jet INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2021, VOL. 67, NO. 3, PP. 319-324
Manuscript received July 28, 2021; revised July, 2021 DOI: 10.24425/ijet.2021.135983

City Backbone Network Traffic Forecasting

Tansaule Serikov, Ainur Zhetpisbayeva, Ainur Akhmediyarova, Sharafat Mirzakulova, Aigerim Kismanova, Aray Tolegenova, and Waldemar Wójcik

***Abstract*—The work considers a one-dimensional time series protocol packet intensity, measured on the city backbone network. The intensity of the series is uneven. Scattering diagrams are constructed. The Dickie Fuller test and Kwiatkowski-Phillips Perron-Shin-Schmitt test were applied to determine the initial series to the class of stationary or non-stationary series. Both tests confirmed the involvement of the original series in the class of differential stationary. Based on the Dickie Fuller test and Private autocorrelation function graphs, the Integrated Moving Average Autoregression Model model is created. The results of forecasting network traffic showed the adequacy of the selected model.**

***Keywords*—time series, packet intensity, Dickie Fuller test, Kwiatkowski-Phillips Perron-Shin-Schmitt test were, forecasting, Integrated Moving Average Autoregression Model**

I. Introduction

THE further evolution of the interconnected worldwide communication network based on packet technology has caused a sharp increase in the amount of data associated with information flows from various human activities.

The ever-increasing amount of information passed through creates a certain complexity for the underlying data transmission network in its processing. On the other hand, modern society requires high transmission speeds of processed information.

Resources of a functioning multiservice network allow you to quickly respond to market changes, quickly collect and deliver the necessary information to consumers, and be updated in a timely manner in accordance with new applications.

As users generate ever-increasing data, forecasting network traffic (data volume) remains an urgent task. 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.

A special property of the time series is that it is a sequence of numerical indicators ordered in time, which characterize the level of state and changes in the phenomenon under study. The time series studied in this work is the packet intensity measured at constant time intervals. This is one-dimensional distribution function.

Time series models explain the behavior of a variable that changes over time, based only on its previous values. Depending on the presence of certain factors, the time series can be stationary and non-stationary.

The levels of stationary series are formed under the influence of random factors that act in different directions and with different intensities. The non-stationary series always has a tendency, which is characterized by nonrandom factors in the processes represented by this time series. A time series is called non-stationary if its characteristics (average value, variance and autocorrelation function) depend on time.

II. Network Traffic Forecasting

The series under investigation displays the number of UDP (User Datagram Protocol) packets per every 10 seconds. Realistically captured data on the backbone of a multiservice network, as a result of traffic tracking for 5 hours, since this is a self-similar traffic. As a result, 287 packets were received. With the help of the Wire Shark program, only UDP packets were removed from the network traffic. This series has 1800 levels. This is a main network traffic (Figure 1).

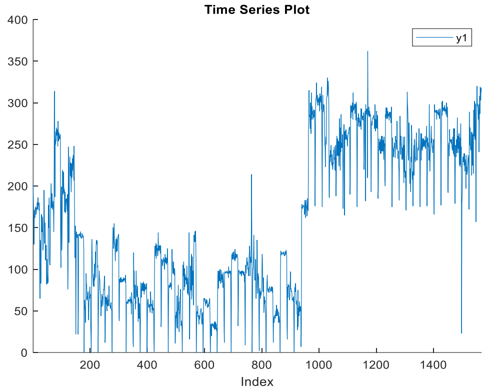


Fig. 1. Time Series Plot

Various methods can be used to recognize the stationarity or non-stationarity of a time series: visual analysis of a graphical representation of the time series for the presence of a trend and a periodic component, the average method, analysis of the time series for the presence of autocorrelation, etc.

Visual analysis of the graphical representation of the time series shows that the series has uneven intensity (the scatter of observations increases and decreases with time), there are ripples in traffic intensity with significant dispersion, there are groups in "packs" in some places, or there are dispersed sections in other time intervals where there are no or few incoming packets. The time series model assumes the

This research has been funded by the Ministry of digital development, innovation and aerospace industry of the Republic of Kazakhstan (Grant No AP06851342).

