SILESIAN UNIVERSITY OF TECHNOLOGY PUBLISHING HOUSE

SCIENTIFIC PAPERS OF SILESIAN UNIVERSITY OF TECHNOLOGY 2023
ORGANIZATION AND MANAGEMENT SERIES NO. 177

APPLICATION OF THE KNOWN SUB-SEQUENCE ALGORITHM TO SELECT THE IMPUTATION METHOD FOR TIME SERIES OF ELECTRIC ENERGY CONSUMPTION

Agnieszka KOWALSKA-STYCZEŃ[1*], Adam SOJDA[2], Maciej WOLNY[3]

[1] Silesian University of Technology, Faculty of Organization and Management, Department of Economics and Informatics; agnieszka.kowalska-styczeń@polsl.pl, ORCID: 0000-0002-7404-9638
[2] Silesian University of Technology, Faculty of Organization and Management, Department of Economics and Informatics; adam.sojda@polsl.pl, ORCID: 0000-0002-3021-4451
[3] Silesian University of Technology, Faculty of Organization and Management, Department of Economics and Informatics; maciej.wolny@polsl.pl, ORCID: 0000-0002-8872-7794
* Correspondence author

Purpose: The key element of effective electricity management is to improve the accuracy of forecasting its consumption. To create a forecast, data on customers' energy consumption in previous periods is required, and the accuracy of the forecasts depends on the quality and availability of data. The acquired historical data is often incomplete and contains missing values. The aim of the article is therefore to choose an appropriate method of imputation of missing values for one-dimensional time series of energy consumption.
Design/methodology/approach: The aim of the article was achieved by using the Known Substring Algorithm (KSSA) to verify the imputation precision. The KSSA algorithm allowed to test of eleven imputation methods, most of which are implemented in the 'ImputeTS' package in R. Based on the RMSE error, the best imputation method was selected for the analyzed series.
Findings: As a result of the analyzes carried out, it was shown that the KSSA algorithm is a good tool for selecting the appropriate imputation method in the case of one-dimensional series of electricity consumption series. Based on the RMSE error, 'auto.arima' turned out to be the best imputation method for the analyzed objects
Research limitations/implications: Future research will concern the use of the KSSA algorithm for a larger number of energy consumption series and with greater variation.
Originality/value: The article presents an important problem of the imputation of missing values in the electricity consumption series. Increasing the accuracy of electricity consumption forecasting depends on the quality of the collected data, which are often incomplete and contain missing values. Therefore, the selection of the appropriate imputation method is so important.
Keywords: Time series, Missing data imputation, Electricity consumption data, Data quality, Missing value.

Category of the paper: Research paper.

http://dx.doi.org/10.29119/1641-3466.2023.177.18

1. Introduction

In this study, we analyze data on electricity consumption because, as Wang et al. (2021) emphasised, economic development leads to an increase in electricity demand and, consequently, generates the need for energy-saving measures, i.e., better energy management systems. Such systems are mainly dedicated to electricity consumers, which is why we analyze data on electricity consumption from individual consumers. To create a forecast, data on a customer's energy consumption in previous periods, i.e., historical data, is required. It should be noted that the accuracy of electricity consumption forecasting depends on the quality of the collected data (Kim et al., 2019; Chen et al., 2017). The obtained historical data often lacks completeness and contains missing values. This is a common issue when data is measured and recorded (Moritz, Bartz-Beielstein, 2017; Sefidian, Daneshpour, 2019). In the case of energy consumption data, these can be communication errors, sensor failures, power outages (Bokde et al., 2018), but also deficiencies due to the lack of readings (values are then not measured).

