecasting, the MLNAN algorithm could be summarized in the following steps (see Fig.8-7):

- Given a dataset, a variety of hypotheses is generated (including nonlinear transformations and taking into account time lags from data history), each of them representing a simple model of dependent variable and a pair of predictors.
- Each competing hypothesis is a hypothesis of entering one additional factor (lag variable), i.e. that this factor is potentially important in the general AR model.

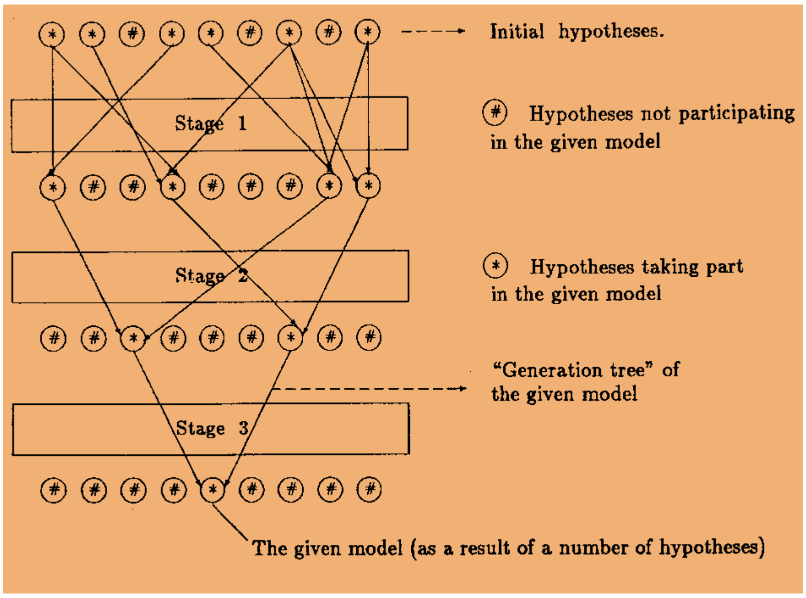

Initial hypotheses.
Stage 1
# Hypotheses not participating in the given model
Stage 2
* Hypotheses taking part in the given model
"Generation tree" of the given model
Stage 3
The given model (as a result of a number of hypotheses)

Fig.8-7 Example of a GMDH algorithm with three stages for MLNAN developing (Motzev & Marchev, 1988).

- The variety of models is limited by the external criteria (for example autocorrelation function) and can be supplemented with heuristic solutions by the researcher.
- The general AR model is "synthesized" during the next stages of the selection procedure, each of them representing a layer of active neurons, where:
  - Variety of hypotheses (partial models), which function as active neurons, is generated using outputs from the previous stage (layer) of selection as functions of two variables, which include indirectly more and more complex combinations of initial inputs, similar to ***DNN***;
  - Estimation of the coefficients in each partial model is done using LS and cross validated by MSE criterion.
  - Each generated hypothesis is a potential "best" model, which similarly to ***GA*** competes with other models at the stage "*fighting for survival*" (Fig.8-6);
  - At the end of each stage, the selected partial models are used as inputs to the next layer.
- Selected hypotheses (partial models) at a given layer are used as inputs for generating new, more complex models at the next stage of selection, from which only the best models that survived are chosen as outputs to the next stage etc.
- The selection procedure ends when satisfactory results are achieved, such as minimum MSE for the testing set, and a number of final models with similar MSE is selected, following the principle of non-finalized solutions which provides "*freedom of choice*" according to Gabor's work (1971). Thus, eventually, the researcher has a set of alternative good models.
- At the end of the procedure the full form of selected equations is restored using an automated backward tracking algorithm, which provides a set of alternative versions of the general AR model.

The choice of the final model is made by the researcher (decision maker), who has one final option to apply additional knowledge about the real-life system, but after having the guarantee that a large number of possible models have been developed, selected and evaluated, and the final choice is based on a small number of the "best" ones.

These characteristics make the proposed MLNAN very useful for addressing the existing model-building problems in TSA and forecasting. For example, overfitting is eliminated by the use of external criterion (cross-validation) for validating the model. The small number of independent measurements (or short time-series) is also not an issue, because the inverted matrix size is always 2x2, due to the pair-wise combinations. This helps in dealing with the problem of multicollinearity as well. The autocorrelation is eliminated by adding lagged time series values as predictors. The procedure is totally automated and because the algorithm has a multi-layer structure a parallel computing can be implemented for its realization. At the same time, at some crucial points of the procedure, the decision maker has options to apply additional insights, knowledge or hypotheses.

ANNs for TSA and Forecasting - discussion

ANNs were successfully applied in time-series analysis and forecasting to build different autoregressive models. The MLNAN described above, was used for building AR models of more than 20 macroeconomic variables with a time lag of up to 5 years (Motzev, 2014), providing in most cases a MAPE% less than 1.5% and an adjusted coefficient of determination ($\mathbf{R}^2$) greater than 0.9. Another detailed study (Onwubolu, 2009) of the predictive performance of two time series forecasting techniques (Elman NN and GMDH-type NN) against the ARIMA also confirmed that GMDH based techniques are able to develop even complex models reliably and achieve lower overall error rates than state-of-the-art methods.

Shahwan and Lemke (2005) explored the usefulness and applicability aspects of ANNs and ***Self-Organizing Data Mining (SODM)*** for short-term forecasting of agricultural commodity prices. Traditional ARIMA models and futures prices served as benchmarks for prediction performance evaluation. They also investigated whether a combination between *Elman neural network* and genetic algorithms generates more predictive model compared to the other forecasting methods. In their conclusions, they pointed out that the potential gain of ***ANNs, ARIMA*** and ***SODM*** seems to depend on the characteristics of the time series under consideration. A more complex time series justifies the use of ANNs. However, the greater flexibility of this model class and its ability to handle nonlinear data patterns comes at the cost of a more demanding specification procedure and computation time, especially when applying GA for optimization purpose, while one advantage of SODM from an application point of view is that it takes by far least efforts both in human (almost nothing) and in computational time. Furthermore, SODM (GMDH-type NN such as MLNAN) provide on the fly after each modeling run a reliable, analytical model that describes and does not overfit the design data. On the other hand, ARIMA needs a lot of manual efforts, theoretical knowledge, and experience in determining the suitable model but it is straightforward in computation.

In contrast to ARIMA and SODM, there is no simple clear-cut method or theory on hand for determining the optimal structure of the ANNs. Creating a reliable neural network model is a trial-and-error process. This means the danger of misspecifying an ANN is higher than for an ARIMA model and SODM. This may erode the potential predictive power of ANNs even if using GA (as their calculations confirmed) for optimizing the network topology of ANN, which can slightly improve its predictive performance.

Finally, it should be noted that ANNs and in particular GMDH-type NN are successfully used not only in time series analysis and forecasting, but also in much more complex cases, such as distributed lag models, autoregressive moving average with exogenous inputs models (ARMAX), vector autoregression models (VAR) and different econometric models, including the most complicated models of simultaneous equations, i.e. multi-input to multi-output models (Madala & Ivakhnenko, 1994; Motzev, 1985; Motzev & Marchev, 1988; Motzev, 2014; Mueller & Lemke, 2003; Onwubolu, 2009). These techniques are discussed in Chapters 9 and 12.

***

SUMMARY AND CONCLUSIONS

A ***time series*** is a sequence of data points, measured typically at successive points in time spaced at uniform time intervals. A few basic techniques of ***Time series analysis*** and ***Forecasting*** were introduced and explored in Chapters 5 and 7**.** Chapter 8 discusses more advanced predictive methods, emphasizing on some important models and their application in contemporary business analysis and management.

- ***Time series analysis*** (***TSA***) comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. ***TSA*** methods are divided as: frequency-domain (spectral analysis) and time-domain methods (auto-correlation analysis); parametric and non-parametric; linear and non-linear methods.
- ***Time series models*** are used for predicting the future behavior of variables based on previously observed data. These models account for the fact that data points taken over time may have an internal structure (such as autocorrelation, trend or seasonal variation) that should be accounted for.

Time series forecasting is the use of a model to predict future values based on previously observed data. Different approaches and techniques are used and the most advanced are discussed in the current chapter:

A) Autoregression and ARIMA Methodology – ***Time series*** models estimate difference equations containing stochastic components. Two commonly used forms of these models are ***autoregressive models (AR)*** and ***moving average (MA)*** models.

- An ***autoregressive process*** operates under the premise that past values have an effect on current values. A process considered ***AR(1)*** is the ***first-order process***, meaning that the current value is based on the immediately preceding value.
- ***MA(q)*** refers to the ***MA*** model of order **q**, and conceptually is a linear regression of the current value of the series against current and previous (unobserved) error terms.
- ***Autoregressive–moving-average (ARMA)*** models provide a parsimonious description of a stationary stochastic process in terms of two polynomials, one for the ***AR*** and the second for the ***MA*** – in ***ARMA(p,q)*** model **p** is the order of the autoregressive part and **q** is the order of the moving average part.
- ***Autoregressive Integrated Moving Average models*** (***ARIMA***) are used to describe non-stationary time series data – in ***ARIMA(p,d,q)*** model parameters **p**, **d**, and **q** are non-negative integers that refer to the order of the autoregressive, integrated, and moving average parts of the model.

- Finding appropriate values of **p** and **q** in the ***ARMA(p,q)*** model can be facilitated by plotting the ***PACF*** (partial autocorrelation functions) for an estimate of **p**, and likewise using the ***ACF*** (autocorrelation functions) for an estimate of **q**.
- ***ARMA*** model is a univariate model. Its extension, the ***Autoregressive–moving-average model with exogenous inputs*** (***ARMAX***) is discussed in Chapter 9.
- ***ARIMA*** in fact is a procedure used to select from a group of forecasting models that best fit to the particular set of time series data. Box and Jenkins proposed a three-stage methodology which does not assume any particular pattern in the historical data of the series to be forecasted. Rather, it uses a three-step iterative approach of model identification, parameter estimation and validation.

B) Time Series Forecasting Using Data Mining Techniques – outside of traditional statistical modelling (stochastic methods) for time series modeling and forecasting, an enormous amount of forecasting is done using ***Data Mining*** techniques.

- Similar to the work of a human brain, ***Artificial Neural Networks (ANNs)*** try to recognize regularities and patterns in the input data, "learn" from experience and then provide generalized results based on their known previous knowledge.
- In a ***multi-layer feed-forward network*** (***FNN***), each layer of nodes receives inputs from the previous layers. The outputs of nodes in one layer are inputs to the next layer.
- In ***Time Lagged*** (***TLNN***) architecture the input nodes are the time series values at some particular time lags.
- The ***seasonal ANN*** (***SANN***) model does not require any preprocessing of raw data. ***SANN*** can learn the seasonal pattern in the time-series, without removing them, contrary to some other traditional approaches, such as the seasonal ***ARIMA***.
- A ***Deep Neural Network (DNN)*** is defined to be an artificial neural network with multiple hidden layers of units between the input and output layers. The extra layers enable composition of features from lower layers, giving the potential of modeling complex data with fewer units than a similarly performing shallow network.
- The most popular ***GMDH-type ANN*** is known as multi-layer inductive procedure which is equivalent to the ANNs with polynomial activation function of neurons.
- ***Multi-stage Selection Algorithms*** are specific ***GMDH-type ANNs*** that use the idea of ***Genetic Algorithms*** (***GA***) and multi-layer organization, similar to ***MLNAN***.

- ***Multi-Layered Net of Active Neurons (MLNAN)*** are very useful in business forecasting for building both multi-input to single-output models (for example different type of regression models) and econometric models of SE (i.e. multi-input to multi-output).
- ***GMDH-type NN*** are successfully used both in time series analysis and forecasting and in much more complex cases, such as distributed lag models, autoregressive moving average with exogenous inputs models (ARMAX), vector autoregression models (VAR) and different econometric models, including the most complicated models of simultaneous equations, as discussed further on in Chapters 9 and 12.

Key Terms

Artificial Neural Networks (ANNs)	***279***
Autocorrelation function (ACF)	***270***
Autoregressive integrated moving average models (ARIMA)	***269, 275***
Autoregressive models (AR)	***269***
Autoregressive moving average (ARMA)	***269, 273***
Data Mining	***279***
Deep learning algorithms, Deep Neural Network (DNN)	***283***
Freedom of choice	***285***
Genetic Algorithms (***GA***), ***Multi-Stage Selection Algorithms***	***285***
GMDH-type ANN, GMDH-type Neural Network, Polynomial NN	***284***
Group Method of Data Handling (GMDH)	***284***
Moving average (MA)	***269, 272***
Multi-layered feed-forward network” (***FNN***)	***280***
Multi-Layered Networks of Active Neurons (MLNAN)	***285***
Partial autocorrelation function (PACF)	***273***
Seasonal ANN	***282***
Seasonal autoregressive integrated moving average models (SARIMA)	***277***
Self-Organizing Data Mining (SODM)	***288***
Time Lagged NN	***282***

CHAPTER EXERCISES

Conceptual Questions:

1. What is an ***autoregressive model (AR)*** and how it works? Discuss.
2. List all the steps we should follow to compute an ***l-step-ahead forecast*** using ***AR*** models and explain how it works.
3. What is ***ARIMA***? Explain Box and Jenkins three-stage methodology.
4. Define what ***Artificial Neural Networks (ANNs)*** are. Discuss their strengths and weaknesses in TSA and forecasting.
5. What is a ***GMDH-type Neural Network?*** Explain its main characteristics.
6. Define what is a ***Multi-Layer Net of Active Neurons (MLNAN)*** and how it works.
7. Explain the similarities and the differences between traditional ANNs and MLNANs in model building and forecasting processes.

Business Applications (continue from Chapter 7):

Open Gretl program and import file SalesData.xslx:

- Set up the time series data for a model with dependent variable "Sales":
- Using ***ARIMA*** methodology develop an ***AR*** model, based on the data patterns detected within the time-series.
- Test the aptness of each model using test statistics provided by Gretl software and perform residual analysis. Do you see any violations of the regression assumptions? If yes return to the previous step and improve the ***AR*** model.
- Compute Sales forecast for the next 12 months.
- Design formulas, similar to the formulas in Part 4 of the Integrative case and compute MAD, MSE, MAPE and MPE for the new model, for a testing dataset of the 12 new monthly forecasts given in spreadsheet Predictions.
- What is the model accuracy? Are there any initial assumptions/reasons leading to this conclusion?

Discuss all findings and write a short report (up to two pages) summarizing your answers.

INTEGRATIVE CASE

HEALTHY FOOD SUPPLY CHAIN & STORES

Part 8: Time Series Analysis and Forecasting

In Chapter 1 we introduced *Healthy Food Stores* – a fast-growing retail food provider with 12 stores in a northwestern state. The company executives decided to study the effect that company advertising dollars have on sales. They hoped that examining collected historical data would reveal relationships that would help determine future advertising expenditures and predict monthly sales volumes for the upcoming quarter.

After identifying basic parameters, along with input (independent) and output (dependent) variables of the forecasting scenario in Part 2, the research team determined the main elements of the forecasting process:

- Forecasting horizon of up to twelve months;
- Quarterly forecast updates, since accuracy decreases as time horizon increases, and sufficient time is needed to implement possible changes;
- Development of different models based on data patterns, if any, and evaluation of their accuracy in order to select the most appropriate one;
- Selection of the best forecasting model with no more than 5% forecasting error.

The benchmark forecast computed in Part 3 (using the baseline of ***one-step naive forecast*** as a reference forecast) provided basic values for the most common measures of accuracy MFE, MAD, MAPE, MPE, MSE and CV(RMSE).

In Part 4, information about the opinion of some important people from the *Healthy Food Stores Company,* concerning this specific case, was collected and the research team applied the Delphi method to top executives group, Sales-force composite to the sales managers from all 12 stores and Scenario writing to the most experienced professionals from Advertising Department. After collecting such valuable information from different sources, in Part 5 the research team made its first steps in Numerical Predictions by developing different basic forecasting models. They created spreadsheets for Naïve techniques (Average model, Random Walk with Drift and Seasonal Naïve Technique), simple Moving Average, Simple Exponential Smoothing (SES) and Triple (Holt-Winters) Exponential Smoothing (TES), which were used to expand the base-line of one-step naïve forecast as reference forecasts.

In Part 6 the research team analyzed the relationships between Sales and all available predictors. After performing multiple correlation and regression analysis, researchers developed reliable forecasting model, representing the real system with certain error. In Part 7, the model was expanded by adding Dummy seasonal variables to analyze the Seasonal effect on company Sales.

In Part 8, the improvement of the forecasting model will be continued. It will be expanded by some advanced Time series analyses and relevant predictive techniques.

Case Questions

1. Open Gretl program and import the original version of file Data.xslx.
2. Set up the time series data for a model with dependent variable “Sales”:
   a) Split the sample into Training (36 observations) and Testing (12 observations) data sets.
   b) Develop different (at least three) ***AR*** models, based on the data patterns detected within the time-series, with the help of ***ARIMA*** methodology.
   c) Test the aptness of each model and perform residual analysis. Do you see any violations of the regression assumptions? If yes return to step a) and develop another ***AR*** model.
   d) Compute Sales forecast for the next 12 months.
3. Use (copy/paste) the formulas designed in Part 3 to compute MFE, MAD, MAPE, MPE, MSE and CV(RMSE) for each new model. Use the given testing dataset of 12 monthly forecasts provided in spreadsheet Errors.
4. Comment and analyze model’s accuracy - how good is the accuracy of these forecasts? Which model, out of all models so far provides the best accuracy? Discuss.
5. What overall recommendations would you make to the research team? Explain.
6. Write a report (at least two pages not counting charts and tables) on the questions above, discussing all important findings and draw relevant conclusions about this part of the Integrative Case.

References

Berry, M., & Linoff, G. (2000). *Mastering Data Mining*. Wiley.

Box, G., & Jenkins, G. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day.

Box, G., Jenkins, G., Reinsel, G., & Ljung, G. (2016). *Time Series Analysis: Forecasting and Control*. Wiley.

Chatfield, C. (1996). *The Analysis of Time Series*. (5th ed.). New York, NY: Chapman & Hall.

Commandeur, J., & Koopman, S. (2007). *Introduction to State Space Time Series Analysis*. Oxford University Press, §10.4.

Faraway, J., & Chatfield, C. (1998). Time series forecasting with neural networks: a comparative study using the airline data. *Applied Statistics, 47*, 231–250.

Gabor, D. (1971). *Perspectives of Planning. Organization of Economic Cooperation and Development*. London: Emp. College of Science and Technology.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feed-forward networks are universal approximators. *Neural Networks, 2*, 359–366.

Madala, H. R., & Ivakhnenko, A. G. (1994). *Inductive Learning Algorithms for Complex Systems Modelling*. Boca Raton, FL: CRC Press Inc.

Motzev, M. (1985). A New Approach for Simulation Models Building. *XVI IFAC/ISSAGA Workshop*. Alma-Ata, USSR.

Motzev, M. (2014). *Predictive Analytics in Business Games and Simulations*. Bielefeld, Germany: W. Bertelsmann Verlag GmbH & Co. KG.

Motzev, M., & Marchev, A. (1988). Multi-Stage Selection Algorithms in Simulation. *Proceedings of XII IMACS World Congress*. Paris, France: vol. 4, pp. 533-535.

Mueller J. A., & Lemke, F. (2003). *Self-Organizing Data Mining: An Intelligent Approach To Extract Knowledge From Data*. Victoria, BC: Trafford Publishing.

Nau, R. (2014). *Slides on seasonal and nonseasonal ARIMA models*. Durham, NC: Duke University. Retrieved from: http://people.duke.edu/~rnau/411arim.htm

Onwubolu, G. (ed.). (2009). *Hybrid Self-Organizing Modeling Systems*. Berlin Heidelberg: Springer-Verlag.

Shahwan, T., & Lemke, F. (2005). Forecasting Commodity Prices for Predictive Decision Support Systems. *Proceedings of EFITA/WCCA joint congress on IT in agriculture*. Vila Real, Portugal: pp. 23-32.

Whittle, P. (1951). *Hypothesis Testing in Time Series Analysis*. Almquist and Wicksell; Whittle, P. (1963). *Prediction and Regulation*. English Universities Press (Republished as: Whittle, P. (1983). *Prediction and Regulation by Linear Least-Square Methods*. University of Minnesota Press).

Yule, G. (1927). On a Method of Investigating Periodicities in Disturbed Series, with Special Reference to Wolfer's Sunspot Numbers. *Philosophical Transactions of the Royal Society of London, A, 226*, 267–298.

Zhang, G., Patuwo, B., & Hu, M. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting, 14*, 35-62.

Chapter 9. Complex Forecasting Techniques

9.1. Econometric Techniques and Models

Econometrics is the application of statistical and mathematical theories to economics for the purpose of testing hypotheses and forecasting future trends. Econometrics takes economic models and tests them through statistical trials. The results are then compared and contrasted against real-life examples.

Econometrics uses tools such as frequency distributions, probability and probability distributions, statistical inference, simple and multiple regression analysis, simultaneous equations models and time series methods. An example of a real-life application of econometrics would be studying the hypothesis that as a family’s income increases, their spending also increases.

A. Basic econometric models

Linear regression

The basic tool for econometrics is the linear regression model described in Chapter 6. In modern econometrics, other statistical tools are frequently used, but linear regression is still the most frequently used starting point for an analysis (Greene, 2012, p.7). Estimating a linear regression on two variables can be visualized as fitting a line through data points representing paired values of the independent and dependent variables. One classical example relates GDP growth to the unemployment rate. This relationship is represented in a linear regression where the change in unemployment rate (**Δ Unemployment**) is a function of an intercept (**β_0**), a given value of GDP growth multiplied by a slope coefficient (**β_1**) and an error term (**ε**):

$$\Delta\, Unemployment = \beta_0 + \beta_1 \mathrm{Growth} + \varepsilon. \qquad (9\text{-}1)$$

The fitted line is found using regression analysis. The unknown parameters **β_0** and **β_1** are estimated using OLS and the model could then be tested for statistical significance as to whether an increase in growth is associated with a decrease in the unemployment, as hypothesized. If the estimate of **β_1** were not significantly different from 0, the test would fail to find evidence that changes in the growth rate and unemployment rate were related.

Production function

A production function relates physical output of a production process to physical inputs or factors of production (see Fig.9-1). Its primary purpose is to address allocative efficiency in the use of factor inputs in production and the resulting distribution of income to those factors, while abstracting away from the technological problems of achieving technical efficiency.

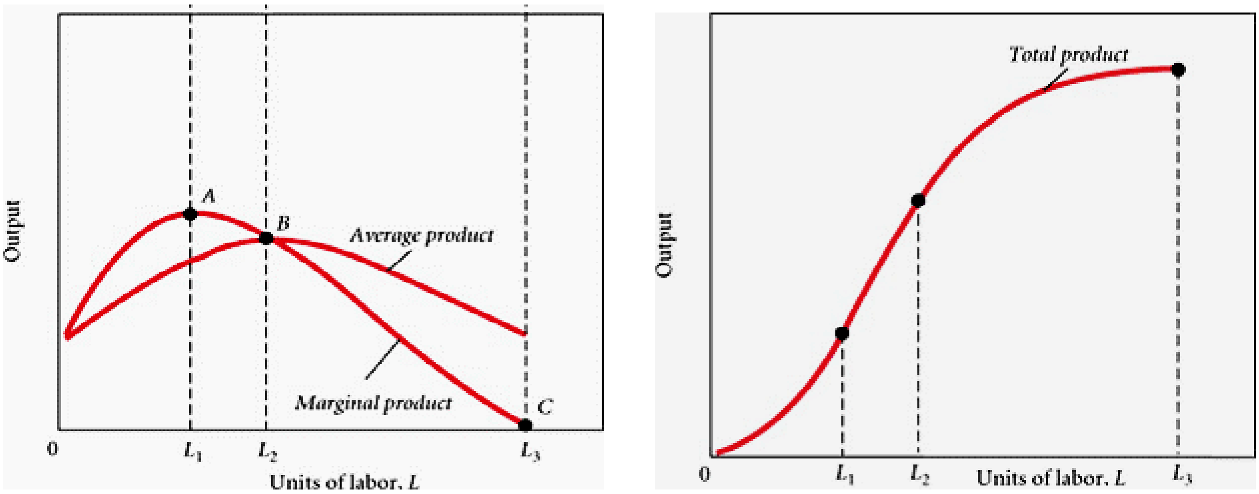


Fig.9-1 Graph of total, average, and marginal product

A production function can be expressed in a functional form as the right side of:

$$\boldsymbol{Q = f(X_1, X_2, \ldots, X_n)} \tag{9-2}$$

where **Q** is the quantity of output (i.e. dependent variable **Y**) and **X** ($\boldsymbol{X_i} = \{\boldsymbol{X_1, X_2, \ldots, X_n}\}$) are the quantities of factor inputs (such as capital, labor, land and/or raw materials).

One formulation, unlikely to be relevant in practice, is as a linear function:

$$Q = a + bX_1 + cX_2 + dX_3 + \cdots \tag{9-3}$$

where ***a, b, c, d*** are parameters that are determined empirically.

Another formulation is as a Cobb-Douglas production function:

$$Q = aX_1^b X_2^c \cdots . \tag{9-4}$$

In its most standard form for production of a single good with two factors, the function is

$$Y = AL^\beta K^\alpha \tag{9-5}$$

where Y = total production (the real value of all goods produced in a year)

L = labor input (the total number of person-hours worked in a year)

K = capital input (the real value of all machinery, equipment, and buildings)

A = total factor productivity

α and β are the output elasticity of capital and labor, respectively. These values are constants determined by available technology.

The Leontief production function applies to situations in which inputs must be used in fixed proportions. Starting from those proportions, if usage of one input is increased without another being increased, the output will not change. This production function is given by:

$$Q = \min(aX_1, bX_2, \dots). \quad \text{or} \quad q = \text{Min}\left(\frac{z_1}{a}, \frac{z_2}{b}\right) \qquad (9\text{-}6)$$

where ***q*** is the quantity of output produced

Z₁ and ***Z₂*** are the utilized quantities of input 1 and input 2 respectively, and

a and ***b*** are technologically determined constants.

Other forms include the constant elasticity of substitution production function, which is a generalized form of the Cobb-Douglas (1928) function, and the quadratic production function. The best form of the equation to use and the values of the parameters (**a, b, c, ... d**) vary from company to company and industry to industry. In a short run production function at least one of the **X**'s (inputs) is fixed. In the long run all factor inputs are variable at the discretion of management.

Supply and demand functions

Supply and demand is an economic model of price determination in a market. It concludes that in a competitive market, the unit price for a particular good will vary until it settles at a point where the quantity demanded by consumers (at current price) will equal the quantity supplied by producers (at current price), resulting in an economic equilibrium for price and quantity (see Fig.9-2).

The ***AD–AS*** or ***aggregate demand–aggregate supply model*** is a macroeconomic model that explains price level and output through the relationship of aggregate demand and aggregate supply. The AD/AS model is used to illustrate the Keynesian model of the business cycle. Movements of the two curves can be used to predict the effects that various exogenous events will have on two variables: real GDP and the price level.

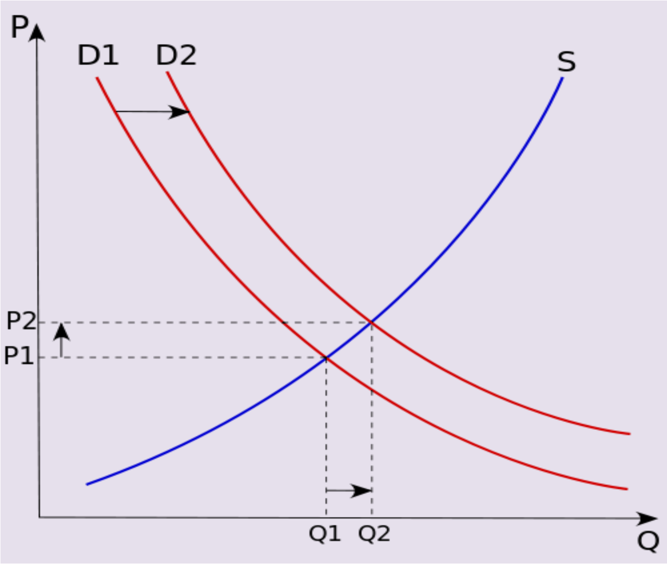


Fig.9-2 Supply and Demand chart

The price P of a product is determined by a balance between production at each price (supply S) and the desires of those with purchasing power at each price (demand D). For example, Fig.9-2 shows a positive shift in demand from D1 to D2 resulting in an increase in price (P) and quantity sold (Q) of the product.

Each curve has its own function, which can be identified with the corresponding equation. The equation for the AD curve in general terms is:

$$Y = Y^d(\tfrac{M}{P}, G, T, Z_1) \tag{9-7}$$

where **Y** is real **GDP**, **M** is the nominal money supply, **P** is the price level, **G** is real government spending, **T** is an exogenous component of real taxes levied, and **Z**$_1$ is a vector of other exogenous variables[1] that affect the location of the **IS** curve (exogenous influences on any component of spending) or the **LM** curve (exogenous influences on money demand) from the **IS–LM** model (Investment Saving–Liquidity Preference Money Supply).

The equation for the aggregate supply curve in general terms for the case of excess supply in the labor market, called the short-run aggregate supply curve, is:

$$Y = Y^s(W/P, \quad P/P^e, \quad Z_2) \tag{9-8}$$

where **W** is the nominal wage rate (exogenous due to stickiness in the short run), **P**^{e} is the anticipated price level (its expected value), and **Z**$_2$ is a vector of exogenous variables that can affect the position of the labor demand curve (the capital stock or the current state of technological knowledge).

Supply and Demand relations in a market can be statistically estimated from price, quantity, and other data with sufficient information in the model. This can be done with ***simultaneous-equation models***, discussed further on in this Chapter.

B. Common Econometric models

Probit regression

A ***probit*** model is a type of regression where the dependent variable can only take two values, for example married or not married. The name is from ***prob***ability + un***it*** (see Freedman, 2009, p.128). The purpose of the model is to estimate the probability that an observation with particular characteristics will fall into a specific one of the categories; moreover, if estimated probabilities greater than 1/2 are treated as classifying an observation into a predicted category, the probit model is a type of binary classification model.

A probit model is a popular specification for an ordinal or a binary response model. It treats the same set of problems as the ***logistic regression***, using similar techniques. The probit model, which employs a probit link function, is most often estimated using the standard maximum likelihood procedure (such an estimation is known as a ***probit regression***).

[1] See explanations given in VAR models in this section.

Logistic regression

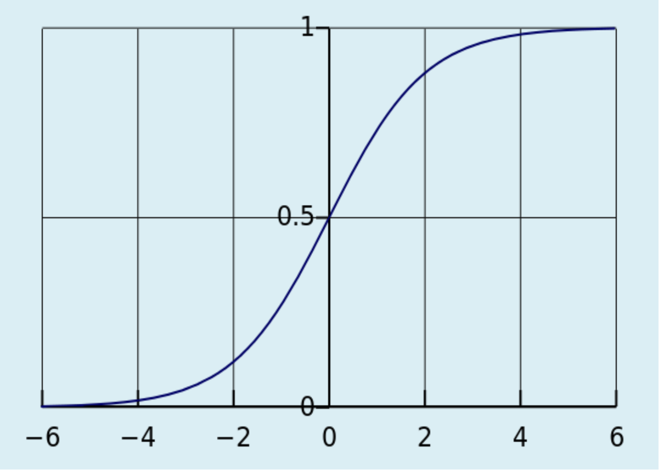


Fig.9-3 The logistic function

Logistic regression, or ***logit regression***, or ***logit model*** is a type of probabilistic statistical classification model (see Freedman, 2009, p.128). It is also used to forecast a binary response from a binary predictor, used for forecasting the outcome of a categorical dependent variable (i.e., a class label) based on one or more predictor variables (features).

It means that the ***logit*** is used in estimating the parameters of a qualitative response model. The probabilities describing the possible outcomes of a single trial are modeled, as a function of the explanatory (predictor) variables, using a logistic function (see Fig.9-3). Frequently "***logistic regression***" is used to refer specifically to the problem in which the dependent variable is binary (i.e. the number of available categories is two), while problems with more than two categories are referred to as ***multinomial logistic regression*** or, if the multiple categories are ordered, as ***ordered logistic regression***.

Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables, which are usually (but not necessarily) continuous, by using probability scores as the predicted values of the dependent variable. Thus, it treats the same set of problems as the ***probit regression*** using similar techniques – the first assumes a logistic function and the second a standard normal distribution function.

Logistic regression can be seen as a special case of a ***generalized linear model*** and thus analogous to linear regression. The model of logistic regression, however, is based on quite different assumptions about the relationship between dependent and independent variables from those of linear regression. In particular, the key differences between these two models can be seen in the following two features of logistic regression. *First*, the conditional mean of the variable follows a Bernoulli distribution rather than a Gaussian distribution, because logistic regression is a classifier. *Second*, the linear combination of the inputs (predictors) is restricted within {0,1} through the logistic distribution function because logistic regression predicts the probability of the instance being positive.

Like other forms of regression analysis, logit makes use of one or more predictor variables that may be either continuous or categorical data. Unlike ordinary linear regression, however, logistic regression is used for predicting binary outcomes of the dependent variable (treating the dependent variable as the outcome of a Bernoulli trial) rather than a continuous outcome. Given this difference, it is necessary that logistic regression takes the natural logarithm of the odds of the dependent variable being a case (known as the logit or log-odds) to create a continuous criterion as a transformed version of the dependent variable. Thus the logit transformation is referred to as the link function in logistic regression, although the dependent variable in logistic regression is binomial, the logit is the continuous criterion upon which linear regression is conducted.

$$g(x) = \ln \frac{F(x)}{1 - F(x)} = \beta_0 + \beta_1 x, \qquad (9\text{-}9)$$

and equivalently:

$$\frac{F(x)}{1 - F(x)} = e^{\beta_0 + \beta_1 x}. \qquad (9\text{-}10)$$

where ***g(x)*** refers to the logit function of some given linear combination ***x*** of the predictors. The equation (9-9) illustrates that the logit (i.e., log-odds or natural logarithm ln of the odds) is equivalent to the linear regression expression.

– ***F(x)*** is the probability that the dependent variable equals a case, given some linear combination ***x*** of the predictors. Equation (9-10) for ***F(x)*** illustrates that the probability of the dependent variable equaling a case is equal to the value of the logistic function of the linear regression. This is important because it shows that the value of the linear regression expression can vary from negative to positive infinity and yet, after transformation, the resulting expression for the probability ***F(x)*** ranges within {0,1}.

– base ***e*** denotes the exponential function**,** $\boldsymbol{\beta_0}$ is the intercept from the linear regression equation (the value of the criterion when the predictor equals zero) and $\boldsymbol{\beta_1}$ is the slope, i.e. the regression coefficient multiplied by some value ***x*** of the predictor

The logit model of success is then fitted using linear regression analysis. The predicted value of the logit is converted back into predicted odds via the inverse of the natural logarithm (the exponential function). Thus, although the observed dependent variable in ***logit*** is a zero-or-one variable, the logistic regression estimates the odds as a continuous variable, that the dependent variable is a success (a case). In some applications, the odds are all that is needed. In others, a specific yes-or-no prediction is needed for whether the dependent variable is or is not a case – this categorical prediction can be based on the computed odds of a success, with predicted odds above some chosen cutoff value being translated into a prediction of a success.

Because the logit model can be expressed as a generalized linear model for 0<p<1, ***OLS*** can suffice, with ***R-squared*** as the measure of goodness of fit in the fitting space. When p=0 or 1, more complex methods are required, such as the maximum likelihood estimation.

Goodness of fit in linear regression models is generally measured using the $\boldsymbol{R^2}$. Since this has no direct analog in logistic regression, various methods such as Deviance and likelihood ratio tests can be used instead.

If the estimated probabilities are to be used to classify each observation of independent variable values as predicting the category that the dependent variable is found in, the various methods for judging the model's suitability in out-of-sample forecasting can also be used on the data that were used for estimation (accuracy, precision and so on – see Chapter 3). The best way to measure a model's suitability is to assess the model against a set of data that was not used to create the model (see Mark & Goldberg, 2001), i.e. the ***cross-validation techniques***.

Vector autoregression (VAR)

Vector autoregression (VAR) is an econometric model used to capture the linear interdependencies among multiple time series. ***VAR*** models generalize the univariate ***autoregression (AR)*** models by allowing for more than one evolving variable. All variables in a VAR are treated symmetrically in a structural sense (although the estimated quantitative response coefficients will generally not be the same). Each variable has an equation explaining its evolution based on its own lags and the lags of the other model variables. ***VAR*** modeling does not require as much knowledge about the forces influencing a variable as do ***structural models with simultaneous equations***, discussed in the next Section. The only prior knowledge required is a list of variables which can be hypothesized to affect each other intertemporally.

A ***VAR*** model describes the evolution of a set of $\boldsymbol{k}$ variables $\boldsymbol{y_t}$ (known as ***endogenous variables***) over the same sample period $\boldsymbol{t}$ ($\boldsymbol{t} = \{1, 2..., T\}$) as a linear function of only their past (lag) values ($\boldsymbol{y_{t-l}}$). The variables are collected in a $\boldsymbol{k\times 1}$ vector $\boldsymbol{y_t}$, which has as the $\boldsymbol{i}^{\text{th}}$ element $\boldsymbol{y_{it}}$, the time $\boldsymbol{t}$ observation of the $\boldsymbol{i}^{\text{th}}$ variable. For example, if the $\boldsymbol{i}^{\text{th}}$ variable is GDP, then $\boldsymbol{y_{it}}$ is the value of GDP at time $\boldsymbol{t}$.

The term ***endogenous*** used above could be explained as follows – a parameter or variable is said to be ***endogenous*** when there is a correlation between the parameter or variable and the error term. ***Endogeneity*** can arise as a result of measurement error, autoregression with autocorrelated errors, simultaneity, and omitted variables. Broadly speaking, either an uncontrolled ***confounder***[2] causing both independent and dependent variables of a model or a loop of causality between the independent and dependent variables of a model leads to ***Endogeneity***.

For example, in a simple supply and demand model, when predicting the quantity demanded in equilibrium, the price is ***endogenous*** because producers change their price in response to demand and consumers change their demand in response to price. In this case, the price is said to have total ***endogeneity*** once the demand and supply curves are known. In contrast, a change in consumer preferences is an ***exogenous*** change on the demand curve.

A **p**th order VAR, denoted ***VAR(p)***, is presented as:

$$y_t = c + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + e_t, \qquad (9\text{-}11)$$

where the ***l***-periods back observation $\boldsymbol{y_{t-l}}$ is the ***l***th lag of $\boldsymbol{y_t}$,

c is a **k×1** vector of constants (intercepts),

Ai is a time-invariant **k×k** matrix and

$\boldsymbol{e_t}$ is a **k×1** vector of error terms which are generally assumed to be independent (i.e. there is no correlation across time – in particular, no serial correlation in individual error terms), identically distributed (***i.i.d.***) random variables, sampled from a normal distribution with zero mean, i.e. $\boldsymbol{e_t} \sim \mathrm{N}(\mathbf{0},\boldsymbol{\sigma}^2)$ where $\boldsymbol{\sigma}^2$ is the variance.

A **p**th order **VAR** is also called a **VAR** with **p** lags. The process of choosing the maximum lag **p** in the VAR model requires special attention because the inference is dependent on the correctness of the selected lag order.

If we consider the ***General Linear Model (GLM)***

$$\mathbf{Y = B\,X + U} \qquad (9\text{-}12)$$

where **Y** is the dependent variables ***y*** matrix with a series of multivariate measurements

X is a design matrix of values of explanatory variables ***x***

B is a matrix containing parameters that are usually to be estimated and

U is a matrix containing errors or noise – the errors are usually assumed to be uncorrelated across measurements, and follow a multivariate normal distribution.

The ***GLM*** incorporates a number of different statistical models: ANOVA, ANCOVA, MANOVA, MANCOVA, ordinary linear regression, t-test and F-test. The ***GLM*** is a generalization of a multiple linear regression model to the case of more than one dependent

[2] Or a ***controlled variable*** (i.e. a variable which is kept constant or monitored to try to minimize its effect on the experiment) that correlates (directly or inversely) with both the dependent variable and the independent variable.

variable ***y***. If **Y, B**, and **U** were column vectors, the matrix equation (9-12) would represent a multiple linear regression.

In terms of the general linear model we can present a ***VAR(p)*** as:

$$\mathbf{Y = B\,Z + U} \tag{9-13}$$

where **Z** is a design matrix of lag values of the dependent variables ***y****ᵢ*.

The ***Multivariate Least Squares (MLS)*** for **B** yields:

$$\hat{B} = YZ'(ZZ')^{-1} \tag{9-14}$$

This estimator is consistent and asymptotically efficient. Furthermore, it is equal to the conditional maximum likelihood estimator (see Zellner, 1962). As the explanatory variables are the same in each equation, the ***MLS*** estimator is equivalent to the ***Ordinary Least Squares (OLS)*** estimator applied to each equation separately.

An estimated ***VAR*** model then can be used for forecasting, and the quality of the forecasts can be judged, in ways that are completely analogous to the methods used in univariate autoregressive modeling discussed in Chapter 8, Section 8.1. A. Autoregressive models (***AR***)

Autoregressive–moving-average model with exogenous inputs (ARMAX)

The notation ***ARMAX(p, q, b)*** refers to the model with ***p*** autoregressive terms, ***q*** moving average terms and ***b*** exogenous inputs terms. This model contains the **AR(*p*)** and **MA(*q*)** models and a linear combination of the last ***b*** terms of a known and external time series $\mathbf{d_t}$. It is given by:

$$X_t = \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \sum_{i=0}^{b} \eta_i d_{t-i}. \tag{9-15}$$

where $\eta_1, \ldots, \eta_b$ are the parameters of the exogenous input $\mathbf{d_t}$.

Some nonlinear variants of models with exogenous variables have been defined, such as the ***Nonlinear Autoregressive Exogenous Model (NARX)*** – a nonlinear autoregressive model which has exogenous inputs. This means that the model relates the current value of a time-series where one would like to predict to both past values of the same series and current and past values of the driving (exogenous) series – that is, of the externally determined series that influences the series of interest. In addition, the model contains an error term, which relates to the fact that knowledge of the other terms will not enable the current value of the time series to be predicted exactly. Such a model can be stated as:

$$y_t = F\left(y_{t-1}, y_{t-2}, y_{t-3}, \ldots, u_t, u_{t-1}, u_{t-2}, u_{t-3}, \ldots\right) + \varepsilon_t \qquad (9\text{-}16)$$

where $\boldsymbol{y_t}$ is the variable of interest, and $\boldsymbol{u}$ is the externally determined variable – in this scheme, information about $\boldsymbol{u}$ helps predict $\boldsymbol{y_t}$, as do previous values of $\boldsymbol{y_t}$ itself

$\boldsymbol{\varepsilon_t}$ is the error term (or noise).

The function **F** is some nonlinear function, such as a polynomial. **F** can be a neural network, a wavelet network, a sigmoid network and so on (see Chapter 11) and the estimation of its unknown parameters is done according to the particular type of **F**.

C. Criticisms of Econometrics

There have been many criticisms of econometrics' usefulness as a discipline and perceived widespread methodological shortcomings in econometric modeling practices. Like other forms of statistical analysis, badly specified econometric models may show a spurious relationship where two variables are correlated but causally unrelated.

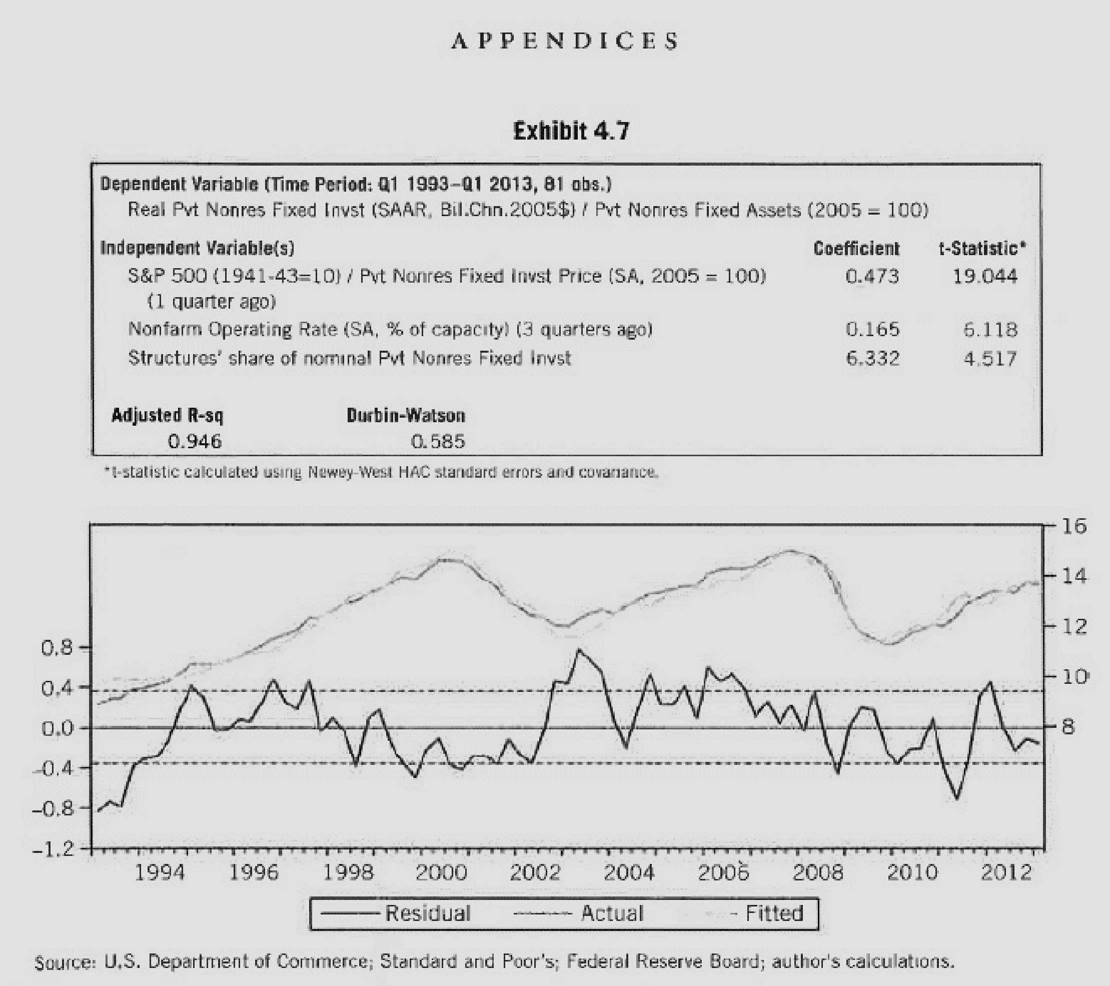
In some cases, economic variables cannot be experimentally manipulated as treatments randomly assigned to subjects. In such cases, economists rely on observational studies, often using data sets with many strongly associated covariates, resulting in enormous numbers of models with similar explanatory ability but different covariates and regression estimates.

One recent example could be found in the book written by Alan Greenspan, the former Chairman of the Board of Governors of the Federal Reserve System, who served for the longest period of time so far (approximately 19 years), for four Presidents of the USA[3]. In his book, the author is using many econometric equations that are questionable since most of them have significant autocorrelation. In such a case, the estimated model violates the assumption that errors are random and independent, and the forecasts may be inefficient (recall Chapter 8).

To test autocorrelation significance, a decision is made, at given significance level ($\boldsymbol{\alpha}$), by comparing the Durbin-Watson test statistic $\boldsymbol{d}$ with the lower and upper critical values ($\mathbf{d_L}$ and $\mathbf{d_U}$). These critical values vary by the level of significance ($\boldsymbol{\alpha}$), the number of observations ($\boldsymbol{T}$), and the number of predictors ($\boldsymbol{k}$) in the regression equation. Their derivation is complex and users typically obtain them from Tables in appendices of statistical texts or online[4]. If computed $\mathbf{d < d_L}$ (given α), then we reject the null hypothesis and accept the $\boldsymbol{H_A}$, which means that there is statistical evidence that the error terms are positively autocorrelated.

[3] See Greenspan, A., The Map and the Territory (Risk, Human Nature, and the Future of Forecasting), The Penguin Press, New York, 2013.

[4] See http://web.stanford.edu/~clint/bench/dwcrit.htm



APPENDICES

Exhibit 4.7

Dependent Variable (Time Period: Q1 1993–Q1 2013, 81 obs.)
Real Pvt Nonres Fixed Invst (SAAR, Bil.Chn.2005$) / Pvt Nonres Fixed Assets (2005 = 100)

Independent Variable(s)	**Coefficient**	**t-Statistic***
S&P 500 (1941-43=10) / Pvt Nonres Fixed Invst Price (SA, 2005 = 100) (1 quarter ago)	0.473	19.044
Nonfarm Operating Rate (SA, % of capacity) (3 quarters ago)	0.165	6.118
Structures' share of nominal Pvt Nonres Fixed Invst	6.332	4.517

Adjusted R-sq	**Durbin-Watson**
0.946	0.585

*t-statistic calculated using Newey-West HAC standard errors and covariance.

Source: U.S. Department of Commerce; Standard and Poor's; Federal Reserve Board; author's calculations.

Fig.9-4 Example of econometric model with significant autocorrelation
(source: Greenspan, 2013, p. 321)

Fig. 9-4 represents one of the models with Durbin-Watson statistics **d** = 0.585. The critical values ($\mathbf{d}_\mathrm{L}$ and $\mathbf{d}_\mathrm{U}$) for 3 independent variables and 81 observations at given significance level ***α = 0.05*** are:

$$\mathbf{d}_\mathrm{L} = 1.58875 \text{ and } \mathbf{d}_\mathrm{U} = 1.68976$$

That is, there is statistical evidence that error terms are positively autocorrelated and the forecasts may be inefficient.

Artificial Intelligence (***AI***) has become important for building econometric models and for use in decision making (see Tshilidzi, 2013). ***AI*** allows economic models to be of arbitrary complexity and also to be able to evolve as the economic environment also changes. For example, artificial intelligence has been applied to simulate the stock market, to model options and derivatives as well as model and control interest rates. Other useful applications of ***Data mining*** techniques as part of ***AI*** are discussed in the next sections.

9.2. Simultaneous Equations Models

A. Structural and reduced form

The traditional and most popular formal language used in econometrics is the ***structural equation model (SEM)***. While ***SEM***s are not the only type of econometric model, they are the primary subject of many econometrics textbooks that we have encountered. The foundations for structural equation modeling in economics were laid by Trygve Magnus Haavelmo (1943). According to Pearl (2014) “Haavelmo's paper, “The statistical implications of a system of simultaneous equations” marks a pivotal turning point, not in the statistical implications of econometric models, as historians typically presume, but in their causal counterparts”. Haavelmo's idea that an economic model depicts a series of hypothetical experiments and that policies can be simulated by modifying equations in the model became the basis of all currently used formalisms of causal inference. In 1989, Haavelmo was awarded the Nobel Prize in Economics "for his clarification of the probability theory foundations of econometrics and his analyses of simultaneous economic structures." (Prokesch, 1989)

According to Haavelmo (1943), “*Measurement of parameters occurring in theoretical equation systems is one of the most important problems of econometrics. If our equations were exact in the observable economic variables involved, this problem would not be one of statistics, but a purely mathematical one of solving a certain system of "observational" equations, having the parameters in question as unknowns. This might itself present complicated and interesting problems, such as the problem of whether or not there is a one-to-one correspondence between each system of values of the parameters and the corresponding set of all values of the variables satisfying the equation system. For example, if we have, simultaneously, a demand curve and a supply curve, the set of possible observations might be just one single intersection point, and knowing that only would not, in general, permit us to draw any inference regarding the slope of either curve. Real statistical problems arise if the equations in question contain certain stochastical elements ("unexplained residuals"), in addition to the variables that are given or directly observable. And some such element must, in fact, be present in any equation which shall be applicable to actual observations (unless the equation in question is a trivial identity). In other words, if we consider a set of related economic variables, it is, in general, not possible to express any one of the variables as an exact function of the other variables only. There will be an "unexplained rest," and, for statistical purposes, certain stochastical properties must be ascribed to this rest, a priori. Personally I think that economic theorists have, in general, paid too little attention to such stochastical formulation of economic theories. For the necessity of introducing "error terms" in economic relations is not merely a result of statistical errors of* *measurement. It is as much a result of the very nature of economic behavior, its dependence upon an enormous number of factors, as compared with those which we can account for, explicitly, in our theories. We need a stochastical formulation to make simplified relations elastic enough for applications.*

This is, perhaps, generally realized among econometricians. But they frequently fail to consider, in full, the statistical implications of assuming a system of such stochastical equations to be simultaneously fulfilled by the data. More specifically, if one assumes that the economic variables considered satisfy, simultaneously, several stochastic relations, it is usually not a satisfactory method to try to determine each of the equations separately from the data, without regard to the restrictions which the other equations might impose upon the same variables. That this is so is almost self-evident, for in order to prescribe a meaningful method of fitting an equation to the data, it is necessary to define the stochastical properties of all the variables involved (e.g., that some of them are given time series, or remain constant, etc.). Otherwise, we shall not know the meaning of the statistical results obtained. Furthermore, the stochastical properties ascribed to the variables in one of the equations should, naturally, not contradict those that are implied by the other equations." (pp.1-2].

