Statistical analysis of data on the traffic intensity of Internet networks for the different periods of time

Olga Malyeyeva[1[0000-0002-9336-41821]],Yurii Davydovskyi[2[0000-0003-2813-41692]],

Viktor Kosenko[3[0000-0002-4905-85083]].

[1,2]National Aerospace University "Kharkiv Aviation Institute", Chkalov str., 17, Kharkiv,
61070, Ukraine
o.maleyeva@khai.edu, davidovskyi2350@gmail.com
[3]State Enterprise "Kharkiv Scientific-Research Institute of Mechanical Engineering
Technology", Sumy str., 130a, Kharkiv, 61023, Ukraine
kosvv@ukr.net

Abstract. The article is devoted to the problems of the analysis of the regularity of changes in the intensity of telecommunications traffic for the different periods of time. A visual analysis of statistics on the intensity of traffic in different countries is conducted during the day, month, year, and five years. With the use of nonparametric criteria (chi-square, Mann-Whitney, Kruskall –Wallis), a pairwise and general comparison of data was conducted to identify their common trends. In addition, with the use of statistical methods of time series processing, data are analyzed for a month and five years. By constructing an autoregression model of the time series, according to observation data, during the month a periodicity of a series with a seasonal lag has been proved. Thus, the existence of a clear periodicity in daily data has been proved and the frequency of the year is shown. The obtained results can be used to model the network parameters taking into account the predicted traffic.

Keywords: traffic analysis, intensity, periodicity, statistical criteria, hypotheses

1 Introduction

The present time can be characterized by the rapid growth of traffic volumes of computer networks. Such a trend is due to the rapid development of such computer technologies as a cloud computing, a cloud storage, "streaming" services for movies, music or games. However, this growth is not a new one for the Internet providers, according to available statistics, the volume of traffic over the past 10 years has increased 8-10 times [1], and the number of people connected to the Internet as a percentage increased from 23% to 55% [2, 3]. All this leads the Internet providers to solve the obvious task of increasing and upgrading the existing computer networks

The primary task when modernizing a computer network is to analyze network traffic, namely its structure and volume [4]. That is why this article is devoted to the analysis of computer network traffic, the detection and description of its laws for a certain time.

Based on the abovementioned, it is obvious that computer traffic has been repeatedly investigated and formalized using different methods and approaches [5, 6]. From the point of view of network modernization, the most worthwhile attention are the following:

- fractal (self-similar) analysis of network properties [7, 8];
- analysis of the source-destination streams of network traffic (Origin-Destination flows) [9]

It should be noted that these methods come to a similar conclusion that computer traffic has the clearly expressed fractal properties that are especially evident in the large time intervals such as weeks, months.

Therefore, the purpose of this work is to analyze the amount of traffic from different networks for different periods of time. It is necessary to find regularities in the time series and to prove the statistical significance of these hypotheses.

The persistent properties of web traffic allow you to build simulation models for forecasting network parameters [10, 11], to model server status and load on communication channels [12 - 15].

An important factor for maintaining the adequacy of the model is the input data for modeling. Special attention should be paid to the quantitative estimates of the volume of traffic, its marginal (maximum and minimum) and the average value for a certain period of time.

2 Materials of the research

An important component of today's computer networks is the collection of statistical data; however, if the specific indicators of Internet service providers in most cases are closed information, then the information collected on the Internet hubs is open for analysis and observation. The first objective of this study is to analyze international traffic on similarity and self-similarity. That is, you should show the similarity of network traffic for some time, as well as demonstrate the independence of the picture of the network dynamics from the specific location of the computer network, and its size.

To accomplish this, we turned to the hubs of such cities in Europe and America as: Frankfurt, Hamburg, Amsterdam, New York. Also, in order to show the relevance of the revealed properties in Ukraine, the relevant statistics collected in Ukraine will be provided.

We analyze the daily load on the network in different countries, that is, with a significantly different average traffic volume. It is necessary to analyze the hypothesis about a certain regularity of the change in the volume of traffic during the day and its independence from the mean value, which shows the independence of the properties of the operation of computer networks from their geographical location.

Below there is a statistics for network traffic in Germany, Ukraine, and America. All graphs are depicted in the form of a diagram with two axes, the X axis depicts a certain period of time (hours, days, weeks), along the Y axis - the amount of data transmitted at the specified time.

Figure 1 shows the web traffic of one of the largest European hubs in Frankfurt [16]. You can see the clearly expressed daily traffic regularity for two days, namely its growth in the so-called "peak hours" (around 8 P.M.), and the decline in traffic at night (around 4 A.M.).

