Part 2: Prediction models

The aim of this part is describing methodologically the approach to predict the travel time at various conditions for the distance between each two virtual stops of the on-demand city transport.

SARIMAX w/ FT model

1. Introduction

The main chosen approach for the chapter is autoregressive moving average models with exogenous input of special class – SARIMAX with Fourier terms. This approach has been selected for the specific problem at hand because its ability to render various set of inputs:

- endogenous (discrete lagged values of the target variable)
- one main seasonal cyclical factor (hour of day)
- additional multiple seasonality cyclical factors (day of week, day of year)
- various exogenous factors (environment disturbances)
- error terms.

In the most general case, the autoregressive models are derived from polynomial models e.g., the chapter turns some attention to the genesis of the particular model.

2. Model derivation

2.1. Polynomial models

A polynomial model (Ljung, 1999) generalizes the concept of transfer functions to express the relationship between input, $u(t)$, output $y(t)$ and noise $e\ (t)$, using the equation:

$$A(q)y(t)=\sum_{i=1}^{nu}\frac{B_i(q)}{F_i(q)}u_i(t-nk_i)+\frac{C(q)}{D(q)}e(t) \qquad \text{<1>}$$

Where:

- A, B, C, D and F are polynomials expressed in the time shift operator q^{-1}.
- u_i is the i -th input,
- nu is the total number of inputs
- nk_i is the i-th lag input, which characterizes the autoregression lag.
- $e(t)$ - white noise dispersion.

Simpler forms of polynomial models, such as ARX, ARMAX, etc. are often used, to simplify the general structure.

There is also the option to introduce an integrator in the noise source so that the overall model has the form:

$$A(q)y(t)=\sum_{i=1}^{nu}\frac{B_i(q)}{F_i(q)}u_i(t-nk_i)+\frac{C(q)}{D(q)}\frac{1}{1-q^{-1}}e(t) \quad <2>$$

To estimate polynomial models, one could use time or frequency domain data. First the order of the model has to be specified as a set of integers that represent the number of coefficients for each polynomial you include in your chosen structure - *na* for *A*, *nb* for *B*, *nc* for *C*, *nd* for *D* and *nf* for *F*, *nk* denoting the input lags is defined as the number of samples before the output corresponds to the input.

The number of coefficients in the denominator polynomials is equal to the number of poles, and the number of coefficients in the numerator polynomials of the model is equal to the number of zeros plus 1. When the dynamics from *u(t)* to *y(t)* contains a delay of *nk* samples, then the first *nk* coefficients of *B* are zeros.

The general polynomial equation is used with respect to the time shift operator q^{-1}*as* the following as discrete time difference equation:

$$y(t)+a_1y(t-T)+a_2y(t-2T)=b_1u(t-T)+b_2u(t-2T) \quad <3>$$

Where:

- *y(t)* is the output,
- *u(t)* is the input, and
- *T* is the sampling time.
- q^{-1} is a time shift operator that compactly represents the difference equations using $q^{-1}\ u\ (t) = u\ (t - T)$:

$$\begin{aligned} &y(t)+a_1q^{-1}y(t)+a_2q^{-2}y(t)=\ b_1q^{-1}u(t)+b_2q^{-2}u(t) \\ &\Leftrightarrow A(q)y(t)=B(q)u(t) \end{aligned} \quad <4>$$

Polynomial models could encompass various configurations. These structural models are subsets of the general polynomial equation <10>. The structures of the model principally differ in how many of the polynomials are included in the structure to account for flexibility of modeling noise dynamics and characteristics.

If the model is already with identified specific structure, then the dynamics and noise may have common or different poles. *A(q)* corresponds to the poles that are common to the dynamic model and the noise model. The use of common poles for dynamics and noise is useful when interference enters the input system. F_i defines the poles unique to the system dynamics, and *D* defines the poles unique to the interference.

Specific and important modifications of polynomial (autoregressive) models:

- ARX - The noise model is reciprocal to A and the noise is related to the dynamic model. ARX does not model noise and dynamics independently.

$$A(q)y(t)=\sum_{i=1}^{nu}B_i(q)u_i(t-nk_i)+e(t) \quad <5>$$

- ARMAX: Extends the structure of the ARX by introducing more terms for noise modeling using the *C* parameters (moving average of white noise).

$$A(q)y(t)=\sum_{i=1}^{nu}B_i(q)u_i(t-nk_i)+C(q)e(t) \qquad \text{<6>}$$

- ARIMAX: Extends the structure of ARMAX to include an integrator in the noise source, $e(t)$ for cases where the disturbance is not stationary.

$$Ay=Bu+C\frac{1}{1-q^{-1}}e \qquad \text{<7>}$$

2.2. ARMAX model

The structure of the ARMAX model (Autoregressive Moving Average with additional input) is:

$$\begin{aligned}&y(t)+a_1y(t-1)+...+a_{na}y(t-na)=\\&=b_1u(t-nk)+...+b_{nb}u(t-n_k-n_b+1)+c_1e(t-1)+\ldots+c_{nc}e(t-n_c)+e(t)\end{aligned} \qquad \text{<8>}$$

A more compact way to write the difference equation is

$$A(q)y(t)=B(q)u(t-n_k)+C(q)e(t) \qquad \text{<9>}$$

Where:

- $y(t)$ - output at time t
- n_a - Number of poles
- n_b - Number of zeros plus 1
- n_c - Number of *C* coefficients
- n_k - Number of input samples that occurred before the input affected the output, also called *dead time* in the system
- $y(t-1)\ldots y(t\text{-}n_a)$ - previous outputs on which the current output depends
- $u(t\text{-}n_k)\ldots u(t\text{-}n_k\text{-}n_b+1)$ - Previous and delayed inputs on which the current output depends
- $e(t-1)\ldots e(t\text{-}n_c)$ - value of white noise interference

The parameters na, nb and nc are the orders of the ARMAX model, and nk is the delay. q is the delay operator such that,

$$\begin{aligned}A(q)&=1+a_1q^{-1}+\ldots+a_{na}q^{-na}\\B(q)&=b_1+b_2q^{-1}+\ldots+b_{nb}q^{-nb+1}\\C(q)&=1+c_1q^{-1}+\ldots+c_{nc}q^{-nc}\end{aligned} \qquad \text{<10>}$$

2.3. **ARIMA model**

As defined by Box – Jenkins (1994), the autoregressive integrated moving average (ARIMA) process generates nonstationary series that are integrated by row *D,* denoted by *I*(*D*). A non-stationary *I*(*D*) process is one that can be made stationary by taking *D* differences. Such processes are often called *different-stationary* or *single root* processes.

A sequence that could be modelled as a stationary ARMA (*p, q*) process after splitting *D* times is denoted by ARIMA (*p, D, q*):

$$\Delta D_{yt} = c + \phi_1 \Delta D_{yt-1} + \ldots + \phi_p \Delta D_{yt-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q} \quad <11>$$

where

- $\Delta^D y_t$ means the *D*-th differentiated time series
- ε_t is an uncorrelated innovation process with a mean zero.

In the delay operator notation, $L^i y_t = y_{t-i}$. ARIMA (*p, D, q*) becomes

$$\phi^*(L) y_t = \phi(L)(1-L)^{D_{yt}} = c + \theta(L)\varepsilon_t \quad <12>$$

Here $\phi^*(L)$ is a non-stationary AR-polynomial of the operator with exactly *D* single roots. A factorization of this polynomial is $\phi(L)(1-L)^D$, where $\phi(L)=(1-\phi_1 L - \ldots - \phi_p L^p)$ is a stable polynomial of the lag operator *p* AR (with all roots located outside the single circle). Similarly, $\theta(L) = (1+\theta_1 L + \ldots + \theta_p L^Q)$ is a reversible degree *p* polynomial MA delay operator (with all roots lying outside the circle unit). The signs of the coefficients in a polynomial of AR delay $\varphi(L)$ are opposed to the right side of the equation <20>.

2.4. ARIMAX model (*p, D, q*)

The autoregressive moving average model, including exogenous factors, ARMAX (*p, q*), extends the ARMA model (*p, q*) to include the linear effect that one or more exogenous series have on the stationary series (Wold, 1938) of outcome y_t. The general form of the ARMAX model (*p, q*) is

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-1} + \sum_{k=1}^{r} \beta_k x_{tk} + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} \quad <13>$$

and it has the following condensed form in the notation of the delay operator :

$$\phi(L) y_t = c + x'_t \beta + \theta(L)\varepsilon_t \quad <14>$$

Where the vector x'_t contains the values of *r* exogenous, time-varying predictors at time *t,* with coefficients denoted by *β.*

For ARIMAX we assume the response series y_t to be not stable and we difference it to form a stationary, by specifying the degrees of integration D. So the response series y_t *are* *differenced before* including the exogenous features by the degree of integration D. Then, the ARIMAX(p,D,q) model becomes

$$\phi(L)y_t = c^* + x'_t\beta + \theta^*(L)\varepsilon_t \qquad \text{<15>}$$

Where

$c^* = c/(1 - L)D$

$\theta^*(L) = \theta(L)/(1 - L)D$

Here the interpretation of β is the expected effect a unit increase in the predictor has on the difference between current and lagged values of the response.

2.5. SARIMAX model

The seasonal autoregressive integrated moving average with exogenous inputs (SARIMAX) is an extension of the ARIMAX model class, which fixes its biggest weakness – accounting for seasonality.

A SARIMAX model is written as SARIMAX (p, d, q) (P, D, Q, S) where:

- p is the order of the AR term.
- d is the order of differencing needed to make the data stationary.
- q is the order of the MA term.
- P is the order of the seasonal AR term.
- D is the order of the seasonal differencing needed to make data stationary.
- Q is the order of the seasonal MA term.
- S is the number of periods in a season.

A SARIMAX (p, d, q) (P, D, Q, S) is mathematically represented as (Cools et al., 2005):

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \dots + \beta_k x_{k,t} + \\ + \frac{\left(1-\theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q\right)\left(1-\Theta_1 B^S - \Theta_2 B^{2S} - \dots - \Theta_Q B^{QS}\right)}{\left(1-\phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p\right)\left(1-\Phi_1 B^S - \Phi_2 B^{2S} - \dots - \Phi_P B^{PS}\right)}\varepsilon_t \qquad \text{<16>}$$

where:

- yt denotes the value of the series at time t.
- $X_{1,t}$,$X_{2,t}$,…,$X_{k,t}$ denote observations of the exogenous variables.
- β_0,β_1,…,β_k denote the parameters of the regression part.
- ϕ_1,ϕ_2,…,ϕ_p denote the weight of the nonseasonal autoregressive terms.
- Φ_1,Φ_2,…,Φ_P denote the weight of the seasonal autoregressive terms.
- θ_1,θ_2,…,θ_q denote the weight of the nonseasonal moving average terms.

- $\Theta_1, \Theta_2, \ldots, \Theta_Q$ denote the weight of the seasonal moving average terms.
- B_s denotes the backshift operator such that $B_s y_t = y_{t-s}$.
- ε_t denotes the white noise terms.

2.6. SARIMAX with Fourier Terms

The SARIMAX model is designed to deal with a single seasonality factor. To work for multiple seasonality, it is possible to apply a method called Fourier terms. For two (or more) specific seasonalities to account for, they are added as regressors in the form of trigonometric generated series – as many as needed:

$$y_t = a + \sum_{i=1}^{M} \sum_{k=1}^{K_i} \left[\alpha \sin\left(\frac{2\pi kt}{p_i}\right) + \beta \cos\left(\frac{2\pi kt}{p_i}\right) \right] + N_t \qquad \text{<17>}$$

where N_t is a SARIMAX process.

The seasonal model is modeled by adding Fourier terms that are used as external regressors. This approach is flexible and allows the inclusion of multiple periods. For example, if there are M periods in the data (p_1, p_2, p_3, p_M), there will be different Fourier series corresponding to each of the M periods.

3. Methodical procedure of SARIMAX w/ FT

Here some specifics of the application of the model for the current research are laid out. Defined are the characteristics of the desired data, specific hyperparameters and other features, within a formal algorithm structure. The procedure to implement the SARIMAX w/ FT model goes through three phases: model identification, parameter estimation/optimization, and prediction, with their corresponding sub-phases.

3.1.Identification.

This step uses the data and the knowledge regarding how it was generated to identify the model. There are three approaches to identification – based on prior domain knowledge (white box), based on statistical methods to best suit the data (black box), and combined of the previous two (gray box) – using some domain knowledge to narrow down to subclass of models and select most appropriate mathematical description based on the data. In this study the approach is gray box, and it can be broken down into the following four sub-phases:

3.1.1. Literature and data review.

After literature and data review in this particular study to narrow down the exact class of models to be used, there are three seasonal periods accepted:

- hourly within a day, which will be rendered by the seasonal component of SARIMAX;
- daily within a week

- and daily within a year.

With p_1 including 7 days and p2 including 365 days. For each period "p_i" is selected, the number of members of Fourier (K_i from <27>), to find the best statistical model for a given set of data. Given a set of models, AIC (which is derived from information theory) and BIC (which is derived from Bayesian theory) evaluates the quality of the model compared to other models and thus provides a tool for model selection. The value of K_i is chosen to minimize the AIC and BIC criteria (see below). To find the exact number of Fourier members corresponding to each of the periods, the AIC or BIC values of the SARIMAX model with variable Fourier terms are calculated. The best model that minimizes the AIC or BIC criterion is recorded as Fourier terms.

The total number of coefficients, which are calculated is equal to the sum of seasonal and non-seasonal AR and MA orders. In other words, we consider a total of "P plus Q, plus, p plus q" many coefficients.

3.1.2. Data preparation.

For the purpose of any SARIMAX model differencing operations are applied to make the time series data stationary. It simply means taking the difference between data points and its backward version. Intuitively, this is analogous to calculating the derivative. Augmented Dickey-Fuller test is used for verification.

3.1.3. Model selection.

Autocorrelation functions (ACF) and partial autocorrelation functions (PACF) are used to identify the orders p and q of the terms AR and MA. These functions explain the correlation of a value in the series with its lagged values. Since the identification process could find complex models of autocorrelation without clear interpretations, and in order to avoid any unscientific assumptions Grid Search or Stochastic search is used.

3.2.Optimization.

This step uses the data to train the model, estimate the coefficients and check the residuals to see if they adhere to the assumptions. This step can be divided into the following two steps:

3.2.1. Estimation.

Estimates are made of the parameters, and the best model is chosen based on a criterion. At this phase, a comprehensive search of all combinations of parameters is performed (hyperparameter optimization), along the chosen criterion.

Akaike's Information Criterion (AIC) is one option used for model selection. It measures the goodness of fit of a model while promoting simplicity. The AIC of a model is a relative measure and is meaningful when compared to other models. AIC is calculated as (Chatfield, 2001):

$$AIC = 2m - 2\ln(\hat{L}) \qquad <18>$$

where

- m - denotes the number of independent parameters estimated
- $\hat{L}$ - maximum likelihood

The Bayesian information criterion proposed by Schwarz is another one of the tools used to determine the maximum number of coefficients in the model (Schwarz, 1978), ie. ultimately, the final appearance of the model is determined. In this way, on one hand, a model is obtained that describes the output data as well as possible, and on the other hand, protection against the so-called overfitting.

$$BIC = m.ln(n) - 2.ln(\hat{L}) \qquad <19>$$

where

- n - denotes the number of observations

Note: although there is alternative approach (Claeskens-Hjort, 2008) here both AIC and BIC are setup for minimization.

3.2.2. Residual diagnostics.

The Ljung – Box statistical test is used to check if the residuals are aligned with the assumptions of the modelling technique. The tests on whether the residuals are white noise reveal valuable insights regarding the model. If the residuals are correlated, a more complex model is needed to capture all the information in the data. If residuals do not have a mean of zero, the forecast is biased.

3.3.Forecasting.

In this final phase, input data is used to generate forecasts from the selected and tested model, and the performance of the model is evaluated using forecast accuracy measures AIC, BIC.