Haavelmo's work and analyses established the beginning of the ***Simultaneous Equation*** (***SE***) models – statistical models in the form of a set of linear simultaneous equations – one of the most advanced and at the same time the most complex econometric models. A variable could be dependent in one equation and a regressor in others. Traditionally, all dependent variables that are determined by the model are called ***endogenous*** or ***jointly determined variables***. Those determined from outside are referred to as ***exogenous*** or ***predetermined variables*** (Maddala & Lahiri, 2009, p.357).

Suppose there are ***m*** regression equations of the form:

$$y_{it} = y'_{-i,t}\gamma_i + x'_{it}\beta_i + u_{it}, \quad i = 1, \ldots, m, \qquad (9\text{-}17)$$

where $\boldsymbol{i}$ is the equation number, and $\boldsymbol{t}$ ($t=\{1, 2, \ldots, T\}$) is the observation index

$\boldsymbol{x_{it}}$ is the $\boldsymbol{k_i}\times 1$ vector of exogenous variables $\boldsymbol{x_t}$ $(x_t = \{x_{1t}, x_{2t}, \ldots x_{kt},\})$, ($\boldsymbol{k_i} < \boldsymbol{k}$)

$\boldsymbol{y_{it}}$ is the dependent variable

$\boldsymbol{y_{-i,t}}$ is the $\boldsymbol{m_i}\times 1$ vector of all other endogenous variables which enter the i^{th} equation on the right-hand side – the "$\boldsymbol{-i}$" notation indicates that the vector $\boldsymbol{y_{-i,t}}$ may contain any of the $\boldsymbol{y}$'s except for $\boldsymbol{y_{it}}$ since it is already present on the left-hand side, i.e. $\boldsymbol{m_i} < \boldsymbol{m}$ is the number of dependent variables presented on the right-hand side

$\boldsymbol{u_{it}}$ are the error terms.

The regression coefficients $\boldsymbol{\beta_i}$ and $\boldsymbol{\gamma_i}$ are of dimensions $\boldsymbol{k_i\times 1}$ and $\boldsymbol{m_i\times 1}$ correspondingly.

Vertically stacking the ***T*** observations corresponding to the ***i***th equation, we can write each equation in vector form as:

$$y_i = Y_{-i}\gamma_i + X_i\beta_i + u_i, \quad i = 1, \ldots, m, \tag{9-18}$$

where $\boldsymbol{y_i}$ and $\boldsymbol{u_i}$ are $\boldsymbol{T}$×**1** vectors,

$\boldsymbol{X_i}$ is a $\boldsymbol{T}$×$\boldsymbol{k_i}$ matrix of exogenous regressors, and

$\boldsymbol{Y_{-i}}$ is a $\boldsymbol{T}$×$\boldsymbol{m_i}$ matrix of endogenous regressors on the right-hand side of the ***i***th equation

Finally, we can move all endogenous variables to the left-hand side and write the ***m*** equations jointly in a vector form as:

$$\mathbf{Y\,\Gamma = X\,B + U} \tag{9-19}$$

where $\boldsymbol{Y = [y_1\, y_2 \ldots y_m]}$ is the $\boldsymbol{T}$×$\boldsymbol{m}$ matrix of dependent variables. Each of the matrices $\boldsymbol{Y_{-i}}$ is in fact an $\boldsymbol{m_i}$-columned submatrix of this matrix ***Y***

– The $\boldsymbol{m}$×$\boldsymbol{m}$ matrix **Γ**, which describes the relation between the dependent variables, has a complicated structure. It has values of one on the diagonal, and all the other elements of each column ***i*** are either the components of the vector $\boldsymbol{-\gamma_i}$ or zeros, depending on which columns of ***Y*** were included in the matrix $\boldsymbol{Y_{-i}}$.

– The $\boldsymbol{T}$×$\boldsymbol{k}$ matrix ***X*** contains all exogenous regressors from all equations, but without repetitions (that is, matrix ***X*** should be of full rank). Thus, each $\boldsymbol{X_i}$ is a $\boldsymbol{k_i}$-columned submatrix of the matrix ***X***

– Matrix **B** has size $\boldsymbol{k}$×$\boldsymbol{m}$, and each of its columns consists of the components of vectors $\boldsymbol{\beta_i}$ and zeros, depending on which of the regressors from ***X*** were included or excluded from $\boldsymbol{X_i}$

– $\boldsymbol{U = [u_1\, u_2 \ldots u_m]}$ is a $\boldsymbol{T}$×$\boldsymbol{m}$ matrix of the error terms.

The representation (9-19) is what Haavelmo called the ***structural form*** of the ***SE***. Postmultiplying (9-19) by $\mathbf{\Gamma^{-1}}$, the system can be written in the ***reduced form***:

$$Y = XB\Gamma^{-1} + U\Gamma^{-1} = X\Pi + V. \tag{9-20}$$

The reduced form of a system of ***SE*** is the result of solving the system for the endogenous variables, i.e. to find the reduced form (9-20), we must solve the structural equations for the endogenous variables, which reduces the system (9-19) considerably. As shown in Fig.9-5, the ***SE*** system (9-20) is presented as a function of the exogenous variables only, if any. In econometrics, "***structural form***" models begin from deductive theories of the economy, while "***reduced form***" models begin by identifying particular relationships between variables.

The reduced system (9-20) is a simple general linear model and it can be estimated for example by **OLS**. Unfortunately, the task of decomposing the estimated matrix $\hat{\mathbf{\Pi}}$ into the individual factors **B** and $\mathbf{\Gamma^{-1}}$ is quite complicated, and therefore the reduced form is more suitable for prediction but not inference.

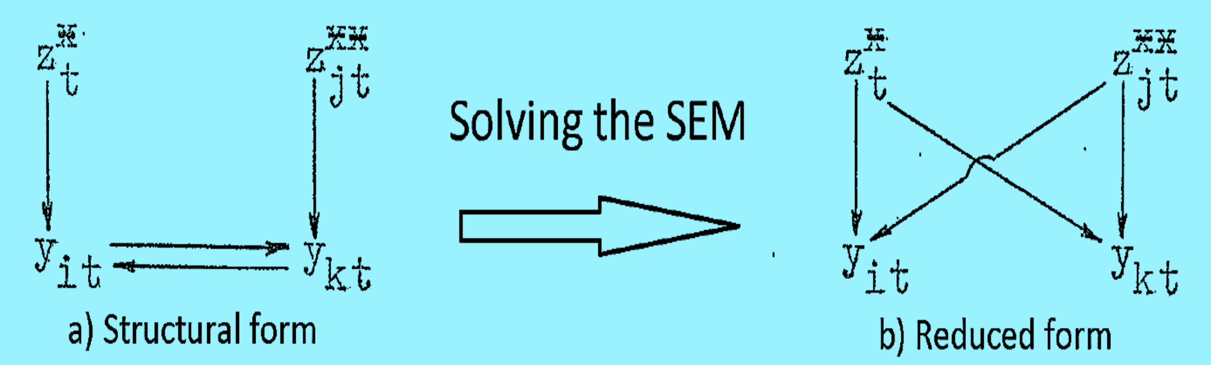


Fig.9-5 Transforming the SE structural form into a reduced form

B. Assumptions

Firstly, the rank of the matrix $\boldsymbol{X}$ of exogenous regressors must be equal to $\boldsymbol{k}$, both in finite samples and in the limit as $\boldsymbol{T} \rightarrow \infty$ (this later requirement means that in the limit the expression should $\frac{1}{T}\boldsymbol{X'X}$ converge to a non-degenerate $\boldsymbol{k}\times\boldsymbol{k}$ matrix). Matrix $\boldsymbol{\Gamma}$ is also assumed to be non-degenerate.

Secondly, error terms are assumed to be serially independent and identically distributed (**i.i.d.**) with $\boldsymbol{u_t} \sim \mathbf{N}(\mathbf{0},\boldsymbol{\sigma}^2)$, where $\boldsymbol{\sigma}^2$ is the variance.

Lastly, the ***identification conditions*** require that the number of unknowns in this system of equations should not exceed the number of equations. More specifically: (a) the ***order condition*** requires that for each equation $\boldsymbol{k_i} + \boldsymbol{m_i} \leq \boldsymbol{k}$, which can be phrased as "*the number of excluded exogenous variables is greater or equal to the number of included endogenous variables*"; (b) the ***rank condition*** is that rank $(\boldsymbol{\Pi_{i0}}) = \boldsymbol{m_i}$, where $\boldsymbol{\Pi_{i0}}$ is a $(\boldsymbol{k} - \boldsymbol{k_i})\times\boldsymbol{m_i}$ matrix which is obtained from $\boldsymbol{\Pi}$ by crossing out those columns which correspond to the excluded endogenous variables, and those rows which correspond to the included exogenous variables (for more details and mathematical explanations see Maddala & Lahiri, 2009, pp.396-397).

C. Parameter identification problem

In statistics and econometrics, the parameter identification problem is the inability in principle to identify the best estimate of the value(s) of one or more parameters in a regression. This problem can occur in the estimation of multiple-equation (***SE***) econometric models, where the equations have variables in common.

Identifiability is a property which a model must satisfy in order for precise inference to be possible. We say that the model is identifiable if it is theoretically possible to learn the true value of this model's underlying parameter after obtaining an infinite number of observations from it. Mathematically, this is equivalent to saying that different values of the parameter must generate different probability distributions of the observable variables. Usually, the model is identifiable only under certain technical restrictions, in which case the set of these requirements is called the ***identification conditions***.

A model that fails to be ***identifiable*** is said to be ***non-identifiable*** or ***unidentifiable*** when two or more parametrizations are observationally equivalent. In some cases, even though a model is non-identifiable, it is still possible to learn the true values of a certain subset of the model parameters. In this case, we say that the model is ***partially identifiable***. In other cases, it may be possible to learn the location of the true parameter up to a certain finite region of the parameter space, in which case the model is ***set identifiable***.

Consider a linear system of **M** equations, with **M > 1**. An equation cannot be identified from the data if less than **M-1** variables are excluded from that equation. In the simplest case of two equations, the parameters of an equation can be identified if *it is known that some variable does not enter into the equation, while it does enter the other equation*. This is a particular form of the order condition for identification. (The general form of the order condition also deals with restrictions other than exclusions.) The order condition is necessary but not sufficient for identification.

The rank condition is a necessary and sufficient condition for identification. In the case of only exclusion restrictions, it must "*be possible to form at least one non-vanishing determinant of order M-1 from the columns of A corresponding to the variables excluded a priori from that equation*" (Fisher, 1966, p. 40), where **A** is the matrix of coefficients of the equations. This is the generalization in matrix algebra of the above mentioned requirement "*that some variable does not enter into the equation, while it does enter the other equation.*"

According to Valavanis (1959, pp.93-94) counting rule, in a ***SE*** model (9-18) with $\boldsymbol{m}$ dependent variables $\boldsymbol{y_{it}}$ and $\boldsymbol{n}$ regressors (both exogenous $\boldsymbol{x_{it}}$ and endogenous $\boldsymbol{y_{-i,t}}$), the equations in ***SE*** (9-17) are:

- ***Exactly identified*** if:

$$\mathbf{n - n_i = m_i - 1} \tag{9-21}$$

where $\mathbf{n_i}$ is the number of all regressors (both exogenous $\boldsymbol{x_{it}}$ and endogenous $\boldsymbol{y_{-i,t}}$) in the *i*th equation only and $\mathbf{m_i}$ is the number of dependent variables $\boldsymbol{y_{it}}$ in the *i*th equation only.

- ***Overidentified*** if**:**

$$\mathbf{n - n_i > m_i - 1} \tag{9-22}$$

- ***Underidentified*** if**:**

$$\mathbf{n - n_i < m_i - 1} \tag{9-23}$$

The parameters are ***underidentified*** (equivalently, ***not identified***) if there are fewer exogenous regressors than there are covariates or, equivalently, if there are fewer excluded exogenous regressors than there are endogenous covariates in the equation of interest.

More generally, the term ***overidentified*** can be used to refer to any situation where a statistical model will invariably have more than one set of parameters that generate the same distribution of observations, meaning that multiple parametrizations are ***observationally equivalent***. In econometrics, two parameter values are considered ***observationally equivalent*** if they both result in the same probability distribution of observable data (Koopmans, 1949).

In science, ***observational equivalence*** is the property of two or more underlying entities being indistinguishable on the basis of their observable implications. Thus, for example, two scientific theories are observationally equivalent if all of their empirically testable predictions are identical, in which case empirical evidence cannot be used to distinguish which is closer to being correct – indeed, it may be that they are actually two different perspectives on one underlying theory[5].

D. Model specification

Two main components are distinguished in an ***SE*** model – the structural form showing potential causal dependencies between endogenous and exogenous variables, and the reduced (measurement) form representing the relations between latent variables and their indicators.

Path diagrams[6] can be viewed as ***SE*** models that contain only the structural part. In specifying pathways in a model, the forecaster can posit two types of relationships: (a) free pathways, in which hypothesized causal (in fact counterfactual) relationships between variables are tested, and therefore are left free to vary, and (b) relationships between variables that already have an estimated relationship, usually based on previous studies, which are considered as fixed in the model.

A forecaster will often specify a set of theoretically plausible models in order to assess whether the model proposed is the best of the set of possible models. The forecaster not only must account for the theoretical reasons for building the model as it is but must also take into account the number of data points and the number of parameters that must be estimated to identify the model. In an identified model a specific parameter value uniquely identifies the model, and no other equivalent formulation can be given by a different parameter value (see the comments about ***observational equivalence*** above).

[5] http://en.wikipedia.org/wiki/Observational_equivalence

[6] ***Path diagrams*** and ***path analysis*** are used to describe the directed dependencies among a set of variables.

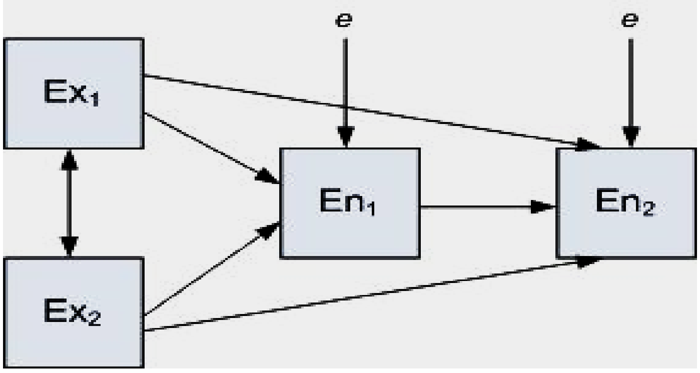


Fig.9-6 Example of Path modeling[7]

A data point is a variable with observed scores, like a variable containing the scores on a question or the number of times respondents buy a TV. The parameter is the value of interest, which typically is a regression coefficient between the exogenous and the endogenous variable. If there are fewer data points than the number of estimated parameters, the resulting model as discussed above is "unidentified", since there are too few reference points to account for all the variance in the model. The solution is to change the model specification, that is to constrain one (or more) of the paths to zero, which means that it is no longer part of the model, i.e. to reduce the size of the matrix of unknown parameters.

For example, as shown in Fig.9-6, two exogenous variables (**Ex**$_1$ and **Ex**$_2$) are modeled as being correlated and as having both direct and indirect (through **En**$_1$) effects on **En**$_2$. **En**$_1$ and **En**$_2$ are the interdependent endogenous variables in the SE model.

In most real models, the endogenous variables are also affected by factors outside the model including the measurement error. The effects of such extraneous variables are depicted by the "**e**" or error terms in the model.

Using the same variables, alternative models are conceivable. For example, it may be hypothesized that **Ex**$_1$ has only an indirect effect on **En**$_2$, deleting the arrow from **Ex**$_1$ to **En**$_2$, and the likelihood or the fit of these two models can be compared statistically.

E. Estimation

Indirect least squares (ILS)

Because of the ***parameter identification problem***, discussed above, ***OLS*** estimation of the structural form (9-19) would yield inconsistent parameter estimates. This problem can be overcome by rewriting the ***SE*** model in the reduced form (9-20).

[7] Source: http://en.wikipedia.org/wiki/Path_analysis_%28statistics%29

The simplest technique to fit an ***SE*** model is known as ***Indirect Least Squares (ILS)***. It is an approach where the coefficients in an ***SE*** model are estimated from the reduced form model (9-20) using ***OLS***. For this, the structural system of equations (9-19) is transformed into the reduced form first. Once the coefficients are estimated the model is put back into the structural form (see Park, 1974).

Two-stage least squares (2SLS)

The most common estimation method for the simultaneous equations model is known as the ***two-stage least squares method*** (***2SLS***), developed independently by Theil (1971) in 1953 and Basmann (1957) in 1957. It is an ***equation-by-equation*** technique, where the endogenous regressors on the right-hand side of each equation are being instrumented with the regressors ***X*** from all other equations. The method is called "two-stage" because it conducts estimation in two steps:

Step 1: Regress $\boldsymbol{Y}_{-i}$ on ***X*** using the equations from the reduced form (9-20) and obtain the predicted values $\hat{Y}_{-i}$

Step 2: Estimate γ_i, β_i by the OLS of y_i on $\hat{Y}_{-i}$ and $\boldsymbol{X_i}$.

If the i^{th} equation in the model (9-18) is written as

$$y_i = \begin{pmatrix} Y_{-i} & X_i \end{pmatrix} \begin{pmatrix} \gamma_i \\ \beta_i \end{pmatrix} + u_i \equiv Z_i \delta_i + u_i, \qquad (9\text{-}24)$$

where $\boldsymbol{Z_i}$ is a $T \times n_i$ ($n_i = m_i + k_i$) matrix of both endogenous and exogenous regressors in the i^{th} equation (often referred to as ***instruments***), and

$\boldsymbol{\delta_i}$ is an n_i ($n_i = m_i + k_i$)-dimensional vector of regression coefficients

Then the ***2SLS*** estimator of $\boldsymbol{\delta_i}$ will be given by

$$\hat{\delta}_i = (\hat{Z}_i' \hat{Z}_i)^{-1} \hat{Z}_i' y_i = (Z_i' P Z_i)^{-1} Z_i' P y_i, \qquad (9\text{-}25)$$

where $P = X(X'X)^{-1}X'$ is the projection matrix onto the linear space spanned by the exogenous regressors ***X***.

The so estimated ***SE*** model then can be used for forecasting either equation-by-equation or using the whole system of equations as one model (such as (9-19) for example), as discussed later on in this Chapter.

Other general techniques for SE models estimation

The ***Limited Information Maximum Likelihood (LIML)*** method was suggested by Anderson & Rubin (1949). It is used when one is interested in estimating a single structural equation at a time (hence its name of limited information), say for observation ***i***:

$$y_i = Y_{-i}\gamma_i + X_i\beta_i + u_i \equiv Z_i\delta_i + u_i \tag{9-26}$$

where $\mathbf{Y}_{-i}$ is the matrix of the endogenous variable(s).

$\boldsymbol{X_i}$ is the matrix of the exogenous variable(s)

$\boldsymbol{Z_i}$ is the matrix of the instruments, i.e. all endogenous and exogenous variables used in the right part of the *i*th equation.

The structural equations for the remaining endogenous variables $\mathbf{Y}_{-1}$ are not specified, and they are given in their reduced form:

$$Y_{-i} = X\Pi + U_{-1} \tag{9-27}$$

The explicit formula for the LIML is:

$$\hat{\delta}_i = \left(Z_i'(I - \lambda M)Z_i\right)^{-1} Z_i'(I - \lambda M)y_i, \tag{9-28}$$

where $\boldsymbol{M = I - X(X'X)^{-1}X'}$

In fact, the ***LIML*** is a special case of the ***K-class estimators***:

$$\hat{\delta} = \left(Z'(I - \kappa M)Z\right)^{-1} Z'(I - \kappa M)y, \tag{9-29}$$

with $\delta = \begin{bmatrix}\beta_i & \gamma_i\end{bmatrix}$ and

$Z = \begin{bmatrix}X_i & Y_{-i}\end{bmatrix}$

It should be noted that several estimators belong to this class. When:

κ=0, this is the ***OLS***;

κ=1, it is the ***2SLS*** (note indeed that in this case $I - \kappa M = I - M = P$ is the usual projection matrix of the ***2SLS***)

κ=λ, this is the ***LIML***

κ=λ-α(n-K), it is known as **instrumental variables** (**IV**) estimator (see Fuller, 1977). Here **K** represents the number of instruments, **n** is the sample size, and **α** is a positive constant to specify. A value of α=1 will yield an estimator that is approximately unbiased.

There are also estimators, like ***Full Information Maximum Likelihood***, the ***Three-Stage Least Squares*** method (***3SLS***) and others (Zellner & Theil, 1977; see also Greene, 2012). Their applications are limited due to complex computational schemes and assumptions. More successful techniques for ***SE*** model building were elaborated in the area of ***ANNs***.

9.3. Complex Model Building and Forecasting Using Self-Organizing Data Mining

A. GMDH Algorithms for Self-Organization of Active-Neuron Neural Networks

Self-organizing modelling introduced in Chapter 3 is based on statistical learning networks, which are networks of mathematical functions that capture complex (both linear and non-linear) relationships in a compact and rapidly executable form. Such networks subdivide a problem into manageable pieces or nodes and then automatically apply advanced regression techniques to solve each of these much simpler problems.

Recent developments of the ***GMDH*** (see Mueller & Lemke, 2003; Onwubolu, (Ed.) 2009 and others) have led to neuronets with active neurons. These neuronets put into effect a twice-multilayered structure, where neurons are multilayered and are connected into a multi-layered structure. This gives the possibility to optimize the set of input variables at each layer, while the accuracy increases. In this way, the accuracy of forecasting and approximation can be increased beyond the limits reached by neuronets with single neurons, or by usual statistical methods. In GMDH approach, which corresponds to the actions of human nervous system, the connections between several neurons are not fixed but flexible (they change) depending on the neurons themselves. During the learning self-organizing process, such active neurons are able to estimate inputs that are necessary to minimize the given objective function of the neuron. This is possible on the condition that every neuron in its turn is a multi-layered unit, such as modelling GMDH algorithm. In this way neuronets with active neurons are effective tools to increase the accuracy of a model and reduce modelling time when dealing with stochastic, non-stationary, small and noisy data samples.

A neural network is designed to handle a particular task. This may involve identification of a relationship, pattern recognition, approximation or extrapolation, i.e. prediction of a random process and repetitive events from information contained in a sample of observations. Each neuron is an elementary system that handles the same task. The objective sought in combining many neurons into a network is to enhance the accuracy in achieving the assigned task through a better use of input data.

The ***Multi-Layered Net of Active Neurons (MLNAN)*** technique described in Chapter 8 is a multilayer ***GMDH*** algorithm for multi-input to single-output models identification. Like other ***Multilayered Iterative GMDH algorithms*** (see Fig. 8-36) it can be used for single equation specification in the ***reduced form*** of the ***SE*** model (9-20).

Here, the elements on a lower layer (or stage of selection) are estimated and the corresponding intermediate outputs are computed and then, using this information as inputs, the parameters of the elements of the next layer (stage of selection) are estimated and so on. At the first layer, all possible pairs of the inputs are considered as potential factors and only the good ones (in the sense of the selection criteria – for example, coefficient of correlation and/or t-Test) are used as inputs for the intermediate models at the second layer as shown in Fig. 9-7. In the succeeding layer(s) all possible pairs of the intermediate models from the preceding layer(s) are connected as inputs to the components of the next layer(s). This means that the output of a component at a processed layer is a potential input, depending on a local threshold value of the selection criteria (coefficient of determination of the partial model, MSE, or others – see Chapter 12), to several other components at the next layer. If additional layers do not provide further improvement of the model accuracy the network self-organization stops.

In case of synthesizing complex models in the ***structural form*** (9-18) of the ***SE*** (i.e. multi-input to multi-output) models, an additional part should be used as an additional layer consisting of the following iterative procedure (Motzev & Marchev, 1988):

- The intermediate models are generated combining already chosen, good equations from the last layer of each endogenous variable, according to the combinatorial algorithm.
- Each of the competing hypotheses is a hypothesis about the significance of entering a given version of a single equation into the ***SEM***.
- Each generated system of ***SE*** is considered a potential model for the system of interest, which competes with others "fighting for survival".

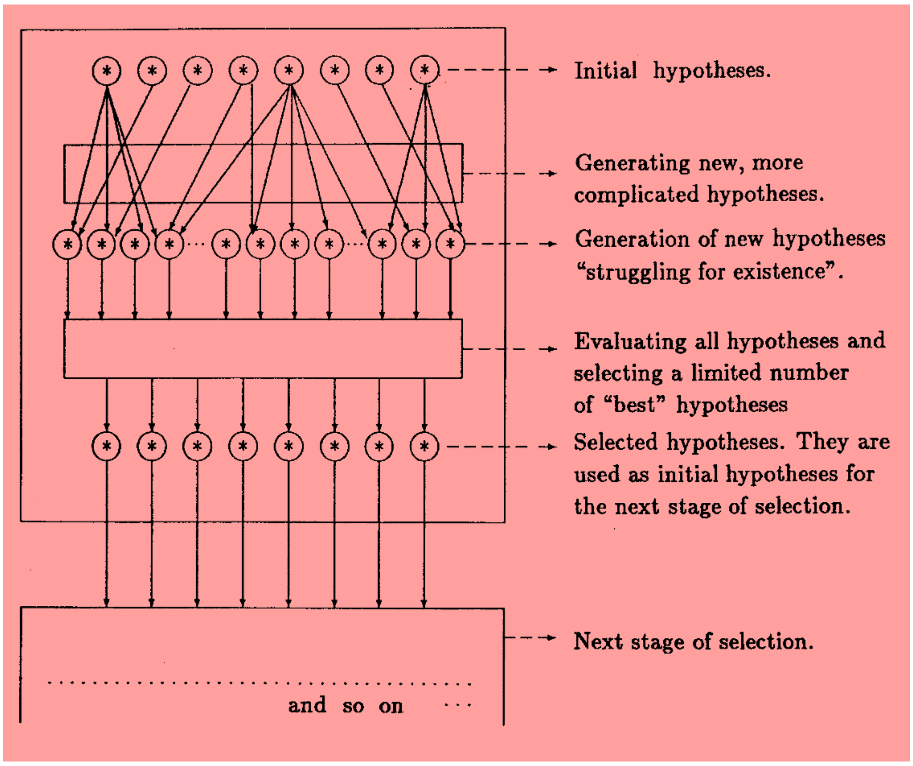

Initial hypotheses.
Generating new, more complicated hypotheses.
Generation of new hypotheses "struggling for existence".
Evaluating all hypotheses and selecting a limited number of "best" hypotheses
Selected hypotheses. They are used as initial hypotheses for the next stage of selection.
Next stage of selection.
and so on ...

Fig.9-7 Hypotheses (representing different models) selection at one layer of the ***Multi-Layered Net of Active Neurons (MLNAN)***

- The evaluation of these competing models is done, using a complex set of criteria – MSE, coefficient of determination, MAPE, and others.
- If the results are unsatisfactory after solving the structural form of the ***SE*** (biased values of the coefficients and/or low accuracy of the ***SE*** model), the procedure returns to the first step.
- Here, the decision maker can then apply some new, a priori knowledge and/or add fresh data observations (if any available), or change the selection criteria. Then a new synthesis of the structural form is done and with the so-updated new set of equations the third part begins again.
- The iterations end when satisfactory results, based on the selection criteria, are achieved.

The final choice of the "best" model is made by the decision maker, who has the option to apply additional insights, qualitative information or knowledge, but after having the guarantee that a large number of possible models have been evaluated and the final choice is based on a small number of good ones. Data mining tools create, store and provide processed data and the business needs these data in the business context. The researcher has to decide the relative importance of the facts generated by ***ANN*** algorithms. The extracted information is useful to a business when it assists decisions which create value or market behavior that provides competitive advantages.

B. Complex Model Forecasting using Self-Organization Data Mining

The algorithms described in the previous section have a large field of applications in model building for analysis and predictions as presented in Mueller & Lemke (2003), Onwubolu (Ed.) (2009), Madala & Ivakhnenko (1994), Motzev & Marchev (1988), Onwubolu (2008), Marchev & Motzev (1985) and others. Such techniques provide opportunities to shorten the design time and reduce the cost and the efforts in model building and forecasting. The results obtained so far in these studies show that GMDH based self-organizing data mining is able to develop reliably even complex models with lower overall error rates than other methods. Increasing model accuracy provides many benefits. For example, it helps decision makers analyze problems more precisely which leads to deeper and better understanding. Moreover, a model with high accuracy will generate better predictions and support managers making better decisions.

A small model should be used as an example to test these algorithms applications with different types of forecasting techniques, described above and also in Chapters 6, 7, 8 and 12. This model is one of the first ***SE*** models of the Bulgarian Economy and was developed at the University of National and World Economy, using theory-driven methods (multiple regression analysis). It is a one-product macro-economic model developed as a system of five simultaneous equations (9-18). The model contains eleven variables:

– five endogenous $\boldsymbol{y_{i,t}}$ ($i = \{1, 2, \ldots 5\}$, $t = \{1, 2, \ldots 15\}$);

– one exogenous $\boldsymbol{x_t}$ ($t=\{1, 2, \ldots 15\}$) and

– five lag variables $\boldsymbol{y_{-i,t-1}}$ ($i = \{1, 2, \ldots 5\}$, $t = \{2, 3, \ldots 15\}$).

Where:

$y_{1,t}$ – National Income produced in Bulgaria (BG) for the period 1961-1975;

$y_{2,t}$ – BG Individual Consumption Expenditures incurred by three institutional sectors, namely households, non-profit institutions serving households and general government;

$y_{3,t}$ – Capital Funds in Industry;

$y_{4,t}$ – Employees in Bulgaria;

$y_{5,t}$ – BG Capital Investments;

$\boldsymbol{y_{-i,t-1}}$ – ($i = \{1, 2, \ldots 5\}$) are their lagged values and

$\boldsymbol{x_t}$ – is the time $\boldsymbol{t} = \{\boldsymbol{1, 2, \ldots T}\}$ as a predictor variable ($\boldsymbol{T}$ is the index set) for the period 1961-1975.

Indirect Least Squares (ILS) was used to estimate unknown coefficients in all equations, which are like (9-17) equation. The achieved model accuracy, measured by root mean squared error normalized (relative) to the mean (i.e. CV(RMSE) of the observed values was 14% (see Marchev & Motzev, 1985).

<u>Improving Model Accuracy</u>

To improve the model accuracy, as already discussed, the forecasters usually try either to add new, "fresh" data or to change model specifications (or both when possible). Since time-series are restricted by the time, we have to wait until next period to add new data. Because of this and to make it possible to compare traditional model building techniques and ***Self-Organizing Data Mining*** (***SODM***), identical time-series data samples and sets of variables were used to modify the existing model specification.

In the process of model improvement, two of the first ***SODM*** prototypes were used – linear version of ***MLNAN*** algorithm described above (Motzev & Marchev, 1988) and the ***KnowledgeMiner*** (Mueller & Lemke, 2003). After selecting the new model specification, the overall accuracy, measured with the same statistics – **CV(RMSE)**, was improved to 3.81% (see Marchev, Motzev & Muller, 1985). This value, compared with the original model error, is almost four times smaller, which means the new model is a more precise and reliable basis for predictions and further analysis of the system of interest.

Table 9.1 Identification tests for all equations in the new model

Equation number	Number of endogenous variables in equation (m_i)	Number of all regressors in equation (n_i)	Number of regressors excluded from equation (n-n_i)	Endogenous variables reduced by one (m_i-1)	Decision - compare (n-n_i) and (m_i-1), i.e. column 4 vs 5
1	2	3	4	5	6
1	3	3	4	2	*Overidentified*
2	2	3	4	1	*Overidentified*
3	4	4	3	3	*Exactly Identified*
4	4	3	4	3	*Overidentified*
5	3	4	3	2	*Overidentified*

Table 9.1 above provides results from parameter identification tests. All equations, but the third one (which is *exactly identified*) are *overidentified,* i.e. they have more than one set of parameters that generate the same distribution of observations, meaning that these multiple parametrizations are *observationally equivalent*. In simple words, there are other potential specifications of these equations, which could be used to improve the model eventually.

More statistics and test results from the new, improved version of the model are provided in Table 9.2 for ***ex-post forecasts*** and Table 9.3, from ***ex-ante*** computations. As mentioned before, ***ex-post forecasts*** are made using later information on the predictors, i.e. ex-post forecasts of sales for each of the 2015 quarters may use the actual observations of demand for each of these quarters, once these have been observed. These are not genuine forecasts, but are useful for studying the behavior of the forecasting model.

Table 9.2 Statistics and test results for ***ex-post*** forecasts

Equation number	Coefficient of multiple correlation (R)	Coefficient of multiple determination (R^2)	F value	CV(RMSE)%	MAPE%
1	0.9949	0.9898	355.81	3.40%	2.69%
2	0.9980	0.9960	936.50	2.16%	1.77%
3	0.9991	0.9982	1848.52	1.76%	1.40%
4	0.9637	0.9287	47.69	0.66%	0.52%
5	0.9522	0.9067	24.30	11.09%	7.13%
For the model	0.9816	0.9635	642.56	3.81%	2.70%

Table 9.3 Statistics and test results for ***ex-ante*** forecasts

Equation number	Coefficient of multiple correlation (**R**)	Coefficient of multiple determination (**R^2**)	**F** value	CV(RMSE)%	MAPE%
1	0.9928	0.9857	254.55	4.03%	3.57%
2	0.9925	0.9851	242.42	4.22%	3.78%
3	0.9993	0.9986	1664.17	3.84%	3.39%
4	0.8288	0.6869	8.05	1.39%	1.11%
5	0.9362	0.8765	17.76	12.76%	8.56%
For the model	0.9499	0.9023	437.39	5.25%	4.08%

The ***ex-ante forecasts***, on the contrary, are made using only the information that is available in advance, i.e. for the same forecasts of sales for the four quarters in 2015 we should only use information that was available before 2015. These are the only genuine forecasts made in advance using whatever information is available at the time.

A comparative evaluation of ***ex-ante*** and ***ex-post*** forecasts help to separate out the sources of forecast uncertainty, revealing whether forecast errors have arisen due to poor forecasts of the predictors or due to a poor forecasting model. In our case, as we can see from the Tables presented here, model statistics and test results change insignificantly and the accuracy change (both for each equation and the model average) is less than 1.5% in most cases. The logical conclusion is, that the differences in ***ex-ante*** and ***ex-post*** forecasts are due to the different data used in forecast computations. It should be noted that this conclusion does not mean that the new model specification is perfect and cannot be improved (as we will show later on in this chapter) – it just proves that the model is stable and reliable and provides a better base for predictions and further analysis, than the original version of the model.

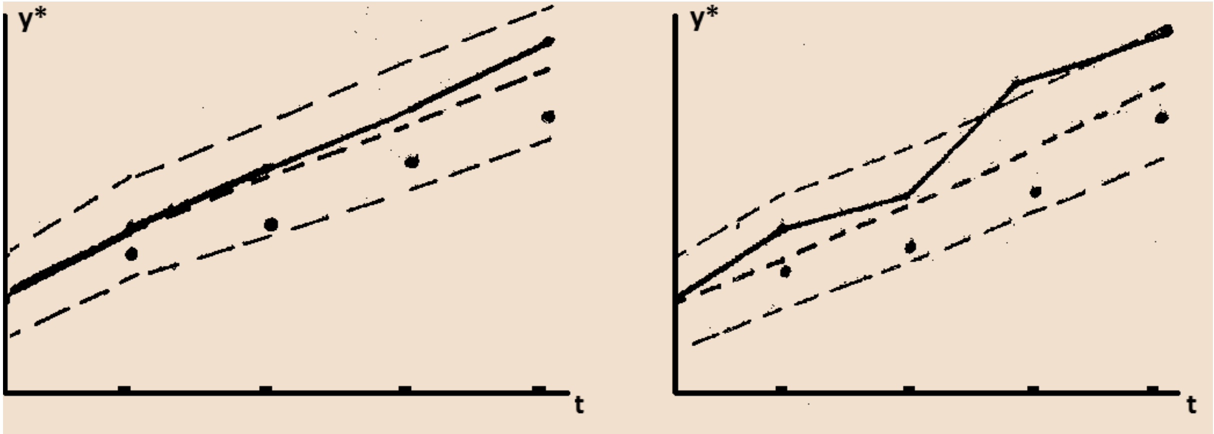


Fig.9-8 ***Ex-ante*** predictions four years ahead for variables $\boldsymbol{y_{1,t}}$ and $\boldsymbol{y_{2,t}}$

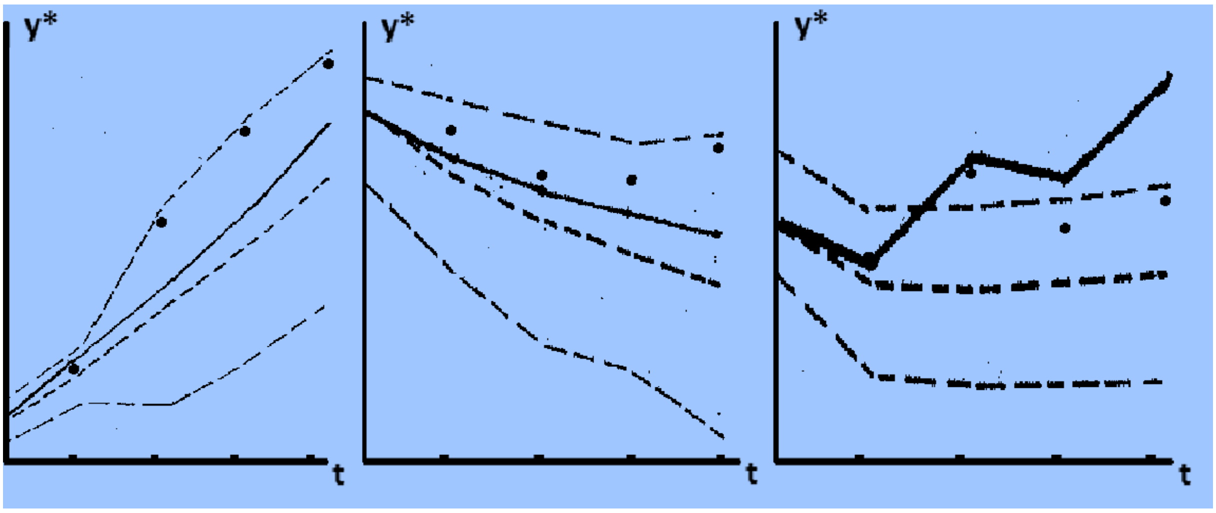


Fig.9-9 ***Ex-ante*** predictions four years ahead for variables $\boldsymbol{y_{3,t}}$, $\boldsymbol{y_{4,t}}$ and $\boldsymbol{y_{5,t}}$

The new improved model was also used to predict ***ex-ante*** all five interdependent variables $\boldsymbol{y_{i,t}}$ for a period of four years ahead (***t={16, 17, … 19}***). Fig.9-8 and Fig.9-9 display the predictions, prediction intervals and the real values observed, where bullets are the real values, dashed lines represent the linear forecast and prediction intervals computed by the ***MLNAN*** and the straight lines are the predictions done using ***KnowledgeMiner*** with nonlinear equations. The comparisons show that, as discussed already in Chapter 8, nonlinear models are not always better – three out of five of the nonlinear prediction sets (for variables $\boldsymbol{y_{3,t}}$, $\boldsymbol{y_{4,t}}$ and $\boldsymbol{y_{5,t}}$) are closer to the real values and two out of five times (for variables $\boldsymbol{y_{1,t}}$ and $\boldsymbol{y_{5,t}}$) the linear predictions are better.

One detail, worth further discussion, is the big jump in variable ($\mathbf{y_{3,t}}$) Capital Funds in Industry after the first period. The reason in fact is that in 1976 there was a change in the BG National Bureau of Statistics methodology for computing the Capital Funds value, which resulted in an increase of about 10% (on average) in the new computed values.

Autocorrelation and Autoregression (AR, VAR and ARMAX) applications

In the new model, a potential issue, identified for equations 2 and 4 (variables $\mathbf{y_{2,t}}$ – BG Individual Consumption Expenditures and $\mathbf{y_{4,t}}$ – Employees in Bulgaria) is the autocorrelation in their residuals at 0.01 level of significance. According to the Durbin-Watson coefficient (**d**) test results are inconclusive. Significant autocorrelation in the residuals[8] was found (see Table 9.4) by using the small sample distribution of the ratio (**Q**), derived by John von Neumann (see Motzev, 1986).

[8] It should be noted that at 0.05 level of significance autocorrelation in the residuals was not detected.

Table 9.4 Comparisons for ***ex-post*** forecasts from stationary and non-stationary models

Equation No.	Coefficient of multiple determination (R^2)		von Neumann test for Autocorrelation (Q)[9]		CV(RMSE)%		MAPE%	
	Stationary	*Non-Stat.*	*Stationary*	*Non-Stat.*	*Stationary*	*Non-Stat.*	*Stationary*	*Non-Stat.*
1	0.9898	0.9920	1.010	0.965	3.40%	2.78%	2.69%	2.35%
2	0.9960	0.9980	*0.735*	1.212	2.16%	1.44%	1.77%	1.23%
3	0.9982	0.9980	1.070	1.069	1.76%	1.63%	1.40%	1.34%
4	0.9287	0.9900	*0.846*	1.081	0.66%	0.81%	0.52%	0.72%
5	0.9067	0.9460	1.020	1.233	11.09%	7.61%	7.13%	6.73%
For the model	xxx		xxx		3.81%	2.85%	2.70%	2.48%

The most probable explanation is the small time lag (only one year) used in this general, aggregated model. The finding was analyzed and tested further on after expanding the model (see next subsection), and here we will use this case to present another, very positive element and a big advantage of the **SODM** techniques.

As mentioned in Chapter 2, most forecasting techniques assume that ***the same underlying causal system that existed in the past will persist into the future***. In fact, managers cannot simply delegate forecasting to models and then forget about it, because unplanned occurrences can wreak havoc with forecasts. For instance, tax increases or decreases, and changes in features or prices of competing products or services can have a major impact on demand and managers must be alert to such occurrences and be ready to override forecasts, which assume a stable causal system.

This assumption sometimes may be a very strong restriction especially in time series forecasting, when most variables and relationships between them are non-stationary (i.e. they are dynamic and change). If this is the case, we can improve predictions by estimating equations in the ***SE*** form with ***dynamic*** (i.e. ***non-stationary***) ***coefficients***.

Thus the ***SE*** model (9-19) will be modified as follows:

$$\mathbf{Y\,\Gamma_t = X\,B_t + U} \tag{9-30}$$

where the $\boldsymbol{m\times m}$ matrix $\mathbf{\Gamma_t}$, which describes the relation between the dependent variables, and the $\boldsymbol{k\times m}$ regressors matrix $\mathbf{B_t}$ have a dynamic structure, i.e. each coefficient $\boldsymbol{\gamma_{i,t}}$ and $\boldsymbol{\beta_{i,t}}$ are non-stationary and change over time ($t=\{1, 2,\ldots, T\}$).

[9] Critical values for Q at 0.01 level of significance are Q′ =0.962 and Q″=3.035

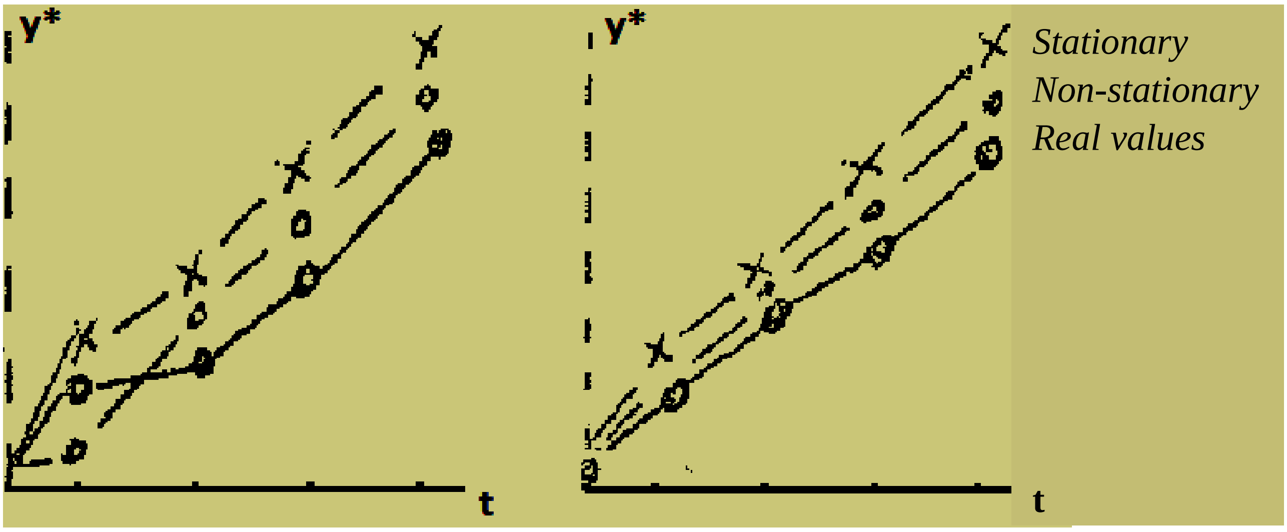


Fig.9-10 ***Ex-ante*** prediction comparisons for four years ahead with stationary and non-stationary models for variables $\boldsymbol{y_{1,t}}$ and $\boldsymbol{y_{2,t}}$

To verify this hypothesis, we should estimate again the new improved version of the ***SE*** model, using the ***MLNAN*** algorithm with non-stationary coefficients (Motzev, 1986). The comparative evaluation of the predictions with stationary and dynamic models in Table 9.4 helps to identify when forecast errors have arisen due to structural changes in the forecasting system. In this example, the first important finding is that there is no more autocorrelation in residuals in the non-stationary version (see von Neumann test **Q** in Table 9.4). It should be expected since the dynamic coefficients estimation in the ***MLNAN*** procedure is done with account of the trend lines for each variable.

Another advantage of the dynamic coefficients is that the model accuracy in ***ex-ante*** predictions is significantly better in all equations but No.3. Because of the methodological and computational changes in the Capital Investments variable ($\mathbf{y_{3,t}}$) mentioned above, it is meaningless to use it in this analysis. As we can see in Table 9.5 and in Fig.9-10 and Fig.9-11 forecast errors are much smaller (both for each equation and the model average). **MAPE** values are about three times less in the non-stationary equations (on average **MAPE** changes from 6.73% in the stationary model to 2.26% in the non-stationary ***SE***) and **CV**(**RMSE**) (the coefficient of variation of the square root of calculated **MSE**) changes from 7.05% to 2.44%.

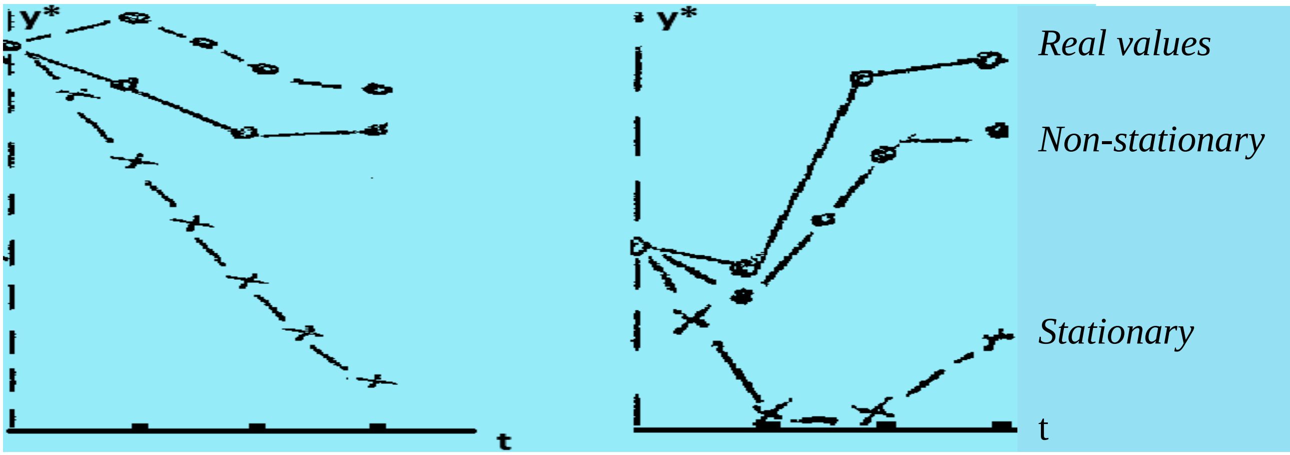


Fig.9-11 ***Ex-ante*** prediction comparisons for three years ahead with stationary and non-stationary models for variables $\boldsymbol{y_{4,t}}$ and $\boldsymbol{y_{5,t}}$

It is worth noting that the ***SODM*** algorithms were used successfully in time-series analysis to build different complex autoregressive models like Distributed lag models, Autoregressive-moving-average with exogenous inputs models (***ARMAX***), Vector autoregression models (***VAR***) and others. As discussed in Chapter 8, detailed studies as Onwubolu (Ed., 2009) and Shahwan & Lemke (2005) of the predictive performance of different time series forecasting techniques confirmed that ***GMDH*** based techniques are able to develop reliably even complex models and achieve lower overall error rates than state-of-the-art methods.

Complex time series justified the use of ANNs. However, the greater flexibility of this model class and its ability to handle nonlinear data patterns comes at the cost of a more demanding specification procedure and computation time, especially when applying genetic algorithms for the purpose of optimization. It is a certain advantage from an application point of view, that ***SODM*** techniques take by far least efforts both in human and in computational time. Furthermore, after each modelling run, ***GMDH***-type ANNs provide on the fly a reliable, analytical model that describes and does not overfit the design data.

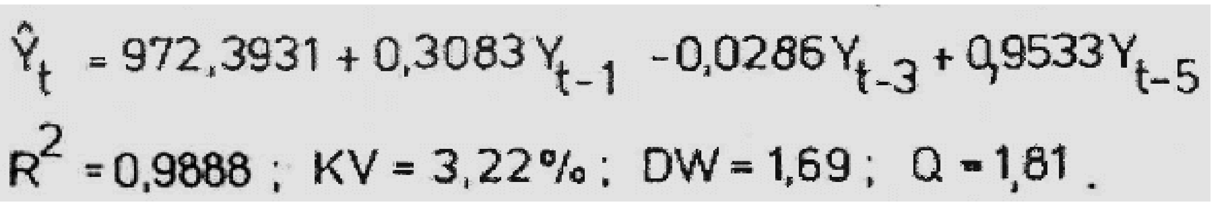
Table 9.5 Comparisons for ***ex-ante*** forecasts from stationary and non-stationary models

Equa-tion No.	**Theil's U-statistics (U)**		**Standard error of the forecast (S)**		**CV(RMSE)%**		**MAPE%**	
	Stationary	*Non-Stat.*	*Stationary*	*Non-Stat.*	*Stationary*	*Non-Stat.*	*Stationary*	*Non-Stat.*
1	0.018	0.014	579.20	463.30	3.58%	2.87%	3.27%	2.80%
2	0.027	0.012	675.30	311.26	5.49%	2.53%	5.31%	2.18%
3	0.001	0.003	95.64	298.28	0.22%	0.69%	0.19%	0.54%
4	0.016	0.003	115.84	27.24	3.22%	0.76%	2.88%	0.75%
5	0.128	0.027	1013.6	238.86	22.74%	5.36%	22.01%	5.03%
Average	xxx		xxx		7.05%	2.44%	6.73%	2.26%

SE models, in fact, comprise the most elements of Distributed lag models, ***ARMAX*** and ***VAR*** models. The examples discussed so far confirmed the hypothesis that ***SODM*** techniques are able to develop reliably even complex models and achieve lower overall error rates than state-of-the-art methods. Another example below proves their high qualities and advantages.

