Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 11, SEPTEMBER 2022

DOI: 10.37943/HFCH4395

Gulnara Bektemyssova
Candidate of Technical Sciences, Associate Professor
Department of Computer Engineering and Information Security
g.bektemisova@gmail.com, orcid.org/0000-0002-0850-0558
International Information Technology University, Kazakhstan

Abdul Rahim Bin Ahmad
Associate Professor, Systems and Networks Department
Faculty (College) of Information Technology
abdrahim.ahmad@gmail.com, orcid.org/0000-0001-6305-0660
Tenaga National University, Malaysia

Sharafat Mirzakulova
PhD, Associate Professor
Department of Digital Technologies and Art
mirzakulova@mail.ru, orcid.org/0000-0002-1400-4729
Turan University, Kazakhstan

Zhanar Ibraeva
Master, Senior Lecturer
Department of Radioengineering, Electronics and Telecommunications
ijanar@mail.ru, orcid.org/0000-0003-3196-696X
International Information Technology University, Kazakhstan

TIME SERIES FORECASTING BY THE ARIMA METHOD

Abstract: The variety of communication services and the growing number of different sensors with the appearance of IoT (Internet of Things) technology generate significantly different types of network traffic. This implies that the structure of network traffic will be heterogeneous, which requires deep analysis to find the internal features underlying the data. A common model for analyzing the processes of a multiservice network is a model based on time series.

Numerous empirical data studies indicate that the packet intensity time series do not belong to the general aggregates of a normal distribution.

The problem of predicting network traffic is still relevant due to managing information that flows into a heterogeneous network.

In this work, the authors studied the time series for stationarity in order to select an appropriate forecasting model. A visual assessment of the series assumed non-stationarity. The Augmented Dickey-Fuller Test is applied, and the measured network traffic is predicted using the ARIMA (Auto-Regressive Integrated Moving Average) statistical method. Results were obtained using the Econometric Modeler Matlab (R2021b) application. The results of the autocorrelation function (ACF) and partial ACF are analyzed, with the help of which the ARIMA model is optimized. As a result of the study, a software algorithm for the ARIMA (0,2,1) model was developed.

Keywords: time series, network traffic, data analysis, forecasting, ARIMA.

DOI: 10.37943/HFCH4395

Introduction

Management of network devices of a functioning multiservice network allows one to respond to an ever-increasing amount of transmitted information and quickly allocates the necessary resource.

Moreover, predicting network traffic remains an urgent task as users generate ever-increasing data. Predictive data provide the necessary information to solve the problem of managing information flows in the network.

Time series modeling is one of the ways to predict them. When modeling a time series, it is usually considered a random process (stochastic) as a statistical phenomenon that develops in time according to the laws of probability theory [1].

The main purpose of time series analysis is to identify and understand patterns underlying data over time and make predictions.

A stationary series is a series whose behavior in the present and future coincides with the behavior in the past, i.e., properties are not affected by changing the origin of time. Non-stationarity means variability in the time of the distribution function. Stationary and non-stationary time series differ in the presence/absence of factors that form the levels of the series.

Literature review and problem statement

In [1], the authors described that forecast data provide the necessary information to solve the problem of managing information flows in the network. Modeling time series is one way to predict them.

Among the statistical approaches, the ARIMA (Auto-Regressive Integrated Moving Average) method allows one to describe non-stationary time series, which are reduced to stationary series by taking differences of some order from the original time series.

The authors of [2] propose several recurrent neural network (RNN) architectures to solve the problem of predicting network traffic.

In [3], the authors described that Autoregressive Integrated Moving Average (ARIMA) Box-Jenkins models combine the autoregressive and moving average models to a stationary time series after the appropriate transformation, while the nonlinear autoregressive (NAR) or the autoregressive neural network (ARNN) models are of the kind of multi-layer perceptron (MLP), which compose an input layer, hidden layer, and an output layer. The study results indicate that the traditional Box-Jenkins model was more accurate than the NAR model in modeling the monthly streamflow of the studied case.

In [4-7], the authors described that, according to forecasters' estimates, there are already more than a hundred forecasting methods, which raises the problem of choosing methods that would give adequate forecasts for the processes or systems under study.

Considering the above review of scientific research, in this paper, we have chosen the time series modeling method for predicting network traffic. As for the use of neural networks, their use is advisable with a large amount of data, but in this work, the number of points is only 18,000.