There are many different techniques that can be used to deal with missing values (Liu et al., 2020; Garcia-Laencina et al., 2010). These include missing values deletion, mean substitution, and model-based imputation. When the data set contains less than approximately 10-15% missing data, it can be simply removed from the data set (Strike et al., 2001). However, as shown by Lin and Tsai (2019), even small amounts of missing data can have a significant impact on the final analysis results. An appropriate approach to handling missing values in our analyzed data is imputation, which is one of the most reliable methods for dealing with missing values (Demirhan, Renwick, 2018). In the literature, various algorithms can be found for replacing missing data with estimated values. The most common data imputation techniques rely on correlations between attributes to estimate values for missing data. These include Multiple Imputation (Rubin, 1987), Expectation-Maximization (Dempster et al., 1977), Nearest Neighbor (Vacek, Ashikaga, 1980), and Hot Deck (Ford, 1983). In the case of electricity consumption data, we often deal with one-dimensional data series where additional attributes are missing. Therefore, imputation algorithms specifically tailored to such data should be applied (Moritz, Bartz-Beielstein, 2017; Kowalska-Styczeń et al., 2022). For example, Bokde et al. (2018) propose the imputePSF method, which is a modification of the pattern sequence based forecasting (PSF) method. Demirhan and Renwick (2018), on the other hand, compare the performance of methods available in the "imputeTS" package, which are dedicated to one-dimensional time series with irregular intervals.

An important element to pay attention to when using imputation methods is the type of data. In the case of electricity consumption data, these are usually one-dimensional time series without additional attributes. The aim of our article is to find the best method of imputation of missing values for one-dimensional electricity consumption series.

As mentioned earlier, effective energy management is very important for electricity trading companies. Such companies often have to buy energy on the wholesale market and then distribute it to individual customers. In this process, there is a need to ensure continuous and accurate balancing of electricity demand and production, i.e. better forecasts of energy consumption. To prepare the forecast, data on energy consumption by customers in previous periods is required. The accuracy of the forecast depends on the data, which often contains missing data. The results of our work may therefore be interesting for energy trading companies that are looking for efficient tools for the imputation of missing values.

After analyzing the available data imputation tools, we chose the Known Sub-Sequence Algorithm (KSSA) proposed by Benavides et al. (2021). The cited authors used this algorithm to assign missing values in the time series for landings of six fish species. We noticed similarities between the series they analyzed and the series of electricity consumption (in both cases these are one-dimensional time series). Our approach allows for the selection of the best imputation method by comparing 12 imputation methods from the "ImputeTS" (Moritz, Bartz-Beielstein, 2017), "forecast" (Hyndman, Khandakar, 2008), and "Rssa" (Golyandina, Korobeynikov, 2014) packages. To select the best method, RMSE and MASE errors are computed between the actual and imputed time series.

We believe that the proposed approach is a good solution that can be used by trading companies.

2. Data structure

The article used historical data on electricity consumption for 6 homes (facilities) in British Columbia (Makonin, 2019). The data for selected facilities were analyzed for missing values. The number of days with missing data, the number of data gaps (where a gap is defined as one or more consecutive days with missing data), the average gap size, the longest data gap, and the percentage of missing data were calculated. Further details can be found in Table 1.

Table 1.
The data structure for the facilities, based on missing data

FACILITY	start of observation	end of observation	number of days	number of days with missing data	number of gaps	average gap size	the longest gap	missing data %
Facility A	January 27, 2015.	January 29, 2018.	1099	57	39	1.46	14	5.19
Facility B	February 21, 2015.	February 20, 2018.	1096	54	43	1.26	6	4.93

Cont. table 1.

Facility C	February 21, 2015.	February 20, 2018.	1096	40	35	1.14	5	3.65
Facility D	November 01, 2017.	February 18, 2019.	475	90	37	2.43	10	18.95
Facility E	November 01, 2017.	June 05, 2018.	217	9	9	1.00	1	4.15
Facility F	July 27, 2017.	April 05, 2020.	984	59	45	1.31	14	6.00

As indicated in Table 1, series of different lengths and structures were selected for analysis. Subsequently, the selected series were analyzed in terms of electricity consumption in kWh. The minimum, maximum, and average energy consumption in each series, along with the standard deviation of consumption and median, were identified. Details are provided in Table 2.

