SILESIAN UNIVERSITY OF TECHNOLOGY PUBLISHING HOUSE

SCIENTIFIC PAPERS OF SILESIAN UNIVERSITY OF TECHNOLOGY 2020
ORGANIZATION AND MANAGEMENT SERIES NO. 142

FORECASTING DEMAND FOR PRODUCTS IN DISTRIBUTION NETWORKS USING R SOFTWARE

Maciej WOLNY[1*], Mariusz KMIECIK[2]

[1] Silesian University of Technology, Faculty of Organization and Management, Zabrze; Maciej.Wolny@polsl.pl,
ORCID: 0000-0002-8872-7794
[2] Silesian University of Technology, Faculty of Organization and Management, Zabrze;
Mariusz.Kmiecik@polsl.pl, ORCID: 0000-0003-2015-1132
* Correspondence author

Purpose: This article addresses the issue of forecasting demand for products flowing in a distribution network conducted from the perspective of a 3PL logistics operator. Its purpose is to present a tool based on the R software, which is to be used for automatic forecasting of time series.
Design/methodology/approach: The work uses the algorithm of automatic forecasting of time series implemented in the "forecast" package. The algorithm was used in a loop for different lengths of the time series to determine the best length of the series. The minimal RMSE value in the training set and in the test set were considered as optimality criteria.
Findings: It is shown that the best time series length is 60 weeks in the considered case.
Originality/value: The procedure for selecting the best time window length for forecasting demand for products in distribution network.

Keywords: logistics operator, forecasting, R software, distribution network.

Category of the paper: Research paper.

1. Introduction

The objective of cooperation in distribution networks is to effectively provide customers with goods and services. Initially, enterprises did not have a strong need to strengthen cooperation with their contractors, while relationships were limited to ordinary transactions. At present, this situation is changing and enterprises are striving to integrate their activities to meet the needs of final customers. This generates the need to create flexible and dynamic market systems. One of the factors that affects the shape of the distribution network are fluctuations in demand for products that are the subject of flow, as well as the growing importance of outsourcing and logistics operators who operate in these networks. The question arises whether

http://dx.doi.org/10.29119/1641-3466.2020.142.8

the logistics operator can make forecasts in the distribution network? The purpose of this article is to present a tool based on the R software for automatic forecasting of time series used from the perspective of the 3PL logistics operator.

R software (https://www.r-project.org/) is a free software environment for statistical calculations, visualization and data analysis. A number of tools in the form of packages have been implemented in this software. One such tool is the "forecasting" package (Hyndman, and Athanasopoulos, 2018). The study uses the functionality of this package in the field of automation of forecasting demand for products in distribution networks.

Forecasting is considered one of key elements of organization management (Bendkowski et al., 2010). It affects, among others, the determination of production capacities, as well as methods of producing and providing services, and thus also affects indirectly elements, such as the number of employees or the level of costs. The need for forecasts in an enterprise arises for two main reasons (Dittman, 2000): uncertainty about the future and delay in the time between the moment the decision is made and its effects. One of the most commonly used forecasting methods for forecasting demand are methods based on time series models. It is assumed that the described stochastic processes are generated by themselves (Grzelak, 2019). Among the models that are used to describe the mentioned processes, special place is occupied by models based on auto-correlation integrating the autoregressive and moving average model, i.e. ARIMA (AutoRegressive Integrated Moving Average).

The use of appropriately selected, flexible and precise forecasting procedures gives the opportunity to obtain good results even in markets where the practice of ordering through intermediaries distorts the demand of end participants (Bendkowski et al., 2010). The environment of the enterprise itself, as well as the appearance and structure of the distribution network also affect the accuracy of the posed forecasts. The structure of the distribution network is conditioned, among others (Kramarz, M., and Kramarz, W., 2012) by: demand characteristics, competition strategy, level of relations, as well as logistics strategy. Forecasting demand is important for the supplier, producer and seller. Forecasts imply the processes and operations that they trigger. It is necessary to plan future demand. Properly adopted forecasts, as well as correctly defined time horizons of inventory management allow to optimally match supplies, eliminate deficiencies and reduce pallet space in the warehouse (Wojciechowski, A., and Wojciechowska, N., 2015). Demand forecasts are necessary for the implementation of basic operational processes.