The aim and objectives of the study

In studying time series, it is important to know the stationary or non-stationary series since they have different statistical characteristics and are estimated differently.

This work aims to analyze, model, and predict the measured series.

To achieve this, the following objectives were set:

– perform a visual evaluation of the series graph;

– perform a test check of the series for stationarity;

– In the case of non-stationarity, choose one of the forecasting models for the non-stationary time series and determine the model parameters;

– building a forecast of future values of the time series.

Materials and methods

We analyzed the measured one-dimensional series (MPEG packet intensity) in this work. The measured series shows the total packets transmitted over the backbone network for each second, and the number of points is 18000. The graph of the measured data is shown in Fig.1. The number of packets received in 5 hours is displayed vertically, and the time (in seconds) is displayed horizontally.

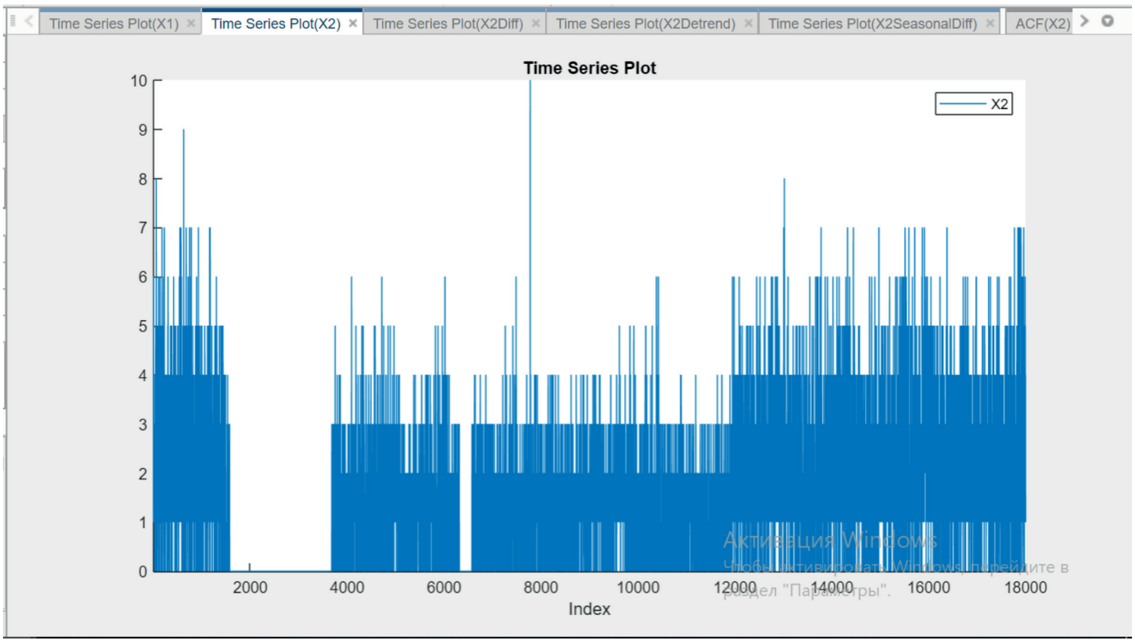
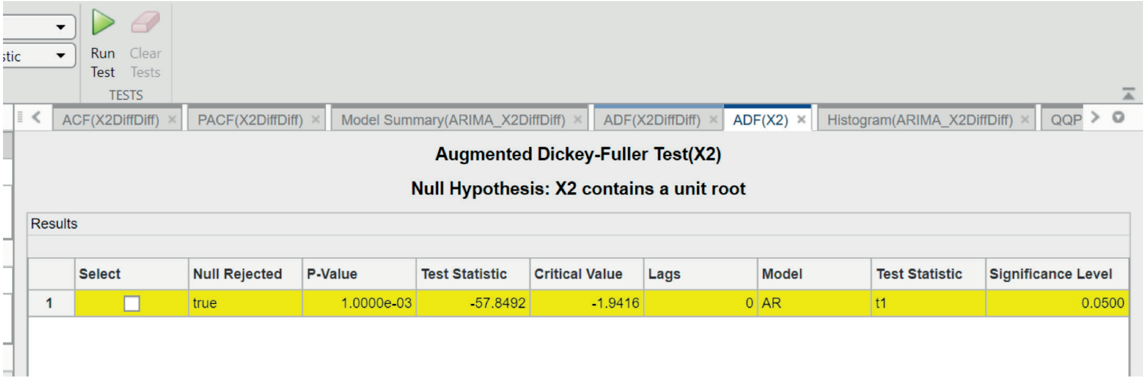


Figure 1. A series of packet transmission intensity

Econometric Modeler (R2021b) performs analysis and modeling time series interactively, and it also performs data transformation, data visualization, hypothesis testing, stationarity test, ARIMA model building, and more.

