УДК 330.43
JEL classification: C44

Zhukovska O. A.
Cand. physics and mathematics, associate professor
ORCID ID: 0000-0003-1110-9696

Shperlov R. V.
ORCID ID: 0000-0003-2430-1011

National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"

FORECASTING HIGH-LOAD SYSTEMS

ПРОГНОЗУВАННЯ НАВАНТАЖЕННЯ HIGH-LOAD СИСТЕМ

In this article, analytical and prognostic model of the information system load was performed, which in metric terms is based on the number of requests to the system per second. The studied statistical data of this indicator outline the main trajectory of its time series, which allows to find patterns of the past and predict the further vector of development. The relevance of this study is a priority challenge in the IT industry, because the scale of information systems, even small companies is so cumbersome that the cost of their maintenance is the lion's share of company revenue. Another significant reason for the urgency is the global crisis provoked by the Covid-19 pandemic. It helped to find new methods of selling goods, taking into account social distance and reducing financial losses. In this regard, most product companies are moving their business to the virtual world, creating a great demand for the formation of their own IT - infrastructures .The presence of predicted values provides business confidence in the stability of their product or technology, which directly affects the cost of implemented software code and future revenue. Timely detection of peak loads or, conversely, subsidence in the system, helps to prevent the principle of "house of cards" and protect dependent family processes in the company. Long-term vision of the time series provides the company with the opportunity to prepare for infrastructure changes and provides a certain period of time to update the technical base or expand the capabilities of their " Cloud Services ". For this purpose, based on statistics built prognosis based integrated moving average or ARIMA - models performed statistically identify the trend by analyzing the autocorrelation function. It was decided to group the data according to the time of day to increase the forecast periods and increase the accuracy of the model forecast. The analysis of autocorrelation and partial autocorrelation consistent differences to identify AR - or MA - processes. Several candidate models for each grouped series were built on the basis of the analysis of the significance of the last nonzero lags. Each model was checked the similarity residues "white noise" and normal distribution coefficients based on using the Q-statistics Ljung-Box. Consolidated forecasts were compared and evaluated to identify the lowest percentage error. The cost of savings for the forecast period was estimated. Based on these data, it was concluded that the importance of conducting research data using such methods of forecasting and data analysis for different vectors of IT companies.

Keywords: information system, high load, ARIMA-model, forecasting, lag.

У статті проведено аналітико-прогностичне моделювання завантаженості інформаційної системи, що в метричному виразі базується на кількості запитів до системи в секунду. Досліджені статистичні дані показника кількості запитів до системи дозволяють знаходити закономірності минулого та прогнозувати подальший вектор розвитку. Актуальність даного дослідження є пріоритетним напрямом в IT-індустрії, адже масштабність інформаційних систем навіть малих компаній є настільки громісткою, що затрати на їх утримання становлять велику долю доходу компаній. Зазначено, що однією значною причиною актуальності є світова криза, спровокована пандемією Covid-19. Саме вона посприяла пошуку нових методів збуту товарів з урахуванням соціальної дистанції та зменшення фінансових втрат. В зв'язку із цим більшість продуктових компаній переводять свій бізнес у віртуальний світ, створюючи неабиякий попит формування власних IT-інфраструктур. Наявність прогнозованих значень забезпечує впевненість бізнесу в стабільності їх продукту чи технології, що напряму впливає на собівартість реалізованого програмного коду та майбутній отриманий дохід. Своєчасність виявлення пікових навантажень або ж навпаки – просідань в системі, допомагають запобігти принципу «карткового домику», та захистити залежні родинні процеси в компанії. Довгострокове бачення руху часового ряду забезпечує компанії можливість підготовки до інфраструктурних змін та надає певний період часу для оновлення технічної бази чи розширення можливостей їх «Cloud Services». З цією метою на основі статистичних даних був побудований прогноз на основі інтегрованого ковзного середнього або ж ARIMA-моделі, проведено статистичне виявлення тренду шляхом аналізу автокореляційних функцій. Було прийнято рішення щодо групування даних відповідно пори доби для збільшення прогнозованих періодів та більшої точності прогнозу моделі. Проаналізовано автокореляційну та часткову автокореляційну функції для часового ряду послідовних різниць для виявлення AR- та MA-процесів. На основі аналізу значущості останніх ненульових лагів побудовано декілька кандидатних моделей для кожного групованого ряду. Кожна модель була провірена на подібність залишків «білому шуму» та нормальності розподілу коефіцієнтів. Зведені прогнози були порівняні та оцінені для виявлення найменшої відсоткової похибки. Було оцінено вартість заощаджень для прогнозованого періоду. На основі цих даних був зроблений висновок, щодо важливості проведення даних досліджень із використанням таких методів прогнозу та аналізу даних для різних векторів діяльності IT-компанії.

