RESEARCH ARTICLE | SEPTEMBER 26 2022

Analysis of the temporal distribution of passenger traffic in road transport for the regional road network

O. Ie

AIP Conf. Proc. 2522, 050007 (2022)
https://doi.org/10.1063/5.0100765

Analysis of the Temporal Distribution of Passenger Traffic in Road Transport for the Regional Road Network

O. Ie[1,2,a)]

[1)]*Ural State University of Railway Transport, 66 Kolmogorov str., 620034 Ekaterinburg, Russia*
[2)]*Ural Federal University named after First President of Russia B. N. Yeltsin, 19 Mir str., 620002 Ekaterinburg, Russia*

[a)]*Corresponding author: Olgaie@mail.ru*

Abstract. The indicators of passenger traffic on road transport for the regional transport network, obtained on the basis of data from online ride-sharing services, are considered in the article. The features and patterns of daily fluctuations in passenger traffic are analyzed. Seasonal and cyclical fluctuations in passenger traffic volumes are modeled using harmonic analysis methods. All necessary statistical procedures are used to identify and evaluate the parameters of the model and to verify its adequacy and accuracy. Short-term forecasts and conclusions of the study are made based on the results obtained.

INTRODUCTION

The study of passenger traffic (PT) in road transport makes it possible to identify the congestion of the transport network in a certain period of time. The formation of PT occurs under the complex influence of various factors, the degree of which is not the same. The need for travel naturally changes by periods and directions of travel. The PT value changes by hours of the day, days of the week, months, seasons of the year. All PT fluctuations must be systematically studied and promptly taken into account for the implementation of sustainable transport planning and technological organization of transport.

The main method for studying the development of passenger transport trends is forecasting. The accuracy of forecasts determines the reality of planning decisions taken. Prediction for today is considered as a mandatory part of the planning process.

Various economic and mathematical methods are used to study the influence of individual factors and their combination on passenger transportation. The accuracy of three representative parametric and nonparametric forecasting models is investigated and compared: Non-parametric K-NN regression model, Gaussian maximum likelihood model and double seasonality HolteWinters exponential smoothing model confirm their goodness to predict the weekly and monthly fluctuations of average daily traffic with varying degrees of performance in [1]. The models of natural development of transport networks influenced by various external and internal factors are considered in papers [2, 3]. The issues of forecasting passenger traffic in the study of duplicate routes were studied on the basis of a comparative analysis of the correspondence of various modes of transport using a synthetic approach in articles [4, 5]. A multiplicative time series model with polynomial smoothing, which can be used for short-term forecasting of passenger traffic for a regional transport network, was constructed in [6]. In paper [7] the spatial and temporal characteristics of PT on different metro lines were analyzed for five working days with normal weather conditions in the city of Nanjing.

Various statistical methods are used to solve forecasting problems. Forecasting method based on trend and time series variability is one such method.

In this paper the modeling of seasonal and cyclical fluctuations in the volume of passenger traffic is carried out using the methods of harmonic analysis.

MATHEMATICAL TOOLS

In the general case the time series Y_t consists of four components: the trend T_t, the seasonal component S_t, the cyclical component C_t, and the random, irregular component ε_t [8]. Here t is a moment in time. Models of relationships between these components can be of different forms: additive, multiplicative, mixed.

Application of Mathematics in Technical and Natural Sciences
AIP Conf. Proc. 2522, 050007-1–050007-8; https://doi.org/10.1063/5.0100765
Published by AIP Publishing. 978-0-7354-4361-7/$30.00

The additive model is applied if the amplitude of seasonal fluctuations remains approximately constant. The multiplicative model is used if the amplitude of seasonal fluctuations changes depending on the change in the general level of the values of the series. Various methods are used to build models of seasonality and cyclicality: the method of applying seasonal dummy variables, the method of autoregressive and moving average, the method of harmonic analysis, *etc.* In most methods the length of the periodic component is usually known (or assumed) in advance or is revealed visually from the series plot and then included in the theoretical model. Unlike other methods, harmonic analysis allows one to recognize periodic oscillations of various lengths and, on this basis, construct a periodic model of the series.