Series stationarity was investigated by the form of the autocorrelation function (ACF) and partial autocorrelation function (PACF) and by performing the Dickey-Fuller test (Fig. 2).



Augmented Dickey-Fuller Test(X2)

Null Hypothesis: X2 contains a unit root

Results

	Select	Null Rejected	P-Value	Test Statistic	Critical Value	Lags	Model	Test Statistic	Significance Level
1	□	true	1.0000e-03	-57.8492	-1.9416	0	AR	t1	0.0500

Figure 2. The results of the Dickey-Fuller test

Results

Visual analysis of the graphical representation of the time series (Fig. 1) and the evidence for a unit root in the Dickey-Fuller test results suggest that the data is not stationary.

Differentiation is an alternative transformation for removing the average trend from a non-stationary series.

This approach is supported in the Box-Jenkins approach to model specification [8]. According to this methodology, the first step in building models is to differentiate the data until it looks stationary.

To conduct statistical analysis, it is necessary to have a stationary time series.

The main methods for forecasting a non-stationary time series include:

- statistical methods;
- new methods based on artificial intelligence (AI).

Among the statistical approaches, the ARIMA (Auto-Regressive Integrated Moving Average) method allows one to describe non-stationary time series, which are reduced to stationary series by taking differences of some order from the original time series. Like taking derivatives, taking the first difference makes the linear trend constant, taking the second difference makes the quadratic trend constant, and so on for higher degree polynomials.

$$\Delta y_t = y_t - y_{t-1} \ , \tag{1}$$

here Δ is the difference operator

$$\Delta^2 y_t = (1 - L)^2 y_t = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) = y_t - 2\, y_{t-1} + y_{t-2} \tag{2}$$

If the data has features of a unit root, then the usage of a non-stationary ARIMA model can be considered.

The Box-Jenkins (BJ) or ARIMA(p,d,q) prediction method, which relies on actual data, has three model parts [8]:

- AR – part of the time series model that describes autoregression, in which the values of the series at the moment can be expressed as a linear combination of the previous values of the same series and a random error that has the “white noise” property (the p parameter is used);
- I – part of the time series model, describing the order of differentiation of the series (the parameter d is used);
- MA is a part of the time series model that describes the current value of the series and is represented as a linear combination of the current and past error values, corresponding to “white noise” in its properties (the q parameter is used).

The analysis of the ACF and PACF data of the bias series is being used to identify the parameters of the ARIMA model.

Below are graphs of the ACF and PACF functions (see Figures 3 and 4). The blue lines on the graphs mark the critical interval within which the ACF and PACF values are considered non-zero. The ACF plot is a histogram of the correlation coefficients between the time series and the lags. In PACF, unlike ACF, the influence of intermediate lags is not considered when calculating partial correlation coefficients. With the help of the ACF and PACF plots of the displacement series, it is possible to select the parameters of the ARIMA model to eliminate the autocorrelation that remains in the difference series [9].