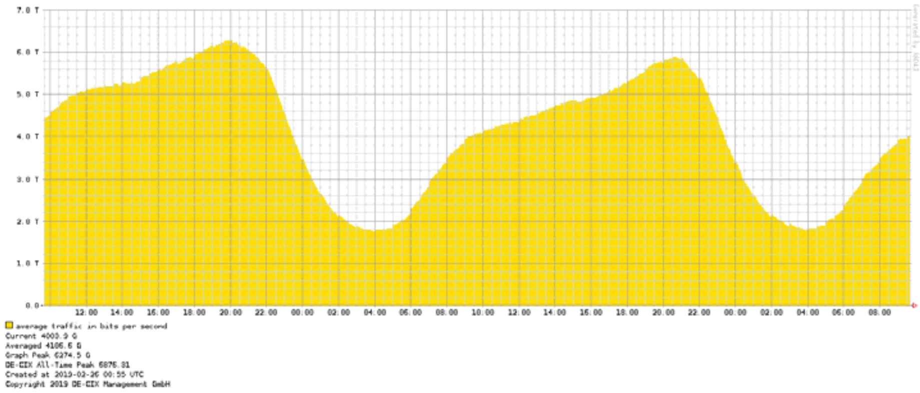


Fig. 1. Two-day web hub traffic in Frankfurt city

Figure 2 depicts web traffic for the same time period in New York [17]. Here you can observe the similar properties, as in the Frankfurt hub. However, additionally there is a peak load at about 12 o'clock.

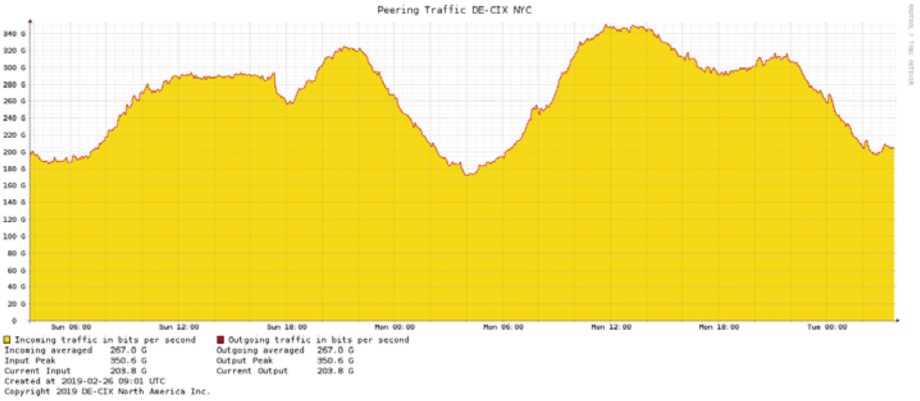


Fig. 2. Two-day web hub traffic in New York City

To illustrate a similar picture in Ukraine, Figure 3 shows the generalized web traffic collected from all hubs of Ukrainian cities [18]. Here the peak charge is centered between 8 P.M. and 9 P.M., and the minimum is between 4 A.M. and A.M.

Although the resource UA-IX allows seeing the statistics only for the current time, the similarity of Web traffic in Ukraine with all the above-mentioned is evident.

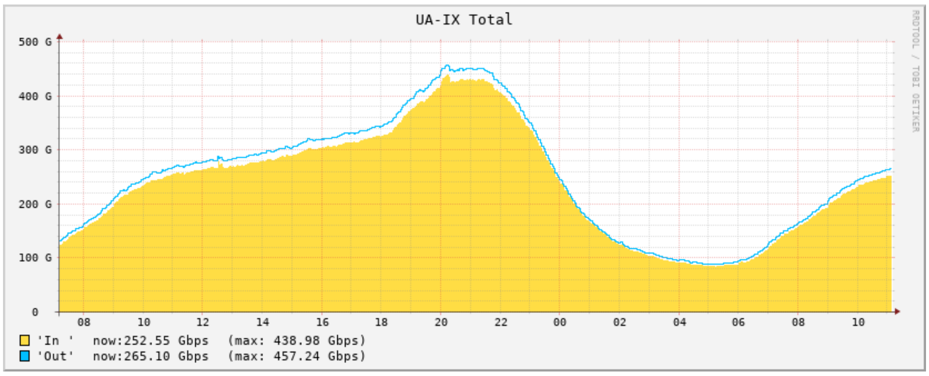


Fig. 3. Daily web traffic in Ukraine

We saw a visual similarity, but noticed some differences. It should be proved that the law is essential, and the differences are random. To revise this hypothesis, it is necessary to compare several graphs, that is, several types of distribution of the random variable of the intensity of traffic for a certain hour.

All Internet hubs that were selected vary not only in geographic distance, but also in traffic volumes. We can conditionally split these hubs into large, medium and small ones. Therefore, the hub in Frankfurt is the largest, with peak loads in several terabytes of information. The Hub of New York City is a mid-day sized and has a peak load of 250-300 gigabytes of data. Ukraine is depicted not by specific cities, but by the total volume of data in the country; therefore, hubs of Ukraine can be considered small, since the total volume of the country does not exceed the amount of data transmitted in New York City, that is, 400 gigabytes in peak hours.

Consider the task of analyzing the similarity of computer network traffic over a period of time. This study aims to demonstrate the similarity of computer traffic over a significant period of time, which in turn allows the use of simulation to predict the state of the computer network with a high probability.

To simulate the behavior of a computer network with the use of simulation models, it is important to use predictable traffic [19]. Therefore, it is necessary to prove that the trends indicated in the preliminary data (Fig. 1-3) are stored for a long time. Using the information gathered in these Internet hubs, it can be shown that relative volatility of network traffic remains unchanged despite the apparent and continuous increase in network traffic.

Figure 4 shows the network traffic of the city of Frankfurt during the month, it is possible to clearly indicate the storage of the above trends in this period of time.