Another very important issue in distribution networks is the frequent use of outsourcing and, in the case of logistics, focusing on contract logistics. This logistics involves entrusting specialized external entities with some additional services. Logistics, in addition to activities related to IT, human resources, accounting and finance is an area often outsourced to other entities. The use of operators' services, despite their long existence on the market, is still one of the market trends. 3PL (Third-Party Logistics) operators are one of the most common forms of operators in current markets. There are 3 main groups of 3PL operators (Żebrucki, 2012):

operators with (or renting) material resources, based on the use of fixed assets, network operators developed thanks to global communication and transport networks (their advantages are mainly information services) and operators building position based on skills and not possessing a logistics infrastructure (similar in nature to 4PL operators, but less complex).

All types of operators provide various management and other activities, which is why it is generally believed that they are entities that would be able to fulfill the leading role in distribution networks (Kawa, 2011). There is also a presumption that logistics operators, as a central link in the distribution network, would be able to take over the task of forecasting, and thus take over from their customers another category of risk in material flows.

Forecasting demand in distribution networks usually refers to a large number of products, often with differing amounts of rotation. The prognostic task may also refer to different time intervals, i.e. it may relate to forecasting the daily size of product releases or weekly volumes. It mainly depends on the planning needs regarding inventory and orders.

The proposed approach and methodology refer to products, for which historical distribution sizes of demand are known. The results presented in the study relate to a typical product, whose distribution sizes of releases is characteristic for the largest group of products distributed by the examined logistics operator. Therefore, the research was also aimed at presenting the most important features of the time series.

2. Proposed methodology

The proposed approach is presented in the example of a typical product operated by one of 3PL logistics operators. This product, during the implementation of the order by the logistics operator, goes to distribution centers or directly to the customer's points of sales (POS). The customer commissioning the fulfillment of logistics activities is a production enterprise, whose main assortment is household chemicals, which includes the analyzed time series. This series applies to SKU (Stock Keeping Units) releases belonging to the household chemicals group. This group accounts for 24.78% of the entire assortment, with a total of over 1,000 SKUs, and is thus the largest manufacturer's assortment group offered (the second group in order is 20.43%, the third is 15.43%, the fourth in order is 9.78% and all others less than 5%). The time series considered is a typical series representing the weekly releases of the above mentioned largest assortment group. The history of releases dates back to 2014 (285 observations).

R software and the "forecasting" package were used to analyze data in the form of time series. The main modeling automation algorithm has been implemented in the auto.arima function (Hyndman, and Khandakar, 2008). The Hyndman-Khandakar algorithm is used to automatically determine the parameters of the ARIMA model using the Akaike information criterion (Akaike, 1974) and Bayes information criterion (Schwarz, and Gideon, 1978).

When making the forecast for subsequent periods, the algorithm procedure presented in Figure 1 was used.