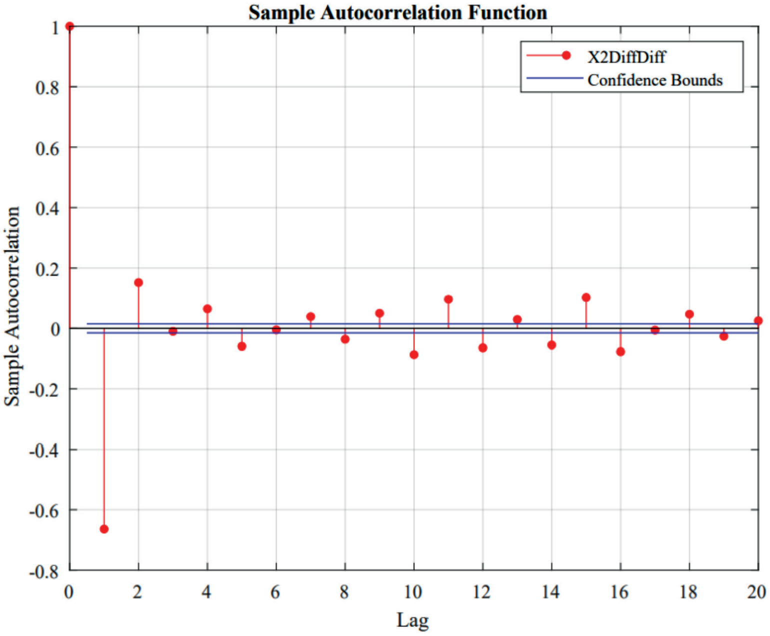


Figure 3. Sample autocorrelation function of X2DiffDiff

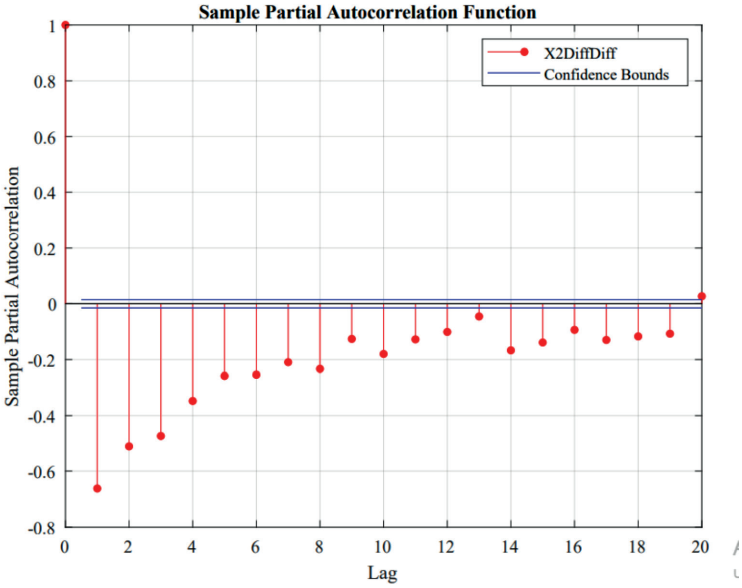


Figure 4. Sample partial autocorrelation function of X2DiffDiff

Analyzing the graphs of ACF and PACF of the second-order difference, it can be stated that the parameter d has already been described. It is equal to two (second difference), and according to the autoregression (AR) process, the levels of the ACF series decay quickly, and the levels of PACF decay gradually. In the case of an autoregressive process, the ACF function would have decayed slowly. Therefore, the parameter p is equal to zero.

For the moving average (MA) process, the ACFs decay sharply after one lag (the last significant lag shows the q parameter), while the PACF function decays gradually. Therefore, the parameter q is equal to one.

The model fitting procedure is based on finding parameters that minimize the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), which can help to reduce fitting in complex models [10].

DOI: 10.37943/HFCH4395

As a result, the integrated model ARIMA (0,2,1) Model (Gaussian Distribution) (ARIMA_X2DiffDiff) is obtained.

Auto regression integrated moving average time series model with the following equation:

$$(1 - L)^{2} y_t = c+(1+\Theta_1 L)\varepsilon_t \tag{3}$$

The Model Estimation values are given in Table 1 and Table 2. The Information Criteria (the best model with the lowest value of the Akaike AIC criterion, which is closely related to the Bayesian BIC criterion).

Table1. Estimation Results

Parameter	Value	Standard Error	t Statistic	P-Value
Constant	0.00033089	0.0004796	0.68992	0.49024
MA{1}	-0.98716	0.00074134	-1331.577	0
Variance	24.3391	0.18873	128.9617	0

The component of the moving average MA{1} has a PValue less than the significance level of 0.05 (0<0.05), so conclusion can be made. The coefficient MA{1} of the moving average is a statistically significant parameter.

Table 2. Goodness of Fit

AIC	108533.1921
BIC	108556.5858

The AIC and BIC criteria consider the degree to which the model conforms to some trade-off between model accuracy and complexity [11].

Figure 5 shows that the red part of the graph (the predicted time series) almost completely overlaps the blue part (the original series). This combined plot of the original series and the predicted data shows that the model is well-fitted. Otherwise, coverage must be partial.

