ABSTRACT
thesis of Ibrayeva Zhanar Bazarbekovna on the topic: «Development of models for network traffic analysis and forecasting», submitted for the degree of Doctor of Philosophy (PhD) in the specialty 6D070400 - «Computer Systems and Software Engineering»

Relevance of the research topic. A multiservice network has been operating in Kazakhstan since 2007. This network is a new generation network NGN (Next Generation Network), which is based on IP protocol with packet switching. Changing the network infrastructure with TDM (Time Division Multiplexing) technology with circuit switching to packet IP (Internet Protocol) has created a modern infrastructure in the field of ICT (info communication technologies) with the provision of services - VoIP (Voice over Internet Protocol), IP VPN (Virtual private network), IPTV (Internet protocol television) and others. The ever-increasing volume of transmitted information creates a certain complexity for the backbone data transmission network in its processing. On the other hand, modern society requires high speeds of processed information transfer.

A modern heterogeneous network generates network traffic with a complex (heterogeneous) structure. A study of measured data shows that they do not have a uniform intensity of packets arriving at serving network devices.

Consequently, there is not only an increase in traffic volumes, but also a change in its structure, so the analysis of network traffic is still an urgent task. To identify and quantify the components of a complex structure - the presence / absence of a trend, periodicity, a random component is the main task of time series analysis.

The growth in the volume of heterogeneous traffic in info communication networks actualizes the issues of ensuring the quality of communication services provided, which in turn requires an appeal to forecasting models.

Modern research has shown that the analysis and prediction of network traffic remains the most important task in traffic management.

Predictive data provide the necessary information to solve the problem of managing information flows in the network and will allow, based on management, to prevent packet loss.

The purpose of the work is the development of models for analysis and prediction of measured network traffic.

To achieve this goal, it was necessary to solve the following tasks:

1. Study of the structure of the time series of empirical data;
2. Conducting experimental studies with classical models of time series forecasting;
3. Development of models for predicting network traffic, considering its heterogeneity.

The object of the research is a time series that displays a set of MPEG protocol packets transmitted over the backbone network for five hours per second (18000 points).

Scientific novelty. Scientific novelty consists in developing and obtaining the following conclusions:

1. Based on the analysis of the structure of actually measured network traffic, a program has been developed that checks the series for stationarity;
2. The parameters of the ARIMA model are determined and the adequacy of the ARIMA(0,2,1) model is proved;
3. A network traffic prediction model based on the NARX (Nonlinear AutoRegressive Network with exogenous inputs) ANN has been developed;
4. Programmatically implemented fuzzy logic models.

Theoretical and practical significance of the work. The theoretical significance of this work lies in the identification of forecasting models that can be used in non-stationary conditions of empirical data. The use of prediction models makes it possible to improve the parameters of the quality of service of the analyzed traffic. The practical significance of the work lies in the development of a network traffic management plan and making the right decisions when managing the proposed forecasting models. The results obtained were tested in LLP "Almaty Institute of Technologies".

Scope and structure of work. includes an introduction, 4 sections, a list of references and appendices.

In the introduction, the substantiation of the relevance of the chosen topic of the dissertation work is given. The purpose, object, subject and tasks of the research work are formulated. The results of the conducted studies are described, their scientific novelty and practical significance are shown.

In the first section of the dissertation, an overview and analysis of the network traffic in a multiservice network were conducted, including the key features of the telecommunications network in Kazakhstan. A review of the main scientific works on network traffic forecasting was also provided. The general characteristics of the problem were presented, and the research tasks were formulated. The classification of stochastic processes, parameters for determining the stationarity and non-stationarity of time series data, and forecasting methods were defined.

A time series is considered non-stationary if its characteristics, such as mean, variance, and autocorrelation function (ACF), depend on time. If the mean of the series shows a linear dependence on time, it indicates the presence of a linear trend in the time series. If the variance of the time series changes over time or exhibits heteroscedasticity (non-uniformity), and if the ACF varies cyclically, it suggests the presence of periodic components in the time series. A non-stationary series always has a trend, which is driven by non-random factors in the processes. Non-stationarity, in a broad sense, means that the correlation function of the series, at a fixed lag or the first moment of the series, or both, change over time. Non-stationarity, in a narrow sense, refers to the variability of the distribution function over time.

Non-stationary series exhibit the following characteristics:

- In the long term, the levels of the series cluster around different mean values.

- The variance of the time series changes from period to period, meaning it is time-dependent.
- The autocorrelation function decreases very slowly.

There are several methods for identifying stationarity:

1. Visual inspection by plotting the time series graphically and checking for the presence of a trend.
2. Examining the presence of autocorrelation, which describes the linear dependence and assumes stationarity of the process.
3. Using tests for trend presence.

