REGRESSION MODEL IN ROAD TRANSPORT SERVICES

Szymon Mitkow
Military University of Technology, Poland
E-mail: szymon.mitkow@wat.edu.pl

Andrzej Świderski
Motor Transport Institute, Poland
E-mail: andrzej.swiderski@its.waw.pl

Received: April 28, 2019
Received revised: August 20, 2019
Accepted for publishing: August 29, 2019

Abstract

The success of a company depends on a number of factors. One of them is the ability to meet customer expectations and match the market needs. Mathematical methods and tools are helpful in assessing demand. The forecasts made should take into account all the factors shaping the demand for goods and services, however, they are often difficult to define, not only because of their large number, but also because of the impact of individual variables which is difficult to determine. In many cases, the number of orders placed is strongly dependent on the time at which they are placed. Needs may vary depending on the time of day, week or year. Then we are dealing with the so-called seasonality, which is a very important matter that needs be taken into account in a company which allows to better adapt the company's activities to customer requirements.

This article describes the seasonality of demand in a company providing domestic road transport services using heavy-duty vehicles. The legitimacy of conducting such analyses and potential benefits were indicated.

Key words: demand forecasting, seasonality, road transport, multiple regression

1. INTRODUCTION

The transport of cargo and people are one of the main needs of today's world. The dynamic growth of industrial production and trade results in the constant growth of the demand for transport services. The transport industry plays a very important role in the market economy, ensuring efficient and effective functioning of all of its elements. Complexity and volatility of transport demand poses a challenge for companies providing such services (Borucka, 2018). The problem lies mainly in the multiplicity of factors influencing demand. These factors shape the demand for transport services (Dittmann, 2000; Mitkow et al., 2018) and affect quality thereof, which translates directly into customer relations (Borucka, 2018). Transport demand determinants are often subject to cyclical, repetitive changes, making it possible not

only to predict them, but also to prepare for them. Mathematical tools and methods come in handy in this respect. The application of the selected one shall be presented in this article.

The method of demand analysis was presented basing on transport data provided by the transport company. Multiple regression, which allows to take into account many exogenous variables, was used for this purpose. The forecast was made on the basis of observations of the demand for transport services made in the last 39 months. The aim of the analysis was to present the effectiveness of the use of mathematical tools that are already successfully used in modeling the vehicle flow (Mitkow et al., 2018), their readiness (Borucka, 2018) or in assessing the impact of selected factors on their efficiency and effectiveness (Świderski et al., 2018).

2. RESEARCH METHOD

Describing the variability of occurring phenomena and processes is possible due to the use of stochastic process models. If time is the domain of the stochastic process, then we are dealing with a time series, i.e. a sequence of information ordered in time. Individual observations are recorded with a precise time step. Then the measurements are a set of observations describing the implementation of the analyzed phenomenon and the changes occurring in it. Full identification of the process requires decomposition of the time series, i.e. extraction of all elements present in it. They can be systematic components such as trends, periodic or cyclical fluctuations as well as random components. A mathematical model is selected on the basis of the dependencies between the diagnosed observations (Bielińska, 2007; Dittmann, 2000). In the presented case, it will be a multiple regression.

Regression models are widely used in many fields of science (Chiou et al., 2015; Pupavac, 2018), including transport. They deal with a number of issues related to it, particularly with regard to the sustainable development of transport services. Examples of this include assessment and forecasting of noise pollution (Gulliver et al., 2015; Dintrans & Préndez, 2013), carbon dioxide emissions (Xu & Lin, 2015; Asumadu-Sarkodie & Owusu 2017; Xie et al., 2017), or the level of energy consumption in transport (Hu et al. 2010; Chai et al., 2016). The literature also contains research on driver behavior (Singh et al., 2019), the use of public transport (Chiou et al., 2015), public transport buses arrival time (Singh et al., 2019; Yu et al., 2017, Bai et al., 2015), etc. In Huang et al. (2017) and Aguero-Valverde et al. (2016) for example, multidimensional spatial models for analyzing the occurrence of road accidents were proposed. Regression models work well in all cases where we are dealing with factors that significantly affect the analyzed phenomenon. The review of literature and empirical data gathered determined the decision to use this model in the present article.

Regression belongs to the group of analytical models that require finding mathematical functional dependencies that reflect the implementation of the process. Its main purpose is to forecast, i.e. to extend the analysis to arguments outside the scope of collected empirical data, and to determine how the studied phenomenon will develop in the future (Statsoft, 2006; Sokołowski, 2016). The estimation

of parameters consists in finding the appropriate trend function, and then describing and extracting seasonal and cyclical fluctuations (if there are any). The simplest form of a trend is the linear function (1), describing the implementation of the process of development of the analyzed phenomenon by means of an exogenous variable, in this case – time t , and a directional coefficient β, which determines the constant growth of the predicted variable in the unit of time (1).

$$f_t = \beta_0 + \beta_1 t \quad (1)$$

Depending on the structure of the studied phenomenon, a trend can also be estimated using other functions, for example exponential, quadratic or power functions. If there are fluctuations in the process, the trend function shall not be sufficient enough (Maciag at al., 2013; Mitkow at al., 2018).