Ключові слова: інформаційна система, високе навантаження, ARIMA-модель, прогнозування, лаг.

Introduction. Productivity of information systems is the most important parameter of the quality of the company's IT department. Most of the largest IT giants state a non-linear increase in the volume of processed requests compared to the growth of the customer base. The dominant role in the formation of the rapid growth plays and is swift digitalization any commercial or public already and processes. The global coronary crisis has also accelerated this process, forcing entrepreneurs to distribute their goods through online stores. The pandemic gave a sharp impetus to the development of new methods of selling to various sectors of the economy. The growth curve of loads on information systems is increasingly taking the form of an exponential function, which complicates the construction of highly loaded information systems.

The term "high load" was defined not so long ago. Experts of various areas of information technology concepts outline the following statement [1]: high load - of the system, when the IT system is no longer cope with the load current, which leads to immediate expansion of infrastructure.

The final clarification of this concept allows us to look at the definition of high workload from a new angle: it is a system that is constantly scalable and has enough resources to work with future workloads. Scalability and sufficient resources provide confidence not only for professionals who maintain the information system, but also for businesses for which IP is a tool for earning and which should never be idle or out of order.

Thus, we can logically conclude that to avoid such situations it is necessary to anticipate future workloads to increase sufficient computing resources and be ready for further scaling, which will significantly reduce the cost of maintenance and upgrade of IP hardware components. According to the architecture, internal structural components and functionality of information systems have a certain distribution among the possible numerical metrics that describe the system load. For heavy system such as "challenge-response" basic load indicator system is called QPS (query per second) - the number of requests to the system and a second. This indicator also has an economic meaning because its maintenance is embedded in the cost of technology or product produced by the company.

Thereby, solving the problem of proper management of computing resources can contribute to additional information that can be obtained from the forecast data on the number of requests to the system based on predictive models. A large number of works dedicated to the development of predictive models to data network, such as [2- 4], shows the feasibility of building predictive models for organizing proper control computing resources. A carefully selected model is able to take into account the most important characteristics of the flow of requests, such as short-term and long-term dependent processes, seasonality and cyclicity over long periods of time.

Setting objectives. The purpose of this article is to build prediction term model to determine the load on the information system in order to further optimize of its physical parameters.

Methodology. The Box-Jenkins methodology is used to build the forecast model, which is based on a set of procedures for determining, correcting and verifying ARIMA models for given time series. The forecast follows directly from the equation of the adjusted model. The main tools used in this article are:

– autoregressive model AR (*p*):

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + \varepsilon_t,$$

where

Y_t – response (dependent variable) at the time t;

$Y_{t-1}, Y_{t-2}, ..., Y_{t-p}$ – review at values intervals $t-1$, $t-2$, …, $t-p$ respectively;

$\phi_0, \phi_1, \phi_2, ..., \phi_p$ – estimated coefficients;

ε_t – error describing the influence of variables that are not considered in the model.

– model with moving average named M A (q):

$$Y_t = \mu + \varepsilon_t - \omega_1 \varepsilon_{t-1} - \omega_2 \varepsilon_{t-2} - \ldots - \omega_q \varepsilon_{t-q},$$

where

ε_t – errors in previous periods of time, which are currently t included in Y_t;

μ – constant process average;

$\omega_1, \omega_2, ..., \omega_q$ – estimated coefficients based on sample observations.

– models of autoregressive and moving average ARMA (p, q):

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \varepsilon_t - \omega_1 \varepsilon_{t-1} - \omega_2 \varepsilon_{t-2} - \ldots - \omega_q \varepsilon_{t-q}.$$

For non-stationary time series used autoregression and integrated and a model and a moving average ARIMA (p, d, q).