Tansaule Serikov, Ainur Zhetpisbayeva, Aigerim Kismanova and Aray Tolegenova are with S.Seifullin Kazakh AgroTechnical university, Nur-Sultan, Kazakhstan (e-mail: tansaule_s@mail.ru; aigulji@mail.ru; akismanova@mail.ru; arai82@bk.ru).

Ainur Akhmediyarova is with Institute of Information and Computational Technologies, Almaty, Kazakhstan(e-mail: aat.78@mail.ru).

Sharafat Mirzakulova is with Turan University, Almaty, Kazakhstan(e-mail: mirzakulova@mail.ru).

Waldemar Wójcik is with Lublin University of Technology, Poland (e-mail: waldemar.wojcik@pollub.pl).

relationship between the current and previous observations. Previous observation of the time series is called a lag. The scattering diagram shows the relationship between observation and lag (Figure 2), which shows that the resulting distribution is not distributed in all four quadrants [1].

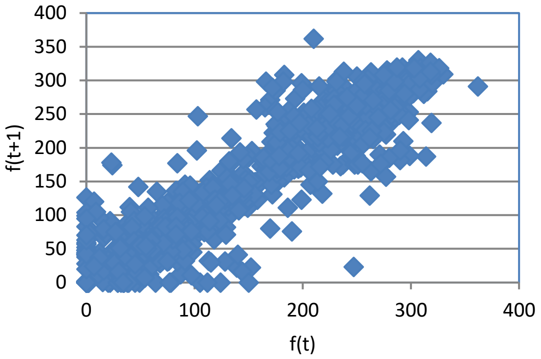


Fig. 2. Scatter diagram of the original series

A uniform distribution of points in all four quadrants would mean independence of neighboring values. The studied data are mainly limited to two quadrants. In this scatterplot diagram $f_{i+1} = f(f_i)$, the slope (bond direction) and the width (bond strength) of an imaginary ellipse are of interest, as it reflects the tightness of the linear relationship between the two measured correlation coefficients. For most intervals, the packet intensities are similar. This suggests that a positive correlation is possible here. Based on the above, it follows that the investigated series is non-stationary. Let's construct a scatter plot of the increments of this series. The transition to increments makes the time series more stationary. To convert the original non-stationary series to stationary, we perform differentiation (taking the finite differences of the values of the series) according to the formula:

$$Y(t)=X(t+1)-X(t). \qquad (1)$$

Figure 3 shows the increments scatter diagram, which differs from the diagram of the initial series and represents a distribution whose points are distributed in all four quadrants, which means the relative independence of neighboring values and these values are collected with a certain dispersion around the zero mathematical expectation.

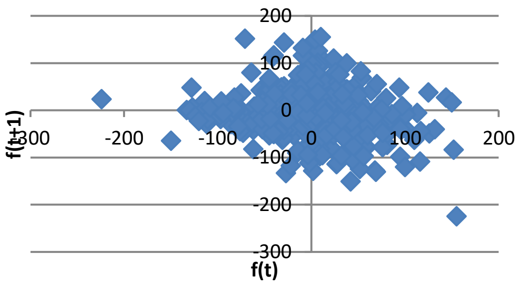


Fig. 3. Scatter plot of increments series

Figure 4 shows a scatter plot of increments that are randomly shuffled. In this case, the obtained scatter diagram differs from the scattering diagram of the increments of the original series in that the points are more evenly distributed in all four quadrants in comparison with the scattering diagram of the increments, where the points are, as it were, slightly larger in the second and fourth quadrants.

To check for the presence of unit roots in the original time series, we will use the Dickie Fuller -test of the Python program.

We will use unit root tests to evaluate stationarity. The concept of "unit root" is an indicator that determines the nature of fluctuations in the system. The system of linear difference equations of the Nth order has N roots. If the absolute value of any of them is greater than 1, the system is approaching "explosion", at least until it meets some constraints, because of which it ceases to be linear. If all roots are less than 1 in absolute value, the system will inevitably strive for its initial equilibrium after any temporary deviations. A root equal to 1 in absolute value, or a unit root, will cause a stable shift in the system, and a series of violations can cause an infinite deviation from the original position. A large number of methods have been developed to statistically check the presence of a unit root.