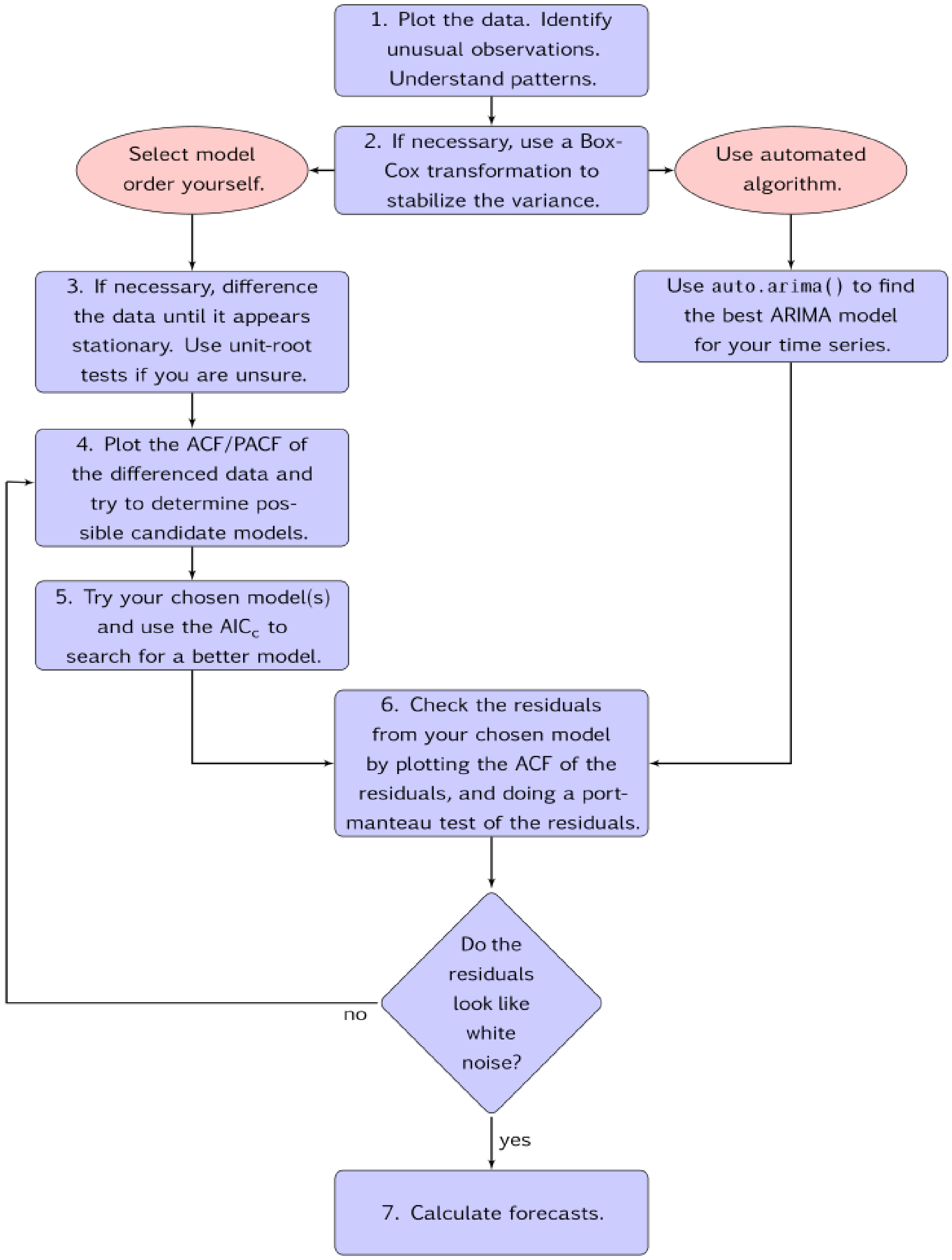
1. Plot the data. Identify unusual observations. Understand patterns.
2. If necessary, use a Box-Cox transformation to stabilize the variance.
Select model order yourself.
Use automated algorithm.
3. If necessary, difference the data until it appears stationary. Use unit-root tests if you are unsure.
Use auto.arima() to find the best ARIMA model for your time series.
4. Plot the ACF/PACF of the differenced data and try to determine possible candidate models.
5. Try your chosen model(s) and use the AIC$_c$ to search for a better model.
6. Check the residuals from your chosen model by plotting the ACF of the residuals, and doing a portmanteau test of the residuals.
Do the residuals look like white noise?
no
yes
7. Calculate forecasts.

Figure 1. General process for forecasting using an ARIMA model. Adapted from: (Hyndman, and Athanasopoulos, 2018).