It is important to note that checking for stationarity is a crucial step in time series analysis, as many statistical methods and models require the time series to be stationary. If the series is found to be non-stationary, appropriate data transformations such as differencing or seasonal adjustment are necessary to obtain a stationary series.

In practice, hypothesis testing criteria are used to verify the stationarity of a series, ensuring the acceptance of the true hypothesis and the rejection of the false hypothesis with a high probability.

The following conclusions were drawn from the first section:

1. Time series can be either stationary or non-stationary. Parametric and non-parametric criteria are used in practice to test the hypothesis of stationarity.
2. The statistical properties of stationary and non-stationary time series differ significantly, requiring the application of different methods for modeling.
3. A model for stationary time series is characterized by time-invariant mean, variance, and autocorrelation.
4. A model for non-stationary time series is multi-component, comprising trend, seasonal, and random components.
5. Non-stationary series can be divided into series with a deterministic trend (TS) and series with a stochastic trend (DS).
6. Achieving stationarity can be accomplished through operations such as trend extraction, seasonal component removal, or integration characterized by the order of successive differences.
7. The Dickey-Fuller test is used to test for DS-series or TS-series.

In the second section, an analysis of numerical characteristics of the measured traffic was conducted. The raw data was collected over a period of five hours, capturing the transmission of packets of the MPEG protocol over the backbone network.

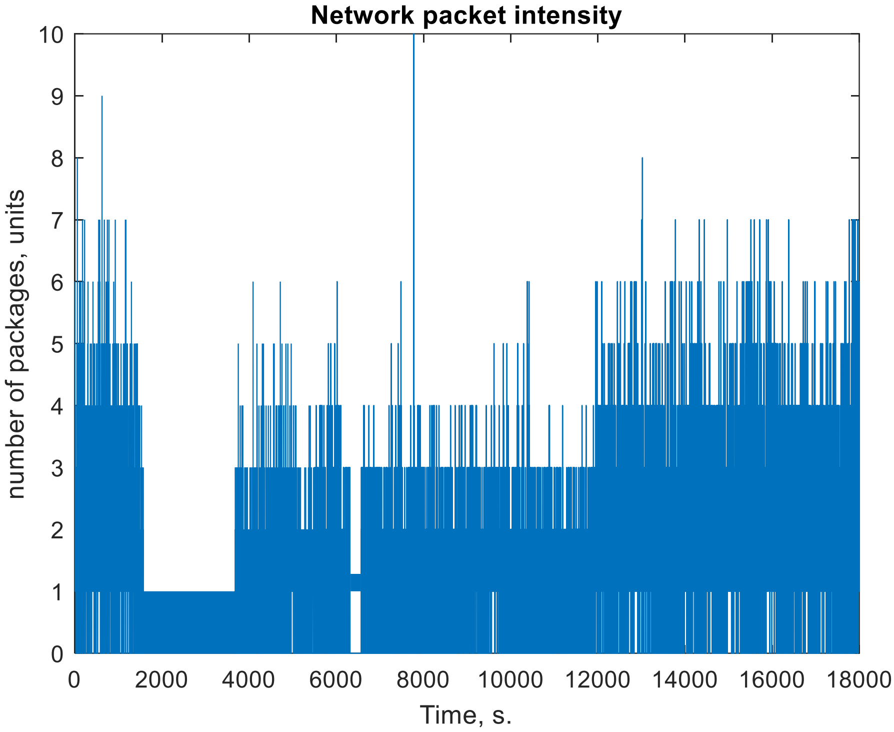

Fig. 1 Transmission Packet Intensity Series.

Statistical estimation of the series for normality was conducted, as it plays an important role in traffic forecasting when using appropriate methods. The hypothesis of the normality of the distribution for the analyzed series was tested, and the following criteria were rejected: modified Kolmogorov-Smirnov criteria, Cramer-Mises, Anderson-Darling, Shapiro-Francia, skewness coefficient, kurtosis, Jarque-Bera, and Giri-D'Agostino. Non-parametric tests indicated that the analyzed series exhibits a trend. As a result of the study, a program was developed in the Python programming environment to test the series for stationarity. In the Matlab programming environment, the original series was examined for unit roots using the Phillips-Perron (PP) test and the Dickey-Fuller test.

The following conclusions were drawn from the second section:

1. The distribution pattern of the analyzed series does not conform to the assumption of normal distribution.
2. Using the developed program and employing non-parametric tests and correlograms, it was identified that the analyzed series exhibits a trend.
3. The investigation using ADF-test, PP-test, and KPSS-test revealed the presence of unit roots.