However, this schedule cannot demonstrate the growth of traffic volumes due to the modernization of computer networks. To do this, you need to use annual reports, or even longer periods of time. For example, the statistics of the city of Frankfurt for the year (Fig. 5) and over 5 years are shown below (Fig. 6). An annual graph shows traffic growth of about 0.5 terabytes.

The best dynamics of growth can be seen in this figure. In just two years the amount of traffic has doubled from 2 terabytes to 4.

Despite the volatile dynamics of computer traffic and its continuous-out growth, one can distinguish user's party behavior.

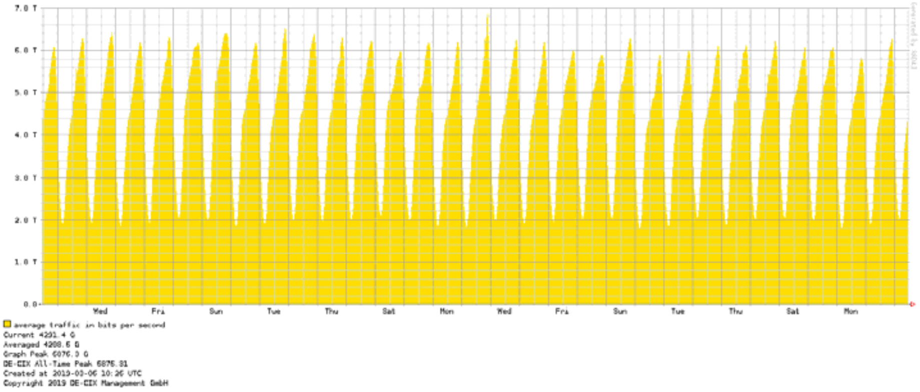


Fig. 4. Monthly web traffic in Frankfurt

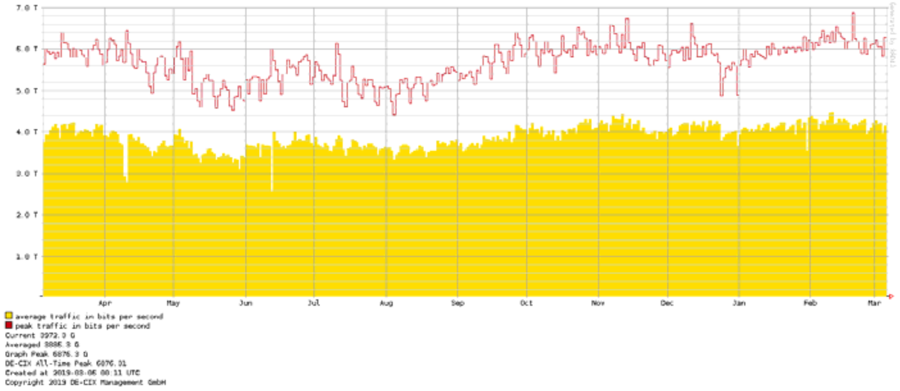


Fig. 5. Annual Web Traffic in Frankfurt

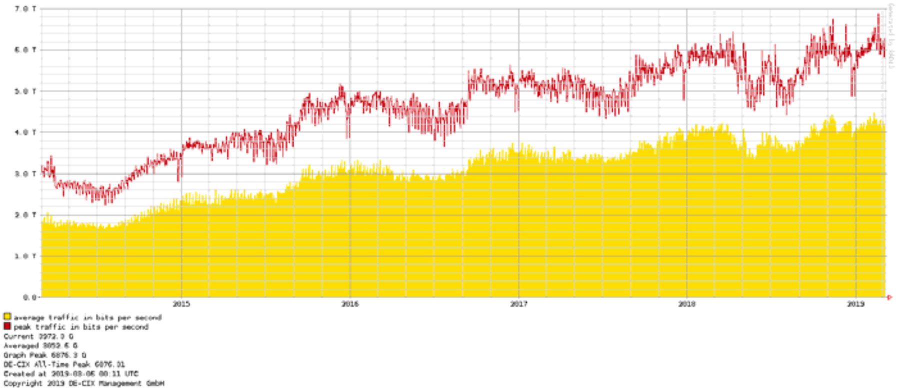


Fig. 6. Five-year web traffic in Frankfurt

3 Results and discussions

According to figures shown in Fig. 1-3, select a daily trend and reduce it to one linear graph (Figure 7). It can be seen that the graphs are similar to changes in volumes at certain parts of the time, but somewhat different.

