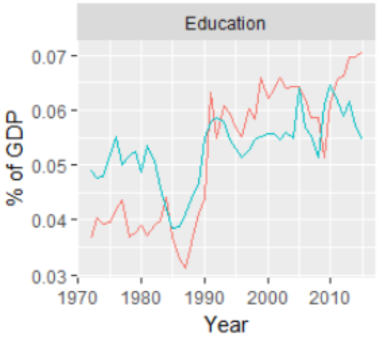
Time series analysis

Angel Marchev, Jr.

Kaloyan Haralampiev

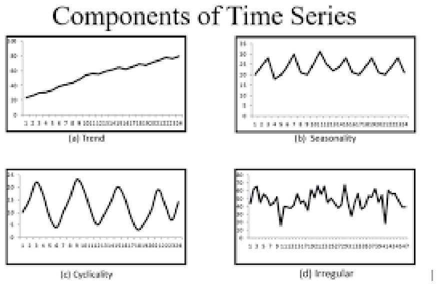
Key topics

Comparability

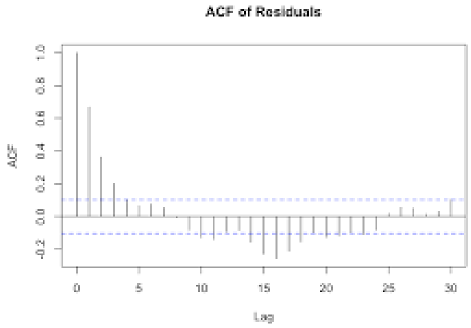


Stationarity

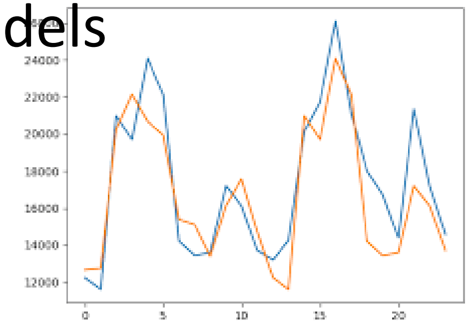
Components



Autocorrelation



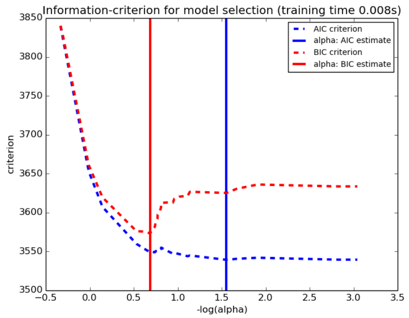
Models



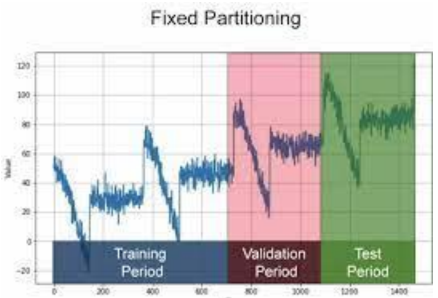
Feature engineering

Date	Value	$Value_{t-1}$	$Value_{t-2}$
1/1/2017	200	NA	NA
1/2/2017	220	200	NA
1/3/2017	215	220	200
1/4/2017	230	215	220
1/5/2017	235	230	215
1/6/2017	225	235	230
1/7/2017	220	225	235
1/8/2017	225	220	225
1/9/2017	240	225	220
1/10/2017	245	240	225

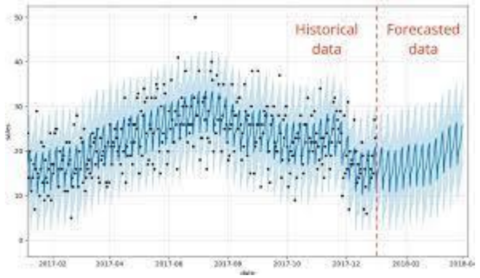
Optimization



Validation



Forecast



Comparability

Basic

- By territory
- By time
- By methodology

Additional

- By prices
- By coverage
- By measurement units

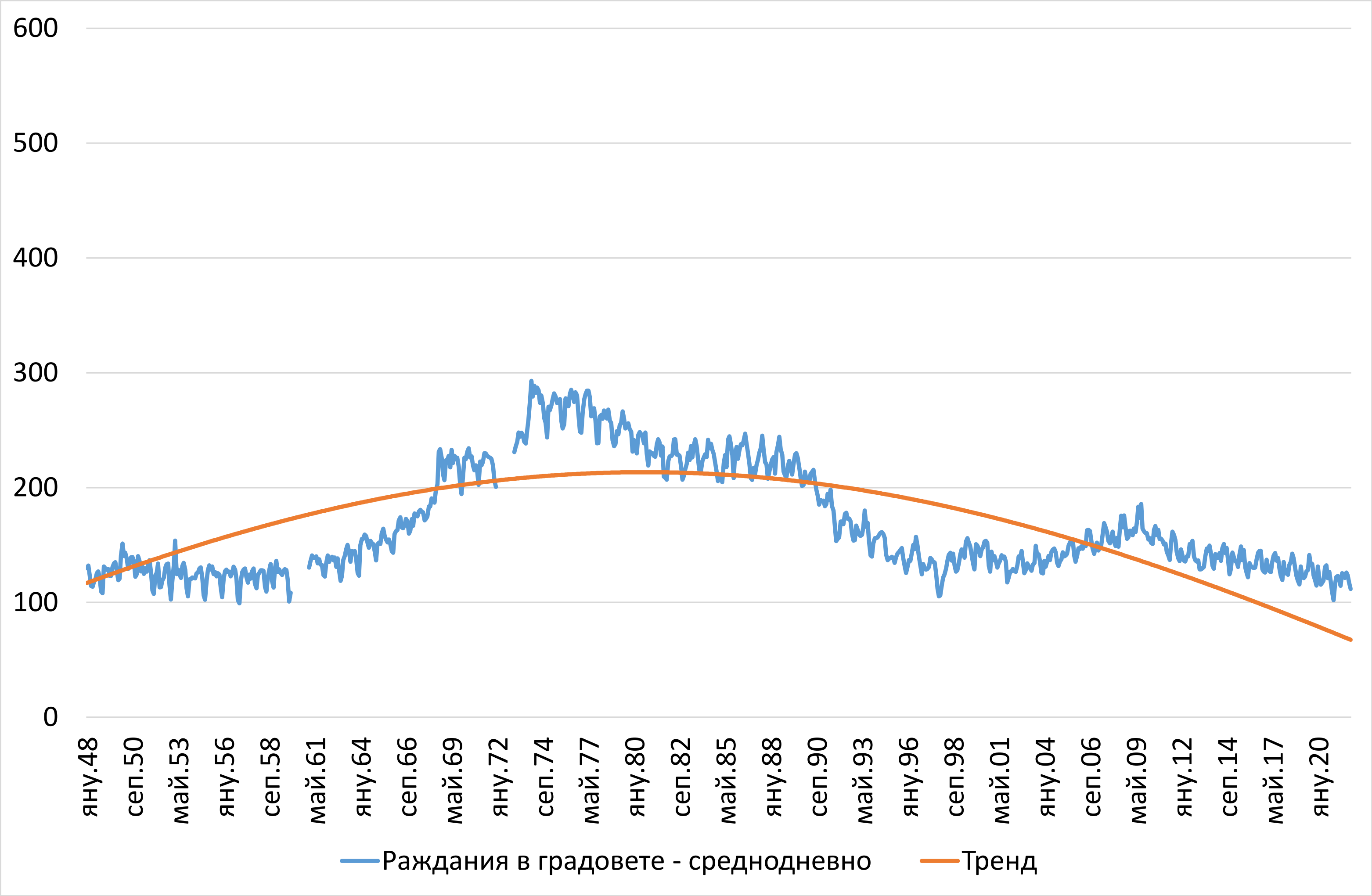
Stationarity

- Constant distribution
- i.e.
- Constant mean
- Constant variance
- etc…

Components of dynamics

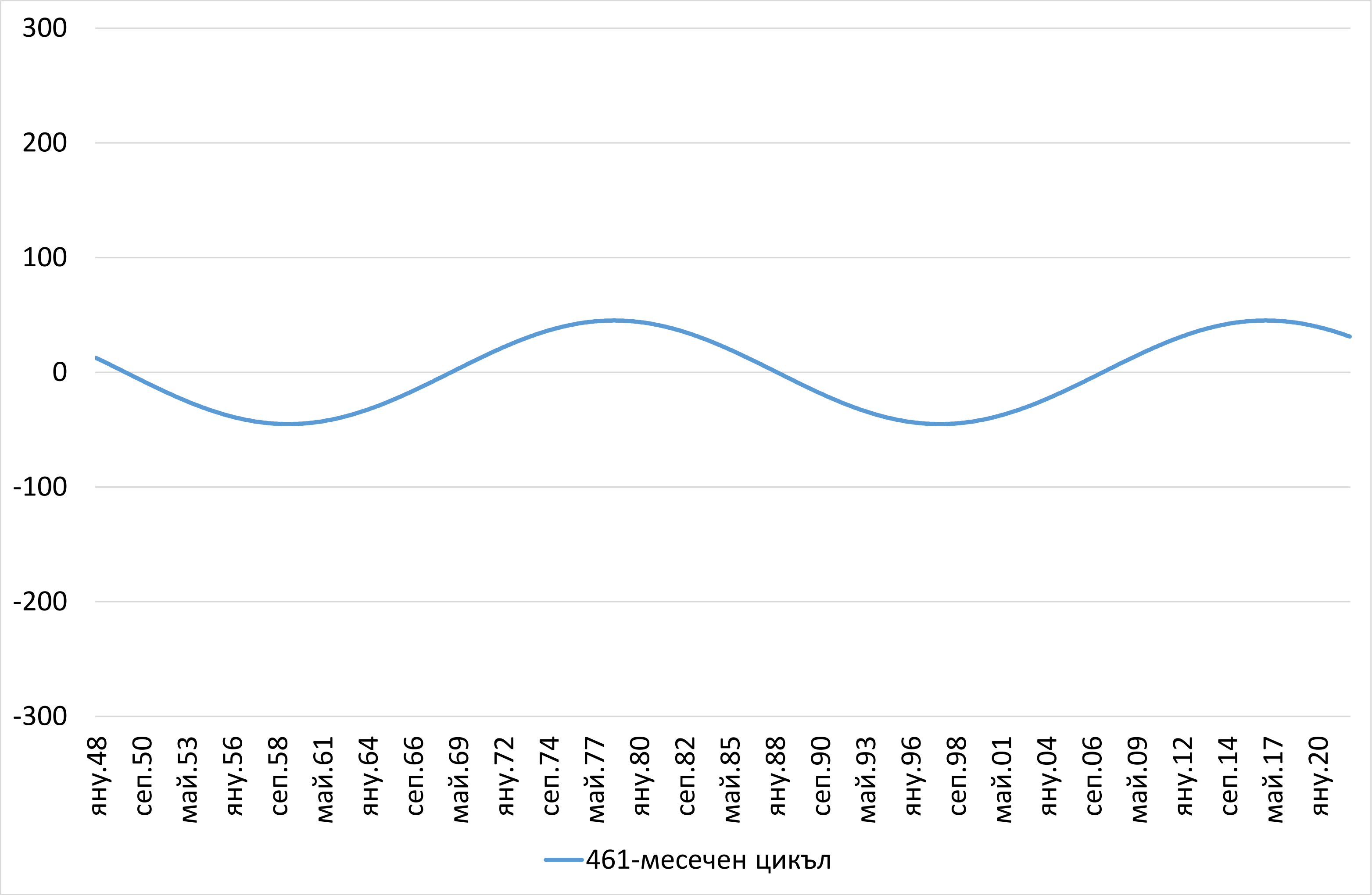
- Trend
- Cycle
- Seasonality
- Residuals