The study uses a part related to the automatic construction of the forecast. The raw data analyzed included the weekly volume of releases of a typical household chemicals product between 01/01/2014 and 30/06/2019. At the same time, various lengths of the time series were considered in terms of their information usefulness for forecasting purposes in order to characterize the impact of the length of the series on the quality of forecasts. Validation of the prognostic model was carried out by dividing the time series into a teaching and testing part. Matches were researched in both parts. The time window of the test part was fixed at four weeks. The results of the tests and analyses carried out are presented in subsequent sections, in accordance with the adopted algorithm. In the first section, statistical analyses and visualizations of the series were performed to understand the data and detect patterns, including the homogeneity of variance using the Fligner-Killeen test (Fligner, Kileen, 1976). Next, automatic forecast package tools were used in the R software to build the model (auto.arima). Then the model was verified by examining the distribution of residuals, mainly for the occurrence of auto-correlation. In addition, the 10 best models were selected in both sets, based on different lengths of test and teaching sets in terms of minimum RMSE values (Root Mean Squared Error) and a visualization of the presented forecast was presented.

3. Research results

Distributions of weekly releases are presented in Figure 2.

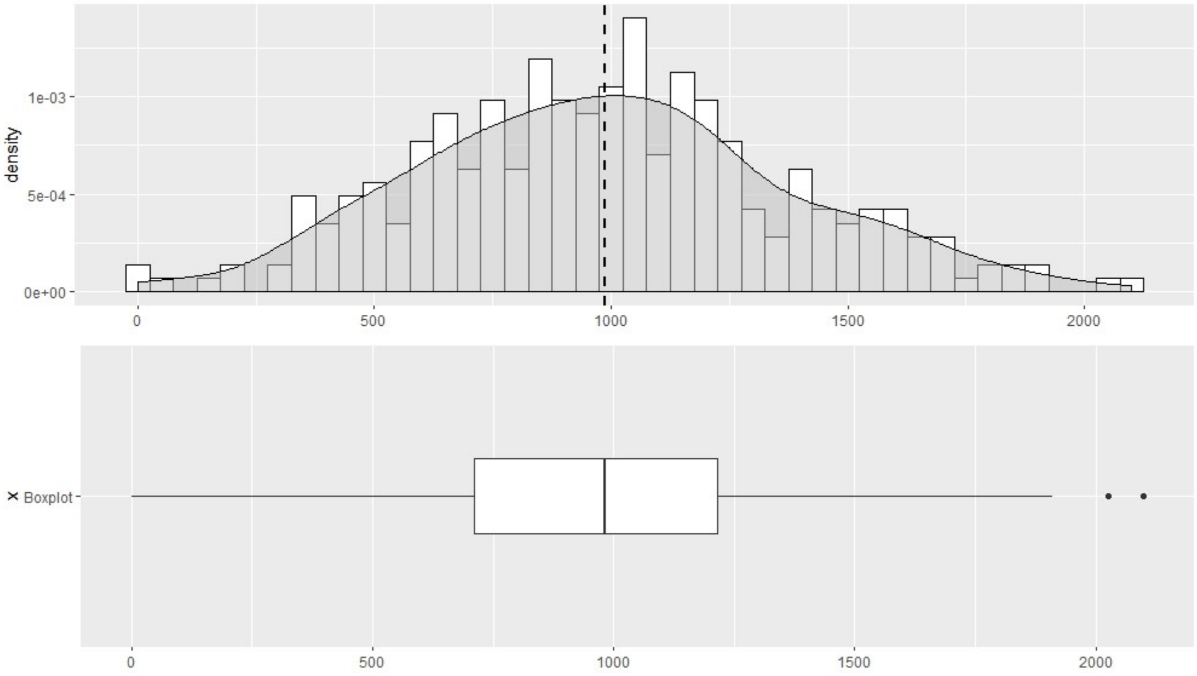

Figure 2. Distribution of the weekly size of releases of the tested product. Own elaboration.

Distribution of the weekly size of releases can be considered symmetrical. The average is 986, median 982, first quartile 710 and third 1216. Figure 3 presents the time series with decomposition into additive components.