During the ***SE*** model building and its improvements, discussed above, the ***MLNAN*** algorithm was used for estimating ***AR*** equations of more than 20 macroeconomic variables with a time lag of up to 5 years (see Motzev, Muller & Marchev, 1986), providing in most cases **MAPE**% less than 1.5%, adjusted coefficient of determination (**R**^{2}) greater than 0.9 and average **CV(RMSE)**% for all equations 4.74%. The ***AR*** models for the five endogenous variables in ***SE*** model are presented below with their statistics **R**^{2}, **CV(RMSE)**% (denoted as **KV**%) and residual autocorrelation tests, where **DW** is the Durbin-Watson coefficient (**d**) and **Q** is von Neumann test:

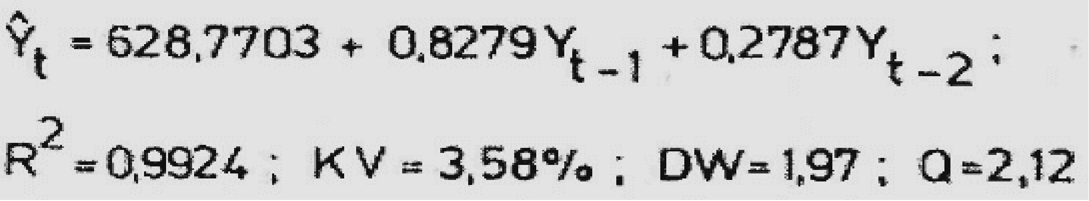
y$_{1,t}$ – National Income produced in Bulgaria:



$$\hat{Y}_t = 972{,}3931 + 0{,}3083\,Y_{t-1} - 0{,}0286\,Y_{t-3} + 0{,}9533\,Y_{t-5}$$

$$R^2 = 0{,}9888\ ;\ KV = 3{,}22\%\ ;\ DW = 1{,}69\ ;\ Q = 1{,}81\ .$$

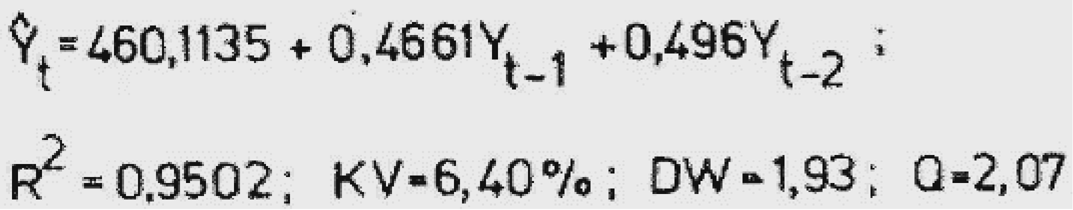
y$_{2,t}$ – BG Individual Consumption Expenditures:



$$\hat{Y}_t = 628{,}7703 + 0{,}8279\,Y_{t-1} + 0{,}2787\,Y_{t-2}\ ;$$

$$R^2 = 0{,}9924\ ;\ KV = 3{,}58\%\ ;\ DW = 1{,}97\ ;\ Q = 2{,}12$$

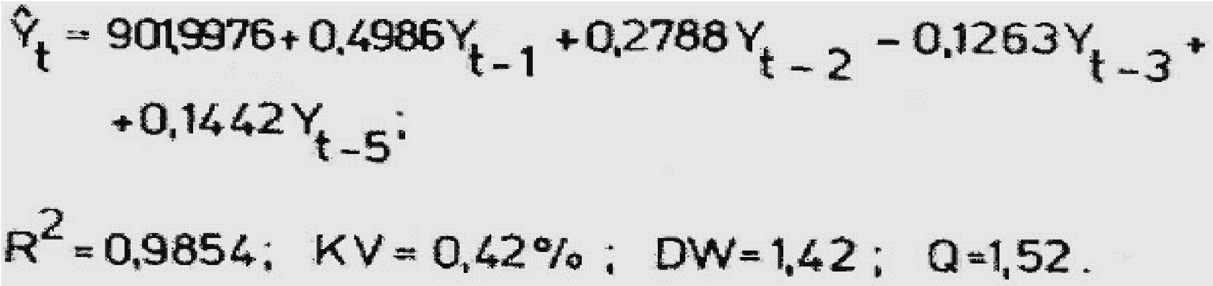
y$_{3,t}$ – Capital Funds in Industry:



$$\hat{Y}_t = 460{,}1135 + 0{,}4661\,Y_{t-1} + 0{,}496\,Y_{t-2}\ ;$$

$$R^2 = 0{,}9502;\ KV = 6{,}40\%\ ;\ DW = 1{,}93;\ Q = 2{,}07$$

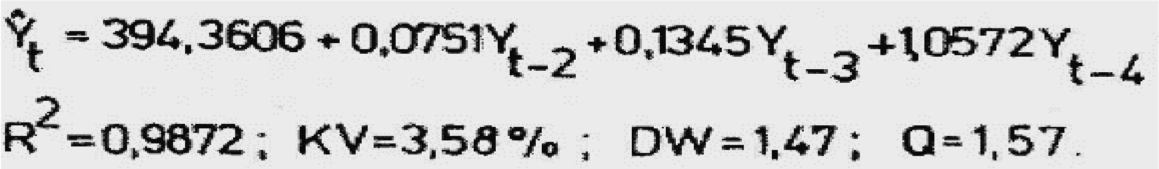
y$_{4,t}$ – Employees in Bulgaria:



$$\hat{Y}_t = 901{,}9976 + 0{,}4986\,Y_{t-1} + 0{,}2788\,Y_{t-2} - 0{,}1263\,Y_{t-3} + \\ + 0{,}1442\,Y_{t-5}\ ;$$

$$R^2 = 0{,}9854;\ KV = 0{,}42\%\ ;\ DW = 1{,}42\ ;\ Q = 1{,}52\ .$$

$y_{5,t}$ – BG Capital Investments:



$$\hat{Y}_t = 394{,}3606 + 0{,}0751Y_{t-2} + 0{,}1345Y_{t-3} + 1{,}0572Y_{t-4}$$

$$R^2 = 0{,}9872;\ KV = 3{,}58\%;\ DW = 1{,}47;\ Q = 1{,}57.$$

Two important conclusions can be made from this example:

(1) The ***MLNAN*** algorithm, as other ***SODM*** techniques, is a cost-effective tool for building AR models with high accuracy and reliability.

(2) ***SE*** models provide a better platform than any other type of models for analysis and predictions of complex systems and processes. Comparing the ***SE*** model endogenous variables' accuracy from Table 9.4 and their ***AR*** models' errors above, we can see that even with a smaller time lag (only one year in ***SE*** versus five in ***AR***) in most cases ***SE*** equations are better and more reliable. Further research (Marchev & Motzev, 1985) with larger time lags in ***SE*** models confirmed this general statement.

Complex Systems Model Building and Forecasting

Another area of application of ***GMDH*** techniques and ***MLNAN*** algorithms in particular is macroeconomic modeling, such as the economies in the USA (Klein, Mueller, & Ivakhnenko, 1980), Germany (Mueller & Lemke, 2003), Bulgaria (Marchev & Motzev, 1985; Marchev, Motzev & Muller, 1985) and other countries (Madala & Ivakhnenko, 1994). The small model (called SIMUR I) discussed in the previous section is the beginning of a series of simulation models of the Bulgarian economy (Marchev, Motzev & Muller, 1985).

The first increase of the complexity of the initial model resulted in the model SIMUR II (Marchev & Motzev, 1985) which is an aggregated macroeconomic model of twelve ***SE***. It contains 42 variables, incl. 12 endogenous, 5 exogenous and 26 lag variables with a time lag of up to 3 years:

Endogenous Variables:

$\mathbf{y_{1,t}}$ – Gross Domestic Product
$\mathbf{y_{2,t}}$ – Net National Income
$\mathbf{y_{3,t}}$ – Gross Consumption, where $\mathbf{y_{3,t}} = \mathbf{y_{2,t}} \text{ x } \mathbf{x_{2,t}}$
$\mathbf{y_{4,t}}$ – Gross Investments, where $\mathbf{y_{4,t}} = \mathbf{y_{2,t}} \text{ x } \mathbf{x_{3,t}}$
$\mathbf{y_{5,t}}$ – Gross Capital Investments
$\mathbf{y_{6,t}}$ – Capital Funds in Industry
$\mathbf{y_{7,t}}$ – Capital Funds in Industry Growth, where $\mathbf{y_{7,t}} = \mathbf{y_{6,t}} - \mathbf{y_{6,t-1}}$
$\mathbf{y_{8,t}}$ – Employees in BG
$\mathbf{y_{9,t}}$ – Personal Income Receipts

y$_{10,t}$ – Personal Consumption

y$_{11,t}$ – Import

y$_{12,t}$ – Export

Exogenous Variables:

x$_{1,t}$ – Labor Force (in thousands);

x$_{2,t}$ – Consumption Index;

x$_{3,t}$ – Investments Index;

x$_{4,t}$ – Price Index (1970 = 100%);

x$_{5,t}$ – Time line (t=1,2,3…28) is representing model's dynamic.

Lagged variables (both endogenous and exogenous) with a time lag of up to 3 years:

z$_{j,t-k}$ – ($j=\{1, 2,..., 26\}$); (***t*** **v** ***T***) and ($k=\{1, 2, 3\}$).

Variables:

Exogenous Endogenous Lag Equation

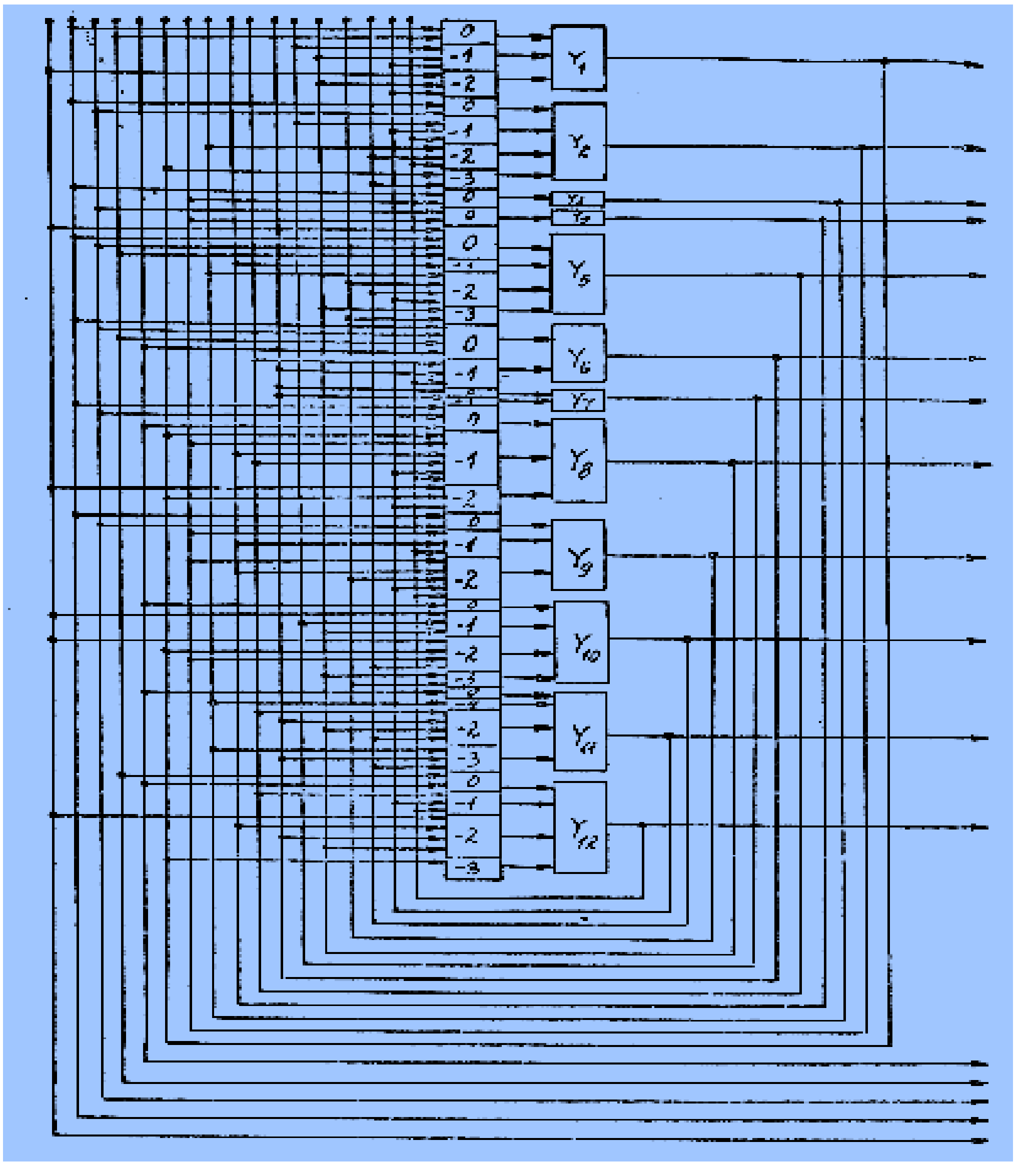
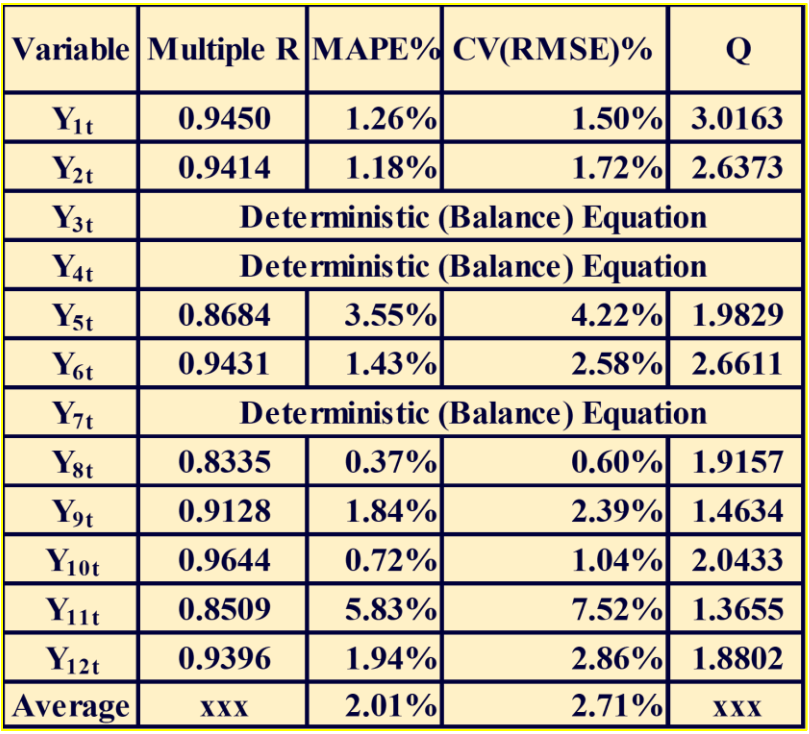


Fig.9-12 System's graph of the generated macroeconomic ***SE*** for the Bulgarian Economy

Table 9.6 Model statistics and accuracy for SIMUR II[10]



Variable	Multiple R	MAPE%	CV(RMSE)%	Q
Y_{1t}	0.9450	1.26%	1.50%	3.0163
Y_{2t}	0.9414	1.18%	1.72%	2.6373
Y_{3t}	Deterministic (Balance) Equation			
Y_{4t}	Deterministic (Balance) Equation			
Y_{5t}	0.8684	3.55%	4.22%	1.9829
Y_{6t}	0.9431	1.43%	2.58%	2.6611
Y_{7t}	Deterministic (Balance) Equation			
Y_{8t}	0.8335	0.37%	0.60%	1.9157
Y_{9t}	0.9128	1.84%	2.39%	1.4634
Y_{10t}	0.9644	0.72%	1.04%	2.0433
Y_{11t}	0.8509	5.83%	7.52%	1.3655
Y_{12t}	0.9396	1.94%	2.86%	1.8802
Average	xxx	2.01%	2.71%	xxx

Fig.9-12 exhibits the structure of the ***SE*** in the new model SIMUR II generated with the use of the ***MLNAN*** algorithm. As should be expected the forecast error of the previous model SIMUR I was reduced significantly – **MAPE%** from 4.08% (refer to Table 9.3) to 2.01% (see Table 9.6) and **CV(RMSE)%** from 5.25% to 2.71%. According to von Neumann statistics (**Q**), at 0.05 level of significance, no autocorrelation in the residuals was detected.

There are two variables in Table 9.6, Gross Capital Investments ($\mathbf{y}_{5,t}$) and Import ($\mathbf{y}_{11,t}$), whose equations' accuracy is relatively worse, in comparison with the rest of the endogenous variables. Their **MAPE%** and **CV(RMSE)%** are higher and their coefficient of multiple correlation **R** is smaller. The latest statistics give a possible explanation of this point – there are potential ***predetermined*** variables excluded from these equations. If such variables are added into the corresponding equation the coefficient of multiple correlation **R** must increase.

Further analyses and experiments with these two variables in the next model of the series, SIMUR III (Motzev & Marchev, 1988), including more detailed relationships and additional ***predetermined*** variables, confirmed the above statement. The **MAPE%** for Gross Capital Investments ($\mathbf{y}_{5,t}$) was reduced from 3.55% to 0.356% and **CV(RMSE)%** from 4.22% to 0.58%. For Import ($\mathbf{y}_{11,t}$) MAPE% reduction is from 5.83% to 1.77% and **CV(RMSE)%** from 7.52% to 2.29%.

[10] Similar model for the German economy will be discussed in Chapter 12.

The model SIMUR II was also used to predict ***ex-ante*** all twelve interdependent endogenous variables $y_{i,t}$ for a period of four years ahead (***t={16, 17, … 19}***). Table 9.7 presents prediction errors for forecasts computed using the ***MLNAN*** algorithm. For more detailed analysis a couple of other statistics are used in addition to **MAPE%** and **CV(RMSE)%,** the **Theil's U-statistics** (see formula 3-11) and the **coefficient of linear correlation** between forecasted and observed values **(R)**:

$$R = \sqrt{\sum_{t=T}^{T+L} y_{i,t} * y^*_{i.t} \Bigg/ \sum_{t=T}^{T+L} y^2_{i.t}} \qquad (9\text{-}31)$$

where $y_{i,t}$ are the observed values of interdependent endogenous variables ($i=\{1, 2,\ldots, 12\}$),
($t=\{T+1, T+2,\ldots, T+L\}$), ($L=4$) and
$y^*_{i,t}$ are the forecasted values for the same time period.

Theil's U-statistics is a normalized measure of total forecast error, which varies between zero and one. By the rule of thumb, for good forecast accuracy, it is desirable that the **U**-statistic is close to zero.

The ***Coefficient of Linear Correlation (R)*** between forecasted and observed values is similar in its interpretation to the regular coefficient of linear correlation. It shows how close the predicted and observed values of the endogenous variables $y_{i,t}$ are. The general rule for a good forecast is to have (**R**) as close as possible to one.

As noted in the previous section, in 1976 there was a change in the BG National Bureau of Statistics methodology for computing the Capital Investments value, which resulted in an increase of about 10% (on average) in the newly computed values. For this reason, predictions for variables Capital Funds in Industry ($\mathbf{y_{6,t}}$) and Capital Funds in Industry Growth ($\mathbf{y_{7,t}}$) were corrected after the first year.

Similar to SIMUR II, models of ***SE*** were developed for the German's economy for the period of 1960-1987 using ***KnowledgeMiner*** software. The comparisons (discussed in more detail in Chapter 12) show that both ***Self-Organizing Data Mining*** techniques generated ***SE*** models which are very close in their accuracy. The Bulgarian model has a slightly smaller error than the German one – MAPE% of 3.90% for the Bulgarian model versus MAPE% of 4.93% for the German's Economy model.

Table 9.7 Prediction errors for forecasts calculated using SIMUR II model

Variable Equation	**MAPE%**				**Average for the model:**			
	1977	**1978**	**1979**	**1980**	**MAPE%**	**CV(RMSE)%**	**R**	**U**
1	8.28%	2.58%	4.52%	3.85%	4.81%	5.10%	0.999	0.0260
2	4.26%	6.45%	10.16%	0.61%	5.37%	6.40%	0.975	0.0040
3	3.71%	6.54%	8.31%	0.62%	4.80%	5.84%	0.977	0.0310
4	5.81%	6.15%	15.67%	0.56%	7.05%	8.93%	0.993	0.0078
5	6.17%	0.35%	6.32%	3.81%	4.16%	4.47%	0.998	0.0022
6	0.48%	0.40%	0.04%	1.36%	0.57%	1.82%	0.996	0.0003
7	1.43%	1.15%	1.44%	1.64%	1.42%	1.77%	0.986	0.0035
8	0.96%	0.83%	0.22%	0.62%	0.66%	0.72%	0.995	0.0001
9	5.24%	6.15%	5.60%	8.23%	6.31%	6.68%	0.995	0.0044
10	0.77%	2.22%	1.20%	3.34%	1.88%	2.31%	0.998	0.0005
11	13.89%	6.26%	6.39%	3.86%	7.60%	7.72%	0.996	0.0056
12	1.03%	3.16%	3.16%	1.18%	2.13%	3.28%	0.995	0.0010
Average	4.34%	3.52%	5.25%	2.47%	3.90%	4.59%	0.992	0.0098

A further step of increasing complexity is achieved with the model SIMUR III (Motzev & Marchev, 1988). It is a complex macro-economic model of thirty-nine ***SE***, which contains more than one hundred variables, incl. 39 endogenous, 7 exogenous and 82 lag variables (time lag up to 5 years). The accuracy for most equations and for the model’s average, measured with **CV(RMSE)%** is less than **1%.** Because of its big size, the large number of variables and specific components this model is not appropriate to be used for detailed discussion and comments in a textbook. More recent applications of ***SODM*** can be found in (Motzev, 2014), (Lemke, 2008) and others.

It is worth noting again, that all results confirm that the ***SODM***, and ***GMDH*** based ***ANN***s in particular, provide opportunities for reducing the time, the cost and the efforts for model building and identification. ***MLNAN*** and similar techniques are certainly able to develop reliably even complex models with lower overall error rates than other methods. These tools are very useful for addressing the model-building problems discussed before. For example, overfitting is eliminated by the use of external criterion (cross-validation) for validating the model. The small number of independent measurements (or short time-series) doesn’t cause problems, because the inverted matrix size is always 2x2 (pair-wise combinations). This feature helps in dealing with the problem of multicollinearity as well. The autocorrelation is eliminated by adding automatically (when needed) lagged variables and so forth. Last, but not least, these techniques are totally automated procedures with strong user-friendly interface which provides opportunities for the forecaster (or decision maker), who at the crucial points of the process has options to apply additional insights, knowledge or hypotheses.

Of course, there are some limitations of predictive models based on data fitting. For example, history cannot always predict the future; using relations derived from historical data to predict the future implicitly assumes certain steady-state conditions or constants in the complex system. This is almost always inaccurate when the system involves people.

Another issue is the “unknown unknowns”. In all data collection, the researcher first defines the set of variables for which data is collected. However, no matter how extensive the researcher considers his selection of the variables, there is always the possibility of new variables that have not been considered or even defined, yet that are critical to the outcome.

It is imperative to conclude that the model outputs must always be evaluated by the forecaster to figure out whether new and useful knowledge of the domain has been discovered. Predictive models create and provide data, but real-life business needs information, i.e. data in the business context. The extracted information is valuable to a business only when it leads to actions that create value or market behavior, which gives a competitive advantage. Eventually, the decision maker must determine the ultimate importance of the information generated by any tool or technique.

***

SUMMARY AND CONCLUSIONS

Chapter 9 discusses complex forecasting models and techniques. There is a big variety of advanced techniques, more or less universal, which could be used in Business Forecasting.

Econometrics is the application of statistical and mathematical theories to economics for the purpose of testing hypotheses and forecasting future trends. Econometric models include:

- ***Linear Regression*** – in modern econometrics, other statistical tools are frequently used, but linear regression is still the most frequently used starting point for an analysis and forecasting.
- ***Production function*** relates physical output of a production process to physical inputs or factors of production. Its primary purpose is to address allocative efficiency in the use of factor inputs in production and the resulting distribution of income to those factors while abstracting away from technological problems.
- ***Supply and demand functions*** are economic models of price determination. They conclude that in a competitive market, the unit price for a particular good will vary until it settles at a point where the quantity demanded by consumers (at current price) will equal the quantity supplied by producers (at current price), resulting in an economic equilibrium for price and quantity.
- A ***probit*** (***prob***ability+un***it***) model is a type of regression where the dependent variable can only take two values, and it is a popular specification for an ordinal or a binary response model. The purpose of the model is to estimate the probability that an observation with particular characteristics will fall into a specific one of the categories.
- ***Logistic regression*** (***logit regression*** or ***model***) is a type of probabilistic statistical classification model. It is also used to forecast a binary response from a binary predictor, used for forecasting the outcome of a categorical dependent variable (i.e., a class label) based on one or more predictor variables (features).
- ***Vector autoregression (VAR)*** is an econometric model used to capture the linear interdependencies among multiple time series. ***VAR*** models generalize the univariate ***autoregression (AR)*** models by allowing for more than one evolving variable. Each variable has an equation explaining its evolution based on its own lags and the lags of the other model variables. The only prior knowledge required is a list of variables which can be hypothesized to affect each other intertemporally.
- The notation ***ARMAX(p, q, b)*** refers to the model with ***p*** autoregressive terms, ***q*** moving average terms and ***b*** exogenous inputs terms. It contains the **AR(*p*)** and **MA(*q*)** models and a linear combination of the last ***b*** terms of a known and external time series $\mathbf{d_t}$.

- ***The nonlinear autoregressive exogenous model (NARX)*** is a nonlinear autoregressive model which has exogenous inputs.

The traditional and most popular formal language used in econometrics are the ***structural equation models (SEM)***:

- ***Simultaneous equation*** **(*SE*)** models are statistical models in the form of a set of linear simultaneous equations – i.e. a variable could be dependent in one equation and a regressor in others. All dependent variables that are determined by the model are called ***endogenous*** or ***jointly determined***. Those determined from outside are referred to as ***exogenous*** or ***predetermined.***
- In econometrics, "***structural form***" models begin from deductive theories of the economy – the reduced form of a system of ***SE*** is the result of solving the structural system for the endogenous variables, i.e. "***reduced form***" models begin by identifying particular relationships between variables.
- ***Identifiability*** is a property which a model must satisfy in order for precise inference to be possible. A model is identifiable if it is theoretically possible to learn the true value of this model's underlying parameter after obtaining an infinite number of observations from it. A model is identifiable only under certain technical restrictions, in which case the set of these requirements is called the ***identification conditions***.
- The parameters are ***underidentified*** (***not identified***) if there are fewer exogenous regressors than there are covariates or, equivalently, if there are fewer excluded exogenous regressors than there are endogenous covariates in the equation of interest.
- ***Overidentified*** refers to any situation where a statistical model will invariably have more than one set of parameters that generate the same distribution of observations, meaning that multiple parametrizations are ***observationally equivalent***.
- In econometrics, two parameter values are considered ***observationally equivalent*** if they both result in the same probability distribution of observable data.
- ***Path diagrams*** can be viewed as ***SE*** models that contain only the structural part.
- ***Indirect Least Squares (ILS)*** is an approach where the coefficients in an ***SE*** model are estimated from the reduced form model using ***OLS***. For this, the structural system of equations is transformed into the reduced form first. Once the coefficients are estimated the model is put back into its structural form.
- ***Two-stage least squares method*** (***2SLS***) is an ***equation-by-equation*** technique, where the endogenous regressors on the right-hand side of each equation are being instrumented with the regressors ***X*** from all other equations.

- ***K-class estimators*** are a group of estimators, when: **κ=0,** this is the ***OLS***; **κ=1,** it is the ***2SLS;*** **κ=λ,** this is the ***LIML;*** **κ=λ-α(n-K),** it is known as **instrumental variables** (**IV**) estimator – **K** represents the number of instruments, **n** is the sample size, and **α** is a positive constant to specify.

Other estimators, like ***full information maximum likelihood*** and ***three-stage least squares method*** (***3SLS***) have limited applications due to complex computational schemes and assumptions. More techniques for ***SE*** model building and estimation were elaborated in the area of ***Data Mining*** such as ***ANNs***. Complex Model Building and Forecasting Using ***Self-Organizing Data Mining (SODM)*** is one of the most successful group of tools in this area.

- ***Self-organising modelling*** is based on statistical learning networks, which are networks of mathematical functions that capture complex (both linear and non-linear) relationships in a compact and rapidly executable form. Such networks subdivide a problem into manageable pieces or nodes and then automatically apply advanced regression techniques to solve each of these much simpler problems:
- ***Combinatorial (COMBI)*** is the basic ***Group method of data handling (GMDH)*** algorithm in ***SODM***. It is based on full or reduced sorting-out of gradually complicated models and their evaluation by external criterion on a testing data set.
- ***Multilayered Iterative GMDH algorithm*** is an algorithm in which the iteration rule remains unchanged from one layer to the next. It should be used when it is needed to handle a big number of variables.
- The key feature of ***Objective System Analysis (OSA)*** algorithm is that it examines systems of algebraic or difference equations, obtained by implicit templates (without goal function). An advantage of the algorithm is that the information embedded in the data sample is utilized better and we can estimate the relationships between variables.
- The ***Multi-Layered Net of Active Neurons (MLNAN)*** technique is a multilayer ***GMDH*** algorithm for multi-input to single-output models identification. Like other ***Multilayered Iterative GMDH algorithms*** it can be used for single equation specification in the ***reduced form*** of the ***SE*** model. Here, the elements on a lower layer are estimated and the corresponding intermediate outputs are computed and then, using this information as inputs, the parameters of the elements of the next layer are estimated and so on.

The ***SODM*** techniques are used successfully in a time-series analysis to build different complex autoregressive models like Distributed lag models, Autoregressive-moving-average with exogenous inputs models (***ARMAX***), Vector autoregression models (***VAR***) and others, including ***SE*** models in their both ***structural*** and ***reduced forms***.

SODM, and ***GMDH*** based ***ANN***s in particular, provide opportunities for shortening the time, the cost and the efforts for model building and identification. ***MLNAN*** and similar techniques are able to develop reliably even complex models with lower overall error rates than state-of-the-art methods. These tools are very useful for addressing model-building problems discussed before. For example, overfitting is eliminated by the use of external criterion (cross-validation) for validating the model. The small number of independent measurements (or short time-series) doesn't cause problems, because the inverted matrix size is always 2x2 (pair-wise combinations). This feature helps in dealing with the problem of multicollinearity as well. The autocorrelation is eliminated by adding automatically (when needed) lagged variables and so forth. Last, but not least, these techniques are totally automated procedures with strong user-friendly interface which provides opportunities for forecaster (or decision maker), who at the crucial points of the process has options to apply additional insights, knowledge or hypotheses.

More details and applications of Business Intelligence and Business Analytics platforms and tools are discussed further on in Chapters 10, 11 and 12.

KEY TERMS

AD-AS (aggregate demand–aggregate supply model)	*299*
Autoregressive–moving-average model with exogenous inputs (ARMAX)	*305*
Coefficient of linear correlation (R) between forecasted and observed values	*331*
Durbin-Watson coefficient (d)	*323*
Econometrics	*298*
Endogenous (jointly determined) variables	*303, 309*
Exogenous (predetermined) variables	*304, 309*
Ex-post, ex-ante forecasts	*321, 322*
General Linear Model (GLM)	*304*
GMDH	*317*
Identifiability, Exactly identified, Overidentified and Underidentified SE	*311, 313*
Indirect least squares (ILS)	*314*
John von Neumann statistics (Q)	*323*
K-class estimators	*316*
KnowledgeMiner	*320, 323*
Limited Information Maximum Likelihood (LIML)	*316*
Logistic regression (logit regression)	*301*
Multi-Layered Net of Active Neurons (MLNAN)	*317*
Multivariate Least Squares (MLS)	*305*
Nonlinear Autoregressive Exogenous Model (NARX)	*305*
Order condition and rank condition in SE identification	*311*
Ordinary Least Squares (OLS)	*305*
Path diagrams	*313*
Probit regression	*300*
Production function	*298*
SE with dynamic (non-stationary) coefficients	*324*
Self-Organising Data Mining (SODM)	*317, 319, 320*
Simultaneous Equations (SE) Models	*308*
Structural and reduced forms of the SE	*…308, 310*
Structural equation model (SEM)	*308*
Two-stage least squares (2SLS)	*315*
Vector autoregression (VAR)	*303*

CHAPTER EXERCISES

Conceptual Questions:

1. List the basic econometric models. Give examples of each and discuss their applications.
2. What are the similarities and the differences between ***probit*** and ***logit regression?***
3. List all types of autoregressive models (***AR***). What are the new elements in ***VAR*** and ***ARMAX*** models? Explain how to distinguish between all these ***AR*** models.
4. What are the two forms of ***Simultaneous equation*** **(*SE*)** models? List and discuss the ***identification conditions***.
5. How ***SODM*** and ***GMDH*** based ***ANN***s address model-building and forecasting problems? Discuss at least three problems.

Business Applications:

Open Gretl program and import file Sales Data.xslx:

- Set up the time series data for a model with dependent variable "Sales":
- Using ***ARIMA*** methodology develop an ***AR*** model, based on the data patterns detected within the time-series.
- Test the aptness of each model using test statistics provided by Gretl software and perform residual analysis. Do you see any violations of the regression assumptions? If yes return to the previous step and improve the ***AR*** model.
- Compute Sales forecast for the next 12 months.
- Design formulas, similar to the formulas in Part 4 of the Integrative case and compute MAD, MSE, MAPE and MPE for the new model, for a testing dataset of the 12 new monthly forecasts given in spreadsheet Predictions.
- What is the model accuracy? Are there any initial assumptions/reasons leading to this conclusion?

Discuss all findings and write a short report (up to two pages) summarizing your answers.

INTEGRATIVE CASE

HEALTHY FOOD SUPPLY CHAIN & STORES

Part 9: Complex Models and Forecasting – 1

In Chapter 1 we introduced *Healthy Food Stores* – a fast-growing retail food provider with 12 stores in a northwestern state. The company executives decided to study the effect that company advertising dollars have on sales. They hoped that examining collected historical data would reveal relationships that would help determine future advertising expenditures and predict monthly sales volumes for the upcoming quarter.

After identifying basic parameters, along with input (independent) and output (dependent) variables of the forecasting scenario in Part 2, the research team determined the main elements of the forecasting process:

- Forecasting horizon of up to twelve months;
- Quarterly forecast updates, since accuracy decreases as time horizon increases, and sufficient time is needed to implement possible changes;
- Development of different models based on data patterns, if any, and evaluation of their accuracy in order to select the most appropriate one;
- Selection of the best forecasting model with no more than 5% forecasting error.

The benchmark forecast computed in Part 3 (using the baseline of ***one-step naive forecast*** as a reference forecast) provided basic values for the most common measures of accuracy MFE, MAD, MAPE, MPE, MSE and CV(RMSE).

In Part 4, information about the opinion of some important people from the *Healthy Food Stores Company*, concerning this specific case, was collected and the research team applied the Delphi method to top executives group, Sales-force composite to the sales managers from all 12 stores and Scenario writing to the most experienced professionals from Advertising Department. After collecting such valuable information from different sources, in Part 5 the research team made its first steps in Numerical Predictions by developing different basic forecasting models. They created spreadsheets for Naïve techniques (Average model, Random Walk with Drift and Seasonal Naïve Technique), simple Moving Average, Simple Exponential Smoothing (SES) and Triple (Holt-Winters) Exponential Smoothing (TES), which were used to expand the base-line of one-step naïve forecast as reference forecasts.

In Part 6 the research team analyzed the relationships between dependent variable Sales and the available predictors. After performing multiple correlation and regression analysis, researchers developed reliable forecasting model, which passed all tests and hypotheses,

representing the real system with certain error. In Part 7, the model was expanded by adding Dummy seasonal variables to analyze the Seasonal effect in company Sales. In Part 8, the improvement of the forecasting model continued (with the help of some advanced Time series analyses and predictive techniques) and few ***AR*** models were built using ***ARIMA*** methodology and *Gretl* software.

The next step of the forecasting process would be to develop new, different, complex models.

Case Questions

1. Open Gretl program and import file Data.xslx containing the results from *Part 6* (Building Multiple Regression Model).
2. Set up the time series data for a multiple regression model with dependent variable "Sales" and all basic predictors as exogenous variables, including the Dummy seasonal variables:
   a) Split the sample into Training (36 observations) and Testing (12 observations) data sets.
   b) Run Multiple Regression analysis and Conduct a test of hypothesis on each of the predictor variables. Eliminate any insignificant predictor at 0.05 significance level.
   c) Rerun the multiple regression equation and test the aptness of the model. Continue eliminating any insignificant regressor at 0.05 level of significance until only important (significant) exogenous variables remain in the model.
   d) Compute Sales forecast for the next 12 months.
3. Use (copy/paste) the formulas designed in Part 3 to compute MFE, MAD, MAPE, MPE, MSE and CV(RMSE) for the new model, for the given testing dataset of 12 monthly forecasts provided in spreadsheet Errors.
4. Comment and analyze new model's accuracy - how good is the accuracy of these forecasts? What model, out of all models so far provides the best accuracy? Discuss.
5. What overall recommendations would you make to the research team? Explain.
6. Write a report (at least two pages not counting charts and tables) on the questions above, discussing all important findings and draw relevant conclusions about this part of the Integrative Case.

References

Anderson, T., & Rubin, H. (1949). Estimator of the parameters of a single equation in a complete system of stochastic equations, *Annals of Mathematical Statistics*, 20(1), 46–63.

Basmann, R. (1957). A generalized classical method of linear estimation of coefficients in a structural equation, *Econometrica*, 25(1), 77–83.

Cobb, C., & Douglas, P. (1928). A Theory of Production, *American Economic Review*, 18 (Supplement), 139–165. (see also Filipe, J., & Adams, F. (2005). The Estimation of the Cobb-Douglas Function: A Retrospective View, *Eastern Economic Journal*, 31(3), 427–445).

Fisher, F. (1966). *The Identification Problem in Econometrics*. Huntington, N.Y. (R.E. Krieger Pub. Co., 1976).

Freedman, D. (2009). *Statistical Models: Theory and Practice*. Cambridge University Press.

Fuller, W. (1977). Some Properties of a Modification of the Limited Information Estimator, *Econometrica*, 45(4), 939–953.

Greene, W. (2012). *Econometric Analysis* (7th ed.). Prentice Hall.

Greenspan, A. (2013). *The Map and the Territory (Risk, Human Nature, and the Future of Forecasting)*. The Penguin Press, New York.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations, *Econometrica*, 11(1), 1-12. (Reprinted in Hendry, D.F., & Morgan M.S. (Eds.). (1995). *The Foundations of Econometric Analysis*. Cambridge University Press, 477-490).

Klein, L., Mueller, J-A., & Ivakhnenko, A. G. (1980). Modeling of the Economics of the USA by Self-organization of the System of Equations, *Soviet Automatic Control*, 13(1), 1-8.

Koopmans, T. (1949). Identification problems in economic model construction, *Econometrica*, 17(2), 125–144.

Lemke, F. (2008). *Parallel Self-organizing Modeling.* http://www.knowledgeminer.com/pdf/performance_yX.pdf

Madala, H. R., & Ivakhnenko, A. G. (1994). Inductive Learning Algorithms for Complex Systems Modelling. Boca Raton, FL: CRC Press Inc.

Maddala, G., & Lahiri, K. (2009). *Introduction to Econometrics* (4th ed.). Willey.

Marchev A., Motzev M., & Muller, J-A. (1985). Applications of Self-Organization Procedures for Business System Models Building, *Automatics*, 1, 37-44.

Marchev, A., & Motzev, M. (1985). Computer Macro-Economic Models for Simulation Experiments, *Systems Analysis and Simulation*, 28(II), (now *Annual Review in Automatic Programming*, 12), 145-150.

Mark, J., & Goldberg, M. (2001, January). Multiple Regression Analysis and Mass Assessment: A Review of the Issues, *The Appraisal Journal*, 89–109.

Motzev, M. (1986). Dynamic Coefficients in Simultaneous Equations Models, *Scientific papers.* 1. Publishing House “Economy”, Bulgaria, 119-148.

Motzev, M. (2014). *Predictive Analytics in Business Games and Simulations*. W. Bertelsmann Verlag GmbH & Co. KG, Bielefeld, Germany.

Motzev, M., & Marchev, A. (1988). Multi-Stage Selection Algorithms in Simulation, *Proceedings of XII IMACS World Congress*, 4, France, 533-535.

Motzev, M., Muller, J-A., & Marchev, A. (1986). Macro-Economic Systems Modelling and Forecasting Using Auto-Regressive Models, *Social Management*, 6, 77-93.

Mueller J. A., & Lemke, F. (2003). *Self-Organizing Data Mining: An Intelligent Approach To Extract Knowledge From Data.* Victoria, BC: Trafford Publishing.

Onwubolu, G. (2008, May). Design of hybrid differential evolution and GMDH networks for modeling and prediction, *Information Sciences*, 178(18), 3616–3634.

Onwubolu, G. (Ed.). (2009). *Hybrid Self-Organizing Modeling Systems*. Springer-Verlag Berlin Heidelberg.

Park, S-B. (1974). On Indirect Least Squares Estimation of a Simultaneous Equation System, *The Canadian Journal of Statistics, 2*(1), 75–82.

Pearl, J. (2014). Trygve Haavelmo and the Emergence of Causal Calculus, *Econometric Theory*. Special issue on Haavelmo Centennial.

Prokesch, S. (1989, October 12). Norwegian Wins Nobel For His Work in Economics, *The New York Times*.

Shahwan, T., & Lemke, F. (2005, July). Forecasting Commodity Prices for Predictive Decision Support Systems, *Proceedings of the EFITA/WCCA joint congress on IT in agriculture*, Portugal, 23-32.

Theil, H. (1971). *Principles of Econometrics*. New York: John Wiley.

Tshilidzi, M. (2013). *Economic Modeling Using Artificial Intelligence Methods*. Springer-Verlag.

Valavanis, S. Econometrics, New York, McGraw-Hill, 1959, pp. 93-94.

von Neumann, J. (1941). Distribution of the ratio of the mean square successive difference to the variance, *Annals of Mathematical Statistics*, 12(4), 367–395.

Zellner, A. (1962). An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias, *Journal of the American Statistical Association*, 57(298), 348–368.

Zellner, A., & Theil, H. (1977). Three-stage least squares: simultaneous estimation of simultaneous equations, *Econometrica,* 30(1), 54–78.

***

CHAPTER 10. FORECASTING, BUSINESS INTELLIGENCE AND BUSINESS ANALYTICS

10.1. Business Intelligence and Business Analytics

In a 1958 article, IBM researcher Hans Peter Luhn (1958) used the term ***Business Intelligence (BI)*** and he defined intelligence as: "*the ability to apprehend the interrelationships of presented facts in such a way as to guide action towards a desired goal*" (p.314). Later in 1989, Howard Dresner (who later became a Gartner Group[1] analyst) proposed ***BI*** as an umbrella term to describe "concepts and methods to improve business decision making by using fact-based support systems." (as cited in Power, 2007, p. 6)

Today, ***Business Intelligence (BI)*** usually refers to skills, processes, technologies, applications and practices used to support decision making. There are many confusions and misinterpretations of ***BI***:

- ***BI*** and ***Decision Support System*** – as BI often aims to support better business decision-making it is sometimes called a decision support system (DSS). Actually, DSS is a class of information systems (including but not limited to computerized systems) that support business and organizational decision-making activities. According to Sol (as cited in Henk et al., 1987, pp.1-2.) the definition and scope of DSS has been migrating over the years. In the 1970s DSS was described as "a computer based system to aid decision making". In the late 1970s, the DSS movement started focusing on "interactive computer-based systems which help decision-makers utilize data bases and models to solve ill-structured problems". In the 1980s, DSS were expected to provide systems "using suitable and available technology to improve effectiveness of managerial and professional activities", and at the end of the 1980s, DSS faced a new challenge towards the design of intelligent workstations.
- ***BI*** and ***Competitive Intelligence*** are often used as synonyms, because they both support decision making. However, BI uses technologies, processes, and applications to analyze mostly internal, structured data and business processes while competitive intelligence is done by gathering, analyzing and disseminating information with or without support of technology and applications. To support decision making BI focuses on all-source information and data (unstructured or structured), mostly external to, but also internal to a company. Fleisher (2003, pp. 56-69) compares and contrasts competitive intelligence with business intelligence, competitor intelligence, knowledge management, market intelligence, marketing research, and strategic intelligence.

[1] Now Gartner Inc, an information technology research and advisory firm headquartered in Stamford, Connecticut

- ***BI*** and ***Data warehousing – BI*** is a term commonly associated with data warehousing. In fact, many of the tool vendors position their products as ***business intelligence software*** rather than data warehousing software. Often BI applications use data gathered from a data warehouse or a data mart. However, not all data warehouses are used for business intelligence nor do all business intelligence applications require a data warehouse. We can summarize, that data warehousing or a data mart system is the backend, or one infrastructural component for achieving business intellignce.
- ***BI*** and ***Computer software solutions*** – the misunderstanding comes from a narrow definition that BI is a broad category of computer software solutions that enables a company or organization to gain insight into its critical operations through reporting applications and analysis tools[2]. BI applications include a variety of components such as tabular reports, spreadsheets, charts, and dashboards. According some authors[3] BI is being fundamentally changed by eXtensible Markup Language and the emerging Web Services model. In the beginning of mid-2000s, ***Business Intelligence 2.0 (BI 2.0)***[4] appeared. ***BI 2.0*** refers to *new tools and software for business intelligence* that enable, among other things, dynamic querying of real-time corporate data by employees, as well as a more web- and browser-based approach to such data, as opposed to the proprietary querying tools that had characterized previous business intelligence software.

In this text we will use a general definition, provided by an online computer dictionary [5]):

Business intelligence (BI) *is a broad category of applications and technologies for gathering, storing, analyzing, and providing access to data to help enterprise users make better business decisions. BI applications include the activities of decision support systems, query and reporting, online analytical processing (OLAP), statistical analysis, forecasting, and data mining.*

BI technologies provide historical, current, and predictive views of business operations. Common functions of Business Intelligence technologies are reporting, online analytical processing, analytics, data mining, business performance management, benchmarking, text mining, and predictive analytics. As figure 10-1 suggests (see Davenport & Harris, 2007, p.7), business intelligence includes both data access and reporting, and analytics, and ***Business***

[2] http://www.informationbuilders.com/business-intelligence.html

[3] http://www.bitpipe.com/bi/bi_overview.jsp

[4] http://intelligent-enterprise. informationweek.com/showArticle.jhtml;jsessionid= VS5YU2BF2QNOXQE1GHOSKH4ATMY32JVN?articleID=197002610

[5] http://searchdatamanagement.techtarget.com/sDefinition/0,,sid91_gci213571,00.html

Analytics** (**BA**)* should be an element of ***BI. However, there are different opinions and usually ***BA*** refers to the skills, technologies, applications and practices for continuous iterative exploration and investigation of past business performance to gain insight and drive business planning. In contrast with ***BI, BA*** focuses on developing new insights and understanding of business performance whereas ***BI*** traditionally focuses on using a consistent set of metrics to both measure past performance and guide business planning (Beller & Barnett, 2009, p.5).

According to Davenport (2007), Shmueli et al. (2007) and other researchers, ***Business Analytics*** can make extensive use of data, statistical and quantitative analysis, explanatory and predictive modeling, and fact-based management to drive decision making. ***BI*** is querying, reporting, OLAP, and alerts, i.e. tools which can answer questions such as: what happened, how many, how often, where, what actions are needed. ***BA*** can answer more sophisticated questions like: what if these trends continue (i.e. what-if analysis & forecasting); what will happen next (i.e. prediction); what is the best that can happen (i.e. optimization) and so on - see Fig.10-1.

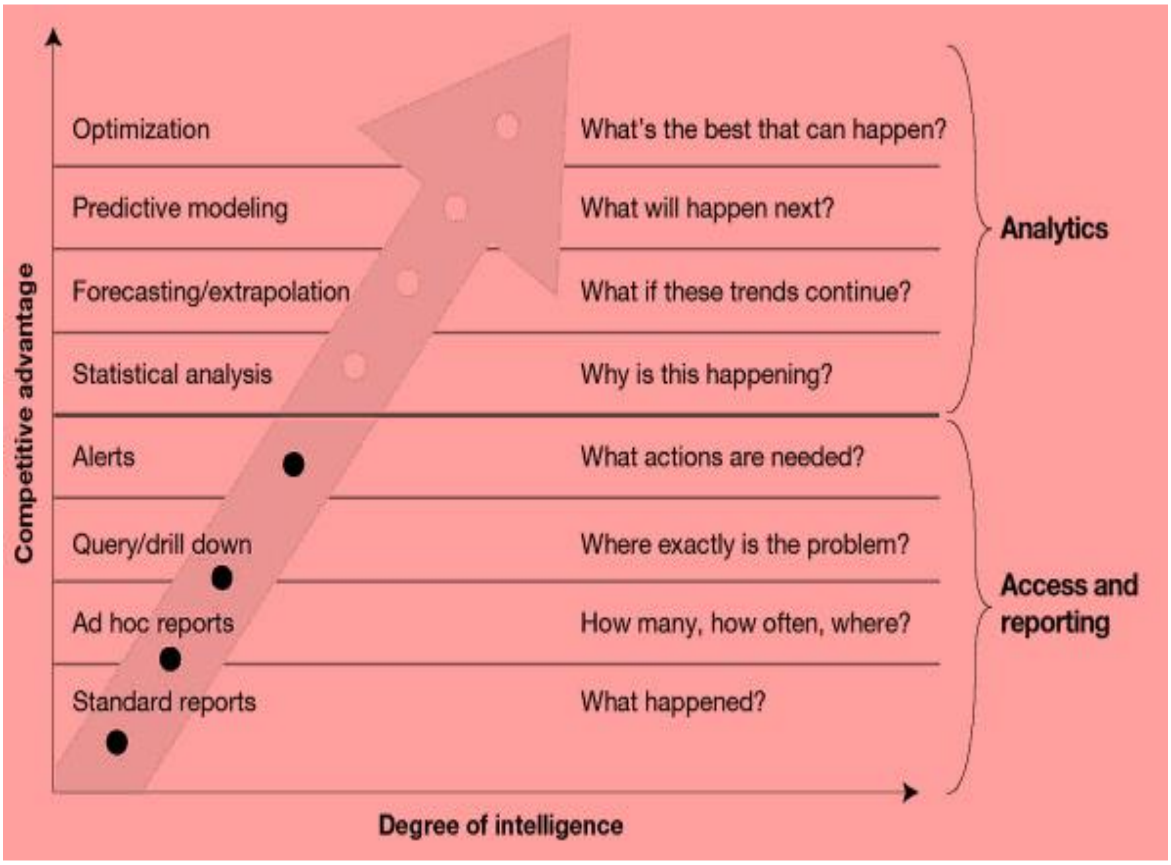


Fig.10-1 Business intelligence and analytics
(Source: Davenport & Harris, 2007, p.8)

We can summarize that ***BI*** usually are related to *information systems* and *software applications* whilst ***BA*** is considered to be more oriented to *analytics*. In spite of what is the most precise definition, both of them have applications in business forecasting and in this textbook, we will discuss only this function, often referred to as *predictive analytics.*

Predictive analytics is an important part of ***BA*** which encompasses a variety of techniques from statistics, data mining and game theory that analyze current and historical facts to make predictions about future events. In business, *predictive models* exploit patterns found in historical and transactional data to identify risks and opportunities. Models capture relationships among many factors to allow assessment of risk or potential risk associated with a particular set of conditions, guiding decision making for candidate transactions.

The approaches and techniques used to conduct predictive analytics can broadly be grouped into *regression techniques* (already discussed in this textbook) and *machine learning techniques*, based on *artificial neural networks*, *genetic algorithms* and other intelligent techniques, which will be discussed in the following chapters.