Trend



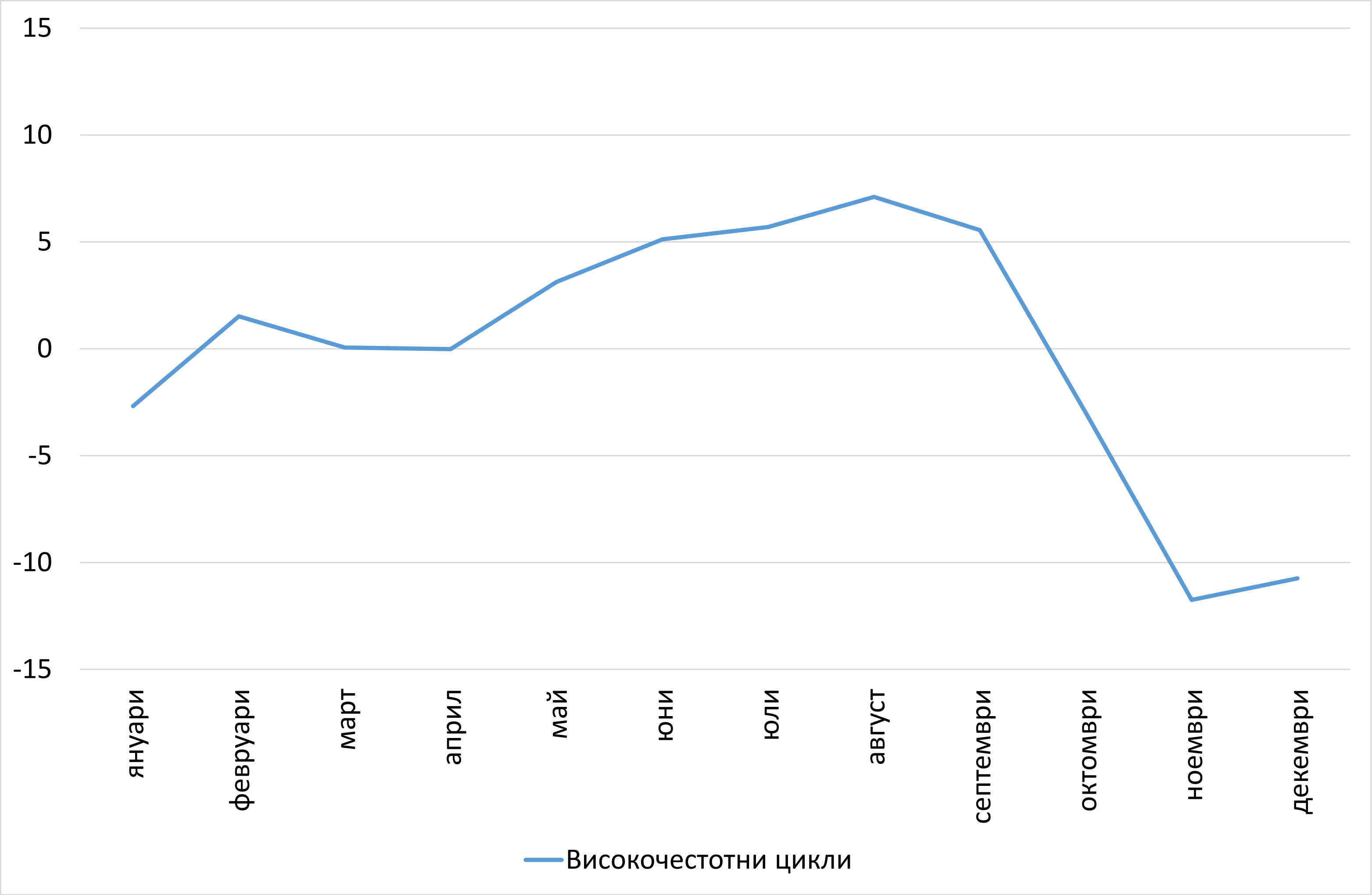
600
500
400
300
200
100
0
яну.48
сеп.50
май.53
яну.56
сеп.58
май.61
яну.64
сеп.66
май.69
яну.72
сеп.74
май.77
яну.80
сеп.82
май.85
яну.88
сеп.90
май.93
яну.96
сеп.98
май.01
яну.04
сеп.06
май.09
яну.12
сеп.14
май.17
яну.20
Раждания в градовете - среднодневно
Тренд

Cycle



300
200
100
0
-100
-200
-300
яну.48
сеп.50
май.53
яну.56
сеп.58
май.61
яну.64
сеп.66
май.69
яну.72
сеп.74
май.77
яну.80
сеп.82
май.85
яну.88
сеп.90
май.93
яну.96
сеп.98
май.01
яну.04
сеп.06
май.09
яну.12
сеп.14
май.17
яну.20
461-месечен цикъл

Seasonality



Autocorrelation

- Autocorrelation function (ACF)

$$R_{y_t, y_{t-i}}$$

- Partial autocorrelation function (PACF)

$$R_{y_t, y_{t-i} | y_{t-j}}, j < i$$

Main models

- Regression
- Autoregression
- Mixed models of regression and autoregression

Regression models

$$\hat{y}_t = f(t)$$

$$\hat{y}_t = f(t, x)$$

Autoregression models

$$\hat{y}_t = f(y_{t-i})$$

$$\hat{y}_t = f\left(y_{t-i}, x_{t-j}\right)$$

Mixed models of regression and autoregression

$$\hat{y}_t = f(t, y_{t-i}, x_{t-j})$$

Feature engineering

Most often operations

- lags
- rolling window statistics
- datetime
- outliers low frequency filter
- Harmonic decomposition

Deriving lagged variables

Variables with a time delay compared to the others. Variable shifted in time.

- used in time series analysis to model the relationships between variables over time
- used to analyze the relationship between a variable and its past values

Methods

- **shift function in pandas**
- **Henkel matrix** - Strongly recommended universal method

```
In [200]: # create a lagged variable with a time shift of 1 day
df['lagged'] = df['value'].shift(1)

print(df)
```

```
   value  lagged
0      1     NaN
1      2     1.0
2      3     2.0
3      4     3.0
4      5     4.0
```

```
import numpy as np

# Generate random time series data with 20 observations
data = np.random.rand(20)

# Define the maximum lag we want to include in our lagged features
max_lag = 5

# Create a Henkel matrix with lagged features
henkel_matrix = np.zeros((len(data), max_lag+1))

for i in range(max_lag+1):
    henkel_matrix[i:len(data), i] = data[0:len(data)-i]
henkel_matrix=henkel_matrix.round(3)
```

```
# Print the Henkel matrix
print(henkel_matrix)
```

```
[[0.74  0.    0.    0.    0.    0.   ]
 [0.497 0.74  0.    0.    0.    0.   ]
 [0.586 0.497 0.74  0.    0.    0.   ]
 [0.061 0.586 0.497 0.74  0.    0.   ]
 [0.617 0.061 0.586 0.497 0.74  0.   ]
 [0.657 0.617 0.061 0.586 0.497 0.74 ]
 [0.859 0.657 0.617 0.061 0.586 0.497]
 [0.569 0.859 0.657 0.617 0.061 0.586]
 [0.905 0.569 0.859 0.657 0.617 0.061]
 [0.834 0.905 0.569 0.859 0.657 0.617]
 [0.568 0.834 0.905 0.569 0.859 0.657]
 [0.847 0.568 0.834 0.905 0.569 0.859]
 [0.026 0.847 0.568 0.834 0.905 0.569]
 [0.818 0.026 0.847 0.568 0.834 0.905]
 [0.961 0.818 0.026 0.847 0.568 0.834]
 [0.207 0.961 0.818 0.026 0.847 0.568]
 [0.57  0.207 0.961 0.818 0.026 0.847]
 [0.954 0.57  0.207 0.961 0.818 0.026]
 [0.237 0.954 0.57  0.207 0.961 0.818]
 [0.474 0.237 0.954 0.57  0.207 0.961]]
```

Rolling window statistics

Sample windows

- used in time series analysis to reduce the dimensionality of the data
- capture relevant patterns over a specific time interval

Method

- defining a fixed-length sample window
- extract a set of features from each window
- size of the sample window is an important hyperparameter
- it should be chosen based on the characteristics of the time series data and the specific prediction problem at hand.

```
# Define the window size for the rolling statistics
window_size = 3
# Calculate rolling mean, standard deviation, and maximum
rolling_mean = series.rolling(window_size).mean()
rolling_std = series.rolling(window_size).std()
rolling_max = series.rolling(window_size).max()
```