The methodology for constructing the ARIMA model of the forecast for the studied time series includes the following main stages [5] (Fig. 1):

- indentification of the trial model;
- evaluation of model parameters;
- diagnostic verification of model adequacy.

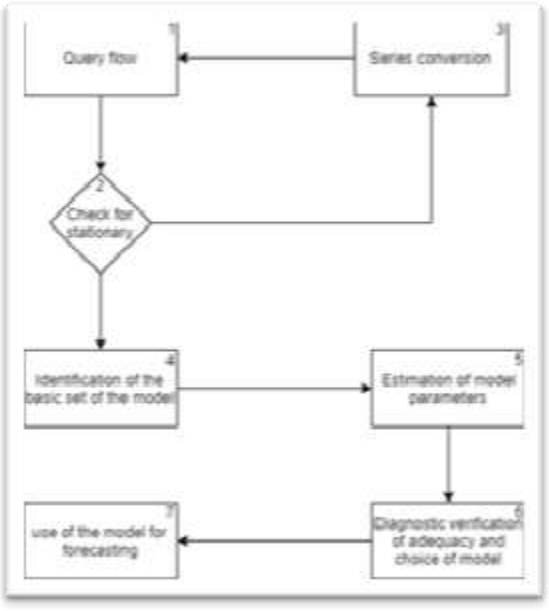


Figure 1 - Block diagram of the selection of the ARIMA model

Results of the research. The study data were used one of the leading companies, Internet providers - advertising in Ukraine. In Figure 2 shows the dynamics of changes

in the value of the aggregate average QPS for the server infrastructure for 3 months of 2020.

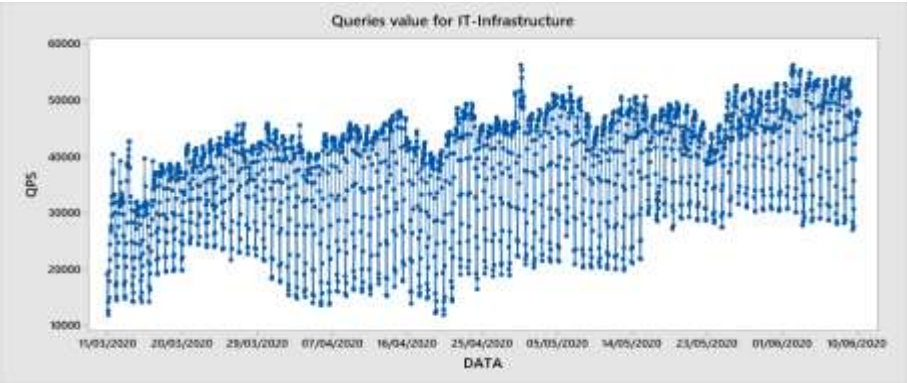


Figure 2 - Graph of the dynamics of QPS

The time series is formed from the round-the-clock hourly average value of the load on the system. One of the most important values in the series is the peak loads that correspond to the upper and lower limits of the series during one day. The values of these boundaries, which formed the next two series, are highlighted for the forecast (Fig. 3). Based on them, the peak loads of the system will be found.