According to harmonic analysis [9], the time series is presented as a set of harmonic oscillatory processes. For each point of this series, the following expression is true

$$y_t = f(t) + \sum_{k=1} \left(a_k \cos\left(kt \frac{2\pi}{n} \right) + b_k \sin\left(kt \frac{2\pi}{n} \right) \right), \;\; t = 1,2,...,n. \tag{1}$$

Here y_t is the actual level of the series at the moment (interval) of time t; $f(t)$ is the aligned level of the series at the same moment in time; a_k, b_k are parameters of the oscillatory process (harmonics) with the number k, together assessing the range (amplitude) of the deviation from the general trend and the shift of oscillations relative to the initial point.

The total number of oscillatory processes that can be distinguished for a series of n levels is $n/2$. Usually limited to a smaller number of the most important harmonics. The parameters of the harmonic with number k are determined by the formulas:

$$\varepsilon_t = y_t - f(t); \tag{2}$$

$$a_k = \frac{2}{n} \sum_{t=1}^{n} \varepsilon_t \cos\left(kt \frac{2\pi}{n} \right), \;\; k = 1,2,..., \frac{n}{2} - 1; \tag{3}$$

$$b_k = \frac{2}{n} \sum_{t=1}^{n} \varepsilon_t \sin\left(kt \frac{2\pi}{n} \right), \;\; k = 1,2,..., \frac{n}{2} - 1; \tag{4}$$

$$a_{n/2} = \frac{1}{n} \sum_{t=1}^{n} \varepsilon_t \cos(\pi t); \;\; b_{n/2} = 0. \tag{5}$$

Each sum term represents the harmonic with a certain period. The first harmonic has a period equal to the length of the period under study. The main period is set equal to 2π. The second harmonic has a period equal to half of the fundamental, the third - one third of the fundamental, *etc.* If there are N observations, then the number of harmonics will not exceed $N/2$.

Periodogram [10], which is a function of data dispersion versus frequency, is used to recognize periodic oscillations of different frequencies. Graphically the periodogram is usually depicted as the dependence of the dispersion of harmonics on their number. Variance accounted one harmonic, is calculated by the following formula:

$$\sigma_k^2 = \frac{C_k^2}{2}, \tag{6}$$

where

$$C_k^2 = a_k^2 + b_k^2.$$

The periodogram view is closely related to the structure of the series and is a good tool for revealing hidden periodicities and pre-selection of harmonics to be included in the model. The presence of seasonal and cyclical fluctuations manifests itself in the form of sharp narrow peaks in the periodogram at the corresponding frequencies.

The F_k statistics and the significance level of each harmonic p_k are calculated to assess the significance of each harmonic [11]:

$$F_k = \frac{\sigma_k^2/2}{\sigma_{\text{res}}^2/(N-(2q+1))}, \tag{7}$$

where σ_{res}^2 is a residual variance, which is determined by the formula

$$\sigma_{\text{res}}^2 = \sum_{k=1}^{N/2} \sigma_k^2 - \sum_{j=1}^{q} \sigma_{k(j)}^2, \tag{8}$$

j is a harmonic numbers that are taken into account in the model, q is a quantity of harmonics that are taken into account in the model.

INITIAL DATA

The information base is necessary for modeling seasonal and cyclical fluctuations in the volume of passenger traffic. Data on the characteristics of a trip on routes can only be obtained from the survey of PT, which is not always possible. The most common survey methods require large material and labor costs. PT on private vehicles is quite difficult to measure. There is no publicly available data on PT volumes. Therefore, ride-sharing services were considered as a source of such data.

To process large amounts of data from online ride-sharing services, which are not only difficult to manually select, but in some cases impossible, the technology of obtaining web data by means of Web Scraping [12] is used. In this paper web scraping is implemented using code written in Python.

During the daily scraping at the same time for 2 months, data was collected on the points of departure and arrival, time of departure and arrival, price of the trip. The settlements of the Sverdlovsk region with a population of more than 20 thousand people were considered as points of departure and arrival, which amounted to 992 requests. The result is a multidimensional array of 10500 records, each of which includes 7 elements. Processing the received amount of data required very significant computing power, so the Wolfram Mathematica program was used for data analysis and modeling.