10.2. Predictive Analytics and Data Mining

The simplest definition of *analytics* is "the science of analysis". A simple and practical definition, however, would be how an entity (let’s say business) arrives at an optimal or realistic decision based on existing data. Business managers may choose to make decisions based on past experiences or rules of thumb, or there might be other qualitative aspects to decision making, but unless there are data involved in the process it would not be considered analytics.

The common definition[6] for predictive analytics is: “*The use of statistics and modeling to determine future performance based on current and historical data. Predictive analytics look at patterns in data to determine if those patterns are likely to emerge again, which allows businesses and investors to adjust where they use their resources in order to take advantage of possible future events.*”

Generally, *predictive analytics* is used to denote *predictive modeling*, *scoring of predictive models*, and *forecasting*. However, people are increasingly using the term to describe related analytical disciplines, such as descriptive modeling and decision modeling or optimization. These disciplines also involve rigorous data analysis, and are widely used in business for segmentation and decision making but have different purposes and the statistical techniques underlying them vary.

[6] See http://www.investopedia.com/terms/p/predictive-analytics.asp

In this book we assume that *Predictive analytics* is an area of analysis that deals with extracting information from data and using it to predict future trends and behavior patterns. The core of predictive analytics relies on capturing relationships between explanatory variables and the predicted variables from past occurrences, and exploiting them to predict future outcomes.

According to *Investopedia Dictionary* quoted above, there are several types of *predictive analytics* methods available. Predictive models look at past data to determine the likelihood of certain future outcomes, while descriptive models look at past data to determine how a group may respond to a set of variables.

Predictive analytics is a decision-making tool in a variety of industries. For example, insurance companies examine policy applicants to determine the likelihood of having to pay out for a future claim based on the current risk pool of similar policy holders, as well as past events that have resulted in payouts. Marketers look at how consumers have reacted to the overall economy when planning a new campaign, and can use shifts in demographics to determine if the current mix of products will entice consumers to make a purchase.

Common applications of analytics include the study of business data using statistical analysis in order to discover and understand historical patterns so that business performance is predicted and improved in the future. What is more, some people use the term to denote the use of mathematics in business. Others hold that field of analytics includes the use of Operations Research, Statistics and Probability. However, it would be erroneous to limit the field of analytics to only statistics and mathematics.

According to Davenport (2007), as we have already mentioned, **analytics** means "*the extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to drive decisions and actions*" (p. 7). Analytics is a subset of what has come to be called ***business inte1ligenci***, i.e. a set of technologies and processes that use data to understand and analyze business performance. Each of these approaches addresses a range of questions about an organization's business activities. The questions that analytics can answer represent the higher-value and the most proactive end of this spectrum as shown in Fig.10-2.

Analytics closely resembles *statistical analysis* and *data mining*, but tends to be based on modeling, involving extensive computation. Some fields within the area of analytics are enterprise decision management, marketing analytics, predictive science, strategy science, credit risk analysis and fraud analytics. In this book, however, we will concentrate and discuss only techniques that can be used in business forecasting.

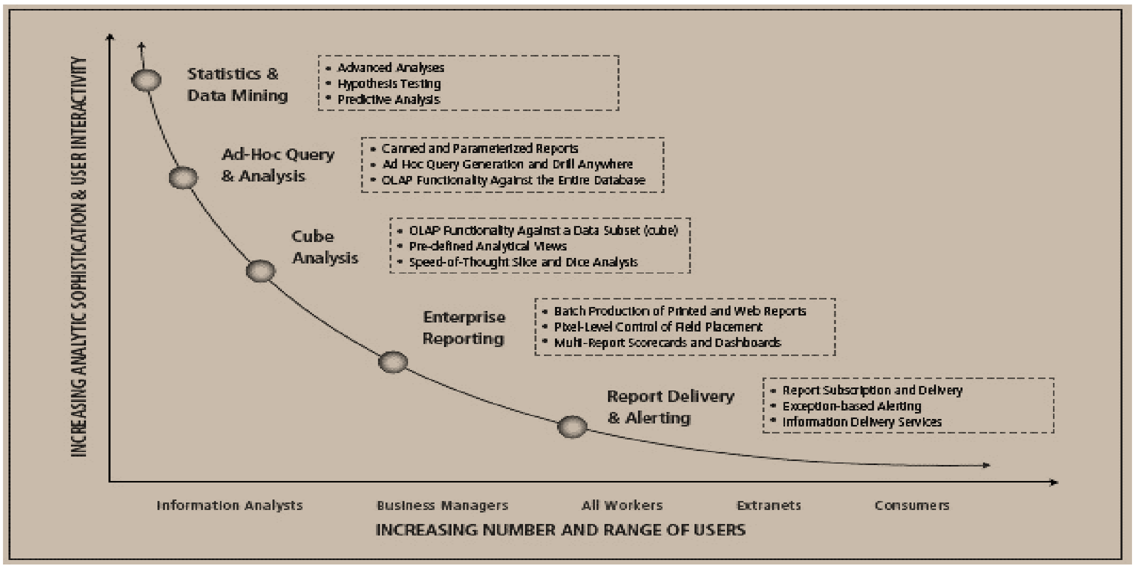

INCREASING ANALYTIC SOPHISTICATION & USER INTERACTIVITY
Statistics & Data Mining
• Advanced Analyses
• Hypothesis Testing
• Predictive Analysis
Ad-Hoc Query & Analysis
• Canned and Parameterized Reports
• Ad Hoc Query Generation and Drill Anywhere
• OLAP Functionality Against the Entire Database
Cube Analysis
• OLAP Functionality Against a Data Subset (cube)
• Pre-defined Analytical Views
• Speed-of-Thought Slice and Dice Analysis
Enterprise Reporting
• Batch Production of Printed and Web Reports
• Pixel-Level Control of Field Placement
• Multi-Report Scorecards and Dashboards
Report Delivery & Alerting
• Report Subscription and Delivery
• Exception-based Alerting
• Information Delivery Services
Information Analysts
Business Managers
All Workers
Extranets
Consumers
INCREASING NUMBER AND RANGE OF USERS

Fig.10-2 Five styles of BI have evolved to support different needs, from advanced professional analysis to basic information consumption
(Source: MicroStrategy White paper, 2002, p.10)

The concept of traditional statistical analysis was already discussed in previous chapters and here we will concentrate on *data mining*. As it often happens in new concepts, definitions about *data mining* also vary.

Turban & Aronson (2001) define ***Data mining*** as a "*term used to describe knowledge discovery in databases. It includes tasks known as knowledge extraction, data archaeology, data exploration, data pattern processing, data dredging, and information harvesting*." (p.148). Berry & Linoff (2000) give a more precise explanation:

> *Data mining is the process of exploration and analysis* (by automatic or semi- automatic means) *of large quantities of data in order to discover meaningful patterns and rules*. (p.8)

The second definition is better because it puts emphasis on large quantities of data since data volumes continue to increase and suggests that the patterns and rules to be found ought to be meaningful. The phrase "by automatic or semi-automatic means" was put in brackets not because it is untrue, without automation it would be impossible to mine the huge quantities of data being generated today (as we will discuss it further on), but because as authors mentioned, we feel there has come to be too much focus on the automatic techniques and not enough on the exploration and analysis. For the purpose of this textbook, we will discuss *data mining* in the context of this definition as *the process of extracting meaningful patterns from large amounts of data.*

Humans have been "manually" extracting patterns from data for centuries, but the increasing volume of data in modern times has called for more automated approaches. The current situation with the abundance of data, coupled with the need for powerful data analysis tools, was described a long time ago by Finlay: "*Without an efficient means of filtering and aggregating data, a manager could be* ***data rich yet information poor***" (as cited in Lucey, 1991, p.16.)

Current trends are the fast-growing, tremendous amount of data (collected and stored in large and numerous data repositories), which has far exceeded our human ability for comprehension without powerful tools. Second, these data archives are seldom visited and, as a result, third, important decision are often made based not on the information-rich data stored in data repositories, but rather on a decision maker intuition. This happens because the decision maker does not have the tools to extract the valuable knowledge embedded in the vast amounts of data.

Early methods of identifying patterns in data include Bayes' theorem (1700s) and Regression analysis (1800s). The proliferation, ubiquity and increasing power of computer technology has increased data collection and storage. As data sets have grown in size and complexity, direct hands-on data analysis has increasingly been augmented with indirect, automatic data processing. This has been aided by other discoveries in computer science, such as neural networks, clustering, genetic algorithms (1950s), decision trees (1960s) support vector machines (1980s) and others.

The urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data was pointed out back in the 1990s of last century by Fayyad, Piatetsky-Shapiro & Smith (1996). There are several reasons which can be cited in support of the growing popularity of data mining today according to Marakas (2003, p. 328):

- The single greatest reason is the ever-increasing volume of data that require processing. The amount of data accumulated by businesses and organizations each day varies according to function and objective. In 2000, a GTE research center report suggests that scientific and academic organizations store approximately one terabyte of new data each day, even though the academic community is not the leading supplier of new data worldwide.
- Another reason for the growing popularity is an increasing awareness of the inadequacy of the human brain to process data, particularly in situations involving multi-factorial dependencies or correlations. Our biases formed by previous experience in data analysis often hold us hostage. As such, our objectivity in data analysis scenarios is often suspicious.

- Finally, a third reason for the growing popularity of data mining is the increasing affordability of machine learning. An automated data mining system can operate at a much lower cost than an army of highly trained (and paid) professional statisticians. Although data mining does not entirely eliminate human participation in problem solving, it significantly simplifies the tasks and allows humans to better manage the process.

An increasingly common synonym for data mining techniques is *knowledge data discovery* (or ***Knowledge Discovery in Databases – KDD***). It should be noted that ***KDD*** applies to all activities and processes associated with discovering useful knowledge from aggregate data. Using a combination of techniques including statistical analysis, neural and fuzzy logic, multidimensional analysis, data visualization, and intelligent agents, ***KDD*** can discover highly useful and informative patterns within the data that can be used to develop predictive models of behavior or consequences in a wide variety of knowledge domains.

The need for distinction between the ***KDD*** process and the data-mining step (within the process – see Fig.10-3) was mentioned in the very early articles in this area. According to Fayyad et al. (1996) "*Data mining is a step in the KDD process that consists of applying data analysis and discovery algorithms that produce a particular enumeration of patterns (or models) over the data*" (p.39).

KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process. ***Data mining* (*DM*)** is the application of specific algorithms for extracting patterns from data.

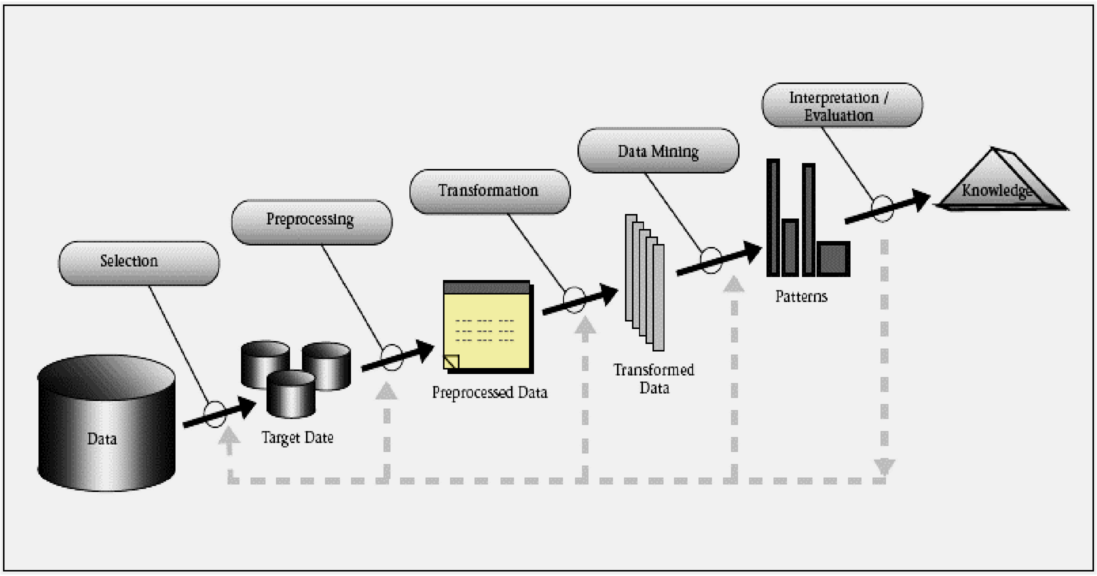


Fig. 10-3 Knowledge discovery in databases process
(Source: Fayyad et al., 1996, p.41)

Additional steps in the *KDD* process, such as data preparation, data selection, data cleaning, incorporation of appropriate prior knowledge, and proper interpretation of the results of mining, are essential to ensure that useful knowledge is derived from the data. Blind application of *data-mining* methods, criticized as *data dredging* in the statistical literature, can be a dangerous activity, easily leading to the discovery of meaningless and invalid patterns. Similarly, ***Data Mining*** should be considered as a step in the forecasting process, discussed in Chapter 3.

DM involves the following activities in extracting meaningful new information from the data: *Classification, Estimation, Prediction, Affinity grouping or association rules, Clustering, Description and visualization.* According to Berry & Linoff (2000, p.8) the first three tasks – classification, estimation, and prediction—are all examples of *directed* **DM**. In *directed* **DM**, the goal is to use the available data to build a model that describes one particular variable of interest in terms of the rest of the available data. The next three tasks are examples of *undirected* **DM**. In *undirected* **DM**, no variable is singled out as the target; the goal is to establish some relationship among all the variables.

Directed data mining is a top-down approach, used when we know what we are looking for. This often takes the form of predictive modeling where we know exactly what we want to predict. In this case the model is considered as a *black box* (Fig. 10-4), i.e. it is not important what the model is doing, we just want the most accurate result possible. We are using already known examples, such as prospects who already received an offer (and either did or did not respond), and we are applying information gleaned from them to unknown examples, such as prospects who have not yet been contacted, to answer questions like “Who is likely to respond to our next offer, based on the history of previous marketing campaigns?”.

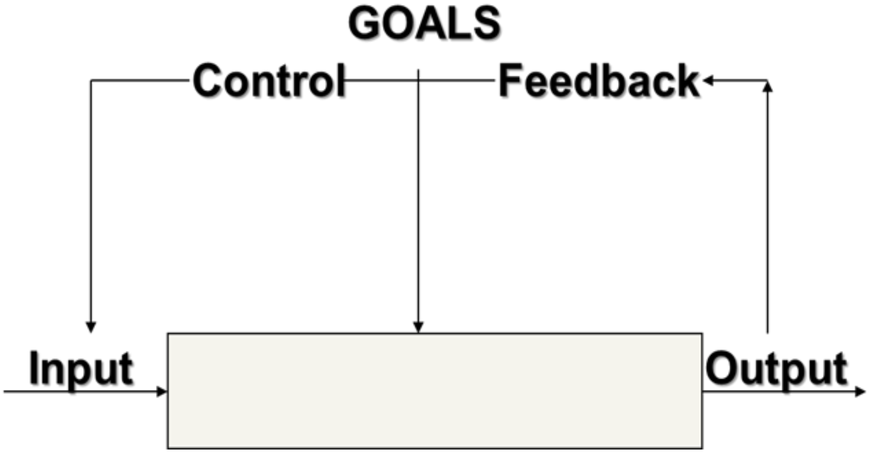


Fig.10-4 The General Systems Model as a Black Box

Undirected data mining is a bottom-up approach that finds patterns in the data and leaves it up to the user to determine whether or not these patterns are important, i.e. it is about discovering new patterns inside the data. These patterns provide insight, and this insight might prove very informative. This form of data mining is represented with *semitransparent boxes* (see Fig. 10-5). Unlike directed ***DM***, here users want to know what is going on, and how the model is coming up with an answer.

Undirected data mining is necessarily interactive. Advanced algorithms can find patterns in the data, but only humans (i.e. managers, professionals etc.) can determine whether the patterns have any significance and what the patterns might mean.

Both undirected data mining and directed data mining are valuable in many data mining efforts. Undirected data mining is often used during the data exploration steps. What is in the data? What does it look like? Are there any unusual patterns? What does the data suggest for customer segmentation? These types of questions are answered using tools that support clustering, visualization, and market basket analysis.

At the same time, some predictive modeling techniques, notably decision trees (see next chapter), explain the models they produce. These techniques sometimes provide important insights in addition to the predictions they make. Two things are happening. An example of directed data mining is that a decision tree makes predictions. An example of undirected data mining is that a person looks at a decision tree and possibly notices an interesting pattern.

It is worth noting, that these two approaches are not mutually exclusive. Data mining efforts often include a combination of both. Even when building a predictive model, it is often useful to search for patterns in the data using undirected techniques. These can suggest new customer segments and new insights that can improve the directed modeling results.

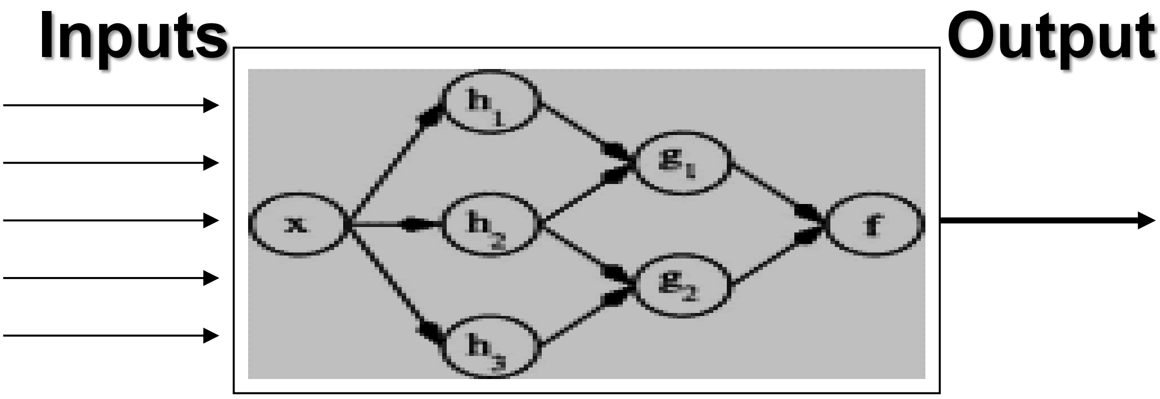


Fig.10-5 Data mining with a semitransparent box

There is a discussion that there should not be a separate heading for prediction, because prediction can be thought of as classification or estimation. The difference is one of emphasis - when data mining is used to classify a phone line as primarily used for Internet access or a credit card transaction as fraudulent, we do not expect to be able to go back later to see if the classification was correct. The classification may be correct or incorrect, but the uncertainty is due only to incomplete knowledge. The computer is or is not used primarily for online business and the credit card transaction is or is not fraudulent. With enough effort, it is possible to check. Predictive tasks are different because data records are classified according to estimated future value. With prediction, the only way to check the accuracy of the classification is to wait and see.

As a matter of fact, any of the techniques used for classification and estimation can be adapted for use in prediction by using training examples where the value of the variable to be predicted is already known, along with historical data for those examples. The historical data is used, to build a model that explains the current observed behavior. When this model is applied to current inputs, the result is a prediction of future behavior.

10.3. Data Mining Process

As we defined above, *Data Mining* is the process of examining large amounts of data in search of hidden patterns and predictive information (mostly in an automated manner), which allows organizations to make better decisions. Data Mining uses database technology, modeling techniques, statistical analysis and machine learning to find hidden patterns and make predictions which elude all but the most expert users and generate scoring or predictive models based on actual historical data.

How does data mining find information that business users and analysts did not already know? How does it find information about what is likely to happen next? In general, data mining platforms assist and automate the process of building and training highly sophisticated data mining models, and applying these models to larger datasets. A white paper published by MicroStrategy (An Architecture for Enterprise Business Intelligence, 2005, pp. 162-173) describes in details the data mining process and as an example, we'll suppose that a credit card company plans to develop a promotional campaign to recruit new customers:

1. Create a predictive model from a data sample – A sample dataset of customers who have responded to past promotional campaigns is extracted from the company data base. This sample contains customer characteristics and trends that potentially can be used to predict "responsiveness" like: Where do they live? What gender are they? What age range do they fall in? What is their income range? What is their marital status? What is their level of education? What are their past purchases? Have they responded to past campaigns? What is their credit history?

Advanced statistical and mathematical techniques like regression analysis and machine learning algorithms (see next chapters) are used to identify the significant characteristics and trends in predicting responsiveness, and a predictive model is created using these as inputs. Note that often only a small subset of all characteristics and trends in the sample dataset are generally used in the model.

2. Train the model against datasets with known results – The new predictive model is applied to additional data samples with known outcomes to validate whether the model is reasonably successful at predicting the known results. In this example it would be data based on historical campaign responses. This gives a good indication of the accuracy of the model. It can then be further trained using these samples to improve its accuracy.

3. Apply the model against a new dataset with an unknown outcome – Once the predictive model is validated against the known data, it is used for scoring, which is defined as the application of a data mining model to forecast an outcome. In the current example, the predictive model is applied to the new customer/prospect database to predict the likelihood of a customer responding to the marketing campaign and will generate a score for each customer that indicates his or her likelihood to respond. This score can be a simple binary result, such as Yes/No, or it could be a number indicating the propensity or confidence in that customer responding, say "97%." In both cases the end result will yield those customers that have a high probability of responding to the marketing promotion.

Figure 13-6[7] represents Data Mining workflow in MicroStrategy, one of the leading BI platforms according Gartner's research (see Fig. 10-8). The "***Create-Train-Apply***" process is typically the domain of the statistician or the data mining analyst. A solid understanding of data mining concepts, statistical concepts, techniques, and data mining tools is necessary in the "Create" and "Train" steps. Applying the predictive model requires less expertise and is available for all business users.

[7] The term PMML used in the chart stands for Predictive Model Markup Language (author's note)

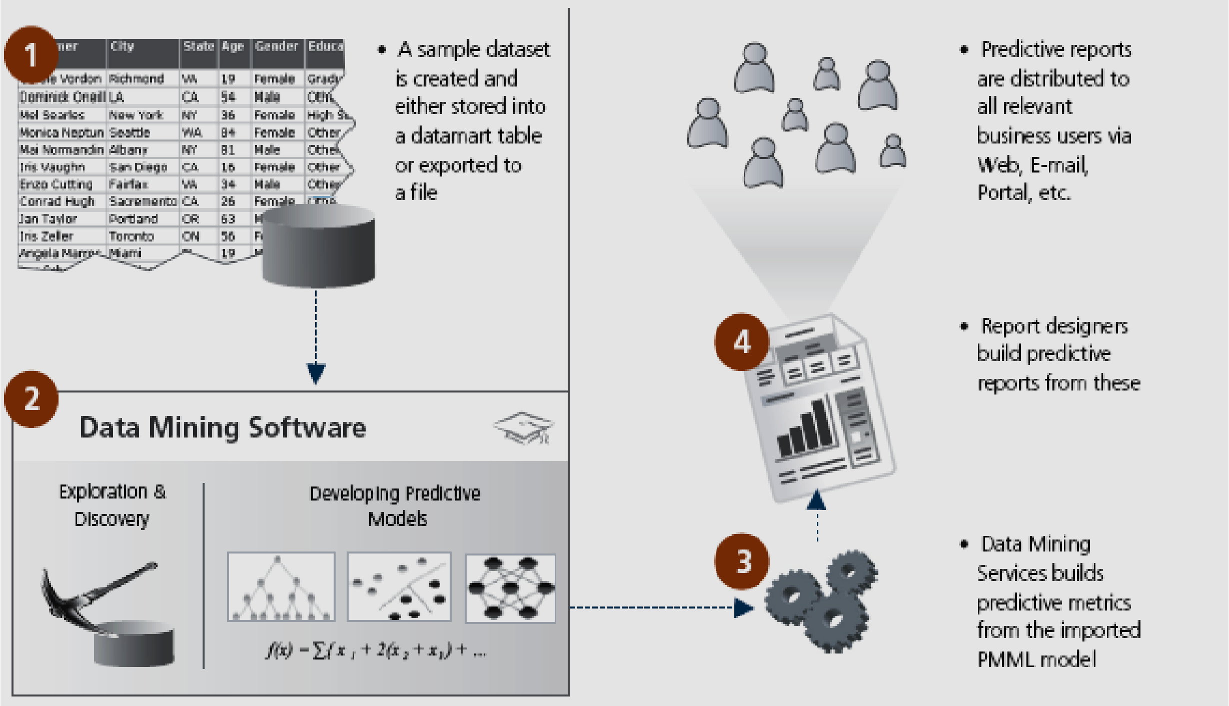

1
• A sample dataset is created and either stored into a datamart table or exported to a file
2
Data Mining Software
Exploration & Discovery
Developing Predictive Models
$f(x) = \sum(x_1 + 2(x_2 + x_3) + ...$
3
• Data Mining Services builds predictive metrics from the imported PMML model
4
• Report designers build predictive reports from these
• Predictive reports are distributed to all relevant business users via Web, E-mail, Portal, etc.

Fig. 10-6 Data Mining Workflow in MicroStrategy platform
(Source: MicroStrategy White paper, 2005, p.171)

If we consider data mining in more details involving selecting, exploring, and building models and using large amounts of data to uncover concealed information, it leads to iterative process. To provide a path for this process to follow, SAS Institute Inc. developed a five-step data mining cycle process known as *SEMMA*: Sample, explore, modify, model, and assess (Reference Configurations, 2007).

The first step in data mining is to create one or more data tables by sampling data from a data warehouse. Mining a representative sample rather than the entire volume reduces the processing time required for the mining process. The next step is to explore the samples visually or numerically for trends or groups. Modifying the data enables the analyst to create, select, and transform one or more variables to focus in a particular direction or to augment the data for clarity. After the data is accessed and modified, analysts can construct models to explain patterns in data. The last step in data mining is to assess the accuracy of the model. This involves testing the model against a holdout sample of data not used in the model-building process, also called testing set of data (see next chapter).

IBM Corp. has a slightly different interpretation of the data mining process and other companies may have their own view as well. In spite of these variations, we can find three main components, which Berry & Linoff (2000, p.93) call the *Three Pillars* of Data Mining*: Data, Modeling Skills* and *Data Mining Techniques*.

All important points (data collection, data organization, descriptive data analysis, etc.) about *Data* as a component of the Business Forecasting process were already discussed in this book. Here, we will mention briefly some data characteristics specific to data mining. Actually, without data business would have to rely on intelligent guesswork which is still often the case even when data is available. The power of data mining is leveraging the data that a company collects to make better informed business decisions.

Data mining has a very simplistic view of data that consists of a single table (or file or view) with well-defined columns (see step 1 in Fig. 13-6). Most algorithms prefer that there are no missing data, and that all the values make sense. Unfortunately, data in the real world does not look like this. It comes from many different sources in many different formats, sometimes incomplete, always dirty. The process of bringing together all the disparate sources of data and extracting useful features from them is the biggest challenge in data mining.

The internal data that a company collects can be a competitive advantage, because competitors do not have access to it. The most voluminous source of data in many industries are the actual transactions recorded by individuals, like each purchase made on a credit card, every web page viewed, each line item recorded at a grocery store check-out, every telephone call made. These are the richest source of information, and at the same time, the most challenging. To be useful for data mining, they must be summarized and yet, it does not work to store presummarized data because the same transactions can usefully be summarized in many different ways.

There is also a lot of data which comes from external sources, such as Demographics, psychographics, and web graphics (information about individuals and households that bureaus glean from many different sources); Data shared within an industry (credit reports, credit scores, and catalog subscriptions); Summary data about geographic areas, store catchment areas, and so on; Purchased external lists that meet some particular criteria; Data shared from strategic business partners.

It is important to know how to work with data for data mining in the real world. There are many issues that arise with data at different stages of data analysis and require specific skills to act in planning and doing data mining as a part of the general forecasting process. We assume that these skills are available now, in some form, after completing all previous chapters in this textbook. Some specific points in data mining techniques and algorithms, like neural networks, decision trees, genetic algorithms, etc. are discussed in the next two chapters.

The set of *modeling skills* needed to build predictive models in data mining in general is the same as in business forecasting process. It is worth noting that the methodology for building effective predictive models, discussed in Chapter 3, is working well for both directed and undirected data mining. Detailed steps and data mining algorithms for building different types of forecasting models, which find patterns in data from the past to make predictions about unknown outcomes are discussed in Chapters 8, 9 and 12.

The predictive modeling process leads to many interesting insights, especially during the data exploration phase or while analyzing how models are working. All elements and features discussed earlier (such as model accuracy and selection, structural identification and so on) are important. The data miner needs to be aware of these factors to judge when and whether predictive models will be effective as far as the purpose of most data mining techniques is to improve business decisions.

Data mining techniques are like anything else a computer does, such as storing files or creating a spreadsheet. The techniques are general approaches to solving problems, and there are usually many ways to approach the technique. Each of these ways is a different algorithm. The algorithms are like recipes with step-by-step instructions explaining what is happening.

The knowledge and skills in statistics accumulated so far in this textbook will work as a background when we try to go into enough detail of the data mining techniques and to convey how they actually work. It is also important to have some understanding of their inner workings to know when to apply them, how to interpret the results, and whether or not they are working. In the following Chapter 11 we are going to discuss techniques that are related to forecasting with enough detail so you can:

- Distinguish between different techniques knowing their advantages and disadvantages;
- Follow the techniques as they are used in the real life business examples;
- Understand which technique is most appropriate for a given business problem;
- Become familiar with important variations.

In addition to major techniques that are found in most comprehensive data mining tools (such as *decision trees, neural networks* and *clustering*), another approach which shows great promises as pointed out in Chapters 8 and 9, based on *Group Method of Data Handling* will be discussed and a few unique hybrid techniques in business forecasting will be described in detail in Chapter 12.

10.4. Business Intelligence Solutions and Platforms

BI and data mining techniques are available on a wide range of computing platforms, from individual desktops, to departmental servers, to the most powerful super computers with parallel data processing. However, as desktops are becoming more powerful, it is often not necessary to purchase expensive hardware to run their intelligent algorithms.

BI usage can be categorized into the following groups:

- **Business operations reporting** – this is the most common form of BI. It often manifests itself in the standard weekly or monthly reports that need to be produced.
- **Forecasting** – it helps to find answers of questions like: What is the expected level of sales if we spend $1000 in advertising? What happened if the price of oil shoots up to $150 a barrel? And so on…
- **Dashboard** – the primary purpose is to convey the information at a glance. For this audience, there is little, if any, need for drilling down on the data. At the same time, presentation and ease of use are very important for a dashboard to be useful.
- **Multidimensional analysis** – this is "slicing-and-dicing" of the data. It offers good insight into the numbers at a more granular level. This requires a solid data base (usually a data warehouse or a data mart) backend, as well as business-savvy analysts to get to the necessary data.
- **Finding correlation among different factors** – this is diving very deep into business intelligence. Questions asked are like, "How do different factors correlate to one another?", "Are there significant time trends that can be anticipated?", etc.

Since the early 1990s, BI applications have evolved dramatically – from operational “green-bar” reports generated by mainframes, to statistical modeling of marketing campaigns, to multi-dimensional ***OLAP***[8] environments for analysts, to dashboards and scorecards for executives – companies began to demand more ways to report on and analyze data. The dramatic expansion of data warehousing combined with the widespread adoption of enterprise applications, such as Enterprise Resource Planning (ERP) and Customer Relationship Management (CRM), as well as the overall increase in computer literacy, fueled this exponential demand for BI reporting and analysis applications.

Fig. 10-7 shows an example of intelligent platforms (*SAS Enterprise Miner and Forecast Server*) developed as a product of more than a-20-year business relationship that is still going

[8] ***OLAP*** – On Line Analytical Processing (author’s note)

strong, between Sun Microsystems and SAS Institute, which has a long, proven track record of delivering open, scalable, and reliable technology solutions.

Looking at the diverse range of BI functionality in the market over the past 15 years, i.e. the historical development of BI applications and BI technology, we can describe *five common styles of BI* that have evolved during the past decade (The 5 Styles of Business Intelligence, 2002), where each style represents a certain characteristic usage and function by end users:

1. **Enterprise Reporting** – Broadly deployed pixel-perfect report formats for operational reporting and scorecards/dashboards targeted at information consumers. *Report writers* were used to generate highly formatted static reports destined for broad distribution to many people.
2. **Cube Analysis** – *OLAP* slice-and-dice analysis of limited data sets, targeted at managers and others who need a safe and simple environment for basic data exploration within a limited range of data. Cube-based BI tools were used to provide simple slice-and-dice analytical capabilities to business managers.
3. **Ad Hoc Query and Analysis** – Full investigative query into all data, as well as automated slice-and-dice OLAP analysis of the entire database – down to the transaction level of detail if necessary. Targeted at information explorers and power users. *Relational OLAP tools* were used to allow power users to query the database for any answer, slice-and-dice the entire database and surf down to the lowest level of transactional information.

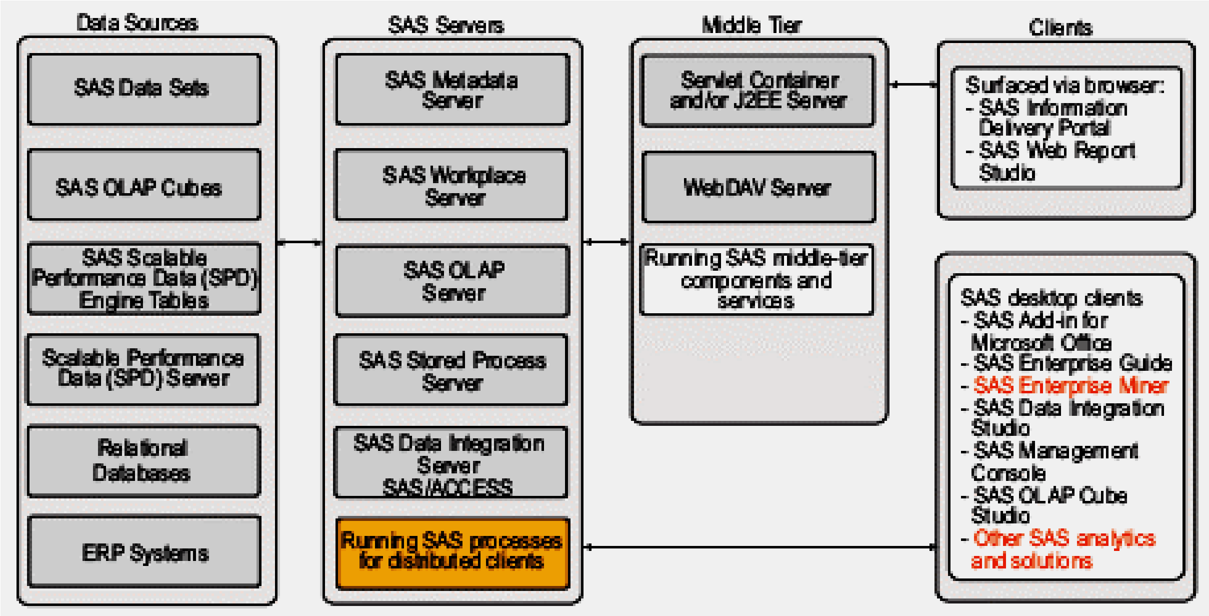

Data Sources
SAS Data Sets
SAS OLAP Cubes
SAS Scalable Performance Data (SPD) Engine Tables
Scalable Performance Data (SPD) Server
Relational Databases
ERP Systems
SAS Servers
SAS Metadata Server
SAS Workplace Server
SAS OLAP Server
SAS Stored Process Server
SAS Data Integration Server SAS/ACCESS
Running SAS processes for distributed clients
Middle Tier
Servlet Container and/or J2EE Server
WebDAV Server
Running SAS middle-tier components and services
Clients
Surfaced via browser:
- SAS Information Delivery Portal
- SAS Web Report Studio
SAS desktop clients
- SAS Add-in for Microsoft Office
- SAS Enterprise Guide
- SAS Enterprise Miner
- SAS Data Integration Studio
- SAS Management Console
- SAS OLAP Cube Studio
- Other SAS analytics and solutions

Fig.10-7 SAS Enterprise Intelligence Platform logical diagram
(Source: Sun Microsystems White paper, 2007, p.6)

4. **Statistical Analysis and Data Mining** – Full mathematical, financial, and statistical treatment of data for the purposes of correlation analysis, trend analysis, financial analysis and projections. It is targeted at the professional information analysts. *Statistical and data mining tools* were used to perform predictive modeling or to discover the cause-and-effect correlation between two metrics.

5. **Alerting and Report Delivery** – Proactive report delivery and alerting to very large populations based on schedules or event triggers in the database. Targeted at very large user populations of information consumers, both internal and external to the enterprise. *Report Distribution engines* were used to send full reports or alerts to large user populations based on subscriptions, schedules or threshold events in the databases.

More information about data processing tools mentioned above is available in Turban (2001) and Marakas (2003). In this textbook we are concentrating on and discussing analytical tools like statistical analysis and data mining.

Different styles of BI could be presented in a two-dimensional space (see Fig. 10-2) where the vertical axis represents the sophistication and interactivity of the analytical processes and the horizontal axis represents scale, or the size of the user population. The most sophisticated and interactive Styles of BI are used by relatively small groups of users consisting of information analysts and power users, for whom data and analysis are their primary jobs. Less interactive styles of BI deliver basic data and results that are applicable to very large user populations ranging from senior executives all the way to staff personnel. This common view can be changed radically with the new paradigm of ***Self-Organizing Data Mining*** approach.

One important view on BI platforms is presented in Gartner's research note. Year-to-year comparisons of vendor positions are not particularly useful given market dynamics (such as emerging competitors, new product road maps, new buying centers) and client concerns and inquiries change almost every year. Gartner Inc. evaluated vendors based on these new market dynamics and reflected the most recent changes in the so called Magic Quadrant criteria evaluation weights for 2021 (see Fig. 10-8).

BI platforms enable users to build applications that help organizations learn, understand, and optimize their business. Gartner defines[9] a BI platform as a software platform that delivers 13 capabilities, organized into three categories of functionality: *integration, information delivery* and *analysis* (one of these capabilities in the third category is *Predictive modeling and data mining*). In 2009, enhancing integration between BI platform components was a major

[9] The Gartner definition of "BI platform" has remained mostly consistent from previous years – in 2010 they added only one capability for search-based BI.

focus of mega vendors digesting their numerous acquisitions. Information delivery continues to be a core focus of most BI projects today, but we see an increasing demand for tools that enable easier and more intuitive analysis to discover new insights.

Gartner's view is that the market for BI platforms will remain one of the fastest growing software markets despite the economic downturn. In tough economic times, when competitiveness depends on the optimization of strategy and execution, organizations continue to turn to BI as a vital tool for smarter, more agile and efficient business. According to Gartner's annual survey of CIO technology priorities, BI remained among the top five priorities in 2009 (and it was No. 1 in each of the previous four years). That said, however, the recession, commoditization and consolidation are expected to reduce BI platform growth from more than 20% in 2008 to single digits in 2009 and beyond. The BI platform market's compound annual growth rate (CAGR) through 2013 was expected to be 6.3%, while the combined BI, analytics and performance management market's CAGR was expected to be 8.1% through 2013.

This is a global view of Gartner's opinion of the main software vendors that should be considered by organizations seeking to develop BI applications. As they recommended, buyers should evaluate vendors in all four quadrants, not assuming that only highly rated organizations can deliver successful BI implementations.

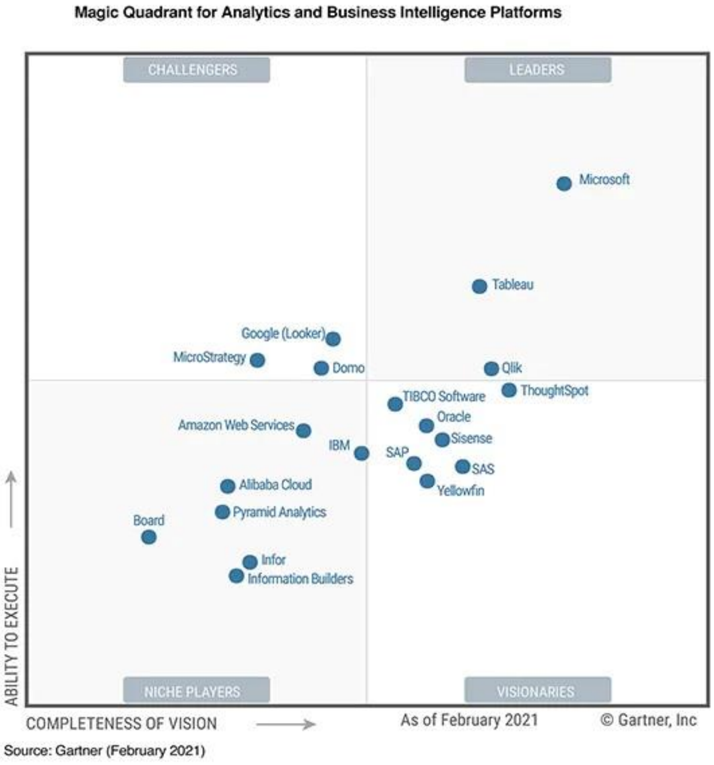


Fig.10-8 Magic Quadrant for Business Intelligence Platforms in 2021
(Source: Magic Quadrant for Business Intelligence Platforms, Gartner Inc., 2021)

10.5. Data Mining Integrated with Business Intelligence Applications

As mentioned earlier, data mining software assists and automates the process of building and training highly sophisticated data mining models. Once a data mining model has been created and tested, it is applied to a new dataset. The process of computing this predictive model to produce a final outcome is called "*scoring*" in BI and data mining. There are three main approaches to integrating predictive insight into a BI application. Scoring (calculating the predictions) can be performed by any one of the three components of enterprise BI applications (see Fig. 10-9) – data mining tool, database, or the BI application.

All three approaches are viable methods for deploying data mining results throughout the enterprise. Determining which approach to use depends greatly on the business need for predictive analysis, and the IT infrastructure and philosophy.

The starting point for most data mining implementations is to use the data mining tool for scoring. Although it is very common for the data mining analyst to provide scores in standalone flat files or spreadsheets, integrating scored results into databases has long been a common practice.

When scoring is required on a real time basis, or when predictive models are created, and changed faster than scores can be calculated and stored in the database, one of the other approaches must be adopted. If the database supports data mining, deploying models in the database is a possible next step. If the BI Platform contains data mining capabilities, deploying models directly in BI applications can speed the adoption of predictive analysis by business users.

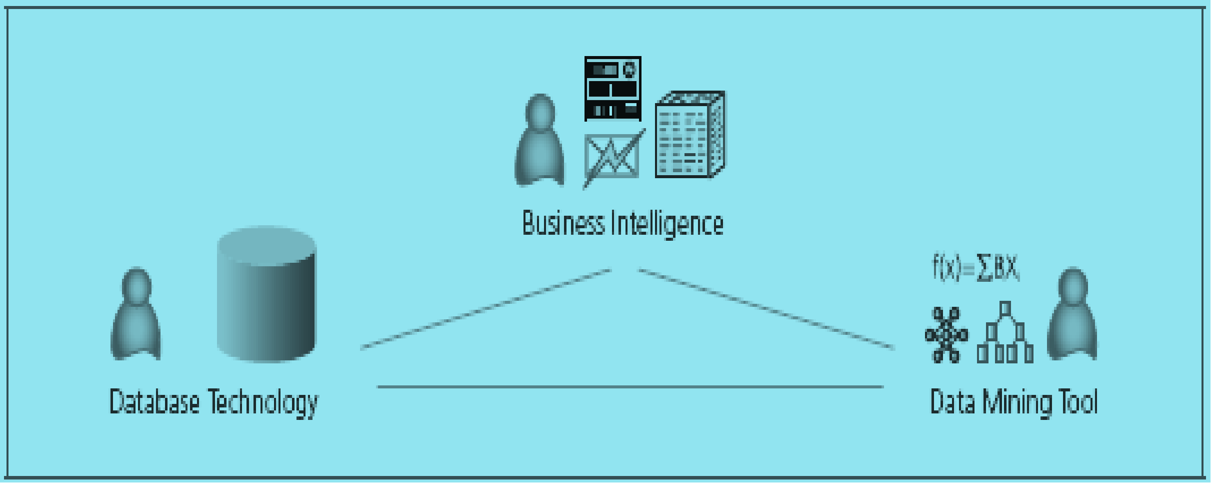


Fig. 10-9 Three main components of enterprise BI applications
(Source: MicroStrategy White paper, 2005, p.165)

With each approach, a different application is responsible for the model scoring. While each approach has its advantages and disadvantages, it is up to the analyst and IT administrator to determine which approach is most suitable for their environment and their use cases.

Some *early examples of data mining integrated with BI applications* are reported in the February 1996 edition of Forbes magazine (Novack, 1996):

Example 1. Marriott Club International, the nation's largest seller of vacation timeshare condos, slashed the amount of junk mail it has to send out to get a response. Before that the company problem was that sending advertisements to million names from their database at great expense, they received a minimal response. The company decided to identify the customers on their list who were most likely to respond using data mining to detect patterns by combing through the digitized customer files.

Marriott started with names, mostly of hotel guests. Digging into a trove of motor vehicle records, property records, warranty cards and lists of people who have bought by mail, or on the Web, computer software enriched the prospect list. It added such facts as the customers' ages, their children’s ages and estimated income, what cars they drive and whether they golf. Then Marriott system used a neural network to figure out who is most likely to respond to a mailed flier.

Using these clues, Marriott was able to cast its net a little more narrowly and catch more fish. Data mining has increased the response rate to Marriott's direct mail time-share pitches to certain hotel guests. In addition, the company reported significant savings on their mail costs.

Example 2. First Commerce Corp. of Louisiana used to plumb its database for insights through a series of queries. "I would ask how many people have installment loans and what's the combination of other services they have," recalls the company Senior Vice President. "Then I'd ask about their ages and demographics." From there on, it was all a hunch.

In 1993 the company bought the same HNC Software neural network system used by Marriott Club International. First Commerce fed data on 2,000 current and recently departed customers into a workstation. The net tested 70 variables in numerous combinations and constructed a model of likely leavers. The bank then used the model to select high-balance customers it might lose. It offered them a new money market account to entice them to stay.

Neural nets are just one of a flood of new tools, from Arbor Software, Oracle, SAS Institute and others, designed to help human managers extract insights from masses of data. The beauty of a neural net is that it will test dozens of variables that humans, with their preconceptions and time limits, won't test. The net's drawback is that it doesn't spell out just what characteristics make the individuals score as hot prospects for a product.

Another example could be found in the white paper IBM provides the foundation for supply chain management by Arvato (2010). It is interesting because Cognos (Cognos Incorporated) was an Ottawa, Ontario-based company making business intelligence and performance management software. Founded in 1969, at its peak Cognos employed almost 3,500 people and served more than 23,000 customers in over 135 countries until being acquired by IBM[10] on January 31, 2008.

> "Our leading position in the logistics services market is based to a great extent on the reporting support provided by IBM Cognos solutions." says Jochen Bremshey, Vice President IS & T Entertainment Services, Arvato services.

Arvato services, a subsidiary of *arvato Bertelsmann AG*, handles all inventory management flows on behalf of its customers. With around 5,000 employees, where required, Arvato logistics specialists can cover a company's entire supply chain all over the world. Its service spectrum encompasses order processing, procurement, warehousing, distribution, communication and financial aspects. Companies ranging from medium-sized firms to large global groups call on Arvato's logistics services. Located at various sites throughout Europe, the Entertainment Services division manages the distribution of entertainment media, mainly audio CDs, DVDs and video games. Major players such as Sony Music, Warner Music and Paramount rely on Arvato systems and services to ensure that their products are available in sales outlets in line with demand. Arvato departments ensure that supplies of the required entertainment media reach these outlets – including 300 Media Market, 185 Saturn, 300 Carrefour, 370 ASDA, 70 Kaufhof and more than 200 Karstadt stores, as well as over 3,700 Esso filling stations in Western and Southern Europe – in good time and according to their individual requirements.

The entertainment media market is extremely fast moving. Today's tops are tomorrow's flops, while unknowns are suddenly thrust into the limelight. At the same time, customers want everything immediately, even items from the back catalogue. This represents a complex challenge for supply chain management.

The key to addressing both requirements is a very responsive supply chain system that is capable of capturing all movement of goods while automatically ensuring appropriate replenishments according to predefined parameters. An overview of all products and sales must always be available to identify any peaks or falls in demand as they occur and take appropriate countermeasures. Every day, Arvato receives around 3.6 million electronic data records from

[10] Riley, D. (2007, Nov 12). *Acquisitions: IBM Buys Cognos, Microsoft Buys Musiwave*. Techcrunch. https://techcrunch.com/2007/11/12/acquisitions-ibm-buys-cognos-microsoft-buys-musiwave/

different source systems. This information needs to be standardized and automatically transmitted to the replenishment scheduling systems. Product and supply chain managers also need an accurate database to support their tactical and strategic analyses and decision-making.

Meeting the Challenge of Retail Inventory Management, in cooperation with IBM Cognos, Arvato services has developed a supply chain management solution called R.I.M. (Retail Inventory Management), which offers automated management of entertainment product supplies, as well as comprehensive reporting and a range of flexible analyses. In 2001 a Vendor Managed Inventory (VMI) system was developed, which has continuously been upgraded and is currently running based on IBM Cognos 8 BI platform. It provides decision-makers at Arvato with information consolidated across different source systems, as well as a wide range of analysis and evaluation tools.

The R.I.M. system responds rapidly, automatically and according to demand, managing replenishments based on sales variations in stores. Top titles are identified and sufficient stocks ensured to meet demand, while a lower priority is placed on slower moving titles. Since R.I.M. also considers financial parameters, the system can ensure the retail outlet has not exceeded its credit limit. Supply chain planning encompasses factors, such as marketing campaigns that may affect additional sales areas, or the expected lifecycle of a product. Digital capture of all goods and networking of all systems involved ensures end-to-end monitoring of the supply chain.

Another R.I.M. component offers Arvato customers a range of flexible analysis options. Multidimensional OLAP data cubes are generated based on PoS data and reports are created using IBM Cognos 8 BI. Variable statistics and ad-hoc analyses are used to map demand and inventory planning trends at sales outlet level, by region, division, artist, or according to any required keyword. Different presentation options such as graphs and diagrams ensure clear and transparent results, even for inexperienced users. This ensures that product and supply chain managers have access to the required information to support their tactical and strategic decision-making, answering questions like “Were sales of Christmas music higher than expected in December?” and respond appropriately.

IBM Cognos technology and expertise are creating a comprehensive approach to supply chain management, in line with the IBM vision of smart planning: completely digitized, networked from end to end, and controlled intelligently. This results in a win-win situation. Manufacturers and sales outlets benefit from automated, targeted, and therefore more cost-effective supplies with low return volumes, as well as from well-founded and responsive planning of sales assortments. Customers can find exactly what they are looking for at their sales outlet, guaranteeing good entertainment for all.