	Original data	Rolling mean	Rolling standard deviation	Rolling maximum
0	0.076313	NaN	NaN	NaN
1	0.264040	NaN	NaN	NaN
2	0.675782	0.338712	0.306631	0.675782
3	0.068876	0.336233	0.309826	0.675782
4	0.806467	0.517042	0.393585	0.806467
5	0.705469	0.526937	0.399894	0.806467
6	0.756620	0.756185	0.050500	0.806467
7	0.018057	0.493382	0.412437	0.756620
8	0.089027	0.287901	0.407471	0.756620
9	0.579511	0.228865	0.305734	0.579511
10	0.527292	0.398610	0.269375	0.579511
11	0.970188	0.692330	0.242044	0.970188
12	0.485930	0.661137	0.268444	0.970188
13	0.957106	0.804408	0.275888	0.970188
14	0.128065	0.523700	0.415809	0.957106
15	0.372937	0.486036	0.425935	0.957106

Datetime index operations

Re-scaling

- manipulating the index of DataFrame to a new scale of dates

```
import pandas as pd

# create a DataFrame with a datetime index
date_rng = pd.date_range(start='1/1/2020', end='1/20/2020', freq='D')
df = pd.DataFrame(date_rng, columns=['date'])
df['data'] = np.random.randint(0,100,size=(len(date_rng)))

# change the frequency to weekly and take the mean of each group
df = df.set_index('date')
weekly_df = df.resample('W').mean()
weekly_df
```

	data
date	
2020-01-05	59.600000
2020-01-12	70.857143
2020-01-19	42.857143
2020-01-26	95.000000

Datetime index operations

Re-framing

- fill in the missing dates with some specified fill value.

```
# fill in the missing dates with NaN values
df = df.set_index('date')
df_new = df.asfreq('D')
df_new
```

	data
date	
2020-01-01	54.0
2020-01-02	67.0
2020-01-03	42.0
2020-01-04	NaN
2020-01-05	60.0
2020-01-06	22.0
2020-01-07	99.0

Datetime index operations

Extracting datetime features

- using the full datetime string to brake down into features

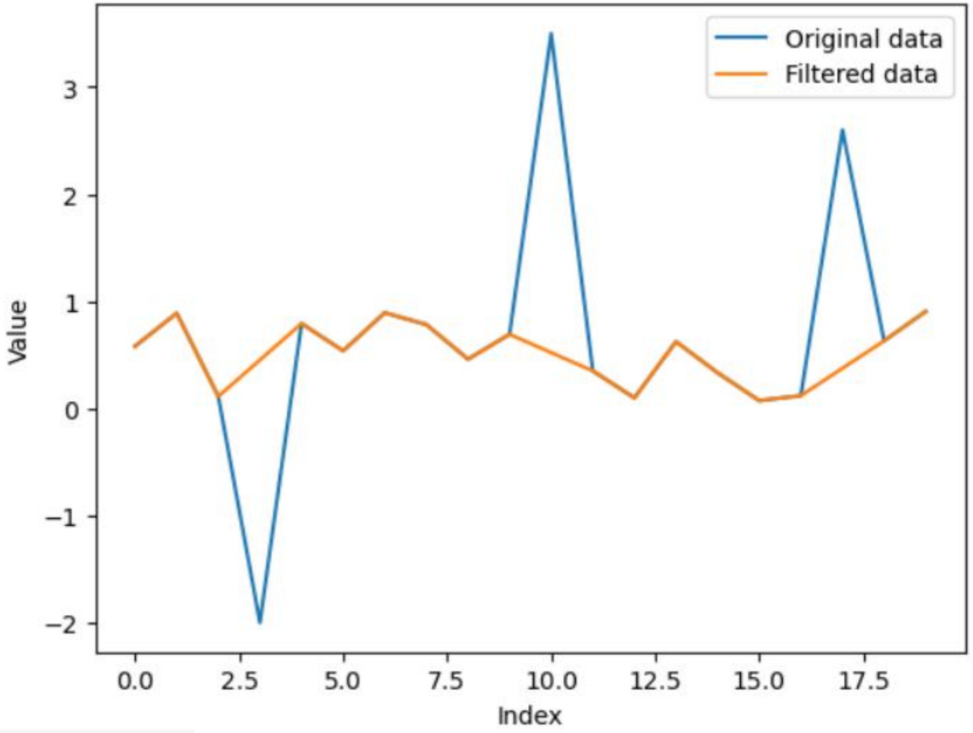
```
# Convert the data to a Pandas Series with DatetimeIndex
series = pd.Series(data, index=date_range)

# Extract calendar and time base features from the index
year = series.index.year
month = series.index.month
day = series.index.day
hour = series.index.hour
minute = series.index.minute
```

	Date	Data	Year	Month	Day	Hour	Minute
0	2022-01-01 00:00:00	0.114295	2022	1	1	0	0
1	2022-01-01 01:00:00	0.499400	2022	1	1	1	0
2	2022-01-01 02:00:00	0.316746	2022	1	1	2	0
3	2022-01-01 03:00:00	0.901192	2022	1	1	3	0
4	2022-01-01 04:00:00	0.531030	2022	1	1	4	0
5	2022-01-01 05:00:00	0.792617	2022	1	1	5	0
6	2022-01-01 06:00:00	0.100412	2022	1	1	6	0
7	2022-01-01 07:00:00	0.187317	2022	1	1	7	0
8	2022-01-01 08:00:00	0.786790	2022	1	1	8	0
9	2022-01-01 09:00:00	0.497147	2022	1	1	9	0
10	2022-01-01 10:00:00	0.138009	2022	1	1	10	0

Outliers low frequency filter

- Similar to panel data case
- but it could be implemented to be a streaming process
- IQR




```
# Convert the data to a Pandas Series
series = pd.Series(data)

# Calculate the first and third quartiles
q1 = series.quantile(0.25)
q3 = series.quantile(0.75)

# Define the filter based on the interquartile range (IQR)
iqr = q3 - q1
filter = (series >= q1 - 1.5*iqr) & (series <= q3 + 1.5*iqr)

# Filter the data
filtered_data = series[filter]
```

Harmonics decomposition

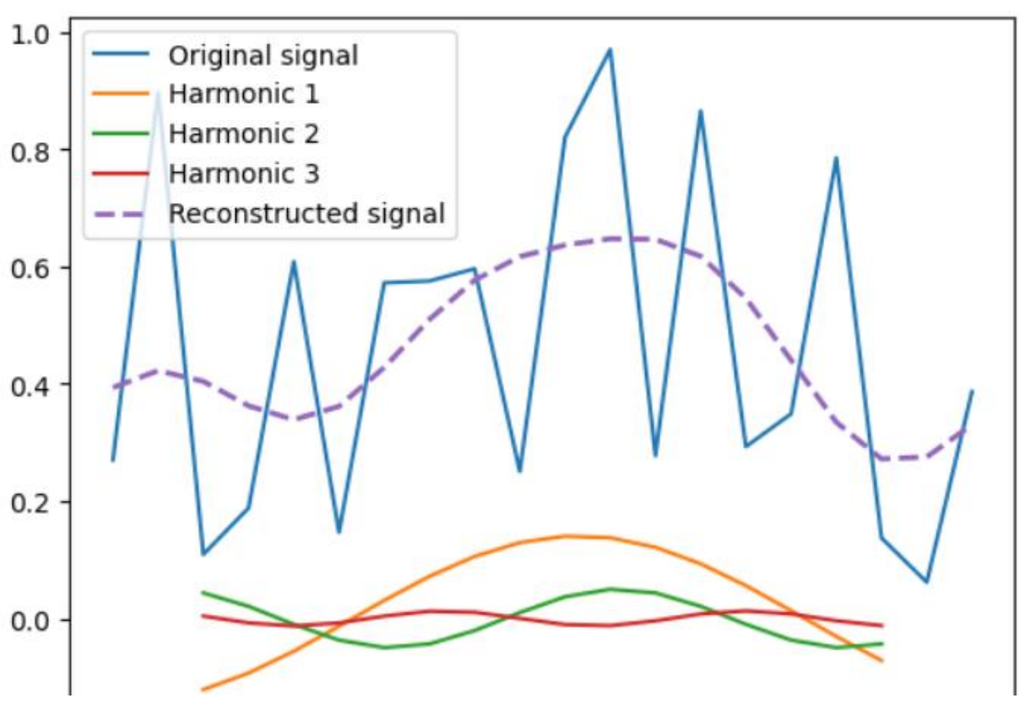
Extract seasonality from a time series, decomposing them into its trend, seasonal, and residual components.

Fourier

- Decompose a signal into its frequency components
- based on the Fourier series
- any periodic function can be represented as a sum of sine and cosine waves of different frequencies, phases, and amplitudes
- the time series data is first transformed into the frequency domain using a Fourier transform
- The amplitudes and phases of these waves are then estimated using a least-squares regression

```
# Calculate the Fourier coefficients for each harmonic separately
num_harmonics = 3
all_coeffs = np.fft.fft(series)
coeffs = []
for i in range(1, num_harmonics+1):
    coeffs.append(np.zeros(len(all_coeffs), dtype=complex))
    coeffs[-1][i] = all_coeffs[i]
    coeffs[-1][-i] = all_coeffs[-i]

# Reconstruct the signal using the first 3 harmonics
reconstructed_coeffs = np.zeros(len(all_coeffs), dtype=complex)
for i in range(num_harmonics):
    reconstructed_coeffs += coeffs[i]
reconstructed_signal = np.fft.ifft(reconstructed_coeffs).real
reconstructed_signal += series.mean()
```