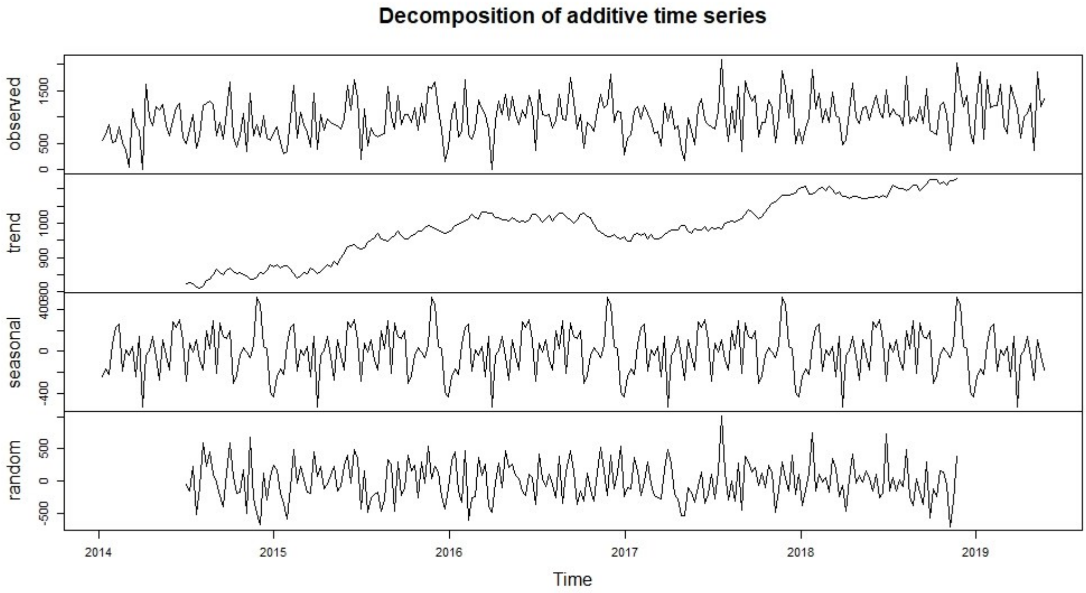

Figure 3. Decomposed time series of the weekly size of releases of the researched product. Own elaboration.

The visual analysis presented in Figure 3 of the time series indicates significant variation in value (coefficient of variation is 39.62%) and constant variance over time. The Fligner-Kileen test was used to examine the homogeneity of variance. It takes into account the last two sub-periods, each of which included: 36 weeks (p-value = 0.420), 40 weeks (p-value = 0.380), 60 weeks (p-value = 0.4387), 90 weeks (p-value = 0.420) and 120 weeks (p-value = 0.380). No seasonality other than annual was identified in the series. The time series was divided into two adequate sets: training (maximum 281 weeks) and test (4 weeks).

For the construction of automated forecasts, a training set for window size from 10 weeks to 281 weeks was used. The 10 best models due to the lowest RMSE value in the range of the training set are presented in Table 1.

Table 1.
The 10 best models due to the lowest RMSE value for the training set

Window size (weeks)	Model	Training set RMSE	Test set RMSE	p-value (Ljung-Box test)
105	ARIMA(0,0,0)(0,1,0)[52]	308.87	454.95	0.097
107	ARIMA(0,0,0)(0,1,0)[52]	309.41	454.95	0.046
106	ARIMA(0,0,0)(0,1,0)[52]	310.32	454.95	0.051
108	ARIMA(0,0,0)(0,1,0)[52]	315.25	454.95	0.274
111	ARIMA(0,0,0)(0,1,0)[52] with drift	318.04	524.72	0.216
109	ARIMA(1,0,1)(0,1,0)[52]	318.70	469.99	0.311
60	ARIMA(3,0,2) with non-zero mean	323.05	193.27	0.366
112	ARIMA(0,0,0)(0,1,0)[52] with drift	324.06	515.95	0.366
110	ARIMA(1,0,1)(0,1,0)[52]	324.23	488.63	0.156
121	ARIMA(0,0,0)(0,1,0)[52] with drift	328.15	520.86	0.008