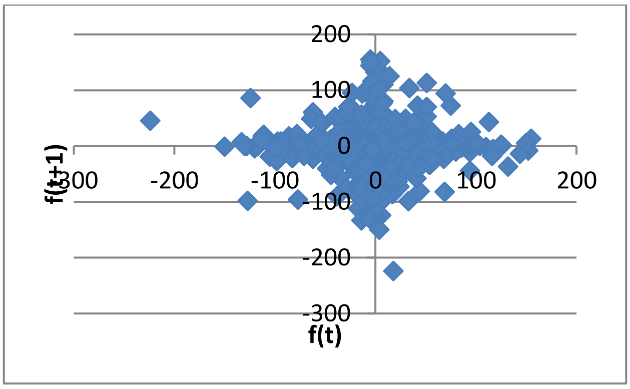


Fig. 4. Scatter plot of a series of shuffled increments in random order

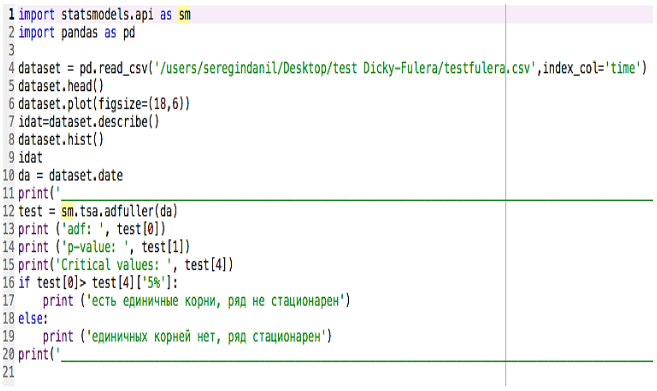
We will use unit root tests to evaluate stationarity. The concept of "unit root" is an indicator that determines the nature of fluctuations in the system. The system of linear difference equations of the Nth order has N roots. If the absolute value of any of them is greater than 1, the system is approaching "explosion", at least until it meets some constraints, because of which it ceases to be linear. If all roots are less than 1 in absolute value, the system will inevitably strive for its initial equilibrium after any temporary deviations. A root equal to 1 in absolute value, or a unit root, will cause a stable shift in the system, and a series of violations can cause an infinite deviation from the original position. A large number of methods have been developed to statistically check the presence of a unit root.

In [2], it is described that the ADF-test (Dickey-Fuller test) checks for the presence of a unit root in order to identify the type of stationarity / non-stationarity of a series. Moreover, the null hypothesis H_0 corresponds to a series of type DS, and the alternative hypothesis corresponds to a series of type TS. The authors of the test are American statisticians David Alan Dickey and Wayne Arthur Fuller. This test was proposed by them in 1979. The essence of the ADF-test is as follows: if a unit root was obtained in the verification in the autoregressive model of a time series, this means the integration of the time series.

In this case, the following hypothesis is used:

- H₀ $\theta = 1$ - the series is non-stationary: it contains a unit root, belongs to DS-series and is described by a random walk process.
- H₁: $\theta < 1$ - the row is stationary – it does not contain a unit root and it is described by a stationary first-order autoregressive process.

In the Python program in the statsmodels module, which provides classes and functions for evaluating many different statistical models and conducts statistical tests, there is an *adfuller()* function for studying statistical data, the time series of the UDP protocol is examined for the presence of unit-roots. The source code for the stationarity estimation program in the Python program is shown in Figure 5.