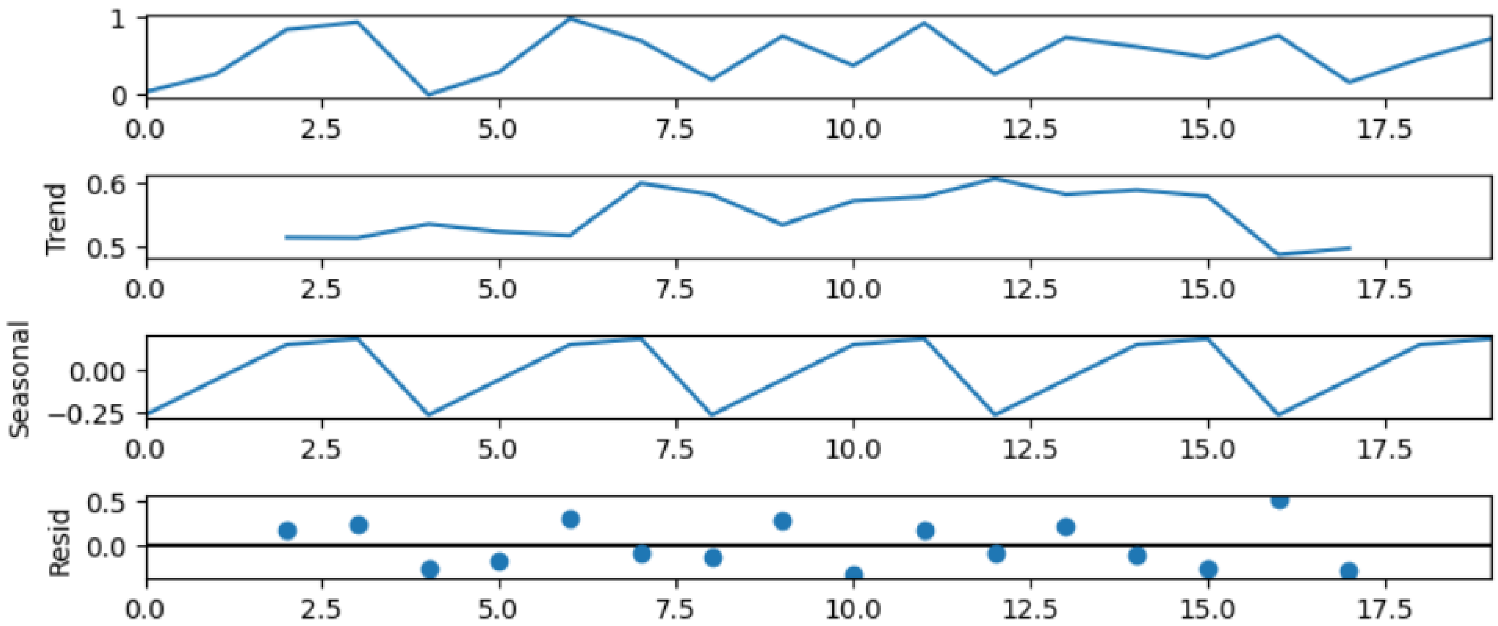
Harmonics decomposition

Seasonality analysis

- uses the classical time series decomposition method based on moving averages

```
# Perform the decomposition
decomposition = sm.tsa.seasonal_decompose(series, model='additive', per

fig=decomposition.plot();
fig.set_size_inches((8, 3.5));
fig.tight_layout();
```



Approaches for estimation of coefficients

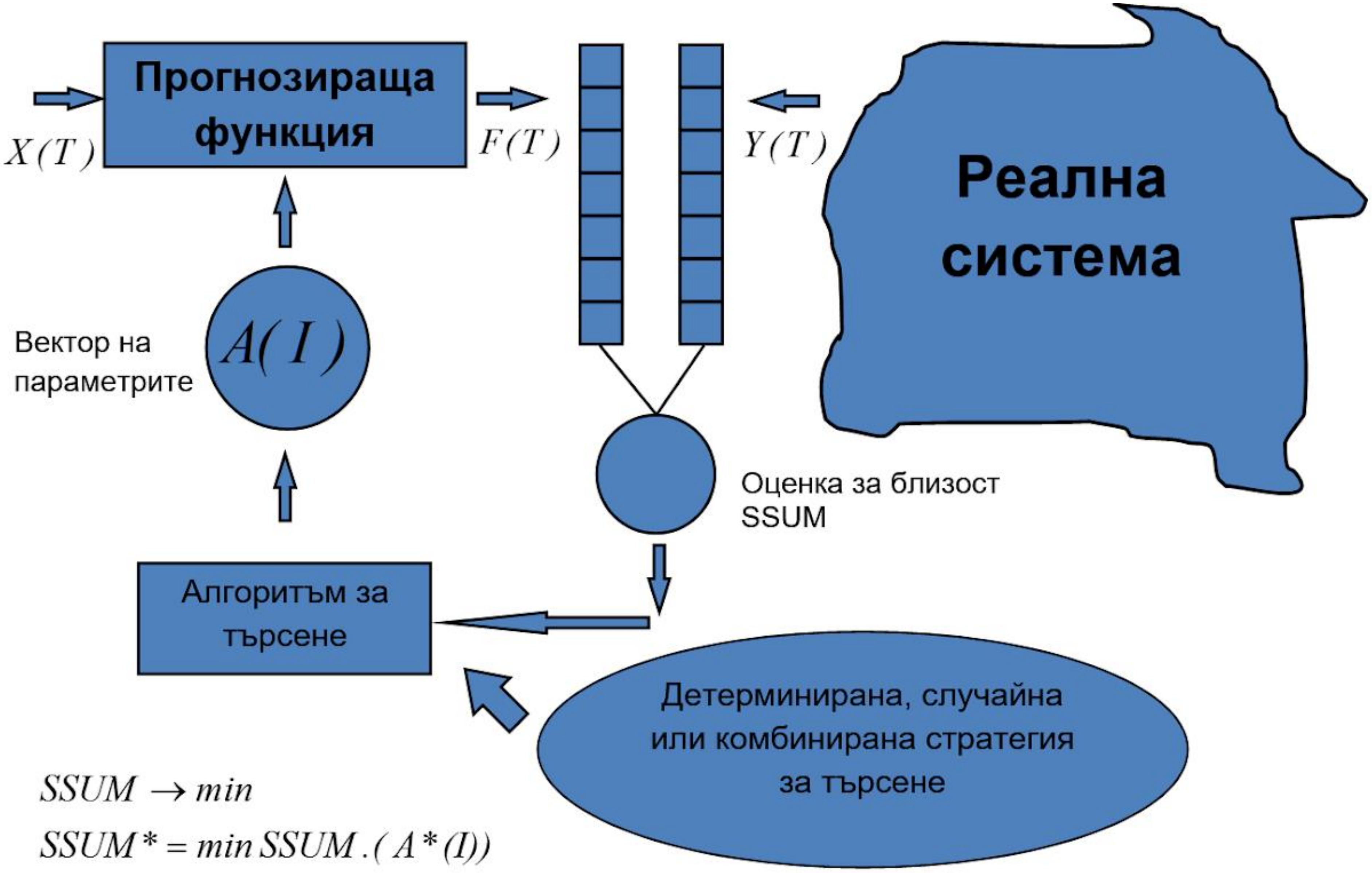
- Analytical
  - Ordinary least squares (OLS)
  - Maximum likelihood (ML)
  - Bayesian