Fluctuations may occur due to a number of factors. They are determined by the repetitive rhythm of the phenomena: daily, weekly, monthly, etc. thus these are seasonal fluctuations, or they shape long-term, rhythmic fluctuations of the value of the series around the developmental trend, i.e. cyclical fluctuations. To estimate the parameters of such a model the classical least squares method or the maximum likelihood estimate can be used.

Regression enables assigning the value of a variable dependent to specific values of independent variables in an analytical way. Depending on the need, it can take different forms. Its simplest type is simple linear regression, describing the dependencies between variables using a straight line (2).

$$\hat{y} = \beta_0 + \beta_1 x + \varepsilon \quad (2)$$

where:

β_1 – directional coefficient,

β_0 – absolute term (point of intersection with the axis of ordinates),

x – independent variable,

y – dependent variable (endogenous, predicted),

ε – random error.

If there are more exogenous variables, the multiple regression model may be used, which takes the form (3):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon \quad (3)$$

where:

β_0 – absolute term,

β_i – model parameters – regression coefficients

x – independent variable,

y – dependent variable (endogenous, predicted),

ε – random error.

Regression factors describe how, on average, the value of the dependent variable y will change if the value of the independent variable x, to which they refer, will change by a unit, assuming a fixed level of the other independent variables (Mitkow et al., 2018; Sokołowski, 2016).

The situation is quite simple if the independent variables are of a quantitative nature. However, often in the analysis of economic phenomena it is not the case, and variables – as in the analyzed example – are of a qualitative nature. Then their number is limited and they cannot be treated in the way accepted for continuous variables in regression, as they have no economic sense, and the calculated model coefficients

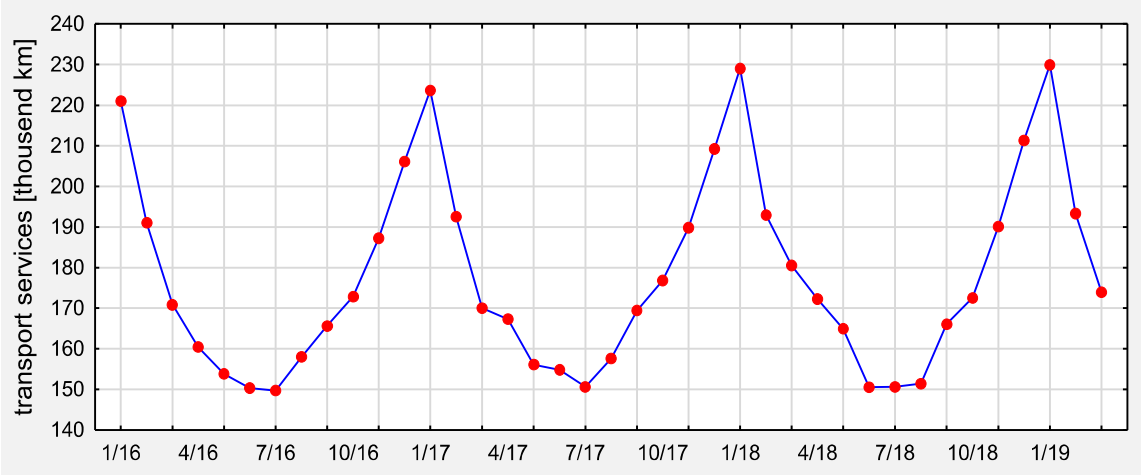
have no economic interpretation. In order to create a regression model, it is necessary to re-code them. Such a qualitative or discrete variable, having several or more categories, is coded to the appropriate q number of binary (zero-one) variables, which are used in the regression equation. However, in order to be able to apply the least squares method to model estimation it is necessary to use q-1 of artificial variables, because the introduction of q number of variables, i.e. in the number equal to exogenous variables, will provoke a linear dependence between the regressors, and the $X'X$ matrix will be singular (Mitkow et al., 2018). This is due to the fact that binary variables sum up to unit. Such a phenomenon is described in econometric literature as a *dummy variable trap* related to binary variables (Bielińska, 2000; Mitkow et al., 2018). Such a model is not estimable and therefore the number of artificial variables must always be one unit less than the q number of categories (levels) identified for a given attribute (feature). Only then is the estimated model correct and consists of an absolute term β_0, the sum of the products of the structural parameters and the binary variables Dk number k=1;;q, representing seasonality, and a random component (4).

$$y = \beta_0 + \beta_1 t + \delta_1 D_1 + \cdots + \delta_k D_k + \varepsilon \quad (4)$$

3. TEST SUBJECT

The subject of the analysis was the company's transport operations over the last 39 months. The company owns 23 semi-trucks. The course of the transport operations is shown in Fig. 1.