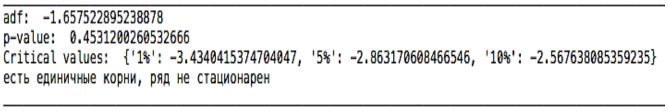
```
import statsmodels.api as sm
import pandas as pd

dataset = pd.read_csv('/users/seregindanil/Desktop/test Dicky-Fulera/testfulera.csv',index_col='time')
dataset.head()
dataset.plot(figsize=(18,6))
idat=dataset.describe()
dataset.hist()
idat
da = dataset.date
print('________________________________________________________________
test = sm.tsa.adfuller(da)
print ('adf: ', test[0])
print ('p-value: ', test[1])
print('Critical values: ', test[4])
if test[0]> test[4]['5%']:
    print ('есть единичные корни, ряд не стационарен')
else:
    print ('единичных корней нет, ряд стационарен')
print('________________________________________________________________

```

Fig. 5. Source code of the program

As a result, the assumption that the series is non-stationary was confirmed. The tables of the Dickey-Fuller test are calculated for significance levels of 1%, 5%, 10% with the corresponding empirical values, while the calculated value is -1.658 and all critical values are less than the reference ones, namely: –3.434; –2.863 and –2.5676, it indicates that it is impossible to reject the hypothesis that the time series under study has the character of random walk (Figure 6).



```
adf:  -1.657522895238878
p-value:  0.4531200260532666
Critical values:  {'1%': -3.4340415374704047, '5%': -2.863170608466546, '10%': -2.567638085359235}
есть единичные корни, ряд не стационарен
```

Fig. 6. The result of the test

Vershinina [3] describes that for reliability of results it is common to use not one, but several tests when series are analyzed for their belonging to the class of stationary or non-stationary ones.

The KPSS test (Kwiatkowski – Phillips – Schmidt – Shin) was developed in 1992, it is based on linear regression and contains three components[4]:

- deterministic trend (σt);
- random walk (c_t);
- null hypothesis (H₀);
- alternative hypothesis (Hₐ);
- stationary error (u_{1t}).

$$y_t = c_t + \sigma t + u_{1t}\,, \tag{1}$$

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

$$u_{2t} \;\sim\; i.i.d(0, \sigma^2)\,, \tag{3}$$

$$H_0 : \sigma^2 = 0\,, \tag{4}$$

$$H_a : \sigma^2 > 0\,, \tag{5}$$

The KPSS test has Null Hypothesis: y1 is trend stationary (the series belongs to the TS series), and the alternative hypothesis is the non-stationary series (presence).

Table 1 shows the KPSS test parameters when checking the initial series in the Matlab program, and table 2 shows the test results.

TABLE I
TEST PARAMETERS

	Lags	Include Trend	Significance Level
1	0	true	0,05
2	0	false	0,05
3	1	false	0,05
4	2	false	0,05
5	1	false	0,05

TABLE II
TEST RESULTS

	Null Rejected	P-Value	Test Statistic	Critical Value
1	true	0,01	16,4938	0,146
2	true	0,01	80,0526	0,463
3	true	0,01	41,0774	0,463
4	true	0,01	27,8283	0,463
5	true	0,01	41,0774	0,463

In this case, the null hypothesis is rejected, since the value of the statistics of the KPSS test is more than the critical value – the series is non-stationary [4].

Since both tests rated the time series as non-stationary, therefore, it is really not stationary. The levels of stationary series are formed under the influence of random factors, acting in different directions and with different intensity. In 1938 Wold proved, that any stationary, in a broad sense, random process can be represented as a linear combination of white noises [5,6].

Figure 7 shows the integrated first-order series obtained by taking the first difference of the original series and the series with a biased period equal to one (differentiation). There is no trend in this series.

All methods of forecasting time series, in general, are divided, depending on the definition of the parameters of the approximating function by past values, into classes: local and global.

At the same time, it is noted that global methods have received priority development and use. They are based on statistical analysis - this is the use of linear models - autoregressive, moving average, ARMA, etc. Local methods are based on the local approximation of LA. As a result of the development of the theory of nonlinear dynamics, new methods were developed (SSA, LA and SSA-LA).

Nonlinear statistical models have also been developed, which are subdivided into two groups: parametric and nonparametric.

Parametric methods make large assumptions about the mapping of input variables to output variable and, in turn, are faster to train, require less data, but may not be as powerful.

Nonparametric methods make little or no assumptions about the objective function and, in turn, require much more data, are slower to train and have higher model complexity, but can lead to more powerful models.

