ARTICLE

DATA ANALYSIS IN DEMAND FORECASTING: A CASE STUDY OF POETRY BOOK SALES IN EUROPE

Kolková, A.

Andrea Kolková / VŠB – Technical University Ostrava, Faculty of Economy, Department of Business Administration, 7. listopadu 2172/15, 708 00 Ostrava, Czech Republic. Email: andrea.kolkova@vsb.cz

Abstract

Logistics concepts are becoming less functional nowadays. The successful concept of just-in-time manufacturing seems untenable for 21st-century Europe. New ideas and new practices are emerging. One of the major trends in contemporary logistics is the demand-driven enterprise. Managing supply, fields and other processes in a company based on demand requires accurate demand forecasting at the relevant time. Relevant data are necessary for this forecasting. In the time of big data, the problem is not collecting data but evaluating them correctly. Data analytics is of great interest in business. The aim of the paper is to verify the possibilities of demand forecasting using traditional statistical methods (exponential smoothing, ARIMA), more advanced statistical methods (TBATS, etc.), methods based on artificial neural networks and a hybrid method. This is in conjunction with thorough data preparation, especially data normality testing. My research question is to provide, on data that contain a number of outliers, the results and the accuracy of the models when the data are or are not normalized. Demand forecasting for poetry books was chosen for the case study. The data were obtained from Google Trends data, i.e., searches for the topic of poetry for the period from 1 September 2013 to 31 September 2023. The results showed that the selected data contain a number of outliers that recur at regular intervals and are the result of a logical order of demand. The expected result was that data normalization increases the accuracy of the model. A method based on artificial neural networks provided significantly more accurate results. However, the resulting estimated underlying trend remained very similar. The article thus opens a discussion about the necessity of excluding outlying observations in time series where outliers exist at regular intervals.

Implications for Central European audience: The article poses fundamental scientific questions for Central Europe. Logistics concepts are developing rapidly in this area and Europe is the creator of significant innovations. Europe is currently facing great competition from foreign companies and new logistics concepts are becoming a necessity. Therefore, many European companies now feel the need to change their logistics processes. One of them may be to switch to an adaptive enterprise based on demand. The practical implications are also based on data from Europe.

Keywords: Demand forecasting; demand drive adaptive enterprise; normality test; exponential smoothing; ARIMA; neural networks; hybrid model

JEL Classification: M21; M15; C53

Introduction

The development of logistics is accelerating with the changing world, and especially in Central Europe, we have to adapt to modern logistics management. The COVID-19 pandemic and the war in Ukraine and Israel have shown that the traditional logistics procedures promoted in the past, such as just-in-time manufacturing, are no longer sufficient. Logistics is becoming a rapidly developing scientific field that responds flexibly to changes in the globalized world. In contrast to just-in-time manufacturing, the concepts of demand-driven adaptive enterprise, omnichannel logistics, or change of logistics processes in connection with the circular and shared economy are beginning to gain more traction. These tendencies are most noticeable in Central Europe.

The concept of a demand-driven adaptive enterprise requires not only demand forecasting methods with high accuracy but also fast calculations using relevant data. In the era of big data, the problem is not to collect data, but to evaluate them correctly. Data analytics is gaining interest in the business.

More and more new technologies are used for demand forecasting calculations. For example, Wang and Chen (2022) used adaptive neural networks to implement tourism demand forecasting. At the same time, the methods of artificial neural networks are constantly being improved and long short-term memory, which is still considered to be the most modern method, is being replaced by more and more accurate models (Anupam & Lawal, 2023). Big data are used by scientists not only in the academic sphere but also in practice. Scientific works also examine the effects of predictive analytics of big data on supply chain performance (Shafique et al., 2024). Conventional forecasting is being replaced by predictive analytics and its knowledge can create competition (Korayim et al., 2024). Big data and artificial intelligence can be one of the most important resources for improving business (Sivarajah et al., 2017).

The concept of demand-driven adaptive enterprise is also identified as one of the steps of the systematic process of business operations through the theory of constraints (Orue et al., 2023). Accurate demand forecasting is important for the concept of a demand-driven company. Scientists are still developing more accurate models or investigating their accuracy (Kolková & Rozehnal, 2022); models are created not only in the academic sphere but also in practice (Kolková & Ključnikov, 2022). An important part of scientific research is often overlooked: data preparation. Scientists take it for granted. The aim of the present article is to verify the possibilities of forecasting demand using traditional statistical methods (exponential smoothing), more advanced statistical methods (ARIMA, TBATS) and methods based on artificial neural networks and hybrid models. All will be verified following data preparation, especially data normality testing. Our research question is to provide the results and accuracy of the models on data that contain a number of outliers in case the data are or are not normalized.

Google Trends (GT) data were chosen for the analysis, namely the search for the topic of poetry for the period from 1 September 2013 to 31 September 2023. In the last decade, the use of GT data has proven to be useful. There are many studies that forecast demand based on GT data (Kolková & Ključnikov, 2021; Dinis et al., 2017). Scientific studies have examined the link between GT data and employment or unemployment (Baker & Fradkin, 2017),

Volume **13** | Issue **5** | 2024
https://doi.org/10.18267/j.cebr.371

consumption (Morgavi, 2020), trade (Gonzales et al., 2020), digitization (Pisu et al., 2020), or prices in the construction industry (Cournède et al., 2020). Recently, Google Trends data have also been used to assess the impact of the COVID-19 crisis (Abay et al., 2020; Doerr & Gambacorta, 2020).

The R language will be used for the calculation, especially the packages forecast according to Hyndman & Athanasopoulos (2021), forecastHybrid (Shaub & Ellis, 2020), Prophet (Taylor & Letham, 2018) and the Python language.