4. Algorithm for implementation of the model

With business understanding and data understanding done already in general for the project, each applied model should be explained in terms of data preparation, data modeling and model validation. In this section these phases are particularly explained algorithmically for SARIMAX w/ FT and while at the same time several etalon models are developed. The software implementation is realized in Matlab (R2020a, Econometrics toolbox, Curve Fitting Toolbox) and the full code could be seen in the GitHub repository at https://github.com/InnoAir-Reserchers-Group .

The input data used in this particular modeling procedure has two sources – traffic data for times of arrival for 4 consecutive stops of a bus line in Sofia (with some additional details about the time schedule and bus stop stays) and also weather data. All data is collected between 10 January 2020 and 30 July 2021.

Several preliminary operations on the data are conducted, so that it would resemble a pseudo time-series:

- For each day there are about 13 bus courses, and this defines the time-frame observation unit of the time-series as the data for a given course (row data) along various features (column data).
- Some of the observed bus-courses were not completed in the data set – either because of particular but exceptional travel conditions, or because of data recording mishaps. Either case the incomplete courses have been dropped from further analysis – these constitute about 4% of the data.
- In order to avoid boundary effects in the data, the non-completed first and last day of data are removed, so that the data for the study contains whole-day observations.
- The weather data is pre-selected and synchronized as to correspond the dates & times of the traffic data.

4.1. Data preparation implementation

4.1.1. Data import.

The input data is delivered as datetime features in MS Excel format. Which is read into MATLAB data format (note that MATLAB date numbers start with 1 = January 1, 0000 A.D., hence there is a difference of 693960 relative to the 1900 date system of MS Excel). For practical purposes the data is read both as numerical values and as date string value.

There are 3 groups of data read for each bus stop:

– Time of arrival (i.e., variable stop_1_355, where _1 corresponds to the serial number of a stop, given for the current study and _355 as the official bus tope code).

– Time scheduled to arrive (i.e., sched_1_355 with mnemonics similar to time of arrival above).

– Time of stay at particular stop, estimated with granularity of 0.5 minute (i.e., stay_stop1 for stop 1).

Apart from the above data, as a data key the serial number of each bus course is also retained from the data.

4.1.2. Define target variables.

– The target variable is defined as the time to travel to next stop, and it is calculated as time differences of time of arrivals in consecutive bus tops. So, for the 4 bus stops under study, there are 3 times of arrival (i.e., time_stop2, where 2 corresponds the difference between time of arrival of stop1 and stop 2).

– Similarly calculated are the scheduled time to travel to next stop.

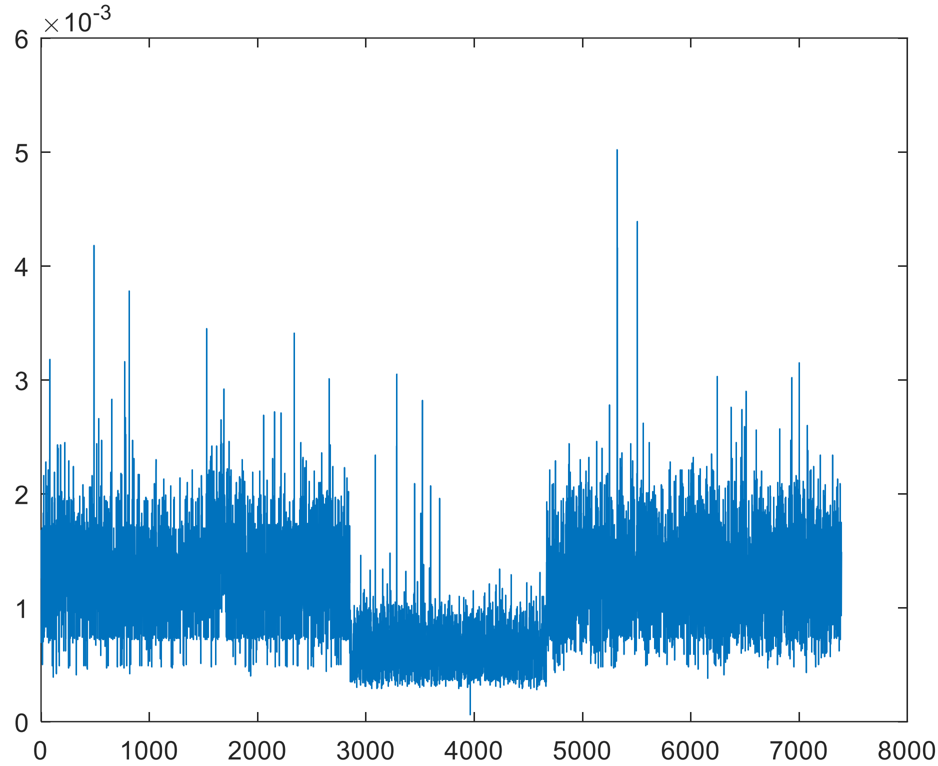

Fig. 4.1.2-1 Time of arrival for stop 2

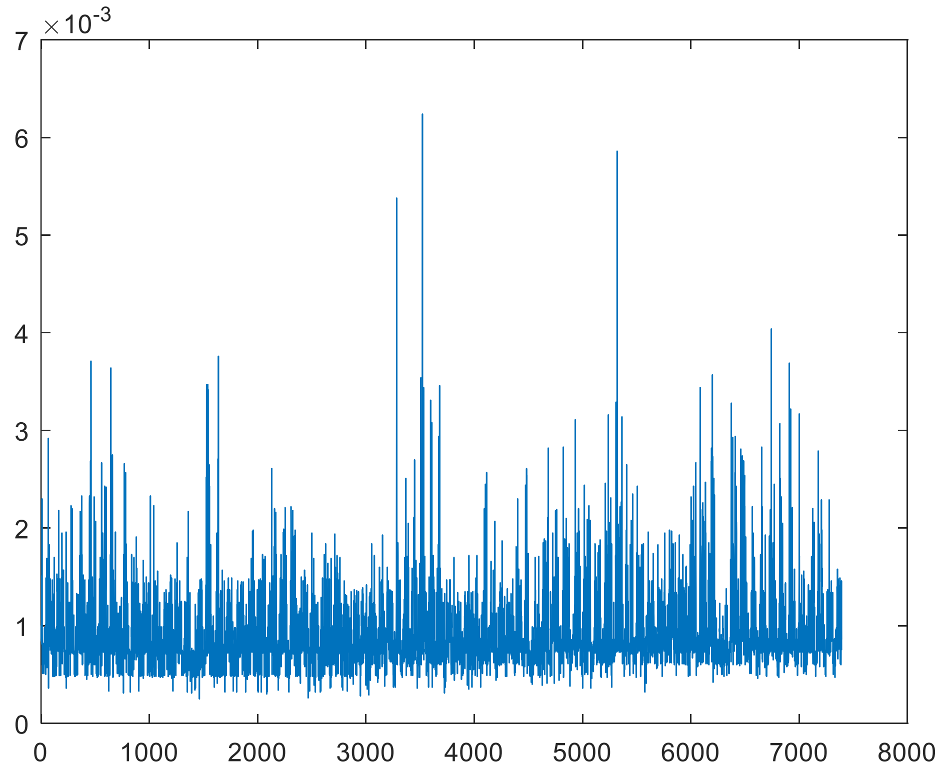

Fig. 4.1.2-2 Time of arrival for stop 3

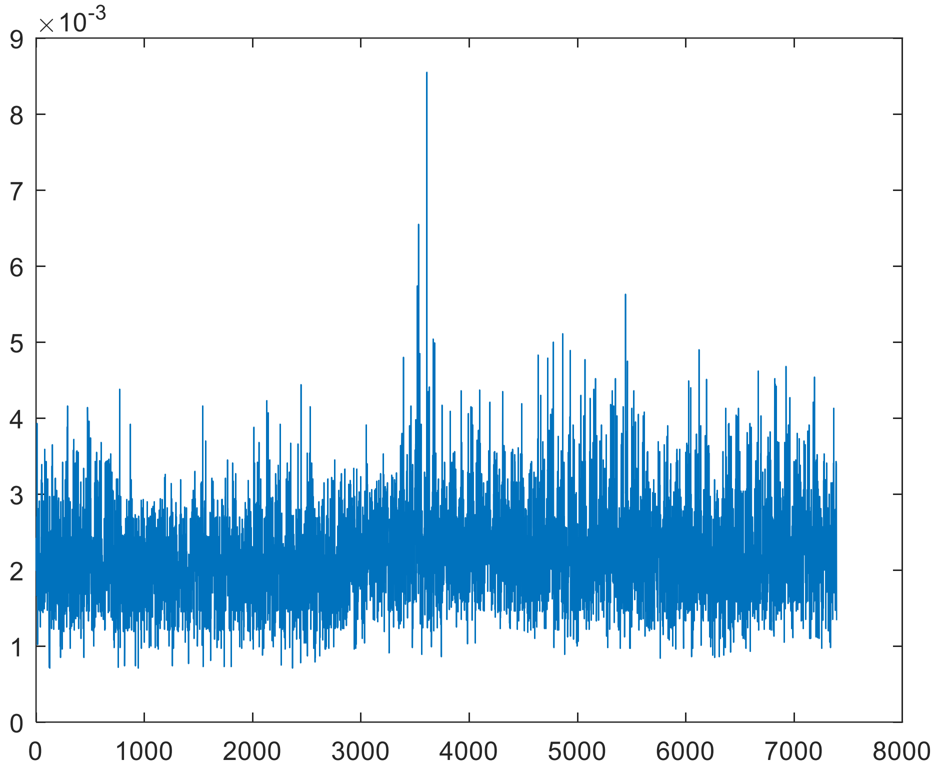

Fig. 4.1.2-3 Time of arrival for stop 4

– In order to convert the timeseries into trend stationary the first differences are calculated, which actually become the actual target variables (i.e., diff_stop2 is time of travel to stop 2.)

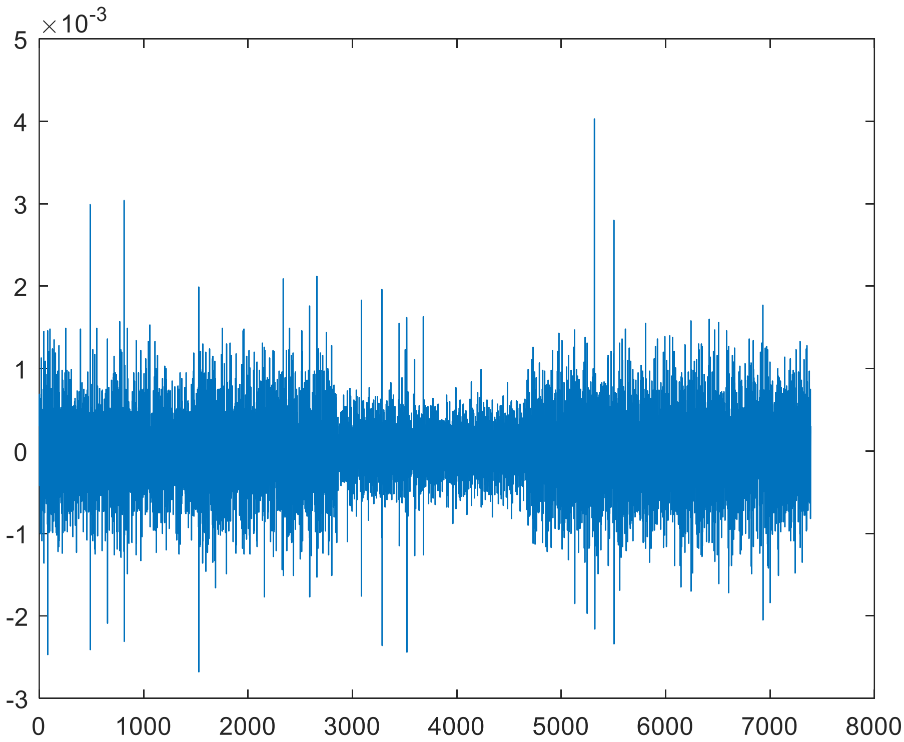

Fig. 4.1.2-4 Time of travel for stop 2

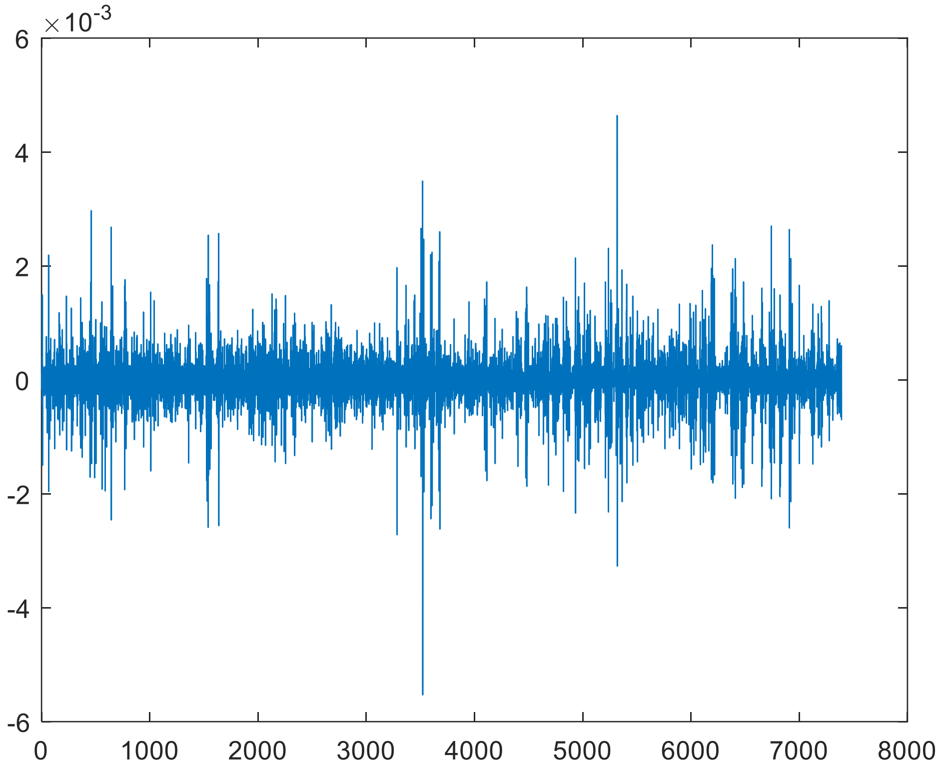

Fig. 4.1.2-5 Time of travel for stop 3

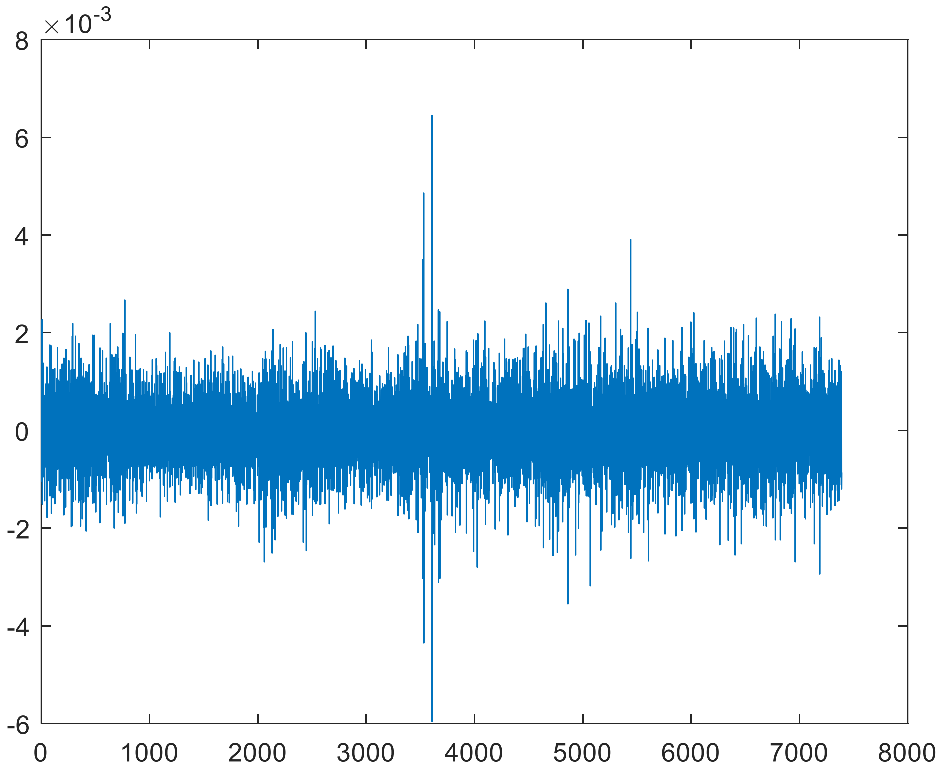

Fig. 4.1.2-6 Time of travel for stop 4

4.1.3. Data exploration.

At this point some exploratory data analysis is a must so as to identify specific characteristics and / or criteria to define outliers.