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 method allows one to describe non-stationary time series, which are reduced to stationary series by taking differences of a certain order from the original time series.

Box-JenKins (BJ) forecasting method or Autoregressive integrated moving average ARIMA (p, d, q) forecasting method is an extension of ARMA model for not-stationary time series, which can be made stationary by taking differences of some order from the initial time series. One can consider ARIMA as "filter" trying separate a signal from nosie, and then signal extrapolates to obtain forecasts in future. ARIMA model is based on actual data and has three parts of the model [7, 8, 9]:

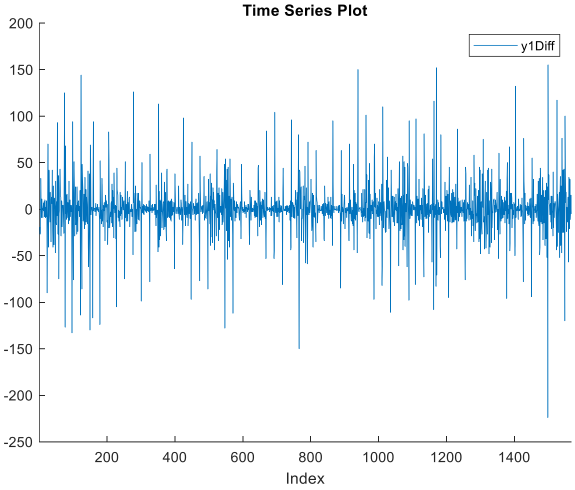


Fig. 7. Time Series Plot Diff

- AR is the 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 previous values of the same series and a random error with the "white noise" property (parameter *p* is used);
- I - part of the time series model that describes the order of differentiation of the series (the parameter d is used, which is 1 based on the above data);
- MA is the part of the time series model that describes the current value of the series and is presented as a linear combination of the current and past error values, corresponding in its properties to "white noise" (parameter q is used).

To select the parameters p and q, we will use the autocorrelation function (ACF) and the private autocorrelation function (PACF) to understand the models AR (p) and MA (q).

In contrast to ACF, the PACF does not take into account the influence of intermediate lags in the calculation of particular correlation coefficients. Therefore, ChAKF gives a more "clean Figure" of the dependence of the series on the lag.

When modeling a time series, it is usually considered as a random process (stochastic), as a statistical phenomenon that develops in time according to the laws of probability theory.

As research tools, the Econometric Toolbox application was used to simulate the process using statistical methods, the MS Excel program for scatter diagrams and the Python and Econometric Toolbox programs were applied to the original series to identify the nonstationarity properties of the series.

Time series models explain the behavior of a variable over time based only on its previous values. Moreover, depending on the presence of certain factors, the time series can be stationary and non-stationary.

To confirm the nonstationarity of the initial time series, we additionally use the KPSS test, which can detect in the process the presence of a random walk, which will lead to systematic deviations from the trend in some parts of the series.

Among the statistical approaches, the ARIMA method allows one to describe non-stationary time series, which are reduced to stationary series by taking differences of a certain order from the original time series.

The component of the moving average MA {2} has PValue less than the significance level of 0.05 (5.7169e-67 <0.05), then we can conclude that the coefficient MA {2} of the moving average is statistically significant and should be used.

AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) take into account the degree of model fit, not some trade-off between model accuracy and complexity. In this case, the model fitting procedure is based on finding parameters that minimize the AIC and BIC, which can help to reduce the fit in complex models. This procedure for finding parameters was proposed by Michael Halls-Moore. AIC was proposed by Akaike in 1974 and BIC by Schwartz in 1978. The AIC is based on a generalization of the maximum likelihood principle and, as a result, assumes that the random disturbance is Gaussian.

Analysis of the structure of the actually measured time series and its forecasting show the complexity of the structure of the series, but still there is a possibility of its statistical analysis using the ARIMA method (0,1,2).