The first section of the paper contains methodology and data. First, the forecasting methods are described. Accuracy is explained in section 1.2. The next section describes the demand forecasting process used in this study. Section 1.4 contains detailed data analysis, including extrapolation description analysis, data decomposition and normality testing. This is followed by a section summarizing the results. The discussion section opens up further research questions that this study can initiate. The last section brings the conclusion.

1 Methodology and Data

The first section defines the methods used in the case study. The next section lists the accuracy used. Finally, the data are analysed, data normality testing is performed and data normalization follows. This article uses Google Trends data, which are freely available from Google at the website:

https://trends.google.com/trends/explore?cat=22&geo=CZ&hl=cs

1.1 Forecasting methods

In this work, both conventional statistical methods, commonly used in forecasts, and new and modern methods are used. Conventional statistical methods include the naïve method and the seasonal naïve (snaïve) method. These methods are used as a benchmark. When decomposing the data in section 1.4.2, it was found that the data contain a seasonal component; therefore, the seasonal naïve method was chosen as the benchmark. However, different methods are used for forecasting.

Other statistical methods that are often used include ETS (Error, Trend, Seasonal) (Brown, 1959; Holt, 2004; Winters, 1960), which can be defined using formulas (1), (2), (3) and (4),

$$y_t = l_{t-1} + b_{t-1} + s_{t-m} + \varepsilon_t, \text{ where} \quad (1)$$

$$l_t = l_{t-1} + b_{t-1} + \alpha\varepsilon_t \quad (2)$$

$$b_t = b_{t-1} + \beta\varepsilon_t \quad (3)$$

$$s_t = s_{t-m} + \gamma\varepsilon_t, \text{ where} \quad (4)$$

l_t is the estimate level, b_t is the estimated trend, t and s_t express the seasonality, α, β and γ are weight coefficients.

Another group of methods used in this article is ARIMA (autoregressive integrated moving average), first defined by Box & Jenkins (1976). The ARIMA model with the seasonal component can be expressed in the form ARIMA (*p, d, q*) *(P, D, Q*), where *p* represents the

degree of the autoregressive part, *d* the degree of differentiation and *q* the degree of the moving average part. ARIMA can be calculated as follows:

$\varphi(B)w_t = \theta(B)\varepsilon_t$, where (5)

$w_t = \Delta^d y_t$ (6)

In this article, ARIMA is calculated in the R program based on the package according to Hyndman et al. (2002). The automated function auto.arima is used. The results describe the most accurate ARIMA model for individual data.

In addition, more complex and modern methods are used, usually providing higher accuracy of results (Kolková, 2020). Artificial neural networks are the method used nowadays for demand forecasting (Swaminathan & Venkitasubramony, 2024; Kolková & Ključnikov, 2022; Kolková, 2019). The output of a neuron is calculated when the sum of the inputs to the neuron x_i multiplied by their specific weights w_i exceeds a certain value, which we call the distortion. The neuron can be described as follows:

$y_j = f\left(\sum_{i=1}^{m} x_i \cdot w_{j,i} - b_j\right)$ (7)

where x_i is the specific value on the *i*-th input, $w_{j,i}$ are the weights of this input, b_j is bias, *m* is the total number of inputs, *f* is a transformation function and the output values of *y*. The model is calculated according to Hyndman et al. (2002).

Hybrid models used in previous studies (Kolkova & Rozehnal, 2022) are represented by the *forecastHybrid* model (Shaub & Ellis, 2020). This model is a combination of ARIMA, ETS and Theta, TBATS and artificial neural networks. The whole model is then defined as follows:

$y(i) = \sum_{m=1}^{n} w_m \cdot f_m(i)$ (8)

where w_m is weight for each *m* of the *n* model, and $f_m(i)$ is the individual model for the time horizon *i*. This model is called *forecastHybrid* and is published in a package in the statistical program R (Shaub & Ellis, 2020).

Another of the hybrid models is the Theta model (Assimakopoulos & Nikolopoulos, 2000). Based on the concept of modification of local fluctuations of the time series using the coefficient theta (denoted by the Greek letter *θ*), this coefficient is applied to the second difference of the time series, as follows:

$X''_{new} = \theta \cdot X''_{data}$, where (9)

$X''_{data} = X_t - 2 \cdot X_{t-1} + X_{t-2}$ (10)

1.2 Accuracy calculation

Accuracy methods can be divided into many groups. Hyndman and Athanasopoulos (2021) divided accuracy calculation methods into scale-dependent measures, measures based on percentage errors and scaled errors. In Hyndman and Koehler (2006), accuracy was also calculated using measures based on relative errors and relative measures.

Scale-dependent errors are used in this article. We employ mean error (ME), mean scaled error (MSE), root mean squared error (RMSE) and mean absolute error (MAE). These accuracies are dependent on the scale of the data used and should not be used to compare datasets with different scales. They can be quantified as follows:

Volume **13** | Issue **5** | 2024
https://doi.org/10.18267/j.cebr.371

$$MSE = \frac{1}{M}\sum_{p=n+1}^{N}\left(y_p - \widehat{y_p}\right)^2 \quad (11)$$

$$RMSE = \sqrt{MSE} \quad (12)$$

$$MAE = \frac{1}{M}\sum_{p=n+1}^{N}\left|y_p - \widehat{y_p}\right| \text{ nebo } MAE = mean|e_t| \quad (13)$$