Figure 4. Demand for transport services in 2016-2019, in thousands of kilometers



Source: own study

The chart clearly shows a high seasonality of transport operations. This is confirmed by the calculated, selected measures of descriptive statistics presented in the table below (tab. 1).

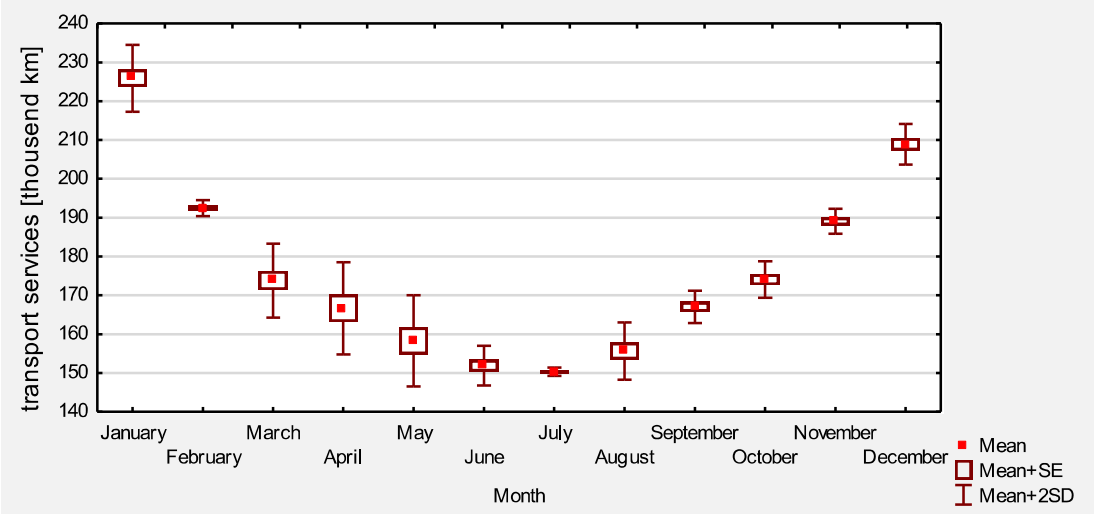
Table 1. Basic measures of descriptive statistics in individual months

Month	Mean [Thousands of km.]	Median [Thousands of km.]	Minimum [Thousands of km.]	Maximum [Thousands of km.]	Standard deviation [Thousands of km.]	Coefficient of variation [%].
January	225.9	226.3	221.0	230.0	4.3	1.9
February	192.5	192.7	191.0	193.3	1.0	0.5
March	173.8	172.4	170.0	180.5	4.8	2.7
April	166.7	167.3	160.4	172.2	6.0	3.6
May	158.3	156.1	153.8	165.0	5.9	3.7
June	151.9	150.6	150.3	154.8	2.6	1.7
July	150.4	150.7	149.7	150.7	0.6	0.4
August	155.7	157.6	151.4	158.0	3.7	2.4
September	167.1	166.1	165.7	169.4	2.1	1.2
October	174.1	172.9	172.5	176.8	2.4	1.4
November	189.1	189.8	187.2	190.1	1.6	0.9
December	208.9	209.2	206.2	211.3	2.6	1.3

Source: own study

The specificity of the company's activity causes that the greatest transport needs occur in the winter months, from November to March. The lowest results were recorded in the warm months, i.e. in the period from May to August. Clear differences between individual months are well illustrated by the frame graph in figure 2.

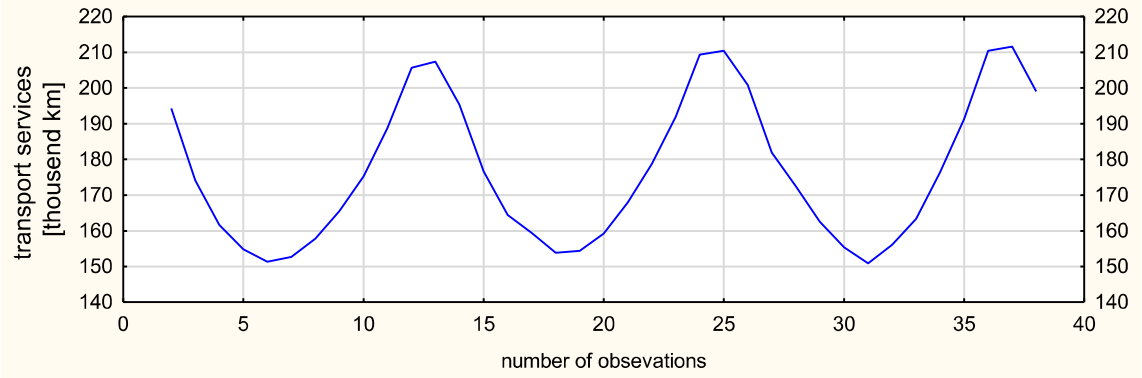
Figure 5. Average transport needs in individual months of the analyzed period