MAIN RESULTS

Since the purpose of the work is to study the temporal distribution of passenger traffic, the next step is to group trips by days of the week. The statistics of the region's daily passenger traffic depending on the day of the week is shown in Figure 1.

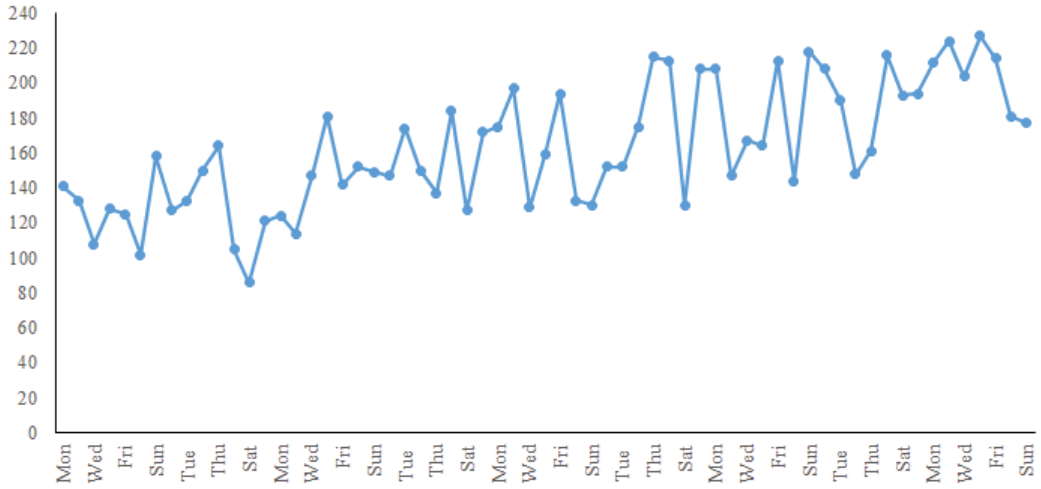

FIGURE 1. Distribution of passenger traffic in the Sverdlovsk region by road transport by days of the week

Already at the stage of graphical analysis, it is possible to determine the presence of a trend and seasonal component. Cyclic fluctuations on the chart are not visually detected. To determine their presence and period, it is necessary to exclude the trend and seasonal fluctuations from the series. The amplitude of seasonal fluctuations is approximately the same, so it is advisable to build an additive model of the following form:

$$Y_t = T_t + S_t + C_t + \varepsilon_t. \tag{9}$$

TABLE I. Correlogram of the time series of passenger traffic indicators.

Lag	Autocorrelation coefficient of levels	Correlogram
1	0.507	*****
2	0.417	****
3	0.514	*****
4	0.448	****
5	0.242	**
6	0.322	***
7	0.566	******
8	0.336	***
9	0.411	****
10	0.512	*****
11	0.434	****
12	0.337	***
13	0.244	**
14	0.339	***
15	0.248	**

First of all, let us define a trend and exclude it from the series. Analysis of trend models shows that a 5th degree polynomial can be chosen as a working model:

$$T_t = -0.0000033t^5 + 0.0005412t^4 - 0.0320106t^3 + 0.827084t^2 - 7.3286859t + 143.5967802.$$

It is reasonable to assume that the passenger traffic indicators in the current period depend on the passenger traffic indicators of the previous periods. Let us determine the degree of closeness of such a connection between the levels of the series using the autocorrelation coefficients [11]

$$r_\tau = \frac{\sum_{t=\tau+1}^{n} (y_t - \bar{y}_{1\tau})(y_{t-\tau} - \bar{y}_{2\tau})}{\sqrt{\sum_{t=\tau+1}^{n} (y_t - \bar{y}_{1\tau})^2 \sum_{t=\tau+1}^{n} (y_{t-\tau} - \bar{y}_{2\tau})^2}}, \tag{10}$$

where τ is the magnitude of the shift (lag), which determines the order of the autocorrelation coefficient,

$$\bar{y}_{1\tau} = \frac{1}{n-\tau} \sum_{t=\tau+1}^{n} y_t, \;\; \bar{y}_{2\tau} = \frac{1}{n-\tau} \sum_{t=\tau+1}^{n} y_{t-\tau}.$$

The following results are obtained (see Table 1).