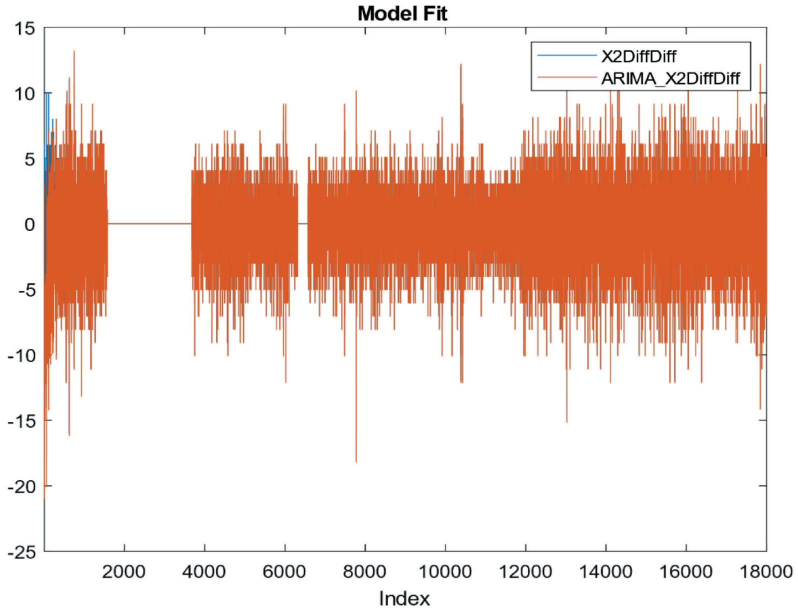


Figure 5. Plot the fit of model ARIMA_X2DiffDiff time series X2

Figures 6,7,8 show the residuals of the time series. Residuals should be uncorrelated, homoscedastic, and normally distributed with constant mean and variance. If the residuals are not normally distributed, innovation distribution must be changed to a Student's t. Figure 6 shows that the time series are distributed symmetrically about 0, and the residual is white noise.