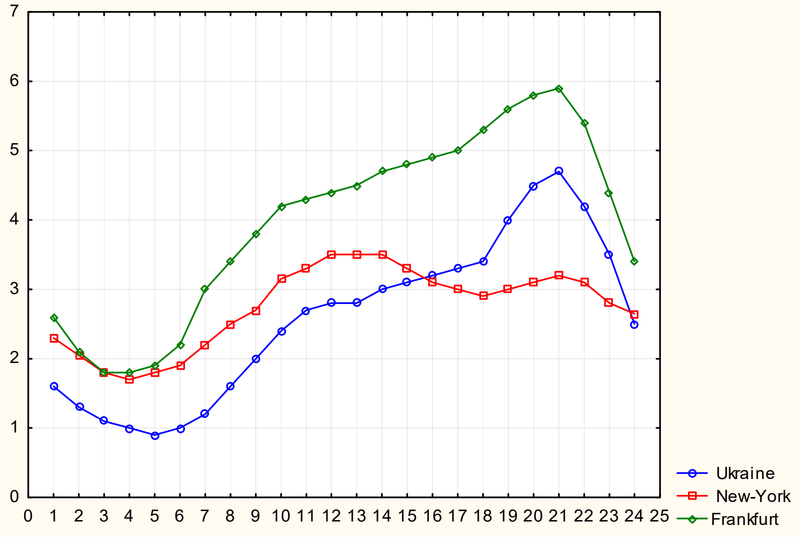


Fig. 7. Outgoing comparable graphs of intensity of traffic during the day

To make the analysis invariant relative to the absolute value of traffic, we normalize the output timelines, that is, let us give the average value of traffic to zero value (Fig. 8). It is seen that the second graph (data in New York) is slightly different. To compare the intensity distribution over the course of the day, use the non-parametric chi-square [21, 22] criterion, making three pairs of comparisons:

(Ukraine-Frankfurt): chi-square = 2,019, df = 23, p = 1,000
(Ukraine-New York): chi-square = 17,300, df = 23, p = 0.794
(New York-Frankfurt): chi-square = 11.670, df = 23, p = 0.975.

In the first comparison, the data coincide very well. The hypothesis of the randomness of convergence is confirmed (the level of significance approximates to one). In the second case, the discrepancy is significant (p = 0,794), the error is about 21%. But the third comparison also gave a meaningful result (p = 0,975).

Thus, under certain assumptions, we can conclude that the trend of changing the intensity of traffic for a long time is statistically significant. To predict the maximum load and simulate loading, you can use averaged traffic, but average values are calculated with the certain weights: the New York data graph should be considered with less weight, according to the value of the degree of confidence received. Averaged traffic (with output) is shown in Fig. 8.

Since the chi-square criterion does not provide a confident answer to the incidence of discrepancies in relation to all three charts, we will apply another verification procedure.

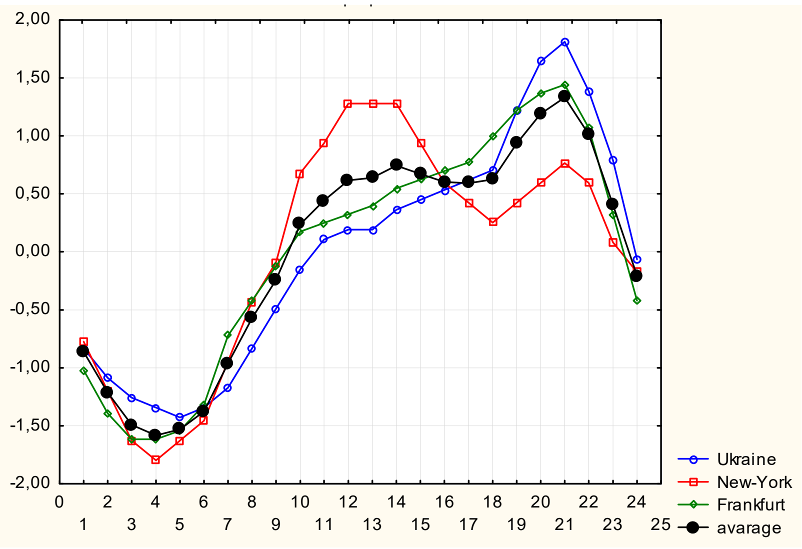


Fig. 8. Normalized linear graphs of traffic intensity and averaged overnight traffic

For a pairwise comparison of the three graphs, apply the Mann-on-Whitney criterion for independent samples [21]. The comparison results are given in Table 1, 2. The indicated criteria are significant.

Table 1. Results of the Mann-Whitney's Criteria for samples (Ukraine-New York)

	Sum rank gr 1	Sum rank gr 2	U	Z	p-value	Z - adjust.	p-value	N gr.1	N gr.2	2-sided exact p
Traf 1-2	584	592	284	-0,072	0,942	-0,072	0,942	24	24	0,943

Table 2. Results of the Mann-Whitney's Criteria for samples (Ukraine - Frankfurt)

	Sum rank gr 1	Sum rank gr 2	U	Z	p-value	Z - adjust.	p-value	N gr.1	N gr.2	2-sid. exact p
Traf 1-3	585	591	285	-0,05	0,958	-0,051	0,958	24	24	0,959

The results of the comparison of the samples (New York-Frankfurt) fully coincide with the results of Table 2, because the selected criterion is ranked, and the calculated ranges in these groups coincide, despite different values of the output data.

As can be seen from the results of the estimates of the significance of the criterion, all comparisons are based on the randomness of the difference in groups: confidence level p = 0,942, and 2-sided exact p = 0,943; in the second comparison p = 0.958, and 2-sided exact p = 0.959. These values confirm the hypothesis of the significance of the difference. So, in the first comparison, the error is about 5.6%, in the other two about 4%.

Now let's conduct simultaneous comparison of three samples with the help of the criterion of Kruskal-Wallis (Table 3).

Table 3. Results of the Kruskal-Wallis criterion for three samples

H (2, N= 72) = 0,008 p = 0,996			
Cod	N	Sum Ranks	Mean Rank
1	24	869,0000	36,20833
2	24	877,0000	36,54167
3	24	882,0000	36,75000

The results of this test fully confirm the hypothesis of the incident of discrepancies, level of trust is close to one (p = 0,996).