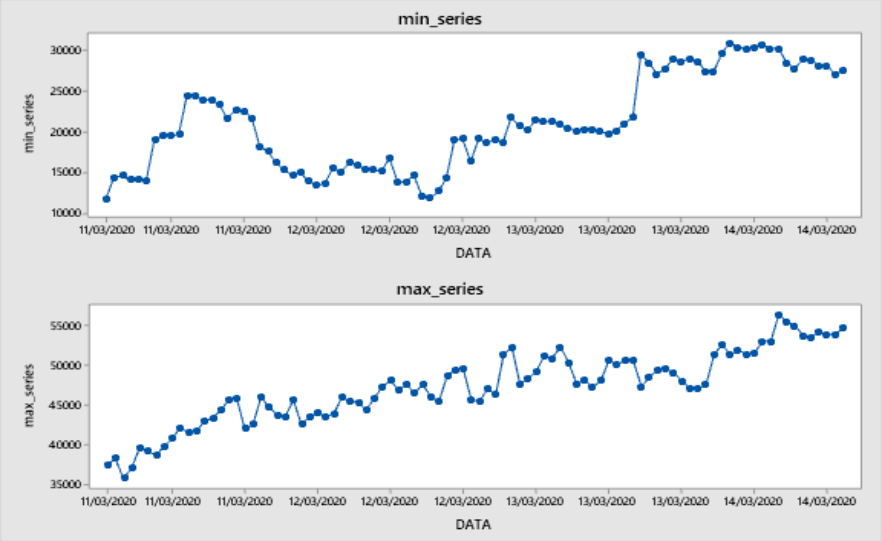


Figure 3 - Upper and lower bound of the time series

From visual analysis chart (Fig. 4) we conclude about attenuation of autocorrelation second function (ACF), indicating the transience of time series and no seasonal component . In this case, to move to a stationary series, it is necessary to take successive differences *d*.

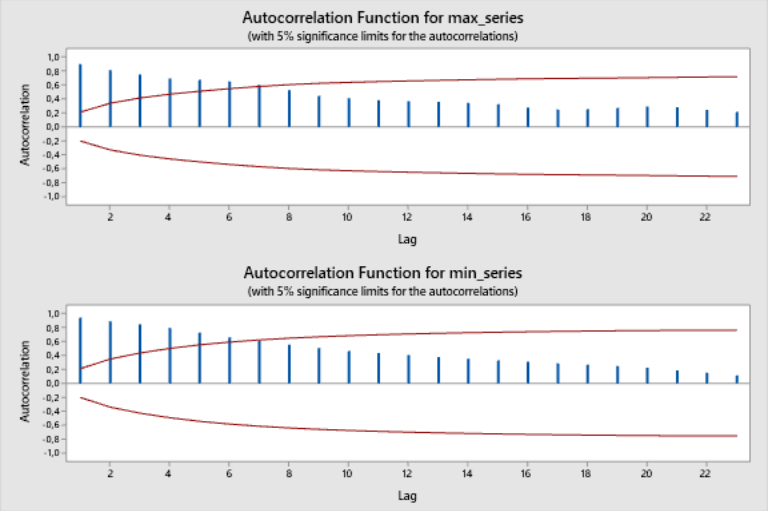


Figure 4 - Autocorrelation functions

Visual analysis of ACF and PACF allows us to conclude whether the time series can be considered a pure AR or MA process, or whether it is a mixed ARMA process. According to the initial data of the time series, the autocorrelation functions slowly fade, and the partial ones fade sharply. In this case, we are talking about the mixed nature of the process of each individual series, where the order of autoregression is determined by the number of the last statistically significant lag PACF, and the order of the moving average – ACF, respectively.

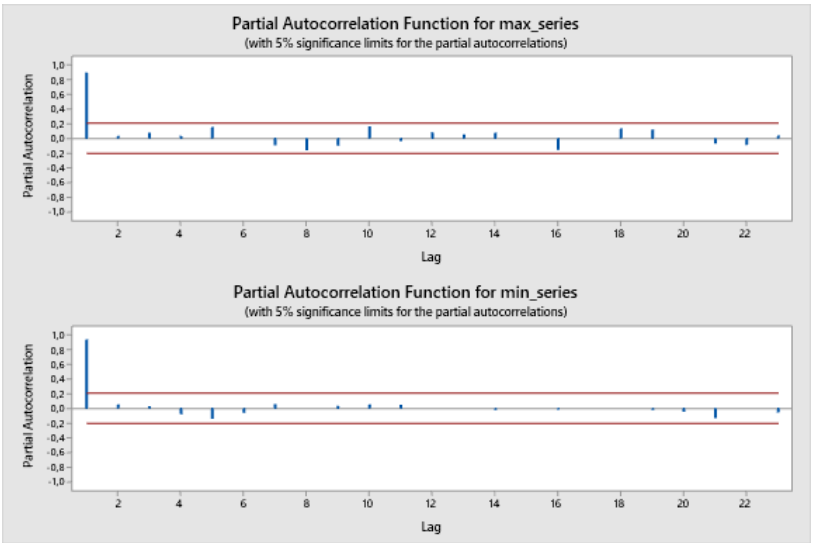
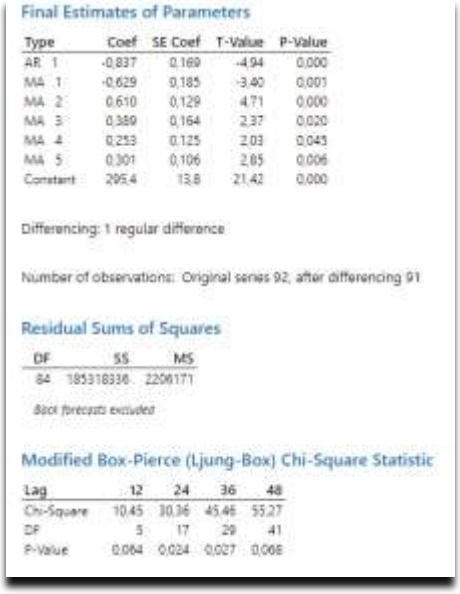