The third section is dedicated to the frequency-time analysis of time series. The analyzed series was subjected to Singular Spectrum Analysis (SSA), which is a contemporary tool for analyzing the structural components of a time series. The SSA decomposition method breaks down the time series into a set of summable components that are grouped and interpreted as trend, periodicity, and noise. This method emphasizes the separability of the underlying components and can easily distinguish periodicities occurring at different time scales, even in highly noisy time series data.

Furthermore, the series was examined for harmonic components and decomposed into a low-frequency component (trend) and periodic components along with high-frequency noise. The decomposition of the time series reveals that it is non-stationary and consists of a trend, harmonic components, and noise.

In the fourth section, the characteristics of forecasting non-stationary time series are described. Among the statistical forecasting approaches, the ARIMA (Auto-Regressive Integrated Moving Average) method allows for modeling non-stationary time series. If the data exhibits unit root characteristics, the possibility of using an ARIMA model can be considered. The tests used in the second section confirmed the presence of a unit root.

The first step in model construction is differencing the data until they appear stationary. Through the selection of numerous model parameters and visual analysis, an integrated ARIMA(0,2,1) model was obtained.

The Auto-Regressive Integrated Moving Average model for the time series can be represented by the following equation:

$$(1 - L)^2 y_t = c+(1+\Theta_1 L)\varepsilon_t \qquad (1)$$

The equation describes an ARIMA model in the form of a difference equation, where y_t represents the current value of the time series, L is the lag operator (i.e., a shift by one time period), c is a constant, ε_t is the random error at time t, and Θ_1 is the parameter of the first-order moving average. $(1 - L)^2$ represents the second-order difference operator, which is applied to the value of the time series y_t.

This equation describes the relationship between the current value of the time series, its past values, random errors, and model parameters. It can be used to forecast future values of the time series based on its past values and model parameters.

Among the AI-based methods, the NARX (Nonlinear AutoRegressive exogenous Network) model has been chosen in this study, which is well-suited for modeling nonlinear systems and fuzzy logic algorithms.

To perform the forecasting of the nonstationary time series, the Neural Network Toolbox (NNT) software package from MathWorks is selected, which operates under the Matlab system.

NARX (Nonlinear Autoregressive with exogenous inputs) is a recurrent neural network that can be used for forecasting network traffic. This network belongs to the class of multilayer perceptrons with feedback, which have input, hidden, and output layers.

For forecasting network traffic, NARX utilizes historical data on traffic and other related variables such as time of day, day of the week, holidays, etc. These data are preprocessed, normalized, and then used for training the network.

Training NARX involves adjusting the weights between the nodes of the input, hidden, and output layers. During the training process, the network gradually optimizes these weights to minimize the forecasting error.

After completing the training, the network can be used to forecast network traffic based on new data on time and other related variables.

In the NARX model with external inputs, the original time series is fed into the neural network, processed according to the weight coefficients, and then the output data are fed back into the network, thus replacing the backpropagation procedure. This allows for considering the obtained weight coefficients from the initial training during the retraining of the network, which in turn improves its accuracy.

In the field of digital signal processing, dynamic neurons described by difference or differential equations have gained wide popularity. One of the simplest dynamic neurons is the Hopfield neuron. The NARX network uses a modified Hopfield neuron whose state is determined by a more distant history.

The dynamics of the NARX model can be described as follows:

$$y(n+1)=F(y(n),...,y(n-q+1),u(n),...,u(n-q+1)) \tag{2}$$

where F – a nonlinear function, which approximates of its arguments during the training process.

q – delay.

The NARX network is a multilayer network with feedforward and feedback connections, where the output data is passed through a time-delay vector. The sigmoid function is used as the activation function. Each layer of the network transforms the input feature space into another space, possibly with a different dimension. This nonlinear transformation occurs until the classes become linearly separable by the neurons of the output layer. All layers of the neural network, except the input and output layers, provide the network with the ability to model nonlinear phenomena. The Levenberg-Marquardt algorithm was used for training the neural network, and the mean squared error (MSE) is used to evaluate its performance.

The capabilities of neural networks can be enhanced through information processing technologies based on fuzzy sets and fuzzy inference. In this study, neuro-fuzzy forecasting algorithms such as Chen and Cheng algorithms, as well as fuzzy clustering algorithms, were investigated. The Chen method showed low accuracy in the predicted data, with a high numerical value of MSE, 9.078161087779954. On the other hand, the Cheng fuzzy forecasting method had higher performance, indicated by a lower numerical value of MSE=1.2359176594533703. This result suggests that the Cheng fuzzy time series method is sufficiently good for forecasting time series. The differences between these methods lie in the stages of forming fuzzy sets, and each group of fuzzy relationships has weights.

In conclusion, the main results of the work, the conclusions of the dissertation and future steps in the study of this direction are outlined.