– One such characteristic is clearly visible out of the previous figures 4.1.2-1 through 4.1.2-6 – there is a special behavior for time of arrival to stop 2 (and correspondingly to time of travel to stop 2) for a large subperiod of time dating between 1 August 2020 and 31 December 2020. This will be addressed later with feature engineering and also would be needed for clearing the fitting data.

– Additional analysis is done on histograms of times of arrival and times of travel for each bus stop. The main take here is to define reasonable criteria for outliers, which would also be addressed with feature engineering.

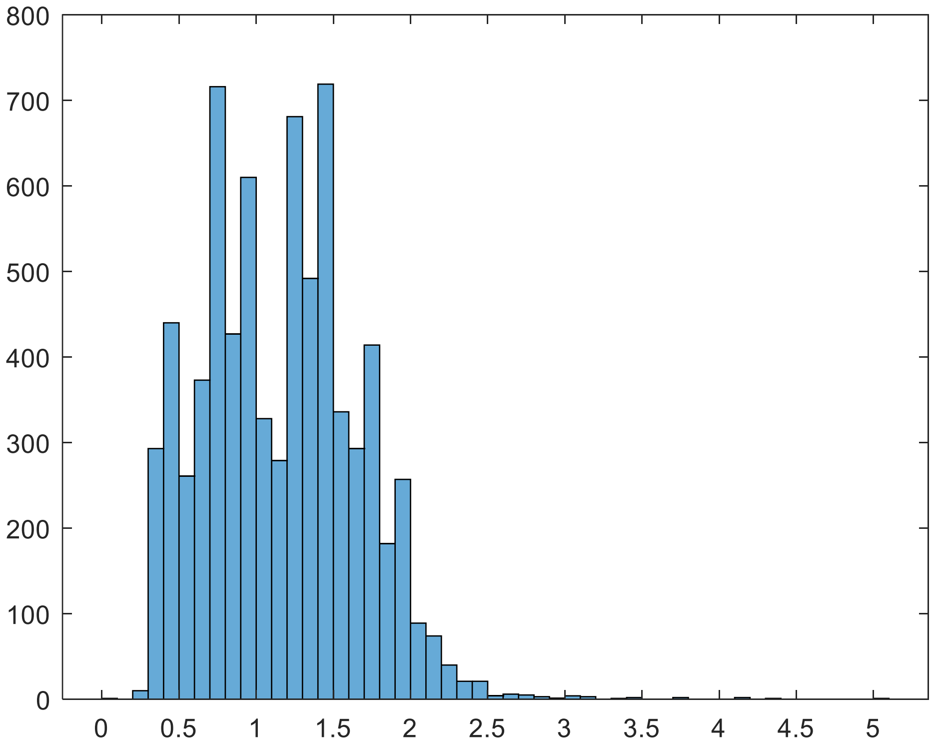

Fig. 4.1.3-1 Histogram of times of arrival for stop 2

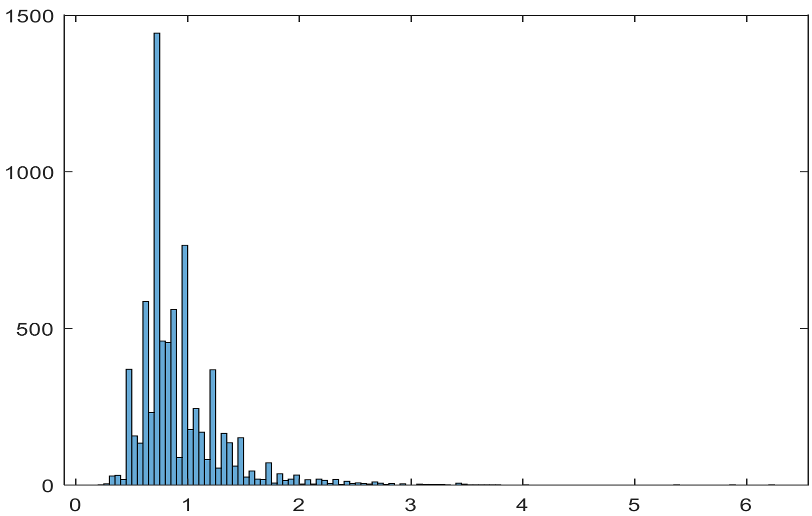

Fig. 4.1.3-2 Histogram of times of arrival for stop 3

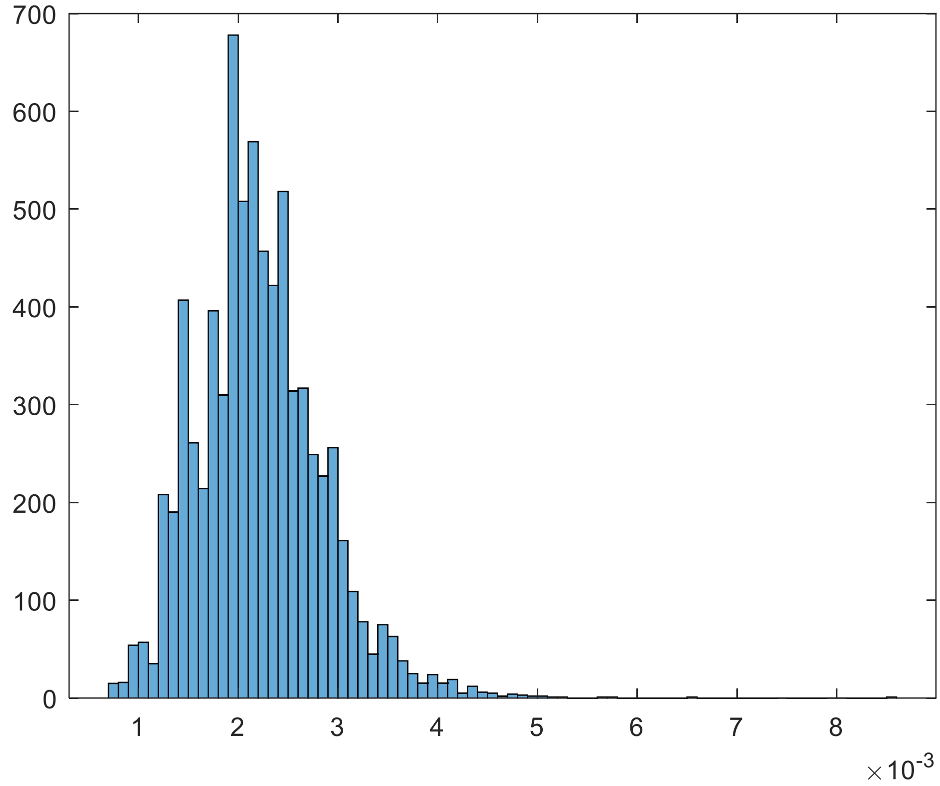

Fig. 4.1.3-3 Histogram of times of arrival for stop 4

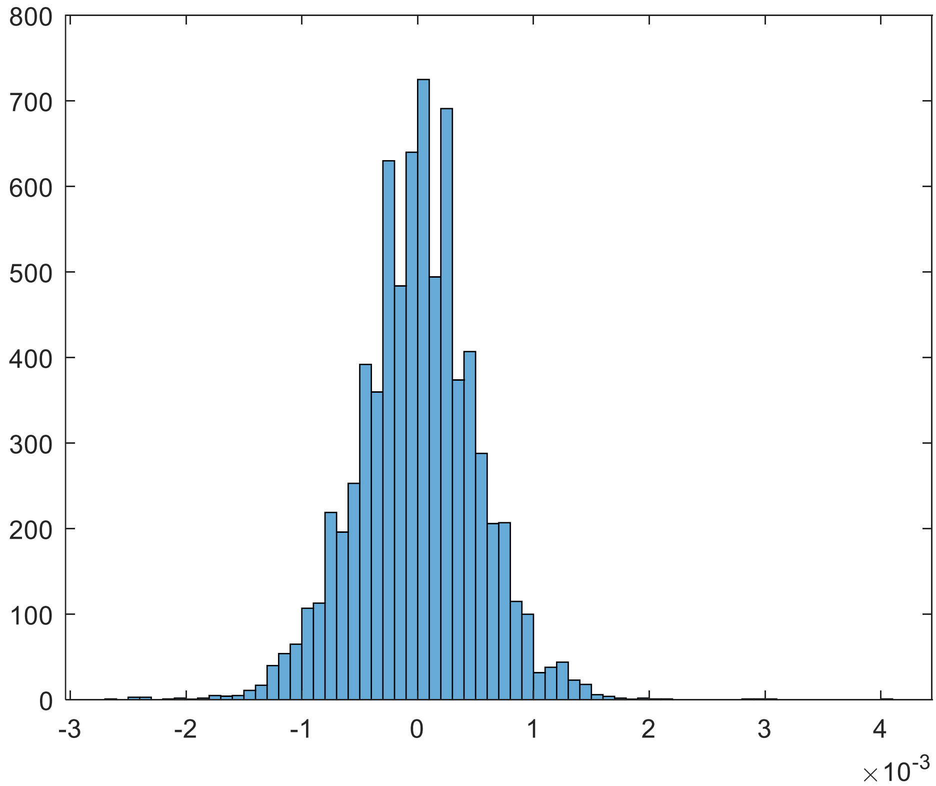

Fig. 4.1.3-4 Histogram of times of travel to stop 2

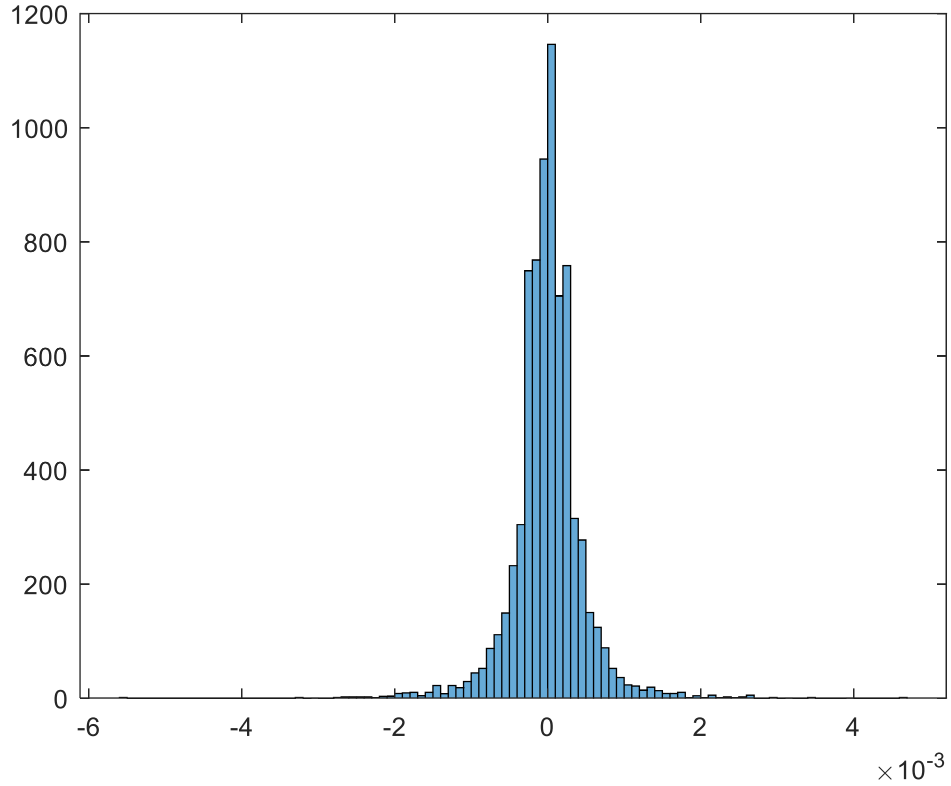

Fig. 4.1.3-5 Histogram of times of travel to stop 3

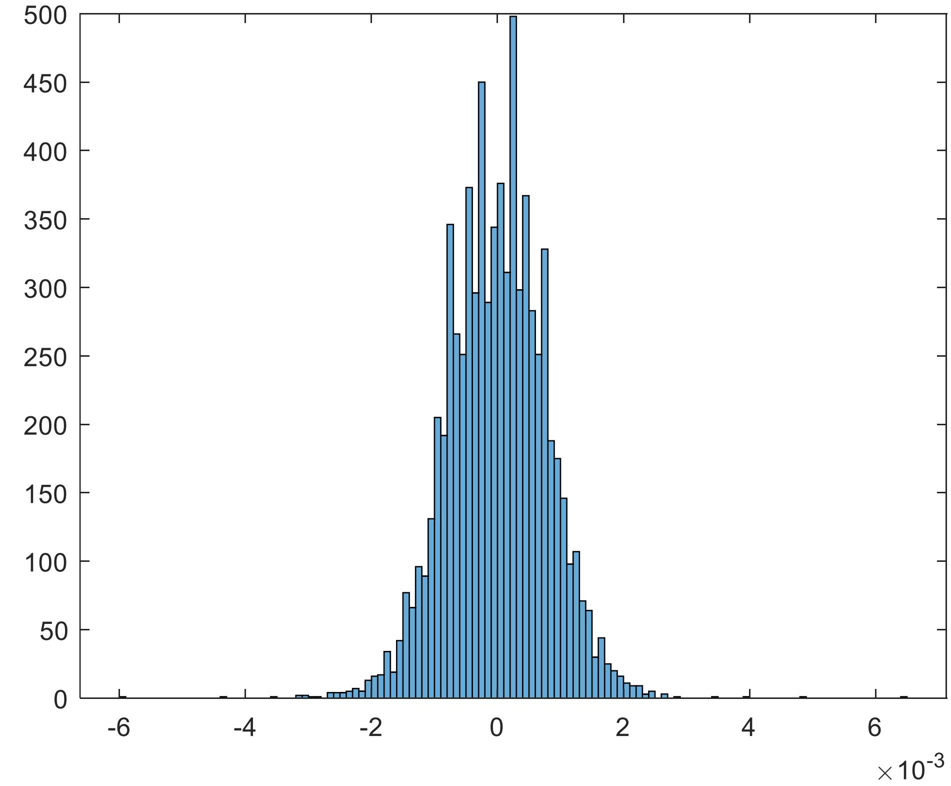

Fig. 4.1.3-6 Histogram of times of travel to stop 4

4.1.4. *Feature engineering.*

These operations for feature engineering are based on the traffic data transformations, which produce mainly dummy type or serial number type features.

– In order to check for weekly cyclicity a day of week feature is created by coding Monday to Sunday as 1 to 7.

– In order to check for day of year cyclicity a feature is created by coding 1 January to 31 December as 1 to 366 (or 365 depending on leap year calendar).

– In order to check for intraday hourly cyclicity a feature is created by coding starting hour for each arrival time by stop as 0 to 23 (note that actually the hours vary from 5 to 23).

– A comparison feature to calculate delay of time of arrival in corresponding to the official schedule as a simple difference is calculated. Note that on some occasions negative delay could occur.

– To reflect the above-mentioned special behavior for time of travel to stop2 a dummy feature is defined as 1 between 1-Aug-2020 and 31-Dec-2020. In case of other special characteristics similar features may also be calculated.