Table 2.
Data structure for facilities by electricity consumption

FACILITY	minimum value in series	maximum value in series	average daily consumption in kWh	standard deviation of energy consumption	median daily consumption
Facility A	11.23	41.89	22.2	5.1	21.30
Facility B	5.8	38.31	16.0	5.3	15.13
Facility C	6.57	50.83	21.8	7.6	21.00
Facility D	0.69	46.28	13.9	8.6	11.72
Facility E	1.63	17.12	5.5	2.7	5.11
Facility F	1.11	66.83	13.1	10.7	9.84

The distribution of missing data in the consumption series is shown in Figure 1. The figure displays daily data from 6 objects, with missing values marked in red.

As can be observed in Figure 1, time series of electrical energy consumption can exhibit different characteristics. In particular, the distributions of missing data vary greatly. Based on the analysis of the plots presented in Figure 1, it can be assumed that the mechanism of missing data is random (Liu et al., 2020; Kowalska-Styczeń et al., 2022; Sefidian, Daneshpour, 2019).

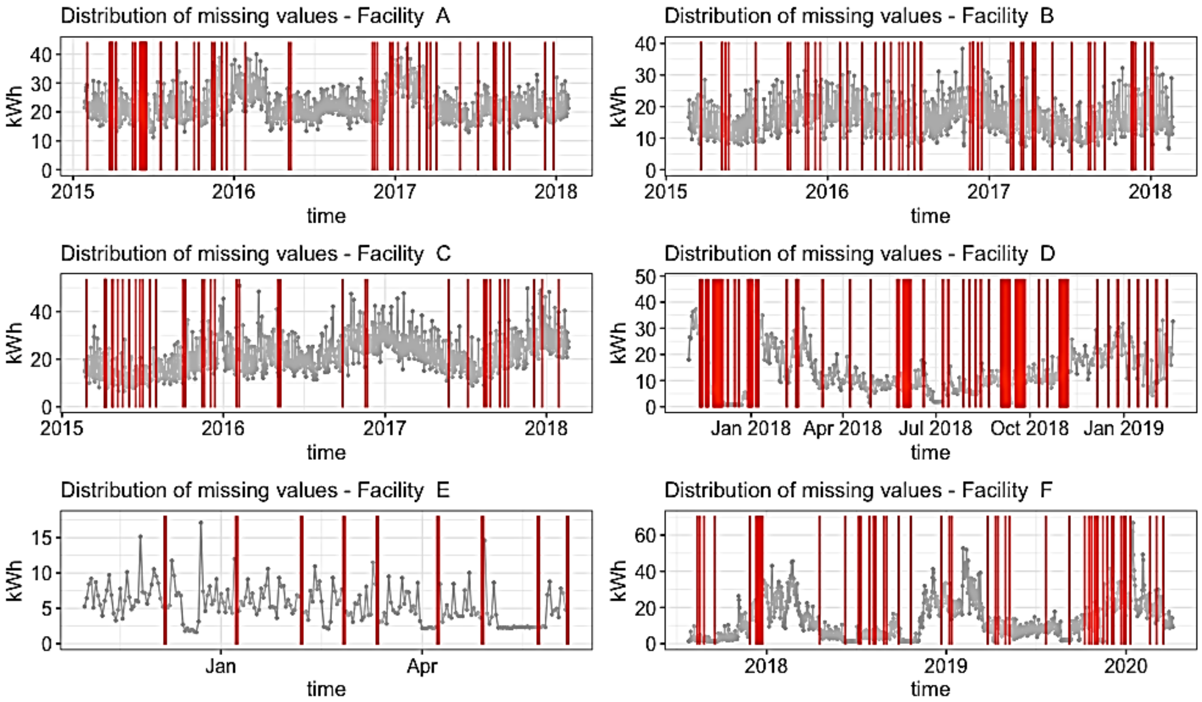

Figure 1. Time series with marked missing data.

Source: own study.

3. Methods

In this article, we employ the Known Sub-Sequence Algorithm (KSSA), which is based on Multiple Imputation. Originally, this algorithm was used for imputing missing values in one-dimensional time series of marine catch offloading for six fish species in the Colombian Pacific (Benavides et al., 2021). Since electrical energy consumption data is often in the form of one-dimensional time series, the use of the KSSA algorithm seems appropriate for our research.