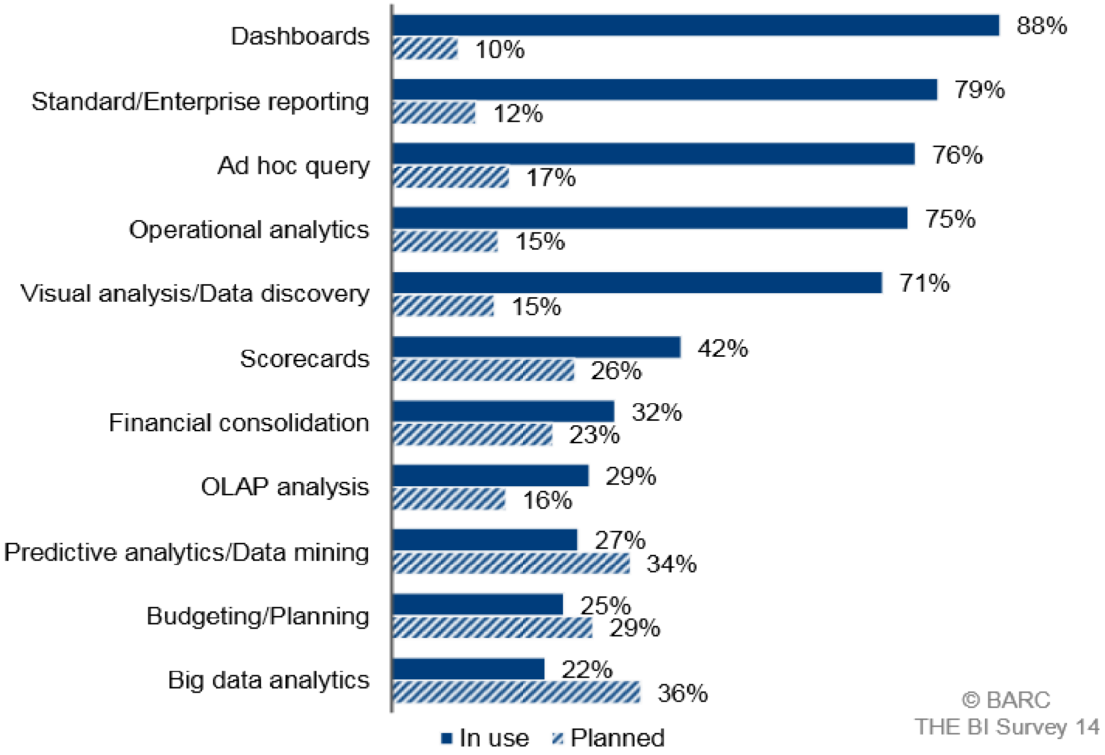
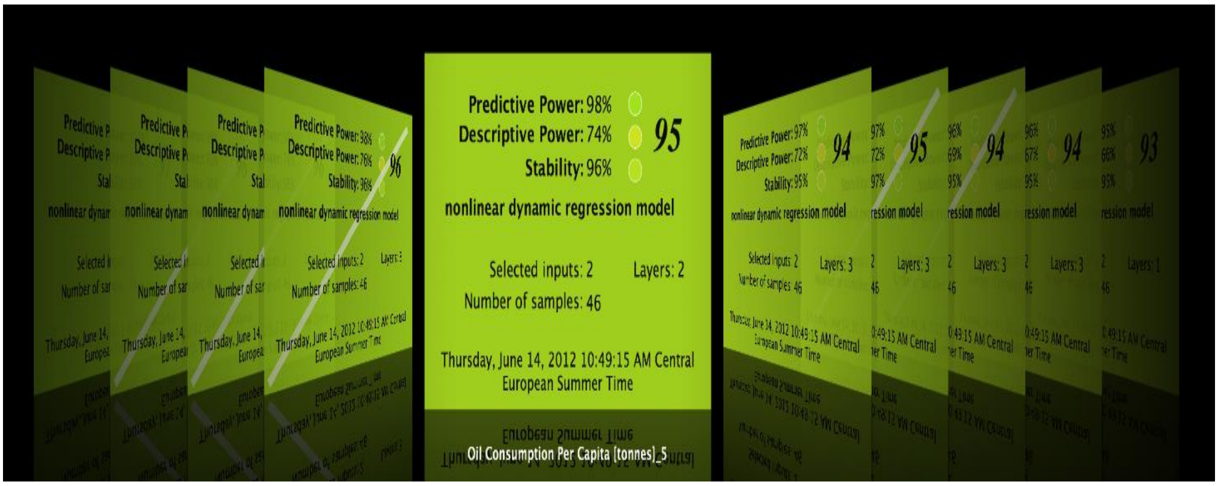


Fig. 10-10 Three main components of enterprise BI applications
(Source: The BI Survey 14, 2015, p.6)

As one of the largest recent surveys of BI applications, The BI Survey 14 (2015) shows today more and more companies are willing to use analytics platforms, such as Predictive analytics, Big Data analytics and Data Mining. As we can see in Fig.10-10, the fact that companies are just planning to use these, rather than they actually use them, in our opinion, is mainly because of lack of appropriate computing platforms.

Historically, using predictive analytics tools, as well as understanding the results they delivered, required advanced skills. However, modern predictive analytics tools are no longer restricted to specialists. As more organizations adopt predictive analytics into decision-making processes and integrate it into their operations, they are creating a shift in the market toward business users as the primary consumers of the information.

Business users want tools they can use on their own. Vendors are responding by creating software that removes the mathematical complexity, provides user-friendly graphic interfaces, and/or builds in short cuts that can, for example, recognize the kind of data available and suggest an appropriate predictive model. Predictive analytics tools have become sophisticated enough to adequately present and dissect data problems, so that any data-savvy information worker can utilize them to analyze data and retrieve meaningful, useful results. For example, tools like ***KnowledgeMiner*** software can present findings (see Fig. 10-11) using simple charts, graphs, and scores that indicate the likelihood and/or the level of possible outcomes.



Predictive Power: 98%
Descriptive Power: 74%
Stability: 96%
95
nonlinear dynamic regression model

Selected inputs: 2 Layers: 2
Number of samples: 46

Thursday, June 14, 2012 10:49:15 AM Central European Summer Time

Oil Consumption Per Capita [tonnes]_5

Fig. 10-11 Example of KnowledgeMiner software user-friendly interactive output
(Source: http://www.knowledgeminer.eu)

Such platforms provide opportunities for shortening the time, reducing the cost and the efforts in predictive model building, as well as for increasing model accuracy (Mueller & Lemke, 2003). This assists decision makers in analyzing the problem more precisely, in deeper and better understanding of the problem, generating better predictions and eventually making better decisions. Similar platforms, successfully used in business forecasting, and some real-life applications will be discussed in detail in Chapter 12.

***

SUMMARY AND CONCLUSIONS

Chapter 10 introduces the most advanced techniques that could be used in Business Forecasting:

- ***Business intelligence (BI)*** *is a broad category of applications and technologies for gathering, storing, analyzing, and providing access to data to help enterprise users make better business decisions. BI applications include the activities of decision support systems, query and reporting, online analytical processing (OLAP), statistical analysis, forecasting, and data mining.*
- ***Business Analytics*** *(**BA**)* should be an element of ***BI***, however, there are different opinions and usually ***BA*** refers to the skills, technologies, applications and practices for continuous iterative exploration and investigation of past business performance to gain insight and drive business planning.
- ***Predictive analytics*** is an important part of ***BA*** which encompasses a variety of techniques from statistics, data mining and game theory that analyze current and historical facts to make predictions about future events. In business, *predictive models* exploit patterns found in historical and transactional data to identify risks and opportunities.

In this book we assume that *Predictive analytics* is an area of analysis that deals with extracting information from data and uses it to predict future trends and behavior patterns. Traditional statistical analysis was already discussed in previous chapters and Chapter 10 concentrates on *data mining*:

- *Data mining is the process of exploration and analysis* (by automatic or semi-automatic means) *of large quantities of data in order to discover meaningful patterns and rules, i.e. process of extracting meaningful patterns from large amounts of data.*
- *Knowledge Discovery in Databases* refers to the overall process of discovering useful knowledge from data, and *data mining* refers to a particular step in this process, i.e. the application of specific algorithms for extracting patterns from data.
- *Data mining* involves the following activities in extracting meaningful new information from the data: *Classification, Estimation, Prediction, Affinity grouping or association rules, Clustering, Description and visualization.*

- *Directed data mining* is a top-down approach, used when we know what we are looking for. This often takes the form of predictive modeling where we know exactly what we want to predict.
- *Undirected data mining* is a bottom-up approach that finds patterns in the data and leaves it up to the user to determine whether or not these patterns are important, i.e. it is about discovering new patterns inside the data.

These two approaches are not mutually exclusive. Data mining efforts often include a combination of both. Even when building a predictive model, it is useful to search for patterns in the data using undirected techniques. These can suggest new insights that can improve the directed modeling results.

- *Data Mining* is the process of examining large amounts of data in search of hidden patterns and predictive information (mostly in an automated manner), which allows organizations to make better decisions. There are a few general steps:

1. Create a predictive model from a data sample
2. Train the model against datasets with known results
3. Apply the model against a new dataset with an unknown outcome

The "Create-Train-Apply" process is typically the domain of statisticians or data mining analysts. A solid understanding of data mining concepts, statistical concepts, techniques, and data mining tools is necessary in the "Create" and "Train" steps. Applying the predictive model requires less expertise and is available for all business users.

In Chapter 11 we are going to discuss techniques that are related to forecasting with enough details so the reader can:

- Distinguish between different techniques knowing their advantages and disadvantages.
- Follow the techniques as they are used in the real life business examples.
- Understand which technique is most appropriate for a given business problem.
- Become familiar with important variations.

In Chapter 11, some major techniques that are found in most comprehensive data mining tools such as *decision trees, neural networks* and *clustering* will be discussed. Another approach known as *Self-Organizing Data mining* based on *Group Method of Data Handling* will also be talked over. It shows great promises as already pointed out in Chapters 8 and 9. In Chapter 12, a few unique *Self-Organizing Data mining* techniques in business forecasting will be studied in detail.

KEY TERMS

Ad Hoc Query and Analysis	***361***
Alerting and Report Delivery	***362***
Business Analytics (BA)	***347***
Business Intelligence (BI)	***345, 346***
Business Intelligence 2.0 (BI 2.0)	***346***
Business intelligence software	***346***
Business operations reporting	***360***
Competitive Intelligence	***345***
Computer software solutions	***346***
Create-Train-Apply	***356***
Cube Analysis	***361***
Dashboard	***360***
Data mining	***350, 352***
Data warehousing	***346***
Decision Support System	***345***
Directed data mining	***353***
Enterprise Reporting	***361***
Knowledge Discovery in Databases (KDD)	***352***
KnowledgeMiner	***368***
Multidimensional analysis	***360***
OLAP	***360***
Predictive analytics	***348***
Self-Organizing Data Mining	***362***
Statistical Analysis	***362***
Undirected data mining	***354***

CHAPTER EXERCISES

Conceptual Questions:

1. Define ***Business intelligence (BI)*** and ***Business Analytics*** *(**BA**)*. What are the differences between these two areas of research?
2. What is ***Predictive analytics?*** How does it related to ***BI*** and ***BA***?
3. What is ***Data Mining***? What is the general purpose of ***Data Mining*** tools?
4. Is ***Data Mining*** a synonym of ***Knowledge Discovery from Data/Databases?*** List all major activities in extracting meaningful information from data – explain at least three of them.
5. What are the general steps in ***Data Mining***? List and discuss briefly the "Create-Train-Apply" process.

Business Applications:

Open Gretl program and the file Sales Data from the previous Chapter 9:

- Set up time series data for an ***ARMAX*** model with dependent variable "Sales" and the given predictors adding time lags for both dependent and independent variables accordingly.
- Test the aptness of the model using test statistics provided by Gretl software and perform residual analysis. Do you see any violations of the regression assumptions? If yes return to the previous step and improve the model.
- Is it possible to compute Sales forecast for the next 12 months? What additional data are necessary to perform these calculations?
- Given the expected values for the predictors in spreadsheet Predictors compute Sales forecasts for the next 12 months.
- Using the formulas developed in *Part 9* (**Complex Models and Forecasting – I**) compute MAD, MSE, MAPE and MPE for the new model, for a testing dataset of the 12 new monthly forecasts given in spreadsheet Predictions.
- Compare the forecasting errors for the previous ***AR*** and the current ***ARMAX*** model.
- What model provide better accuracy? Are there any reasons leading to this conclusion?

Discuss all findings and write a short report (up to two pages) summarizing your answers.

INTEGRATIVE CASE

HEALTHY FOOD SYPPLY CHAIN & STORES

Part 10: Complex Models and Forecasting – 2

In Chapter 1 we introduced *Healthy Food Stores* – a fast-growing retail food provider with 12 stores in a northwestern state. The company executives decided to study the effect that company advertising dollars have on sales. They hoped that examining collected historical data would reveal relationships that would help determine future advertising expenditures and predict monthly sales volumes for the upcoming quarter.

After identifying basic parameters, along with input (independent) and output (dependent) variables of the forecasting scenario in Part 2, the research team determined the main elements of the forecasting process:

- Forecasting horizon of up to twelve months;
- Quarterly forecast updates, since accuracy decreases as time horizon increases, and sufficient time is needed to implement possible changes;
- Development of different models based on data patterns, if any, and evaluation of their accuracy in order to select the most appropriate one;
- Selection of the best forecasting model with no more than 5% forecasting error.

The benchmark forecast computed in Part 3 (using the baseline of ***one-step naive forecast*** as a reference forecast) provided basic values for the most common measures of accuracy MFE, MAD, MAPE, MPE, MSE and CV(RMSE).

In Part 4, information about the opinion of some important people from the *Healthy Food Stores Company*, concerning this specific case, was collected and the research team applied the Delphi method to top executives group, Sales-force composite to the sales managers from all 12 stores and Scenario writing to the most experienced professionals from Advertising Department. After collecting such valuable information from different sources, in Part 5 the research team made its first steps in Numerical Predictions by developing different basic forecasting models. They created spreadsheets for Naïve techniques (Average model, Random Walk with Drift and Seasonal Naïve Technique), simple Moving Average, Simple Exponential Smoothing and Triple Exponential Smoothing, used to expand the base-line of one-step naïve forecast as reference forecasts.

In Part 6 the research team analyzed the relationships between dependent variable Sales and the available predictors. After performing multiple correlation and regression analysis, researchers developed reliable forecasting model, which passed all tests and hypotheses,

representing the real system with certain error. In Part 7, the model was expanded by adding Dummy seasonal variables to analyze the Seasonal effect in company Sales. In Part 8, the improvement of the forecasting model continued (with the help of some advanced Time series analyses and predictive techniques) and few simple ***AR*** models were build using ***ARIMA*** methodology and *Gretl* software.

The next step of the forecasting process was to develop models using complex techniques. In *Part 9* new ***AR*** models were build using ***ARIMA*** methodology and *Gretl* software, which were analyzed and compared with the previous forecasting models. In the current, second part of this complex forecasting step different ***ARMAX*** models would be developed.

Case Questions

1. Open Gretl program and import file Data containing the results from *Part 9* (**Complex Models and Forecasting – 1**).
2. Set up time series data for an ***ARMAX*** model with dependent variable "Sales" and all basic predictors as exogenous variables adding time lags for both dependent and independent variables:
   a) Split the sample into Training (36 observations) and Testing (12 observations) data sets.
   b) Run Multiple Regression analysis and Conduct a test of hypothesis on each of the predictor variables. Eliminate any insignificant predictor at 0.05 significance level.
   c) Rerun the multiple regression equation and test the aptness of the model. Continue eliminating any insignificant regressor at 0.05 level of significance until only important (significant) exogenous variables remain in the model.
   d) Under what conditions it is possible to compute Sales forecast for the next 12 months? What additional assumptions are necessary to perform these calculations?
   e) Compute Sales forecast for the next 12 months using the Testing data set.
3. Use (copy/paste) the formulas designed in Part 3 to compute MFE, MAD, MAPE, MPE, MSE and CV(RMSE) for the new model, for the given testing dataset of 12 monthly forecasts provided in spreadsheet Errors.
4. Comment and analyze model's accuracy - how good is the accuracy of these forecasts? What model, out of all models so far provides the best accuracy? Discuss.
5. What overall recommendations would you make to the research team? Explain.
6. Write a report on the questions above, discussing all important findings and draw relevant conclusions about this part of the Integrative Case.

References

An Architecture for Enterprise Business Intelligence. (2005). *A White Paper*, MicroStrategy. http://www.microstrategy.com/Publications/Whitepapers

Beller, M., & Barnett, A. (2009). *Next Generation Business Analytics Technology Trends*. Lightship Partners LLC. http://www.docstoc.com/docs/7486045/Next-Generation-Business-Analytics-Technology-Trends

Berry, M., & Linoff, G. (2000). *Mastering Data Mining.* Wiley.

Davenport, T., & Harris, J. (2007). *Competing on Analytics: The New Science of Winning*. Harvard Business School Press.

Fayyad, U., Piatetsky-Shapiro, Gr., & Smyth, P. (1996, Fall). From Data Mining to Knowledge Discovery in Databases, *American Association for Artificial Intelligence Magazine*, 37-54. http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf

Fleisher, C., & Blenkhorn, D. (2003). *Controversies in Competitive Intelligence: The Enduring Issues*, Westport, CT: Praeger.

Henk G. et al. (1987). Expert systems and artificial intelligence in decision support systems, *Proceedings of the Second Mini Euroconference (The Netherlands, 1985),* 1-2. Springer.

Gartner Magic Quadrant for Analytics and Business Intelligence Platforms. (February, 2021). Gartner Inc. https://www.gartner.com/en/documents/3996944

IBM provides the foundation for supply chain management by Arvato. (2010, January). IBM. ftp://ftp.software.ibm.com/common/ssi/pm/ab/n/imc14491gben/IMC14491GBEN.PDF

Lucey, T. (1991). *Management Information Systems* (6th ed.). DP Publications Lim.

Luhn, H. (1958). A Business Intelligence System. *IBM Technical Journals*, 2 (4), 314.

Magic Quadrant for Business Intelligence Platforms. (2020). *Gartner RAS Core Research Note G00173700*. https://www.sisense.com/gartner-magic-quadrant-business-intelligence/

Marakas, G. (2003). *Decision Support Systems in the 21st Century*. Prentice-Hall.

Mueller J. A., & Lemke, F. (2003). *Self-Organizing Data Mining: An Intelligent Approach To Extract Knowledge From Data.* Victoria, BC: Trafford Publishing.

Novack, J. (1996, February 12). The data miners. *Forbes*, 157(3), 96-97.

Power, D. (ed.). (2007). *A Brief History of Decision Support Systems*. (ver. 4.0) *Decision Support Systems Resources*. http://dssresources.com/history/dsshistory.html

Reference Configurations. (2007). *A White Paper*, SAS® ANALYTICS, Sun Microsystems Inc. http://www.sas.com/partners/directory/sun/SAS-RC-AI-1107.pdf

Shmueli, G. et al. (2007). Predictive vs. Explanatory Modeling in IS Research, *Proceedings of the Conference on Information Systems and Technology*, Seattle, WA, USA. http://www.citi.uconn.edu/cist07/5c.pdf

The 5 Styles of Business Intelligence: INDUSTRIAL-STRENGTH BUSINESS INTELLIGENCE. (2002). *A White Paper*, MicroStrategy. http://www.microstrategy.com/Publications/Whitepapers

The BI Survey 14 by BARC. (2015). *A White Paper*, Bitpipe at TechTarget. http://docs.media.bitpipe.com/io_12x/io_121843/item_1116964/The%20BARC%20BI%20Survey%202014.pdf

Turban, E., & Aronson, J. (2001). *Decision Support Systems and Intelligent Systems.* Prentice-Hall.

***

Chapter 11. Business Forecasting and Data Mining

11.1. Knowledge Discovery from Data

At the end of the 20th century Fayyad, Piatetsky-Shapiro and Smyth (1996) made the statement: "*Across a wide variety of fields, data are being collected and accumulated at a dramatic pace*" (p. 37). Many more researchers and scientists (Devlin, 1997, Mueller & Lemke, 2003, and others) agreed with the statement about the ever-increasing volume of data that require processing. Today, there is a significant need to discover hidden and valuable information from massive amounts of data for decision making. This information includes, for example, new key facts, general tendencies and relationships, significant and useful patterns of information, etc. To extract this information a new generation of techniques and tools, which are able to help humans intelligently and automatically to analyze both large and small data sets, is required.

The traditional method of turning data into knowledge relies on manual analysis and interpretation. For example, in the health-care industry it is common for specialists to periodically analyze current trends and changes in health-care data on a quarterly basis. Specialists then provide a report detailing the analysis to the sponsoring health-care organization, which becomes the basis for future decision making and planning for health-care management. Whether in science, marketing, finance, health care, retail, or other fields, the classical approach to data analysis relies fundamentally on one or more analysts becoming intimately familiar with the data and serving as an interface between the data and end users.

For these (and many other) cases, this form of manual probing of a data set is slow, expensive, and highly subjective. In fact, as data volumes grow dramatically, this type of manual data analysis is becoming completely impractical in many domains. Databases are increasing in size in two ways: first, the number N of records or objects in the database and second, the number of fields or attributes to an object. Nowadays data warehouses containing about $N = 10^9$ or more objects are becoming increasingly common and it is not only in the astronomical sciences.

Who could be expected to digest millions of records, each having tens or hundreds of fields? We believe that this job is certainly not one for humans; hence, analysis work needs to be automated, at least partially. The need to scale up human analysis capabilities to handle the large number of bytes, which we can collect, is both economic and scientific. Businesses use data to gain competitive advantage, increase efficiency, and provide more valuable services to customers. Data we capture about our environment are the basic evidence we use to build theories and models of the universe we live in. Since computers have enabled humans to gather more data than we can digest, it is natural to turn to computational techniques to help us reveal meaningful patterns and structures from the massive volumes of data.

These techniques and tools are the subject of data mining, and within knowledge discovery they have to turn information located in the data into valuable information for successful decision making. ***Knowledge discovery from data* (*KDD*)** is an interactive and iterative process of solving several major subtasks and decisions like data selection and preprocessing, choice and application of data mining algorithms, and analysis of the extracted knowledge. ***Data mining techniques*** in ***KDD*** help researchers in analyzing the massive amounts of data and turning information located in the data into successful decisions.

Knowledge discovery has evolved, and continues to evolve, from the intersection of research fields such as machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, data visualization, and high-performance computing. The unifying goal is extracting high-level knowledge from low-level data in the context of large data sets.

In *Business Forecasting*, the ***KDD*** helps researchers to develop good and reliable models. As a matter of fact, ***KDD*** process has a wide range of applications and business forecasting is just one of them.

Frawley, Piatetsky-Shapiro and Matheus (1992) presented a framework for knowledge discovery (see Fig.11-1) and Brachman and Anand (1996) gave a practical view of the knowledge discovery process, emphasizing the interactive and iterative nature of the process. Here, we broadly outline some of its basic steps (see also Fig. 10-3):

- First is *developing an understanding of the application domain* and the relevant prior knowledge and identifying the goal of the ***KDD*** process from the customer’s viewpoint.
- Second is *creating a target data set*: selecting a data set, or focusing on a subset of variables or data samples, on which discovery is to be performed.
- Third is *data cleaning and preprocessing*. Basic operations include removing noise if appropriate, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, and accounting for time-sequence information and known changes.
- Fourth is *data reduction and projection*: finding useful features to represent the data depending on the goal of the task. With dimensionality reduction or transformation methods, the effective number of variables under consideration can be reduced, or invariant representations for the data can be found.

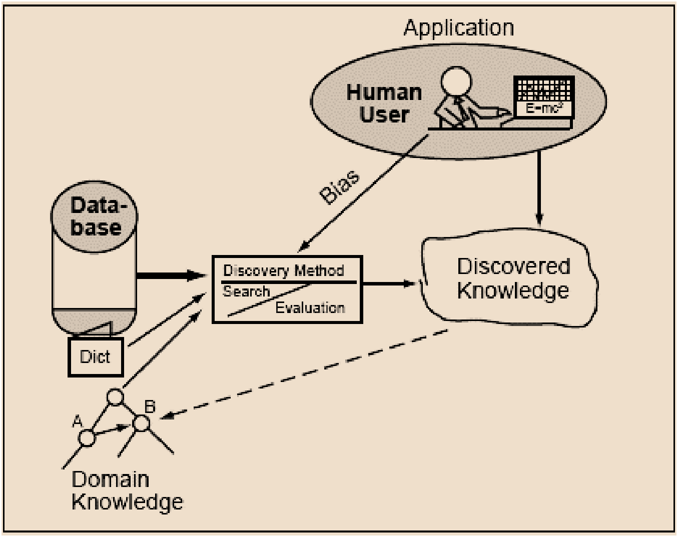


Fig.11-1 A Framework for Knowledge Discovery in Databases (***KDD***)
(Source: Frawley et al.,1992, p.61)

- Fifth is *matching the goals* of the ***KDD*** process (step 1) to a particular data-mining method. For example, summarization, classification, regression, clustering, and so on.
- Sixth is *exploratory analysis and model and hypothesis selection*: choosing the data mining algorithm(s) and selecting method(s) to be used for searching for data patterns. This process includes deciding which models and parameters might be appropriate (for example, models of categorical data are different than models of vectors over the reals) and matching a particular data-mining method with the overall criteria of the ***KDD*** process (for example, the end user might be more interested in understanding the model than its predictive capabilities).
- Seventh is *data mining*: searching for patterns of interest in a particular representational form or a set of such representations, including classification rules or trees, regression, and clustering. The user can significantly aid the data-mining method by correctly performing the preceding steps.
- Eighth is *interpreting mined patterns*, possibly returning to any of steps 1 through 7 for further iteration. This step can also involve visualization of the extracted patterns and models or visualization of the data, given the extracted models.
- Ninth is *acting on the discovered knowledge*: using the knowledge directly, incorporating the knowledge into another system for further action, or simply documenting it and reporting it to interested parties.

Data mining (DM) itself includes not only a straightforward utilization of a single analytical technique but also consists of processes which are appropriate for many methods and techniques depending on the nature of the inquiry. This set of methods contains data visualization, tree-based models, neural networks, methods of mathematical statistics (like clustering and discriminant analysis, regression and correlation analysis), and methods of artificial intelligence. In the spectrum of modeling methods are also included methods of knowledge extraction from data using self-organizing modeling technology which will be discussed in Chapter 12.

This process includes as well checking for and resolving potential conflicts with previously believed (or extracted) knowledge. The ***KDD*** process can involve significant iteration and can contain loops between any two steps. The basic flow of steps, without the potential multitude of iterations and loops, is illustrated in Fig.11-1. Most previous work on ***KDD*** has focused on step 7, the data mining process. However, the other steps are as important for the successful application of ***KDD*** in practice.

Having defined the basic notions and introducing the ***KDD*** process, we now focus on the data-mining component, which has, by far, received the most attention in the literature. *Data mining techniques* are like anything else a computer does, such as storing files or creating a spreadsheet. The ***techniques*** are general approaches to solving problems, and there are usually many ways to approach the technique. Each of these ways is a different algorithm. The ***algorithms*** are like recipes with step-by-step instructions explaining what is happening.

The knowledge and skills in statistics accumulated so far in this book will work as a background when we go into detail of the data mining techniques and to explain how they actually work. It is also important to have some understanding of their inner workings, to know when to apply them, how to interpret the results, and whether or not they are working. The purpose of the following two chapters is to explain the techniques with enough detail so that one can:

- Distinguish between different techniques, knowing their advantages and disadvantages;
- Follow the techniques as they are used in the real life business examples;
- Understand which technique is most appropriate for a given business problem;
- Become familiar with important variations.

The major techniques we are going to discuss first are the ones that are found in most comprehensive data mining tools: *clustering, decision trees* and *neural networks*. The emphasis here is on ***Self-Organizing Data Mining*** using the ***Group Method of Data Handling***. This unique approach already introduced in the previous chapters will be discussed in detail in Chapter 12 and original ***MLNAN*** algorithms will be used as examples in business forecasting.

These techniques are also available on a wide range of computing platforms (see the next topic), from individual desktops, to departmental servers, to the most powerful super computers with parallel data processing. However, as desktops are becoming more powerful, it is often not necessary to purchase expensive hardware to run data mining algorithms.

The techniques mentioned above represent the most commonly used groups of algorithms in real-life applications. Undoubtedly, each technique has many different algorithms and implementations – in fact, almost every tool has some nuance that makes its implementation a little different from the next tool. In spite of this, just as it is possible to learn how to drive a car, and generalize it to any car, it is possible to learn how to use neural networks or decisions trees, and generalize them to any tool.

It is true that data mining has a broad reach, but it is not possible to cover every algorithm used in the business world. Additionally, the minor variations have much less effect on data mining results than other issues, such as preparing the data and building the right models.

11.2. Data Mining Techniques

Because ***data mining (DM)*** is viewed as a technical subject, people often get the notion that mastering data mining is largely a matter of studying advanced algorithms and learning the techniques for applying them. This technical understanding is actually only one small component of the mastery one seeks. It is, however, a very important one! Without at least an understanding of the most important ***DM*** algorithms, users will not be able to understand when one technique is called for and when another would be more suitable. Users also need to understand what is going on inside a model in order to understand how best to prepare the set used to build it and how to use various model parameters to improve results.

Usually, the level of understanding needed to make good use of ***DM*** algorithms does not require detailed study of machine learning or statistics. In fact, only a basic understanding of the principal algorithms is essential for anyone wishing to master the art of ***DM***. First of all, each user should distinguish between ***DM*** techniques and the algorithms used to implement them. The term *technique* refers to a conceptual approach to extracting information from data. An *algorithm* contains the step-by-step details of a particular way of implementing a technique. For example, automatic cluster detection is a technique that can be implemented using self-organizing maps, simple ***k-means***, Gaussian ***k-means***, and a number of other algorithms.

Data mining is a method of searching data for unexpected patterns or relationships using a variety of tools and algorithms. Fayyad et al. (1996) identified six tasks as follow:

- *classification*: learning a function that maps (classifies) a data item into one of several predefined classes;
- *regression*: learning a function that maps a data item into a real-valued prediction variable;
- *clustering*: identifying a finite set of categories or clusters to describe the data;
- *summarization*: finding a compact description for a subset of data;
- *dependency modeling*: finding a model that describes significant dependencies between variables;
- *change and deviation detection*: discovering the most significant changes in the data from previously measured or normative values.

Data Mining uses database technologies, modeling techniques, statistical analysis and machine learning to find hidden patterns of relationships, to generate forecasting models based on actual historical data and to make predictions. The purpose of data mining platforms is to assist and automate the process of building and training highly sophisticated models, and apply these models to make predictions and perform what-if analysis and simulations, which support and help make better decisions.

Each data mining technique has many different algorithms and implementations – in fact, almost every tool has some nuance that makes its implementation a little different from the next tool. In spite of this, just as it is possible to learn how to drive a car, and generalize it to any car, it is possible to learn how to use neural networks or decisions trees, and generalize to any tool. It is also true that data mining has a broad reach, but it is not possible in one book to cover every algorithm used in the business world. In addition, as many authors mention, the minor variations have much less effect on data mining results than other issues, such as preparing the data and building the right models.

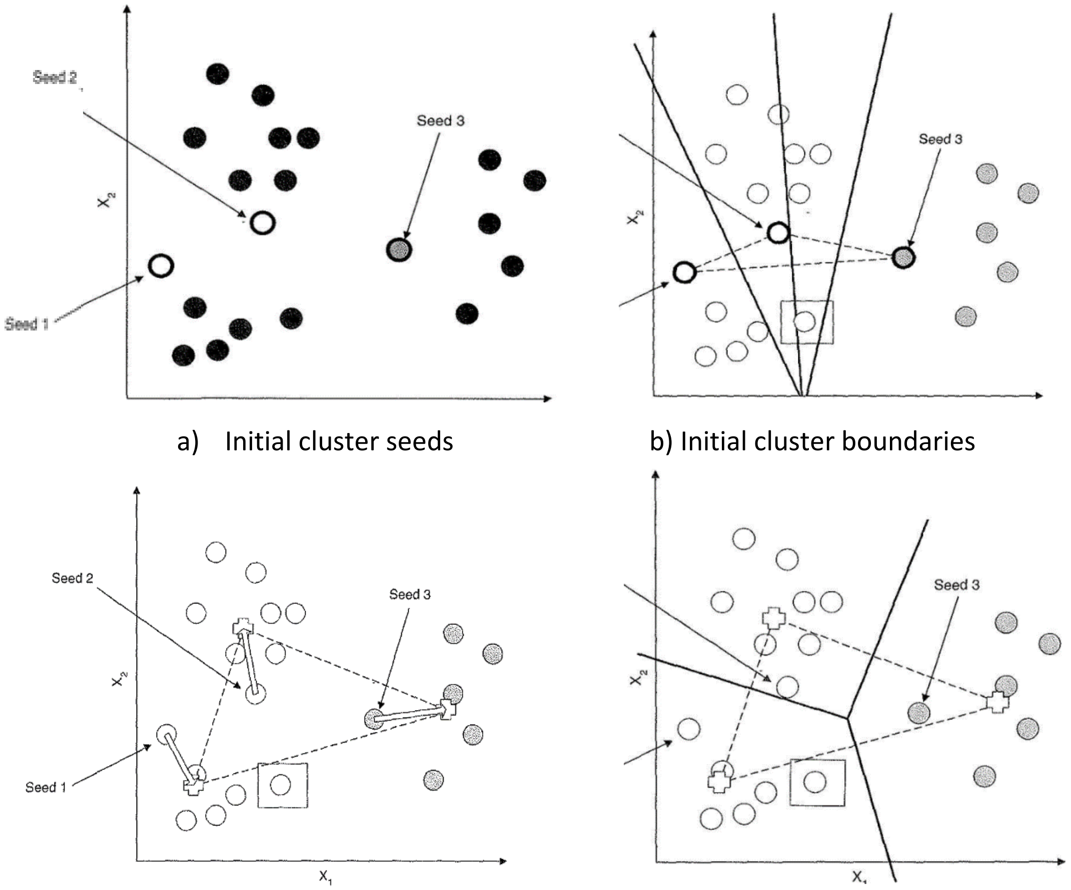
The major groups of techniques that are found in most comprehensive data mining tools (see Fayyad et al., 1996) are ***clustering, decision trees*** and ***artificial neural networks (ANNs)***.

A. Clustering

There are many mathematical approaches to find clusters in data (Fig.11-2), and whole text books are devoted to the subject. Some methods, called *divisive methods*, start by considering all records to be part of one big cluster. That cluster is then split into two or more smaller clusters, which are themselves split until eventually each record has a cluster all to itself. At each step in the process, some measure of the value of the splits is recorded so that the best set of clusters can be chosen in the end. Other methods, called *agglomerative method*s, start with each record occupying a separate cluster, and iteratively combine clusters until there is one big one containing all the records. There are also *self-organizing maps*, a specialized form of neural network that can be used for cluster detection and others (Madala & Ivakhnenko, 1994).

For example, ***k-means*** is a clustering algorithm, which is available in a wide variety of commercial ***DM*** tools and is more easily explained than most of them. It works best when the input data is primarily numeric. Consider an analysis of supermarket shopping behavior based on loyalty card data. Simply take each customer and create a field for the total amount purchased in various departments in the supermarket over the course of some period of time – diary, meat, cereal, fresh produce, and so on. This data is all numeric, so ***k-means*** *clustering* can work with it quite easily and the algorithm will find clusters of customers with similar purchasing patterns.

This algorithm divides a data set into a predetermined number of clusters. That number is the "***k***" in the phrase ***k-means***. The mean is an average value, which in this case refers to the average location of all of the members of a particular cluster. But what does it mean to say that cluster members have a location when they are records from a database?



a) Initial cluster seeds b) Initial cluster boundaries

c) After one iteration d) New cluster assignments

Fig.11-2 Example of Automatic Cluster Detection

The answer comes from geometry. To form clusters, each record is mapped to a point in "record space." The space has as many dimensions as there are fields in the records. The value of each field is interpreted as a distance from the origin along the corresponding axis of the space.

In order this geometric interpretation to be useful, the fields must all be converted into numbers and the numbers must be normalized so that a change in one dimension is comparable to a change in another. Records are assigned to clusters through an iterative process (see Fig.11-2) that starts with clusters centered at essentially random locations in the record space and moves the cluster *centroids* (another name for the cluster means) around until each one is actually at the center of some cluster of records. This process is best illustrated through diagrams. To draw it easier, we show the process in two dimensions, but bear in mind that in practice the record space will have many more dimensions, because there will be a different dimension for each field in the records. We don't need to worry about drawing the clusters, because there are other ways of understanding them.

In the first step as shown in Fig.11-2 a) ***k*** data points are selected to be the seeds, more or less arbitrarily. Each of the seeds is an embryonic cluster with only one element. In this example, ***k*** is 3. In the second step, we assign each record to the cluster whose centroid is nearest. Drawing the boundaries between the clusters is easy if you recall from high school geometry that given two points, X and Y, all points that are equidistant from X and Y fall along a line that is half way along the line segment that joins X and Y and perpendicular to it. In the illustration, the initial seeds are joined by dashed lines and the cluster boundaries constructed from them are solid lines. In three dimensions these boundaries would be planes and in N dimensions they would be hyperplanes of dimension N-1.

As we continue to work through the ***k-means*** algorithm, one should pay particular attention to the fate of the point with the box drawn around it. On the basis of the initial seeds, it is assigned to the cluster controlled by seed number 2 because it is closer to that seed than to either of the others. It is on seed 3's side of the perpendicular line separating seeds 1 and 3, on seed 2's side of the perpendicular line separating seeds 2 and 3, and on the seed 2's side of the perpendicular line separating seeds 1 and 3.

At this point, every point has been assigned to exactly one of the three clusters centered around the original seeds. The next step is to calculate the centroids of the new clusters. This is simply a matter of averaging the positions of each point in the cluster along each dimension. Remember that ***k-means*** clustering requires that the data values be numeric. Therefore, it is possible to calculate the average position just by taking the average of each field. If there are 200 records assigned to a cluster and we are clustering based on four fields from those records, then geometrically we have 200 points in a four-dimensional space. The location of each record is described by four fields, with the form [x_1, x_2, x_3, x_4]. The value of x_1 for the new centroid is the average of all 200 x_1s and similarly for x_2, x_3, and x_4.

In Fig.11-2 c) the new centroids are marked with crosses. The arrows show the motion from the position of the original seeds to the new centroids of the clusters. Once the new clusters have been found, each point is once again assigned to the cluster with the closest centroid. Figure 11-2 d) shows the new cluster boundaries – formed, as before, by drawing lines equidistant between each pair of centroids. Notice that the point with the box around it, which was originally assigned to cluster number 2, has now been assigned to the cluster number 1. The process of assigning points to a cluster and then re-calculating centroids continues until the cluster boundaries stop changing. Happily, for most data sets, the cluster boundaries are set after a handful of iterations.

Since automatic cluster detection is an undirected technique, it can be applied without prior knowledge of the structure to be discovered. On the other hand, the clusters that are automatically detected have no other natural interpretation than the one for a given mapping of records to a geometric coordinate system. Therefore, it can be hard to put the results to practical use as some records are close to one another.

By choosing different distance measures, automatic clustering can be applied to almost any kind of data. For instance, there are measures of the distance between two passages of text that can be used to cluster newspaper articles into subject groups. Most clustering software, however, uses the Euclidean distance formula we all once learned in school – the one where you take the square root of the sum of the squares of the displacements along each axis. This means that non-numeric variables must be transformed and scaled before they can take part in the clustering. Depending on how these transformations are done, the categorical variables may dominate the clustering or be completely ignored.

In the ***k-means*** method, the original choice of a value for *k* determines the number of clusters that will be found. If this number does not match the natural structure of the data, the technique will not obtain good results. Unless there is some a *priori* reason to suspect the existence of a certain number of clusters, the researcher will probably want to experiment with different values for ***k***. Each set of clusters must then be evaluated. In general, the best set of clusters is the one that does the best job of keeping the distance between members of the same cluster small and the distance between members of adjacent c1usters large. For descriptive data mining, though, the best set of clusters may be the one that shows some unexpected pattern in the data.

The strength of automatic cluster detection is that it is an undirected knowledge discovery technique, but each strength usually has a corresponding weakness. If we don't know what we are looking for, we may not recognize it when we find it! The clusters generated by the automated clustering algorithms (whether ***k-means*** or any other algorithm) are not guaranteed to have any practical value. Once the clusters have been created, it is up to the user to interpret them.

The most frequently used approaches to applying clusters are:

- Building a decision tree with the cluster label as the target variable and using it to derive rules explaining how to assign new records to the correct cluster.
- Using visualization to see how the clusters are affected by changes in the input variables.
- Examining the differences in the distributions of variables from cluster to cluster, one variable at a time.

When to use Cluster Detection: we should use cluster detection when we suspect that there are natural groupings that may represent groups of customers or products that have a lot in common with each other. These may turn out to be naturally occurring customer segments for which customized marketing approaches are justified. More generally, clustering is often useful when there are many competing patterns in the data, making it hard to spot any single pattern. Creating clusters of similar records reduces the complexity within clusters so that other data mining techniques are more likely to succeed.

B. Decision Trees

Decision trees are a wonderfully versatile tool for *data mining*. ***Decision trees*** seem to come in nearly as many varieties as actual trees in a tropical rain forest. And, like deciduous and coniferous trees, there are two main types of ***decision trees***:

– ***Classification trees*** label records and assign them to the proper class. Classification trees can also provide the confidence that the classification is correct. In this case, the classification tree reports the class probability which is the confidence that a record is in a given class.

– ***Regression trees*** estimate the value of a target variable that takes on numeric values. So, a regression tree might calculate the amount that a donor will contribute or the expected size of claims made by an insured person.

All of these trees have the same structure. When a tree model is applied to data, each record flows through the tree along a path determined by a series of tests such as “is field 3 greater than 27?” or “is field 4 red, green, or blue?” until the record reaches a leaf or terminal node of the tree. There it is given a class label based on the class of the records that reached that node in the training set or, in the case of regression trees, assigned a value based on the mean (or other mathematical function) of the values that reached that leaf in the training set.

There are various algorithms for ***decision trees*** which produce trees that differ from one another in the number of splits allowed at each level of the tree, the way those splits are chosen when the tree is built, and the way the tree growth is limited to prevent overfitting. Although these variations have led to many doctoral theses, for the purposes of this text they are not very interesting. Today’s ***DM*** software tools typically allow the user to choose among several splitting criteria and pruning rules, and to control parameters such as minimum node size and maximum tree depth, allowing one to approximate any of these algorithms.

The discussion on clustering described how the fields in a record can be viewed as the coordinates of that record in a multidimensional record space. That geometric way of thinking is useful when talking about ***decision trees*** as well. Each branch of a decision tree is a test on a single variable that cuts the space into two or more pieces. For concreteness and simplicity, suppose a simple example where there are only two input variables, X and Y. These variables take on values from 0 to 100. Each split in the tree is constrained to be binary. That is to say, at every node in the tree, a record will go either left or right based on some test of either X or Y.

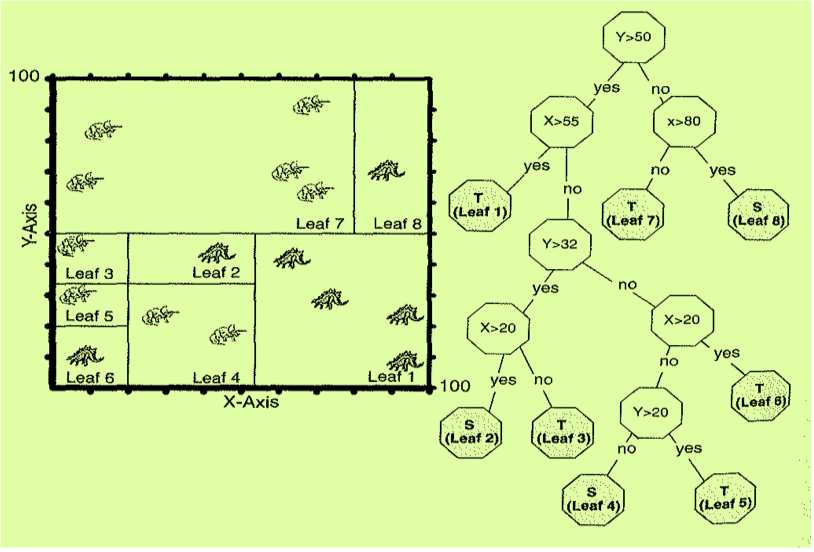


Fig.11-3 Decision tree cuts the space into boxes
(Source: Berry & Linoff, 2000, p.112)

The Decision tree in Fig.11-3 has been grown until every box is completely pure in the sense that it contains only one species of dinosaur. Such a tree is fine as a description of this particular arrangement of stegosauruses and triceratopses, but is unlikely to do a good job of classifying another similar set of big reptiles.

For some training sets, it is possible to build a decision tree that correctly classifies every single record. This is possible when the training set contains no examples of records whose input variables have the same values, but whose target variables belong to different classes. Although such a tree provides a good description of the training data, the tree is unlikely to generalize new data sets. That is why the test set is used to prune the tree once it has been grown using the training data set.

Why do we need to do this? A tree that precisely describes the data from which it was derived is unlikely to generalize well to another sample drawn from the same population. This problem is known as ***overfitting**, a* topic we will return to in the next and other sections. However, ignoring that for the moment, how would we use this tree to classify an unknown dinosaur for which X=40 and Y=75? Starting at the root node, we go to the right because the Y-value is greater than 50. Then, since the X-value is not greater than 80, we classify the unknown dinosaur as a Triceratops. Equivalently, by looking at the box chart we can see that the point (40, 75) is clearly in a box containing only triceratopses.

Decision trees are built through a process known as ***recursive partitioning***. *Recursive partitioning* is an iterative process of splitting the data up into partitions – and then splitting it up some more. Initially, all of the records in the training set (the pre-classified records that are used to determine the structure of the tree) are together in one big box. The algorithm then tries breaking up the data, using every possible binary split on every field. So, if age takes on 72 values, from 18 to 90, then one split is everyone who is 18 and everyone older than 18. Another is everyone who is 18 or 19, and everyone who is 20 or older, and so on... The algorithm chooses the split that partitions the data into two parts that are purer than the original. This splitting or partitioning procedure is then applied to each of the new boxes. The process continues until no more useful splits can be found.

The graph in Fig. 11-4 shows how to make a ***pruning decision***, using an approach which bases the pruning decision on the actual performance of the tree when data is plentiful (as it is rarely the case in the academic environments where algorithms are developed, but it is frequently true in the commercial world where they are applied). The performance of the tree and all of its subtrees is measured on a separate set of pre-classified data, called the test set (recall Chapter 3 and the "*external complement*" in Fig.3-11). With a single test set, the algorithm can prune back to the subtree that minimizes the error on the test set. With multiple test sets, we can even more directly address the issue of model generality by selecting the subtree that performs most consistently across several test sets.

It is important to point out some of the consequences of choosing ***decision trees***. First, because every split in a decision tree is a test on a single variable, ***decision trees*** can never discover rules that involve a relationship between variables. This puts a responsibility on the researcher to add derived variables to express relationships that are likely to be important.

For example, a loan database is likely to have fields for the initial amount of the loan and the remaining balance, but neither of these fields is likely to have much predictive value in isolation. The ratio of the outstanding balance to the initial amount carries much more helpful information, but a decision tree will never discover a single rule based on this ratio unless it is included as a separate variable.

In some situations, the way that ***decision trees*** handle numeric input variables can cause valuable information to be lost. When a split is chosen only the rank order of the observations comes into play. For the most part, this does not cause any problems, but under certain circumstances, information carried in the distribution of the values will be lost. Some decision tree algorithms first bin all numeric variables and then treat them as if they were categorical. This process can destroy information.

One advantage to the way ***decision trees*** treat numeric inputs is that they are not sensitive to scale differences between the inputs, nor to outliers and skewed distributions. This means that data preparation is less of a burden with ***decision trees*** than it is with ***ANNs*** or ***clustering***.

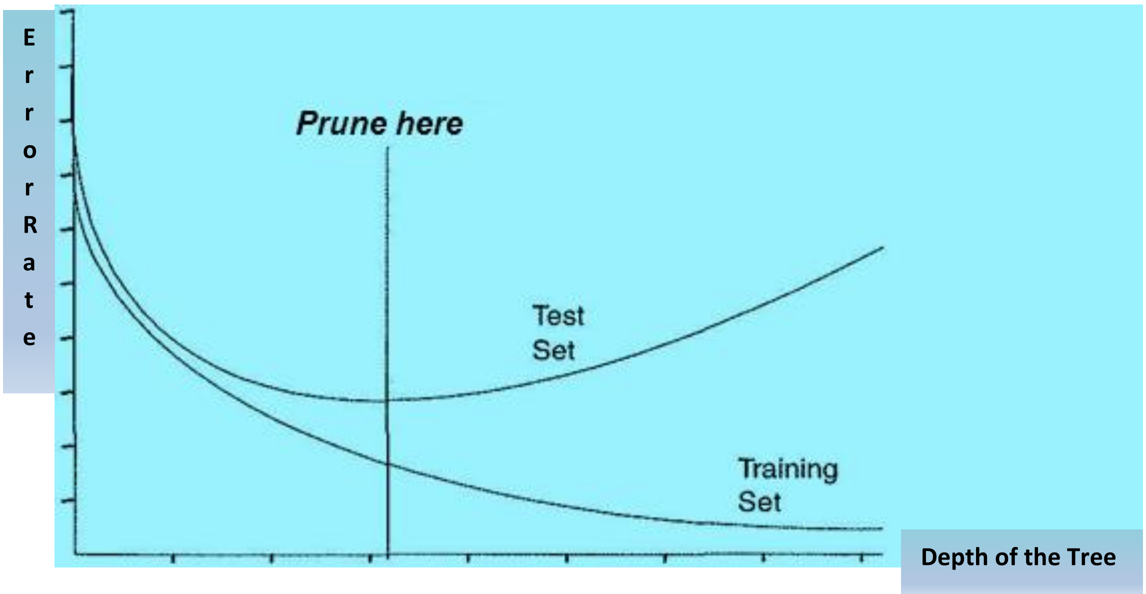


Fig.11-4 Error rate on training set and test set as tree complexity increases

The handling of categorical variables can also cause problems. Depending on the algorithm employed, categorical variables may be split on every value taken on by the variable, leading to a very bushy tree that soon runs out of records on which to base further splits. Other algorithms find ways to group class labels into a small number of larger classes by combining classes that yield similar splits. Since the number of possible groupings grows very large, very fast as the number of classes grows, an exhaustive search of all combinations quickly becomes impractical. Software products use various shortcuts to pare down the space to be searched, but the clustering process can still be quite time consuming. Categorical target variables that take on many values are also problematic.

Decision trees are error-prone when the number of training examples per class gets small. This can happen rather quickly in a tree with many levels and/or many branches per node because trees are very sensitive to the density of the outcomes.

Decision-tree building algorithms put the field that does the best job of splitting at the root node of the tree (and the same field may appear at other levels in the tree as well). It is not uncommon for ***decision trees*** to be used for no other purpose than prioritizing the independent variables. That is, using a decision tree, it is possible to pick the most important variables for predicting a particular outcome because these variables are chosen for splitting high in the tree. Usually this is better than using the *Stepwise regression* described in Chapter 4, but still not enough to address all issues. ***Self-organizing Data Mining*** discussed in the previous chapters provide many more options as a permanent solution of these problems.

Another useful consequence of the way that important variables float to the top is that it becomes very easy to spot input variables that are doing *too* good a job of prediction because they encode knowledge of the outcome that is available in the training data, but would not be available in the field. There are some curious examples of this, such as discovering that people with nonzero account numbers were the most likely to respond to an offer of credit – less than surprising since account numbers are assigned only after the application has been processed.

Decision tree methods are often chosen for their ability to generate understandable rules, but this ability can be overstated. It is certainly true that for any particular classified record, it is easy to simply trace the path from the root to the leaf where that record landed in order to generate the rule that led to the classification – and most decision tree tools have this capability. Many software products can output a tree as a list of rules in SQL, pseudocode, or pseudo-English. However, a large complex decision tree may contain hundreds or thousands of leaves. Such a tree is less likely than an ***ANN*** to communicate anything intelligible about the problem as a whole.

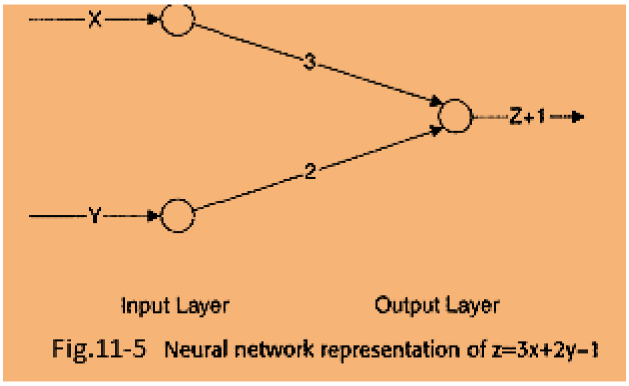


Fig.11-5 Neural network representation of z=3x+2y-1

When to use Decision Trees: Decision-tree methods are a good choice when the data mining task is classification of records or prediction of outcomes. We should use ***decision trees*** when the goal is to assign each record to one of a few broad categories. ***Decision trees*** are also a natural choice when the goal is to generate rules that can be easily understood, explained, and translated into SQL or a natural language.

11.3. Artificial Neural Networks

Artificial Neural Networks (ANNs), as quoted by many researchers, "are at once the most widely known and the least understood of the major data mining techniques". Much of the confusion stems from overreliance on the metaphor of the brain that gives the technique its name. The people who invented ***ANNs*** were not statisticians or data analysts. They were machine learning researchers interested in mimicking the behavior of natural neural networks such as those found in fruit flies, earthworms, and human beings. The vocabulary these machine learning and artificial intelligence researchers used to describe their work – *"perceptrons"*, *"neurons"*, *"learning"*, and the like – led to a romantic and anthropomorphic impression of ***artificial neural networks*** among the general public and to deep distrust among statisticians and analysts. Depending on your background, you may be either delighted or disappointed to learn that, whatever the original intentions of the early ***ANNs***, from a *data mining* perspective, ***ANNs*** are just another way of fitting a model to observed historical data in order to be able to make classifications or predictions.