Figure 5 - Partial Autocorrelation functions

Using the Box-Jenkins methodology, we assume that this process will be successfully described by the ARIMA (1,1,5) model for the maximum peak series and the ARIMA (1,1,4) model for the minimum peak series.

The models were built in the Minitab software environment. In Fig.6 we can see a listing of the result of building the ARIMA model (1,1,5).

According to χ^2- statistics of Ljung-Box, the null hypothesis about the randomness of residues is accepted, which indicates a higher level of significance of *p-value*. The autocorrelation of the residuals does not go beyond the confidence intervals (Fig. 7).



Final Estimates of Parameters

Type	Coef	SE Coef	T-Value	P-Value
AR 1	-0,837	0,169	-4,94	0,000
MA 1	-0,629	0,185	-3,40	0,001
MA 2	0,610	0,129	4,71	0,000
MA 3	0,389	0,164	2,37	0,020
MA 4	0,253	0,125	2,03	0,045
MA 5	0,301	0,106	2,85	0,006
Constant	295,4	13,8	21,42	0,000

Differencing: 1 regular difference

Number of observations: Original series 92, after differencing 91

Residual Sums of Squares

DF	SS	MS
84	185318336	2206171

Back forecasts excluded

Modified Box-Pierce (Ljung-Box) Chi-Square Statistic

Lag	12	24	36	48
Chi-Square	10,45	30,36	45,46	55,27
DF	5	17	29	41
P-Value	0,064	0,024	0,027	0,068

Figure 6 - Listing of the result of building a model for the several max. peaks

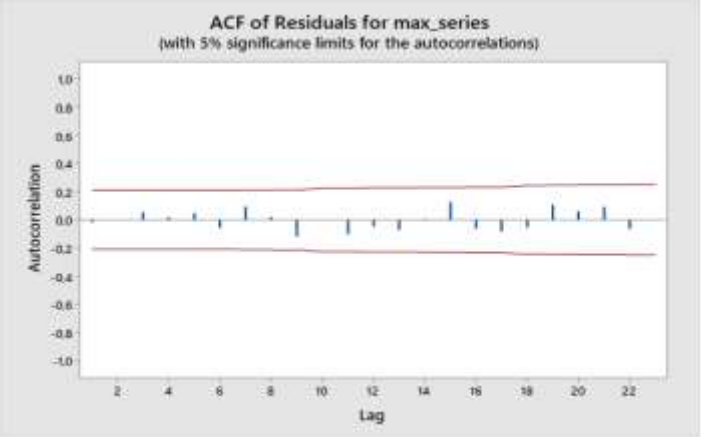


Figure 7 - Autocorrelation function of residues

The adequacy of the predicted model also confirms the normality of the distribution of residues and the location of points in some corridor on the graph of the dependence of residues on the predicted values (Fig. 8).