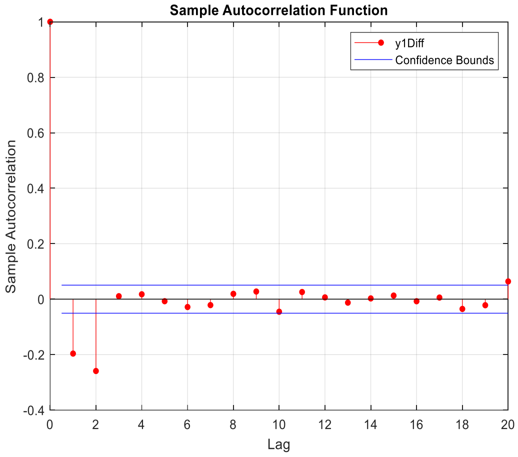


Fig. 8. Sample autocorrelation function Diff

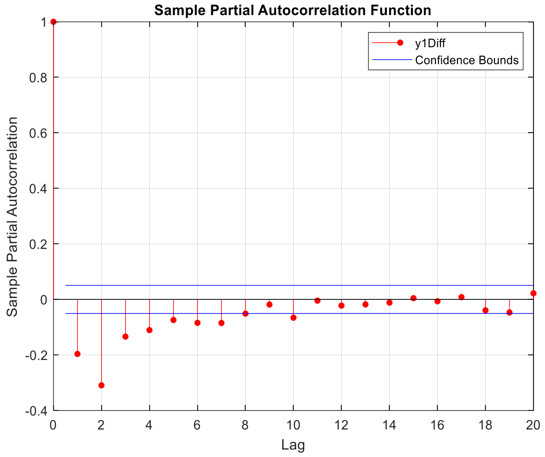


Fig. 9. Sample partial autocorrelation function Diff

The ARIMA method has a very clear mathematical and statistical basis, which makes it one of the most scientifically based models of the entire set of models for forecasting trends in time series.

The only gap is that there are few works devoted to this problem in our country. To date, several articles have been published in Kazakhstan in the field of forecasting tuberculosis and in the field of economics.

Analyzing the ACF and PACF plots of the first-order difference (Figs. 8 and 9), it can be said that according to the autoregression (AR) process, the levels of the ACF series fade out quickly, and the PACF levels gradually fade out. If there was an autoregressive process, the ACF would fade out slowly. For the moving average (MA) process, ACF decays sharply after two lags (the last significant lag is shown by parameter q), and the PACF function gradually fades.

The result is an integrated model with parameter q = 2 – ARIMA (0,1,2) Model (Gaussian Distribution) (ARIMA_y12)

Autoregressive integrated moving average model of time series y1 with the following equation:

$$\left(1-L\right)y_t=\left(1+\theta_1 L+\theta_2 L^2\right)\varepsilon_t \tag{6}$$

Model estimation is shown in Tables 3 and 4.

A combined graph of the initial series and forecast data is shown in Figure 10 and shows that the model is correctly selected. Figure 11 shows a graph of balances.