To illustrate this point and to introduce the various components of a *neural network* (*NN*), it is worth noting that standard linear regression models and many other functions equally devoid of mystery can easily be drawn as neural network diagrams. If we take for example the function z= 3x+2y-l, there are two variable inputs, x and y. For any values of x and y, the function will return a value for z. It might be that this function is a model, based on many observed values of x, y, and z that is now being used to predict values for z given new, previously unobserved values of x and y. In such a case, it is a predictive model just as surely as anything created by a data mining software package. This particular predictive model can be easily represented by the simple *NN* as shown in Fig.11-5.

In neural network terminology, this network has an ***input layer*** and an ***output layer***. Each of the inputs x and y gets its own *unit*, or *network node*. In general, it is not the actual values of the input variables that are fed into the input layer but some transformation of them. Each input unit is connected to the output unit with a *weight*. In this case, the weights are the coefficients 3 and 2. Inside the output unit, the input weights are combined using a ***combination function*** (typically summation, as in this case) and then passed to a ***transfer function*** the result of which is the output of the network. Together, the combination function and the transfer function make up the unit's ***activation function***. The value produced by the output node's activation function is usually some transformation of the actual desired output. In this case, the network outputs ***z+1*** rather than ***z***. Just as a function is applied to the input variables in order to generate suitable inputs to the ***ANN***, a certain function of the network's output is required to translate it back to the actual range of the target variable.

In fact, most ***ANNs*** are not as simple as the one in Fig.11-5. There is usually one or more additional layers of units between the input and output layers. These layers are called ***hidden layers*** and the units in them are *hidden units*. Fig.11-6 represents a *NN* similar to the one in Fig.11-5 but with a hidden layer. With the addition of the hidden layer, the function represented by the network is no longer a simple combination of its inputs. The output value is now calculated by feeding the weights coming from the two hidden units to the activation function of the output unit. The weights produced by the hidden units are themselves functions of the input units, each of which is connected to both units of the hidden layer. All this gets pretty complicated, pretty quickly, which is why no one ever actually writes out a *NN* as an equation, although it is possible.

The network illustrated here is a feed-forward network with a hidden layer. By feed-forward, we mean that data enters at the input nodes and exits at the output nodes without ever looping back on itself. ***ANNs*** like this are also called ***multilayer perceptrons***. If there is a "standard" *NN*, it is the fully connected, feed-forward network with one hidden layer and a single node output layer, but there are many possible variations. Often, there are multiple nodes in the output layer, each estimating the probability of a separate class of the target variable. Sometimes there is more than one hidden layer, or there are direct links from inputs to outputs that skip the hidden layer. There are neural network architectures that include loops and ones where the inputs arrive in waves, not all at the same time. Usually, when ***ANNs*** are discussed, we are referring to fully connected, feed-forward, multilayer perceptrons.

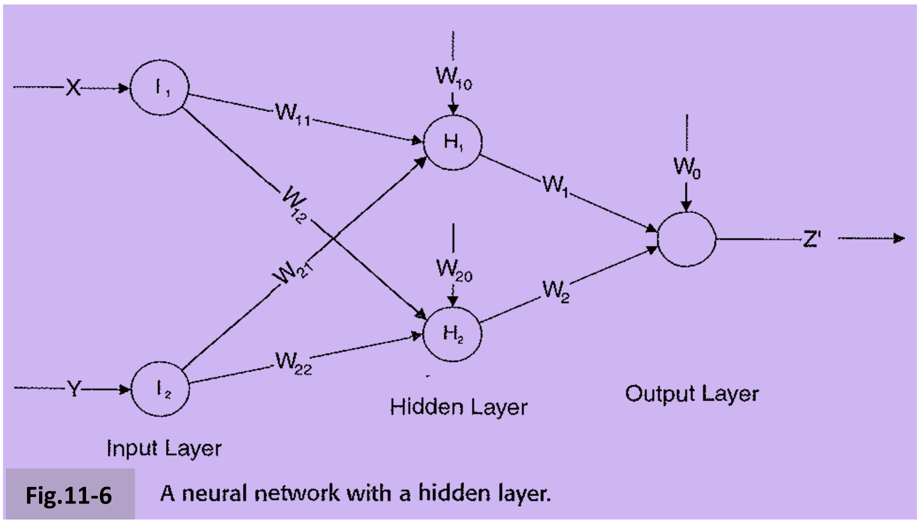


Fig.11-6 A neural network with a hidden layer.

Inside each unit of a *NN*, there is an activation function that consists of a combination function and a transfer function. The combination function is nearly always the weighted sum of the inputs. Transfer functions come in many more flavors.

The graph in Fig. 11-7 shows a linear transfer function illustrating the network drawn in Fig.11-6, which represents a linear function. More commonly, the transfer function is ***sigmoidal*** (S-shaped) or bell-shaped. The bell-shaped transfer functions are called ***radial basis functions***. Common sigmoidal transfer functions are the arctangent, the hyperbolic tangent, and the logistic. The nice thing about these S-shaped and bell-shaped functions is that any curve, no matter how wavy, can be created by adding together enough S-shaped or bell-shaped curves. In fact, multilayer perceptrons with sigmoidal transfer functions and radial basis networks are both ***universal approximators***, meaning that they can theoretically approximate any continuous function to any degree of accuracy. Of course, theory does not guarantee that we can actually find the right *NN* to approximate any particular function in a finite amount of time, but it is good to know it and also that ***decision trees*** are not universal approximators.

The sigmoidal transfer functions used in the classic multilayer perceptrons have several properties that provide sometimes big advantages. The shape of the curve means that no matter how extreme the input values, the output value is always constrained to a known range (–1 to 1 for the hyperbolic tangent and the arctangent, 0 to 1 for the logistic). For moderate input values, the slope of the curve is nearly constant. Within this range, the sigmoid function is almost linear and exhibits almost-linear behavior. As the weights get larger, the response becomes less and less linear as it takes a larger and larger change in the input to cause a small change in the output. This behavior corresponds to a gradual movement from a linear model to a nonlinear model as the inputs become extreme.

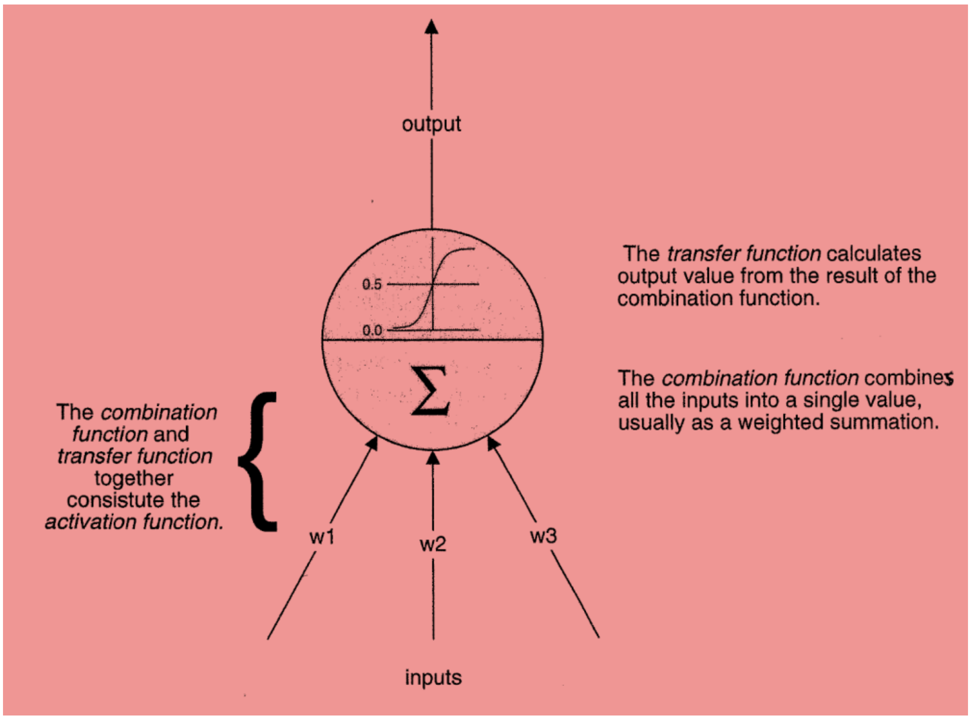

output
The *transfer function* calculates output value from the result of the combination function.
0.5
0.0
Σ
The *combination function* combines all the inputs into a single value, usually as a weighted summation.
The *combination function* and *transfer function* together consistute the *activation function.*
w1
w2
w3
inputs

Fig.11-7 Linear transfer function

Training a neural network is the process of setting the weights on the inputs of each of the units in such a way that the ***ANN*** best approximates the underlying function, or according the data mining terms, does the best job of predicting the target variable. This is an optimization problem and there are whole textbooks dedicated to optimization, but in broad outline most software packages for building neural network models use some variation of the technique known as ***backpropagation***. The term ***backpropagation*** refers to any method of training an *ANN* that involves comparing the expected result for a given set of inputs to the output of the network during a training run, and feeding that difference back through the network to adjust the weights. In general, most of the ***ANNs*** in use today are trained using backpropagation. However, the original backpropagation networks popularized in the 1980s used an optimization method called *steepest descent* to correct the network weights. This turns out to be inefficient and is now generally replaced by other algorithms such as conjugate gradient or modified Newton. Some writers reserve the term “backpropagation networks” for the earlier, less efficient variety and coin new terms for each combination of error estimate and optimization method. This could be somehow confusing, so we will refer to all as “backpropagation methods”.

Training a backpropagation *NN* has three steps:

1. The network gets a training instance and, using the existing weights in the network, it calculates the output or outputs for the instance.

2. Backpropagation then calculates the error, by taking the difference between the calculated result and the expected (actual) result.

3. The error is used to adjust the weights (this is referred to as *feeding the error back through the network*).

Using the error measure to adjust the weights is the critical part of any back-propagation algorithm. In classic backpropagation, each unit is assigned a specific responsibility for the error. For instance, in the output layer, one unit is responsible for the whole error. This unit then assigns a responsibility for part of the error to each of its inputs, which come from units in the hidden layer, and so on, if there is more than one hidden layer. The specific mechanism is not important. Suffice it to say that it is a complicated mathematical procedure that requires taking partial derivatives of the transfer function. More recent techniques adjust all the weights at once, which is one of the things that make them more efficient.

Given the error, how does a unit adjust its weights? It starts by measuring how sensitive its output is to each of its inputs. That is, it estimates whether changing the weight on each input would increase or decrease the error. The unit then adjusts each weight to reduce, but not eliminate, the error. The adjustments for each example in the training set slowly nudge the weights toward their optimal values. The goal is to generalize and identify patterns in the input, not to match the training set exactly. Adjusting the weights is "like a leisurely walk instead of a mad-dash sprint". After being shown enough training examples, the weights on the network no longer change significantly and the error no longer decreases. This is the point where training stops – the network has learned the input.

One of the concerns with any neural network training technique is the risk of falling into something called a local optimum. This happens when the adjustments to the network weights, suggested by whatever optimization method is in use, no longer improve the performance of the *NN* even though there is some other combination of weights, significantly different from those in the *NN*, that yields a much better solution. This is analogous to trying to climb to the top of a mountain and finding that you have only climbed to the top of a nearby hill. There is a tension between finding the local best solution and the global best solution. Adjusting parameters such as the learning rate and momentum helps to find the best solution.

ANNs can produce very good predictions, but in general, they are neither easy to use nor easy to understand. The difficulties with ease of use stem mainly from the extensive data preparation required to get good results from a neural network model. The results are difficult to understand because a *NN* is a complex nonlinear model that does not produce rules.

The biggest drawback of typical ***ANN*** in a business decision support context is that it *cannot explain results*. For many users, understanding what is going on is often as important, if not more important, than getting the best prediction. In situations where explaining rules may be critical, such as denying loan applications, ***ANNs*** are not a good choice. There are many situations, however, when the prediction itself matters far more than the explanation. For example, the neural network models which can spot a potentially fraudulent credit card transaction before it has been completed. An analyst or data miner can study the historical data at leisure in order to come up with a good explanation of why the transaction was suspicious, but in the moments after the card is swiped, the most important thing is to make a quick and accurate prediction.

When to use Artificial Neural Networks: ***ANNs*** are a good choice for most classification and prediction tasks when the results of the model are more important than understanding how the model works. ***ANN*** actually represent complex mathematical equations, with lots of summations, exponential functions, and many parameters. These equations describe the neural network, but are quite opaque to human eyes. The equation is the rule of the network, and it is useless for our understanding.

Typical ***ANN*** does not work well when there are many hundreds or thousands of input features. Large numbers of features make it more difficult for the network to find patterns and can result in long training phases that never converge to a good solution. Here, ***ANN*** can work well with decision tree methods. ***Decision trees*** are good at choosing the most important variables and these can then be used for training a network.

The primary lesson that one should take away from this chapter is that no one *data mining technique* is right for all situations and a possible solution is in the unification of different techniques. One possible solution could be the so called ***Statistical Learning Networks*** in ***Self-organizing Data Mining*** (Mueller & Lemke, 2003), such as ***MLNAN*** already discussed in the previous chapters. In Chapter 12 these techniques will be presented in detail. The chapter will also discuss how they can help researchers to analyze massive amounts of data and to turn information located in the data into successful decisions.

***

SUMMARY AND CONCLUSIONS

In Chapter 11 the connection between ***Business Forecasting*** and ***Data Mining (DM)*** is discussed. There is a big variety of specific advanced approaches in theory and practice, which could be used successfully in *Business Forecasting*.

- ***Knowledge discovery from data (KDD)*** is an interactive and iterative process of solving subtasks and decisions like data selection and preprocessing, choice and application of data mining algorithms, and analysis of the extracted knowledge.
- ***Data mining techniques*** in ***KDD*** help researchers in analyzing the massive amounts of data and turning information located in data into successful decisions.
- In ***Business Forecasting***, ***KDD*** helps researchers to develop good and reliable models. ***KDD*** has a wide range of applications and business forecasting is just one of them.
- ***Data mining*** itself includes not only a straightforward utilization of a single analytical technique but also consists of processes which are appropriate for many methods and techniques depending on the nature of the inquiry. This set of methods contains data visualization, tree-based models, neural networks, methods of mathematical statistics (like clustering and discriminant analysis and regression and correlation analysis), and methods of artificial intelligence, such as genetic algorithms, self-organizing techniques and others.

Data mining is a method of searching data for unexpected patterns or relationships using a variety of tools and algorithms. Fayyad et al., (1996) identified six tasks as follow:

- *classification*: learning a function that maps (classifies) a data item into one of several predefined classes;
- *regression*: learning a function that maps a data item into a real-valued prediction variable;
- *clustering*: identifying a finite set of categories or clusters to describe the data;
- *summarization*: finding a compact description for a subset of data;
- *dependency modeling*: finding a model that describes significant dependencies between variables;
- *change and deviation detection*: discovering the most significant changes in the data from previously measured or normative values.

Data mining techniques are like anything else a computer does, such as storing files or creating a spreadsheet. The *techniques* are general approaches to solving problems, and there are usually many ways to approach the technique. Each of these ways is a different algorithm. The *algorithms* are like recipes with step-by-step instructions explaining what is happening:

- ***Clustering:*** There are many mathematical approaches to finding clusters in data. Some methods, called *divisive methods*, start by considering all records to be part of one big cluster. That cluster is then split into two or more smaller clusters, which are themselves split until eventually each record has a cluster all to itself. At each step in the process, some measure of the value of the splits is recorded so that the best set of clusters can be chosen in the end. Other methods, called *agglomerative method*s, start with each record occupying a separate cluster, and iteratively combine clusters until there is one big one containing all the records. There are also *self-organizing maps,* a specialized form of neural network that can be used for cluster detection and others...
- ***Cluster Detection*** should be used when we suspect that there are natural groupings that may represent groups of customers or products that have a lot in common with each other. These may turn out to be naturally occurring customer segments for which customized marketing approaches are justified. More generally, clustering is often useful when there are many competing patterns in the data, making it hard to spot any single pattern. Creating clusters of similar records reduces the complexity within clusters so that other data mining techniques are more likely to succeed.
- ***Decision trees*** are a wonderfully versatile tool for ***DM***. ***Decision trees*** seem to come in nearly as many varieties as actual trees in a tropical rain forest. And, like deciduous and coniferous trees, there are two main types of ***decision trees***:

  – *Classification trees* label records and assign them to the proper class. Classification trees can also provide the confidence that the classification is correct. In this case, the classification tree reports the class probability, which is the confidence that a record is in a given class.

  – *Regression trees* estimate the value of a target variable that takes on numeric values. So, a regression tree might calculate the amount that a donor will contribute or the expected size of claims made by an insured person.
- ***Decision-tree methods*** are a good choice when the data mining task is classification of records or prediction of outcomes. We should use ***decision trees*** when the goal is to assign each record to one of a few broad categories. ***Decision trees*** are also a natural choice when the goal is to generate rules that can be easily understood, explained, and translated into SQL or a natural language. One advantage to the way

decision trees treat numeric inputs is that they are not sensitive to scale differences between the inputs, nor to outliers and skewed distributions. This means that data preparation is less of a burden with ***decision trees*** than it is with ***ANN*** or clustering.

Artificial Neural Networks (ANNs) are at once the most widely known and the least understood of the major data mining techniques. From a *data mining* perspective, ***ANNs*** are just another way of fitting a model to observed historical data in order to be able to make classifications or predictions:

- In *neural network* terminology, an ***ANN*** has an *input layer* and an *output layer.*
- Each of the inputs gets its own *unit,* or *network node*. In general, it is not the actual values of the input variables that are fed into the input layer but some transformation of them.
- Each input unit is connected to the output unit with a *weight.*
- Inside the output unit, the input weights are combined using a *combination function* (typically summation) and then passed to a *transfer function* the result of which is the output of the network.
- Together, the combination function and the transfer function make up the unit's *activation function.* The value produced by the output node's activation function is usually some transformation of the actual desired output. Just as a function is applied to the input variables in order to generate suitable inputs to the ***ANN***, a certain function of the network's output is required to translate it back to the actual range of the target variable.
- Most ***ANNs*** usually have one or more additional *layers of units* between the *input* and *output layers*. These layers are called *hidden layers* and the units in them are *hidden units.*

Training a neural network is the process of setting the weights on the inputs of each of the units in such a way that the ***ANN*** best approximates the underlying function, or put in data mining terms, does the best job of predicting the target variable. This is an optimization problem and there are whole textbooks dedicated to optimization, but in broad outline most software packages for building neural network models use some variation of the technique known as *backpropagation:*

- ***Backpropagation*** refers to any method of training an ***ANN*** that involves comparing the expected result for a given set of inputs to the output of the network during a training run, and feeding that difference back through the network to adjust the weights.

- Training a backpropagation *NN* has three steps:

  1. The network gets a training instance and, using the existing weights in the network, it calculates the output or outputs for the instance.

  2. Backpropagation then calculates the error, by taking the difference between the calculated result and the expected (actual) result.

  3. The error is used to adjust the weights (this is referred to as *feeding the error back through the network*).

<u>When to use Artificial Neural Networks:</u>

- ***ANNs*** can produce very good predictions, but in general, they are neither easy to use nor easy to understand. The biggest drawback of typical ***ANN*** in a business decision support context is that they *cannot explain results*.
- ***ANNs*** are a good choice for most classification and prediction tasks when the results of the model are more important than understanding how the model works. ***ANNs*** actually represent complex mathematical equations, with lots of summations, exponential functions, and many parameters. These equations describe the neural network, but are quite opaque to human eyes. The equation is the rule of the network, and it is useless for our understanding.
- Typical ***ANN*** do not work well when there are many hundreds or thousands of input features. Large numbers of features make it more difficult for the network to find patterns and can result in long training phases that never converge to a good solution. Here, ***ANNs*** can work well with decision tree methods. ***Decision trees*** are good at choosing the most important variables and these can then be used for training a network.

The primary lesson that one should take away from this chapter is that no one *data mining technique* is right for all situations and a possible solution is in the unification of different techniques. One possible solution could be the so called ***Statistical Learning Networks*** in ***Self-organizing Data Mining***, such as ***MLNAN*** already discussed in the previous chapters. In next chapter these techniques will be presented in detail. Chapter 12 will also discuss how these specific techniques can help researchers to analyze massive amounts of data and to turn information located in the data into successful decisions.

KEY TERMS

CHAPTER EXERCISES

Conceptual Questions:

1. Define ***Knowledge discovery from data (KDD)***. What are the differences between ***Knowledge discovery from data*** and ***Data Mining***?
2. What is the purpose of ***Data Mining*** techniques in ***KDD?*** How does it related to ***Business Forecasting***?
3. List and describe all major six tasks in ***Data mining*** as identified by Fayyad et al.?
4. What are the main groups of ***Data mining techniques?*** Explain when is appropriate to use ***Cluster Detection*** and/or ***Decision trees?***
5. What are the ***Artificial Neural Networks (ANNs)*** like?
6. How do we ***Train a neural network?*** Define ***backpropagation*** according to the ***ANN*** building process?
7. Explain the three steps in training a ***backpropagation NN***. What is the purpose of *feeding the error back through the network*

Business Applications:

1. Open the file SalesData.xlsx in ***MS Excel***.
2. Open ***Knowledge-Miner (yX) for Excel*** and import the opened file SalesData from MS Excel.
3. Prepare Data Table in ***Knowledge-Miner (yX) for Excel*** for building an ***AR*** model***:***
   - Set up the time series data for a dependent variable “Sales”;
   - Select time lags of up to 12 periods within the time-series.
4. Self-organize the ***AR*** model and compute Sales forecast for the next 12 months.
5. Export the results in MS Excel worksheet:
   - Analyze the output of the mining process. Are there any potential improvements to be made? If yes return to step 1 and improve the ***AR*** model.
   - Design formulas, similar to the formulas in Part 4 of the Integrative case and compute MAD, MSE, MAPE and MPE for the new model, for a testing dataset of the 12 new monthly forecasts given in spreadsheet Predictions.
   - What is the model accuracy? Are there any initial assumptions/reasons leading to this conclusion?

Discuss all findings and write a short report (up to two pages) summarizing your answers.

INTEGRATIVE CASE

HEALTHY FOOD SYPPLY CHAIN & STORES

Part 11: Self-Organizing Data Mining Forecasts – 1

In Chapter 1 we introduced *Healthy Food Stores* – a fast-growing retail food provider with 12 stores in a northwestern state. The company executives decided to study the effect that company advertising dollars have on sales. They hoped that examining collected historical data would reveal relationships that would help determine future advertising expenditures and predict monthly sales volumes for the upcoming quarter.

After identifying basic parameters, along with input (independent) and output (dependent) variables of the forecasting scenario in Part 2, the research team determined the main elements of the forecasting process:

- Forecasting horizon of up to twelve months;
- Quarterly forecast updates, since accuracy decreases as time horizon increases, and sufficient time is needed to implement possible changes;
- Development of different models based on data patterns, if any, and evaluation of their accuracy in order to select the most appropriate one;
- Selection of the best forecasting model with no more than 5% forecasting error.

The benchmark forecast computed in Part 3 (using the baseline of ***one-step naive forecast*** as a reference forecast) provided basic values for the most common measures of accuracy MFE, MAD, MAPE, MPE, MSE and CV(RMSE).

In Part 4, information about the opinion of some important people from the *Healthy Food Stores Company*, concerning this specific case, was collected and the research team applied the Delphi method to top executives group, Sales-force composite to the sales managers from all 12 stores and Scenario writing to the most experienced professionals from Advertising Department. After collecting such valuable information from different sources, in Part 5 the research team made its first steps in Numerical Predictions by developing different basic forecasting models. They created spreadsheets for Naïve techniques (Average model, Random Walk with Drift and Seasonal Naïve Technique), simple Moving Average, Simple Exponential Smoothing and Triple Exponential Smoothing, used to expand the base-line of one-step naïve forecast as reference forecasts.

In Part 6 the research team analyzed the relationships between dependent variable Sales and the available predictors. After performing multiple correlation and regression analysis, researchers developed reliable forecasting model, which passed all tests and hypotheses,

representing the real system with certain error. In Part 7, the model was expanded by adding Dummy seasonal variables to analyze the Seasonal effect in company Sales. In Part 8, the improvement of the forecasting model continued (with the help of some advanced Time series analyses and predictive techniques) and few simple ***AR*** models were build using ***ARIMA*** methodology and *Gretl* software.

In *Parts 9* & *10* (**Complex Models and Forecasting**), new and more complex ***AR*** and ***ARMAX*** models were build using ***ARIMA*** methodology and *Gretl* software.

The next step would be to build Time series (***AR***) models as ***ANN***. These ***AR*** models would be self-organized using Data Mining platform ***Knowledge-Miner (yX) for Excel.***

Case Questions

1. Open Data.xslx file in ***MS Excel*** and select Data worksheet.
2. Open ***Knowledge-Miner (yX) for Excel*** and import the worksheet Data from MS Excel.
3. Prepare Data Table in ***Knowledge-Miner (yX) for Excel*** for building an ***AR*** model***:***
   - Set up the time series data for a dependent variable “Sales”;
   - Select time lags of up to 12 periods within the time-series.
4. Self-organize the ***AR*** model and compute Sales forecast for the next 12 months.
5. Export the results in MS Excel worksheet and rename it to MLNANAR (use the same file Data.xslx) and analyze the output of the mining process. Are there any potential improvements to be made? If yes return to step 1 and improve the ***AR*** model.
6. Use (copy/paste) the formulas designed in Part 3 to compute MFE, MAD, MAPE, MPE, MSE and CV(RMSE) for the new model, for the given testing dataset of 12 monthly forecasts provided in spreadsheet Errors.
7. Comment and analyze model’s accuracy - how good is the accuracy of these forecasts? What model, out of all models so far provides the best accuracy? Discuss.
8. What overall recommendations and in particular about Self-organizing Auto Regression models would you make to the research team? Explain.
9. Write a report (at least two pages not counting charts and tables) on the questions above, discussing all important findings and draw relevant conclusions about this part of the Integrative Case.

References

Berry, M., & Linoff, G. (2000). *Mastering Data Mining.* Wiley.

Brachman, R. & Anand, T. (1996). The Process of Knowledge Discovery in Databases: A Human-Centered Approach, *Advances in Knowledge Discovery and Data Mining*. The MIT Press, 37–58.

Devlin, B. (1997). *Data Warehouse from Architecture to Implementation*. Addison-Wesley.

Fayyad, U., Piatetsky-Shapiro, Gr., & Smyth, P. (1996, Fall). From Data Mining to Knowledge Discovery in Databases, *American Association for Artificial Intelligence Magazine*, 37-54. http://www.kdnuggets.com/gpspubs/aimag-***KDD***-overview-1996-Fayyad.pdf

Frawley, W., Piatetsky-Shapiro, Gr. & Matheus, C. (1992, Fall) Knowledge Discovery in Databases: An Overview, *AI Magazine*, 13(3), 57-70. https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1011/929

Madala, H. R., & Ivakhnenko, A. G. (1994). *Inductive Learning Algorithms for Complex Systems Modelling*. Boca Raton, FL: CRC Press Inc.

Mueller J. A., & Lemke, F. (2003). *Self-Organizing Data Mining: An Intelligent Approach To Extract Knowledge From Data.* Victoria, BC: Trafford Publishing.

***

CHAPTER 12. SELF-ORGANIZING DATA MINING FOR MODEL BUILDING AND FORECASTING

12.1. GMDH based Self-Organizing Data Mining Algorithms

In the second half of the 20th century, the Russian scientist Alexey Ivakhnenko (1968) introduced the ***Group Method of Data Handling (GMDH)***[1] as an inductive approach to model building based on self-organization principles. ***GMDH*** is a *heuristic self-organizing modeling method* and is particularly useful in solving the problem of modeling multi-input to single-output data. In ***GMDH-based Self-organizing modeling algorithms*** (Madala & Ivakhnenko, 1994), models are generated adaptively from data in the form of networks of active neurons. In this procedure, a repetitive generation of populations of competing models of growing complexity, corresponding validation, and model selection are done until an optimal complex model, neither too simple nor too complex, has been identified (see Fig. 12-1).

The modeling approach grows a tree-like network out of data of input and output variables (seed information) in a pair-wise combination and competitive selection from a simple single unit (neuron) to a desired final solution that does not have a predefined behavior (model). In this approach, neither the number of neurons and the number of layers in the network, nor the actual behavior of each created neuron is predefined. The modeling is self-organizing because the number of neurons, the number of layers, and the actual behavior of each created neuron are adjusting during the process of self-organization.

GMDH is the most successful method in ***Statistical Learning Networks*** (Mueller & Lemke, 2003, p.57) designed to address the common problems of ***ANNs*** as follows: by default, ***ANNs*** are implicit models which have no explanation; ***ANN*** topology designing is a trial-and-error process; in ***ANN*** design, there are no rules how to use the theoretical a priori knowledge, and so on…

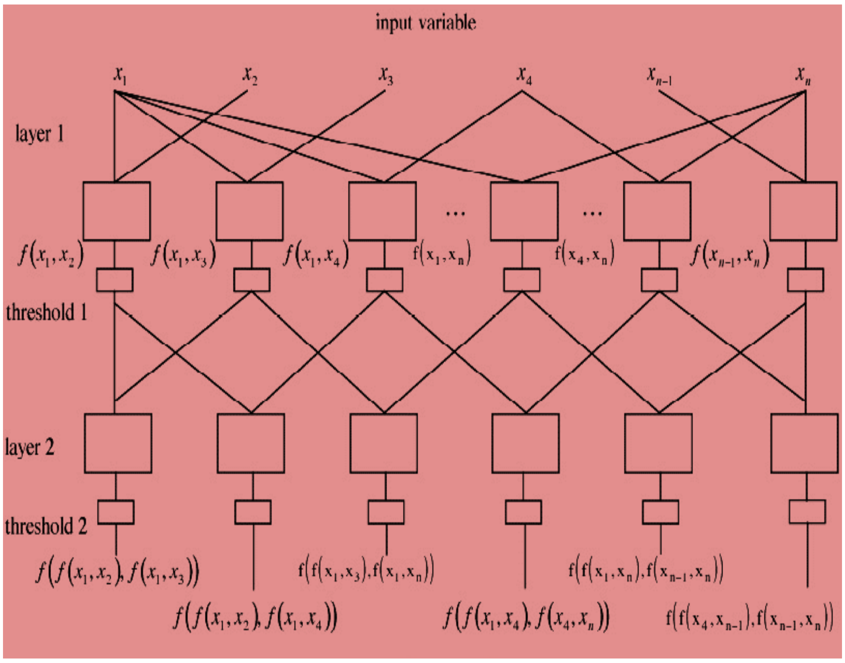

input variable
x_1
x_2
x_3
x_4
x_{n-1}
x_n
layer 1
...
...
$f(x_1,x_2)$
$f(x_1,x_3)$
$f(x_1,x_4)$
$f(x_1,x_n)$
$f(x_4,x_n)$
$f(x_{n-1},x_n)$
threshold 1
layer 2
threshold 2
$f(f(x_1,x_2),f(x_1,x_3))$
$f(f(x_1,x_3),f(x_1,x_n))$
$f(f(x_1,x_n),f(x_{n-1},x_n))$
$f(f(x_1,x_2),f(x_1,x_4))$
$f(f(x_1,x_4),f(x_4,x_n))$
$f(f(x_4,x_{n-1}),f(x_{n-1},x_n))$

Fig.12-1 General scheme of ***GMDH*** self-organizing modeling algorithm

[1] GMDH is a method of inductive statistical learning. See. http://en.wikipedia.org/wiki/Group_Method_of_Data_Handling

Table 12.1. Self-Organizing Data Mining Algorithms

Variables	Parametric	Non-parametric
Continuous	- Combinatorial (COMBI) - Multilayered Iterative (MIA) - Objective System Analysis (OSA) - Harmonic - Two-level (ARIMAD) - Multiplicative-Additive (MAA)	- Objective Computer Clusterization (OCC); - "Pointing Finger" (PF) clusterization algorithm; - Analogues Complexing (AC)
Discrete or binary	- Harmonic Re-discretization	- Based on Multilayered Theory of Statistical Decisions

There are many applications of ***Self-Organizing Data Mining Algorithms.*** The ***Multi-Layered Net of Active Neurons (MLNAN)*** technique described in Chapters 8 is a multilayered ***GMDH*** algorithm for multi-input to single-output models identification. Like other ***Multilayered Iterative GMDH algorithms*** (***MIA***) it can be used for single equation specification in the ***reduced form*** of the ***SE*** model (9-20) as presented in Chapter 9. Table 12.1 displays a summary of the basic type algorithms developed so far. Most of them have been used successfully to address existing problems in model building (Madala & Ivakhnenko, 1994; Ivakhnenko & Müller, 1996; Mueller & Lemke, 2003; Onwubolu, 2008; Motzev, 2010.)

Combinatorial (COMBI) algorithm

When the data sample is transformed to the conditional form as given by Gauss, the various interpolation problems of artificial intelligence, such as pattern recognition, dependence detection, stepwise prediction of random processes, and so on, can be solved by general algorithms, mostly by the combinatorial GMDH algorithm. These algorithms differ mostly in the choice of output variables and modeling space coordinates. In pattern recognition and dependence detection, the algorithms are searching to find a discriminant function, whereas in prediction, they are looking for a model to compute the further values of the dependent variables using current and lagged values of all variables. The interpolation-type models accept current and lagged values as coordinates of modeling space and the future values also can be used.

COMBI is based on full or reduced sorting-out of gradually complicated models and their evaluation by external criterion on a testing data set. The schematic flow of the algorithm is presented in Fig.12-2. The ***COMBI*** algorithm generates models of all possible input variable combinations and regarding the chosen selection criterion selects a final best model from the generated set of models. It is a complete model induction algorithm where all possible models are considered and no one is missed (see an Example in Fig. 12-3).

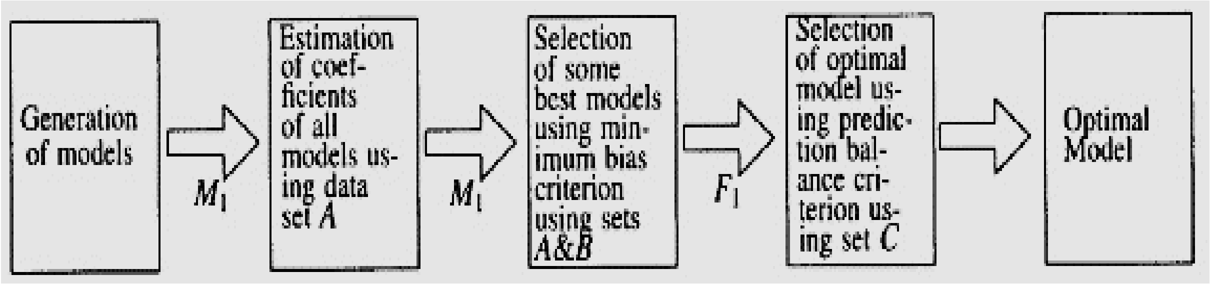

Generation of models
M_1
Estimation of coef-ficients of all models us-ing data set A
M_1
Selection of some best models using min-imum bias criterion using sets $A\&B$
F_1
Selection of optimal model us-ing predic-tion bal-ance cri-terion us-ing set C
Optimal Model

Fig.12-2 Schematic flow of single-layered combinatorial algorithm
(Source: Madala & Ivakhnenko, 1994, p.35)

The disadvantage, however, is that such algorithm can handle effectively only up to 30-40 input variables due to a nonlinear increase of the total number of possible model versions. Since a complex system can easily reach a few hundred input variables, other alternative algorithms were elaborated.

One possible way of improvement is to apply a Recursive scheme for faster combinatorial sorting. The recursive technique is convenient to use in constructing models of partial polynomials of gradually increasing complexity that begin with a single argument. This type of approach is called "method of bordering" (Madala & Ivakhnenko, 1994, p. 37).

Another improvement is the *multilayered structures using combinatorial setup*. One version of a multilayered structure is that the combinatorial algorithm could be implemented at each layer of the multilayered network structure by keeping the limit on the "freedom of choice" at each layer. The unit outputs are fed forward layer by layer as per the threshold measure to obtain the global output response for optimal complexity. This structure is exhibited in Fig.12-4 with three input arguments and three selected nodes at each layer.

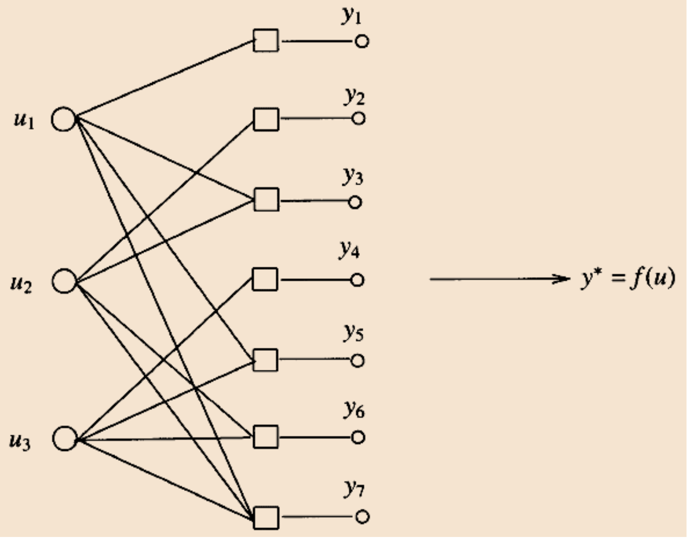


Fig.12-3 Example of a single-layered layout of combinatorial structure
(Source: Madala & Ivakhnenko, 1994, p.34)

In the first layer, all models containing single inputs (nodes) are estimated and some of the best are selected given certain external criteria, and passed on to the next layer as inputs. In the second layer, different inputs are selected and added to these models which improves the response given the external criteria. This continues until the model accuracy deteriorates. In contrast to the original multilayered setup, the outputs of the units are not passed on. The multilayered error is not passed on because of the retainment of the original basis functions, i.e. their number of inputs coincides with the layer number, and the total number of layers cannot exceed *m*.

The important feature of this algorithm is its realization of the recursive procedure for successive estimation of coefficients of the partial models according to the ***LS*** method.

Multilayered Iterative Algorithm (MIA)

The *Combinatorial algorithm* described in the previous subsection may likewise be called a *multilayered* or *iterative* one, despite the fact that the iteration rule grows more complicated with every layer. This is possibly the reason why it is customary to call the ***Multilayered Iterative GMDH algorithm*** an algorithm in which the iteration rule remains unchanged from one layer to the next. It should be used when it is needed to handle a big number of variables.

The network structures in ***GMDH*** differ as per the interconnections among the units and their hierarchical levels. Multilayered algorithms use a multilayered network structure with linearized input arguments and generate simple partial functions. Regarding the computational aspects in the process, the multilayered network procedures are more repetitive in nature. It is important to consider the algorithm in modules and facilitate repetitive characteristics.

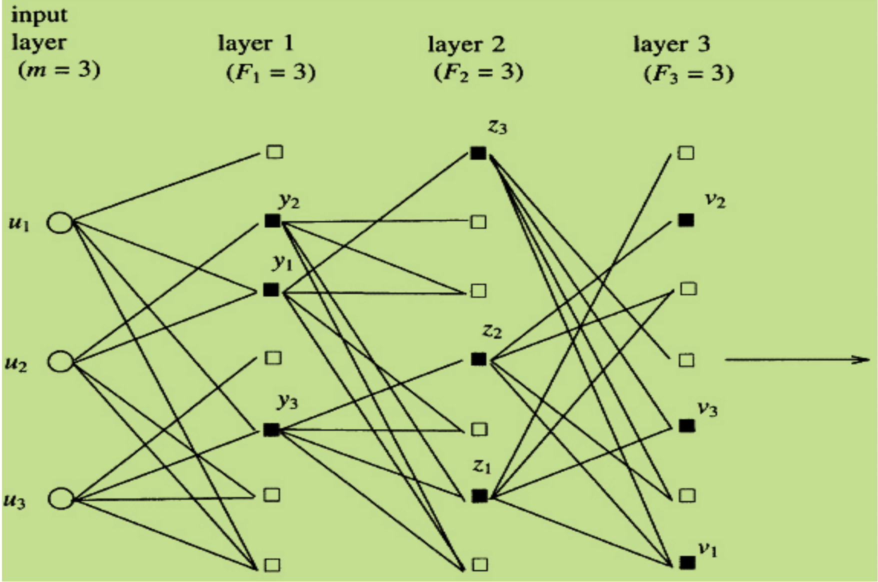


Fig.12-4 Multilayered structure with restricted combinatorial set up at each layer (Source: Madala & Ivakhnenko, 1994, p.39]

A multilayered network is a parallel-bounded structure that is built up on the basis of the connectionism[2], which is given in the basic iterative algorithm with linearized input variables and information in the network flows only forward. Each layer has a number of simulated units depending upon the number of input variables. Two input variables are passed on through each unit. For example, X_i and X_j are passed on through k^{th} unit and build a summation function. Weights are estimated using the training set. At the threshold level, error criterion is used to evaluate this function using the validation (test) set. If there are m input variables, the first layer generates M_1 ($=C^2_m$) functions. Here, F_1 ($F_1<M_1$) units as per the threshold values are made "on" to the next layer. Outputs of these functions become inputs to the second layer and the same procedure takes place in the second layer. It is further repeated in successive layers until a global minimum on the error criterion is achieved (if the expected global minimum of error is not achieved, the heuristic specifications must be considered for alteration). The partial function that achieves the global minimum is considered as an optimal model under the given assumptions and specifications.

An example of the multilayered network structure with five input arguments and five selected nodes at each layer is presented in Fig.12-5.

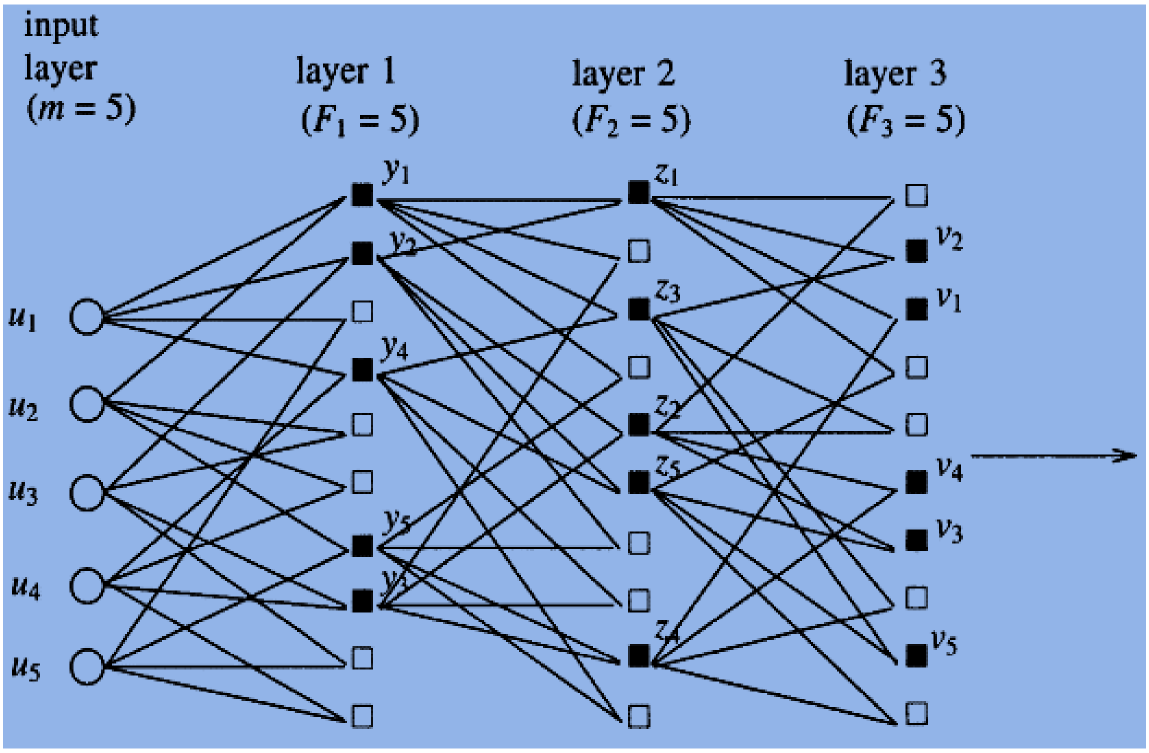


Fig.12-5 Multilayered network structure with five input arguments and selected nodes

[2] An artificial intelligence cognitive approach in which multiple connections between nodes (equivalent to brain cells) form a massive interactive network in which many processes take place simultaneously. Certain processes in this network, operating in parallel, are grouped together in hierarchies that bring about results such as thought or action.

Objective System Analysis (OSA) algorithm

Both algorithms described in the previous subsections are suitable for the ***equation-by-equation*** techniques. The key feature of ***OSA*** algorithm is that it examines systems of algebraic or difference equations, obtained by implicit templates (without goal function). An advantage of the algorithm is that the information embedded in the data sample is utilised better and we can estimate the relationships between variables.

In discrete mathematics, the term template refers to a graph indicating which of the delayed arguments are used in setting up conditional and normal Gauss equations. The key feature of the ***OSA*** algorithm, as presented in Fig. 12-6, is that it uses implicit templates, and an optimal model is therefore found as a system of algebraic or differential equations. An advantage of this algorithm is that the number of regressors is increased and as a consequence, the information embedded in the data sample is utilised better. A disadvantage is that it calls for a large amount of calculations in order to solve the system of equations and a greater number of candidate models have to be searched and evaluated. The amount of search can be reduced by using a constraint in the form of an auxiliary precision criterion.

In setting up the system of equations, one then discards the poorly forecasting equations for which the forecast *variation accuracy criterion* is less than the unity (narrowing operation):

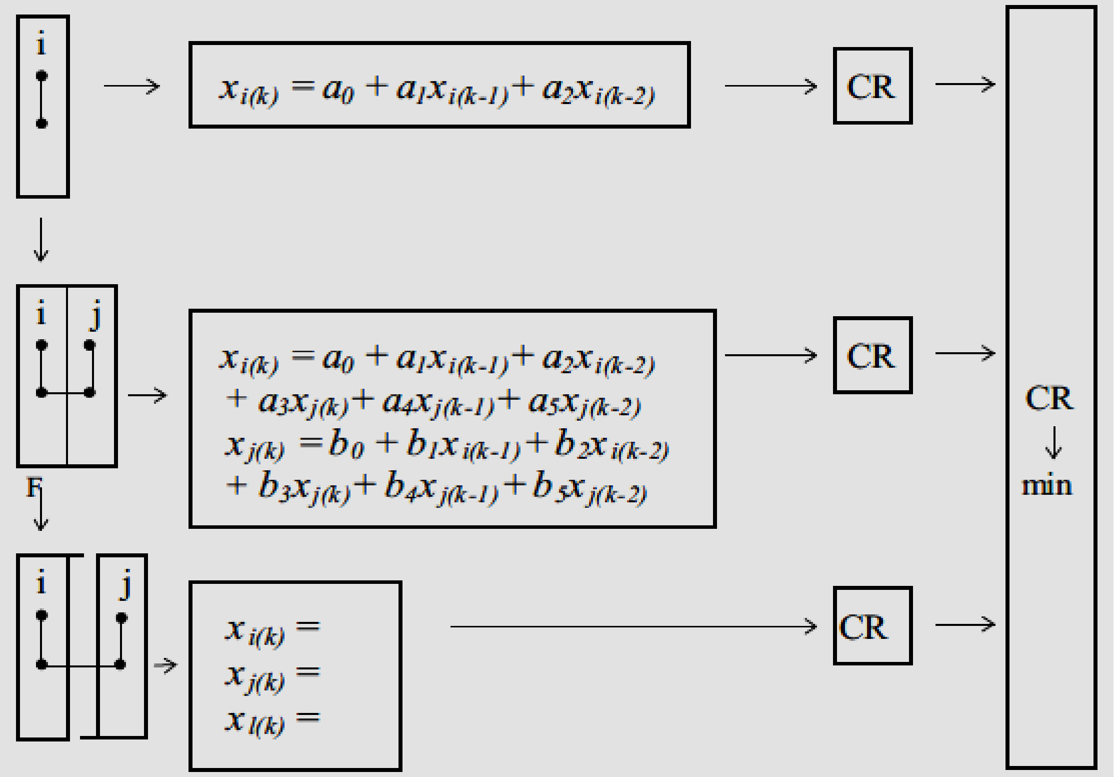

i
$x_{i(k)} = a_0 + a_1x_{i(k-1)} + a_2x_{i(k-2)}$
CR
i
j
$x_{i(k)} = a_0 + a_1x_{i(k-1)} + a_2x_{i(k-2)} + a_3x_{j(k)} + a_4x_{j(k-1)} + a_5x_{j(k-2)}$
$x_{j(k)} = b_0 + b_1x_{i(k-1)} + b_2x_{i(k-2)} + b_3x_{j(k)} + b_4x_{j(k-1)} + b_5x_{j(k-2)}$
CR
F
i
j
$x_{i(k)} =$
$x_{j(k)} =$
$x_{l(k)} =$
CR
CR
↓
min

Fig.12-6 Objective System Analysis (OSA) algorithm
(Source: Ivakhnenko & Müller, 1996, p.14)

$$\delta_i^2 = \frac{\sum_1^N (y_i - \hat{y}_i)^2}{\sum_1^N (y_i - \bar{y})^2} \rightarrow min \qquad (12\text{-}1)$$

where:

y_i - are the variable values in the table of data observations;
$\hat{y}_i$ - is the value calculated according to the model and
$\bar{y}$ is the mean value.

This criterion is recommended in the literature in order to evaluate the success of an approximation or of a forecast (Ivakhnenko & Müller, 1996, p.15). When $\delta^2 < 0.5$, the result of modeling is considered good, when $0.5 < \delta^2 < 0.8$ it is considered satisfactory, when $\delta^2 > 1.0$, modeling is considered to have failed, and the model yields misinformation.

OSA (also known as *Self-organization of systems of networks,* or *Autonomous systems of equations)* is the task of modeling complex systems which usually have more than one output variables. The goal of modeling systems of equations (***SE*** models) using *GMDH* is to self-organize a model for each output variable and to identify the *structural form* (9-19), i.e. the interdependence structure between the system variables separating them into endogenous and exogenous according to their corresponding model purpose. After modeling a system of *m* equations, ***OSA*** selects one (or a few alternative versions) best autonomous system consisting of *m** equations (*m** < *m*) necessary to describe the system completely. Here, both the number *m** of equations the best system consists of and its composition of variables are completely identified by the algorithm using the system error criterion.

All *m** variables of the so-identified best system are considered as endogenous variables of the ***SE*** system. All remaining variables which may be part of the ***SE*** model are either exogenous or are identified as exogenous due to an insufficient data basis. This could be a disadvantage sometimes, when the forecasting goal is based on a pre-specified structure (in terms of its endogenous variables) of the model. More details and other ***GMDH*** algorithms are presented in (Aksenova & Yurachkovsky, 1988; Ivakhnenko, Ivakhnenko, & Müller, 1994; Ivakhnenko & Müller, 1996) and others[3].

The main peculiarity of a GMDH algorithm is that, when it uses continuous data with noise, it selects as optimal the simplified non-physical model. Only for accurate and discrete data do the algorithms point out the so-called physical model which is the simplest optimal model from all unbiased ones.

The convergence of multilayered GMDH algorithms is proven and it is also proven that a shortened non-physical model is better than a full physical model (Aksenova & Yurachkovsky,

[3] See http://www.gmdh.net/articles/index.html

1988). For noisy and continuous data for prediction and approximation solving, more simplified Shannon's non-physical models are even more accurate. It can be noted, that this conclusion has place in model selection on the basis of model entropy maximization (Akaike's approach), in average risk minimization (Vapnik's approach) and in other modern approaches. The only way to get non-physical models is to use sorting-out GMDH algorithms. The regularity of the optimal structure of forecasting models changes according to the general indexes of data indeterminacy (such as noise level, data sample length, design of experiment, number of informational variables) as shown in (Ivakhnenko & Müller, 1996).