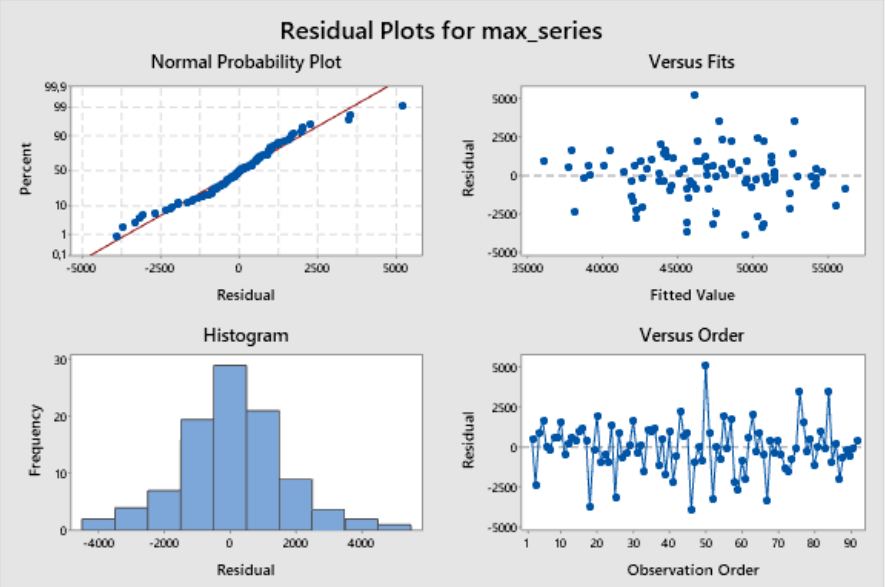
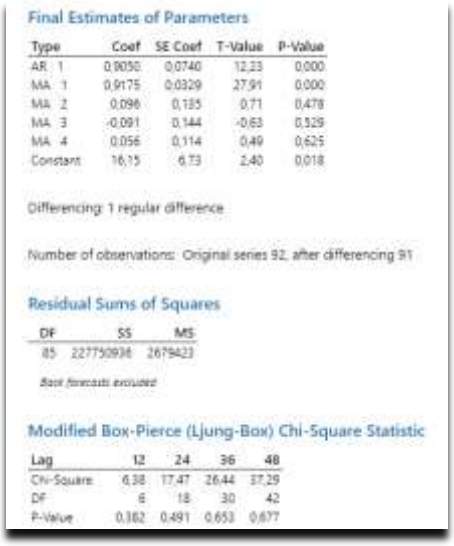


Figure 8 - Distribution of residues

The result of constructing the ARIMA model (1, 1, 4) for several minimum peaks is shown in Fig. 9.



Final Estimates of Parameters

Type	Coef	SE Coef	T-Value	P-Value
AR 1	0,9050	0,0740	12,23	0,000
MA 1	0,9175	0,0329	27,91	0,000
MA 2	0,096	0,135	0,71	0,478
MA 3	-0,091	0,144	-0,63	0,529
MA 4	0,056	0,114	0,49	0,625
Constant	16,15	6,73	2,40	0,018

Differencing: 1 regular difference

Number of observations: Original series 92, after differencing 91

Residual Sums of Squares

DF	SS	MS
85	227750936	2679423

Back forecasts excluded

Modified Box-Pierce (Ljung-Box) Chi-Square Statistic

Lag	12	24	36	48
Chi-Square	6,38	17,47	26,44	37,29
DF	6	18	30	42
P-Value	0,382	0,491	0,653	0,677

Figure 9 - Listing of the result of building a model for several min peak

The analysis of residues (Fig. 10) corresponds to the normal distribution, and the dependence of the correlation coefficients of the residues on the estimated values indicates the absence of time influence and the residues have a horizontal structure. Autocorrelation of model residues (Fig. 11) to χ^2 – statistics of Ljung-Box demonstrate random error behaviour.