Source: own study

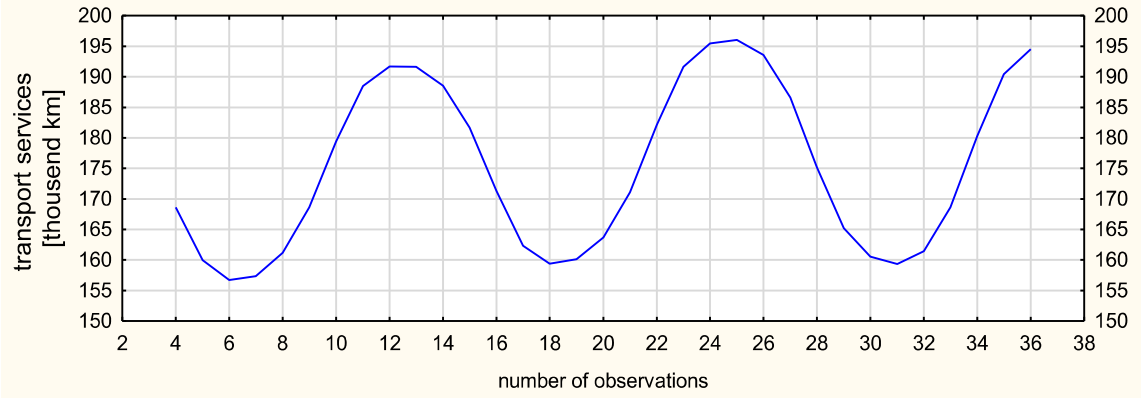
The strong dependence between the services provided and the calendar date should be taken into account in the mathematical model of demand. The graph of transport needs in the studied period (Fig. 1) also suggests the occurrence of a development trend, however, seasonal fluctuations make it somewhat difficult to distinguish it visually. In order to confirm the occurrence of a trend, a mechanical method of determining the trend using moving averages was used. Simple moving averages (3-period, 6-period and 12-period) were determined, the graphs of which are shown in Fig. 3, 4 and 5.

Figure 6. 3-period moving averages



Source: own study

Figure 7. 6-period moving averages

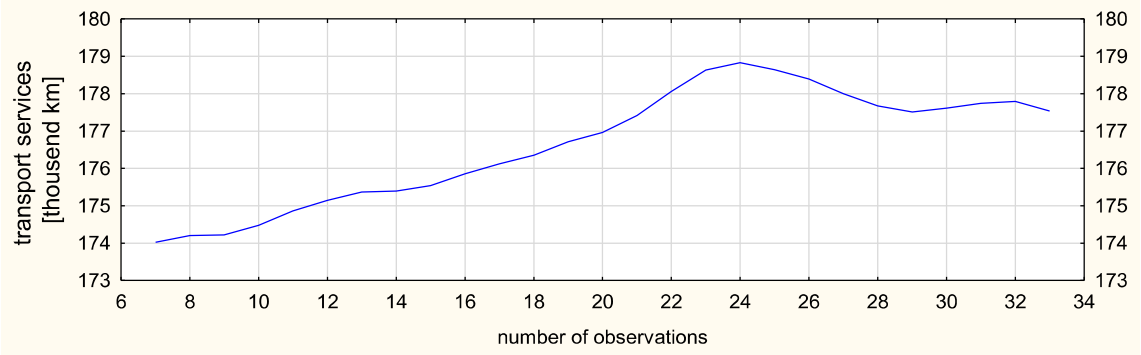


Source: own study

The graph of 6-period moving averages shows the existence of a trend and its increasing character, which is confirmed by the graph of 12-period moving averages. This means that the mathematical model should include long-term trend and short-term seasonality for individual months, expressed in binary variables in a number of one less than the number of months. Thus, the model will consist of the absolute term, the trend, and the sum of eleven products of structural parameters and binary variables D_k for $k\epsilon\{1, 11\}$ seasonality, as presented in formula (5).

$$y = a_0 + a_1 t + b_1 D_1 + \cdots + b_k D_k \tag{5}$$

Figure 8. 12-period moving averages



Source: own study

The literature proposes the estimation of the model reduced by the variable with the lowest or highest indication. Then the remaining variables refer to the maximum or minimum level of the studied phenomenon. In the analyzed case, the highest value recorded in January was selected. This means that the seasonal parameter estimates will refer to the "January" level. Therefore, all parameter values for the individual months will be negative, as each parameter will be lower than the value for January. The results of estimation of parameters of multiple regression function and errors in estimation are presented in Table 2.