The special peculiarities of GMDH are as follow:

1) External supplement – following Beer (1959), only the external criteria, calculated on new independent information can produce the minimum of sorting-out characteristic. Because of this, data sampling is divided into parts for model training and testing (evaluation).

2) Freedom of choice – according to Gabor's work (1971), in multilayered GMDH algorithms not one but F (F>1) best results of choice are to be conveyed from one layer to the next layer to provide "freedom of choice".

3) The rule of layers complication – partial models (forms of a mathematical description for iteration) should be simple, without quadratic equations.

4) Additional model definition – in cases, when the choice of optimal physical model is difficult, because of noise level or oscillations of criterion minima characteristic, auxiliary discriminating criterion is used (Belogurov, 1990). The choice of the main criterion and constrains of sorting-out procedure is the main heuristic of GMDH.

5) All algorithms have multilayered structure and parallel computing can be implemented for their implementation.

6) All questions that arise about the type of algorithm, criterion, variables set, etc. should be addressed by the minimum of criterion value.

The main criteria used in GMDH are cross-validation, regularity and balance of variables (recall Chapter 3). Estimation of their effectiveness (investigation of noise immunity, optimality and adequateness) and their comparison with other criteria is done in detail in (Belogurov, 1990) and others (see Ivakhnenko & Müller, 1996).

The conditions, under which GMDH algorithm produces the minimum value are as follow:

a) criterion of model choice is to be external, based on additional fresh information, which was not used for model construction;

b) data sample cannot be too long. Such data sample produces the same form of characteristic as the exact data sample without noises;

c) when difference type balance criterion is used, either small noise is necessary or the variables in the data sample cannot not be exactly measured (Sawaragi et al., 1979).

What makes the GMDH algorithms different from other algorithms of structural identification, genetic and best regression selection algorithms, consists of the following main peculiarities (Ivakhnenko & Müller, 1996):

- *usage of external criteria*, which are based on data sample dividing (cross-validation) and are adequate to the problem of forecasting models building by reducing the requirements of the initial information volume;
- *much more diversity of structure generators usage* like in regression algorithms of the ways of full or reduced sorting of structure variants and of *original multilayered procedures*;
- *better level of automation* – it is only necessary to enter an initial data sample and the type of external criterion;
- *automatic adaptation* of optimal model complexity and external criteria to the level of noises or statistical violations – the effect of noise immunity causes robustness of the approach;
- *implementation of the principle of inconclusive decisions* in the process of gradual models complication.

It is important to point out the difference between the original ***GMDH algorithms*** and the "***algorithms of GMDH type***" (Farlow, 1984). The first ones work using the minimum of an external criterion, i.e. objective choice of the optimal model. The external criterion is calculated using a new data sample, which is not used for model training (cross-validation). To make an objective choice, selection is done without thresholds or coefficients in criterion. It is recommended to calculate the criteria two times: first to find the best models at each layer of selection for structure identification and second time to find the optimal model. Selection procedure is stopped when minimal criterion value is reached.

The second group is GMDH-type algorithms which could be expressed as "*the more complex the model – the more accurate it is*". For this purpose, it is necessary to set up definite threshold or to point out coefficients of weight for the members of the internal criterion formula, i.e. to find out the optimal model in a subjective way. Unfortunately, in almost all GMDH type software (like ModelQuest, NeuroShell and others) and research works in USA and Japan this deductive approach is used, which is not effective for short or noised data samples.

12.2. Artificial Neural Networks versus GMDH (Statistical Learning Networks)

Problems of complex systems modeling (functions approximation and extrapolation, identification, pattern recognition and forecasting of random processes) can be solved in general by deductive logical-mathematical or by inductive sorting-out (self-organization) methods. Deductive ones have advantages in the cases of rather simple modeling problems, when the theory of the object being modeled is known and therefore it is possible to develop a model from physically based principles employing the user's knowledge of the process.

It should be noted that self-organization does not replace a good domain theory. Inclusion of some well-known a priori information widens the basic scheme of self-organizing modeling by knowledge extraction from data and scientific theory (see Fig.12-7). However, self-organization very often provides the only way to get any knowledge from a complex system or to add some new aspects to existing theoretical fragments.

The inductive approach does not eliminate the experts or take them away from the computer, but rather assigns them a special position. Experts indicate the selection criterion of a very general form and interpret the chosen models. They can influence the result of modeling by formulating new criteria. The computer becomes an objective referee for scientific controversies, if criteria ensemble is coordinated among experts, who take part in the discussion.

Problems of complex objects modeling such as analysis and prediction of stock market, cannot be solved by deductive logical-mathematical methods with needed accuracy. In this case knowledge extraction from data, i.e. to derive a model from experimental measurements, has advantages in cases of rather complex objects, being only little a priori knowledge or no definite theory particularly for objects with fuzzy characteristics on hand.

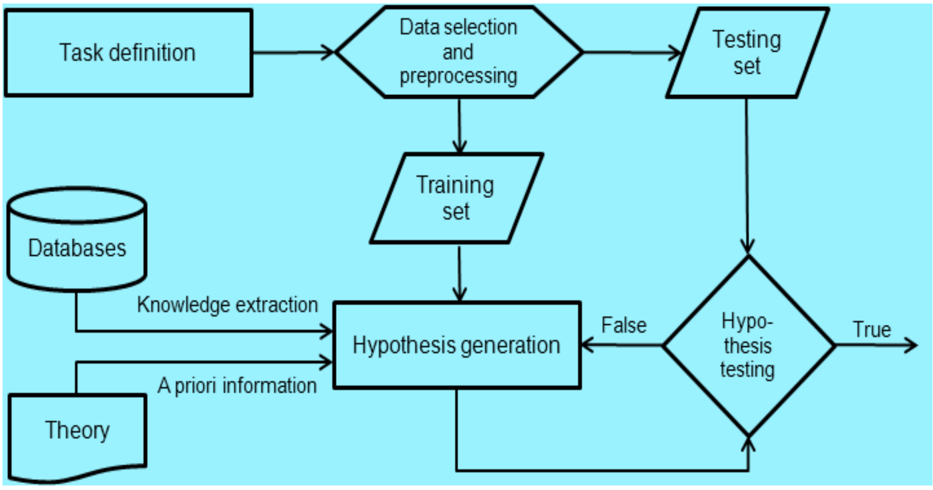


Fig. 12-7 Self-organizing modeling with a priori information

The task of knowledge extraction from data is to select mathematical description from data. However, the users cannot be in command of the required knowledge about mathematical models design and ***ANNs*** architecture. In mathematical statistics it is necessary to have a priori information about the structure of the mathematical model.

In ***ANNs*** the user estimates this structure by choosing the number of layers and the number and transfer functions of nodes of a neural network. This requires not only knowledge about the theory of ***ANNs***, but also knowledge of the object nature and time. Besides this the knowledge from systems theory about the systems modeled is not applicable without transformation in neural network world and the rules of translation are usually unknown.

GMDH type ANNs can overcome these problems - it can pick out knowledge about object directly from data sampling. The GMDH is an inductive sorting-out method, which has advantages in case either of rather complex objects and/or no definite theory. GMDH algorithms can find the only optimal model using full sorting-out of model-candidates and evaluating them by external criteria of accuracy.

It is obvious that in designing ANNs one cannot overcome the development of a human brain because it has an astronomical number of neurons (up to10^7 elements). Therefore, it is inevitable to use more complicated neurons, for example blocks acting according to the GMDH algorithms. The main difference between the GMDH neurons and the perceptrons is that, in the perceptron, the hidden elements of a second layer are not divided into clusters during training. In perceptrons, where there is no such a division of hidden elements, the information about the cluster on which the biggest signal is received is lost. The GMDH neurons can be united into repeatedly multilayered networks (see Fig.12-8) to increase the efficiency of solution of the artificial intelligence problems (accuracy and unbiasedness).

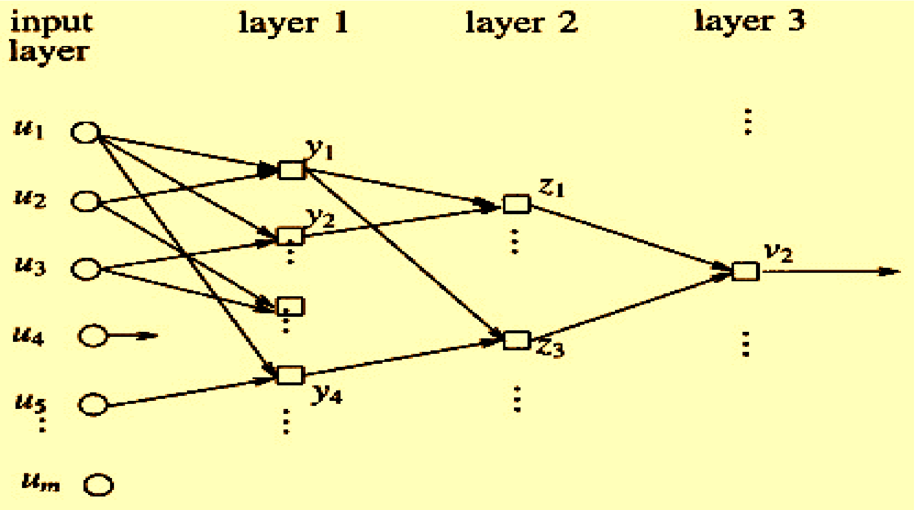


Fig.12-8 Example of *GMDH ANN* representing the output flow to unit two of layer three

Table 12.2. Neural networks versus Self-organizing modeling (Ivakhnenko & Müller, 1996).

	ANNs	**Statistical Learning Networks**
Data analysis	universal approximator	structure identificator
Analytical model	indirect by approximation	direct
Architecture	unbounded network structure; experimental selection of adequate architecture demands time and experience	bounded network structure [1]; adaptively synthesised structure
A-priori-Information	without transformation in the world of ANNs not usable	can be used directly to select the reference functions and criteria
Self-organisation	deductive, given number of layers and number of nodes (subjective choice)	inductive, number of layers and of nodes estimated by minimum of external criterion (objective choice)
Parameter estimation	in a recursive way; demands long samples	estimation on training set by means of maximum likelihood techniques, selection on testing set (extremely short)
Feature	result depends on initial solution, time-consuming technique, necessary knowledge about the theory of neural networks	existence of a model of optimal complexity, not time-consuming technique, necessary knowledge about the task (criteria) and class of system (linear, non-linear)

Table 12-2 presents a comparison of both methodologies, traditional ANNs and Statistical Learning Networks (i.e. Self-organizing modeling) in regards with their application to data analysis. Results obtained by statistical learning networks and especially GMDH algorithms are comparable with results obtained by ANNs (Müller & Lemke, 1995). In distinction to neural networks, the results of GMDH algorithms are explicit mathematical models obtained in a relative short time on the base of extremely short samples. The well-known problems of an optimal (subjective) choice of the neural network architecture are addressed in the GMDH algorithms by means of an adaptive synthesis (objective choice) of the architecture. GMDH algorithms could be used to estimate networks of the right size with a structure evolved during the estimation process to provide a parsimonious model for the particular desired function. Such algorithms, combining in a powerful way the best features of neural nets and statistical techniques, discover the entire model structure in the form of a network of polynomial functions, differential equations and others. Models are selected automatically based on their ability to solve the task such as approximation, identification, prediction, and classification.

12.3. Self-Organization of Nets of Active Neurons

The present stage of computer technology allows a new approach in ***ANNs***, which increases the accuracy of traditional modeling algorithms. Such approach can solve complex problems – we can use the GMDH algorithms as complex neurons, where the self-organization processes are well studied.

Only by this inductive self-organizing method for small, inaccurate or noisy data samples optimal non-physical model, accuracy of which is higher and structure is simpler than structure of usual full physical model can be found. GMDH algorithms are the examples of complex ***active neurons***, because they choose the effective inputs and corresponding coefficients of them by themselves, in process of self-organization.

The problem of neuronet links structure self-organization is solved in a rather simple way. Each neuron is an elementary system that handles the same task. The objective sought in combining many neurons into a network is to enhance the accuracy in achieving the assigned task through a better use of input data. As already noted, the function of active neurons can be performed by various recognition systems, notably by Rosenblatt's two-layer perceptrons – such neural network achieves the task of pattern recognition. In self-organization of ANNs the exhaustive search is first applied to determine the number of neuron layers and the sets of input and output variables for each neuron. The minimum of the discriminating criterion suggests the variables for which it is advantageous to build a neural network and how many neuron layers should be used. Thus, the theory of self-organizing ANNs is similar in many respects to that of active neurons.

A neural network is designed to handle a particular task. This may involve relation identification (approximation), pattern recognition, or a forecast of random processes and repetitive events from information contained in a sample of observations.

Active neurons are able, during the self-organizing process, to estimate which inputs are necessary to minimize the given objective function of the neuron. In a neuronet with such neurons, we shall have twofold multilayered structure: neurons themselves are multilayered, and they will be united into a multilayered network. They can provide generation of new features of special type (the outputs of neurons from previous layer) and the choice of effective set of factors at each layer of neurons. The output variables of previous layers are very effective secondary inputs for the neurons of next layer. First layer of ***active neurons*** acts similar to Kalman’s filter (Ivakhnenko & Müller, 1996) – output set of variables repeated the input set but with filtration of noises. The number of ***active neurons*** in each layer is equal to the number of variables given in the initial data sample.

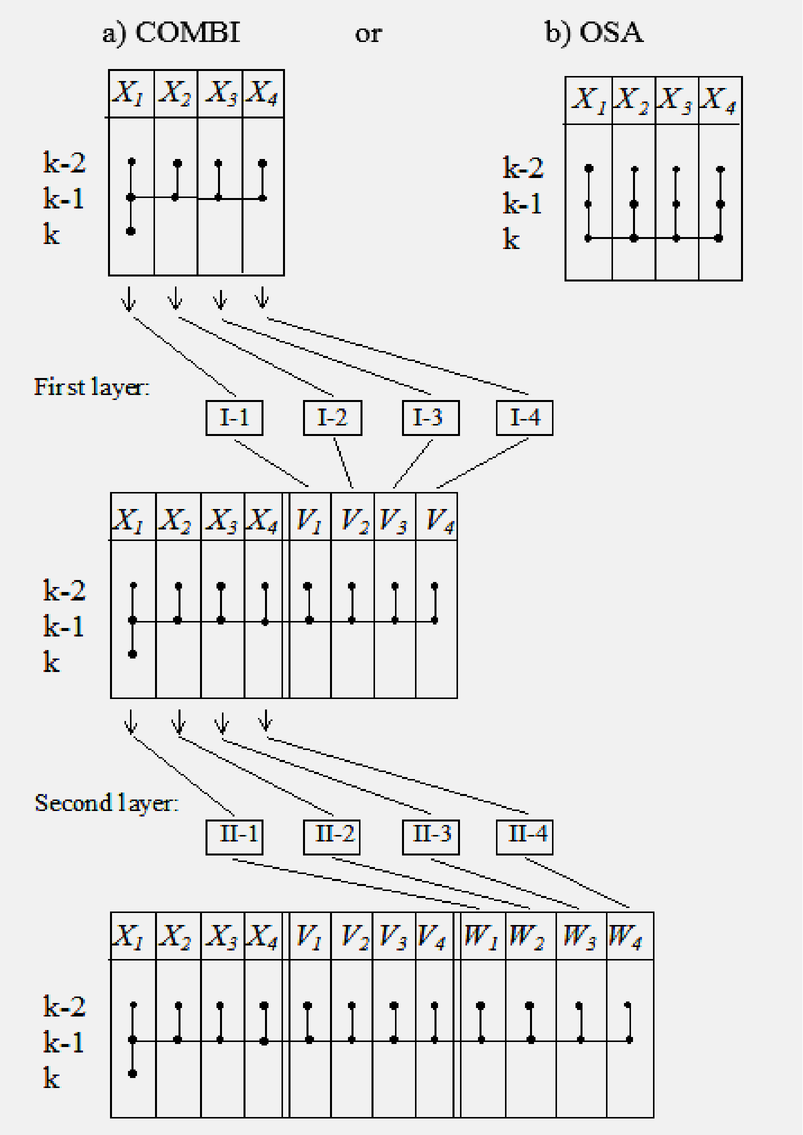


Fig.12-9 Schematic arrangement of the first two rows of a neural network
(Source: Ivakhnenko, Ivakhnenko, & Müller, 1994, p.15)

Neuronet structure is presented in Fig.12-9. Solely including the output calculated variables from each previous layer of neurons affects sample extension. The samples show the form of the discrete template used to teach the first neurons of a layer by the Combinatorial GMDH algorithm. The algorithm will identify which of the proposed arguments should be taken into consideration and will help to estimate the connectivity coefficients.

Initially, we construct the first layer of neurons in the network. We will then be able to determine how accurate the forecast will be for all variables. For this purpose, we use a discrete template that allows a delay of one or two periods for all variables. Then we add a second (third, fourth etc.) layer to the neural network, as shown in Fig.12-9. Further on, we continue doing so as long as it improves the forecast or decrease external criterion value.

For each neuron, the extended definition procedure to one model (out of the five closest to the optimal one) is applied. For optimal models the forecast variation criterion is calculated. It may be inferred, that there is no need to construct a neural network in order to form a forecast for those variables, for which variation criterion value takes on the least value in the first layer. It is advisable to use a neural network to form a forecast for the variables, for which the variation criterion takes on the least value in the last layers of neurons.

The equations for the neurons of the network define the connections that must be implemented in the neural network; in this way they help achieve the task of structural self-organization of the neural network. For brevity, the data sample in the above example is extended in only one way: tile output variables of the first layer are passed on as additional variables to the second, third, etc. layer of neurons. It is possible to compare different schemes of data sample extension by external criterion value.

The task for self-organization of such networks of ***active neurons*** by selection is to estimate the number of layers of ***active neurons*** and the set of possible potential inputs and outputs of every neuron. The sorting characteristic ‘number of neuronet layers - variables, given in data sample’ defines the optimum number of layers for each variable separately. Neuronets with ***active neurons*** should be applied to raise the accuracy of short-term and long-term forecasts.

In self-organization, the layers of neurons are extended as long as this improves the accuracy of the solution yielded by the neural network. Self-organization of each neuron taken separately uses the differential balance criterion or the regularity precision criterion. Since the exhaustive-search curve approaches its minimum in a gradual manner the criteria of models close to the optimal one differ only slightly from one another in value. This is why an extended definition algorithm should be used, because instead of one, several of the best models are selected. From them, only one that complies with another variation discriminating criterion is chosen.

The final choice of the “best” model is made by the researcher who has one final option to apply additional, sometimes qualitative information or knowledge, but after having the guarantee that a big number of possible models have been evaluated and the final choice is based on a small number of good ones. Model outputs must always be evaluated by the researcher to figure out whether new and useful knowledge of the domain has been discovered. Models provide data, but the real-life business needs information, i.e. data in the business context. The extracted information is valuable to a business only when it leads to actions that create value or market behavior that gives a competitive advantage.

12.4. Self-Organizing Data Mining Platforms

Data mining techniques require powerful software, designed and elaborated for this specific aim. Developing your own computer program in this area is a big project, which takes a lot of resources, time and highly qualified researchers. To implement algorithms, similar to the ***Multi-Layered Net of Active Neurons (MLNAN)*** described in the previous chapters, a prototype of an information system was developed long time ago (see Motzev & Marchev, 1988; Motzev & Marchev, 1989). Unfortunately, it is too large and too complex — its' total volume is about 30 thousand program lines in the PL/1 language. It was designed for an IBM 4331 computer under the VM-370 operating system. At present this system has a multitude of abilities and parameters which are assigned in interactive mode which in fact often hampers more than aids the unprepared user. Moreover, it ties him to an outdated operating system and an expensive computer.

One possible solution in this case is to unify an algorithm like ***MLNAN*** with existing Data mining software. One of the leading software platforms in self-organizing data mining[4] is ***KnowledgeMiner*** program. ***KnowledgeMiner*** is a self-organizing tool for modeling and predictions, which implements GMDH, Analog Complexion (***AC***) and Fuzzy Rule Induction.

The GMDH implementation employs active neurons and thus provides networks of active neurons at the lowest possible level. It can be used to create linear & nonlinear, static & dynamic time series models, multi-input/single-output and multi-input/multi-output models as systems of equations even from small and noisy data samples. The model outputs are presented analytically (as equations with estimated coefficients) and the systems of equations are presented graphically as well, by a system graph reflecting the interdependent structure of the system (see equations 9.17 to 9.20).

In case of fuzzy objects modeling & prediction, different procedures can be used. In Analog Complexion a model consists of a composition of similar patterns. Since several most similar patterns are always used to form a model and a prediction by a synthesis, a confidence interval is produced too to present the uncertainty in predictions. This is of special importance when using predictions for decision support.

Fuzzy Rule Induction (***FRI***) is used to create fuzzy or logical rules or systems of rules from fuzzy or Boolean data. Complex systems are described & presented qualitatively and the models generated are usually easy to interpret.

It should be mentioned that the MLNAN algorithm is more flexible and allows, in case of

[4] http://www.knowledgeminer.eu/

conflict, researcher intervention to prevent that some of the endogenous variables are transformed into exogenous, which will reduce the model and the information at the output.

Example: Using KnowledgeMiner software to build a forecasting model

A short example of building an input-output forecasting model with KnowledgeMiner will illustrate the overall process more clearly. Assume that a data set of a national economy is available and our goal is to use some of the variables from it to model one output variable. This variable for example could be:

$\mathbf{x_{1,t} = y_t}$ is the U.S. Gross National Product

The potential input variables, which could be used, are:

$\mathbf{x_{2,t}}$ – U.S. Income Receipts

$\mathbf{x_{3,t}}$ – U.S. Personal Income

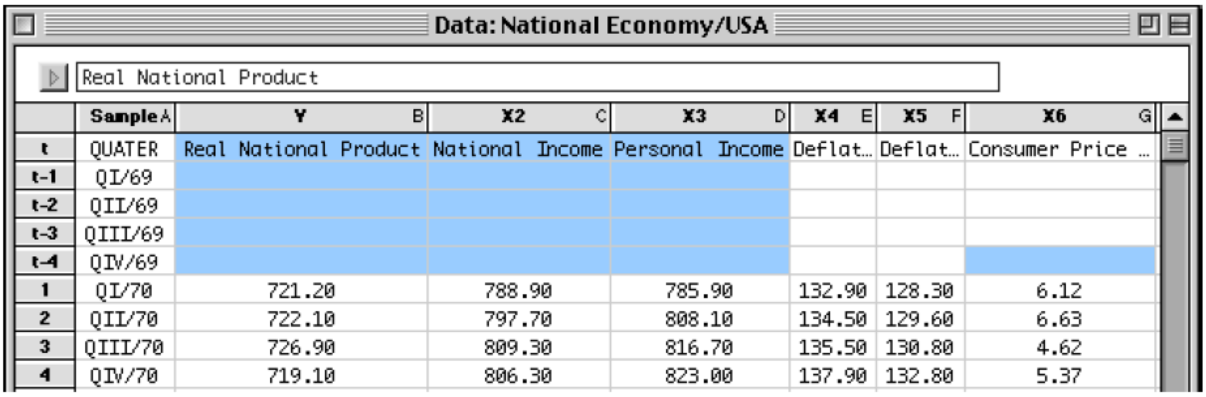
and their lagged values

$\mathbf{x_{j,t-k}}$ – j = 2, 3; k=1, 2, 3 &4

There is also a variable Consumer Price Index with a time lag of 4 only ($\mathbf{x_{6,t-4}}$).

Fig.12-10 shows how these input & output variables could be identified in a spreadsheet. First, by clicking in the first row of column X_1 the output variable is defined. The header of the column changes from X_1 to Y. Then, the input variables are identified by selecting the corresponding cells – columns indicate a variable number and rows indicate their time lag accordingly. Irregular definitions (if any) would be detected and corrected later on during the model building process.

In this way, any combination of input & output variables can be selected for a given data set without changing & modifying the initial data table manually. After that KnowledgeMiner constructs the information matrix automatically.



Data: National Economy/USA

Real National Product

	Sample A	Y B	X2 C	X3 D	X4 E	X5 F	X6 G
t	QUATER	Real National Product	National Income	Personal Income	Deflat…	Deflat…	Consumer Price …
t-1	QI/69						
t-2	QII/69						
t-3	QIII/69						
t-4	QIV/69						
1	QI/70	721.20	788.90	785.90	132.90	128.30	6.12
2	QII/70	722.10	797.70	808.10	134.50	129.60	6.63
3	QIII/70	726.90	809.30	816.70	135.50	130.80	4.62
4	QIV/70	719.10	806.30	823.00	137.90	132.80	5.37

Fig.12-10 Input-output variables and their time lags defined easy by selecting corresponding cells in the spreadsheet
(Source: Müller & Lemke, 1995, p.155)

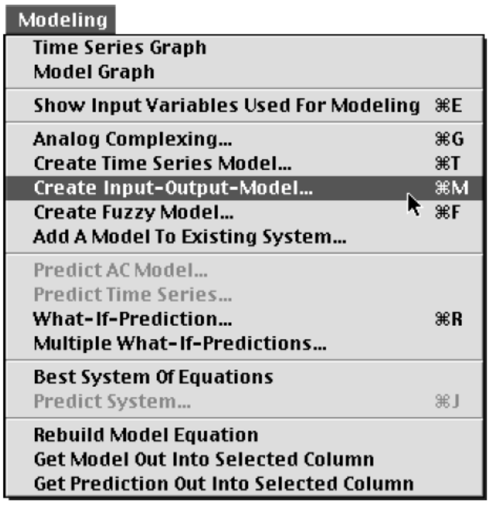
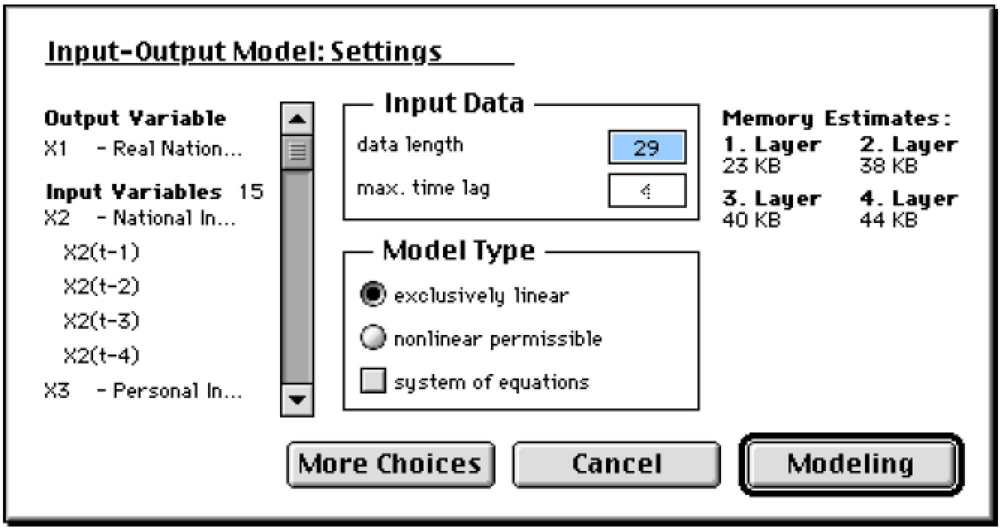

Modeling
Time Series Graph
Model Graph
Show Input Variables Used For Modeling ⌘E
Analog Complexing... ⌘G
Create Time Series Model... ⌘T
Create Input-Output-Model... ⌘M
Create Fuzzy Model... ⌘F
Add A Model To Existing System...
Predict AC Model...
Predict Time Series...
What-If-Prediction... ⌘R
Multiple What-If-Predictions...
Best System Of Equations
Predict System... ⌘J
Rebuild Model Equation
Get Model Out Into Selected Column
Get Prediction Out Into Selected Column

Fig.12-11 Modeling Menu of KnowledgeMiner
(Source: Müller & Lemke, 1995, p.156)

The Modeling Menu opens by selecting “Create Input-Output Model” (see Fig.12-11) and a dialog window appears as shown in Fig.12-12.

At the left side of the window, the output variable and all selected input variables are listed and can be reviewed. Then, it’s necessary to define the time horizon one would like to use and the maximum time lag for dynamic systems (or it’s just zero for static systems). If time lags have been already defined at the stage of variables selection (as shown in Fig.12-10 for this example) they don’t need to be redefined in this window. Next, the general model type should be selected – linear or nonlinear. When nonlinear is chosen KnowledgeMiner will not necessarily create a nonlinear model due to active neurons. If the detected optimal model is linear, it will also be selected as the best model.



Input-Output Model: Settings

Output Variable
X1 - Real Nation...

Input Variables 15
X2 - National In...
X2(t-1)
X2(t-2)
X2(t-3)
X2(t-4)
X3 - Personal In...

Input Data
data length 29
max. time lag 4

Model Type
● exclusively linear
○ nonlinear permissible
□ system of equations

Memory Estimates:

1. Layer 23 KB	2. Layer 38 KB
3. Layer 40 KB	4. Layer 44 KB

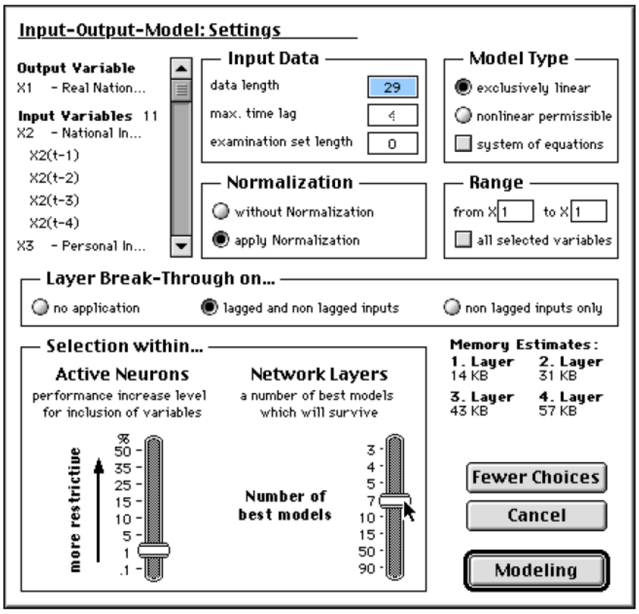
Fig.12-12 Dialog window for setting up the modeling process
(Source: Müller & Lemke, 1995, p.156)

Finally, it must be decided whether a system of equations should be generated or not. A system of equations would consist of all considered variables as output variables (here these are $\mathbf{x_{1,t}}$, $\mathbf{x_{2,t}}$, $\mathbf{x_{3,t}}$ and $\mathbf{x_{6,t}}$) and input variables accordingly (including time lags). When clicking on "Memory" option, the approximately required memory of the first four network layers will be listed for the chosen settings. To start the model building process one should click on the "Modeling" tab.

Alternatively, an expanded dialog window (Fig.12-13) can be used when clicking on "More Options" button.

Here, a third data subset, the examination set can be identified if a greater length than zero is entered. This data set is used for true out-of-sample model performance testing during model building. It is done on data not used yet for both learning and/or testing of created potential model candidates. The performance measure is referenced later on when selecting the best model candidates within a layer as another discriminating criterion. The examination data set should always be selected from the last data observations while reducing the available data sets for learning and testing models. Therefore, the examination set length should be very small.

Next, it is possible to set up a range of variables within the chosen settings, which will be applied automatically. In this example, the choice "from X1 to X6" would generate input-output models for the variables $\mathbf{x_{1,t}}$, $\mathbf{x_{2,t}}$, $\mathbf{x_{3,t}}$ and $\mathbf{x_{6,t}}$ sequentially on corresponding set of input & output variables. In such case, simply checking the "all selected variables" control would have the same effect. Usually, it should also be checked when creating systems of equations in a single modeling run.



Input-Output-Model: Settings

Output Variable
X1 - Real Nation...

Input Variables 11
X2 - National In...
X2(t-1)
X2(t-2)
X2(t-3)
X2(t-4)
X3 - Personal In...

Input Data
data length: 29
max. time lag: 4
examination set length: 0

Model Type
● exclusively linear
○ nonlinear permissible
□ system of equations

Normalization
○ without Normalization
● apply Normalization

Range
from X 1 to X 1
□ all selected variables

Layer Break-Through on...
○ no application
● lagged and non lagged inputs
○ non lagged inputs only

Selection within...

Active Neurons
performance increase level for inclusion of variables

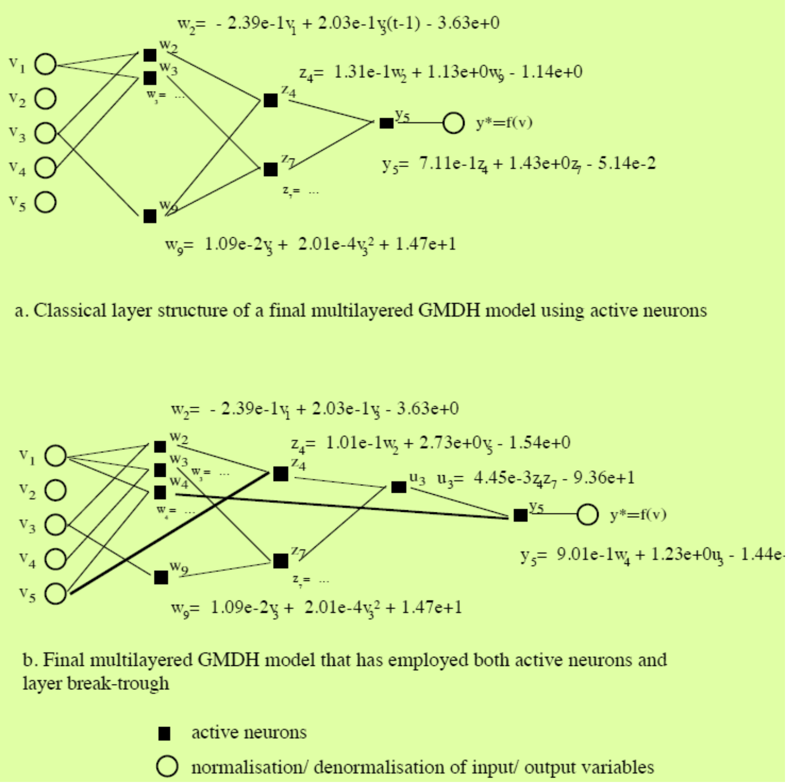
Network Layers
a number of best models which will survive

Memory Estimates:
1. Layer 14 KB
2. Layer 31 KB
3. Layer 43 KB
4. Layer 57 KB

Fig.12-13 Expanded dialog window for setting up the modeling process
(Source: Müller & Lemke, 1995, p.156)

The "Layer Break-Through on…" option could be used to define the acceptable degree of freedom for network structure synthesis. If "no application" is chosen the network will evolve in the classical way (see Fig.12-14.a). Otherwise, either lagged or non-lagged inputs or non-lagged inputs will only be provided for all layers of network synthesis (Fig. 12-14.b). Note that using "Layer Break-Through on…" option significantly increases memory requirements and computing time, but also may improve significantly modeling results. Therefore, it is the recommended choice.

Finally, there are two slider controls (like scroll boxes), which adjust the selection of best models candidates at different levels. The left one controls self-organization of active neurons by defining a threshold value of a performance gain, which a created model candidate must pose after validation (compared with a previously selected intermediate best model) to become the new best reference model for that neuron. The greater this threshold value is the more restrictive active neuron self-organization will be, and its transfer function will be composed of the most significant input variables and terms. Usually, thresholds between 1% (linear models) and 15% (nonlinear models) are good.


$w_2 = -2.39e\text{-}1v_1 + 2.03e\text{-}1v_3(t\text{-}1) - 3.63e{+}0$
v_1
v_2
v_3
v_4
v_5
w_2
w_3
$w_3 = \ldots$
w_9
$z_4 = 1.31e\text{-}1w_2 + 1.13e{+}0w_9 - 1.14e{+}0$
z_4
z_7
$z_7 = \ldots$
y_5
$y^* = f(v)$
$y_5 = 7.11e\text{-}1z_4 + 1.43e{+}0z_7 - 5.14e\text{-}2$
$w_9 = 1.09e\text{-}2v_3 + 2.01e\text{-}4v_3^2 + 1.47e{+}1$

a. Classical layer structure of a final multilayered GMDH model using active neurons

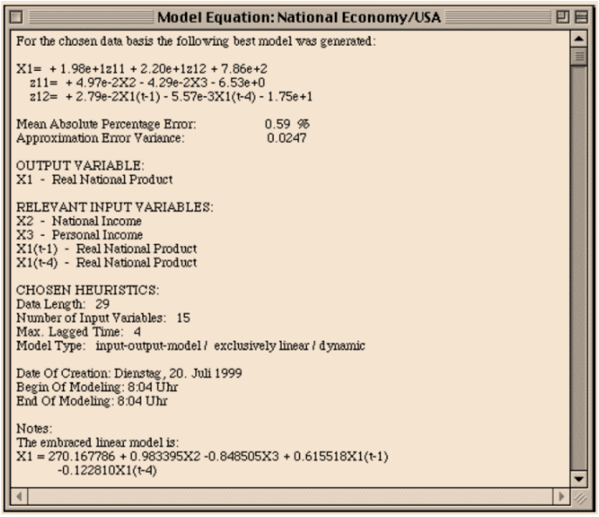
$w_2 = -2.39e\text{-}1v_1 + 2.03e\text{-}1v_3 - 3.63e{+}0$
v_1
v_2
v_3
v_4
v_5
w_2
w_3
$w_3 = \ldots$
w_4
$w_4 = \ldots$
w_9
$z_4 = 1.01e\text{-}1w_2 + 2.73e{+}0v_5 - 1.54e{+}0$
z_4
z_7
$z_7 = \ldots$
u_3 $u_3 = 4.45e\text{-}3z_4z_7 - 9.36e{+}1$
y_5
$y^* = f(v)$
$y_5 = 9.01e\text{-}1w_4 + 1.23e{+}0u_3 - 1.44e$
$w_9 = 1.09e\text{-}2v_3 + 2.01e\text{-}4v_3^2 + 1.47e{+}1$

b. Final multilayered GMDH model that has employed both active neurons and layer break-trough

■ active neurons

○ normalisation/ denormalisation of input/ output variables

Fig.12-14 Classical GMDH network structure and layer break-through structure (Source: Müller & Lemke, 1995, p.148)



```
For the chosen data basis the following best model was generated:

X1=  + 1.98e+1z11  + 2.20e+1z12  + 7.86e+2
  z11=  + 4.97e-2X2 - 4.29e-2X3 - 6.53e+0
  z12=  + 2.79e-2X1(t-1) - 5.57e-3X1(t-4) - 1.75e+1

Mean Absolute Percentage Error:          0.59  %
Approximation Error Variance:             0.0247

OUTPUT VARIABLE:
X1  -  Real National Product

RELEVANT INPUT VARIABLES:
X2  -  National Income
X3  -  Personal Income
X1(t-1)  -  Real National Product
X1(t-4)  -  Real National Product

CHOSEN HEURISTICS:
Data Length:   29
Number of Input Variables:   15
Max. Lagged Time:   4
Model Type:   input-output-model /  exclusively linear / dynamic

Date Of Creation: Dienstag, 20. Juli 1999
Begin Of Modeling: 8:04 Uhr
End Of Modeling: 8:04 Uhr

Notes:
The embraced linear model is:
X1 = 270.167786 + 0.983395X2 -0.848505X3 + 0.615518X1(t-1)
        -0.122810X1(t-4)
```

Fig.12-15 Model equation and Report for the model created
(Source: Müller & Lemke, 1995, p.158)

The right slider should be used to define an appropriate number of best models of a layer which will survive as inputs for the following layer(s). This choice dramatically influences memory requirements and an appropriate number means to find a compromise between memory and computing time and a comfortable freedom of choice. Here, a small number of models are already good, since the best models of all layers (plus the initial input variables) are also used at any layer.

The self-organization can begin again by clicking on the "Modeling" button. When the modeling process is done, the created models are presented both graphically and analytically (see Fig.12-15). Along with the model equations other important information (model error for example) is reported in the same window. The models are added to the model base and after that could be used for predictions.

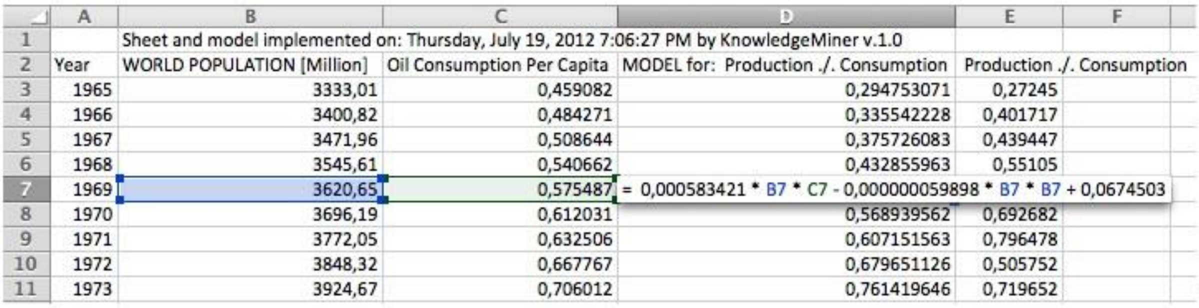
For more than 20 years KnowledgeMiner Software creates knowledge mining tools that enable anyone to use data modeling to quickly visualize new possibilities. It has been doing researching and consulting in many fields[5]. It has been used very intensely in modeling and prediction of toxicological and eco-toxicological hazards and risks of chemical compounds for regulatory purposes with the goal to substitute and minimize the still widely used animal tests by computer models. Another project which exclusively and extensively uses KnowledgeMiner Software tools is focused on climate change modeling and prediction. Sales and demand predictions, macro- and micro-economic modeling, budget and resource planning, energy consumption analysis and prediction, medical diagnosis, traffic prediction, etc. are further examples where KnowledgeMiner tools have been applied to.

[5] See http://knowledgeminer.eu/solutions.html)

The main advantages of the inductive ***KnowledgeMiner*** approach are:

- Only minimal (may be uncertain a priori) information about the system is required, i.e. even if the user has no experience in modeling, data analysis or designing a neural network he/she will be able to build a model, analyze and predict complex relationships in almost any kind of system.
- A very fast and effective learning process for a personal computer. That means the user can solve problems on his/her desktop in a reasonable time, which one may have never thought as possible.
- Modeling short and noisy data samples, i.e. a user can deal with a problem as it is and does not have to construct artificial conditions for the modeling method to get it work.
- Output of an optimally complex model. The user can expect to get a model at the end of the automated modeling process, which must not be overfitted. Overfitted models are not able to predict inherent relationships between variables.
- Output of an analytical model as a transparent explanation component (i.e. a white box). In this way, immediately after model building, the user can evaluate the synthesized model analytically and use it to explain the relationships in real system.

KnowledgeMiner (yX) for Excel is a knowledge mining tool that works with data stored in MS Excel for building predictive and descriptive models from this data autonomously and easily (see Fig.12-16). It supports all major releases of Microsoft Excel for Mac computers. The modeling engine of ***KnowledgeMiner (yX) for Excel*** implements unique modeling technologies which are built on the principles of the self-organization, i.e. learning from noisy data an unknown relationship between output and input of any given system in an evolutionary way from a very simple model to an optimal complex one which generalizes well.



	A	B	C	D	E	F
1		Sheet and model implemented on: Thursday, July 19, 2012 7:06:27 PM by KnowledgeMiner v.1.0				
2	Year	WORLD POPULATION [Million]	Oil Consumption Per Capita	MODEL for: Production ./. Consumption	Production ./. Consumption	
3	1965	3333,01	0,459082	0,294753071	0,27245	
4	1966	3400,82	0,484271	0,335542228	0,401717	
5	1967	3471,96	0,508644	0,375726083	0,439447	
6	1968	3545,61	0,540662	0,432855963	0,55105	
7	1969	3620,65	0,575487	= 0,000583421 * B7 * C7 - 0,000000059898 * B7 * B7 + 0,0674503		
8	1970	3696,19	0,612031	0,568939562	0,692682	
9	1971	3772,05	0,632506	0,607151563	0,796478	
10	1972	3848,32	0,667767	0,679651126	0,505752	
11	1973	3924,67	0,706012	0,761419646	0,719652	

Fig.12-16 *Example:* Implemented single model in Excel.
(Source: http://knowledgeminer.eu/about.html)

The main characteristic of ***KnowledgeMiner (yX) for Excel*** could be summarized as follow:

- Unique self-organizing modeling technologies implemented in the cross-platform (yX) modeling engine to make modeling most objective and easy-to-use.
- An analytical model that describes the data is produced and available to the user for implementing directly in Excel.
- High-dimensional modeling which scales to data sets of very small (< 20) to large number of samples and from small to large number of potential inputs (<= 50,000) and no any prior variables selection is necessary. Table 12.3 shows the ranges within it is which designed to work on as compared to other common data mining methods.
- Original approaches to noise immunity and evaluation of models derived from data to improve reliability of models and to minimize the risk of getting invalid models from badly sized data sets. This point needs some additional explanations:
- A new feature implemented in KnowledgeMiner (yX) for Excel is on-the-fly evaluation of self-organized regression models based on the concept of noise immunity. Moreover, a new model quality measure - Descriptive Power - is introduced, which takes into account the risk of obtaining a chance model from, only, a given data set.
- A key problem in knowledge discovery from data is final evaluation of generated models. To know whether the obtained model is likely to adequately reflect an input-output relationship that exists in reality or if it is just a chance model with non-causal correlations is essential for applying models obtained by data mining in real systems. However, it is not possible to get this information only out of the modeling algorithm the model was generated with. Some new, external information is required as described above.
- **64-bit parallel software** – ultra-fast processing of very computation intensive algorithms by 64-bit, multi-core/multi-processor, and vector processing support. The software automatically scales to the number of CPU-cores found at runtime to take full advantage of current and future multi-core hardware. The basic rule: Modeling speed grows with every core and/or processor available at runtime.

Insights, the newest version of ***KnowledgeMiner for Excel***, implements vector processing (single instruction - multiple data parallelism (SIMD)), multi-core and multi-processor support (multiple instruction - multiple data parallelism (MIMD)) for high-performance computing. It scales automatically to the hardware of the computer to take full advantage of multi-processor and multi-core based Mac computers.

Table 12.3 Common modeling technologies and their applicability to different data set dimensions (Source: http://knowledgeminer.eu/about.html)

Sample Size n	Small number of inputs m (m < 50)	Medium number of inputs m (49 < m < 500)	Large number of inputs m (499 < m < 50K)
Very Small (n < 30)	I, [R]	I	I
Small (29 < n < 200)	I, [R, ML]	I, [R]	I
Medium (199 < n < 10K)	I, R, ML	I, [R, ML]	I
Large (10K < n < 1M)	I, R, ML	I, [R, ML]	I

Where:

- ML - Neural Networks, Support Vector Machines and other Machine Learning methods.
- R - Statistical regression methods.
- (yX) - Self-organizing high-dimensional modeling with KnowledgeMiner Insights.
- [.] - Can be applied under certain conditions only.

A major concern for models built from any data-driven modeling technology is model reliability, and the risk of obtaining an invalid model grows fast with increasing number of input variables and with increasing model complexity. Therefore, original research into noise immunity of high-dimensional state space modeling **(**Lemke, 2008) has been performed by KnowledgeMiner Software for many years. This research resulted in highly improved noise immunity algorithms compared to traditional modeling approaches. These new algorithms have been implemented in KnowledgeMiner for the first time. They additionally allow a new model evaluation approach, which helps and supports the user in a powerful way to assess obtained models.

Based on its independent, cross-platform core modeling engine, self-organizing, parallel, high-dimensional modeling with on-the-fly model evaluation, KnowledgeMiner provides a unique and original tool. It also allows non-experts to build reliable predictive analytical models of complex systems from the desktop with unprecedented power, ease-of-use, and ease-of-applicability likewise.

12.5. Forecasting Applications of Self-Organizing Data Mining

A framework for Complex Model Building using Self-Organizing Data Mining based on the ***MLNAN*** algorithm was designed in (Motzev & Marchev, 1988). There are three main parts developed for synthesizing simulation models in the form of a system of simultaneous equations. The general structure of the procedure is presented in Fig.12-17.

In the First part, on the basis of existing real data, grouped in a table of observations of the researched system, a researcher makes the first choice of significant factors, which will be included in the model.

The Second part is the ***MLNAN*** procedure for multi-stage selection of each equation (like 9-17) of the model. The form of the equation is not predetermined. The task of finding the form and coefficients of each equation is divided in many simple tasks for identification of coefficients of equations with two variables. This approach allows computing statistically significant coefficients even from a limited number of observations (small size samples). The form of the equation is determined in a few consecutive stages of selection. Based on a given number of selections, an average minimum error for the generation or others, the selection procedure ends when certain conditions or results, which the researcher wants to obtain, are achieved.

At the end of the Second part the full form of the best equations is automatically restored and for each equation (describing a given connection in the model) a certain number of alternative versions are saved.

In the Third part of the procedure, on the basis of the results from Part 2, a synthesis of the full model of simultaneous equations (9-19) is done. The variety of alternative versions of the model is generated by combining the best equations already chosen from Part 2. Each of the competing hypotheses represents a hypothesis for inclusion in the model of a given variant of the separate equation. Each generated system of equations is considered as a potential model of the analyzed system. The evaluation of these competing models is done using many statistical criteria – coefficient of correlation, coefficient of variation, MAPE, forecasting error and others.

The final choice of the best model is made by the researcher, who has the opportunity to use some additional insights, but after evaluation of a large number of possible models and selection of a small number of good ones through the Automated Procedure.

The Prototype developed initially contains few interrelated computer programs (see Motzev & Marchev, 1989):

- "ANALYSIS" — used for preliminary analysis of variables' dynamics in the model and relationships between them.
- "SELECTION" — used for synthesizing the equations in the designed model through multi-stage selection procedure.
- "TUNNING" — used for automatic generation & identification of unknown coefficients in each equation and for automatic selection of equation modules in the model through competition between different versions of the model.
- "SIMUL" — used for complex evaluation of the synthesized model adequacy, using a certain number of predefined criteria as well as for conducting different simulation experiments with the model.

As mentioned in section 12.3 above, this Prototype was too large and too complex, and developed for an outdated computer system IBM 4331 under the VM-370 operating system. In a similar automated procedure all programs but SIMUL could be substituted with the KnowledgeMiner for Excel software (see Motzev, 2018), described in the two previous sections.

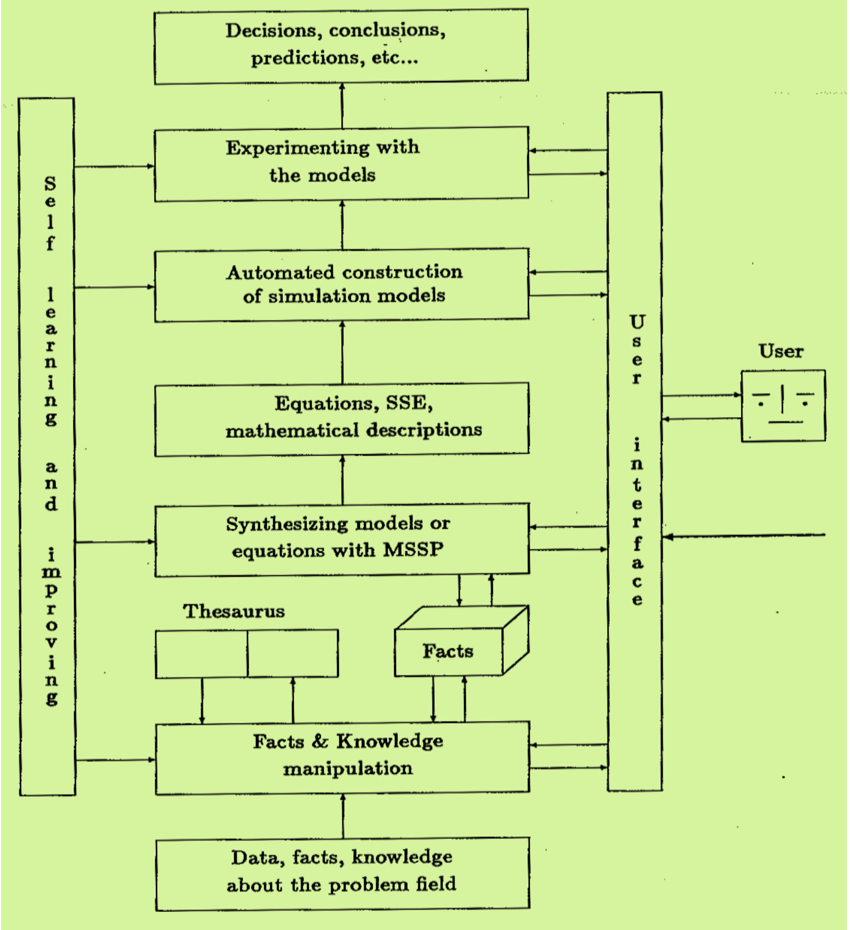

Decisions, conclusions, predictions, etc...
Experimenting with the models
Automated construction of simulation models
Equations, SSE, mathematical descriptions
Synthesizing models or equations with MSSP
Thesaurus
Facts
Facts & Knowledge manipulation
Data, facts, knowledge about the problem field
Self learning and improving
User interface
User

Fig.12-17 General structure (bottom-up) of a framework using ***MLNAN***

It should be mentioned that the program “SIMUL” (original software designed and developed exclusively by the author of this text) provides some tools not covered by KnowledgeMiner for Excel. The latter has an excellent module for complex evaluation of the synthesized model, its adequacy and reliability, but “SIMUL” provides more options for conducting different simulation experiments with the model. Updating this program and making it compatible with KnowledgeMiner software could be a useful project in the future.