The study of the autocorrelation function and correlogram makes it possible to analyze the structure of the series. The autocorrelation coefficient of the order of 7 is of the greatest importance, so we make the assumption that the series contains seasonal fluctuations with a frequency of 7 points in time.

The seasonal component is excluded from the detrended series using a centered moving average calculated over 7 levels of the series. The resulting graph of smoothed cyclical fluctuations is shown in Figure 2.

The data in Figure 2 indicate that there is a cyclic component with unstable amplitude in the time series under study.

For the obtained detrended series, we calculate the Fourier coefficients using formulas (3)-(5). Results of calculation the harmonic coefficients and other parameters necessary for identification of the model are given in Table 2.

As can be seen from Table 2, only four harmonics are significant at $p < 0.05$: 6th, 8th, 18th and 20th.

The eighth harmonic reflects seasonal effects with a period of 7 days and an oscillation frequency of 0.13, it is the main component in the seasonality model. Similarly, the sixth harmonic with a period of 10.5 days and a frequency of 0.10 is the main component in the cyclic model. The frequency 0.29 will be a multiple of this frequency, which corresponds to the 18th harmonic. In addition, the 20th harmonic with a frequency of 0.32 is also a component in the cycling model. Therefore, the model of cyclic oscillations will include the 6th, 18th and 20th harmonics.

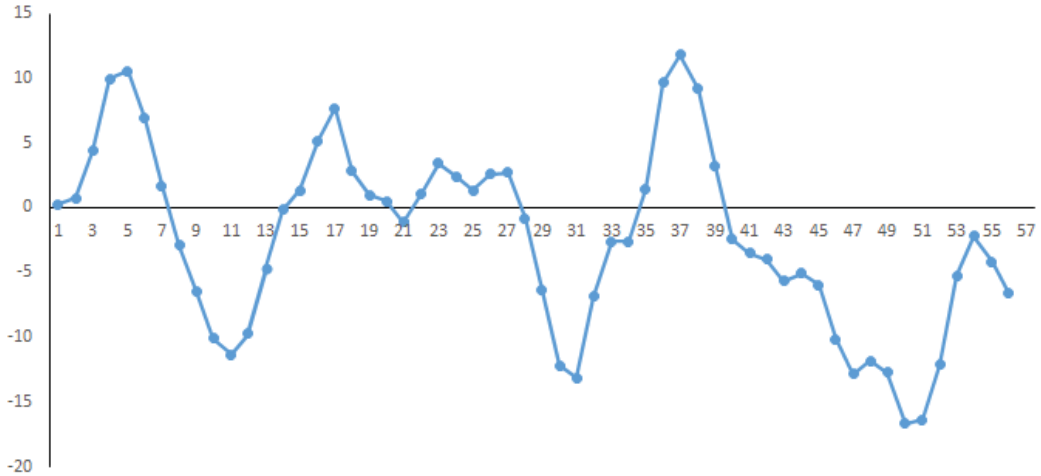

FIGURE 2. Graph cyclic fluctuations

Thus, we obtain the following models of the trend, seasonal and cyclical components:

$$\begin{aligned}
&T_t = -0.0000033t^5 + 0.0005412t^4 - 0.0320106t^3 + 0.827084t^2 - 7.3286859t + 143.5967802;\\
&S_t = 5.71\cos\left(\frac{16\pi}{63}t\right) + 10.85\sin\left(\frac{16\pi}{63}t\right);\\
&C_t = 1.03\cos\left(\frac{12\pi}{63}t\right) - 10.95\sin\left(\frac{12\pi}{63}t\right) - 2.15\cos\left(\frac{36\pi}{63}t\right) + 14.36\sin\left(\frac{36\pi}{63}t\right)\\
&-9.94\cos\left(\frac{40\pi}{63}t\right) + 5.38\sin\left(\frac{40\pi}{63}t\right).
\end{aligned} \tag{11}$$

The resulting additive model of the series has the following accuracy indicators: coefficient of determination is equal $R^2 = 0.867$, the mean square error is equal 18.69. These indicators, as well as the graphs of the initial and model values of the series levels, presented in Figure 3, indicate the high accuracy of the constructed model. Consequently, this series model can be used to build a point forecast as a whole for the indicator under study.