- Iterative…

- but…

- All of these are in fact optimization

Optimization



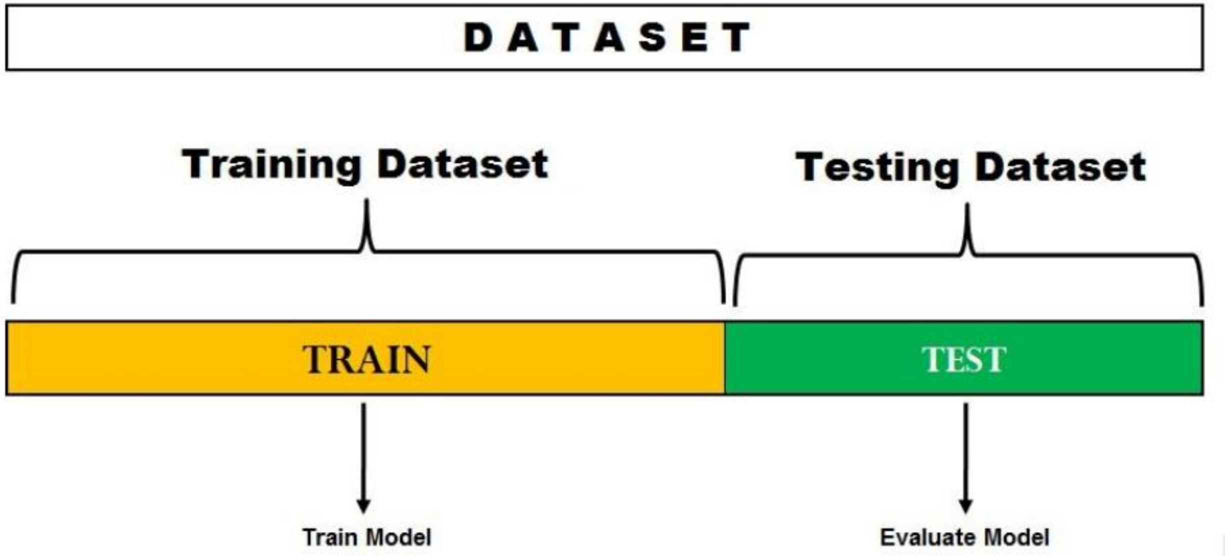
$$SSUM \rightarrow min$$

$$SSUM^* = min\, SSUM.(A^*(I))$$

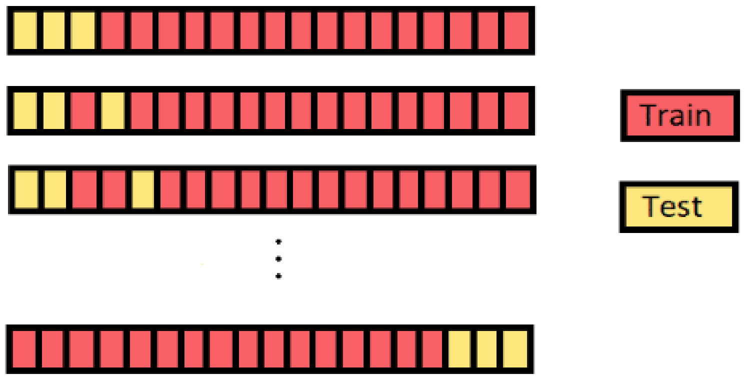
Validation

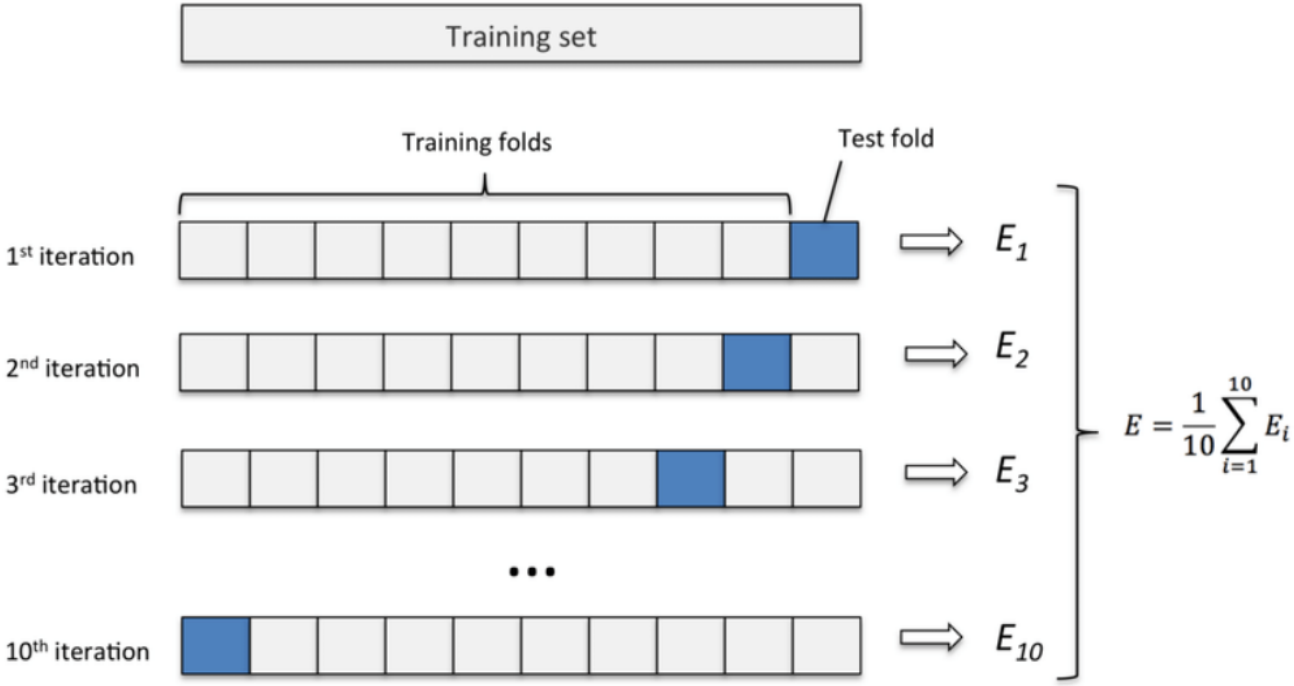
- Split sample validation
  - Training set
  - Test set

  - Validation subset/method



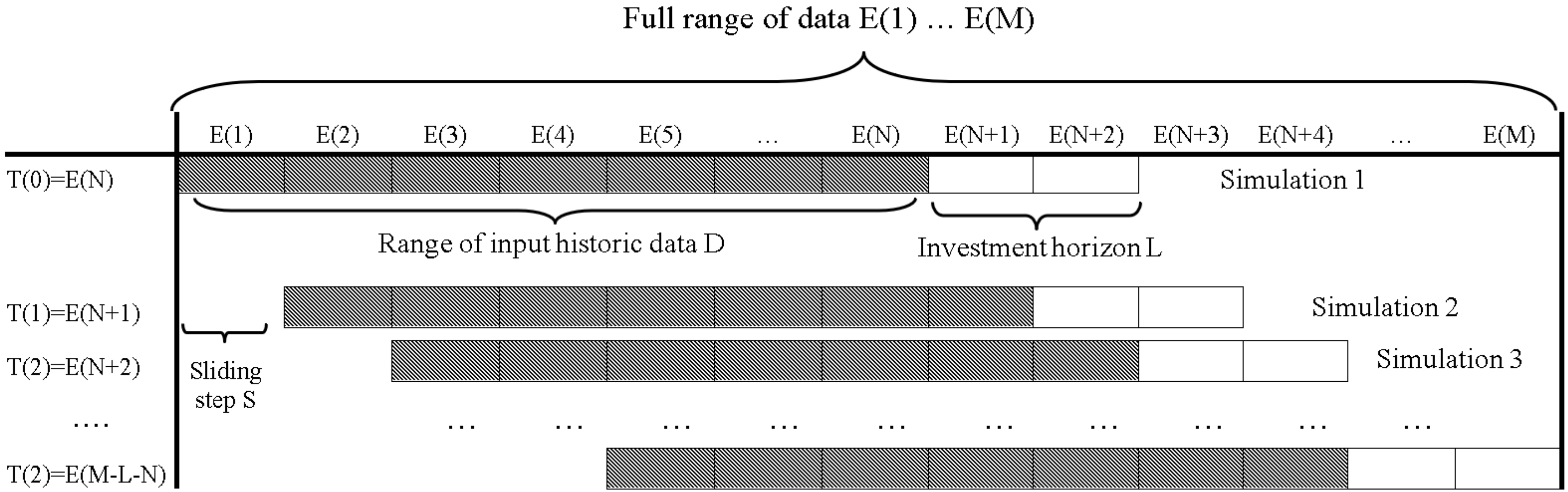
Leave P-out Cross Validation





Validation

- Validation with moving window



Overfitting

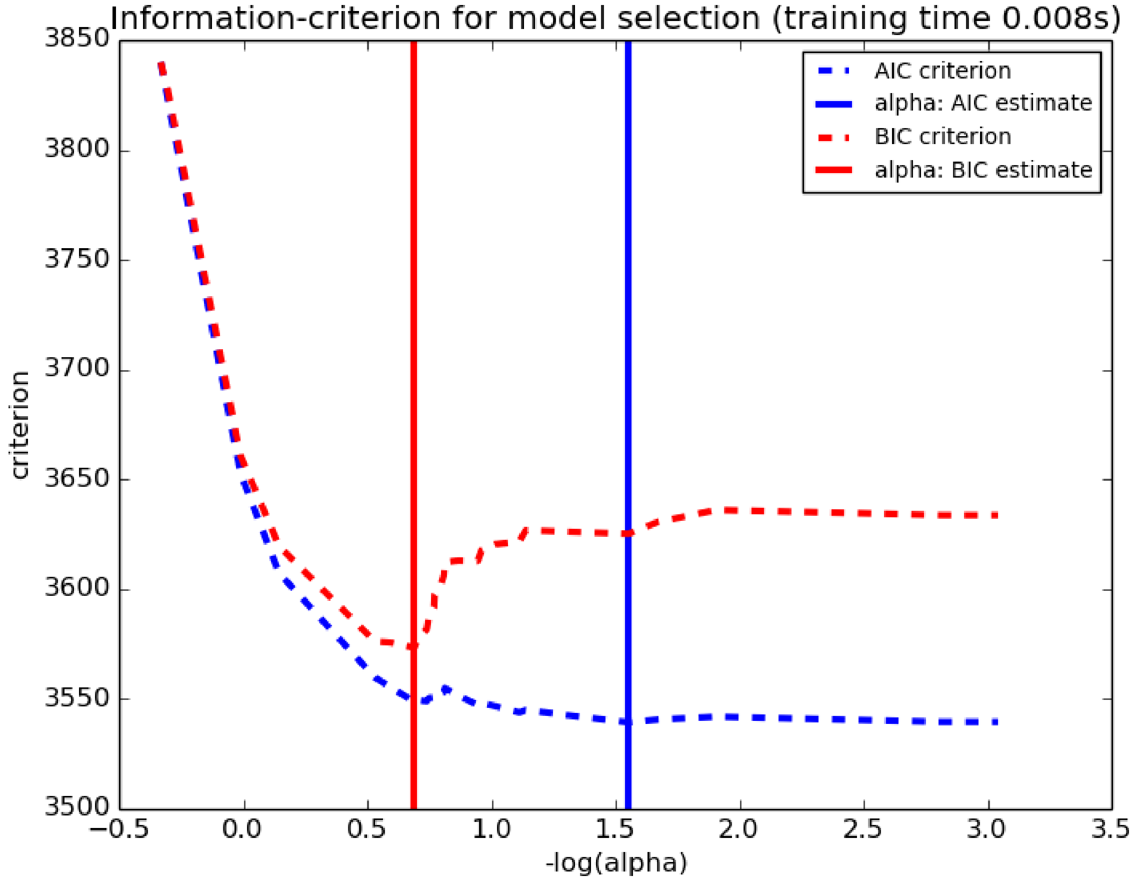
- Akaike information criterion (AIC)