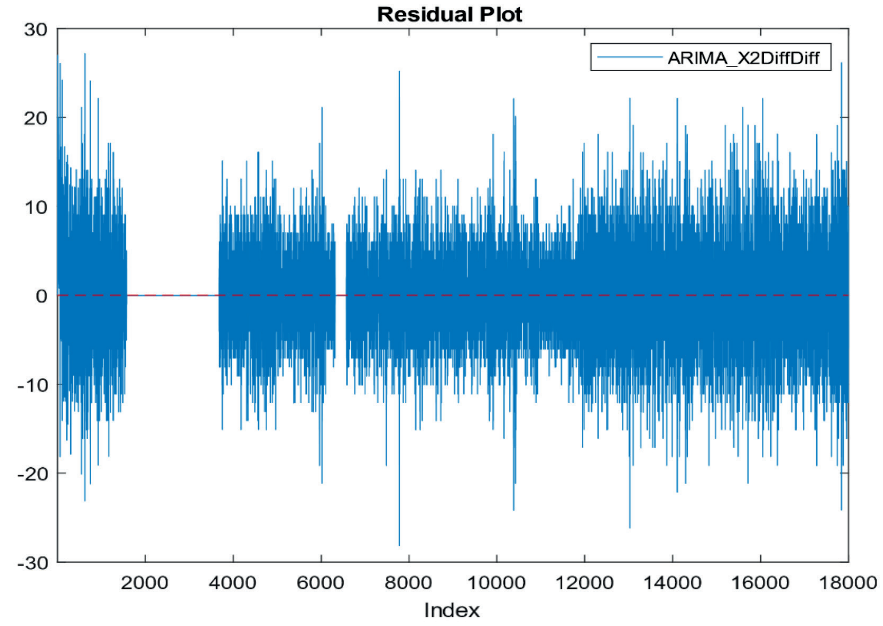


Figure 6. Plot of the residuals of model ARIMA_X2DiffDiff

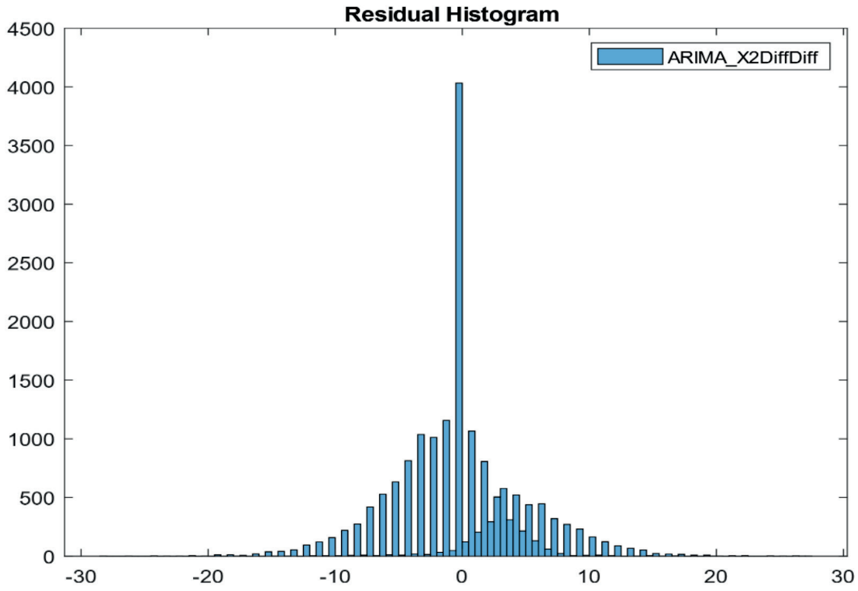


Figure 7. Histogram of residuals of the ARIMA_X2DiffDiff model

Fig.7 shows that the residuals are reasonably normally distributed and uncorrelated.

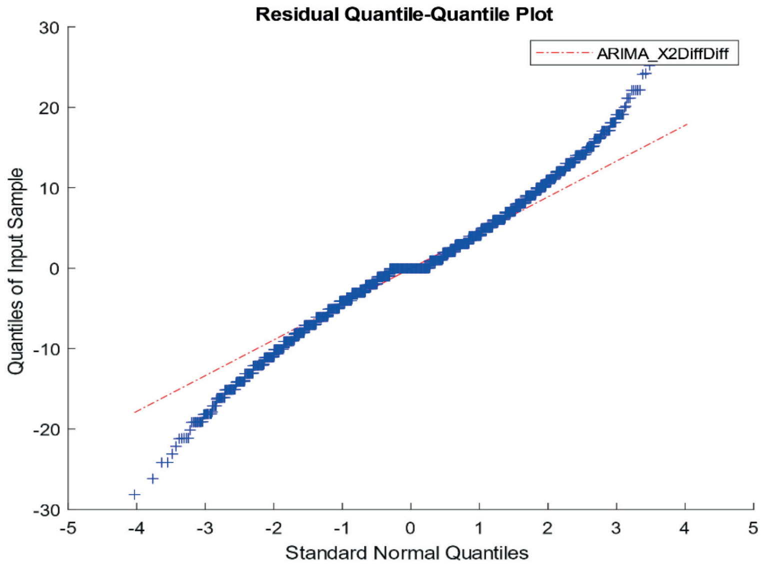


Figure 8. Quantile-quantile plot of the residuals of the ARIMA_X2DiffDiff model

The quantile-quantile plot (QQ-plot, Figure 8) of the residuals of the ARIMA model shows no obvious violations of the normality assumption.

Residual diagnostics include evaluating the model assumptions and investigating whether one must respecify the model to address other data properties. Model assumptions to assess include checking whether the residuals are centered on zero, normally distributed, homoscedastic, and serially uncorrelated. If the residuals do not demonstrate all these properties, then you must determine the severity of the departure, whether to transform the data, and whether to specify a different model.

Analyzing the graphs (Figures 6,7,8) it can be noted that the residuals of the series have a normal distribution with an average value close to zero.