Based on the entire above, one can conclude that the structure of web traffic is similar throughout the world, regardless of the geographical location of the computer network and the volume of web traffic. Thus, it can be used in constructing simulation models of computer networks, as inputs to achieve model adequacy and forecast accuracy.

The next step in the study is to analyze the change in traffic over the course of a month. The graph of its change in accordance with Fig. 4 is shown in Fig. 9.

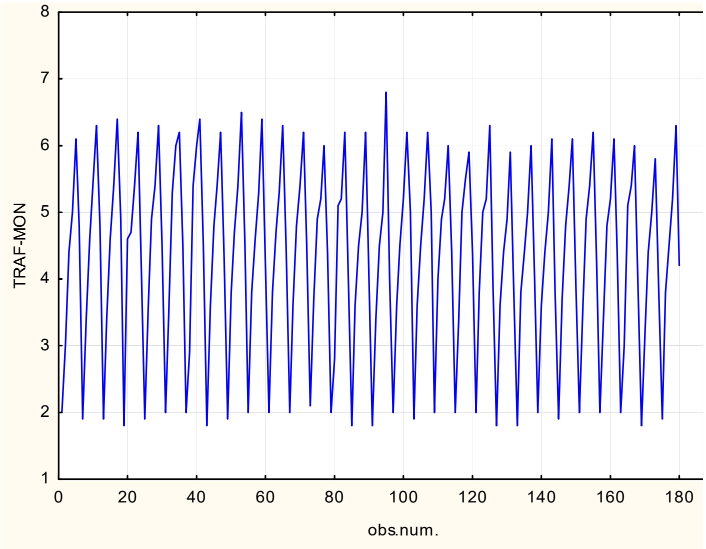


Fig. 9. Change the intensity of traffic for a month

Output data form a stationary time series (under the condition of converting the normalization of values to normalized values) [23]. Frequency of changes is not in doubt. Let us prove this by constructing the model of auto regression taking into account the seasonal component (Table 4) [24, 25]. A model with three parameters was selected: q - autoregressive parameter, Ps (1) and Ps (2) - seasonal parameters of auto regression. It is seen that the estimates of the parameters are very precise (the significance levels p are close to zero).

Table 4. Estimation of parameters of autoregressive data model for five years

Model (1,0,0)(2,0,0) Seasonal lag: 24 MS Residual = 0,3041				
Param.		Asimpt. - Std.error	Asimpt. - t(116)	p
p(1)	0,7932	0,0473	16,7417	0,0000
Ps(1)	0,7400	0,0753	9,8263	0,0000
Ps(2)	0,2361	0,0756	3,1202	0,0021

Thus, the resulting model accurately describes the time series of observations. The periodicity of a series with a seasonal lag of 24 hours is confirmed.

The randomness of the model errors is confirmed by the residue graph, which is close to the normal distribution, as illustrated by the normal probabilistic graph (Fig. 10). It is seen that the remnants are not substantially deviating from the straight line, only some deviations in the upper part of the graph are observed.

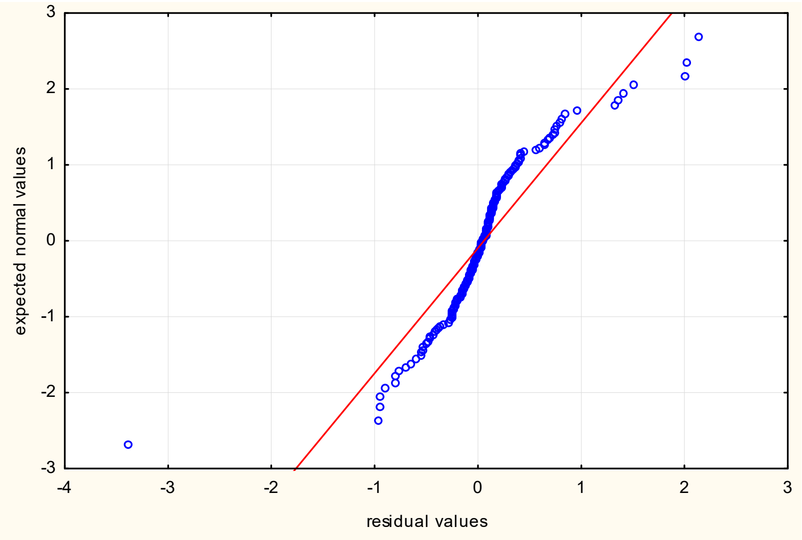


Fig. 10. Normal probabilistic distribution graph of residues

The frequency of the analyzed time series is also confirmed by the graph of the auto regression function.

The third stage of the study is the analysis of traffic changes over the year based on the data presented in Fig. 5 (data on the maximum load was applied). A linear graph of data, forming a time series, is presented in Fig. 11. It should be noted that the data are presented from March to February of the year following the year.

It is seen that the range is not stationary. For its analysis, a number of significant transformations should be carried out [26]. When applying the exponential smoothing of a range you can see the tendency of its change, which should pay attention to the months of maximum load. One can assume that there is an annual periodicity of traffic changes. It can be analyzed only on the basis of data for several years (Fig. 6). The time range of values of maximum loads in the network is shown in Fig. 12.