The operation of the KSSA algorithm can be described as follows:

1. the true missing data (MD) in the original time series under investigation is imputed using a single imputation algorithm, x;
2. the time series is divided into time segments (year, quarter, month, etc.);
3. moving and expanding windows are set in each segment, simulating chunks of missing data of varying sizes in (known) sub-sequences that do not contain original missing data;
4. imputing simulated chunks of missing data in each segment using a single x-algorithm and calculating the errors (RMSE, MASE "metrics" of the R package) between the actual and imputed values in the time series;
5. repeating steps 1-4 n times using bootstrapping;
6. repeating steps 1-5 for all algorithms;

7. repeating steps 1-6 for all time series;
8. all results are saved in a final data frame containing the size of missing data, RMSE, and MASE for n runs of each algorithm and time series.

KSSA utilizes missing value imputation methods mainly from the "ImputeTS" package in R (Moritz, Bartz-Beielstein, 2017), which include:

- "State space representation of an ARIMA model" (auto.arima) - imputations using an ARIMA model (Hyndman, Khandakar, 2008);
- "State space representation of a structural model" (StructTS) - imputations using a structural time series model (Harvey, 1990; Durbin, Koopman, 2001);
- "Seasonal decomposition with Kalman smoothing" (seadec) - imputations on seasonally adjusted series using Kalman smoothing (Aravkin et al., 2017) and then considering the seasonal component in imputations (Cleveland et al., 1990);
- "Linear interpolation" (linear_i) - imputations using linear interpolation;
- "Spline interpolation" (spline_i) - imputations using cubic spline interpolation (Hall, Meyer, 1976);
- "Stineman interpolation" (stine_i) - imputations using Stineman interpolation (Stineman, 1980);
- "Simple moving average" (simple_ma) - imputations using simple moving average of neighboring observations;
- "Linear moving average" (linear_ma) - imputations using linearly weighted moving average;
- "Exponential moving average" (exponential_ma) - imputations using exponentially weighted moving average;
- "Last observation carried forward" (locf) - imputation by replacing the missing data with the last known observation;
- "Seasonal and trend decomposition with Loess" (stl) - imputations using time series decomposition with local smoothing (Cleveland et al., 1990).

The description of our procedure, where we utilize the KSSA algorithm, is as follows:

1. Perform a time series analysis - calculate the percentage of missing data.
2. Determine one of the evaluation criteria for imputation methods: correlation coefficient, RMSE, MASE, SMAPE - RMSE is chosen.
3. Choose one of the available methods for missing data imputation: auto.arima, StructTS, seadec, linear_i, spline_i, stine_i, simple_ma, linear_ma, exponential_ma, locf, stl.
4. For the selected method and time series, set the parameters of the KSSA algorithm:
   - Time series with missing data.
   - Imputation method - chosen in step 3.
   - Number of segments - 5 segments.

- Number of iterations - 100 observations.
- Percentage of missing data - consistent with the percentage determined for each object.

5. Run the KSSA algorithm and obtain the imputation results.
6. Repeat steps 3-5 for different imputation methods.
7. Based on the established criterion, select the best imputation method.

The chosen error measure, RMSE (root-mean-squared error), calculates the square root of the average squared difference between the actual and imputed values. It measures the average deviation of the actual variables from the imputed values.

4. Results and discussion

The procedure proposed in section 2 was used to analyze electricity consumption for 6 objects (6 houses). As was shown earlier, the analyzed time series of electricity consumption show different characteristics (especially the distribution of missing data is very diverse).

The purpose of the analysis is to find the best imputation method. The chosen error measure, RMSE (mean squared error), measures the average deviation of real variables from imputed values. The smaller the RMSE, the better the method of imputation of missing values.

The results of applying the proposed method are shown in Figures 2-7. The figures show the RMSE errors for each object and each of the 11 imputation methods for missing values.