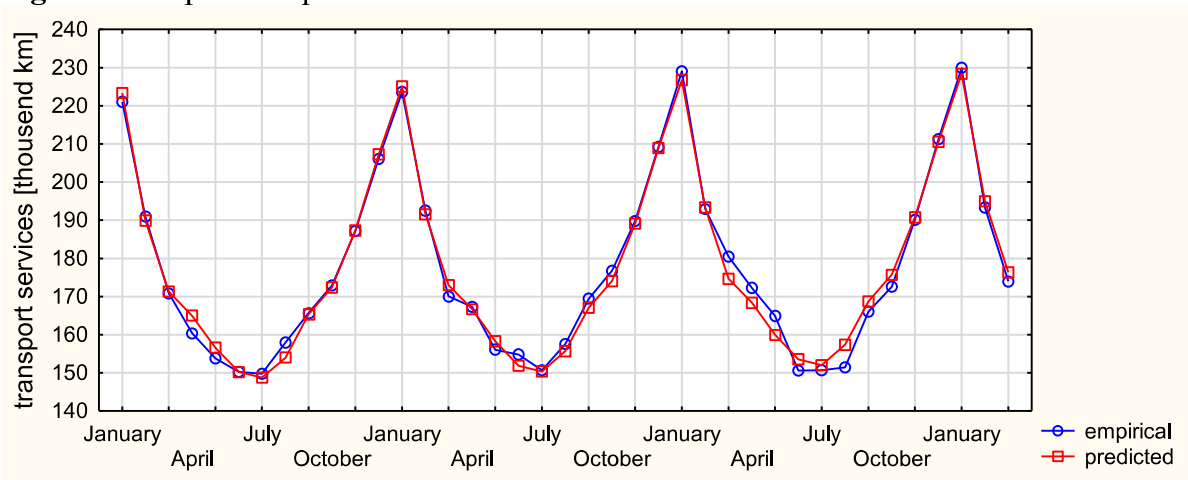
Table 2. Results of the estimation of the multiple regression model

	R= 0.99, R^2= 0.98 Adjusted R2= 0.98 F(12,26)=182,08, p<0,0000 SE: 3.11			
	b	Std. error of b	t(26)	p
Absolute term	223.25	1.78	125.53	0.00
t	0.14	0.05	3.10	0.00
February	-33.60	2.20	-15.25	0.00
March	-52.35	2.20	-23.75	0.00
April	-58.84	2.38	-24.69	0.00
May	-67.36	2.38	-28.29	0.00
June	-73.88	2.38	-31.05	0.00
July	-75.53	2.38	-31.75	0.00
August	-70.38	2.38	-29.58	0.00
September	-59.14	2.38	-24.84	0.00
October	-52.27	2.38	-21.93	0.00
November	-37.42	2.39	-15.68	0.00
December	-17.71	2.39	-7.41	0.00

Source: own study

The adjusted coefficient of determination, which determines what percentage of the variation of the dependent variable (Y - endogenous) is explained by the independent variable (X - exogenous) is satisfactory and amounts to 98%, errors in parameter estimation are low and do not exceed 3%. Moreover, the quality of the model is evidenced by the fact that all estimated parameters are statistically significant, which is also satisfactory due to the substantive interpretation of the model and the lack of need to remove insignificant exogenous variables. The graph of empirical and forecast values is presented in Fig. 6.

Figure 9. Graph of empirical and forecast values



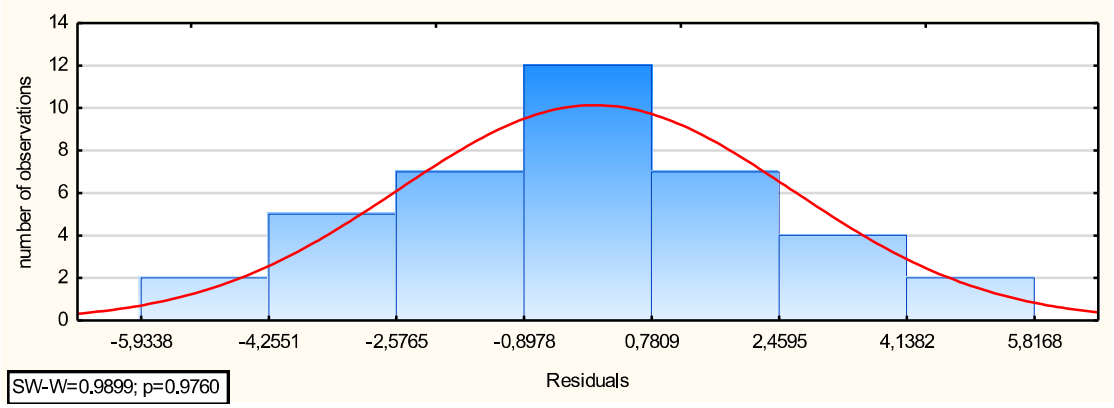
Source: own study

4. MODEL DIAGNOSTICS

The basis for assessing the accuracy of matching the theoretical function to the empirical data is the analysis of the differences between the empirical and theoretical values known as the residuals of the model. Confirmation of the correctness of the regression model therefore requires verifying the basic assumptions concerning the residuals, which include the checking the normality of their distribution and the existence of significant dependencies of the autocorrelation function.