– Outliers may become important information since there might be a pattern in their distribution. Analyzing the histograms (fig. 4.1.3-1 through 4.1.3-6) a dummy feature is calculated to be equal to 1 if the corresponding times of arrival and times to travel for each stop comply with the criteria shown in Table 4.1.4-1.

Table 4.1.4-1 Criteria for outliers

Variable	Bottom criterion	Top criterion
Time of arrival at stop 2	time_stop2 < 0.0003	time_stop2 > 0.0025
Time of arrival at stop 3	time_stop3 < 0.0003	time_stop3 > 0.0020
Time of arrival at stop 4	time_stop4 < 0.0009	time_stop4 > 0.0037
Time of travel to stop 2	diff_stop2 < -0.0015	diff_stop2 > 0.0015
Time of travel to stop 3	diff_stop3 < -0.0016	diff_stop3 > 0.0017
Time of travel to stop 4	diff_stop4 < -0.0021	diff_stop4 > 0.0021

– Another possible feature to be included is the previous stops times to account for the interconnection among the various bus stop times of arrival within the same bus course: for stop 2 there is no data to add; for stop 3 times of arrival at stop 2 are added; for stop 4 times of arrival for stop 2 and stop 3 are added.

– Similarly previous stops stay times are added where possible: nothing to add for stop 2; stays at stop 2 added for stop 3; stays for stop 2, stop 3 added for stop 4.

4.1.5. *Data enrichment.*

The operations for data enrichment are based on exogenous data addition and transformation. The new features are mainly dummy type or measurement type.

– For official holidays and weekends a dummy feature is calculated as equal to 1 for every Saturday and Sunday but also for the list of holidays (03-Mar-2020, 19-Apr-2020, 01-May-2020, 06-May-2020, 24-May-2020, 06-Sep-2020, 22-Sep-2020, 24-Dec-

2020, 25-Dec-2020, 26-Dec-2020, 01-Jan-2021, 03-Mar-2021, 19-Apr-2021, 01-May-2021, 06-May-2021, 24-May-2021).

– In order to check for special behavior during the official lockdown period in Bulgaria a dummy feature is defined as 1 between 13-Mar-2020 and 13-May-2020.

– Wind-chill factor is imported from the available weather data, measured in degrees centigrade and is a combined factor to account for the windy conditions as well.

– Humidity feature is imported from the available weather data, measured in percentage as the concentration of water vapor present in the air. The humidity indicates the likelihood for precipitation, dew, or fog to be present.

– Clouds feature is imported from the available weather data, measured in percentage as relative share of sky above city of Sofia covered by clouds.

– A new dummy feature is calculated aggregation of presence of rainfall and snowfall in the previous hour, mainly aiming at accounting for road conditions.

4.1.6. Fourier terms analysis.

The core of the SARIMAX w/ FT methods are its Fourier terms. To deal with multiple seasonality factors, they are modeled by adding Fourier terms that are used as external regressors. This approach is flexible and allows inclusion of multiple periods. There would be different Fourier series corresponding to each of the seasonal periods.

For each time-of-travel data set, there are seven seasonal periods, each of them modelled by up to 8 Fourier terms. For each of the periods, the number of Fourier terms is chosen to find the best statistical model. Given a set of models, the quality of the model is compared to other models and the best of them is selected for use.

The seasonal periods used are:

- Course within day – on average about 14 courses per day
- Hour of day – about 18 hours a day
- Day of week – seven days
- Course within week – about 91 courses per week
- Month within year – 12 months per year
- Day of year –366 days per year (note: 2020 is a leap year)
- Day of dataset – 560 days in the dataset

– Modeling Fourier terms for stop 2.

The best results from the modeling Fourier terms for stop 2 are shown in Table 4.1.6-1. Visible from the table is that most of the periods do not result in good results (R-square ~ 0 and Adjuster R-square <0) even though the table only presents their best reasonable modification. The Periods which have some (but not at all strong) impact are Daily courses (fig. 4.1.6-1), Hours within the day (fig. 4.1.6-2), and to some extend – weekly courses (fig. 4.1.6-3). The models are fitted after filtering out the outliers as defined above.

Table 4.1.6-1 Best fitted Fourier terms by seasonal period for stop 2

Fit name	Fit type	SSE	R-square	Adj R-sq	RMSE	# Coeff
Daily course	fourier7	0.0016	0.0817	0.0798	4.7156e-04	16
Day hour	fourier7	0.0016	0.0834	0.0815	4.7112e-04	16
Round day	fourier8	0.0018	1.1472e-04	-0.0022	4.9212e-04	18
Week day	fourier8	0.0018	6.9249e-05	-0.0023	4.9213e-04	18
Weekly course	fourier4	0.0017	0.0025	0.0013	4.9126e-04	10
Year day	fourier8	0.0018	1.9968e-04	-0.0021	4.9210e-04	18
Year month	fourier8	0.0018	7.0415e-05	-0.0023	4.9213e-04	18

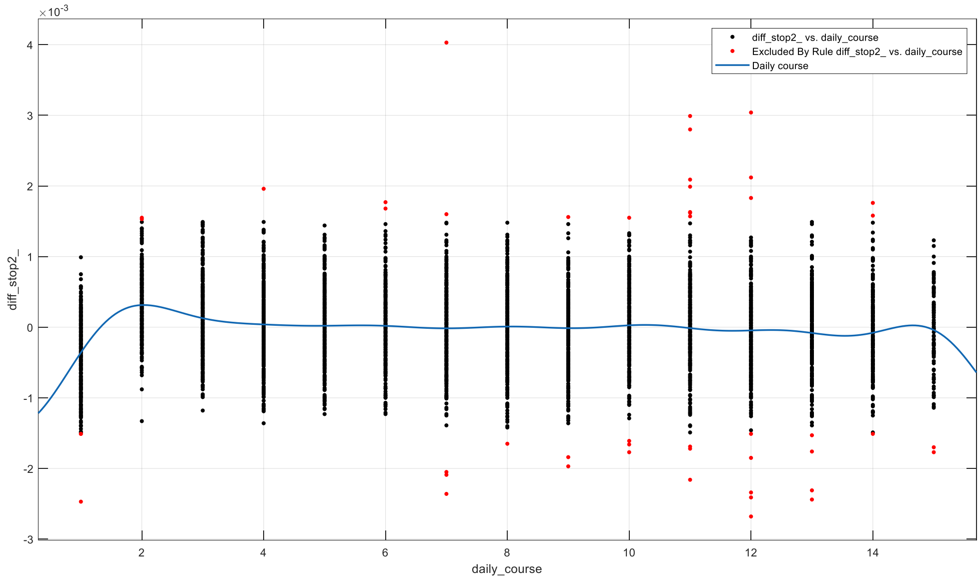

Figure 4.1.6-1 Best fitted Fourier term with period daily course for stop 2

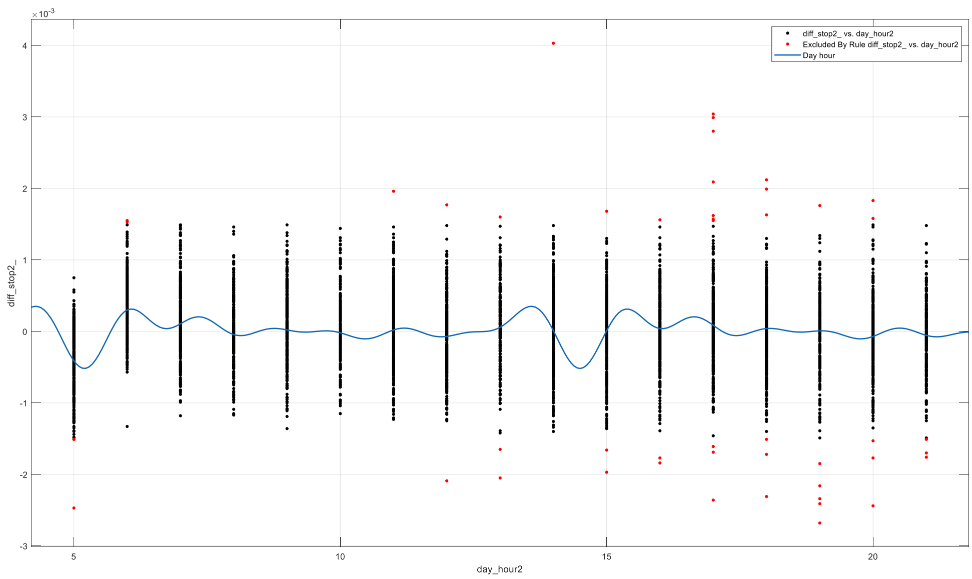

Figure 4.1.6-2 Best fitted Fourier term with period hour within the day for stop 2

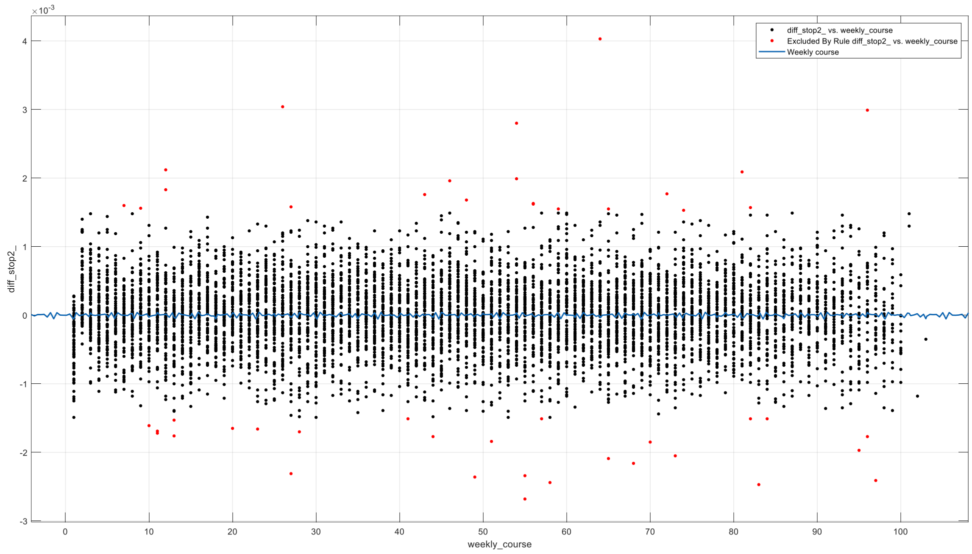

Figure 4.1.6-3 Best fitted Fourier term with period weekly course for stop 2

– Modeling Fourier terms for stop 3.

The best results from the modeling Fourier terms for stop 3 are shown in Table 4.1.6-2. Visible from the table is that most of the periods do not result in good results (R-square ~ 0 and Adjuster R-square <0) even though the table only presents their best reasonable modification. The Periods which have some (but not at all strong) impact are Daily courses (fig. 4.1.6-4), Hours within the day (fig. 4.1.6-5), and to some extend – weekly courses (fig. 4.1.6-6). The models are fitted after filtering out the outliers as defined above.

Table 4.1.6-2 Best fitted Fourier terms by seasonal period for stop 3

Fit name	Fit type	SSE	R-square	Adj R-sq	RMSE	# Coeff
Daily course	fourier7	9.7371e-04	0.0466	0.0446	3.6834e-04	16
Day hour	fourier6	9.4703e-04	0.0727	0.0710	3.6320e-04	14
Round day	fourier8	0.0011	4.0974e-04	-0.0020	3.8839e-04	18
Week day	fourier8	0.0010	4.8271e-04	-0.0019	3.7719e-04	18
Weekly course	fourier8	0.0010	0.0106	0.0082	3.7528e-04	18
Year day	fourier8	0.0010	3.8816e-04	-0.0020	3.7721e-04	18
Year month	fourier8	0.0010	9.1551e-05	-0.0023	3.7726e-04	18

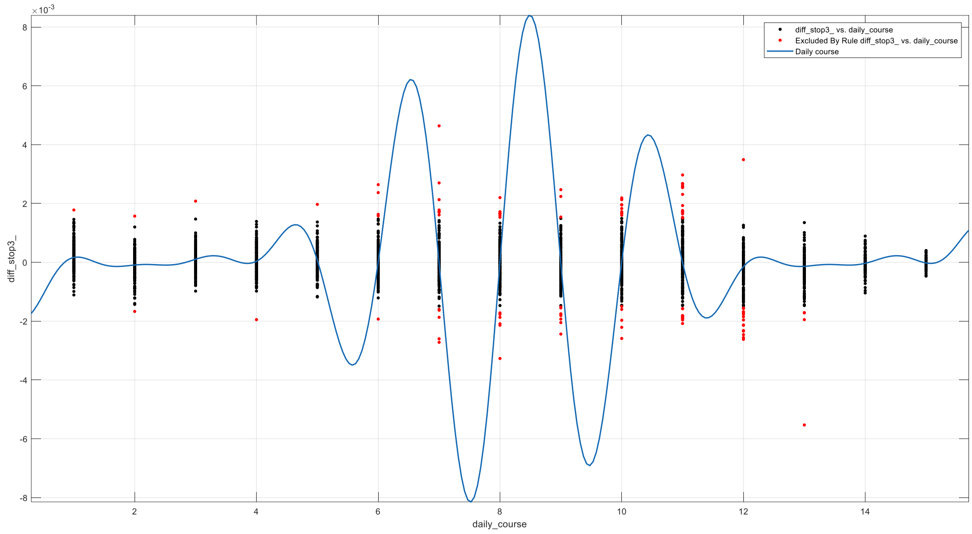

Figure 4.1.6-4 Best fitted Fourier term with period daily course for stop 3

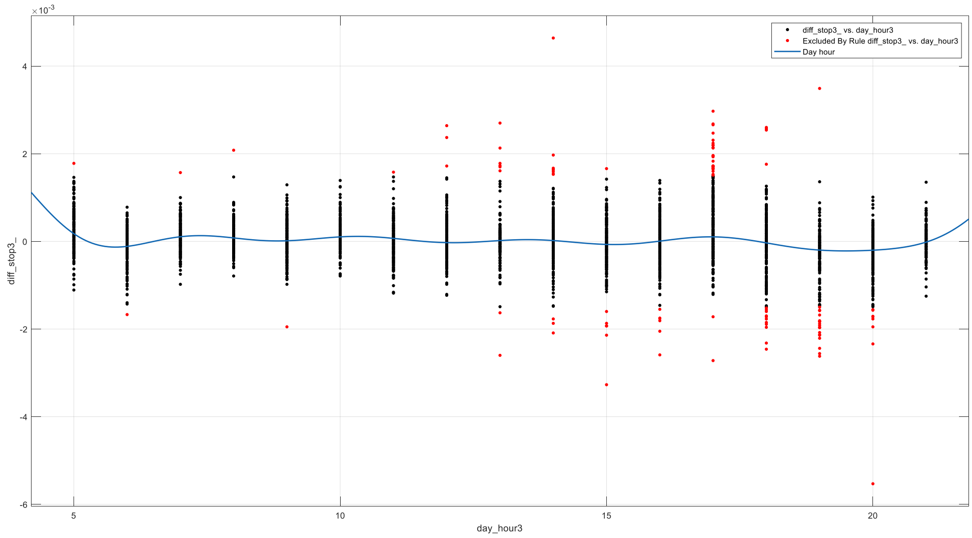

Figure 4.1.6-5 Best fitted Fourier term with period hour within the day for stop 3

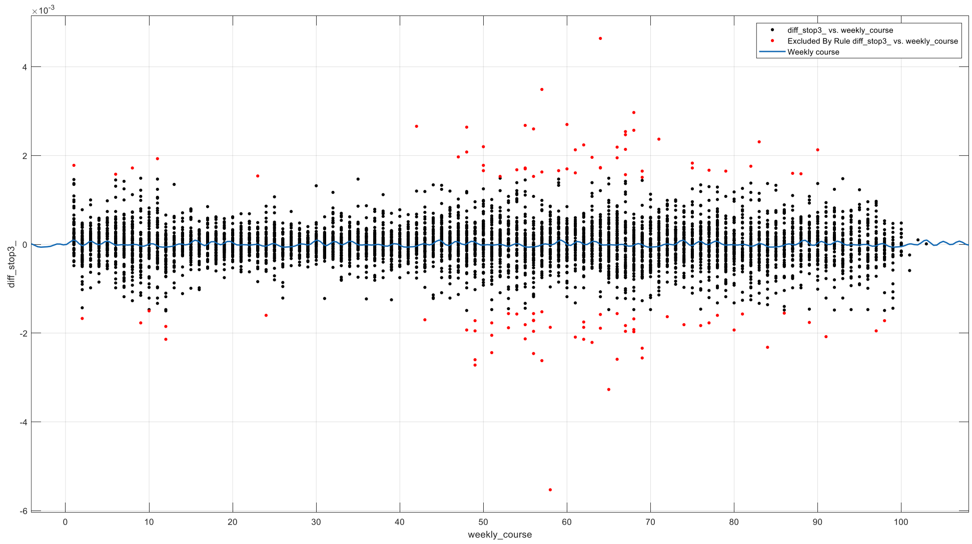

Figure 4.1.6-6 Best fitted Fourier term with period weekly course for stop 3

– Modeling Fourier terms for stop 4.