$$AIC = 2k - 2.\ln(\hat{L})$$

- Bayesian information criterion (BIC) or Schwarz information criterion

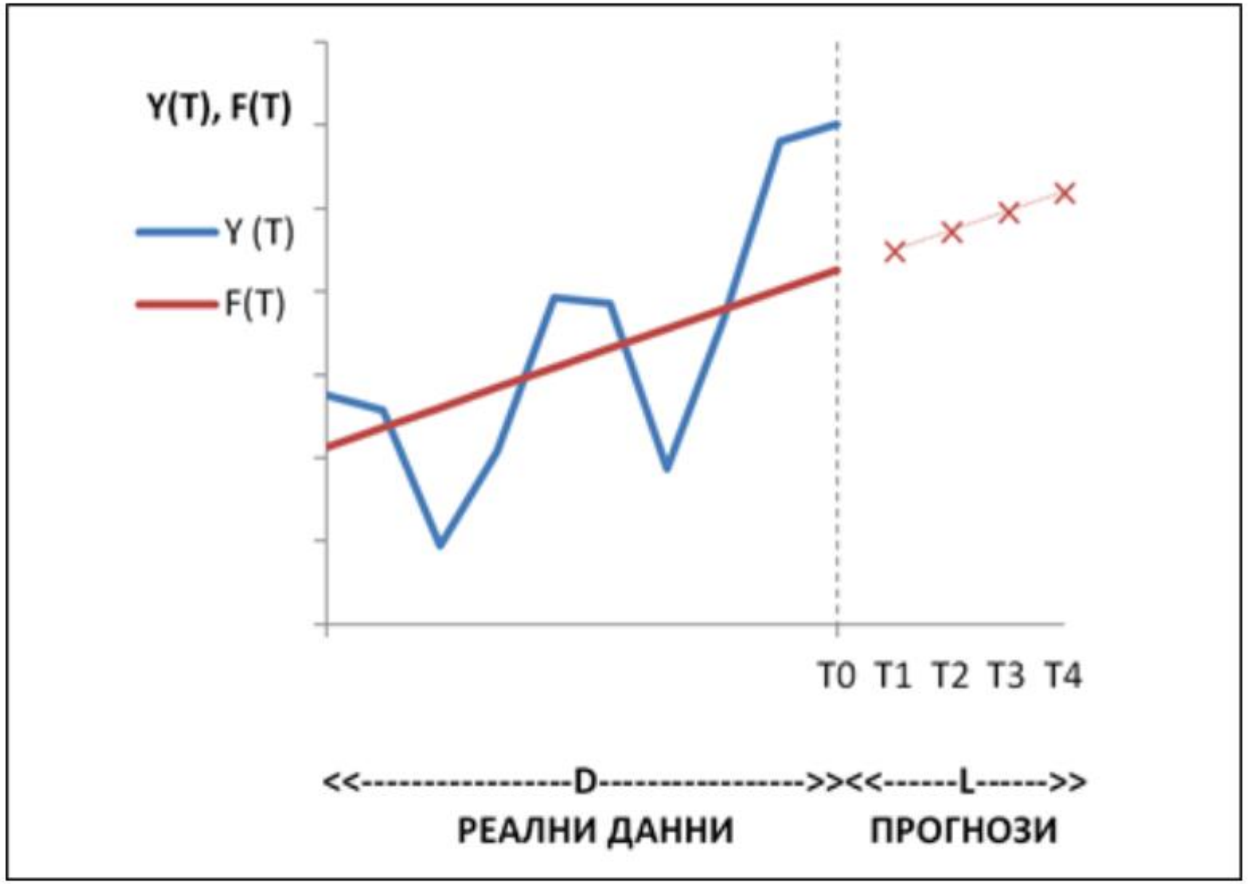
$$BIC = k.ln(n) - 2.\ln(\hat{L})$$

$$AIC/BIC = min$$



Forecasts

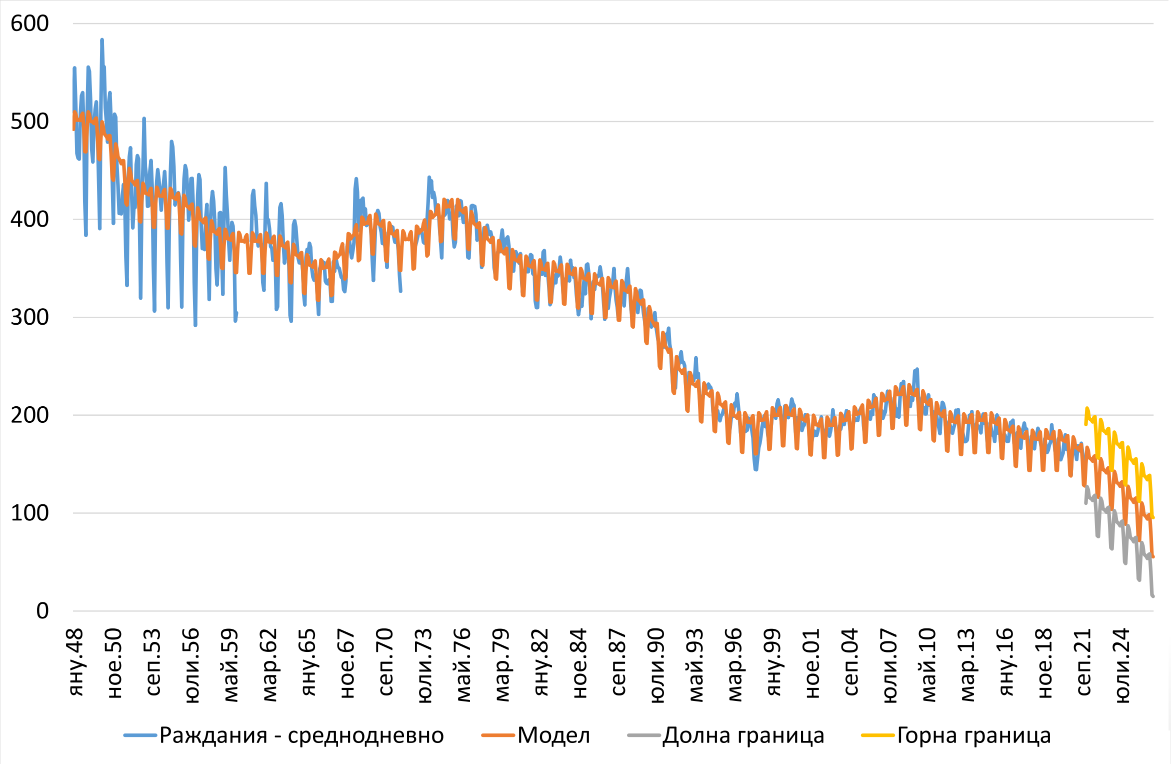
- Extrapolation/interpolation
- Analytical forecasts
- Target forecasts

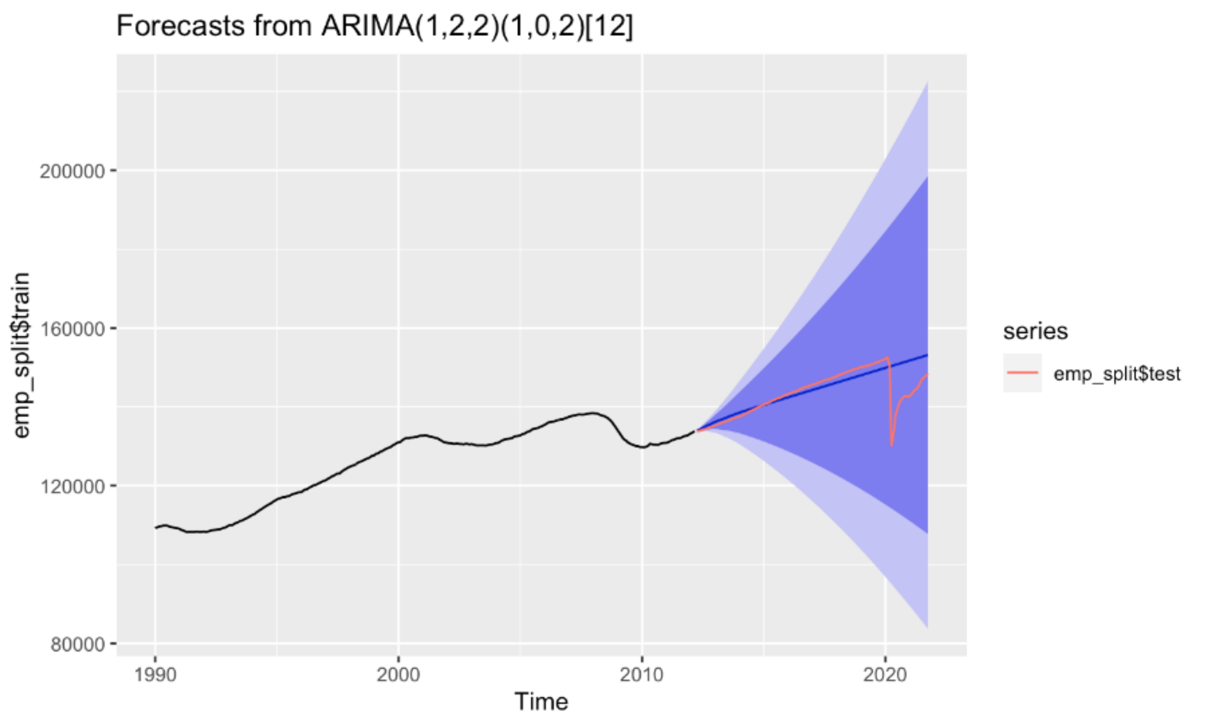


Forecast horizon

- Short term, medium term, long term
- Depending of time series length

Confidence interval of forecast





Evaluation of forecast

- Mean squared error (MSE)

$$MSE = \frac{\sum(y - \hat{y})^2}{n}$$

- Root mean squared error (RMSE)

$$RMSE = \sqrt{MSE} = \sqrt{\frac{\sum(y - \hat{y})^2}{n}}$$

- Mean absolute error (MAE)

$$MAE = \frac{\sum |y - \hat{y}|}{n}$$

Bayesian estimation of model coefficients

$$y_i = f(t_i) + \varepsilon_i$$

$$P(a_k\sigma^2|DI) = \frac{P(D|a_k\sigma^2 I)P(a_k\sigma^2|I)}{\sum_{l=1}^{m}[P(D|a_l\sigma^2 I)P(a_l\sigma^2|I)]}$$

$$P(a_k\sigma^2|I) = const$$

$$P(a_k\sigma^2|DI) = \frac{P(D|a_k\sigma^2 I).const}{\sum_{l=1}^{m}[P(D|a_l\sigma^2 I).const]} = \frac{P(D|a_k\sigma^2 I)}{\sum_{l=1}^{m} P(D|a_l\sigma^2 I)} \propto P(D|a_k\sigma^2 I)$$

$$= \prod_{i=1}^{n}\left[\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(\varepsilon_i-0)^2}{2\sigma^2}}\right] = \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^n}e^{-\frac{\sum_{i=1}^{n}\varepsilon_i^2}{2\sigma^2}}$$

Bayesian estimation of model coefficients

$$P(a_k \sigma^2 | DI) = max \rightarrow P(D | a_k \sigma^2 I) = max$$

$$\left| \begin{array}{l} \dfrac{dP(D|a_k\sigma^2 I)}{da_k} = 0 \\ \dfrac{dP(D|a_k\sigma^2 I)}{d\sigma^2} = 0 \end{array} \right.$$

$$\left| \begin{array}{c} \displaystyle\sum_{i=1}^{n} \left[y_i \frac{df(t_i)}{da_k} \right] = \sum_{i=1}^{n} \left[f(t_i) \frac{df(t_i)}{da_k} \right] \\ \sigma^2 = \dfrac{\sum_{i=1}^{n} \varepsilon_i^2}{n} \end{array} \right.$$