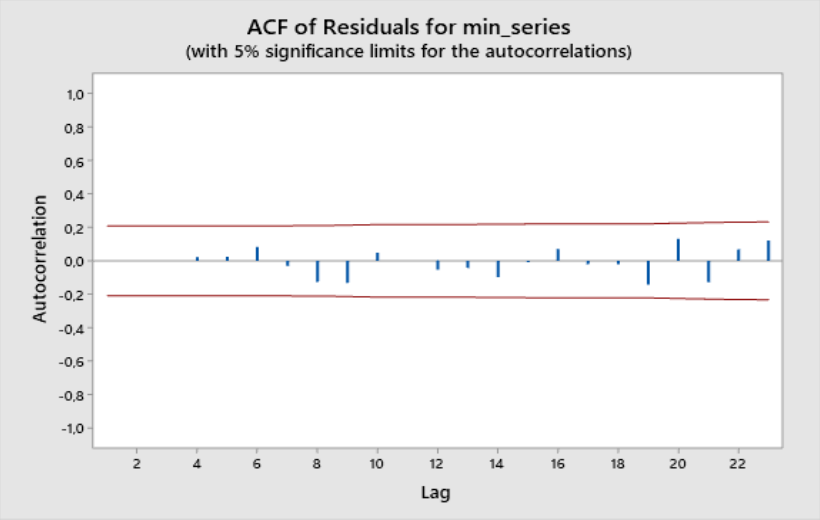


Figure 10 - Autocorrelation function of residues

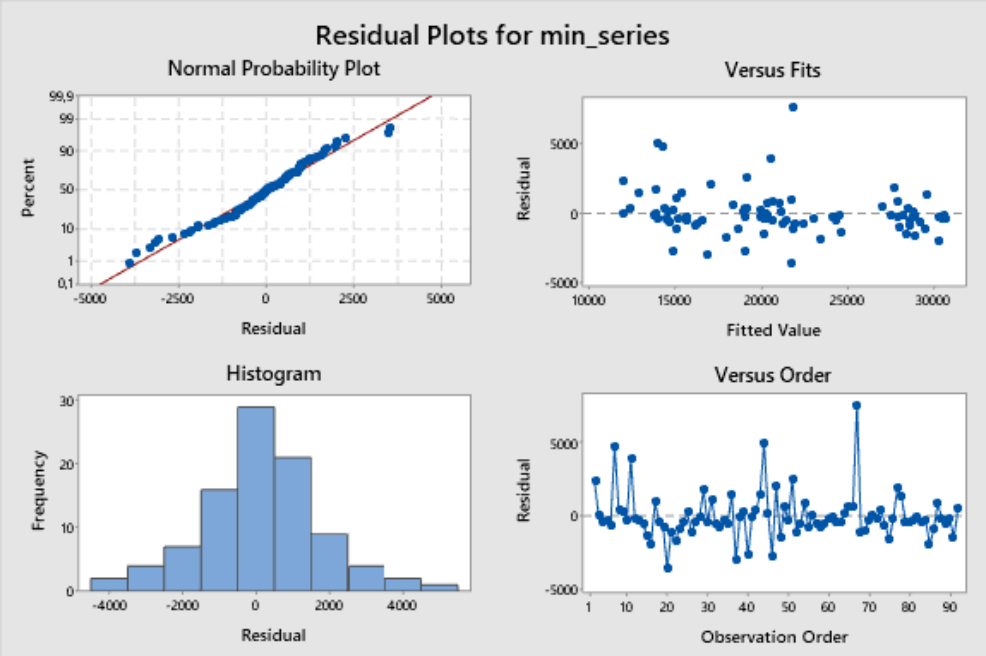


Figure 11 - Distribution of residues

Since the models for time series are adequate, since the residuals are random values, we can build a forecast. As the long-term forecast is economically impractical due to sharp fluctuations of exogenous factors, it was decided to limit the forecast to the next four working days (Fig. 12).