Figure 7 shows the histogram of the distribution of residuals which indicates that the distribution is close to normal, as confirmed by the calculated statistics of the Shapiro-Wilk test, the calculated value of which is 0.9899 and the test probability is p=0.976, which means that there are no grounds to reject the H0 hypothesis at the level of significance alpha=0.05, indicating that the distribution of the variable is close to normal.

Figure 10. Histogram of the residuals of the model

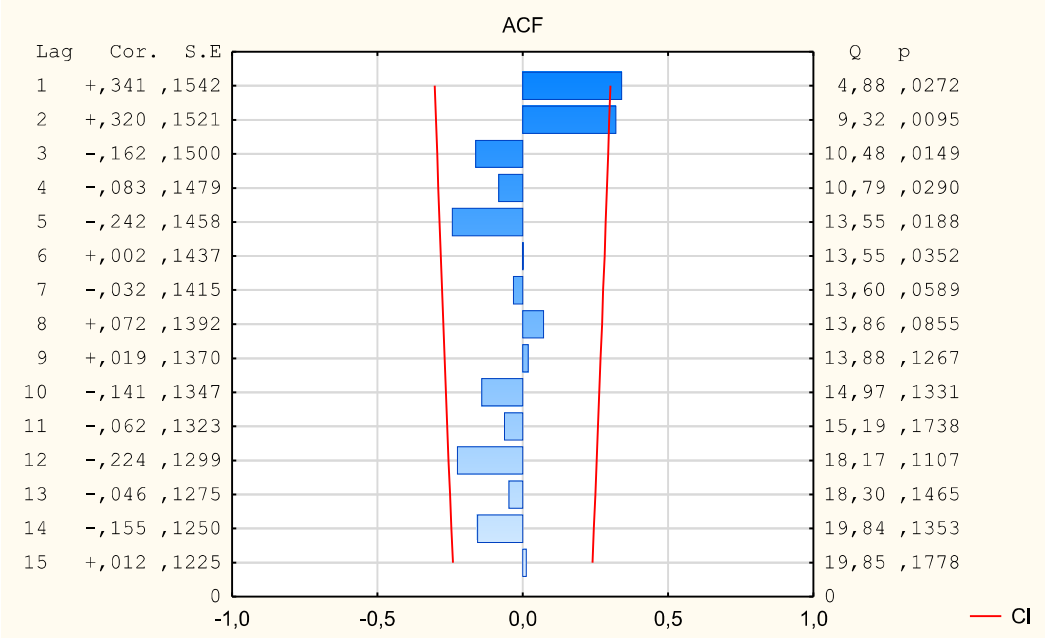


Source: own study

Another factor determining the correctness of the model is the lack of correlation between all the residuals' values, i.e. the confirmation that there are dependencies which were unexplained by the model. For this purpose, the graphs of autocorrelation and partial autocorrelation function, presented in Fig. 8 and Fig. 9 were analyzed.

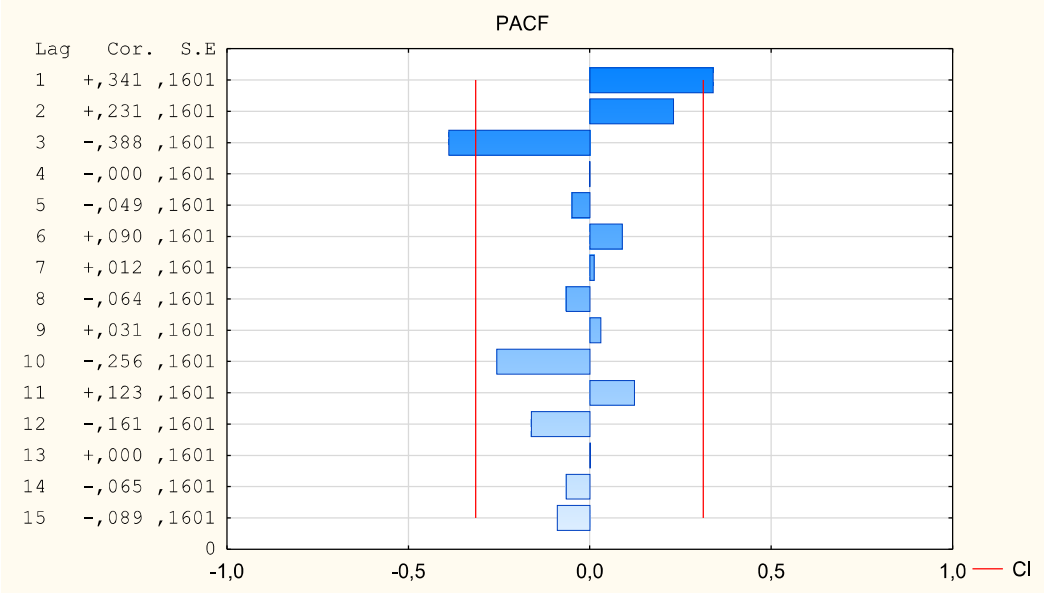
It turns out that all values of functions are statistically insignificant, which confirms the correctness of the model's construction.