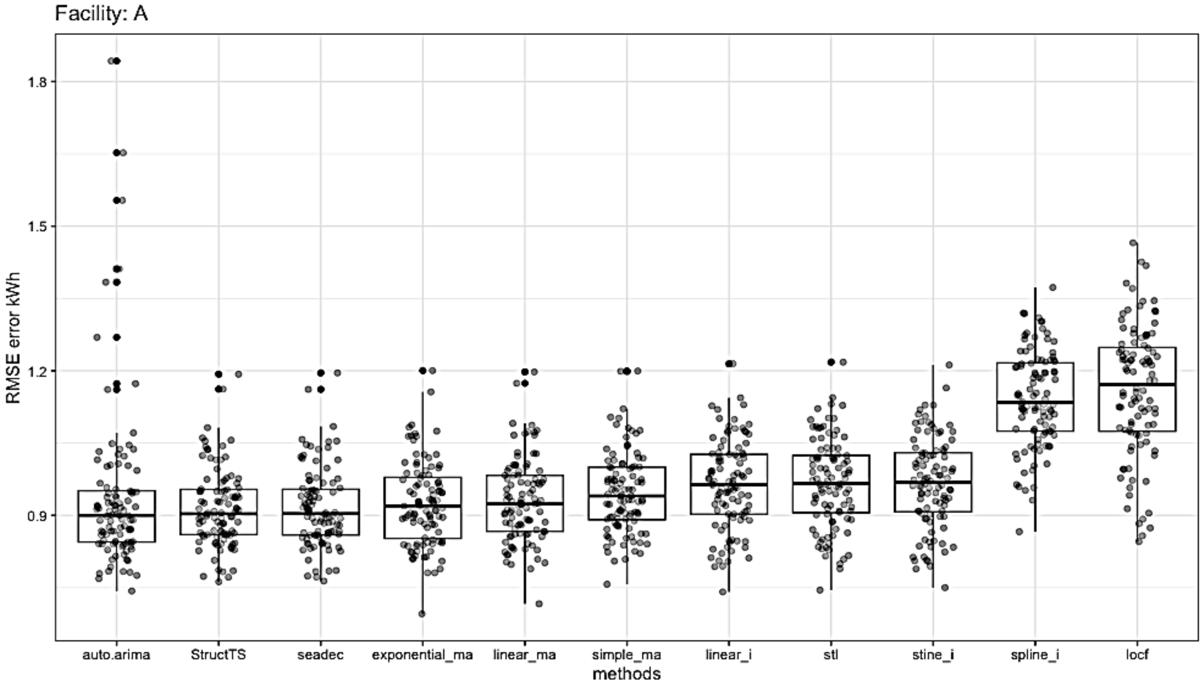

Figure 2. Distributions of RMSE error values for different imputation methods for Facility A.

Source: own study.