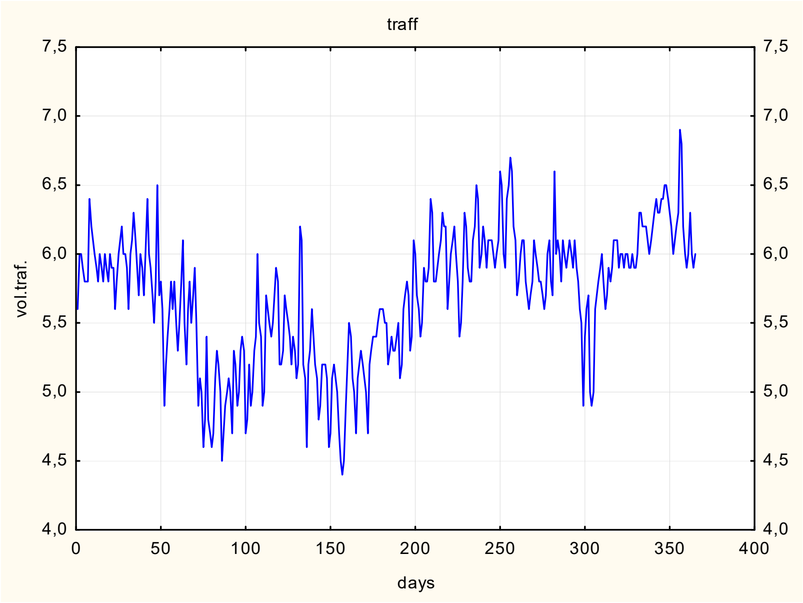


Fig. 11. Time range of traffic intensity throughout the year

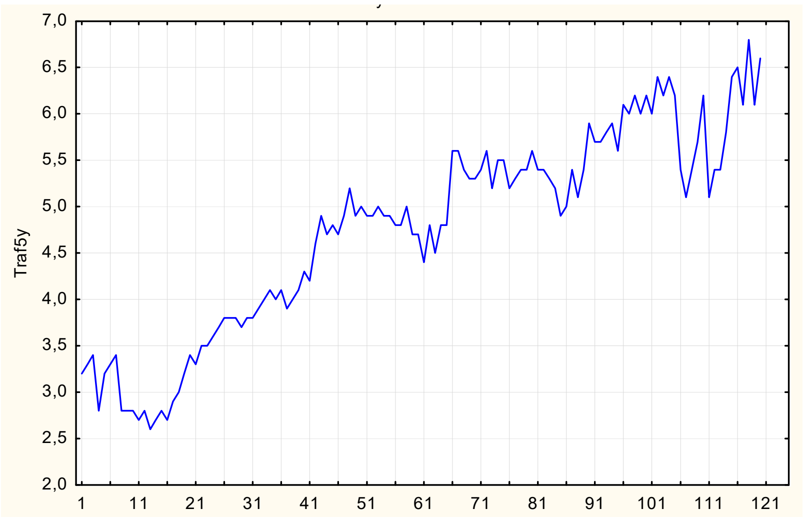


Fig. 12. Time range of maximum traffic loads for five years

The frequency of this range during the year visually expressed is not clear. In order to check its periodicity, using the autocorrelation model, it is necessary to increase the number of stationary ones. To do this, the shift was applied to one lag (Figure 13).

A model of auto regression with three parameters was selected (Table 5), here additionally Qs is the seasonal parameter of the sliding average. It is seen that the parameters are estimated fairly accurately (the levels of significance are close to zero), the function of autocorrelation (Figure 14) also confirms the correctness of the model. The figure shows that for 15 logs the value of autocorrelation is random: both positive and negative values do not exceed the line of the permissible limit. One can conclude that the seasonal annual frequency is significant. That is, it is a pattern that the lowest traffic is observed in June and August, and the highest in February (this is evident from Figure 12).