The last column of Table 1 presents the p-values from the Ljung-Box test. The first 6 models with the lowest RMSE value take into account the annual seasonality, with differentiation value in a year-to-year range. At the same time, in terms of parameters, they are identical, taking into account only the constant average level of the examined variable. It should also be noted that only for the time window of the last 60 weeks, the RMSE value is lower during the test period than in the training period. The 10 best models due to the criterion of the minimum RMSE value during the test period are presented in Table 2.

Table 2.
The 10 best models due to the lowest RMSE value in the test set

Window size	Model	Training set RMSE	Test set RMSE	p-value (Ljung-Box test)
60	ARIMA(3,0,2) with non-zero mean	323.05	193.27	0.366
93	ARIMA(1,0,3) with non-zero mean	361.295	212.098	0.676
90	ARIMA(2,0,3) with non-zero mean	340.855	263.854	0.573
81	ARIMA(2,0,3) with non-zero mean	342.806	266.769	0.650
95	ARIMA(3,0,1) with non-zero mean	372.481	267.774	0.292
102	ARIMA(0,0,3) with non-zero mean	372.286	400.393	0.768
62	ARIMA(3,0,0) with non-zero mean	358.431	400.952	0.377
25	ARIMA(0,0,1) with non-zero mean	382.036	425.203	0.391
24	ARIMA(0,0,1) with non-zero mean	389.924	425.486	0.411
228	ARIMA(3,1,2)(1,0,0)[52]	370.042	426.587	0.947

The best model, due to the minimum RMSE value in the test set, was built on a time window of 60 weeks. At the same time, it is one of the best models in terms of the minimum RMSE value in the training set. The results of examining the residuals of this model are presented in Figure 4.