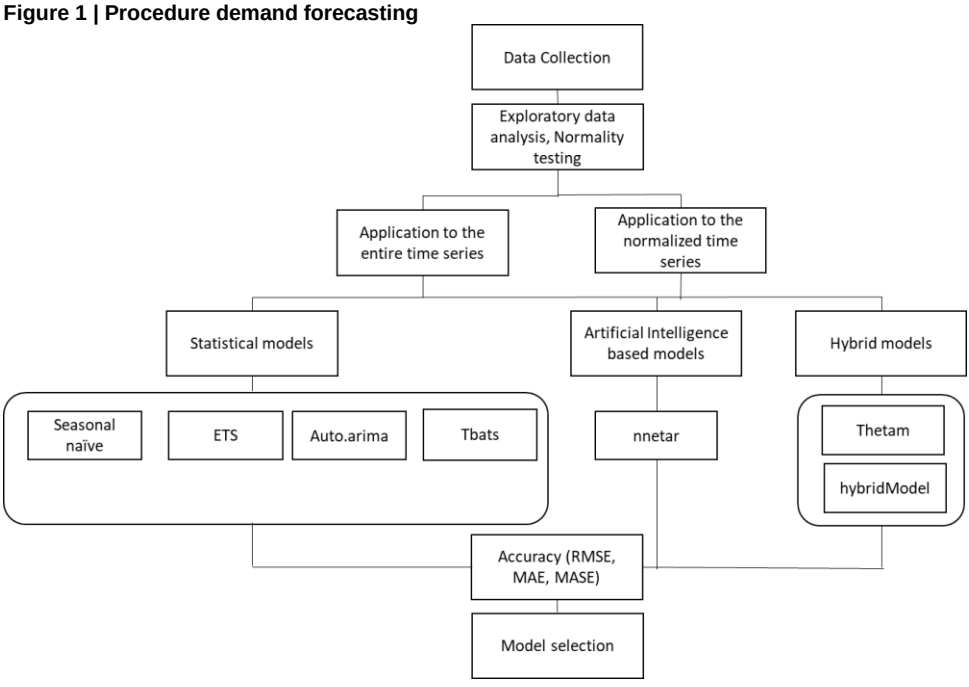
These accuracies have the same unit as the original time series. The parameters of the accuracy measures based on percentage errors are already independent of the scale. Their use is precisely to compare measures of accuracy across different datasets. It is not necessary to use this accuracy in this article, but it is also calculated here for the sake of interest. Mean absolute percentage error (MAPE) is among the least common indicators of this type.

$$MAPE = \frac{1}{M}\sum_{p=n+1}^{N}\frac{\left|y_p - \widehat{y_p}\right|^2}{y_p} \quad (14)$$

1.3 Demand forecasting

As follows from the previous text, demand forecasting is not just a calculation of the selected forecast method but a set of consecutive steps. The demand forecast procedure may vary and may be specific to different studies, for example, Kolková & Ključnikov (2022), Xie et al. (2023) or Wu et al. (2023). In this article, the procedure according to Figure 1 is used. Data collection is followed by data preparation, such as exploratory data analysis and normality testing. Data collection is an integral part of demand forecasting and is carried out in section 1.4. Forecasting methods are calculated according to the methodology described in section 1.1 and accuracy is calculated according to section 1.2. These two procedures are implemented on normalized data and on the original time series. Then, the results are compared.

Figure 1 | Procedure demand forecasting



Source: Own processing

1.4 Data analysis

GT data provide indexes of search data volume that measure search intensity according to the formula:

$$\frac{keyword\ search\ count}{total\ search\ count} \tag{15}$$

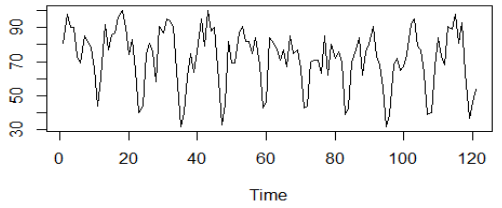
The calculation is based on location and period. In GT data, these indexes can be used by keyword, keyword category or topic. Of course, keyword-based queries are subject to language specifics. Categories and topics are harmonized across languages. That is why categories are used in this work as well. About 1200 categories can be found in GT data, which assign individual searches according to a probability algorithm (Varian & Choi, 2009).

The main disadvantage of GT data is that they have been collected since 2004. The number of Google search users has increased significantly since then, so the relative intensity of searches is decreasing for most categories. To eliminate this shortcoming, data from 2013 to 2023 are used in this study. In this period, the user base is approximately similar. It goes without saying that the big data used in GT data must be statistically described and possibly even processed.

1.4.1 Extrapolation description analysis of data

First, it is necessary to thoroughly describe the data. Their development contains certain patterns and these repeat over time. The development of the data is shown in Figure 2.

Figure 2 | GT data graph for book poetry



Source: Own processing

The data used represent search interest relative to the highest point of the chart for a given region and time. A value of 100 represents the highest popularity of the term and the highest number of purchases. A value of 50 means that the term had half the popularity. A score of 0 means that not enough data were collected for the term or there was minimal interest in searching for this term.

The data were described statistically. The results of the extrapolation description analysis are shown in Table 1.

Table 1 | Extrapolation description analysis

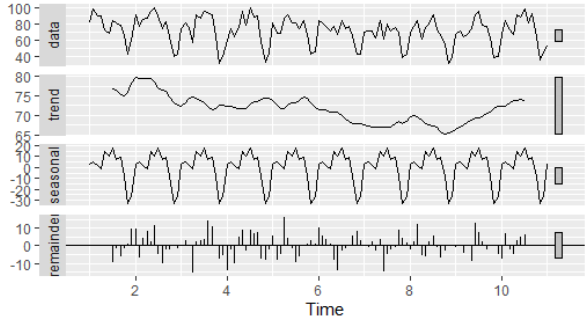
	Book of poetry		Book of poetry
Minimum	32	**Maximum**	100
1st quartile	64	**Variance**	292.8833
Median	75	**Skewness**	-0.6432905
Mean	72	**Kurtosis**	2.657988
3rd quartile	84		

Source: Own processing

1.4.2 Data decomposition

Data decomposition is also necessary for appropriate data visualization. An additive decomposition was performed. The result is expressed in Figure 3.