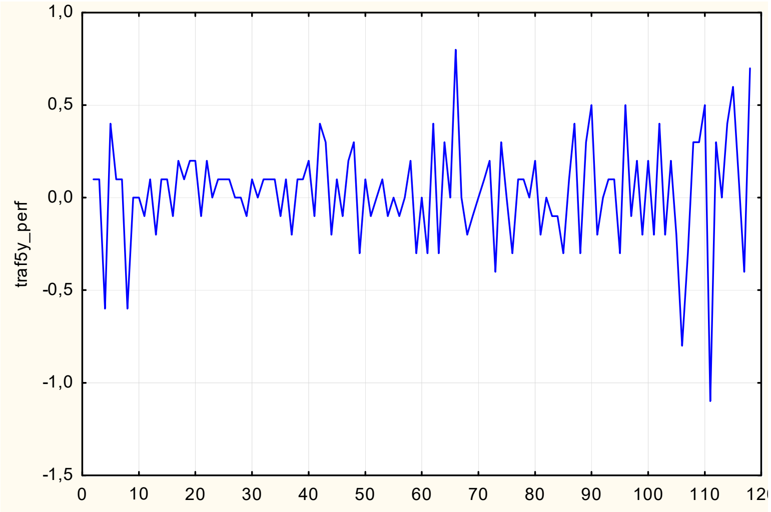


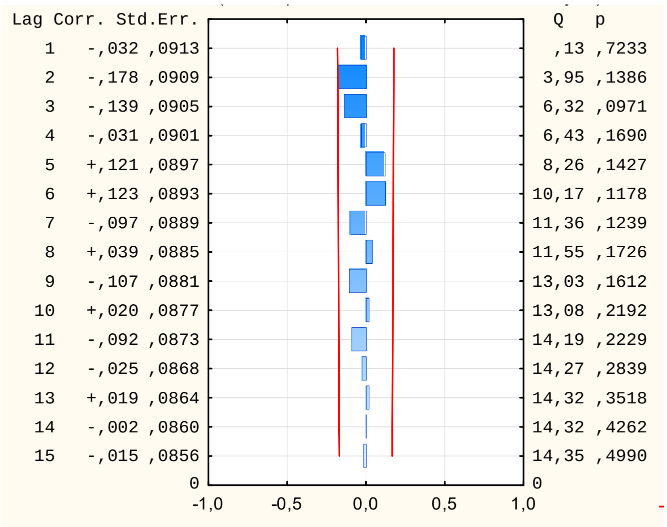
Fig. 13. Converted time range of maximum traffic loads for five years

Table 5. Estimation of parameters of autoregressive data model for five years

Transform.: D (1) Model (0,1,1) (1,0,1) Seasonal lag: 24 MS Residual = 0,0662				
Param.		Asimpt. - Std.error	Asimpt. - t(116)	p
q(1)	0,4865	0,0906	5,366837E+00	0,0000
Ps(1)	0,9998	0,0000	2,734817E+19	0,0000
Qs(1)	0,6742	0,1017	6,623994E+00	0,0000

According to these same data one can investigate the trend of increasing traffic intensity (Fig. 15). The trend was obtained by smoothing with the help of the weighted least squares method.

The resulting trend can be used to predict the maximum load on the network for future years [27, 28]. At the same time, it should be made possible not to take into account the sharp changes in the trend associated with qualitative changes (for example, due to the explosion of new technologies) in the field of telecommunications in the future.



Lag	Corr.	Std.Err.	Q	p
1	-,032	,0913	,13	,7233
2	-,178	,0909	3,95	,1386
3	-,139	,0905	6,32	,0971
4	-,031	,0901	6,43	,1690
5	+,121	,0897	8,26	,1427
6	+,123	,0893	10,17	,1178
7	-,097	,0889	11,36	,1239
8	+,039	,0885	11,55	,1726
9	-,107	,0881	13,03	,1612
10	+,020	,0877	13,08	,2192
11	-,092	,0873	14,19	,2229
12	-,025	,0868	14,27	,2839
13	+,019	,0864	14,32	,3518
14	-,002	,0860	14,32	,4262
15	-,015	,0856	14,35	,4990

Fig. 14. Autocorrelation function for auto regression model

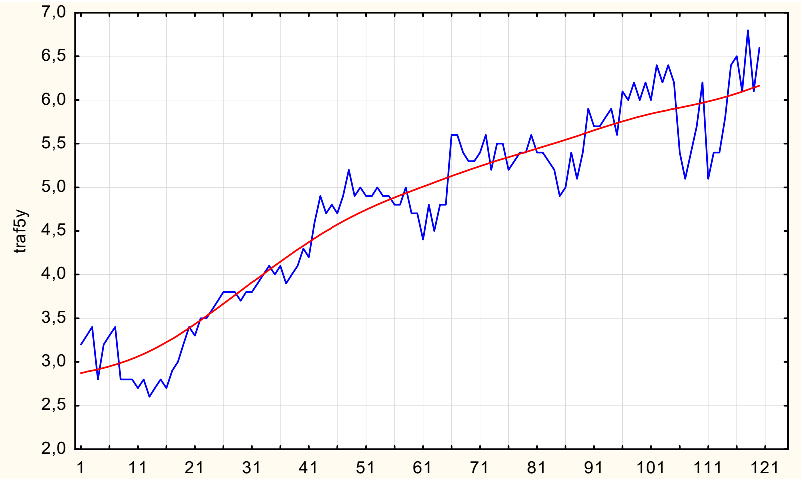


Fig. 15. Trend increase traffic for five years

4 Conclusion

The task was to find and analyze the patterns in the on-load of computer networks by studying international traffic at different time intervals (day, month, year and five years). The following results are obtained:

1) Analyzed the daily load of computer networks. Comparison of international traffic on similarity and self-similarity with the help of three statistical criteria is carried out. It is concluded that the trend of changing the intensity of traffic over the course of the day is statistically significant. The daily periodicity is proved for a cer-

tain time and the hypothesis of the independence of the network dynamics from the specific location of the computer network and its size is confirmed. Consequently, we can conclude that the structure of web traffic is similar throughout the world, regardless of the geographical location of the computer network and the volume of traffic. Thus, you can use averaged traffic as input data when simulating the loading of computer networks.

2) Traffic data is analyzed within a month. Taking into account the considerable number of observations, methods of analysis of time series were used. By constructing an auto-regression model of the time series, according to observations of observations during the month, the frequency of a series with a seasonal lag of 24 hours has been proved.

3) Traffic data for the year is analyzed. By smoothing random data, a graph of intensity changes was received, from which no periodicity was visually revealed.

4) The observation of intensity of traffic for five years is analyzed. The seasonal annual frequency of traffic intensity was proved on the basis of an autoregressive model with three parameters. In addition, on the basis of five-year observations, a traffic growth trend has been obtained, which can be used to predict the maximum load in the future for years to come.