The best results from the modeling Fourier terms for stop 4 are shown in Table 4.1.6-3. Visible from the table is that most of the periods do not result in good results (R-square ~ 0 and Adjuster R-square <0) even though the table only presents their best reasonable modification. The Periods which have some (but not at all strong) impact are Daily courses (fig. 4.1.6-7) and Hours within the day (fig. 4.1.6-8) The models are fitted after filtering out the outliers as defined above.

Table 4.1.6-3 Best fitted Fourier terms by seasonal period for stop 4

Fit name	Fit type	SSE	R-square	Adj R-sq	RMSE	# Coeff
Daily course	fourier7	0.0030	0.0155	0.0134	6.5846e-04	16
Day hour	fourier8	0.0030	0.0249	0.0225	6.5539e-04	18
Round day	fourier8	0.0039	1.4147e-04	-0.0022	7.3723e-04	18
Week day	fourier8	0.0031	4.3880e-04	-0.0020	6.6357e-04	18
Weekly course	fourier8	0.0030	0.0030	5.6296e-04	6.6271e-04	18
Year day	fourier8	0.0030	0.0010	-0.0014	6.6338e-04	18
Year month	fourier8	0.0031	3.2334e-04	-0.0021	6.6361e-04	18

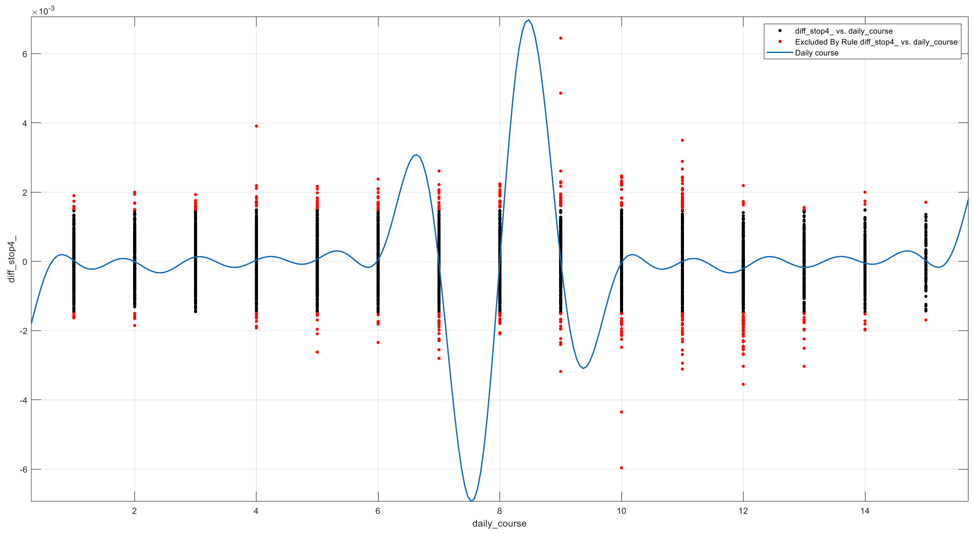

Figure 4.1.6-7 Best fitted Fourier term with period daily course for stop 4

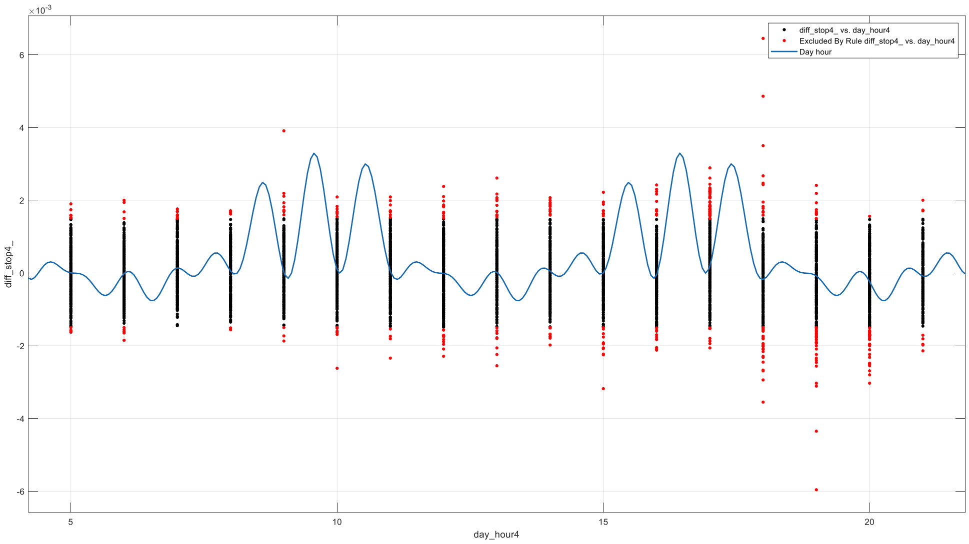

Figure 4.1.6-8 Best fitted Fourier term with period hour within the day for stop 4

4.1.7. Forming data sets.

As a terminal phase of data preparation, finalizing operations are conducted on the data so as to prepare useful time series data frame:

– Due to the previous operation of first difference the target feature is now one observation short, this is why a 0 is added at the beginning of the series.
– Collect all data for modeling for each stop in one timetable for multivariate time series data, stored in a MATLAB table
– As a regularization measure each dataset is split in train and validation samples, where the test consists of all the last week of data (starting from 23-Jul-2021) and the train consists of all the rest.

Table 4.1.6-1 Sample from data for stop 2

Time	clouds_all	daily_course	daily_course_ft2	day_hour_ft2	delay_stop2	diff_stop2_	day_hour2	feels_like	holiday_weekend	humidity	lockdowns	m_data_stop2	outliers2	pour	round_day_data	time_mode2	week_day	weekly_course	weekly_course_ft2	year_day	year_month
2020-01-10 05:30	1	1	-0.00037	-0.00040	-0.0007	0.0000	5	-4.3	0	86	0	737800.23	0	0	737800	0	5	1	-0.00006	10	1
2020-01-10 06:43	1	2	0.00031	0.00029	-0.0002	0.0003	6	-2.5	0	79	0	737800.28	0	0	737800	0	5	2	0.00001	10	1
2020-01-10 07:47	1	3	0.00013	-0.00004	0.0007	0.0007	7	0.2	0	69	0	737800.32	0	0	737800	0	5	3	0.00003	10	1
2020-01-10 09:00	0	4	0.00004	0.00002	-0.0001	-0.0004	9	3.8	0	59	0	737800.38	0	0	737800	0	5	4	0.00001	10	1
2020-01-10 10:30	1	5	0.00002	0.00002	-0.0003	0.0004	10	4.8	0	55	0	737800.44	0	0	737800	0	5	5	0.00000	10	1

Table 4.1.6-2 Sample from data for stop 3

Time	clouds_all	daily_course	daily_course_ft3	day_hour_ft3	delay_stop3	diff_stop3_	day_hour3	feels_like	holiday_weekend	humidity	lockdowns	m_data_stop3	outliers3	pour	stay_stop2	round_day_data	time_stop2	week_day	weekly_course	weekly_course_ft3	year_day	year_month
2020-01-10 05:31	1	1	0.00014	0.00023	-0.0012	0.0000	5	-4.3	0	86	0	737800.23	0	0	0	737800	0.0007	5	1	0.00008	10	1
2020-01-10 06:44	1	2	-0.00010	-0.00012	-0.0007	0.0001	6	-2.5	0	79	0	737800.28	0	0	1	737800	0.0010	5	2	0.00000	10	1
2020-01-10 07:49	1	3	0.00007	0.00009	0.0003	0.0000	7	0.2	0	69	0	737800.33	0	0	1	737800	0.0017	5	3	0.00004	10	1
2020-01-10 09:01	0	4	0.00000	0.00001	-0.0007	-0.0002	9	3.8	0	59	0	737800.38	0	0	0	737800	0.0013	5	4	0.00000	10	1
2020-01-10 10:31	1	5	0.00031	0.00007	-0.0006	0.0003	10	4.8	0	55	0	737800.44	0	0	1	737800	0.0017	5	5	0.00006	10	1
2020-01-10 12:54	0	6	-0.00050	-0.00003	-0.0002	0.0004	12	9.3	0	36	0	737800.54	0	0	1	737800	0.0013	5	6	0.00000	10	1

Table 4.1.6-3 Sample from data for stop 4

Time	clouds_all	daily_course	daily_course_ft4	day_hour_ft4	delay_stop3	diff_stop3_	day_hour3	feels_like	holiday_weekend	humidity	lockdowns	m_data_stop3	outliers3	pour	stay_stop2	round_day_data	time_stop2	time_stop3	week_day	weekly_course	year_day	year_month
2020-01-10 05:34	1	1	0.00015	0.00008	-0.0006	0.0000	5	-4.3	0	86	0	737800.23	0	0	0	1	737800	0.0007	0.0009	5	1	10
2020-01-10 06:48	1	2	0.00006	-0.00027	0.0004	0.0004	6	-2.5	0	79	0	737800.28	0	0	1	0	737800	0.0010	0.0010	5	2	10
2020-01-10 07:52	1	3	0.00008	0.00034	0.0004	-0.0002	7	0.2	0	69	0	737800.33	0	0	1	2	737800	0.0017	0.0010	5	3	10
2020-01-10 09:04	0	4	-0.00003	0.00084	-0.0006	0.0000	9	3.8	0	59	0	737800.38	0	0	0	0.5	737800	0.0013	0.0008	5	4	10
2020-01-10 10:34	1	5	-0.00001	0.00071	-0.0006	-0.0001	10	4.8	0	55	0	737800.44	0	0	1	0	737800	0.0017	0.0011	5	5	10
2020-01-10 12:56	0	6	-0.00011	0.00000	-0.0006	-0.0003	12	9.3	0	36	0	737800.54	0	0	1	0	737800	0.0013	0.0015	5	6	10
2020-01-10 15:19	20	7	0.00104	0.00001	0.0001	0.0010	15	8.5	0	47	0	737800.64	0	0	0	1	737800	0.0014	0.0011	5	7	10

4.2. Data modeling.

4.2.1. Endogenous feature engineering.

Next using the autocorrelation function (ACF) and partial autocorrelation function (PACF) the moving average (MA) and the auto regression (AR) lags significance is determined.

– ACF is analysis for the number of significant MA lags. It turns out for all stops we have the simplest possible pattern of MA=1 (e.g., 1 significant lag back).