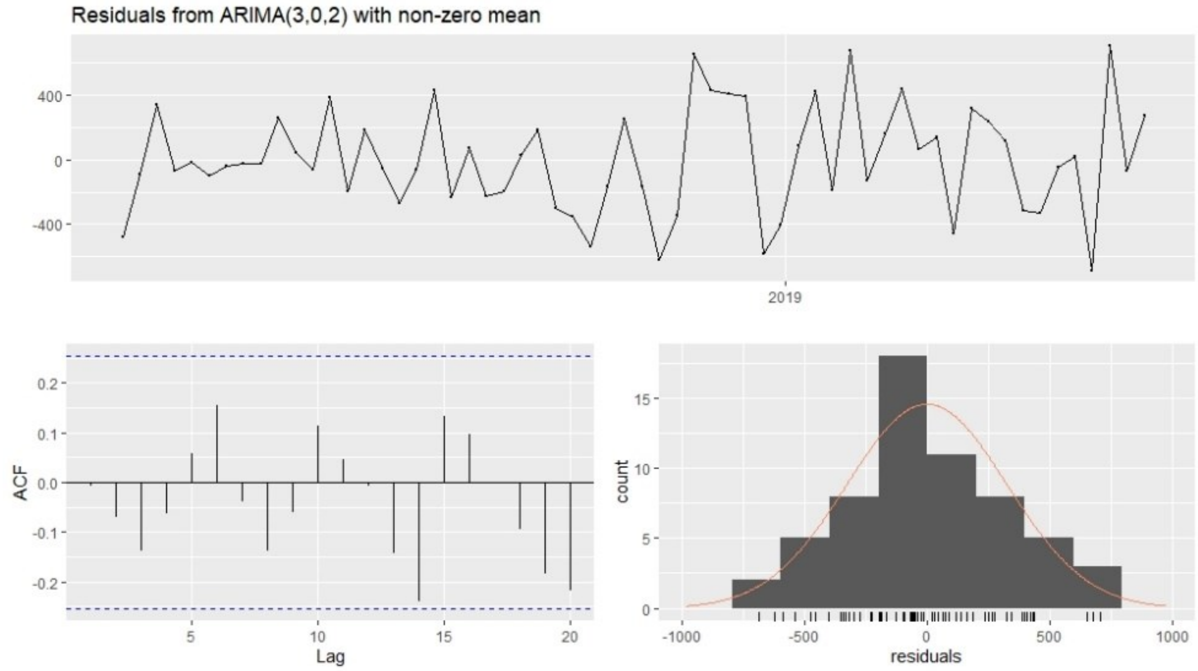

Figure 4. Results of the analysis of the residual size model of weekly releases for 60 weeks. Source: own elaboration.

Auto-correlation analysis of residuals (ACF) indicates, as well as the p-value from the Ljung-Box test indicates, that there are no significant auto-correlation of model residuals. Noteworthy is the fact that the distribution of residuals shown in the histogram indicates symmetry and “normality” of the distribution of residuals.

The forecast built on the basis of the described selection is presented in Figure 5.

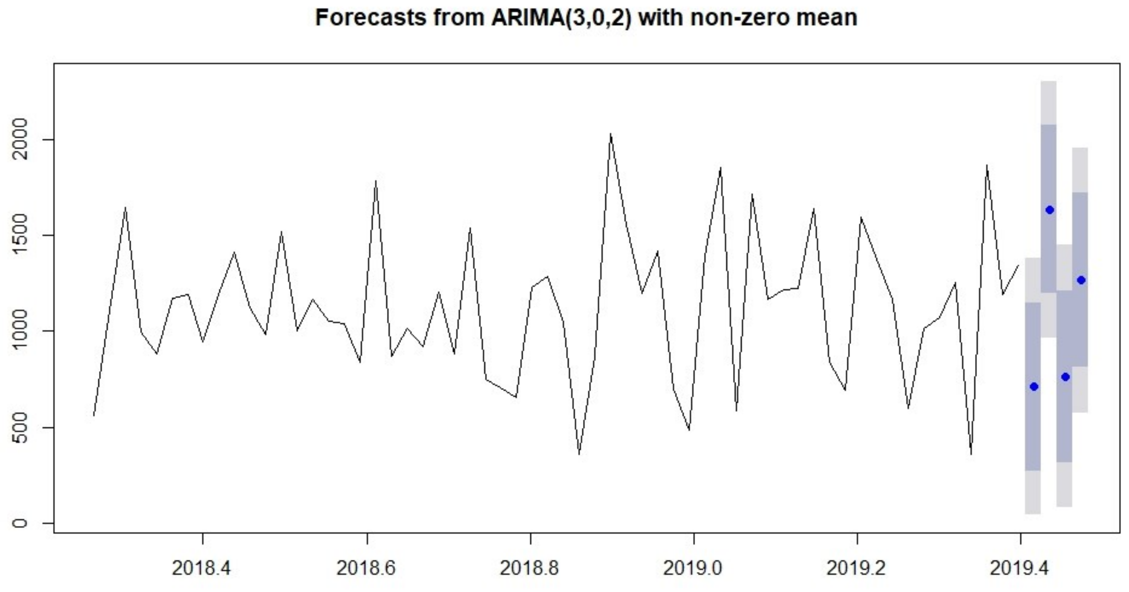

Figure 5. Forecast results. Source: own elaboration.

4. Conclusions

The approach to automatic forecasting of time series presented in the article was based on the algorithm created by Hyndman and Athanasopoulos and assumed the automatic forecast using the auto.arima function in the R software. The time series reflecting the weekly size of releases of the assortment group, typical for flows in the considered distribution network, were taken for analysis. This series concerned releases of a household chemicals product.

The high variability of the series, stability of variance, as well as the symmetrical distribution of weekly releases gave a basis for considerations on the selection of the appropriate length of the series, and thus making a forecast using it. Series lengths were tested in windows from 10 to 281 weeks, and the appropriate length was indicated based on the window selection, taking into account the lowest RMSE value in both the test and training set. One of the best models, taking into account the considered criteria, is a model built on a time window of 60 weeks. Based on this model, the forecast for the previously specified horizon was made.