Bayesian estimation of model coefficients

$$P(D|a_k\sigma^2 I) = \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^n} e^{-\frac{\sum_{i=1}^{n}\varepsilon_i^2}{2\sigma^2}} = \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^n} e^{-\frac{n\sigma^2}{2\sigma^2}} = \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^n} e^{-\frac{n}{2}}$$

$$= \frac{1}{\left(\sqrt{2\pi e\sigma^2}\right)^n}$$

$$BIC = m.ln(n) - 2.ln\frac{1}{\left(\sqrt{2\pi e\sigma^2}\right)^n} = m.ln(n) - 2.ln(2\pi e\sigma^2)^{-\frac{n}{2}}$$

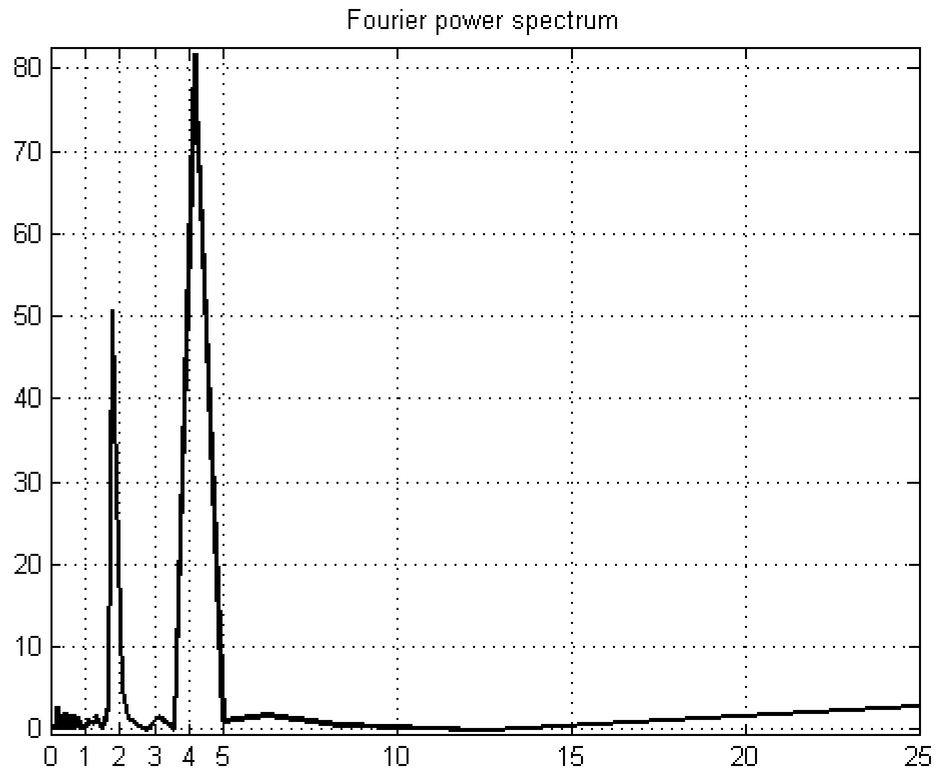
$$= m.ln(n) + n.ln(2\pi e\sigma^2)$$

Periodogram analysis

$$f(t) = a_0 + \sum_{j=1}^{\infty}\left(a_j cos\frac{2\pi j}{n}t + b_j sin\frac{2\pi j}{n}t\right)$$

$\omega_j = \frac{2\pi j}{n}$ - frequency

$T_j = \frac{2\pi}{\omega} = \frac{n}{j}$ - period



[An Interactive Introduction to Fourier Transforms (jezzamon.com)]

Bayesian periodogram analysis

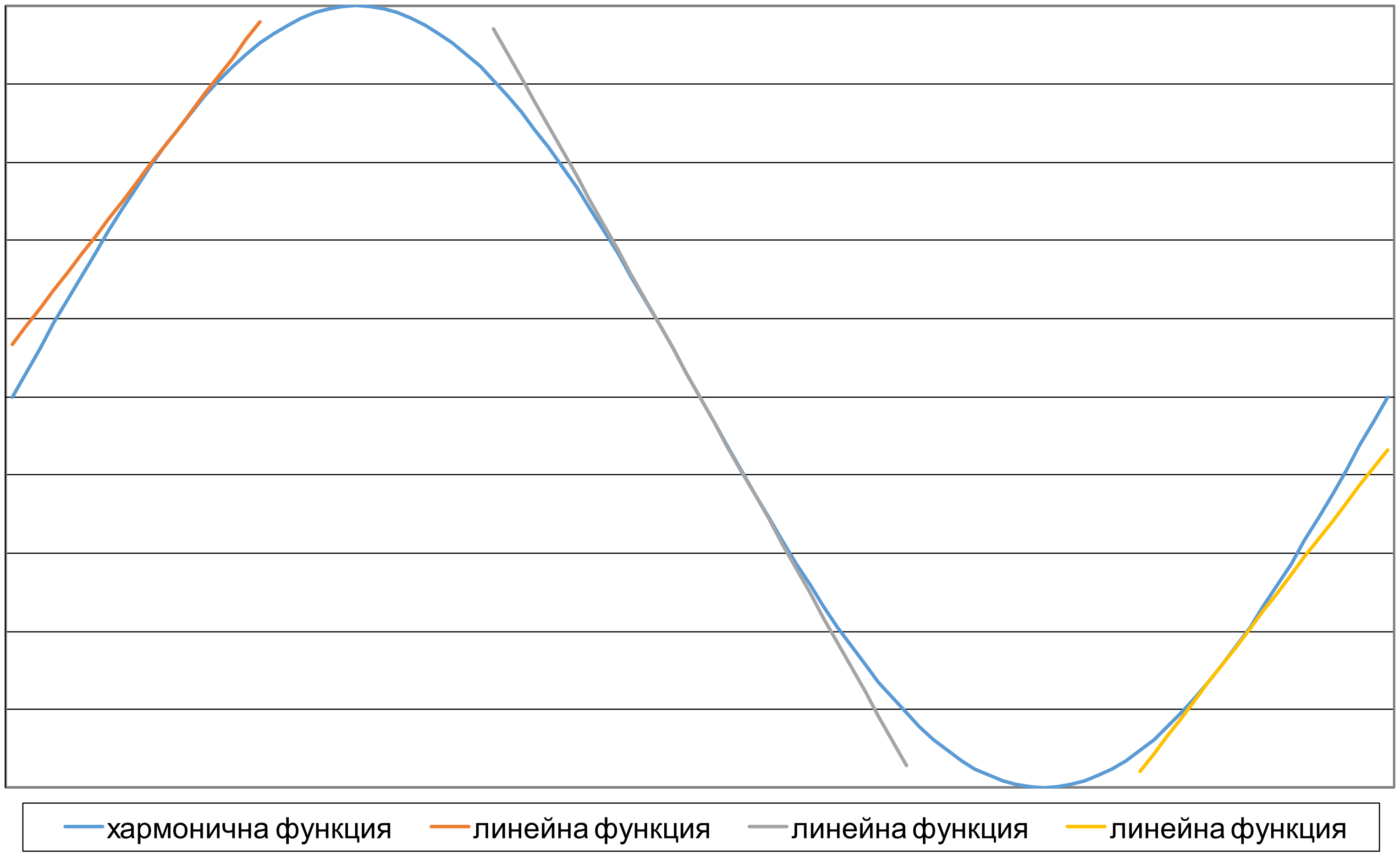
$$f(t) = \sum_{j=1}^{\infty}\left(a_j cos\frac{2\pi t}{T_j} + b_j sin\frac{2\pi t}{T_j}\right)$$

$$a.\cos(x) + b.\sin(x) = \sum_{i=0}^{\infty}\left[(-1)^i\frac{x^{2i}}{(2i)!}\left(a + b\frac{x}{2i+1}\right)\right]$$

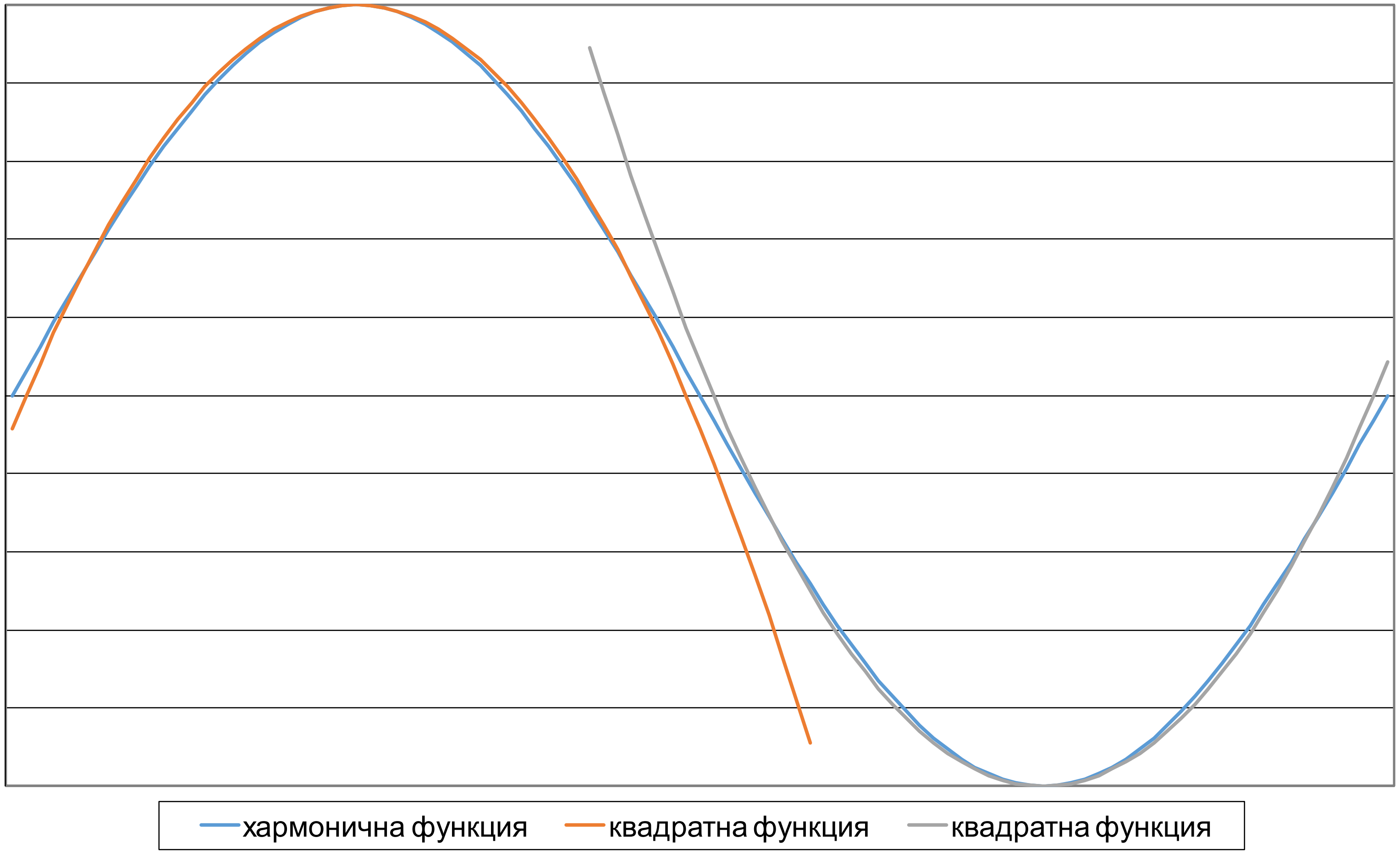
$$a.cos(x) + b.sin(x) = a + bx + \sum_{i=1}^{\infty}\left[(-1)^i\frac{x^{2i}}{(2i)!}\left(a + b\frac{x}{2i+1}\right)\right]$$