Figure 3 | Decomposition of additive time series



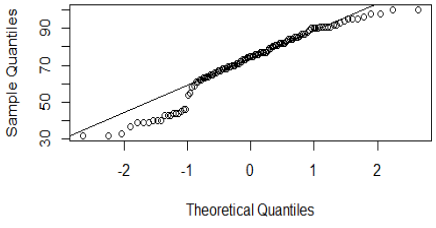
Source: Own processing

The first graph is a line graph of the examined data. The next three graphs form the decomposition of these data. For better visual appearance, the data on the y-axis are not to the same scale. In the trend graph, it can be seen that at first, less than 9 years of data form a slightly decreasing or stagnant trend. In recent years, we can see a clearly growing trend and, therefore, increasing numbers of searches for terms on the topic of poetry in the Czech Republic. The seasonal trend, which is in the third graph, really reflects that holiday dates are outliers, and in the months of July and August, the number of searches for information about poetry decreases significantly. The fourth graph then contains a random component that is part of all data. The decomposition of the data shows that the data contain seasonality; therefore, the seasonal naïve (snaïve) method was chosen as the benchmark.

1.4.3 Normality testing

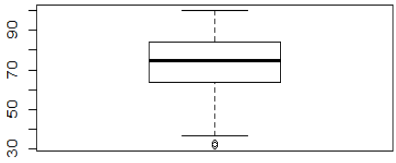
From the analysis of skewness and kurtosis, it is already clear that the data will contain outliers, and it is therefore necessary to define them. Using the box plot and QQ plot graph, outliers in the lower values of the data are visible (see Figures 4 and 5).

Figure 4 | Q-Q plot



Source: Own processing

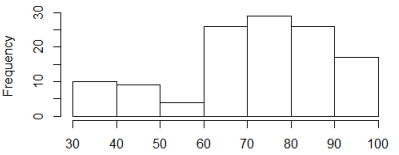
Figure 5 | Box plot with outliers



Source: Own processing

We have two options for dealing with outliers. They can be either excluded them from the dataset or retained. This decision can be supported by testing the normality of the data. If the data do not belong to a normal distribution, it may be due to these outlying observations. If the data are not in a normal distribution, the use of some statistical forecasting methods is impossible or difficult. For visual representation, a Q-Q plot (see Figure 2) and a histogram (see Figure 6) were chosen. From its visual comparison, it is clear that this is not a normal distribution.

Figure 6 | Histogram with outliers



Source: Own processing

According to the graph, it is not possible to confirm normality unequivocally; a test had to be performed. Normality was tested using the Shapiro-Wilk normality test. Using this test, it was found that all the data together were not in a normal distribution, as expressed in the first row of Table 2. Therefore, it was necessary to analyse the data further.

The data were analysed in more detail, focusing on outliers. Two values were extreme outliers and another five were also outliers, but not extreme. The analysis of outliers found that these are always holiday dates. This is probably due to the fact that searching for terms on the topic of poetry is often associated with schooling, including elementary, secondary and university studies. So, it is obvious that the holidays will always register a lower demand for poetry books. Based on this consideration, it would be appropriate to leave these holiday dates out of the analysis. They do not follow the underlying trend that is predicted in this article. Of course, this decision required normality testing. The results of the normality test for all the data are shown in Table 2 and, as already said, they do not belong to a normal distribution. Furthermore, two extreme outliers were removed from the data, but even after removing these extreme values, the data still did not form a normal distribution. After deciding to exclude all holiday dates and thus all outliers due to their insignificance in the dataset, we can already see that the data have a normal distribution.

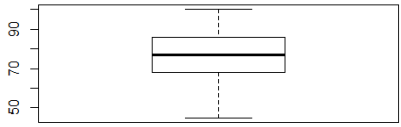
Table 2 | Shapiro–Will normality test

	W	p-value
Original time series	W = 0.94184	p-value = 5.311e-05
Data without 2 largest outliers	W = 0.94511	p-value = 0.0001106
Data without outliers	W = 0.98253	p-value = 0.1828

Source: Own processing

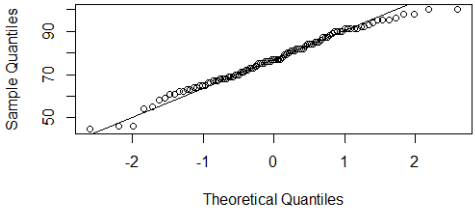
In the following graphs, we can see the box plot (see Figure 6) and the Q-Q plot (see Figure 7) without outliers. The data now have a normal distribution.

Figure 7 | Box plot without holiday values



Source: Own processing

Figure 8 | Q-Q plot without holiday values



Source: Own processing

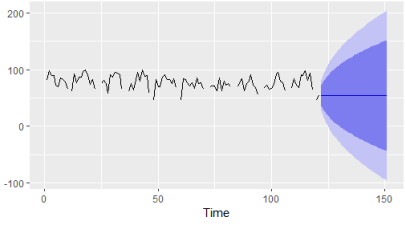
For demand forecasting, normalized data, i.e., data without outliers, will be used. Holiday values are excluded from these data, which differ from searches or demand in the current year.

For comparison, demand forecasting will be calculated on all the data. Subsequently, the accuracy of the forecast and the resulting trend predicted by demand forecasting using both datasets will be compared.