Approbation of work. The main results on the topic of the dissertation are presented in 15 papers, 7 articles in journals recommended by the Committee for Quality Assurance in the Sphereof Education of the Ministry of Science and Higher Education of the Republic of Kazakhstan, 1 article in an international scientific publication included in the Scopus database and 7 materials of international foreign conferences:

Scientific publications: The results obtained on the topic of the dissertation are presented in the following publications:

1. Serikov T., Zhetpisbayeva A., Mirzakulova S., Zhetpisbayev K., **Ibraeva Zh.**, Tolegenova A., Soboleva L., Zhumazhanov B. (2021). Application of the NARX neural network for predicting a one-dimensional time series. Eastern-European Journal of Enterprise Technologies. Vol. 5 №4 (113) Pp.12-19. Scopus: 56%, WoS: Q2
2. Г.У.Бектемысова, **Ж.Б.Ибраева**, А.Е.Кулакаева, Б.А.Кожахметова. Анализ измеренного сетевого трафика на стационарность. Вестник КазАТК, Вестник Казахской Академии Транспорта и Коммуникации им. М.Тынышпаева, ISSN 2790-5802, -№ 3 (122), -2022г. -С.: 302-308
3. G. Bektemyssova, Abdul R., Sh.Mirzakulova, **Zh.Ibraeva**. Time series forecasting by the Arima method. Scientific Journal of Astana IT University, ISSN (P): 2707-9031 ISSN (E): 2707-904X, Volume 11, September 2022, -P.: 14-23
4. Г.У.Бектемысова, **Ж.Б.Ибраева**. Исследование временного ряда на стационарность. Вестник НИА РК, Вестник Национальной инженерной академии РК, №4(86), 2022, -С.: 20-27
5. G.U. Bektemyssova, **Zh.B. Ibraeva**, Abd Rakhim Akhmad. Fuzzy model for time series forecasting. Scientific Journal of Astana IT University.Scientific Journal of Astana IT University. ISSN (P): 2707-9031 ISSN (E): 2707-904X VOLUME 13, MARCH 2023, -P.: 93-102
6. Бектемысова Г.У., **Ибраева Ж.Б.** Возможности применения искусственного интеллекта в строительстве. Вестник КазГАСА. – 2018. – Том 68, выпуск 2. – с. 205-212.
7. Бектемысова Г.У., **Ибраева Ж.Б.,** Луганская С.П., Миркасимова Т.Ш. инструментов MATLAB для анализа больших данных по энергоэффективности зданий. Вестник КБТУ. – 2019. – Том 16, выпуск 3. – с. 324-328
8. Аймагамбетова З.Т., **Ибраева Ж.Б.** Аспекты формирования комфортной городской среды, Вестник КазГАСА. – 2021. – Том 81, выпуск 3. с. 15-20
9. Aimagambetova Z.T, Bektemyssova G.U, **Ibraeva Zh.B.** Buildings energy consumption modeling methods. STCCE-2020 IOP Conf. Series: Materials Science and Engineering (Scopus). 2020. Volume 890, Kazan, Russian Federation doi:10.1088/1757-899X/890/1/012144, p.14-23
10. **Ибраева Ж.Б.,** Айтжанов Д., Каттабек А. Анализ и оптимизация сетевого трафика. Международный журнал информационных и коммуникационных технологий, Спец выпуск, май, 2022, С.:166-170

11. **Ибраева Ж.Б.,** Мирзакулова Ш.А. Analysis of a one-dimensional time series for a trend. Материалы международной научной конференции молодых ученых, IMA-2022, Суми-Нур-Султан, апрель, 2022, P.:258-259
12. **Ибраева Ж.Б.,** Мирзакулова Ш.А. Network traffic analysis using Leybourne-Mccabe test. Материалы международной научной конференции молодых ученых, IMA-2022, Суми-Нур-Султан, апрель, 2022, P.:259-261
13. **Ибраева Ж.Б.,** Миркасимова Т.Ш. Мәліметтерді басқарудың ертеңі мен бүгіні. Сборник материалов международной научно-методической конференции. Современные концепции науки и образования. Алматы, 2017. стр. 56-59
14. Sh.Mirzakulova, **Zh.Ibraeva.** Clustering Time Series Data. Материалы международной научной конференции молодых ученых, IMA-2022, Суми-Нур-Султан, апрель, 2023, P.:398-399
15. Доскен Б., **Ибраева Ж.Б.** Компьютерное моделирование усилителя сигналов. Материалы международной научной конференции молодых ученых, IMA-2022, Суми-Нур-Султан, апрель, 2023, P.:400-401

Authorship certificates:

1. Разработка модели прогнозирования с использованием статистического метода Auto-Regressive Integrated Moving Average.
   № 32481 от «9» февраля 2023 года.
2. Нечеткая модель прогнозирования временного ряда.
   № 35224 от «27» апреля 2023 года