$$a.cos(x) + b.sin(x) = a + bx - \frac{a}{2}x^2 - \frac{b}{6}x^3 + \sum_{i=2}^{\infty}\left[(-1)^i\frac{x^{2i}}{(2i)!}\left(a + b\frac{x}{2i+1}\right)\right]$$

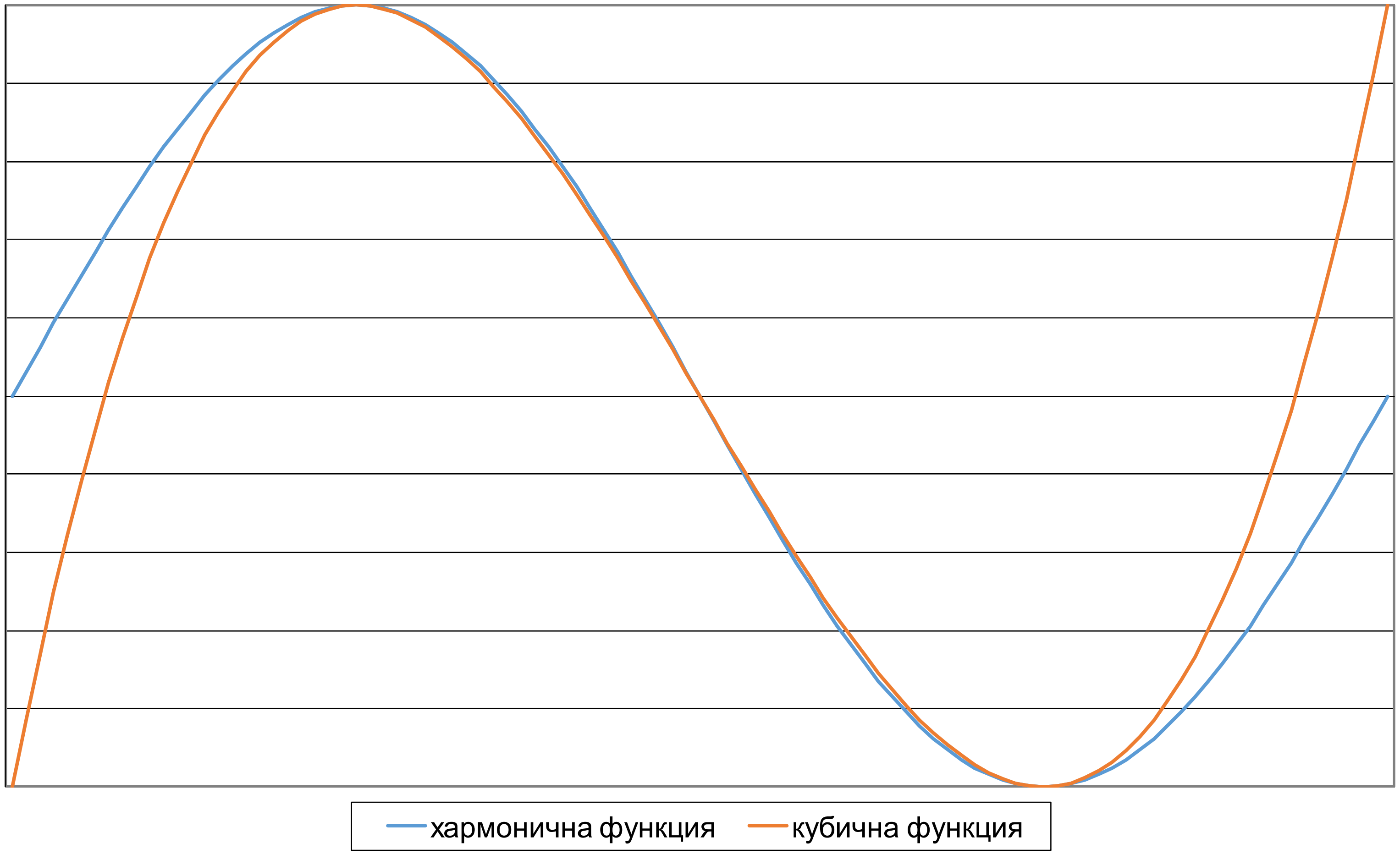
Approximation of linear function



Approximation of quadratic function



Approximation of cubic function



Components of dynamics

- If $T_j > t_n - t_1 + 1 \rightarrow Trend$

- If $T_j < \frac{2}{3}(t_n - t_1 + 1) \rightarrow Cycle$

- If $\frac{2}{3}(t_n - t_1 + 1) < T_j < t_n - t_1 + 1 \rightarrow$ $Grey\ zone\ (insufficient\ data)$

- If $T_j \leq 12\ months\ \rightarrow Seasonality$

Bayesian periodogram analysis

$$f(t_i) = a_1 cos\frac{2\pi t_i}{T_1} + b_1\, sin\frac{2\pi t_i}{T_1} + a_2 cos\frac{2\pi t_i}{T_2} + b_2\, sin\frac{2\pi t_i}{T_2}$$

$$f(t_i) = \sum_{j=1}^{w}\left[A_{1j} sin\left(\frac{2\pi t_i}{T_{1j}} + \varphi_{1j}\right) + A_{2j} sin\left(\frac{2\pi t_i}{T_{2j}} + \varphi_{2j}\right)\right]$$

$$f(t_i) = \sum_{j=1}^{w}\left[A_{1j} sin\frac{2\pi(t_i - t_{0,1j})}{T_{1j}} + A_{2j} sin\frac{2\pi(t_i - t_{0,2j})}{T_{2j}}\right]$$

Confidence intervals of forecast

$$P\left(\hat{f}_i \middle| f_i I\right) = \frac{C_{\hat{f}_1}^{f_1} C_{\hat{f}_2}^{f_2} \ldots C_{\hat{f}_m}^{f_m}}{C_{N+m-1}^{N-n}}$$

$$N = n + 1$$

$$\hat{f}_k = f_k + 1$$

$$\hat{f}_i = f_i; i \neq k$$

$$P\left(\hat{f}_i \middle| f_i I\right) = \frac{C_{f_1}^{f_1} C_{f_2}^{f_2} \ldots C_{f_k+1}^{f_k} \ldots C_{f_m}^{f_m}}{C_{n+1+m-1}^{n+1-n}} = \frac{C_{f_k+1}^{f_k}}{C_{n+m}^{1}} = \frac{f_k + 1}{n + m}$$

Confidence intervals of forecast

$$\frac{f_2 + 1}{n + 3} = P$$

$$f_2 = P(n + 3) - 1$$

$$\frac{f_{1,3} + 1}{n + 3} = \frac{1 - P}{2}$$

$$f_{1,3} = \frac{1 - P}{2}(n + 3) - 1$$

Confidence intervals of forecast

$$\varepsilon_{[f_1]} \le \varepsilon_{LL} \le \varepsilon_{[f_1]+1}$$
$$\varepsilon_{LL} = \varepsilon_{[f_1]} + \left(\varepsilon_{[f_1]+1} - \varepsilon_{[f_1]}\right)(f_1 - [f_1])$$
$$y_{n+i,LL} = f(t_{n+i}) + \varepsilon_{LL}$$

$$\varepsilon_{n-[f_3]} \le \varepsilon_{UL} \le \varepsilon_{n-[f_3]+1}$$
$$\varepsilon_{UL} = \varepsilon_{n-[f_3]+1} - \left(\varepsilon_{n-[f_3]+1} - \varepsilon_{n-[f_3]}\right)(f_3 - [f_3])$$
$$y_{n+i,UL} = f(t_{n+i}) + \varepsilon_{UL}$$

References

- Атанасов, А. 2018. Статистически методи за анализ на динамични редове. София: Издателски комплекс – УНСС
- Величкова, Н. 1981. Статистически методи за изучаване и прогнозиране развитието на социално-икономическите явления. София: „Наука и изкуство“
- Харалампиев, К. Бейсовски подход за оценка на хармонични колебания с променлива амплитуда. Научна конференция с международно участие „Авангардни научни инструменти в управлението”, Равда, 2014
- Bretthorst, G., 1988. Bayesian Spectrum Analysis and Parameter Estimation. Berlin, Heidelberg: Springer-Verlag
- Schwarz, G. 1978. Estimating the Dimension of a Model. The Annals of Statistics, Vol. 6, No. 2, 461-464
- Stone, M. 1977. An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike’s Criterion. Journal of the Royal Statistical Society. Series B (Methodological) , 1977, Vol. 39, No. 1 (1977), pp. 44-47
- Stone, M. 1979. Comments on Model Selection Criteria of Akaike and Schwarz. Journal of the Royal Statistical Society. Series B (Methodological) , 1979, Vol. 41, No. 2 (1979), pp. 276-278