It is not necessary to include all available data. The results presented in Table 1 and Table 2 indicate that information on the shaping of the studied phenomenon results from its history covering less than two years (60 weeks). Most of the presented models were built taking into account about a two-year time window. Thus, the direction of further research in this area will include verification of the hypothesis regarding a two-year time window for similar products.

The ability to automatically predict time series can give a logistics operator a number of benefits, which are related to, among others, the improvement of the quality of planning processes related to the possibility of predicting future demand, as well as bring many benefits to the operator's customers and the entire distribution network, in which demand can be predicted with greater accuracy. Of course, issues related to the future of the solution and its impact on the distribution network are conditioned not only by a properly constructed forecasting tool, but also by numerous changes in the way cells are managed in the network, reconfiguration of this network and numerous changes in information flow processes. Therefore, the issue of forecasting in the distribution network from the perspective of logistics operators remains a subject of research.

Acknowledgements

This paper was financed from the resources of the Silesian University of Technology, project no. BK-235/ROZ-1/2020 (13/010/BK_20/0042).

References

1. Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19(6)*, pp. 716-723. doi:10.1109/TAC.1974.1100705.
2. Bendkowski, J., Kramarz, M., and Kramarz, W. (2010). *Metody i techniki ilościowe w logistyce stosowanej. Wybrane zagadnienia*. Gliwice: Wydawnictwo Politechniki Śląskiej.
3. Burnham, K.P., and Anderson, D.R. (2004). Multimodel inference: understanding AIC and BIC in Model Selection. *Sociological Methods & Research, 33*, pp. 261-304. doi:10.1177/0049124104268644.
4. Dittman, P. (2000). *Metody prognozowania sprzedaży w przedsiębiorstwie*. Wrocław: Wydawnictwo Akademii Ekonomicznej im. Oskara Langego.

5. Fligner, M.A., and Kileen, T.J. (1976). Distribution-Free Two-Sample Tests for Scale. *Journal of the American Statistical Association, 71*, pp. 210-213.
6. Grzelak, M. (2019). Zastosowanie modelu ARIMA do prognozowania wielkości produkcji w przedsiębiorstwie. *Systemy Logistyczne Wojsk, 50*.
7. Hyndman, R.J., and Athanasopoulos, G. (2018). *Forecasting: principles and practice*. Melbourne: OTexts. Retrieved from http://www.OTexts.com/fpp2, 2019.07.24.
8. Hyndman, R.J., and Khandakar, Y. (2008). Automatic Time Series Forecasting: The Forecast Package for R. *Journal of Statistical Software*, *27*, pp. 1-220. https://doi.org/10.18637/jss.v027.i03.
9. Kawa, A. (2011). *Konfigurowanie łańcucha dostaw. Teoria, instrumenty i technologie*. Poznań: Wydawnictwo Uniwersytetu Ekonomicznego.
10. Kramarz, M., and Kramarz, W. (2012). Struktura sieci dostaw – sieciowe łańcuchy dostaw wyrobów hutniczych. In: J. Pyka (Ed.), *Nowoczesność przemysłu i usług – nowe wyzwania. Praca zbiorowa* (pp. 300-310). Katowice: Towarzystwo Naukowe Organizacji i Kierownictwa. Oddział w Katowicach.
11. Schwarz, G.E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6(2)*, pp. 461-464. doi:10.1214/aos/1176344136.
12. Wojciechowski, A., and Wojciechowska, N. (2015). Zastosowanie klasycznych metod prognozowania popytu w logistyce dużych sieci handlowych. *Zeszyty Naukowe Uniwersytetu Szczecińskiego, 41*, pp. 545-554.
13. Żebrucki, Z. (2012). *Badania form partnerstwa logistycznego między przedsiębiorstwami*. Gliwice: Wydawnictwo Politechniki Śląskiej.