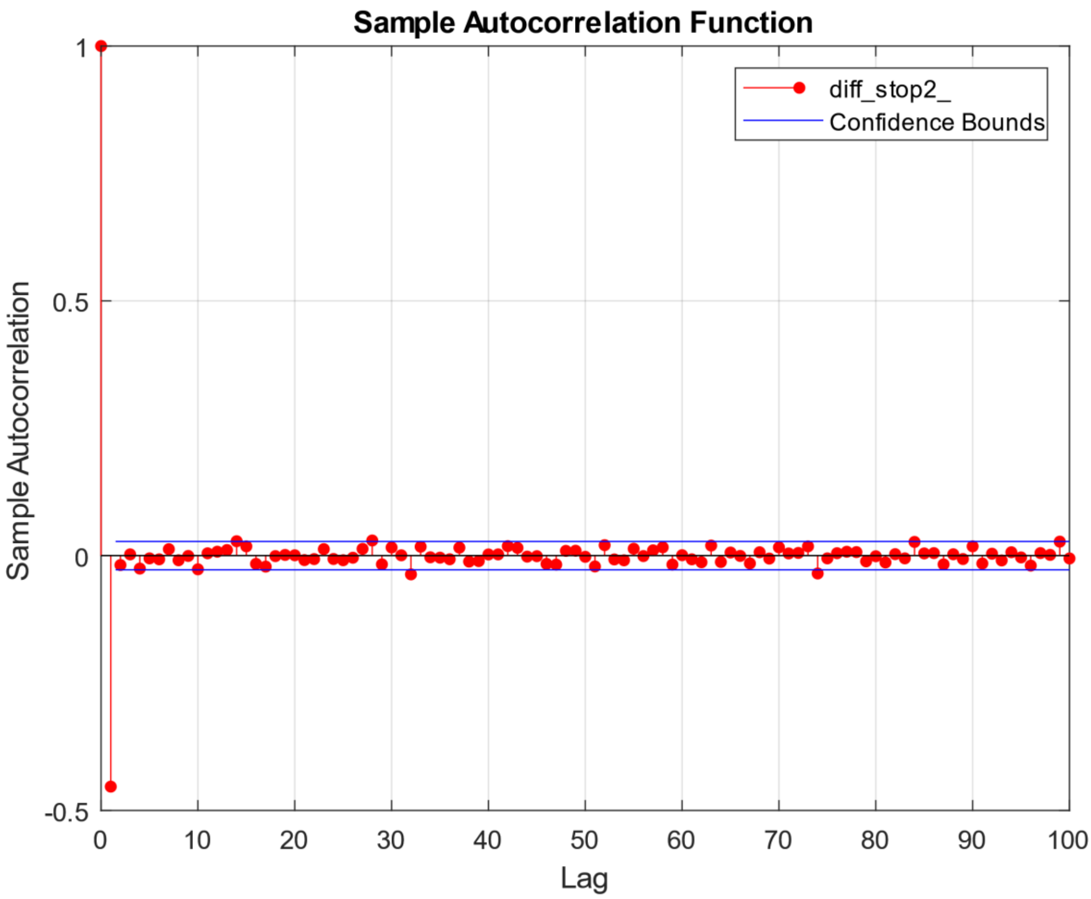

Figure 4.2.1-1 Sample autocorrelation function of stop 2

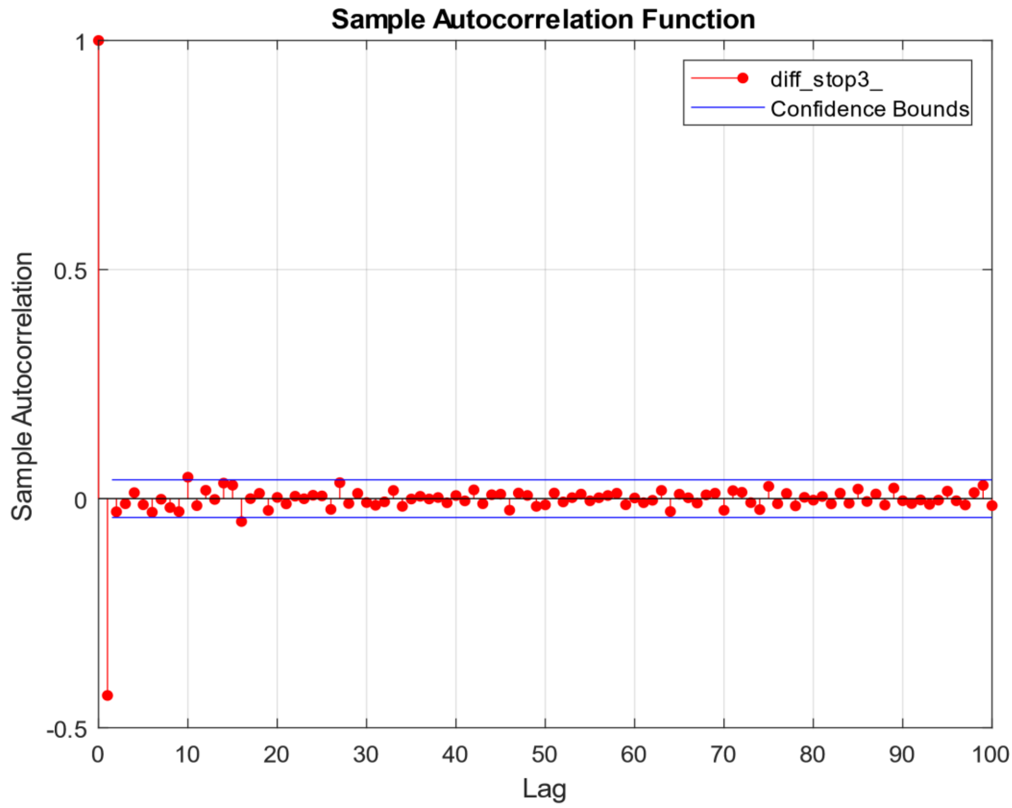

Figure 4.2.1-2 Sample autocorrelation function of stop 3

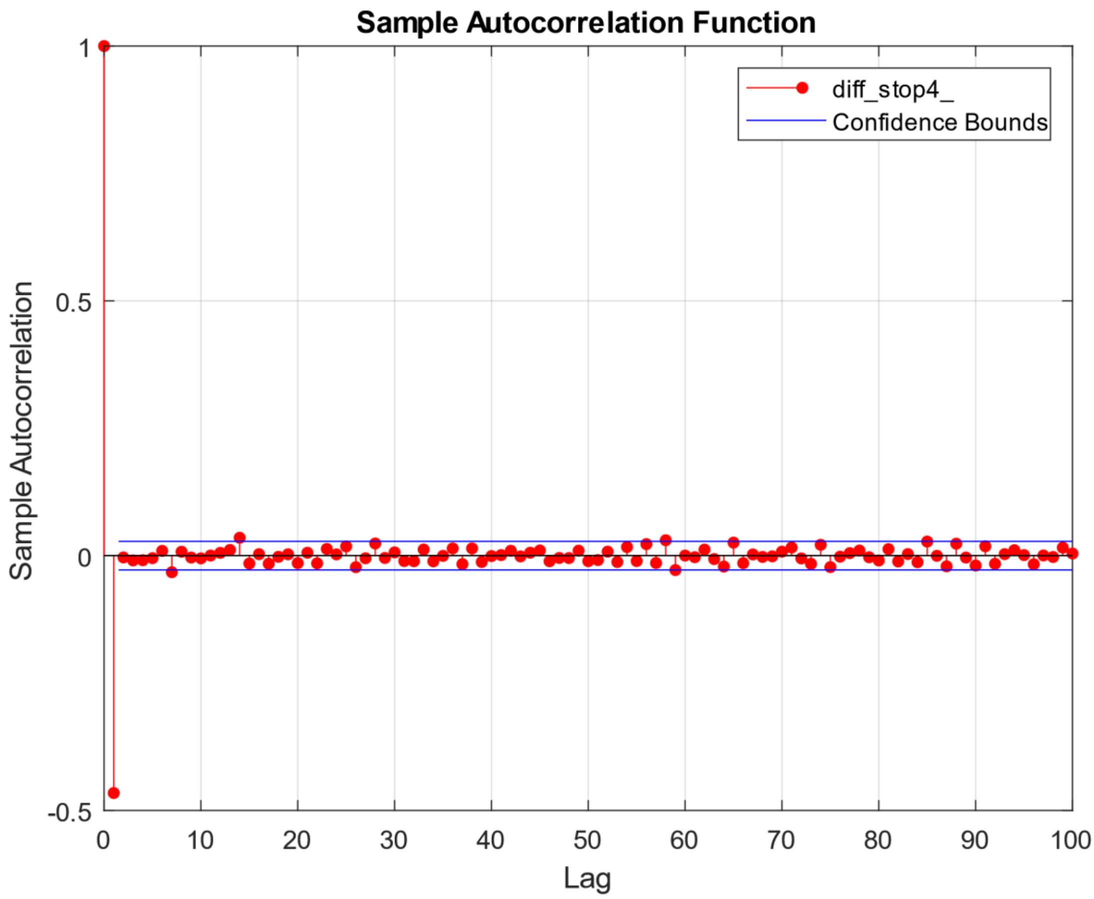

Figure 4.2.1-3 Sample autocorrelation function of stop 4

– PACF is analysis for the number of significant AR lags. It turns out that the results are close – 14 significant lags for stop 2 and 13 significant lags for both stop 3 and stop 4.

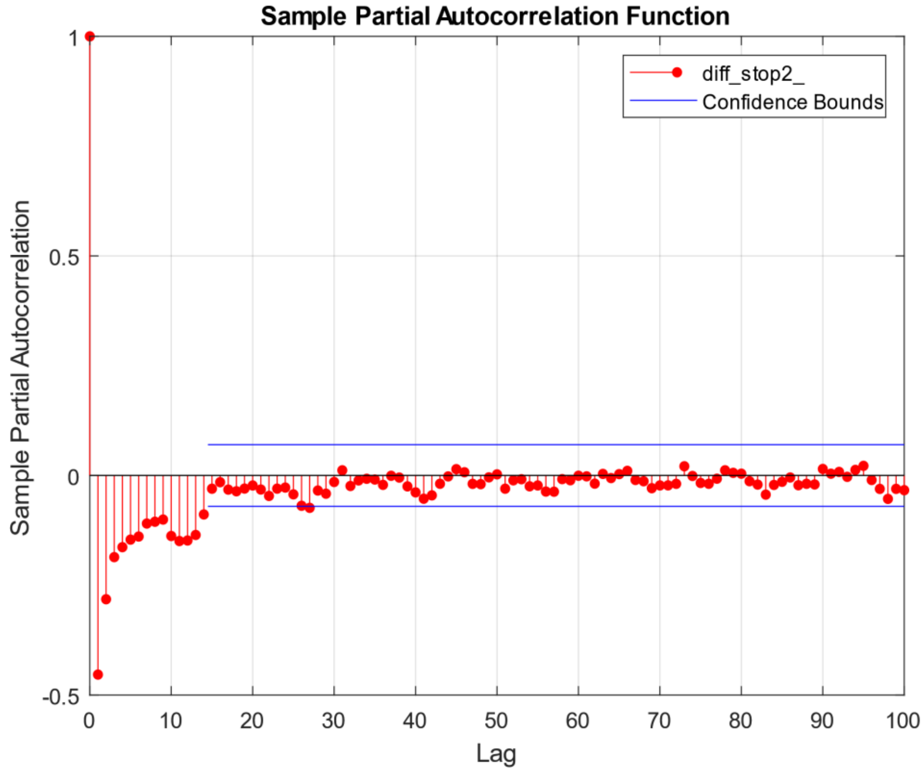

Figure 4.2.1-4 Sample partial autocorrelation function of stop 2

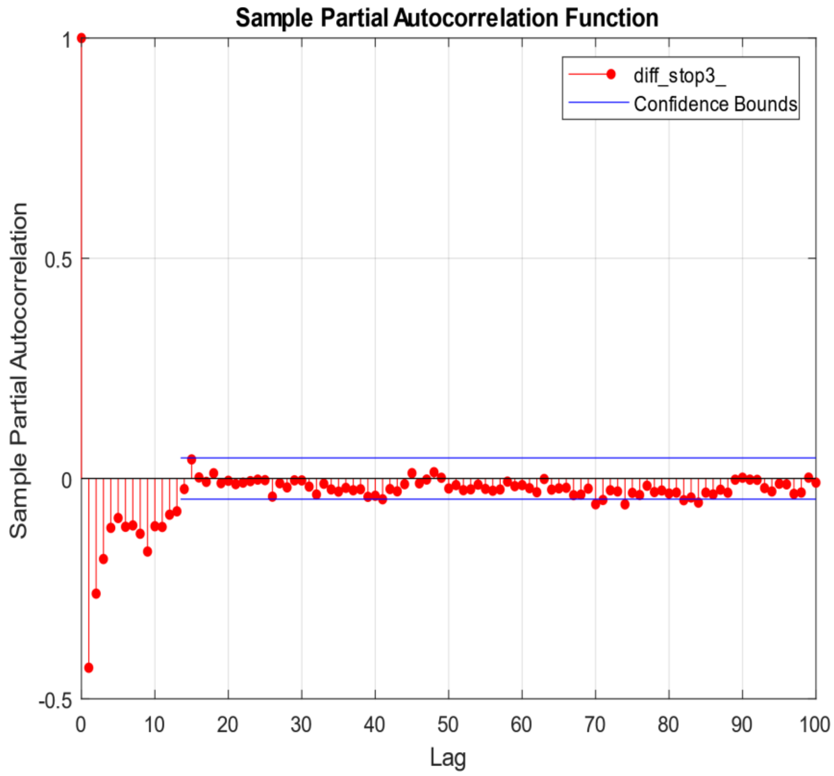

Figure 4.2.1-5 Sample partial autocorrelation function of stop 3

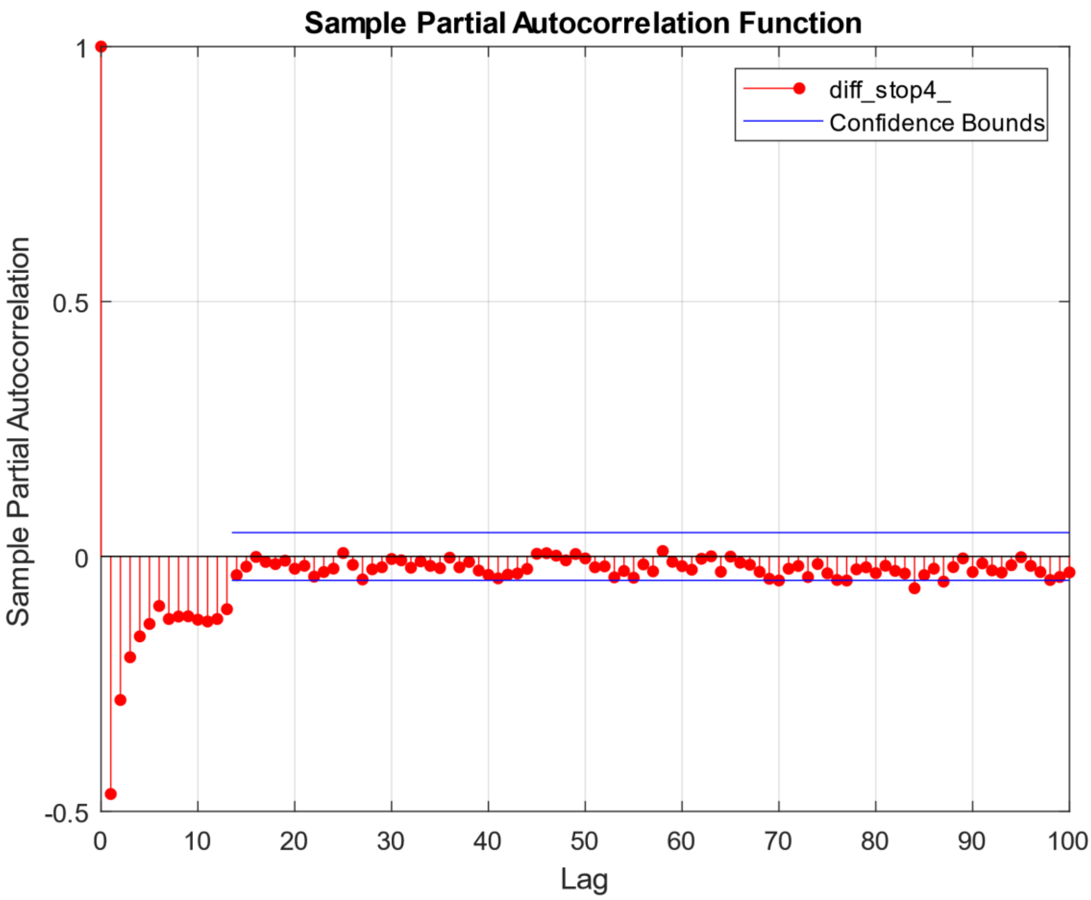

Figure 4.2.1-6 Sample partial autocorrelation function of stop4

4.2.2. Statistical tests

- Augmented Dickey-Fuller Test

Null Hypothesis: predicted feature contains a unit root

$$y_t = c + \delta t + \phi y_{t-1} + \beta_1 \Delta y_{t-1} + \ldots + \beta_p \Delta y_{t-p} + \varepsilon_t$$

$$H_0 : \phi = 1$$

$$H_a : \phi < 1$$

Test Parameters: Lags: 0; Model: AR; Test Statistic: t1; Significance Level: 0.05

Table 4.2.2-1 ADF Test Results

Feature	Null Rejected	P-Value	Test Statistic	Critical Value
Stop 2	true	0.001	-139.2992	-1.9416
Stop 3	true	0.001	-135.2616	-1.9416
Stop 4	true	0.001	-141.5354	-1.9416

- KPSS Test

Null Hypothesis: the predicted feature is trend stationary

$$y_t = c_t + \delta t + u_{1t}$$

$$c_t = c_{t-1} + u_{2t}$$

$$u_{2t} \sim i.i.d(0, \sigma^2)$$

$$H_0 : \sigma^2 = 0$$

$$H_a : \sigma^2 > 0$$

Test Parameters: Lags: 0; Include Trend: true; Significance Level: 0.05

Table 4.2.2-2 KPSS Test Results

Feature	Null Rejected	P-Value	Test Statistic	Critical Value
Stop 2	false	0.1	0.0001678	0.146
Stop 3	false	0.1	0.00010895	0.146
Stop 4	false	0.1	9.0571e-05	0.146

4.2.3. Model estimation.

- SARIMAX(14,2,1,1,0,1) for stop 2

Seasonal ARIMA model of time series for stop 2 using exogenous predictors. The model uses the following equation:

$$(1 - \phi_1 L - ... - \phi_{14} L^{14})(1 - L) y_t = c + X_1\beta_1 + ... + X_{18}\beta_{18} + (1 + \theta_1 L)(1 + \Theta_1 L)\varepsilon_t$$

Table 4.2.3-1 Estimation Results

Parameter	Value	Standard Error	t Statistic	P-Value
Constant	0.00016817	3.8781e-05	4.3365	1.4474e-05
AR{1}	-1.6257	0.029694	-54.7483	0
AR{2}	-2	0.049076	-40.7541	0
AR{3}	-2.1924	0.06193	-35.4019	1.598e-274
AR{4}	-2.288	0.069019	-33.1505	5.579e-241
AR{5}	-2.2936	0.072021	-31.846	1.4938e-222
AR{6}	-2.2215	0.072602	-30.5988	1.2709e-205
AR{7}	-2.0609	0.070974	-29.0368	2.2616e-185
AR{8}	-1.8489	0.06649	-27.8066	3.6078e-170
AR{9}	-1.5969	0.059805	-26.7027	4.3848e-157
AR{10}	-1.3269	0.052137	-25.4498	7.103e-143
AR{11}	-1.0345	0.043688	-23.6793	5.8876e-124
AR{12}	-0.73653	0.033797	-21.7927	2.7225e-105
AR{13}	-0.43347	0.022708	-19.0889	3.1238e-81
AR{14}	-0.1703	0.011617	-14.66	1.1632e-48
MA{1}	-0.035087	0.015753	-2.2273	0.02593
SMA{1}	-0.035087	0.015768	-2.2251	0.026073
Beta(clouds_all)	9.087e-07	1.6514e-07	5.5026	3.7416e-08
Beta(daily_course)	1.5837e-05	6.4743e-06	2.4462	0.014437