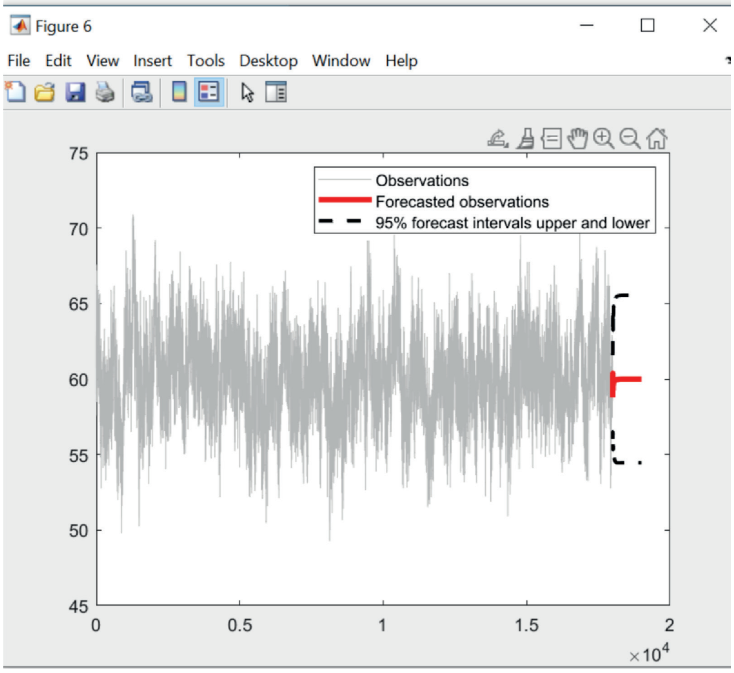


Figure 6
File Edit View Insert Tools Desktop Window Help
Observations
Forecasted observations
95% forecast intervals upper and lower
75
70
65
60
55
50
45
0
0.5
1
1.5
2
×10^4

Figure 9. Forecasting by the ARIMA (0,2,1) model

Considering all the above, we can say that the resulting ARIMA predictive model is adequate. Another important goodness-of-fit check of the model is predictive-performance assessment. Figure 9 shows forecasted observations with taking into account the upper and lower confidential intervals with 95%.

Discussion of results

Visual analysis of the graphical representation of the time series (Fig. 1) shows that the series has uneven intensity. Therefore, it can be supposed that the series is non-stationary.

Figure 2 shows that the series X2 contains a unit root, therefore, the series is not stationary.

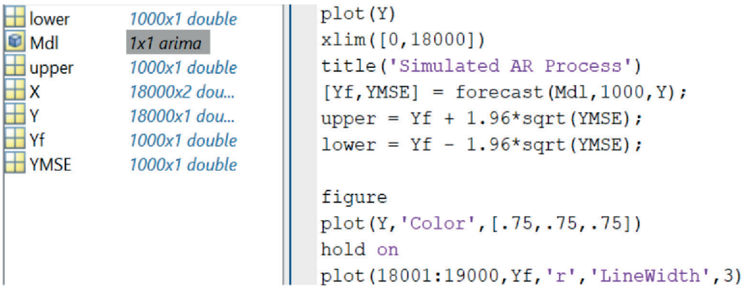
The ACF plot (Fig. 3) builds a histogram of the correlation coefficients between the time series and the lag itself. PACF (Fig. 4) does not consider the influence of lags when calculating partial correlation coefficients.

Plot the fit of model ARIMA (Fig. 5) shows that we have chosen the right model with the right parameters and adequate predictions. The blue part of the graph is barely visible, indicating a perfect forecast.

The plots of residuals (Fig. 6,7,8) show that the series' plots have a normal distribution. According to the ARIMA method, one of the steps in building models is to difference the data from the original time series until it looks stationary. As is required, these graphs look stationary with an average value close to zero.

ARIMA model (Fig. 9) shows the adequate forecast with 95% forecast intervals.

The program code of the ARIMA (0,2,1) model is shown in Figure 10.



lower	1000x1 double
Mdl	1x1 arima
upper	1000x1 double
X	18000x2 dou...
Y	18000x1 dou...
Yf	1000x1 double
YMSE	1000x1 double

```
plot(Y)
xlim([0,18000])
title('Simulated AR Process')
[Yf,YMSE] = forecast(Mdl,1000,Y);
upper = Yf + 1.96*sqrt(YMSE);
lower = Yf - 1.96*sqrt(YMSE);

figure
plot(Y,'Color',[.75,.75,.75])
hold on
plot(18001:19000,Yf,'r','LineWidth',3)
```

Figure 10. Program code for prediction

Figure 10 shows a program code for forecasting traffic values for 1000 steps ahead for the selected time series section, which consists of 18000 points.

Conclusion

Since 2007, a new generation NGN network (Next Generation Network) based on IP (Internet Protocol) has been operating in the Republic of Kazakhstan. The present work is based on the classic ARIMA method for predicting the future values of empirical data in the field of telecommunications. This work aims to predict the subsequent levels of a series to control information flows in the network to avoid overloads and losses.