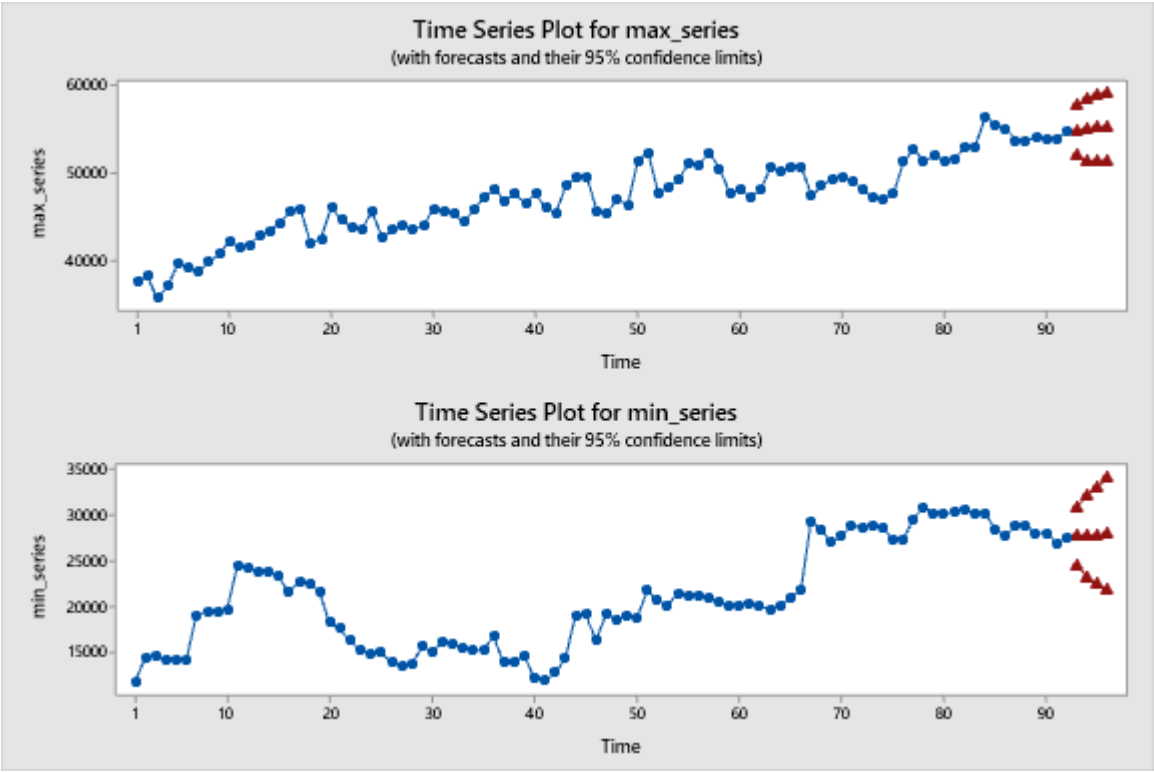


Figure 12 - Predicted values

Conclusions. The constructed forecast outlines the main limits of the technical infrastructure, which allows you to adjust the operation of components with reduced computing power , which in turn affects the final cost of maintenance or lease of computer equipment for the information system . Proper selection of the model for a particular information system will not only save, but also make it impossible to crash systems, which can lead to losses for the company. The main task of this forecast is to prevent daily peaks, for which the parameters of the information system will be configured accordingly. Based on the forecasting results, it is possible to determine the effective number of operating components of the system, which affects the cost of renting host machines. Every ten thousand requests to the system are served by two physical machines. Server rental costs from $ 699 per month relative to configuration, or $ 23.3 per day. By selecting the number of servers that will serve requests during the day, you can reduce their number by half at minimum loads. But you need to take into account the fact that the dynamics of the load is smooth, and for security it will be safer to reduce the number of servers by a third. With full readiness for peak load, 16 servers are running. Accordingly, during minimum load can run 10 servers for 4 hours, which will reduce the cost of rent for 16 servers from 373 dollars to 360 dollars per day, respectively for the two days of savings will amount to $ 26 equivalent to the cost of the server daily.

References:

1. Uchebnik po vysokim nagruzkam [Elektronnyi resurs]. – URL: https://www.dropbox.com/s/k2q0od22h7so751/HL-academy.pdf?dl=0
2. Kryukov A. ARIMA – model prognozirovaniya znacheniy trafika // *Journal "Informatsioniy tekhnologii i vichislitelniy sistemy"*. 2011. S. 42-44.
3. Doronina A. I. Time-series models the example of oil products consumption in France // *Financial University under the Government of the Russian Federation.* 2012. 21 s.
4. Dimas Setyo. Time Series Analysis for Predicting the Total Patient Treated in Blora Health Center Using Minitab Software 14 // *FMIPA UNNES Publisher*. 2011.
5. Dubrova T. A. Statystychni metody prohnozuvannya // *Journal YUNYTY-DANA*. 2003. 206 s.
6. Boks Dzh. Analiz chasovykh ryadiv // *Prohnoz ta upravlinnya*. 1974. 406 p.
7. Lyuys K.D. Metody prohnozuvannya ekonomichnykh pokaznykiv // *Finansy ta statystyka.*1986. 133 s.
8. Hadijah., Forecasting of Operational Reserve with Program Approach Using Minitab Arima // *Journal THE WINNERS.* 2016. *№94.* pp. 13–19.