2 Results

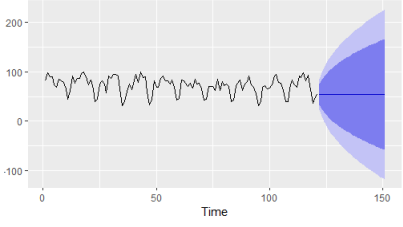
The results are described for the original time series first, and then for the normalized data. The seasonal naïve method was chosen as the benchmark. The forecast results are shown in Figures 9 and 10. It is clear from the results that the demand for poetry books has a stagnant trend. This applies when using the original time series, i.e., data after excluding outliers (see Figures 9 and 10). This forecast is only very general and insufficient for company logistics management. The forecast is used only as a benchmark.

Figure 9 | Forecast from seasonal naïve method (normalized series)



Source: Own processing

Figure 10 | Forecast from seasonal naïve method (all time series)



Source: Own processing

The accuracy results are shown in Table 3. Accuracy improved when outliers were used. The improvement is noticeable for all the monitored accuracies.

Table 3 | Seasonal naïve method accuracy

Accuracy	RMSE	MAE	MAPE
Original time series	16.00286	12.90833	20.49864
Normalized time series	13.93908	11.125	15.08387

Source: Own

The results of the ETS method are shown in Table 4. When using data without outliers, the forecasting method is more accurate according to all the accuracy parameters. When examining the data resulting from the forecasts, the results are again a stable stagnant demand. Given that it is already evident from the decomposition that demand has seasonal tendencies, it is clear that the information on long-term stagnation is insufficient.

Table 4 | Parameters, ETS forecast

ETS	Smoothing parameters	Initial states	Sigma	AIC	AICc	BIC
Original time series ETS(A,N,N)	alpha = 0.9883	l = 81.1937	16.0694	1256.288	1256.493	1264.675
Normalized time series ETS(M,N,N)	alpha = 0.0519	l = 80.423	0.1559	1017.218	1017.456	1025.180

Accuracy	RMSE	MAE	MAPE
Original time series	15.93603	12.77695	20.29823
Normalized time series	12.10385	9.628115	13.61999

Source: Own processing

According to the methodology defined above, the ARIMA models were further calculated and the best model was selected using an automated procedure in the R language. The ARIMA(1,0,3) model with non-zero mean, with the parameters listed in Table 5, was found to be the most suitable when using the original time series.

Table 5 | Parameters, ARIMA forecast ARIMA(1,0,3) using original time series

Coefficients	ar1	ma1	ma2	ma3	mean
	0.6951	-0.0395	-0.3139	-0.3253	71.8550
std. error	0.1287	0.1357	0.0943	0.0772	1.3053
	sigma^2	**log likelihood**	**AIC**	**AICc**	**BIC**
	183.4	-484.78	981.57	982.31	998.34

Source: Own processing

Using a normalized time series, the best-fit model was ARIMA(2,0,2) with a non-zero mean with parameters listed in Table 6.

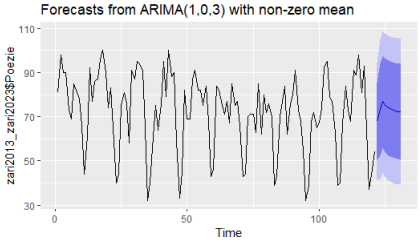
Table 6 | Parameters, ARIMA forecast ARIMA(3,0,1) using normalized time series

Coefficients	ar1	ar2	Ar3	Ma1	mean
	0.9009	0.0902	-0.3771	-0.6616	75.9619
std. error	0.1407	0.1478	0.1106	0.1226	0.9597
	sigma^2	**log likelihood**	**AIC**	**AICc**	**BIC**
	119.9	-399.4	810.8	811.54	827.57

Source: Own processing

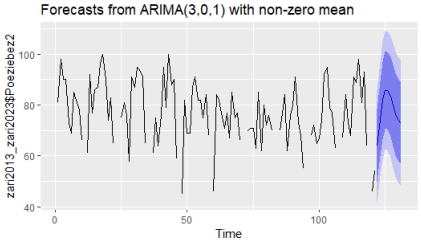
The models forecast demand movements in the chosen future horizon. When using the original time series, the forecast was first increasing demand, followed by a small decline and stagnant demand. When using a normalized series, growth is again first predicted in a relatively small forecast horizon. Then, there is a comparatively deeper drop in demand, which lasts throughout the forecast horizon. We can see the differences in demand forecasts in Figures 11 and 12.

Figure 11 | Forecast from original time series



Source: Own processing

Figure 12 | Forecast from normalized time series



Source: Own processing

The accuracy parameters were calculated according to the methodology presented in the previous section. Again, the accuracy of the model when using the normalized data is higher in all the parameters. The results are shown in Table 7.

Table 7 | Accuracy, ARIMA models

Accuracy	RMSE	MAE	MAPE
Original time series	13.2605	10.58643	16.96686
Normalized time series	10.68473	9.122588	12.35824

Source: Own processing

Another selected method is the BAT method. Again, the parameters of the best-fitting model were selected based on an automated process in the R language. The best-fitting model for the original time series was BATS(1,{2,3},-,-) with the parameters listed in Table 8.

Volume **13** | Issue **5** | 2024
https://doi.org/10.18267/j.cebr.371

Table 8 | Parameters, TBATS forecast BATS(1, {2,3}, -, -) using original time series

BATS(1, {2,3}, -, -)	Alpha	AR coefficients	MA coefficients	Sigma	AIC
	0.2546683	1.652788 -0.884301	-1.409452 0.180459 0.304204	12.31798	1211.967

Source: Own processing

For normalized data, i.e., without outliers, the model BATS(1, {2,2}, 1, -) was chosen, with the parameters listed in Table 9.