Parameter	Value	Standard Error	t Statistic	P-Value
Beta(daily_course_ft2)	0.35412	0.075609	4.6836	2.8192e-06
Beta(day_hour2)	-7.2087e-06	5.2729e-06	-1.3671	0.17158
Beta(day_hour_ft2)	0.44677	0.075596	5.9099	3.4223e-09
Beta(delay_stop2)	0.01994	0.0013653	14.6048	2.6172e-48
Beta(feels_like)	-8.6367e-07	7.8123e-07	-1.1055	0.26893
Beta(holiday_weekend)	1.3174e-06	1.6567e-05	0.079521	0.93662
Beta(humidity)	-4.2995e-06	3.9227e-07	-10.9606	5.9118e-28
Beta(lockdowns)	-6.579e-06	1.4484e-05	-0.45423	0.64966
Beta(outliers2)	0.00037892	1.8757e-05	20.2016	9.4791e-91
Beta(pour)	9.8153e-05	1.9257e-05	5.0971	3.4491e-07
Beta(time_mode2)	3.8645e-05	2.4848e-05	1.5552	0.11989
Beta(week_day)	3.9372e-06	3.6761e-06	1.071	0.28416
Beta(weekly_course)	7.0687e-08	2.0339e-07	0.34755	0.72818
Beta(weekly_course_ft2)	0.064911	0.20632	0.31461	0.75306
Beta(year_day)	1.5374e-07	5.2893e-07	0.29067	0.77131
Beta(year_month)	-4.6157e-06	1.5872e-05	-0.29082	0.77119
Variance	1.7058e-07	3.2089e-08	5.3158	1.0621e-07

Best goodness of Fit is at AIC: -92819.6321; BIC: -92571.4486

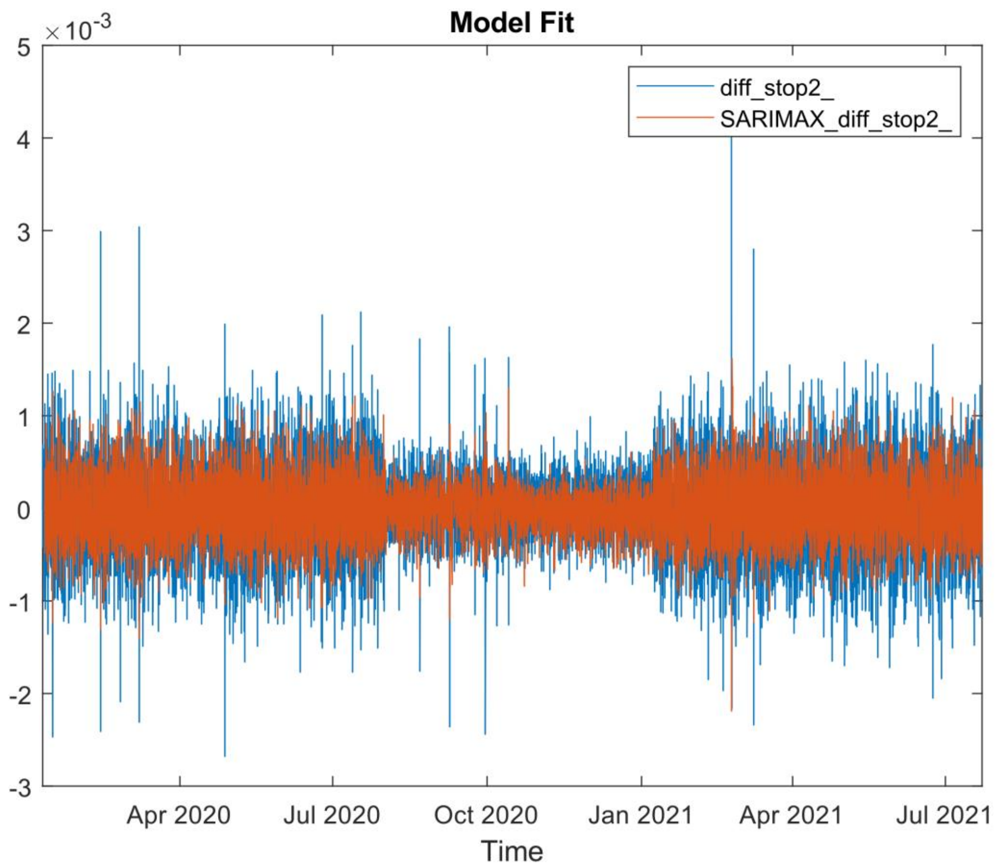

Figure 4.2.3-1 Plot the fit of model SARIMAX and time series for stop 2

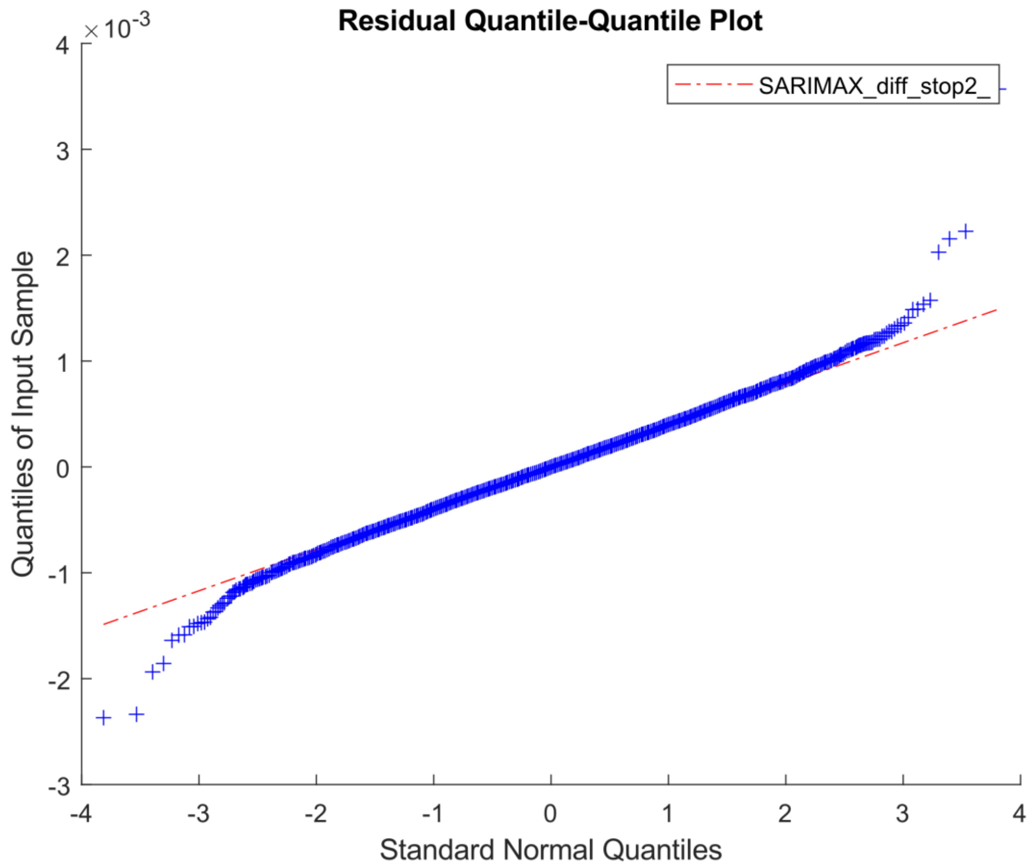

Figure 4.2.3-2 Quantile-quantile plot of the residuals of model SARIMAX for stop 2

- SARIMAX(13,2,1,1,0,1) for stop 3

Seasonal ARIMA model of time series stop 3 using exogenous predictors. The model uses the following equation:

$$(1-\phi_1 L - \dots - \phi_{13} L^{13})(1-L)y_t = c + X_1\beta_1 + \dots + X_{19}\beta_{19} + (1+\theta_1 L)(1+\Theta_1 L)\varepsilon_t$$

Table 4.2.3-2 Estimation Results

Parameter	Value	Standard Error	t Statistic	P-Value
Constant	-3.8919e-05	4.2363e-05	-0.9187	0.35825
AR{1}	-1.5974	0.013704	-116.5636	0
AR{2}	-1.9179	0.02083	-92.0761	0
AR{3}	-2.0595	0.025769	-79.9229	0
AR{4}	-2.0545	0.029229	-70.2904	0
AR{5}	-1.9672	0.031695	-62.0666	0
AR{6}	-1.8525	0.033371	-55.5143	0
AR{7}	-1.6846	0.033245	-50.6719	0
AR{8}	-1.49	0.03211	-46.4043	0
AR{9}	-1.2765	0.028746	-44.4053	0
AR{10}	-0.99312	0.024657	-40.2773	0
AR{11}	-0.69827	0.020005	-34.9053	6.1884e-267
AR{12}	-0.40643	0.01392	-29.1984	2.0308e-187
AR{13}	-0.16085	0.007032	-22.8733	8.5635e-116

Parameter	Value	Standard Error	t Statistic	P-Value
MA{1}	0.097659	0.0082432	11.8472	2.2238e-32
SMA{1}	0.097659	0.0082519	11.8347	2.5832e-32
Beta(clouds_all)	3.7062e-07	1.5554e-07	2.3828	0.01718
Beta(daily_course)	2.2364e-05	5.3688e-06	4.1655	3.1063e-05
Beta(daily_course_ft3)	-0.006855	0.008082	-0.84818	0.39634
Beta(day_hour3)	-1.803e-05	4.3117e-06	-4.1816	2.8946e-05
Beta(day_hour_ft3)	0.69059	0.046371	14.8925	3.6846e-50
Beta(delay_stop3)	0.0012732	0.00096803	1.3153	0.18841
Beta(feels_like)	-5.6629e-07	7.3072e-07	-0.77497	0.43836
Beta(holiday_weekend)	8.6267e-05	1.9079e-05	4.5216	6.1364e-06
Beta(humidity)	-1.9044e-06	3.9947e-07	-4.7673	1.8668e-06
Beta(lockdowns)	3.1312e-05	1.6801e-05	1.8637	0.062365
Beta(outliers3)	0.00065702	1.3592e-05	48.34	0
Beta(pour)	1.4492e-05	1.5805e-05	0.91694	0.35917
Beta(stay_stop2)	0.00032139	8.1263e-06	39.5489	0
Beta(time_stop2)	-0.02096	0.0075395	-2.78	0.0054359
Beta(week_day)	5.0298e-06	3.8108e-06	1.3199	0.18687
Beta(weekly_course)	-9.9998e-08	2.189e-07	-0.45682	0.6478
Beta(weekly_course_ft3)	0.57635	0.11711	4.9216	8.5841e-07
Beta(year_day)	-3.5386e-07	5.6883e-07	-0.62208	0.53389
Beta(year_month)	1.6657e-05	1.7546e-05	0.94935	0.34244
Variance	1.1579e-07	2.577e-08	4.4934	7.0099e-06

Best goodness of Fit is at AIC: -95661.5808; BIC:-95413.3924

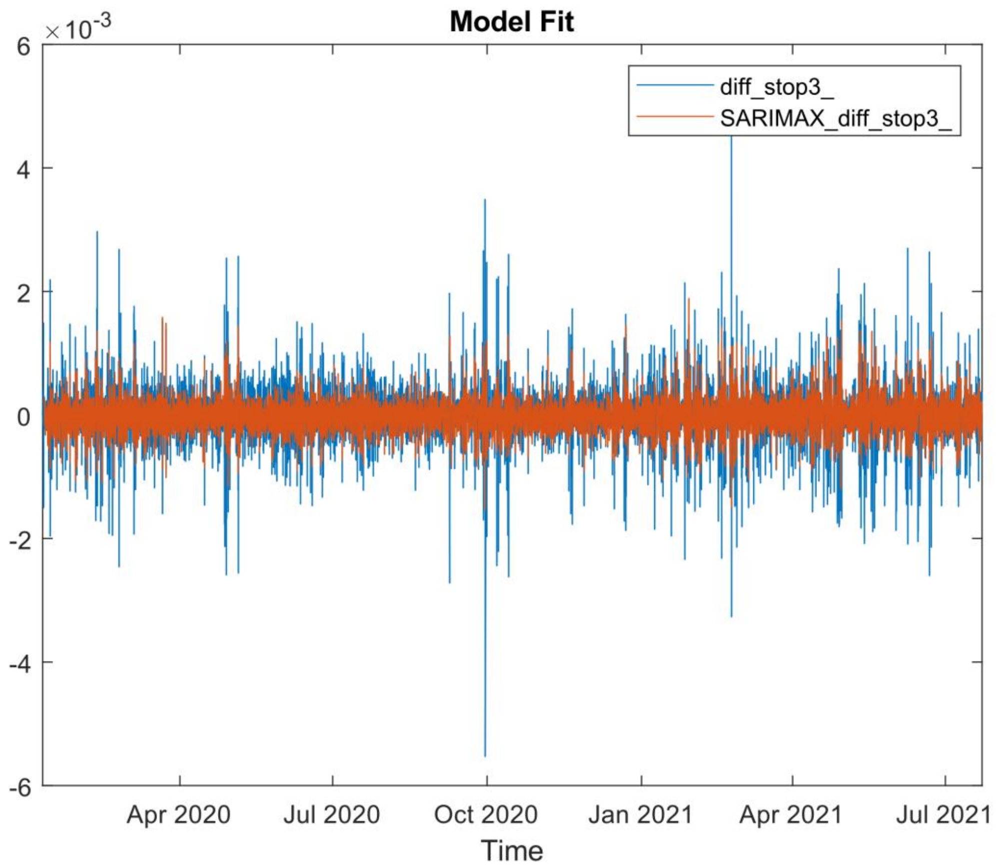

Figure 4.2.3-3 Plot the fit of model SARIMAX and time series stop 3

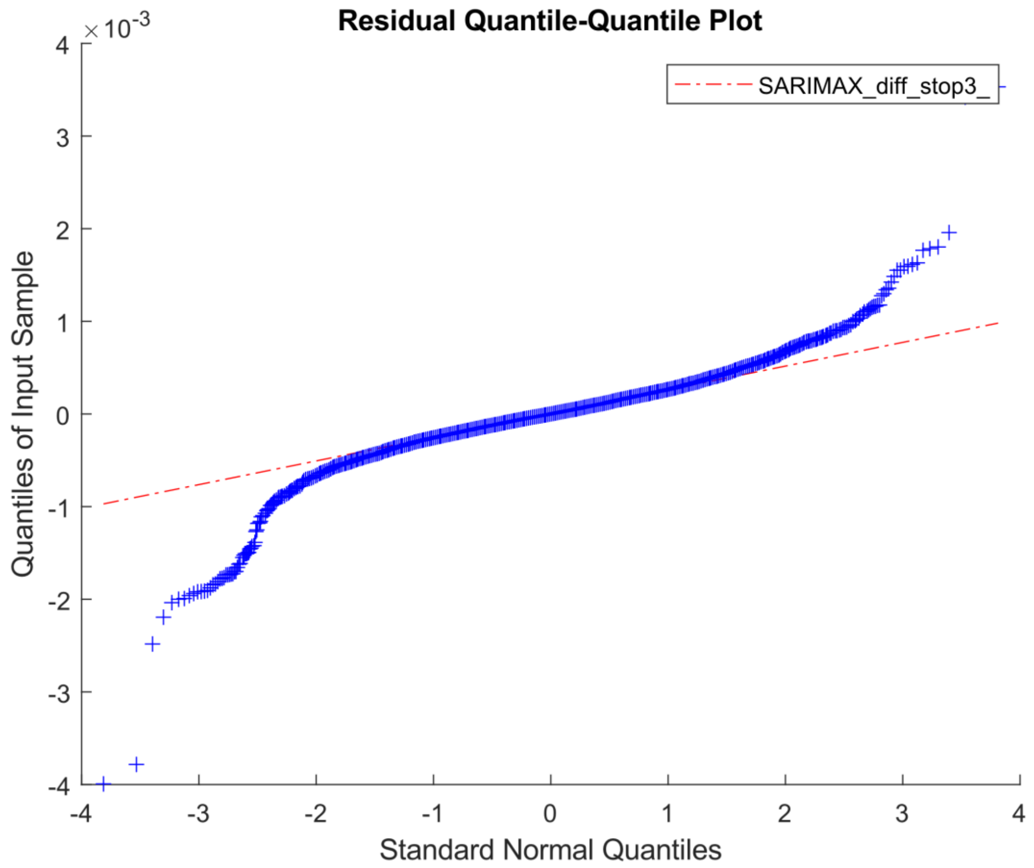

Figure 4.2.3-4 Quantile-quantile plot of the residuals of model SARIMAX for stop 3