Thus, the researches are characterized by a scientific novelty, which consists in the fact that the hypothesis of self-similarity, periodicity and the presence of a trend in the intensity of traffic of modern computer networks has been proved for the first time, which allows scientifically reasonably predict maximum network load at its modernization.

Further research is to be carried out in the direction of forecasting of the maximal load on the network and simulation of network parameters with consideration of predicted traffic.

References

1. Reva, A., Davydovskyi, Y.: Method of the network topology transformation to quasihomogeneous structure. Radioelectronic and computer systems, 2, 43-51 (2018), doi: 10.32620/reks.2018.2.05
2. Internet World Stats: Usage and Population Statistics. https://www.internetworldstats.com/stats.htm
3. Zhuang, Y., Cappos, J., Rappaport, T., McGeer, R.: Future Internet Bandwidth Trends: An Investigation on Current and Future Disruptive Technologies. Secure Systems Lab, Dept. Comput. Sci. Eng., Polytech. Inst. New York Univ., New York, NY, USA, Tech. Rep. TR-CSE-2013-0411/01/2013 (2013)
4. Lavrut, O. O.: The quality control of information flows in the telecommunication system of critical use. Systems of Arms and Military Equipment, 4 (40), 89–93 (2014)
5. Stepanov, S. N.: Basics of teletraffic multiservice networks. Eco-Trendz (2010)
6. Paramonov, A. I.: Models of traffic flows for M2M networks. Telecommunications and radio engineering, 4, 11–16 (2014)
7. Leland, W., Taqqu, M., Willinger, W., Wilson, D.: On the self-similar nature of Ethernet traffic. IEEE/ACM Transactions on Networking, 1, 1-15 (1994)
8. Peters, E. E.: Fractal analysis of financial markets. Application of Chaos Theory in Investment and Economics. Internet-trading, Moscow (2010)

9. Lakhina, A., Papagiannaki, K., Crovella, M., Diot, C.: Structural Analysis of Network Traffic Flows. ACM SIGMETRICS Performance Evaluation Review, 32(1), 61-72 (2004)
10. Cisco: Cisco Visual Networking Index: Forecast and Trends, 2017–2022 White Paper. Cisco VNI, (2018)
11. Bunimovich, L., Smith, D., Webb, B.: Specialization Models of Network Growth. Journal of Complex Networks, (2018), doi: 10.1093/comnet/cny024
12. Mestres, A., Alarcon, E., Cabellos-Aparicio, A.: Understanding the Modeling of Computer Network Delays using Neural Networks. Big-DAMA@SIGCOMM (2018), doi: 10.1145/3229607.3229613
13. Kosenko, V.: Principles and structure of the methodology of risk-adaptive management of parameters of information and telecommunication networks of critical application systems. Innovative technologies and scientific solutions for industries, 1 (1), 45–51 (2017) doi: https://doi.org/10.30837/2522-9818.2017.1.046
14. Kosenko, V., Persiyanova, E., Belotskyy, O., Malyeyeva, O.: Methods of managing traffic distribution in information and communication networks of critical infrastructure systems. Innovative technologies and scientific solutions for industries, (2 (2), 48-55 (2017) doi: 10.30837/2522-9818.2017.2.048
15. Kuchuk, G. A., Kirillov, I. G., Pashnev, A. A.: Modeling of traffic of multiservice distributed telecommunication network. Processing Systems Information, 9 (58), 50–59 (2006)
16. DE-CIX Frankfurt statistics. https://www.decix.net/en/locations/germany/ frankfurt/statistics
17. DE-CIX New York statistics. https://www.de-cix.net/en/locations/united-states/new-york/statistics
18. UI-EX. Ukrainian Internet Exchange. Statistics. https://www.ix.net.ua/pro-kompaniyu/statystyka
19. Poshtarenko V. M., Andreev A. Yu., Amal M.: Ensuring the quality of service at critical sites of the multiservice network. Bulletin of the NTU "KhPI", 60, 94–100 (2013)
20. Sultanov, A. Kh., Sultanov, R. R.: Method of assessing the quality of service indicators of hierarchical multiservice networks. Bulletin USATU, Vol. 12, 1 (30), 175–181 (2009).
21. Borovikov, V. P.: STATISTICA. The art of analyzing data on a computer. Piter, SPg. (2001)
22. Blake, A. P. Kapetanios, G.: Pure Significance Tests of the Unit Root Hypothesis Against Nonlinear Alternatives. Journal of Time Series Analysis. Vol. 24, 3, 253–267 (2003)
23. Kendal, M.: Time Series. Finance and Statistics, Moscow (1981)
24. Box, J., Jenkins, G.: Time series analysis, forecast and management. Mir, Moscow (1974)
25. Borovikov, V. P.: Prediction in the STATISTICA system in the Windows environment. Fundamentals of the theory and intensive practice on the computer. Finance and statistics, Moscow (2000)
26. Orlov, Yu. N., Osminin, K. P.: Non-stationary time series. Forecasting methods with examples of analysis of financial and commodity markets. Librokom, Moscow (2011)
27. Lukashin, Yu. P.: Adaptive methods for short-term forecasting of time series. Finance and Statistics, Moscow (2003)
28. Hanke, D. E., Wichern, D. W., Reitsch, A. J.: Business Forecasting. Williams, Moscow (2003)