We will make a short-term forecast of the expected indicators of passenger traffic for 3 days.

The predicted value Y_{p_t} of the time series level in the additive model is the sum of the trend, seasonal and cyclical components. Using the constructed polynomial model, we find the values of the trend component:

$$\begin{aligned}
T_{64} &= -0.0000033\cdot 64^5 + 0.0005412\cdot 64^4 - 0.0320106\cdot 64^3\\
&\quad + 0.827084\cdot 64^2 - 7.3286859\cdot 64 + 143.5967802 = 207.392;\\
T_{65} &= -0.0000033\cdot 65^5 + 0.0005412\cdot 65^4 - 0.0320106\cdot 65^3\\
&\quad + 0.827084\cdot 65^2 - 7.3286859\cdot 65 + 143.5967802 = 202.550;\\
T_{66} &= -0.0000033\cdot 66^5 + 0.0005412\cdot 66^4 - 0.0320106\cdot 66^3\\
&\quad + 0.827084\cdot 66^2 - 7.3286859\cdot 66 + 143.5967802 = 196.192.
\end{aligned}$$

Seasonal components take on the following values

$$\begin{aligned}
S_{64} &= 5.71\cos\left(\frac{16\pi}{63}\cdot 64\right) + 10.85\sin\left(\frac{16\pi}{63}\cdot 64\right) = 4.562;\\
S_{65} &= 5.71\cos\left(\frac{16\pi}{63}\cdot 65\right) + 10.85\sin\left(\frac{16\pi}{63}\cdot 65\right) = -7.319;\\
S_{66} &= 5.71\cos\left(\frac{16\pi}{63}\cdot 66\right) + 10.85\sin\left(\frac{16\pi}{63}\cdot 66\right) = -1.259.
\end{aligned}$$

TABLE II. Calculation results of harmonic parameters.

Harmonic number k	Period	Frequency	Coefficient a_k	Coefficient b_k	Variance σ_k^2	Statistics F_k	Significance p
0			-2.99		4.47	0.35	0.70
1	63.00	0.02	-3.38	3.08	10.47	0.82	0.43
2	31.50	0.03	-0.29	3.15	5.01	0.39	0.67
3	21.00	0.05	3.07	2.78	8.58	0.67	0.50
4	15.75	0.06	-5.97	-3.10	22.62	1.77	0.17
5	12.60	0.08	-4.77	-1.08	11.94	0.93	0.39
6	10.50	0.10	1.03	-10.95	60.47	4.73	0.01
7	9.00	0.11	-3.20	5.36	19.50	1.53	0.21
8	7.88	0.13	5.71	10.85	75.16	4.79	0.01
9	7.00	0.14	-3.54	1.88	8.02	0.63	0.53
10	6.30	0.16	-0.96	-2.91	4.70	0.37	0.68
11	5.73	0.17	-1.05	-3.21	5.69	0.45	0.63
12	5.25	0.19	-5.00	5.99	30.42	2.38	0.09
13	4.85	0.21	2.92	1.66	5.65	0.44	0.63
14	4.50	0.22	-3.42	4.78	17.27	1.35	0.25
15	4.20	0.24	2.36	2.38	5.62	0.44	0.64
16	3.94	0.25	-2.97	-7.37	31.58	2.47	0.09
17	3.71	0.27	3.86	-5.04	20.17	1.58	0.20
18	3.50	0.29	-2.15	14.36	105.38	8.25	0.00
19	3.32	0.30	6.06	2.27	20.96	1.64	0.19
20	3.15	0.32	-9.94	5.38	63.85	5.00	0.01
21	3.00	0.33	1.76	-0,28	1.58	0.12	0.88
22	2.86	0.35	0.57	0.13	0.17	0.01	0.99
23	2.74	0.37	-3.18	0.90	5.45	0.43	0.64
24	2.63	0.38	0.35	-3.97	7.95	0.62	0.53
25	2.52	0.40	1.30	-2.28	3.44	0.27	0.76
26	2.42	0.41	-6.56	0.56	21.69	1.70	0.18
27	2.33	0.43	9.12	2.27	44.20	3.46	0.06
28	2.25	0.44	2.10	-3.86	9.66	0.76	0.46
29	2.17	0.46	-0.98	-2.01	2.50	0.20	0.82
30	2.10	0.48	-4.88	5.76	28.48	2.23	0.11
31	2.03	0.49	0.16	0.00	0.01	0.00	1.00