- SARIMAX(13,2,1,1,0,1) for stop 4

Seasonal ARIMA model of time series diff_stop4_ using exogenous predictors. The model uses the following equation:

$$(1-\phi_1 L - \ldots - \phi_{13} L^{13})(1-L)y_t \;=\; c \;+\; X_1\beta_1 + \ldots + X_{20}\beta_{20} + (1+\theta_1 L)(1+\Theta_1 L)\varepsilon_t$$

Table 4.2.3-3 Estimation Results

Parameter	Value	Standard Error	t Statistic	P-Value
Constant	-5.9846e-05	7.1073e-05	-0.84203	0.39977
AR{1}	-1.6573	0.024511	-67.6147	0
AR{2}	-2.0454	0.039631	-51.6107	0
AR{3}	-2.2328	0.049351	-45.2425	0
AR{4}	-2.2707	0.055077	-41.2272	0
AR{5}	-2.189	0.058097	-37.6784	1.123e-310
AR{6}	-2.0267	0.058737	-34.5053	6.684e-261
AR{7}	-1.8326	0.057298	-31.9838	1.8297e-224
AR{8}	-1.5899	0.054185	-29.3422	3.0015e-189
AR{9}	-1.3136	0.048178	-27.2656	1.0843e-163
AR{10}	-1.023	0.040523	-25.2457	1.2637e-140
AR{11}	-0.72879	0.031723	-22.9736	8.5642e-117
AR{12}	-0.44074	0.021671	-20.3377	5.9628e-92
AR{13}	-0.18034	0.010905	-16.5375	1.9718e-61

Parameter	Value	Standard Error	t Statistic	P-Value
MA{1}	0.030396	0.014087	2.1577	0.030954
SMA{1}	0.030396	0.014088	2.1576	0.030961
Beta(clouds_all)	6.804e-07	2.6007e-07	2.6162	0.0088916
Beta(daily_course)	8.0429e-06	8.3066e-06	0.96825	0.33292
Beta(daily_course_ft4)	0.023897	0.0115	2.078	0.037708
Beta(day_hour4)	-1.3383e-05	6.5464e-06	-2.0444	0.040914
Beta(day_hour_ft4)	0.18492	0.0273	6.7736	1.2559e-11
Beta(delay_stop4)	0.036343	0.0019026	19.1021	2.4258e-81
Beta(feels_like)	-1.5467e-06	1.1243e-06	-1.3756	0.16894
Beta(holiday_weekend)	0.00015858	2.7836e-05	5.697	1.2196e-08
Beta(humidity)	-3.9677e-06	6.2762e-07	-6.3218	2.5854e-10
Beta(lockdowns)	6.8149e-05	2.7774e-05	2.4537	0.014138
Beta(outliers4)	0.00052314	2.4305e-05	21.5238	9.3269e-103
Beta(pour)	7.5714e-05	3.0917e-05	2.4489	0.014328
Beta(stay_stop2)	4.8359e-05	1.8702e-05	2.5858	0.009715
Beta(stay_stop3)	0.00028079	1.4047e-05	19.9892	6.8372e-89
Beta(time_stop2)	-0.014008	0.018107	-0.77363	0.43915
Beta(time_stop3)	0.039206	0.022243	1.7626	0.077973
Beta(week_day)	3.9385e-06	6.1096e-06	0.64465	0.51916
Beta(weekly_course)	1.0663e-07	3.394e-07	0.31418	0.75339
Beta(year_day)	2.5331e-07	8.8738e-07	0.28545	0.7753
Beta(year_month)	-3.048e-08	2.7184e-05	-0.0011212	0.99911
Variance	3.8601e-07	4.3229e-08	8.9295	4.279e-19

Best goodness of Fit is at AIC-86884.136; BIC:-86629.0535

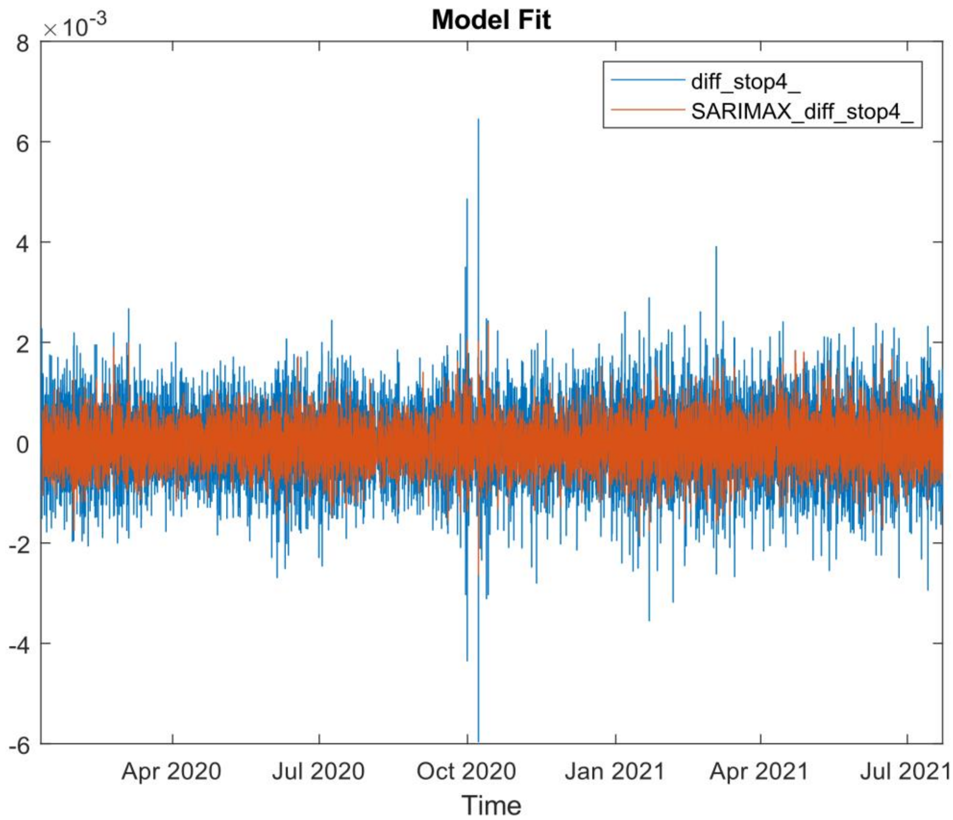

Figure 4.2.3-5 Plot the fit of model SARIMAX and time series for stop 4

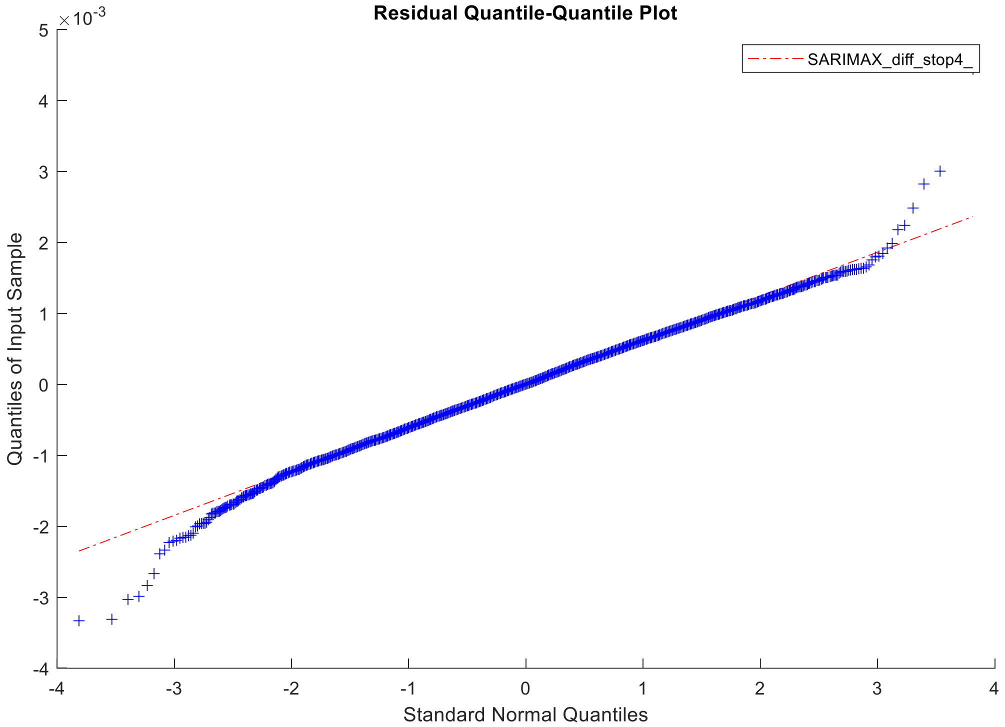

Figure 4.2.3-6 Quantile-quantile plot of the residuals of model SARIMAX for stop 4

4.3. Model validation.

4.3.1. Out of sample prediction.

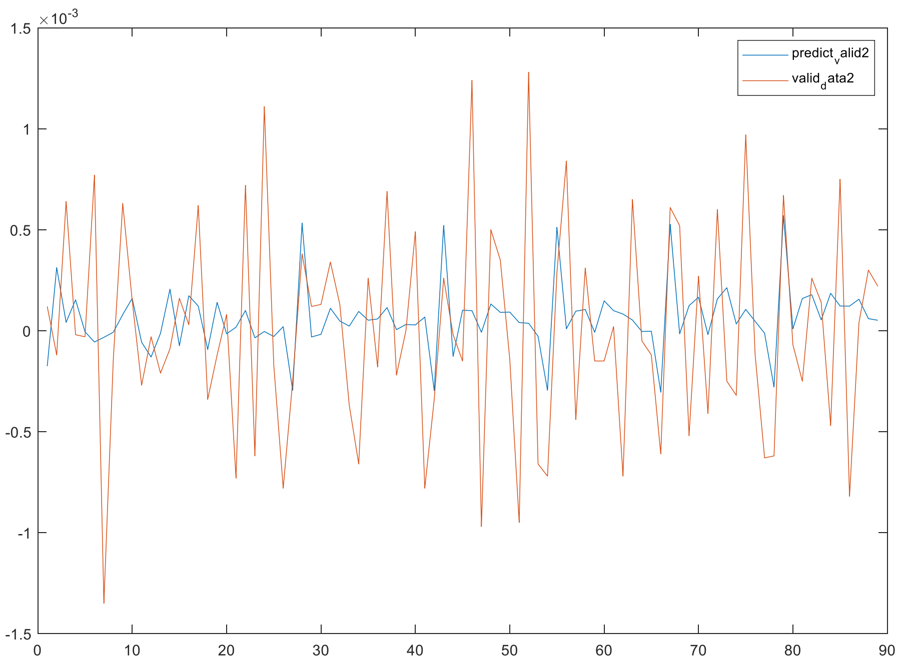

Figure 4.3.1-1 Out of sample prediction and test set for stop 2

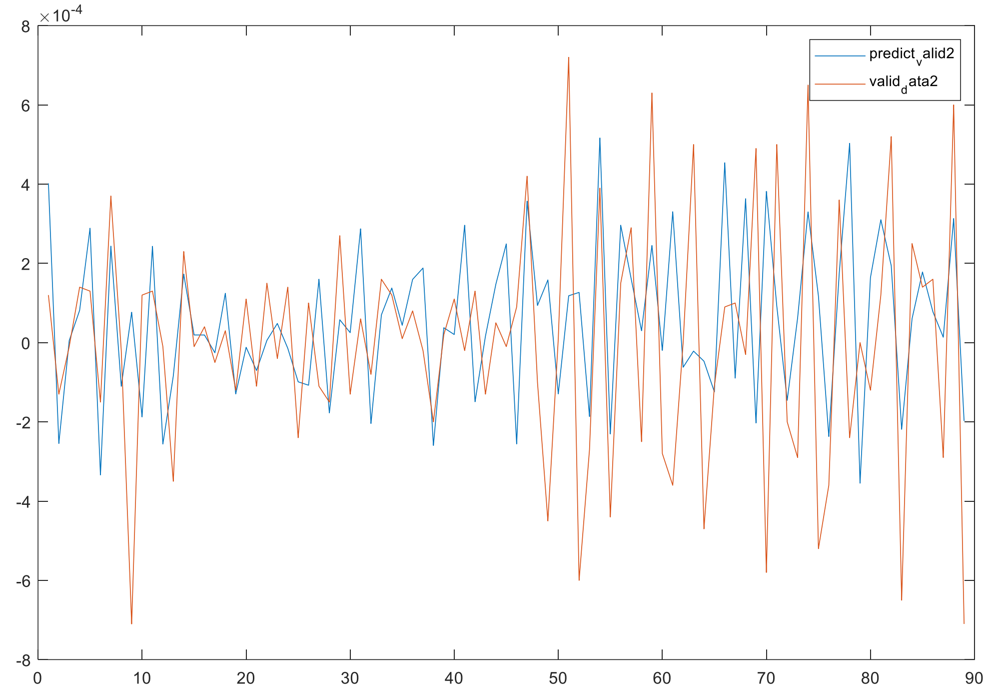

Figure 4.3.1-2 Out of sample prediction and test set for stop 3

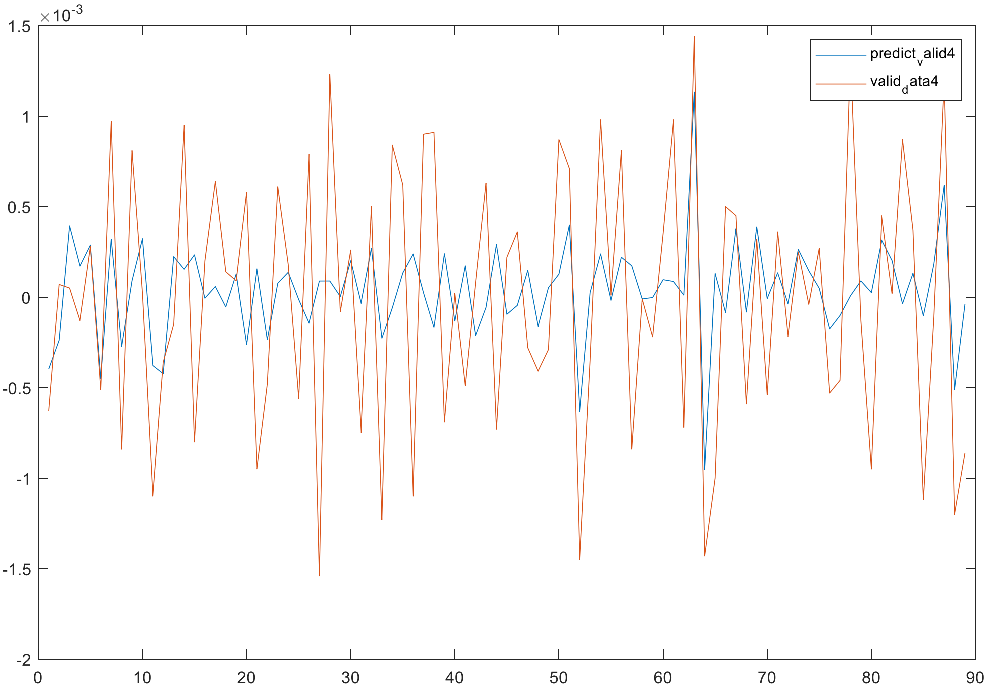

Figure 4.3.1-3 Out of sample prediction and test set for stop 4

4.3.2. Relative mean square error.

Table 4.3.2-1 Relative mean square error in seconds

	Stop 2	Stop 3	Stop 4
RMSE (seconds)	43.16	26.95	54.82