The procedure described above has a large field of applications. In general, they could be summarized in two major groups. The first one is model building for analysis and predictions (Motzev, 2010) and the second one is model building for model-based business games (Motzev, 2011). As mentioned before, such technique provides opportunities for shortening the time, cost and efforts for model development. Moreover, the results obtained in these studies confirmed that GMDH based self-organizing data mining is able to reliably develop even complex models with lower overall error rates than other methods.

The procedure was tested both for developing simultaneous equations (SE) models like (9-19) and for synthesizing individual equations (9-18), like Autoregressive models for example, when running only Parts 1 & 2 of the procedure.

To evaluate the quality of performance of Self-Organizing Data Mining, two types of comparisons were made: **Type A** – between traditional models and models developed with the MLNAN; **Type B** – between models developed by different Self-Organizing Data Mining techniques, KnowledgeMiner and the MLNAN original prototype.

Comparison Type A was already discussed in detail in Chapter 9.3 using a small model (SIMUR I) of the Bulgarian economy (Marchev & Motzev, 1985). Indirect OLS was used to estimate unknown coefficients in equations (9-17) and the model accuracy, measured by mean squared error relative to the mean (see Chapter 3) **CV(RMSE)** = 14%.

Same data and set of variables were used to build a new model, using the MLNAN algorithm and the automated procedure, described above. The new model accuracy (measured with the same statistics was improved to **CV(RMSE)** = 3.81%. The brief comparison shows that the new model has much better accuracy (almost four times smaller) and thus provides more reliable base for simulations and further analysis of the system of interest.

Another significant improvement of the ***ex-ante*** predictions was done by estimating equations in the ***SE*** form with ***dynamic*** (i.e. ***non-stationary***) ***coefficients***. In this case **MAPE** values are about three times less in the non-stationary equations (on average **MAPE** changes from 6.73% in the stationary model to 2.26% in the non-stationary ***SE***) and **CV**(**RMSE**) (the coefficient of variation of the square root of calculated **MSE**) changes from 7.05% to 2.44%.

It is worth noting that the ***MLNAN*** algorithm was used successfully in time-series analysis to build different complex autoregressive models like Distributed lag models, Autoregressive-moving-average with exogenous inputs models (***ARMAX***), Vector autoregression models (***VAR***) and others. As mentioned in Chapter 9.3, during the ***SE*** model building and its improvements the ***MLNAN*** algorithm was used for estimating ***AR*** equations of more than 20 macroeconomic variables in Bulgaria with a time lag of up to 5 years (Motzev, Muller, & Marchev, 1986), providing in most cases **MAPE**% less than 1.5%, adjusted coefficient of determination (**R^2**) greater than 0.9 and average **CV(RMSE)**% for all equations 4.74%.

Two important conclusions can be made from these examples:

(a) The ***MLNAN*** algorithm, as other ***GMDH based SODM*** techniques, is a cost-effective tool for building AR models with high accuracy and reliability.

(b) ***SE*** models, as should be expected from general logic point of view, provide better platform for analysis and predictions of complex systems and processes than any other type of models. Comparing the ***SE*** model endogenous variables' accuracy and their ***AR*** models' errors, we can see that even with a smaller time lag (only one year in ***SE*** versus five in ***AR***) in most cases ***SE*** equations are better and more reliable. Further research (Motzev & Lemke, 2016) with larger time lags in ***SE*** models confirmed this general statement.

Comparison Type B using models created with similar data mining techniques as the ***MLNAN*** prototype and the ***KnowledgeMiner*** software. Since it was impossible to use the same set of data the results could be used only for general conclusions. Despite this fact, as the following information shows, both models provide high reliability and accuracy and the model developed by the author is slightly better, which of course requires further tests and examinations for final conclusions.

Model 1: National Economy of Germany (Mueller & Lemke, 2003)

One important task given to economic sciences is to improve the quality of planning and management at all national economic levels. In economic research and plans & strategies development, analysis of economic systems (i.e. making studies on the level of developments, achieved so far and on existing deviations from a plan made before), and the prediction of economic indexes (i.e. determination of potential development possibilities of the economy under study) are gaining more and more importance. The purpose of such studies is to create appropriate preconditions for expected future developments, and to find the rules and factors of influence, causing these developments. The following example shows the status quo prediction of 13 important variables of the German national economy for a time horizon of two years.

Information used:

Data set contains 28 Time series of yearly observations (t = 1960 – 1987) for 13 variables:

$\mathbf{x_{1,t}}$ – Gross Domestic Product;
$\mathbf{x_{2,t}}$ – Population;
$\mathbf{x_{3,t}}$ – Gross Wages per Employee;
$\mathbf{x_{4,t}}$ – Unemployed people;
$\mathbf{x_{5,t}}$ – Employment Vacancies;
$\mathbf{x_{6,t}}$ – Total number of Employees in Germany;
$\mathbf{x_{7,t}}$ – Savings;
$\mathbf{x_{8,t}}$ – Cash Circulation;
$\mathbf{x_{9,t}}$ – Personal Consumption;
$\mathbf{x_{10,t}}$ – State Consumption;
$\mathbf{x_{11,t}}$ – Investments;
$\mathbf{x_{12,t}}$ – Export – Import;
$\mathbf{x_{13,t}}$ – Credits.

There was no any additional data preprocessing and variables were not divided into endogenous or exogenous a priory, although $\mathbf{x_{12}}$ certainly expresses the influence of other national economies.

Solution:

a) GMDH

Given the task of macroeconomic modeling, apparently dynamic models are needed. Because of this reason and since the noise dispersion of most variables is small, GMDH and the Fuzzy Rule Induction can be used here. Linear models are recommended for GMDH, because in general, this type of models is the most appropriate for control and prediction. In case that these models do not predict well, nonlinear models can be built up for the specific variables. Auto-regressive models can also be used, if the information available is insufficient for input-output system models or time lag is presented.

The only parameters needed to build this macroeconomic model are the sample size (time-series of 28 observations), the maximum time lag (up to 4 years) for model dynamics and the general model type (linear system of equations). Other parameters usually required to set up an ANN (such as when the process has to stop, penalty terms, learning rules or topology settings) are not necessary here, since the automated self-organizing data mining identifies all of them.

Using the given data set (a data table containing 13 columns and 28 rows) and the chosen system dynamic (the time lag) of up to 4 years, the information matrix for model building is constructed automatically by the software in the background. It consists of 64 columns (for 12 non lagged and 52 lagged variables) and 24 rows, and all values are already normalized automatically. It means that, for instance, for the variable $\boldsymbol{x_1}$ the following model will be created out of this information:

$$x_{1,t} = f(x_{2,t}, x_{3,t}, \dots, x_{13,t}, x_{1,t-1}, x_{2,t-1}, \dots, x_{13,t-1}, x_{1,t-2}, \dots, x_{13,t-4}).$$

In this way, a linear dynamic system of equations with 64 input variables (predictors) and 13 output (dependent) variables was created (self organized) autonomously using GMDH:

$$\underline{x}_t = A\underline{x}_t + \sum_{j=1}^{4} B_j \underline{x}_{t-j}, \quad \underline{x}_t = (x_{1,t}, x_{2,t}, \dots, x_{13,t})$$

Then the so builded model was used for predicting all dependent variables ex ante 2 years ahead in a single step.

b) Fuzzy Rule Induction

Here, all data need fuzzification first. For this purpose, five fuzzy predicates were used:

negative big (NB_)
negative small (NS_)
zero
positive small (PS_)
positive big (PB_)

Original data in the input vector $\underline{x} = (x_1, x_2, \dots x_p)$, $(p=13)$ was transformed into corresponding fuzzy vectors $x^j_p = (x^1_p, x^2_p, \dots x^5_p)$ with $x^j_p = \mu_{A^j}(x_p)$. The fuzzy membership functions $\mu_{A^j}(x_p)$ used here are of Lambda type. The mean based fuzzification results in a data set of 65 linguistic variables.

Again, it was decided to create a dynamic system of fuzzy rules with a time lag of up to four years using the transformed data set of 65 fuzzy variables. That is, the task is to create 65 rules from an information matrix of 256 linguistic input variables, and each rule is composed of several a priori unknown linguistic variables connected by AND/OR/NOT operators.

After predicting the system of fuzzy rules two years ahead, the predictions for the initial 13 dependent variables are computed using the corresponding defuzzification models. To determine these models, the GMDH is used to identify the following transformation:

$$x_{i,t} = f(x^1_{i,t}, x^2_{i,t}, \dots, x^5_{i,t}), \quad i=(1, 2, \dots, 13)$$

Results:

a) GMDH

The system of equations was generated in a single run of the KnowledgeMiner. Fig. 12-18 presents the structure (System Graph) of the generated system model automatically created by the software. At the end of the procedure, an analytical model (regression equation) is available for each dependent (output) variable. Since the final layer of the ANN shows coefficients which are obtained after many intermediate estimations, they are automatically transformed by the program back into the original data space.

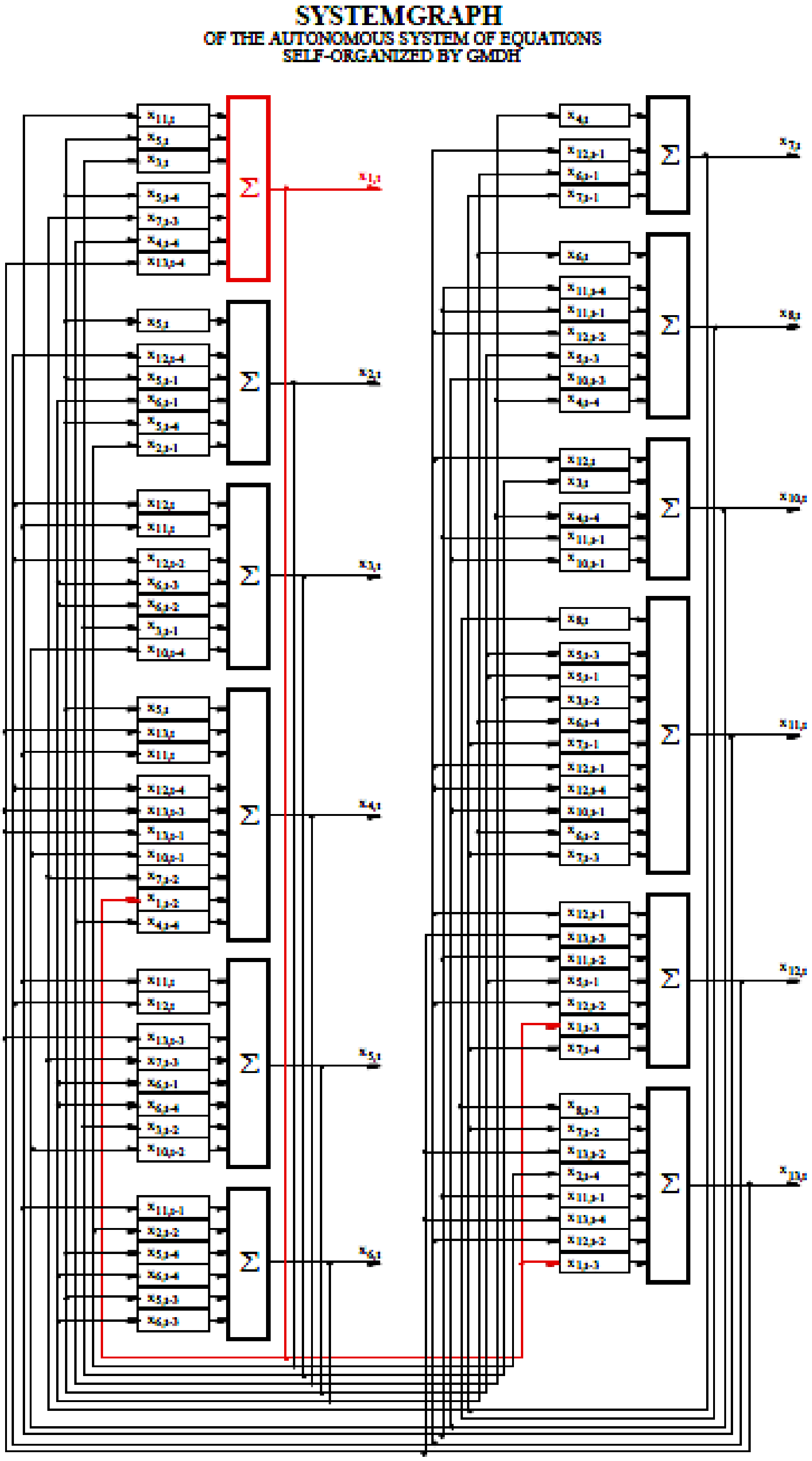
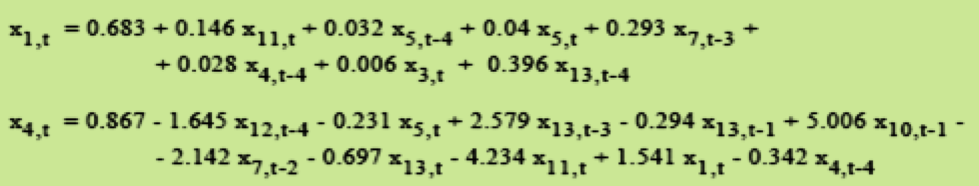


Fig.12-18 System graph of the generated macroeconomic system of equations for the German Economy

For example, the equations generated for the Gross Domestic Product and for the Unemployed People are as follow:



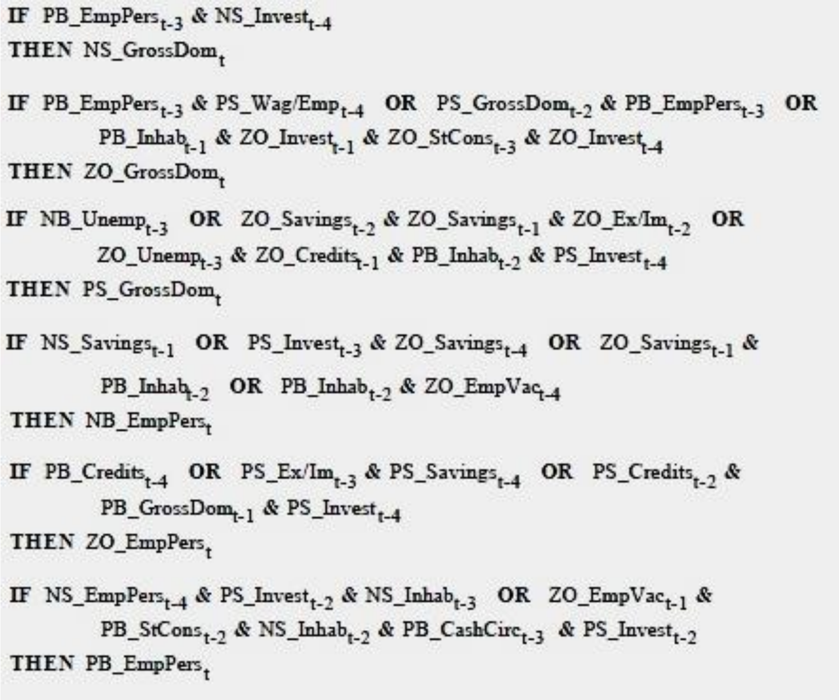
$$x_{1,t} = 0.683 + 0.146\, x_{11,t} + 0.032\, x_{5,t-4} + 0.04\, x_{5,t} + 0.293\, x_{7,t-3} + \\ + 0.028\, x_{4,t-4} + 0.006\, x_{3,t} + 0.396\, x_{13,t-4}$$

$$x_{4,t} = 0.867 - 1.645\, x_{12,t-4} - 0.231\, x_{5,t} + 2.579\, x_{13,t-3} - 0.294\, x_{13,t-1} + 5.006\, x_{10,t-1} - \\ - 2.142\, x_{7,t-2} - 0.697\, x_{13,t} - 4.234\, x_{11,t} + 1.541\, x_{1,t} - 0.342\, x_{4,t-4}$$

It is evident that only a subset of the most relevant variables is included in the models making them provident and robust. Table 12.4 presents the prediction errors when predicting the system two years ahead.

b) Fuzzy Rule Induction

Although the dimension of the modeling task was much larger here, the system of fuzzy rules was created for about the same time as with the GMDH. This is because there are no parameters to estimate and the structure of the transfer functions are faster to optimize. Here, for the Gross Domestic Product and the Unemployed People the following rules were synthesized:



IF PB_EmpPers$_{t-3}$ & NS_Invest$_{t-4}$
THEN NS_GrossDom$_t$

IF PB_EmpPers$_{t-3}$ & PS_Wag/Emp$_{t-4}$ **OR** PS_GrossDom$_{t-2}$ & PB_EmpPers$_{t-3}$ **OR**
PB_Inhab$_{t-1}$ & ZO_Invest$_{t-1}$ & ZO_StCons$_{t-3}$ & ZO_Invest$_{t-4}$
THEN ZO_GrossDom$_t$

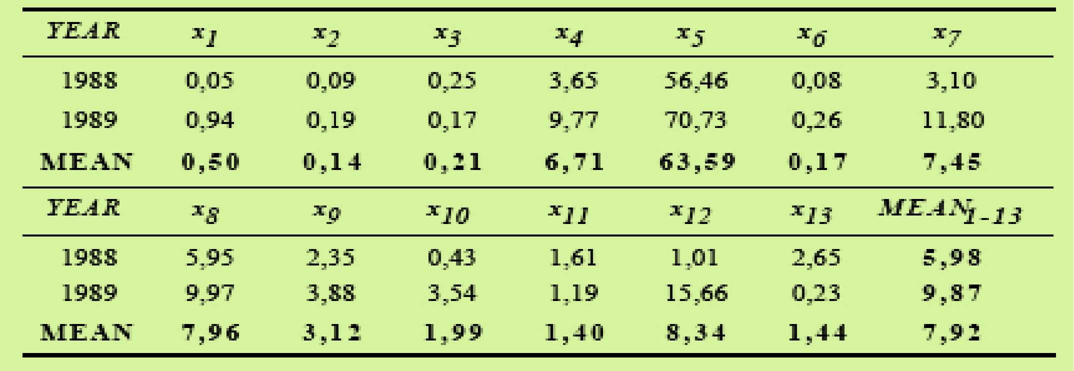
IF NB_Unemp$_{t-3}$ **OR** ZO_Savings$_{t-2}$ & ZO_Savings$_{t-1}$ & ZO_Ex/Im$_{t-2}$ **OR**
ZO_Unemp$_{t-3}$ & ZO_Credits$_{t-1}$ & PB_Inhab$_{t-2}$ & PS_Invest$_{t-4}$
THEN PS_GrossDom$_t$

IF NS_Savings$_{t-1}$ **OR** PS_Invest$_{t-3}$ & ZO_Savings$_{t-4}$ **OR** ZO_Savings$_{t-1}$ &
PB_Inhab$_{t-2}$ **OR** PB_Inhab$_{t-2}$ & ZO_EmpVac$_{t-4}$
THEN NB_EmpPers$_t$

IF PB_Credits$_{t-4}$ **OR** PS_Ex/Im$_{t-3}$ & PS_Savings$_{t-4}$ **OR** PS_Credits$_{t-2}$ &
PB_GrossDom$_{t-1}$ & PS_Invest$_{t-4}$
THEN ZO_EmpPers$_t$

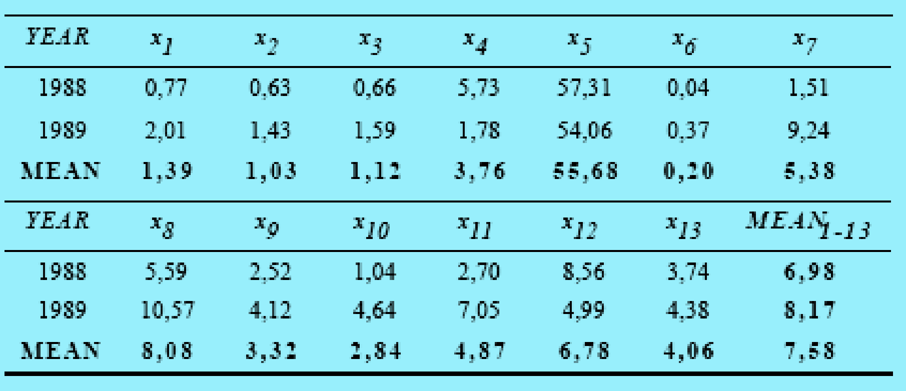
IF NS_EmpPers$_{t-4}$ & PS_Invest$_{t-2}$ & NS_Inhab$_{t-3}$ **OR** ZO_EmpVac$_{t-1}$ &
PB_StCons$_{t-2}$ & NS_Inhab$_{t-2}$ & PB_CashCirc$_{t-3}$ & PS_Invest$_{t-2}$
THEN PB_EmpPers$_t$

Table 12.4 Absolute percentage prediction errors (MAPE) for the generated GMDH system model



YEAR	x_1	x_2	x_3	x_4	x_5	x_6	x_7
1988	0,05	0,09	0,25	3,65	56,46	0,08	3,10
1989	0,94	0,19	0,17	9,77	70,73	0,26	11,80
MEAN	**0,50**	**0,14**	**0,21**	**6,71**	**63,59**	**0,17**	**7,45**
YEAR	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	*MEAN*$_{1-13}$
1988	5,95	2,35	0,43	1,61	1,01	2,65	5,98
1989	9,97	3,88	3,54	1,19	15,66	0,23	9,87
MEAN	**7,96**	**3,12**	**1,99**	**1,40**	**8,34**	**1,44**	**7,92**

Table 12.5 MAPE when applying the system of fuzzy rules



YEAR	x_1	x_2	x_3	x_4	x_5	x_6	x_7
1988	0,77	0,63	0,66	5,73	57,31	0,04	1,51
1989	2,01	1,43	1,59	1,78	54,06	0,37	9,24
MEAN	1,39	1,03	1,12	3,76	55,68	0,20	5,38
YEAR	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	$MEAN_{1-13}$
1988	5,59	2,52	1,04	2,70	8,56	3,74	6,98
1989	10,57	4,12	4,64	7,05	4,99	4,38	8,17
MEAN	8,08	3,32	2,84	4,87	6,78	4,06	7,58

For comparisons, Table 12.5 displays the prediction results for the fuzzy system model and Table 12.6 shows the effect of synthesizing predictions from both methods.

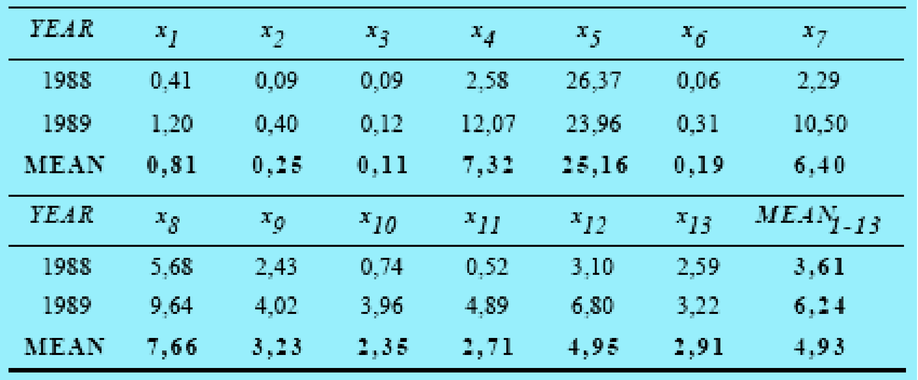
As the results show, Self-organizing Data Mining techniques (in this case GMDH and FRI) can extract some useful knowledge for both analysis and predictions. Of course, results depend on available information. In this example, only a subset of relevant macroeconomic variables was used. Moreover, some of them are influenced by governmental decisions, which can hardly be predicted. Employment vacancies (X_5) is an example here. An auto-regressive model with a time lag of 10 has predicted this variable with 25% and 1% error (mean of 13%), indicating that X_5 cannot be described precisely from the given data set.

Model 2: The National Economy of the Republic of Bulgaria SIMUR II (Marchev & Motzev, 1985) discussed in Chapter 9.3. It is a very similar aggregated macroeconomic model in the form of 12 SE (9-17) for a similar period of time. It contains 42 variables, incl. 12 endogenous, 5 exogenous and 26 lag variables with a time lag of up to 3 years.

Fig.12-19 displays the structure of the generated SE, synthesized with the ***MLNAN*** algorithm and the automated procedure using ***KnowledgeMiner (yX) for Excel*** software.

For the need of comparisons both models were used to calculate forecasts for two years ahead. Table 12.7 presents the prediction errors for calculated forecasts with ***MLNAN***.

Table 12.6 MAPE for calculated predictions from both methods



YEAR	x_1	x_2	x_3	x_4	x_5	x_6	x_7
1988	0,41	0,09	0,09	2,58	26,37	0,06	2,29
1989	1,20	0,40	0,12	12,07	23,96	0,31	10,50
MEAN	0,81	0,25	0,11	7,32	25,16	0,19	6,40
YEAR	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	$MEAN_{1-13}$
1988	5,68	2,43	0,74	0,52	3,10	2,59	3,61
1989	9,64	4,02	3,96	4,89	6,80	3,22	6,24
MEAN	7,66	3,23	2,35	2,71	4,95	2,91	4,93

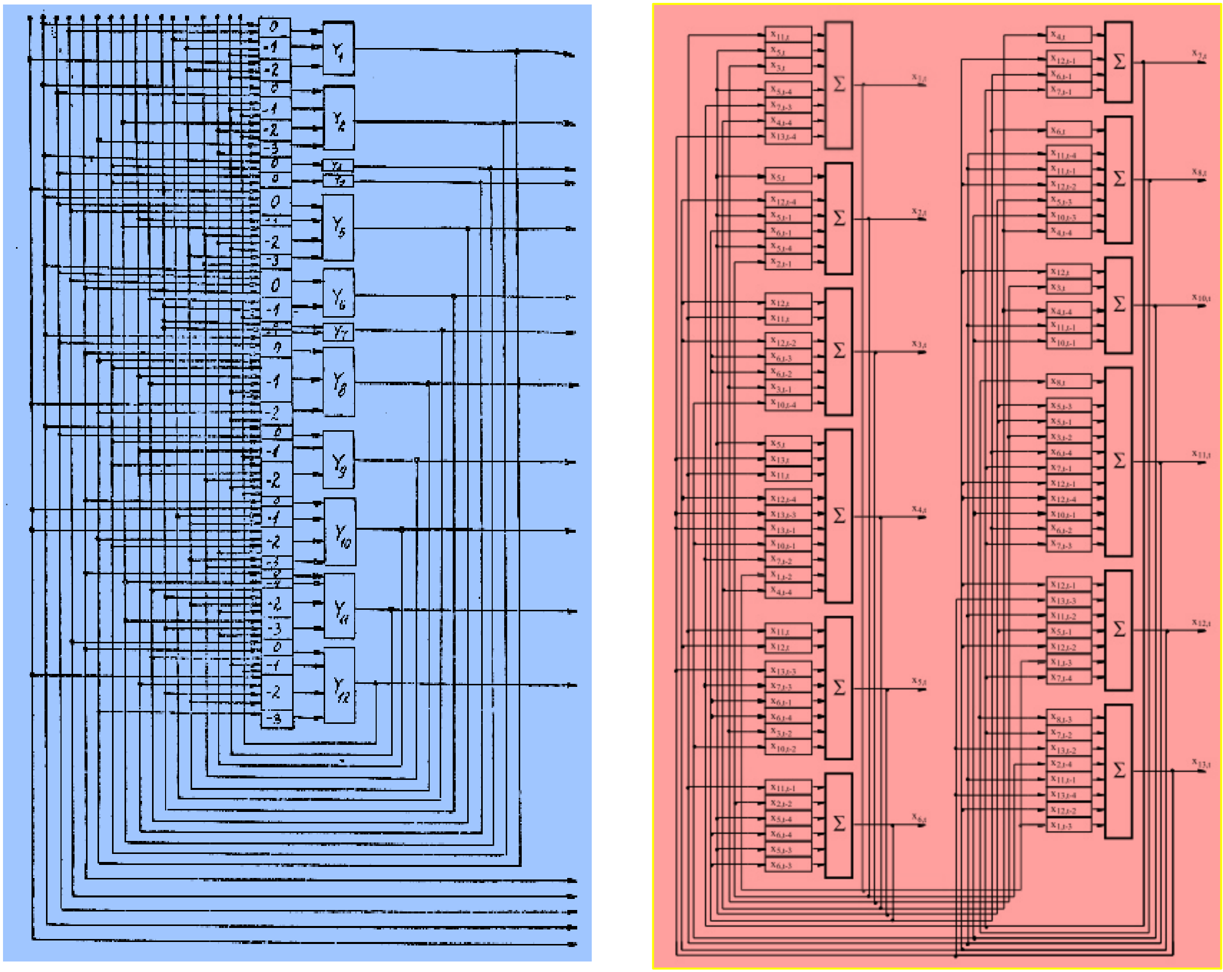


Fig.12-19 System graphs of the generated macroeconomic system of equations for the Bulgarian and the German Economy

A comparison between Tables 12.6 and 12.7 shows that in general the Bulgarian model has smaller error than the German model of National Economy – Mean % Differences Between Predictions & Real Data 4.29% or MAPE=4.61% (it’s not clear in the publication (Mueller & Lemke, 2003) which of the two statistics was used) for the German model versus MAPE=3.90% for the Bulgarian model. This almost insignificant difference proves that both ***Self-Organizing Data Mining*** techniques generated ***SE*** models which are very close in their accuracy.

Further analysis and simulations with the next, more detailed and bigger model of the series SIMUR III, a complex macro-economic model of 39 SE with 39 endogenous, 7 exogenous and 82 lag variables with a time lag up to five years (see Motzev & Marchev, 1988), confirms the above statement. In SIMUR III, the average MSE% < 1%.

Table 12.7 Prediction errors for calculated forecasts

Variable	*% Differences Between Predictions & Real Data*			*MAPE*	*MSE* (%)
	1988	**1989**	**Average %**		
Y_{1t}	2.58%	3.85%	3.22%	3.48%	5.10%
Y_{2t}	6.45%	0.61%	3.53%	3.73%	6.40%
Y_{3t}	6.54%	0.62%	3.58%	3.84%	5.85%
Y_{4t}	6.15%	0.56%	3.36%	3.59%	8.93%
Y_{5t}	0.35%	3.81%	2.08%	2.30%	4.47%
Y_{6t}	0.04%	1.36%	0.70%	1.10%	1.82%
Y_{7t}	16.51%	17.60%	17.06%	17.80%	10.12%
Y_{8t}	0.83%	0.62%	0.73%	1.15%	0.72%
Y_{9t}	6.15%	8.23%	7.19%	7.43%	6.68%
Y_{10t}	2.22%	3.34%	2.78%	2.98%	2.31%
Y_{11t}	6.26%	3.86%	5.06%	5.45%	7.72%
Y_{12t}	3.16%	1.18%	2.17%	2.48%	3.28%
Mean	4.77%	3.80%	4.29%	4.61%	5.28%

Another proof of evidence about GMDH based predictive techniques qualities is the Global Energy Forecasting Competition (GEFCom2014) organized by the IEEE Power & Energy Society and the University of North Carolina at Charlotte. Based on a 12-week rolling real-world forecasting scenario in four categories - electric load, electricity price, wind and solar power forecasting - for up to 10 sites, ***Insights***' self-learning forecasting skills ended up in top five positions in all categories (see https://www.knowledgeminer.eu/pr/pr_020315.html).

Just recently, the ***MLNANs*** have been used in Business Forecasting and Predictive Analytics class at the Walla Walla University School of Business. For three years in a row, many predictive models have been developed using more or less complex techniques. In summary, the predictions done with the MLNAN always have the smallest errors (i.e. the highest accuracy) as presented in Tables 12.8 and 12.9.

Table 12.8 Sales predictions accuracy using different models in 2018

Best Model	**Second Best**	**Third Best**
MLNAN	***Triple Exponential***	***Multiple Autoregression***
MASE = 0.0414	MASE = 0.0627	MASE = 0.0908
MPE = 1.42%	MPE = -0.57%	MPE = 2.03%
MAPE = 1.42%	MAPE = 1.76%	MAPE = 2.58%
CV(RMSE) = 1.56%	CV(RMSE) = 2.45%	CV(RMSE) = 3.17%

Table 12.9 Sales predictions accuracy using different models in 2019

Best Model	Second Best	Third Best
MLNAN:	***Multiple Regression with Time and Dummy Seasonal Variable***	***Triple Exponential***

MASE: 0.0446
MPE = 1.55%
MAPE = 1.55%
CV(RMSE) = 1.56%
	MASE = 0.0508
	MPE = -1.09%
	MAPE = 1.59%
	CV(RMSE) = 1.56%
		MASE = 0.0627
		MPE = -0.57%
		MAPE = 1.76%
		CV(RMSE) = 2.45%

All results so far have proved the advantages of utilizing ***MLNANs*** in business simulations. ***Self-organizing Data Mining*** techniques such as ***MLNANs*** really provide opportunities in both shortening the design time and reducing the cost and efforts in simulations model building, as well as reliably developing even complex models with high level of accuracy.

***

SUMMARY AND CONCLUSIONS

Chapter 12 discusses ***Self-organizing Data Mining*** and its applications in *Model building* and *Business Forecasting*. ***Group Method of Data Handling (GMDH)*** was introduced in 1968 by Alexey Ivakhnenko as an inductive approach to model building based on self-organization principles. In ***GMDH-based Self-organizing modeling algorithms***, models are generated adaptively from data in the form of networks of active neurons in a repetitive generation of populations of competing models of growing complexity, corresponding validation, and selection model until an optimal complex model that is not too simple and not too complex have been realized.

A) There are many applications of ***Self-Organizing Data Mining Algorithms:***

- ***Combinatorial (COMBI) algorithm*** is based on full or reduced sorting-out of gradually complicated models and their evaluation by external criterion on a testing data set. It generates models of all possible input variable combinations and regarding the chosen selection criterion selects a final best model from the generated set of models. Its disadvantage is that such algorithm can handle effectively only up to 30-40 input variables due to a nonlinear increase of the total number of possible model versions:

  - One possible way of improvement is to apply a Recursive scheme for faster combinatorial sorting. The recursive technique is convenient to use in constructing models of partial polynomials of gradually increasing complexity that begin with a single argument. This type of approach is called "method of bordering."

  - Another improvement is the ***multilayered structures using combinatorial setup*.** One version of a multilayered structure is that the combinatorial algorithm could be implemented at each layer of the multilayered network structure by keeping the limit on the "freedom of choice" at each layer. The unit outputs are fed forward layer by layer as per the threshold measure to obtain the global output response for optimal complexity.
- ***Multilayered Iterative GMDH algorithm*** is an algorithm in which the iteration rule remains unchanged from one layer to the next. It should be used when it is needed to handle a big number of variables. The network structures in ***GMDH*** differ as per the interconnections among the units and their hierarchical levels. Multilayered algorithms use a multilayered network structure with linearized input arguments and generate simple partial functions. Regarding the computational aspects in the

process, the multilayered network procedures are more repetitive in nature. It is important to consider the algorithm in modules and to facilitate repetitive characteristics.

- ***Objective System Analysis (OSA) algorithm*** examines systems of algebraic or difference equations, obtained by implicit templates (without goal function). An advantage of the algorithm is that the information embedded in the data sample is utilised better and we can estimate the relationships between variables. ***OSA*** is also known as *Self-organization of systems of networks (autonomous systems of equations)*. This is the task of modeling complex systems which usually have more than one output variables. The goal of modeling systems of equations (***SE*** models) using *GMDH* is to self-organize a model for each output variable and to identify the ***structural form***, i.e. the interdependence structure between the system variables separating them into endogenous and exogenous according to their corresponding model purpose. After modeling a system of *m* equations, ***OSA*** selects one (or a few alternative versions) best autonomous system consisting of *m** equations (*m** < *m*). All *m** variables of the so-identified best system are considered as endogenous variables of the ***SE*** system. All the remaining variables which may be part of the ***SE*** model are either exogenous or are identified as exogenous due to an insufficient data basis. Sometimes, when the forecasting goal is based on a pre-specified structure (in terms of its endogenous variables) of the model, this could be a disadvantage.

B) There are special peculiarities of ***GMDH***:

1) External supplement – following Beer principle, only the external criteria, calculated on new independent information, can produce the minimum of sorting-out characteristic. Because of this data sampling is divided into parts for model training and evaluation.

2) Freedom of choice – according to Gabor's work (1971), in multilayered GMDH algorithms not one but F (F>1) best results of choice are to be conveyed from one layer to the next layer to provide "freedom of choice".

3) The rule of layers complication – partial models (forms of a mathematical description for iteration) should be simple, without quadratic equations;

4) Additional model definition – in cases, when the choice of optimal physical model is difficult, because of noise level or oscillations of criterion minima characteristic, auxiliary discriminating criterion is used. The choice of the main criterion and the constrains of sorting-out procedure is the main heuristic of GMDH;

5) All algorithms have multilayered structure and parallel computing can be implemented for their implementation;

6) All questions that arise about the type of algorithm, criterion, variables set, etc. should be addressed by the minimum of criterion value - The main criteria used in GMDH are cross-validation, regularity and balance of variables.

C) ***GMDH algorithms*** are different from other algorithms of structural identification, genetic and best regression selection algorithms because of the following main properties:

- *usage of external criteria*, which are based on data sample dividing (cross-validation) and are adequate to the problem of forecasting models building by reducing the requirements of the initial information volume;
- *much more diversity of structure generators usage* like in regression algorithms of the ways of full or reduced sorting of structure variants and of *original multilayered procedures*;
- *better level of automation* – it is only necessary to enter an initial data sample and the type of external criterion;
- *automatic adaptation* of optimal model complexity and external criteria to the level of noises or statistical violations – the effect of noise immunity causes robustness of the approach;
- *implementation of the principle of inconclusive decisions* in the process of gradual models complication.

D) In ***ANNs*** the user estimates this structure by choosing the number of layers and the number and transfer functions of nodes of a neural network. This requires not only knowledge about the theory of ***ANNs***, but also knowledge of the object nature and time. Besides this the knowledge from systems theory about the systems modeled is not applicable without transformation in the neural network world and the rules of translation are usually unknown.

E) ***GMDH type ANNs*** can overcome these problems - it can pick out knowledge about object directly from data sampling. The GMDH is an inductive sorting-out method, which has advantages in case either of rather complex objects and/or no definite theory. GMDH algorithms can find the only optimal model using full sorting-out of model-candidates and evaluating them by external criteria of accuracy:

- In distinction to neural networks, the results of GMDH algorithms are explicit mathematical models obtained in a relatively short time based on extremely short samples.
- The well-known problems of an optimal (subjective) choice of the neural network architecture are addressed in the GMDH algorithms by means of an adaptive synthesis (objective choice) of the architecture.

- GMDH algorithms could be used to estimate networks of the right size with a structure evolved during the estimation process in order to provide a parsimonious model for the particular desired function. Such algorithms, combining in a powerful way the best features of neural nets and statistical techniques, discover the entire model structure in the form of a network of polynomial functions, differential equations and others. Models are selected automatically based on their ability to solve tasks such as approximation, identification, prediction, and classification.

F) Only by this inductive self-organizing method for small, inaccurate or noisy data samples the optimal non-physical model can be found, the accuracy of which is higher and the structure is simpler than the structure of a usual full physical model. GMDH algorithms are the examples of complex active neurons, because they choose by themselves the effective inputs and their corresponding coefficients in the process of self-organization.

G) Data mining techniques require powerful software, designed and elaborated for this specific aim. Developing your own computer program in this area is a big project, which takes a lot of resources, time and highly qualified researchers. One of the leading software platforms in self-organizing data mining is *KnowledgeMiner* program, a self-organizing tool for modeling and predictions, which implements GMDH, Analog Complexion (***AC***) and Fuzzy Rule Induction. KnowledgeMiner (yX) for Excel is a knowledge mining tool that works with data stored in MS Excel for building predictive and descriptive models from this data. It supports all major releases of Microsoft Excel for Mac computers.

H) Experiments with ***Self-Organizing Data Mining*** techniques show that they can generate many different types of forecasting models (***SE***, *Distributed lag models*, *Autoregressive-moving-average with exogenous inputs models*, *Vector autoregression models* and others) in a cost-effective way with high accuracy and reliability.

I) Recent proof of evidence about ***GMDH*** based predictive techniques qualities is the Global Energy Forecasting Competition (GEFCom2014) organized by the IEEE Power & Energy Society and the University of North Carolina at Charlotte. Based on a 12-week rolling real-world forecasting scenario in four categories - electric load, electricity price, wind and solar power forecasting - for up to 10 sites, ***Insights***' self-learning forecasting skills ended up in the top five positions in all categories.

KEY TERMS

CHAPTER EXERCISES

Conceptual Questions:

1. What is ***Self-organizing Data Mining?*** Why it is important in *Model building* and *Business Forecasting*?
2. Explain the main characteristics of ***Group Method of Data Handling (GMDH).*** What makes this approach so unique in model building and forecasting?
3. What are the major groups of ***GMDH-type*** algorithms?
4. What are the differences between traditional ***ANNs*** and ***Statistical Learning Networks (GMDH type ANNs)?*** List and discus at least three of them.
5. Define ***Self-Organization of Nets of Active Neurons*** and the ***Multi-Layer Net of Active Neurons (MLNAN)*** in particular. What are the general steps in this approach?
6. How ***GMDH type ANNs*** can overcome existing problems in model building and business forecasting? Discuss some forecasting applications of this special type of ***ANNs***.
7. What is ***KnowledgeMiner (yX) for Excel?*** Explain its main unique features.

Business Applications:

1. Open the file SalesData.xlsx in ***MS Excel***.
2. Open ***Knowledge-Miner (yX) for Excel*** and import the file SalesData from MS Excel.
3. Prepare Data Table in ***Knowledge-Miner (yX) for Excel*** for building an ***ARMAX*** model with dependent variable “Sales” and the given predictors:
   - Set up the time series data for a dependent variable “Sales”;
   - Set up the time series data for all given predictors;
   - Select time lags of up to 12 periods within the time-series for both dependent and independent variables accordingly.
4. Self-organize the ***ARMAX*** model and compute Sales forecast for the next 12 months given the expected values for the predictors.
5. Export the results in MS Excel worksheet:
   - Analyze the output of the mining process. Are there any potential improvements to be made? If yes, return to step 1 and improve the ***ARMAX*** model.
   - Design formulas, similar to the formulas in Part 4 of the Integrative case and compute MAD, MSE, MAPE and MPE for the new model, for a testing dataset of the 12 new monthly forecasts given in spreadsheet Predictions.
   - What is the model accuracy?

Discuss all findings and write a short report (up to two pages) summarizing your answers.

INTEGRATIVE CASE

HEALTHY FOOD SUPPLY CHAIN & STORES

Part 12: Self-Organizing Data Mining Forecasts – II

In Chapter 1 we introduced *Healthy Food Stores* – a fast-growing retail food provider with 12 stores in a northwestern state. The company executives decided to study the effect that company advertising dollars have on sales. They hoped that examining collected historical data would reveal relationships that would help determine future advertising expenditures and predict monthly sales volumes for the upcoming quarter.

After identifying basic parameters, along with input (independent) and output (dependent) variables of the forecasting scenario in Part 2, the research team determined the main elements of the forecasting process:

- Forecasting horizon of up to twelve months;
- Quarterly forecast updates, since accuracy decreases as time horizon increases, and sufficient time is needed to implement possible changes;
- Development of different models based on data patterns, if any, and evaluation of their accuracy in order to select the most appropriate one;
- Selection of the best forecasting model with no more than 5% forecasting error.

The benchmark forecast computed in Part 3 (using the baseline of ***one-step naive forecast*** as a reference forecast) provided basic values for the most common measures of accuracy MFE, MAD, MAPE, MPE, MSE and CV(RMSE).

In Part 4, information about the opinion of some important people from the *Healthy Food Stores Company*, concerning this specific case, was collected and the research team applied the Delphi method to top executives group, Sales-force composite to the sales managers from all 12 stores and Scenario writing to the most experienced professionals from Advertising Department. After collecting such valuable information from different sources, in Part 5 the research team made its first steps in Numerical Predictions by developing different basic forecasting models. They created spreadsheets for Naïve techniques (Average model, Random Walk with Drift and Seasonal Naïve Technique), simple Moving Average, Simple Exponential Smoothing and Triple Exponential Smoothing, used to expand the base-line of one-step naïve forecast as reference forecasts.

In Part 6 the research team analyzed the relationships between dependent variable Sales and the available predictors. After performing multiple correlation and regression analysis, researchers developed reliable forecasting model, which passed all tests and hypotheses,

representing the real system with certain error. In Part 7, the model was expanded by adding Dummy seasonal variables to analyze the Seasonal effect in company Sales. In Part 8, the improvement of the forecasting model continued (with the help of some advanced Time series analyses and predictive techniques) and few simple ***AR*** models were build using ***ARIMA*** methodology and *Gretl* software.

In *Parts 9* & *10* (**Complex Models and Forecasting**), new and more complex ***AR*** and ***ARMAX*** models were built using ***ARIMA*** methodology and *Gretl* software and in *Part 11* an ***AR*** model as ANN was self-organized using ***Knowledge-Miner for Excel.***

The final step of the research in this integrative case is to build an ***ARMAX*** model using Data Mining platform ***Knowledge-Miner for Excel*** and its newer, improved version ***Insights.***

Case Questions

1. Open Data.xslx file in ***MS Excel*** and select Data worksheet.
2. Open ***Insights*** software and import the worksheet Data from MS Excel file.
3. Prepare Data Table in ***Knowledge-Miner for Excel/Insights*** for building an ***ARMAX*** model with dependent variable “Sales” and the given predictors:
   - Set up the time series data for a dependent variable “Sales”;
   - Set up the time series data for all given predictors;
   - Select time lags of up to 12 periods within the time-series for both dependent and independent variables accordingly.
4. Self-organize the ***ARMAX*** model and analyze the output of the mining process.
5. What are the properties and the accuracy of the model? Is it possible to improve it? If yes, return to step 3 and design new ***ARMAX*** models until satisfactory results.
6. Compute Sales forecast for the next 12 months. Export the results in MS Excel worksheet and rename it to MLNANARMAX.
7. Use (copy/paste) the formulas designed in Part 3 to compute MFE, MAD, MAPE, MPE, MSE and CV(RMSE) for the new model, for the given testing dataset of 12 monthly forecasts provided in spreadsheet Errors.
8. Comment and analyze model’s accuracy - how good is the accuracy of these forecasts? What model, out of all models so far, provides the best accuracy? Discuss.
9. What overall recommendations and in particular about ***Self-organizing Data Mining*** and ***Multi-Layered Net of Active Neurons*** would you make to the research team?

10. Write a report (at least two pages not counting charts and tables) on the questions above, discussing all important findings and draw relevant conclusions about this part of the Integrative Case.
11. Prepare summary report the whole Integrative Case combining in appropriate way partial reports from all 12 Parts. Draw overall conclusions and recommendations in one final section.
12. Check the final draft for formal research writing style (use APA style) and submit an electronic copy to the class lecturer.

References

Aksenova, T., & Yurachkovsky, Y. (1988). A Characterization at Unbiased Structure and Conditions of Their J-Optimality, Soviet Journal of Automation and Information Sciences, 21(4), 36-42.

Beer, S. (1959). Cybernetics and Management. English University Press, London, p. 280.

Belogurov, V. (1990). A criterion of model suitability for forecasting quantitative processes, Soviet Journal of Automation and Information Sciences, 23(3), 21-25.

Farlow, S. (Ed.). (1984). Self-organizing Methods in Modeling, Statistics: Textbooks and Monographs, 54. Marcel Dekker Inc., New York and Basel.

Gabor, D. (1971). Perspectives of Planning, Organization of Economic Cooperation and Development. Empire College of Science and Technology, London.

Ivakhnenko, A. G. (1968). The Group Method of Data Handling - a Rival of the Method of Stochastic Approximation, *Soviet Automatic Control*, 13(3), 43–55.

Ivakhnenko, A. G., & Müller, J-A. (1996). Recent Developments of Self-Organizing Modeling *Prediction and Analysis of Stock Market*. http://www.gmdh.net/articles/index.html

Ivakhnenko, A. G., Ivakhnenko, G. A., & Müller, J-A. (1994). Self-Organization of Neural Networks with Active Neurons, Pattern Recognition and Image Analysis, 4(2), 85-196.

Lemke, F. (2008) *Parallel Self-organizing Modeling*. http://www.knowledgeminer.com/pdf/performance_yX.pdf

Madala, H. R., & Ivakhnenko, A. G. (1994). *Inductive Learning Algorithms for Complex Systems Modeling*. Boca Raton, FL: CRC Press Inc.

Marchev, A., & Motzev, M. (1985). Computer Macro-Economic Models for Simulation Experiments, Systems Analysis and Simulation, 28(II), (now Annual Review in Automatic Programming, 12), 145-150.

Motzev, M. (2010). Intelligent Techniques in Business Games and Simulations – A Hybrid Approach. In Martin Beran (Ed.) Changing the world through meaningful play, Proceedings of ISAGA World Conference, (WA, USA), 81-86.

Motzev, M. (2011). New Product – An Integrated Simulation Game In Business Education. In Bonds & Bridges, Proceedings of ISAGA World Conference (Poland), 63-75.

Motzev M. (2018). A Framework for Developing Multi-Layered Networks of Active Neurons for Simulation Experiments and Model-Based Business Games Using Self-Organizing Data Mining with the Group Method of Data Handling. In: Lukosch H., Bekebrede G., Kortmann R. (eds) Simulation Gaming. Applications for Sustainable Cities and Smart Infrastructures. ISAGA 2017. Lecture Notes in Computer Science, vol 10825. Springer, Cham., pp 191-201 (original - in English) https://doi.org/10.1007/978-3-319-91902-7_19

Motzev, M., & Lemke, Fr. (2016). Self-Organizing Data Mining Techniques in Model Based Simulation Games for Business Training and Education, Vanguard Scientific Instruments in Management, Vol. 11.

Motzev, M., & Marchev, A. (1988). Multi-Stage Selection Algorithms in Simulation, Proceedings of XII IMACS World Congress, 4, France, 533-535.

Motzev, M., & Marchev, A. (1989). Principles of Multi-Stage Selection in Software Development in Decision Support Systems, Methodology and Software for Interactive Decision Support . Lecture Notes in Economics and Mathematical Systems, 337 IIASA. Springer-Verlag, 181-189.

Motzev, M., Muller, J-A., & Marchev, A. (1986). Macro-Economic Systems Modeling and Forecasting Using Auto-Regressive Models, Social Management, 6, 77-93.

Müller, J-A., & Lemke, F. (1995). Self-Organizing modeling and decision support in economics, Proceedings of the IMACS Symposium on Systems Analysis and Simulation, (135-138). Gordon and Breach Publ.

Mueller J. A., & Lemke, F. (2003). *Self-Organizing Data Mining: An Intelligent Approach To Extract Knowledge From Data.* Victoria, BC: Trafford Publishing.

Onwubolu, G. (2008, May). Design of hybrid differential evolution and GMDH networks for modeling and prediction, Information Sciences, 178(18), 3616–3634.

Sawaragi, Y., Soeda, T. et al. (1979). Statistical Prediction of Air Pollution Levels Using Non-Physical Models, Automatica (IFAC), 15(4), 441-452.

***

INDEX