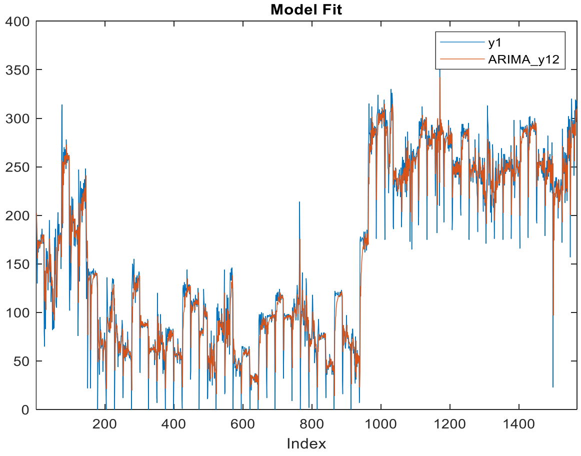


Fig. 10. Plot the fit of model ARIMA_y12 time series y1

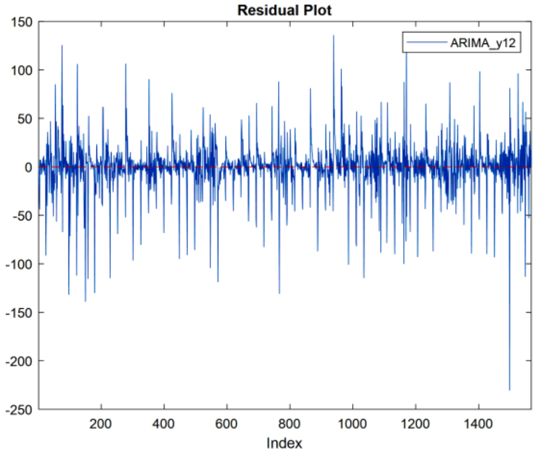


Fig. 11. Plot of the residuals of model ARIMA

TABLE III
ESTIMATION RESULTS

Parameter	Value	Standard Error	TStatistic	PValue
Constant	0	0		
MA{1}	-0,3452	0,016298	-21,1803	1,4509e-99
MA{2}	-0,32033	0,018528	-17,2888	5,7169e-67
Variance	698,8574	10,5872	66,0097	0

TABLE IV
ESTIMATION RESULTS

Parameter	Value
Akaike Information Criterion	14715,9358
Bayesian Information Criterion	14732,0047

The remains of the series have a normal distribution with an average value close to zero. Given all of the above, we can say that the resulting ARIMA forecast model (taking into account confidence intervals) is adequate [10].

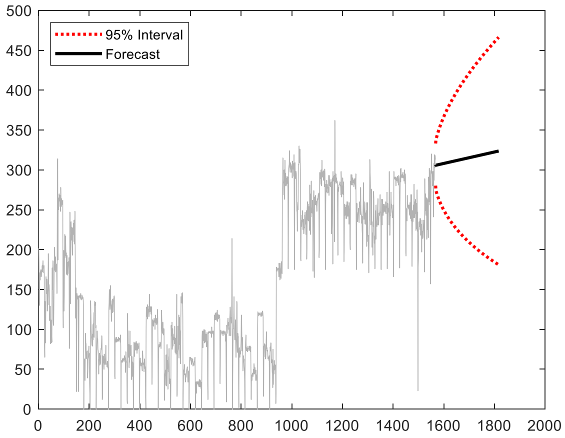


Fig. 12. Forecast of the number of packages on the ARIMA model(0,1,2)

CONCLUSIONS

This study of measured real traffic will be a common contribution to the development of telecommunications science in the Republic of Kazakhstan. This material can be useful in the field of radio engineering electronics and telecommunications. Since 2007, a new generation NGN network (Next Generation Network) based on IP (Internet

Protocol) has been operating in the Republic of Kazakhstan. As for the present work, it is based on the classical ARIMA method for predicting future values of empirical data in the field of telecommunications. The purpose of this work is to predict the subsequent levels of the series for managing information flows in the network in order to avoid congestion and losses.

Visual analysis of the investigated time series showed that it has an uneven intensity with pulsation, has groupings in "packs" in some places or has discharged areas in other time intervals, where there are no or few incoming packets.

The scatter plot of the time series demonstrates the scatter of points in the form of an elongated cloud, limited mainly by one quadrant, which indicates the correlation of points and the assumption of non-stationarity of the series. Non-stationary time series are characterized by the presence of a trend, systematic changes in variance, changing interdependencies between the elements of the time series. As for the scattering diagram of the increments of this series, the points in it are distributed in all four quadrants with a relatively high density of points near zero, which means the relative independence of neighboring values with the absence of a strong linear correlation between the levels of the series of increments.

The original time series was examined by two unit root tests: ADF (Augmented Dickey-Fuller test) and KPSS (Kwiatkowski - Phillips - Schmidt - Shin), which confirmed the involvement of the original series in the class of differential stationary series DS (Difference stationary). The test result confirmed that the series under study is not stationary.

To predict the time series, the ARIMA method (autoregressive integrated moving average) was used, which allows you to make reliable short-term forecasts with a minimum number of parameters. The essence of this method lies in the fact that linear methods can be used to reduce a non-stationary series to a stationary one.