Table 9 | Parameters, TBATS forecast BATS(1, {2,2}, 1, -) using normalized time series

Alpha	Beta	Damping parameter	AR coefficients	MA coefficients	Sigma	AIC
0.07603753	0.03710725	1	1.544293	-1.394948	10.87069	1013.741
			-0.797976	0.444973		

Source: Own processing

However, the forecast results again, similar to ETS or seasonal naïve, only indicate a long-term trend of demand stagnation. This information is insufficient. In addition, the model has less accuracy on data without outliers than the auto.arima model, so it is unsuitable for use in forecasting this demand. Table 10 shows the accuracy of BATS models for original data and for normalized data.

Table 10 | Accuracy, BATS models

Accuracy	RMSE	MAE	MAPE
Original time series	12.31798	9.917763	15.68755
Normalized time series	10.87069	8.715767	11.87327

Source: Own processing

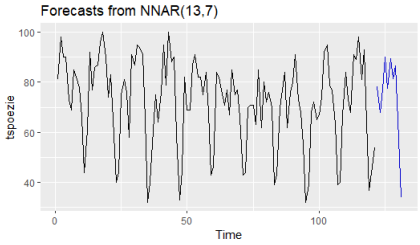
Another calculated model was a model based on artificial neural networks. The nnetar model is analysed according to the methodology described in the previous section. First, a model was calculated based on the original values, which do not belong to a normal distribution. The model was an average of 20 networks, each of which is a 13-7-1 network with 106 weights. Options were linear output units. The model was labelled NNAR(13,7). Sigma^2 estimated as 0.2343.

When we exclude outliers, the results of the model according to artificial neural networks are an average of 20 networks, each of which is a 10-6-1 network with 73 weights. Options were linear output units. The model is labelled NNAR(10,6) and sigma^2 estimated as 2.709.

Figures 13 and 14 show the forecast with the original time series and the normalized time series. While the original time series predicts only a decline at the end of the forecast period, the normalized data predict a decline and subsequent growth. This is the most significant deviation in the forecast that the methods provided. Since the accuracy based on the

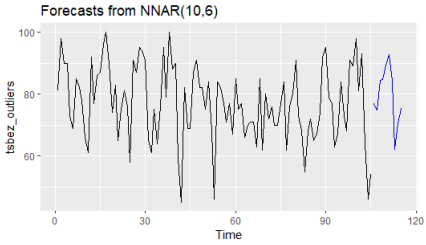
normalized data for the nnetar method was significantly greater, it can be assumed that this forecast is also more realistic.

Figure 13 | Forecast from original time series



Source: Own processing

Figure 14 | Forecast from normalized time series



Source: Own processing

Accuracy was again calculated for comparison; the results are shown in Table 11. From the results, the model with normalized data is clearly more accurate. Models based on artificial neural networks are significantly more accurate than previous models.

Table 11 | Accuracy, nnetar

NAR(13,7)	RMSE	MAE	MAPE
Original time series	8.566658	6.83864	9.763706
Normalized time series	1.58721	1.095748	1.504651

Source: Own

Among the hybrid models, the theta model (Assimakopoulos & Nikolopoulos, 2000) and hybridModel (Shaub & Ellis, 2020) was chosen, both according to the formulas defined in the theoretical part. The theta model was created with the parameters listed in Table 12. The optimization criterion for selecting the model parameters was MSE.

Table 12 | Parameters, theta using original time series

Smoothing parameters	Alpha	Initial states	sigma	AIC	AICc	BIC
	0.988	81.214	16.0694	1256.288	1256.493	1264.675

Source: Own processing

Table 13 | Parameters, theta using normalized

Smoothing parameters	Alpha	Initial states	sigma	AIC	AICc	BIC
	0.0393	80.0972	12.2107	1018.133	1018.370	1026.094

Source: Own processing

The theta model for the original time series only provided information about a stagnant trend. The model did not provide a detailed analysis of the individual parts of the forecast. Also, the accuracy of the models in Table 14 was not great. However, the accuracy of the model is higher than benchmark; thus, it can be included among the successful models.

Volume **13** | Issue **5** | 2024
https://doi.org/10.18267/j.cebr.371

Table 14 | Accuracy, theta

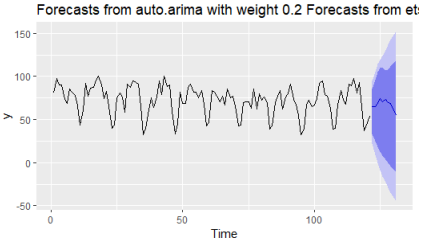
Accuracy	RMSE	MAE	MAPE
Original time series	15.93603	12.77645	20.29759
Normalized time series	12.09387	9.656844	13.67629

Source: Own processing

The hybrid forecast model is comprised of the following models: auto.arima, ets, theta, nnetar, tbats. It was created with the same weight for all the models, namely 0.2.

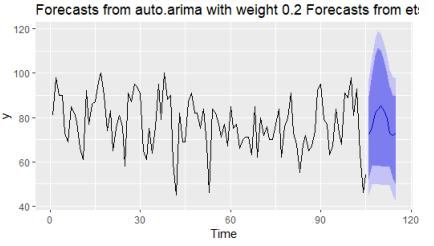
In the forecast, we again see a stagnant trend with visible parts. First, there is growth and it is followed by a decline. This decline is then followed by a short stagnation in the given forecast horizon for the model using normalized data.

Figure 15 | Forecast from original time series



Source: Own processing

Figure 16 | Forecast from normalized time series



Source: Own processing

The results of the accuracy models with both types of data are in Table 15. It can be seen from the table that all the accuracy parameters show a higher accuracy of the methods when using normalized data.