Next, we find the cyclic component:

$$C_{64} = 1.03\cos\left(\frac{12\pi}{63}\cdot 64\right) - 10.95\sin\left(\frac{12\pi}{63}\cdot 64\right) - 2.15\cos\left(\frac{36\pi}{63}\cdot 64\right) + 14.36\sin\left(\frac{36\pi}{63}\cdot 64\right)$$
$$-9.94\cos\left(\frac{40\pi}{63}\cdot 64\right) + 5.38\sin\left(\frac{40\pi}{63}\cdot 64\right) = 18.154;$$

$$C_{65} = 1.03\cos\left(\frac{12\pi}{63}\cdot 65\right) - 10.95\sin\left(\frac{12\pi}{63}\cdot 65\right) - 2.15\cos\left(\frac{36\pi}{63}\cdot 65\right) + 14.36\sin\left(\frac{36\pi}{63}\cdot 65\right)$$
$$-9.94\cos\left(\frac{40\pi}{63}\cdot 65\right) + 5.38\sin\left(\frac{40\pi}{63}\cdot 65\right) = -11.570.$$

$$C_{66} = 1.03\cos\left(\frac{12\pi}{63}\cdot 66\right) - 10.95\sin\left(\frac{12\pi}{63}\cdot 66\right) - 2.15\cos\left(\frac{36\pi}{63}\cdot 66\right) + 14.36\sin\left(\frac{36\pi}{63}\cdot 66\right)$$
$$-9.94\cos\left(\frac{40\pi}{63}\cdot 66\right) + 5.38\sin\left(\frac{40\pi}{63}\cdot 66\right) = -34.551.$$

Thus, the predicted values of the expected indicators of passenger traffic for the next 3 days will be 230, 184 and 160 trips, respectively. Actual values of passenger traffic volumes are equal to 209, 191 and 149 trips for the period under review, respectively.

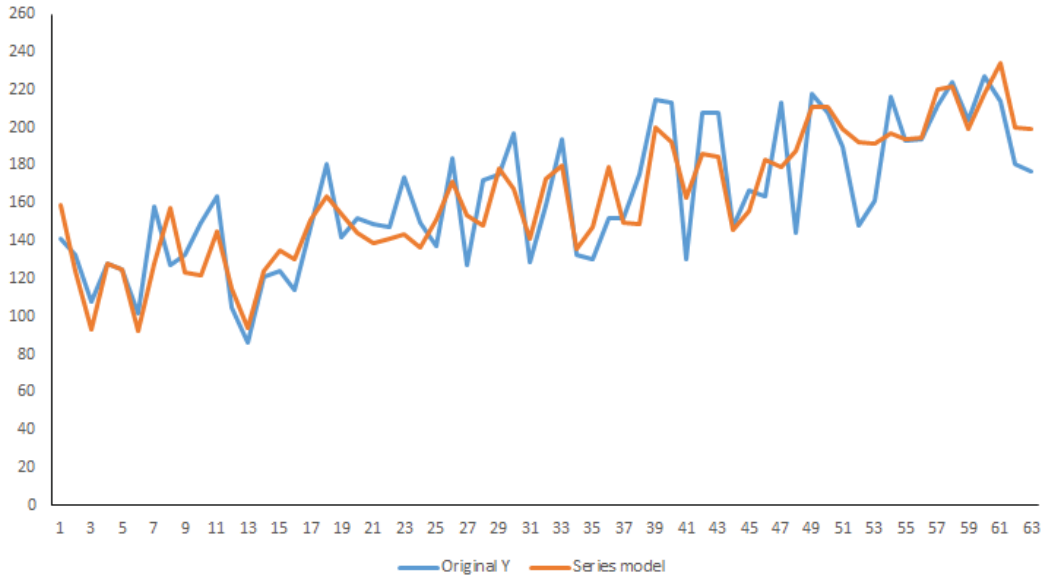

FIGURE 3. Graphs of the original and the model series

Various methods of assessing the quality of a forecast can be applied. We use the weighted absolute percent error (WAPE) as the main error to calculate the prediction accuracy. WAPE is calculated using the formula:

$$WAPE = \frac{\sum |Y_t - \hat{Y}_t|}{\sum Y_t} \cdot 100,$$

where Y_t is the actual value of the time series, and $\hat{Y}_t$ is the predicted value.