Figure 11. Autocorrelation function of the regression model residuals



Source: own study

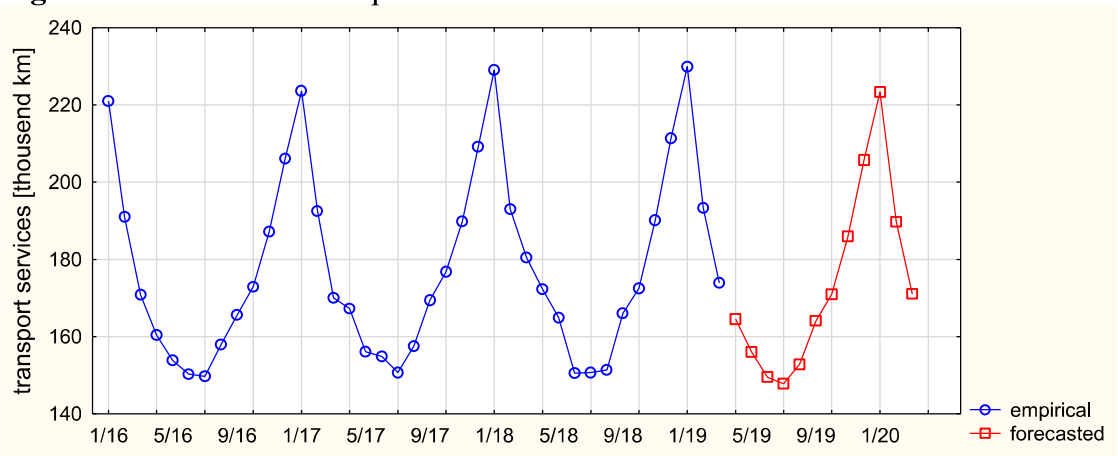
Figure 12. Partial autocorrelation function of the regression model residuals



Source: own study

On the basis of the results obtained, a forecast of transport needs for 2019 was also proposed (fig. 10).

Figure 13. Forecast of transport needs for 2019



Source: own study

After analyzing of the graph in Fig. 10 it seems that the forecast for January is underestimated. However, it should be stressed that each forecast requires monitoring, verification and continuous adjustment to the dynamically changing market. In addition, the forecast plays only an advisory function, allowing to indicate the main directions of change and provides excellent support for management processes. However, it does not provide ready-made decisions.

5. CONCLUSIONS

The analysis showed the possibility of forecasting the transport demand in a company providing transport services. The proposed model revealed not only a clear seasonality of the process, but also the existing upward trend. This information is very important, especially for a company, which, similarly to the studied one, operates its own transport. Firstly, it indicates that there are months in which the transport potential may not be fully utilized. On the other hand, it warns that steadily growing demand may in the near future be the cause of a shortage of vehicles and a failure to fully meet transport needs, especially in the months when these are the highest. This means that it is necessary to review the transport strategy of the company and propose directions that on the one hand, minimize the degree of underutilization of vehicles and on the other hand, protect the interests of the company in the event of demand greater than the transport capacity.

The analysis was limited to the assessment and forecasting of the impact of the calendar month on transport demand. Despite the fact that it provided interesting conclusions concerning the company's operations, which strongly emphasizes its utilitarian character, the demand is influenced by many other factors, which were not included in this study. Therefore, other factors influencing demand, such as market competition, the level of prices and quality of transport services, size and structure of demand for transport of a given type of cargo, impact of the economic situation, should also be analyzed in future studies. This would increase the scale and adaptability of the company's offer to market demand by adjusting transport capacity and planning (if necessary) investment measures, the need for which has already been partially articulated in the study presented.

6. REFERENCES

Aguero-Valverde, J. & Wu, K-F. K. & Donnell, E. T. (2016). A multivariate spatial crash frequency model for identifying sites with promise based on crash types, *Accident Analysis & Prevention*, 87, p. 8-16.

Asumadu-Sarkodie, S. & Owusu, P.A. (2017). The impact of energy, agriculture, macroeconomic and human-induced indicators on environmental pollution: evidence from Ghana, *Environmental Science and Pollution Research,* 24(7), p. 6622–6633.