Table 15 | Accuracy, hybridModel

NAR(13,7)	RMSE	MAE	MAPE
Original time series	10.90396	8.786429	14.08227
Normalized time series	9.062566	7.363862	10.38252

Source: Own

3 Discussion

The accuracy of all the analysed models exceeded the benchmark (seasonal naïve). Thus, it would be possible to adopt all the models and apply them to the practice of forecasting the demand for poetry books. However, the accuracy of the models and their predictions differed when using the original and normalized data. It is evident from the results that the accuracy of the models is greater when using normalized data; in this case, data without outliers. This

is a finding that could have been expected. The increase in accuracy is greatest when using the method of artificial neural networks.

Table 16 | Comparison of forecast models

Model	Accuracy		Original time series forecast	Normalized time series forecast
	RMSE			
	All time series	Normalized time series		
snaïve	16.00286	13.93908	→	→
ETS	15.93603	12.10385	→	→
auto.arima	13.2605	10.68473	↗ ↘ →	↗ ↘
TBATS	12.31798	10.87069	→	→
nnetar	8.566658	1.58721	↗ ↘	↗ ↘ ↗
theta	15.93603	12.09387	→	→
hybridModel	10.90396	9.062566	↗ ↘	↗ ↘ →

Source: Own processing

ETS and theta were the least accurate methods. These methods also only predicted stagnation in demand without a more detailed view. Other methods predicted demand growth followed by a decline. In the auto.arima and hybrid methods, a decline can be followed by stagnation. Due to the short length of the forecast horizon, the period of stagnation is only small.

The method based on artificial neural networks is absolutely the most accurate on normalized data. This method is the only one that assumes demand growth at the end of the forecast horizon.

Differences in the accuracy of models with original data and normalized data, cleaned from the influence of outliers, are present in all used models. Models without outliers are definitely more accurate. However, the resulting forecast did not differ much over the short forecast horizon of 30 days. The methods provided different results on the normalization data only at the end of the forecast horizon. The first two-thirds of the forecast horizon provided almost identical forecasts based on the original and normalized data.

The prognostic procedure has already been analysed in previous studies (Kolková & Navrátil, 2021; Kolková & Ključnikov, 2022). Here, in addition to data preparation, a multi-criteria evaluation of the results is also defined. In accordance with other authors (Korayim et al., 2024), the importance of big data and subsequent demand forecasting is confirmed in this article. Predictive analytics applied here is an integral part of modern business management, which is in line with other studies (Shafique et al., 2024).

The results of this study open questions for further research. In particular, the research question is whether, when forecasting methods are improved using AI tools, it will still be necessary to prepare data for the forecast itself. We may ask whether models that are based on learning, for example, neural networks, could deal with errors in the data by themselves. It is, therefore, a big scientific question whether to always prepare the data and thoroughly clean them of possible outliers or whether to use learning algorithms to create models that shorten the path to the resulting forecast.

Volume **13** | Issue **5** | 2024
https://doi.org/10.18267/j.cebr.371

Conclusion

The objective of this paper was to provide the results and accuracy of models on data that contain a number of outliers when the data are or are not normalized. To meet this objective, poetry book demand data for the period from 1 September 2013 to 31 September 2023 and monthly data were processed. This was a GT data search with the theme of poetry for the Czech Republic.

These data were analysed and several outliers were found. By studying the data, it was found that these outliers are always in the months of June and August, which correspond to the holiday months. In this paper, the predictions using the entire time series, i.e., including the outlying observations, and the predictions using the normalized time series were compared.

The results show that the accuracy of the models is significantly better when using normalized data. However, the demand forecast does not contain significantly different values and predicts the same trends even on data containing all values. It was also appreciated that the use of multiple accuracies is not necessary, as they usually provide the same results. It was calculated that the demand in the predicted horizon will first grow for a short time and then fall and have a stagnant trend. The method with the highest accuracy was based on an artificial neural network called nnetar.

The paper thus opens a discussion on the need to exclude outliers in time series where outliers occur at regular intervals.

Acknowledgement

Funding: There was no special funding towards this study.

Conflict of interest: The authors hereby declare that this article was not submitted or published elsewhere. The authors do not have any conflict of interest.

References

Abay, K., Tafere, K., & Woldemichael, A. (2020). Winners and losers from COVID-19: Global evidence from Google search. *World Bank Policy Research Working Paper*, No. 9268. Washington D. C.: World Bank. https://papers.ssrn.com/abstract=3617347.

Anupam, A., & Lawal, I. (2023). Forecasting air passenger travel: A case study of Norwegian aviation industry. *Journal of Forecasting*, *43*(3), 661–672. https://doi.org/10.1002/for.3051.

Assimakopoulos, V., & Nikolopoulos, K. (2000). The theta model: a decomposition approach to forecasting. *International Journal of Forecasting*, *16*(4), 521–530. https://doi.org/10.1016/S0169-2070(00)00066-2.

Baker, S., & Fradkin, A. (2017). The impact of unemployment insurance on job search: Evidence from google search data. *Review of Economics and Statistics*, *99*(5), 756–768. http://dx.doi.org/10.1162/REST_a_00674.

Box, G., & Jenkins, G. (1976). *Time series analysis, forecasting and control*. San Francisco: Holden-Day.

Brown, R. (1959). *Statistical forecasting for inventory control*. New York: McGraw-Hill.

Cournède, B., V. Ziemann and F. De Pace (2020). Housing amid Covid-19: Policy responses and challenges. *OECD Policy Responses to Coronavirus (COVID-19).* OECD Publishing, Paris.

Dinis, G., Costa, C., & Pacheco, O. (2017). Forecasting British tourist inflows to Portugal using Google trends data. In V. Katsoni, A. Upadhya & A. Stratigea (eds.), *Tourism, culture and heritage in a smart economy* (pp. 483-496). Springer International Publishing. https://doi.org/10.1007/978-3-319-47732-9_32.