Flexible mathematical and statistical and at the same time relatively simple ARIMA model made it possible to select a prediction model for current data based on previous values, which is most suitable for measured data in a real network of packet rates.

A forecast model for the ARIMA (0,1,2) time series has been created, in which the PACF (partial autocorrelation function) has a sinusoidal shape (decays exponentially), and the ACF (autocorrelation function) differs significantly from zero for the lag q = 2. Results of predicting network traffic showed the adequacy of the chosen model.

Analysis of works related to the ARIMA method shows that the number of works on predicting empirical data in the field of telecommunications is very small (this work is one of the first in Kazakhstan) and there are several in the CIS countries. Basically, the works describe either the analysis of mathematical models ARIMA, or recommendations for the implementation of forecasting. The perspective of this research is the experience of researching real data taken on a real network. This study revealed the possibility of predicting non-stationary time series by a statistical method, which can be useful for university students. Moreover, using this method with a minimum number of parameters, a model for predicting a nonlinear series by linear methods was chosen.

This study will help the researcher to reveal that the structure of real data transmitted in the telecommunication network is complex and the ADF test is not enough to assess the lack of stationarity, but it needs to be confirmed using another unit root test. There are not many researchers looking at real data. Thus, a new experience has been obtained in studying real time series in the Republic of Kazakhstan. Important statement: In this research work, tests for stationarity of a unit root were applied for the first time and the ARIMA method was applied for the first time in the field of telecommunications in the Republic of Kazakhstan, and it is important in comparison with the existing literature in the Republic of Kazakhstan and is a completed prototype for society as a whole.The original time series was examined by two unit root tests: ADF and KPSS. The results of the tests confirmed that the test series is not stationary. The ARIMA time series forecast model(0,1,2) was created.

Analysis of the obtained results of network traffic forecasting shows that the obtained forecast values are close to the original statistical data.

References

[1] V. S. Maraev, “Time series visualization Tools in space research. Volume 1”, *Research of science city*, vol. 4, no. 22, 2017

[2] G.G. Kantorovich, “Analysis of temporal rows. Lecture and methodical materials”, *Economic Journal of the Higher School of Economics*, no. 3, 2002, pp. 379-701.

[3] M. S. Vershinina, “Analysis of assumptions about the stationarity of some temporal series”, Collection of the all-Russian conference on mathematics with international participation "IAC-2018", Barnaul: AltSU University, 2018, pp. 172-176.

[4] R. M. De Jong, C. Amsler, and P. Schmidt, “A robust version of the KPSS test, based on indicators”, J. Econometrics, vol. 137, no. 2, 2007, pp. 311–333.

[5] W. Wojcik, T. Bieganski, A, Kotyra, and A, Smolarz, "Application of forcasting algorithms in the optical fiber coal dust burner monitoring system", *Proc. SPIE* 3189, Technology and Applications of Light Guides, (5 August 1997); DOI: https://doi.org/10.1117/12.285618

[6] K. O. Kizbikenov, “Prognostication and temporary series: textbook by K. O. Kizbikenov”, Barnaul:AltSPU, 2017.

[7] V. S. Korolyuk, N. I. Portenko, A. V. Skorokhod, A. F. Turbin (eds.) “Handbook of probability theory and mathematical statistics”, Moscow: Nauka, 2005.

[8] G. Box, G. Jenkins, “Time Series Analysis: Forecasting and Control,” San Francisco: Holden-Day, 1970.

[9] I Rizkya, K Syahputri, R. M.Sari, I. Siregar and J. Utaminingrum, “Autoregressive Integrated Moving Average (ARIMA) Model of Forecast Demand in Distribution Centre,” *Department of Industrial Engineering, Faculty of Engineering, Universitas Sumatera Utara* in IOP Conf. Series: Materials Science and Engineering 598, 2019, 012071.

[10] N.Albanbay, B.Medetov, M. A. Zaks, “Statistics of Lifetimes for Transient Bursting States in Coupled Noisy Excitable Systems,” *Journal of Computational and Nonlinear Dynamics*. vol. 15, no. 12, 2020, DOI: https://doi.org/10.1115/1.4047867