As a result of the study, a visual analysis of the studied time series showed that the series has an uneven packet intensity, which allows for the non-stationarity of the data series.

For the reliability of the assumptions about the non-stationarity of the data series, test experiments were carried out from among the Unit Root. The Augmented Dickey-Fuller test confirmed the presence of a unit root, and hence the series is not stationary.

Further, for forecasting the time series, the ARIMA method (autoregressive integrated moving average) was used to describe non-stationary time series and make reliable shortterm forecasts with a minimum number of parameters. A forecast model is defined for predicted data based on previous values. The results of ACF and PACF were analyzed, based on which differentiation was performed to obtain the correct parameters of the ARIMA model: autoregression parameter - 0, integration order - 2, and moving average parameter - 1. As a result of the study, a software algorithm for the ARIMA (0,2,1) model was developed.

A predictive ARIMA model was obtained with an accuracy of 95% (Fig.9), proving the chosen model's adequacy.

Reference

1. Serikov, T., Zhetpisbayeva, A., Akhmediyarova, A., Mirzakulova, S., Kismanova, A., Tolegenova, A., & Wójcik, W. (2021). City backbone network traffic forecasting. *International Journal of Electronics and Telecommunications*, *67* (3), 319–324. https://doi.org/ 10.24425/ijet.2021.135983
2. Serikov, T., Zhetpisbayeva, A., Mirzakulova, S., Zhetpisbayev, K., Ibraeva, Z., Soboleva, L., Tolegenova, A., & Zhumazhanov, B. (2021). Application of the NARX neural network for predicting a one-dimensional time series. *Eastern-European Journal of Enterprise Technologies*, *5* (4), 12–19. https://doi.org/10.15587/1729-4061.2021.242442
3. Al-Saati, N. H., Omran, I. I., Salman, A. A., Al-Saati, Z., & Hashim, K. S. (2021). Statistical modeling of monthly streamflow using time series and artificial neural network models: Hindiya Barrage as a case study. *Water Practice and Technology*, *16*(2), 681-691. https://doi.org/10.2166/wpt.2021.012
4. Khedkar, S. P., Canessane, R. A., & Najafi, M. L. (2021). Prediction of traffic generated by IoT devices using statistical learning time series algorithms. *Wireless Communications and Mobile Computing*, 1–12. https://doi.org/10.1155/2021/5366222
5. Weerakody, P. B., Wong, K. W., & Wang, G., E. (2021). A review of irregular time series data handling with gated recurrent neural networks. *Neurocomputing*, *441*, 161–178. https://doi.org/10.1016/j.neucom.2021.02.046
6. Sovetov, B. Y., Tatarnikova, T. M., & Tsekhanovskiy, V. V. (2020) Avtoregressionnyye modeli prognozirovaniya setevogo trafika. Materialy konferentsii «Informatsionnyye tekhnologii v upravlenii». [Autoregressive models for predicting network traffic. Proceedings of the conference "Information technologies in management"], Saint-Petersburg Electrotechnical University "LETI" named after V.I. Ulyanov (Lenin).
7. Rizkya, I., Syahputri, K., Sari, R. M., Siregar, I., & Utaminingrum, J. (2019, August). Autoregressive Integrated Moving Average (ARIMA) Model of Forecast Demand in Distribution Centre. In *IOP Conference Series: Materials Science and Engineering* (Vol. 598, No. 1, p. 012071). IOP Publishing. https://doi.org/10.1088/1757-899X/598/1/012071
8. Joshi, M., & Hadi, T. H. (2015). A review of network traffic analysis and prediction techniques. *arXiv preprint arXiv:1507.05722*.
9. Rutka, G. (2008). Network traffic prediction using ARIMA and neural networks models. *Elektronika ir Elektrotechnika*, *84*(4), 53-58.
10. Brockwell, P. J., & Davis, R. A. (Eds.). (2002). *Introduction to time series and forecasting*. New York, NY: Springer New York. https://doi.org/10.1007/0-387-21657-X_8
11. Brownlee, J. (2017). *How to Create an ARIMA Model for Time Series Forecasting in Python*. Machine Learning Mastery. https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/
12. Maltseva, K. (2018). *ARIMA: making predictions based on history.* Foresight. https://www.fsight.ru/blog/arima-stroim-prognoz-na-osnove-istorii/