Bai, C.& Peng, Z-R.& Lu, Q-C. & Sun, J. (2015). Dynamic bus travel time prediction models on road with multiple bus routes. *Computational Intelligence and Neuroscience*, 2015, p. 1-9.

Bielińska, E. (2007). *Prognozowanie ciągów czasowych*. Wydawnictwo Politechniki Śląskiej, Gliwice.

Bitner A. (2007). Konstrukcja modelu regresji wielorakiej przy wycenie nieruchomości, *Acta Scientiarum Polonorum. Administratio Locorum*, 4(6), p. 59-66.

Borucka A. (2018). Risk Analysis of Accidents in Poland Based on ARIMA Model, *Transport Means 2018, Proceedings of the 22nd International Scientific Conference part I, Lithuania*, , p. 162-166.

Borucka A. (2018). Three-state Markov model of using transport means, *Proceedings of the 18th International Scientific Conference, Business Logistics In Modern Management*, Croatia, p. 3-19.

Chai, J. & Lu, Q-Y, Wang, S-Y. & Lai, K.K. (2016). Analysis of road transportation energy consumption demand in China, *Transportation Research Part D: Transport and Environment*, 48, p. 112-124.

Chiou, Y-C. & Jou, R-C. & Yang, C-H. (2015). Factors affecting public transportation usage rate: Geographically weighted regression, *Transportation Research Part A: Policy and Practice*, 78, p. 161-177.

Dintrans, A. & Préndez, M. (2013). A method of assessing measures to reduce road traffic noise: A case study in Santiago, Chile, *Applied Acoustics*, 74(12), p. 1486-1491.

Dittmann P. (2000). *Metody prognozowania sprzedaży w przedsiębiorstwie* Wydawnictwo Akademii Ekonomicznej, Wrocław.

Gulliver, J.& Morley, D. & Vienneau, D. & Fabbri, F. & Bell, M. & Goodman, P. & Beevers, S.&, Dajnak, D. & Kelly, F. J.& Fecht, D. (2015). Development of an open-source road traffic noise model for exposure assessment, *Environmental Modelling & Software*, 74, p. 183-193.

Hu, X. & Chang, S. & Li, J. & Qin,Y. (2010). Energy for sustainable road transportation in China: challenges, initiatives and policy implications**,** *Energy*, 35(11), p. 4289-4301.

Huang, H. & Zhou, H. &Wang, J. & Chang, F. & Ma, M. (2017). A multivariate spatial model of crash frequency by transportation modes for urban intersections, *Analytic Methods in Accident Research*, 14, p. 10-21.

Maciąg A., Pietroń R., Kukla S. (2013). *Prognozowanie i symulacja w przedsiębiorstwie*, Polskie Wydawnictwo Ekonomiczne, Warszawa.

Mitkow Sz., Borucka A. (2018). Mathematical model of travel times related to a transport congestion: an example of the capital city of Poland – Warsaw, *Proceedings of the 18th International Scientific Conference, Business Logistics in Modern Management*, Croatia, p. 501-526.

Pupavac, D. (2018). Employment Analysis In The Logistics Sector Of The Republic Of Croatia, Business Logistics in Modern Management, Josip Juraj Strossmayer University of Osijek, Faculty of Economics, Croatia, 18, p. 617-626.

Singh, G. & Bansal, D. & Sofat, S. (2019). Communication Assisted Dynamic Scheduling of Public Transportation Systems. In: Hemanth J., Fernando X., Lafata P., Baig Z. (eds) *International Conference on Intelligent Data Communication*

Technologies and Internet of Things (ICICI) 2018. ICICI 2018. Lecture Notes on Data Engineering and Communications Technologies, 26, Springer, Cham.

Sokołowski, A. (2016). *Prognozowanie i analiza szeregów czasowych*. Materiały szkoleniowe, StatSoft Polska, Kraków.

StatSoft, *StatSoft Electronic Statistics Textbook*, Kraków, 2006 [http://www.statsoft.pl/textbook/stathome.html]

Świderski A., Borucka A., Jacyna-Gołda I., Szczepański E. (2019). Wear of brake system components in various operating conditions of vehicle in the transport company, *Eksploatacja i Niezawodność – Maintenance and Reliability*, 1(21), p. 1-9.

Xie, R. & Fang, J. & Liu, C. (2017). The effects of transportation infrastructure on urban carbon emissions, *Applied Energy*, 196, p. 199-207.

Xu, B. & Lin, B. (2015). Factors affecting carbon dioxide (CO2) emissions in China's transport sector: a dynamic nonparametric additive regression model, *Journal of Cleaner Production*, 101, p. 311-322.

Yu, H.& Wu, Z.& Chen, D.& Ma, X. (2017). Probabilistic Prediction of Bus Headway Using Relevance Vector Machine Regression, *IEEE Transactions on Intelligent Transportation Systems*, 18(7), p. 1772-1781.