Thus, the prediction error is $WAPE = 7.10\%$.

The prediction accuracy is equal to 92.9%. This result shows that the constructed model very accurately describes the analyzed data.

Also, the reliability of the forecast can be estimated using the mean absolute percent error (MAPE):

$$MAPE = \frac{1}{n} \sum_{t=1}^{n} \frac{|Y_t - \hat{Y}_t|}{Y_t} \cdot 100.$$

The prediction error is $MAPE = 7.03\%$.

The prediction accuracy is equal to 92.97%. This result also confirms that the resulting model accurately describes the actual data.

CONCLUSION

The analysis of passenger traffic allows us to identify the main patterns of their fluctuations for the use of results in planning and organizing transportation. The data presented in this paper suggests a rather high accuracy of the constructed model and its suitability for practical use. The obtained results can be used both to improve the organization of passenger transportation on routes and to reorganize the transport network as a whole.

ACKNOWLEDGMENTS

The investigation is supported by the Russian Federal Program in the framework of the project “Optimization of the transport and logistics system based on modeling the development of transport infrastructure and models of consumer preferences” No 121050500050-5.

REFERENCES

1. A. Pompigna and F. Rupi, Comparing practice-ready forecast models for weekly and monthly fluctuations of average daily traffic and enhancing accuracy by weighting methods, Journal of Traffic and Transportation Engineering (2018), DOI:10.1016/j.jtte.2018.01.002.
2. O. Ie, Simulation modeling of transport systems: software and directions for their improvement, Topical Issues of Modern Economy **5**, 428–439 (2020).
3. A. Martynenko and O. Ie, Modeling the natural development of the intercity road network, Herald of the Ural State University of Railway Transport **2**(46), 4–12 (2020).
4. G. Timofeeva and O. Ie, Application of a synthetic gravity model with exponential-power function of gravity for calculating the splitting of passenger traffic by different types of public transport, Transport of the Urals **4**(67), 3–9 (2020).
5. G. Timofeeva and O. Ie, "Evaluation of origin-destination matrices based on analysis of data on transport passenger flows," in *AMEE'20*, AIP Conference Proceedings 2333 (American Institute of Physics, Melville, NY, 2021), paper 100002, 6p., DOI:10.1063/5.0041801.
6. O. Ie, Analysis of the temporal distribution of passenger traffic in road transport based on ride-sharing data, Transport of the Urals **2**(69), 10–17 (2021).
7. Yu, Bai, C. Guanhua, and Yan, Analysis of space-time variation of passenger flow and commuting characteristics of residents using card data of nanjing metro, Sustainability **11**, 4989 (2019), DOI:10.3390/su11184989.
8. T. Andersen, *The Statistical Analysis of Time Series* (Wiley-Interscience, Hoboken, NJ, 1994).
9. M. Jakubauskas, D. Legates, and J. Kastens, Harmonic analysis of time-series avhrr ndvi data, Photogrammetric Engineering and Remote Sensing **67**, 461–470 (2001).
10. P. Mohan, A. Mangalam, and S. Chattopadhyay, Parametric models of periodogram, Journal of Astrophysics and Astronomy **35** (2014), DOI:10.1007/s12036-014-9240-x.
11. T. Cipra, Statistical analysis of time series, 361–385 (2010), DOI:10.1007/978-3-7908-2593-0-31.
12. A. Rahmatulloh and R. Gunawan, Web scraping with html dom method for data collection of scientific articles from google scholar, Indonesian Journal of Information Systems **2**, 16 (2020), DOI:10.24002/ijis.v2i2.3029.