Doerr, S., & Gambacorta, L. (2020). Identifying regions at risk with Google trends: the impact of Covid-19 on US labour markets. *BIS Bulletins*. https://ideas.repec.org/p/bis/bisblt/8.html.

Gonzales, F., Jaax, A., & Mourougane, A. (2020). *Nowcasting aggregate services trade. A pilot approach to providing insights into monthly balance of payments data*. Paris: OECD.

Holt, C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, *20*(1), 5–10. https://doi.org/10.1016/j.ijforecast.2003.09.015.

Hyndman, R., & Athanasopoulos, G. (2021). F*orecasting: principles and practice*. Melbourne: OTexts.

Hyndman, R., & Koehler, A. (2006). Another look at measures of forecast accuracy. *Internationa*l *Journal of Forecasting*, *22*(4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001.

Hyndman, R., Koehler, A., Snyder, R., & Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, *18*(3), 439–454. https://doi.org/10.1016/S0169-2070(01)00110-8.

Kolková, A. (2019). Application of artificial neural networks for forecasting in business economy. In *Proceedings of the 7th international conference Innovation Management, Entrepreneurship And Sustainability (IMES 2019)* (pp. 359–368). Vysoká škola eknomická v Praze, Prague, Czech Republic.

Kolková, A. (2020). The application of forecasting sales of services to increase business competitiveness. *Journal of Competitiveness*, *12*(2), 90–105. https://doi.org/10.7441/joc.2020.02.06.

Kolková, A., & Ključnikov, A. (2021). Demand forecasting: an alternative approach based on technical indicator Pbands. *Oeconomia Copernican*a, *12*(4), 1063–1094. https://doi.org/10.24136/oc.2021.035.

Kolková, A., & Ključnikov, A. (2022). Demand forecasting: AI-based, statistical and hybrid models vs practice-based models - the case of SMEs and large enterprises. *Economics and Sociology*, 15(4), 39–62. https://doi.org/10.14254/2071-789X.2022/15-4/2.

Kolková, A., & Navrátil, M. (2021). Demand forecasting in Python: Deep learning model based on LSTM architecture versus statistical models. *Acta Polytechnica Hungarica*, *18*(8), 123–141. https://doi.org/10.12700/APH.18.8.2021.8.7.

Kolkova, A., & Rozehnal, P. (2022). Hybrid demand forecasting models: pre-pandemic and pandemic use studies. *Equilibrium*, *17*(3), 699–725. https://doi.org/10.24136/eq.2022.024.

Korayim, D., Chotia, V., Jain, G., Hassan, S., & Paolone, F. (2024). How big data analytics can create competitive advantage in high-stake decision forecasting? The mediating role of organizational innovation. *Technological Forecasting and Social Change*, *199*. https://doi.org/10.1016/j.techfore.2023.123040.

Morgavi, H. (2020). A GARCH model to now-cast private consumption using Google trends data. *OECD Economics Department Working Paper*. Paris: OECD.

Orue, A., Lizarralde, A., Apaolaza, U., & Amorrortu, I. (2023). Designing the process of implementing step three of the theory of constraints in a make-to-order environment: Integrating sales and operation planning. *Journal of Industrial Engineering and Managemen*t, *16*(2), 205–214. https://doi.org/10.3926/jiem.5127.

Pisu, M., Costa, H., & Hwang, H. (2020). Digital platforms and the COVID-19 crisis. *OECD Economics Department Working Paper*. Paris: OECD.

Shafique, M., Yeo, S., & Tan, C. (2024). Roles of top management support and compatibility in big data predictive analytics for supply chain collaboration and supply chain performance. *Technological Forecasting and Social Change*, *199*. https://doi.org/10.1016/j.techfore.2023.123074.

Shaub, D., & Ellis, P. (2020). Package 'forecastHybrid' (p. 27). https://gitlab.com/dashaub/forecastHybrid.

Sivarajah, U., Kamal, M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of big data challenges and analytical methods. *Journal of Business Research*, *70*, 263–286. https://doi.org/10.1016/j.jbusres.2016.08.001.

Swaminathan, K., & Venkitasubramony, R. (2024). Demand forecasting for fashion products: A systematic review. *International Journal of Forecasting*, *40*(1), 247–267. https://doi.org/10.1016/j.ijforecast.2023.02.005.

Taylor, S., & Letham, B. (2018). Forecasting at Scale. *The American Statistician*, *72*(1), 37–45. https://doi.org/10.1080/00031305.2017.1380080.

Varian, H., & Choi, H. (2009). Predicting the present with Google trends. *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.1659302.

Wang, L., & Chen, H. (2022). Tourism demand forecast based on adaptive neural network technology in business intelligence. *Computational Intelligence and Neuroscience*, *2022*, 1–14. https://doi.org/10.1155/2022/3376296.

Winters, P. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science*, *6*(3), 324–342. https://doi.org/10.1287/mnsc.6.3.324.

Wu, C., Li, J., Liu, W., He, Y., & Nourmohammadi, S. (2023). Short-term electricity demand forecasting using a hybrid ANFIS–ELM network optimised by an improved parasitism–predation algorithm. *Applied Energy*, *345*. https://doi.org/10.1016/j.apenergy.2023.121316.

Xie, G., Liu, S., & Li, X. (2023). An interval decomposition-ensemble model for tourism forecasting. *Journal of Hospitality & Tourism Research*. https://doi.org/10.1177/10963480231198539.

The research article passed the double-blind review process. | Received: 31 January 2024**; Revised:** 5 March 2024**; Accepted:** 17 March 2024**; Available online:** 24 July 2024**; Published in the regular issue:** 31 December 2024.