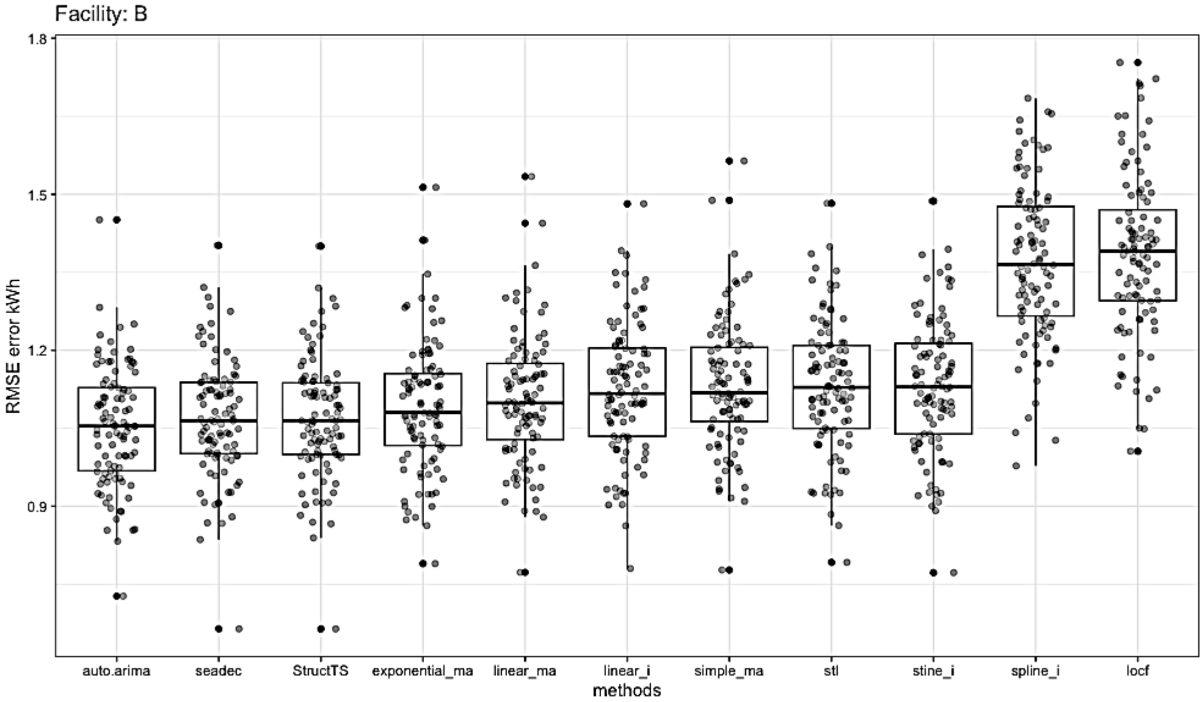

Figure 3. Distributions of RMSE error values for different imputation methods for Facility B.
Source: own research.

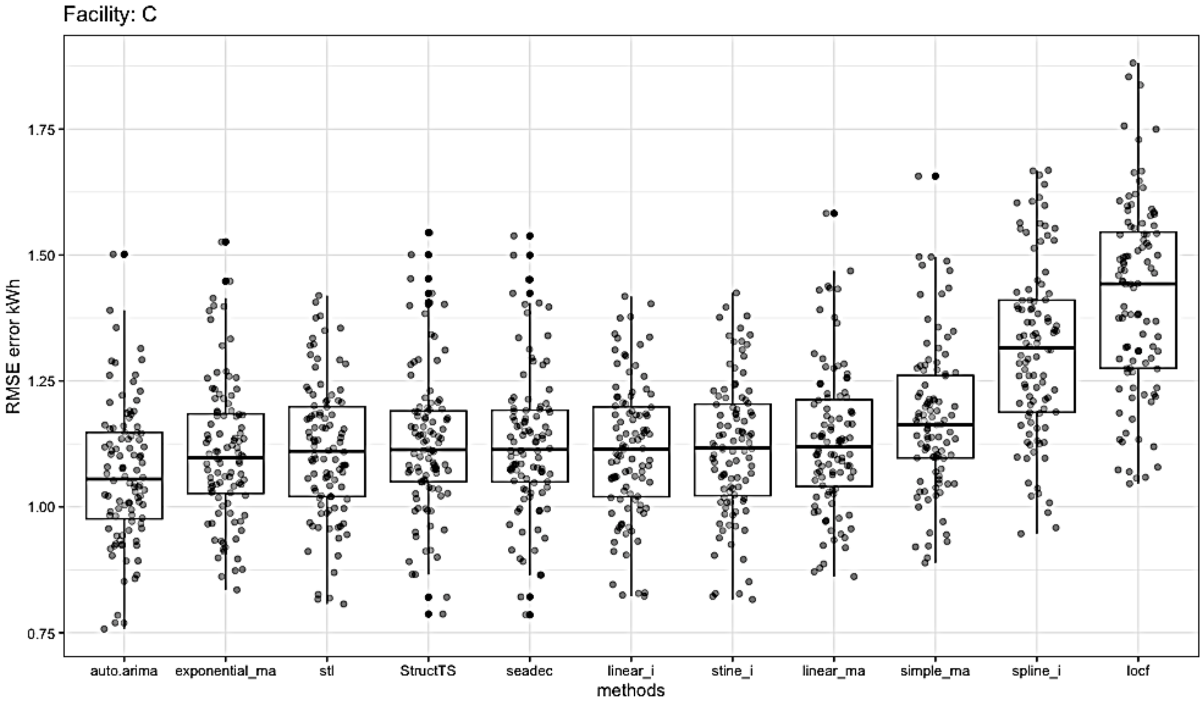

Figure 4. Distributions of RMSE error values for different imputation methods for Facility C.
Source: own research.

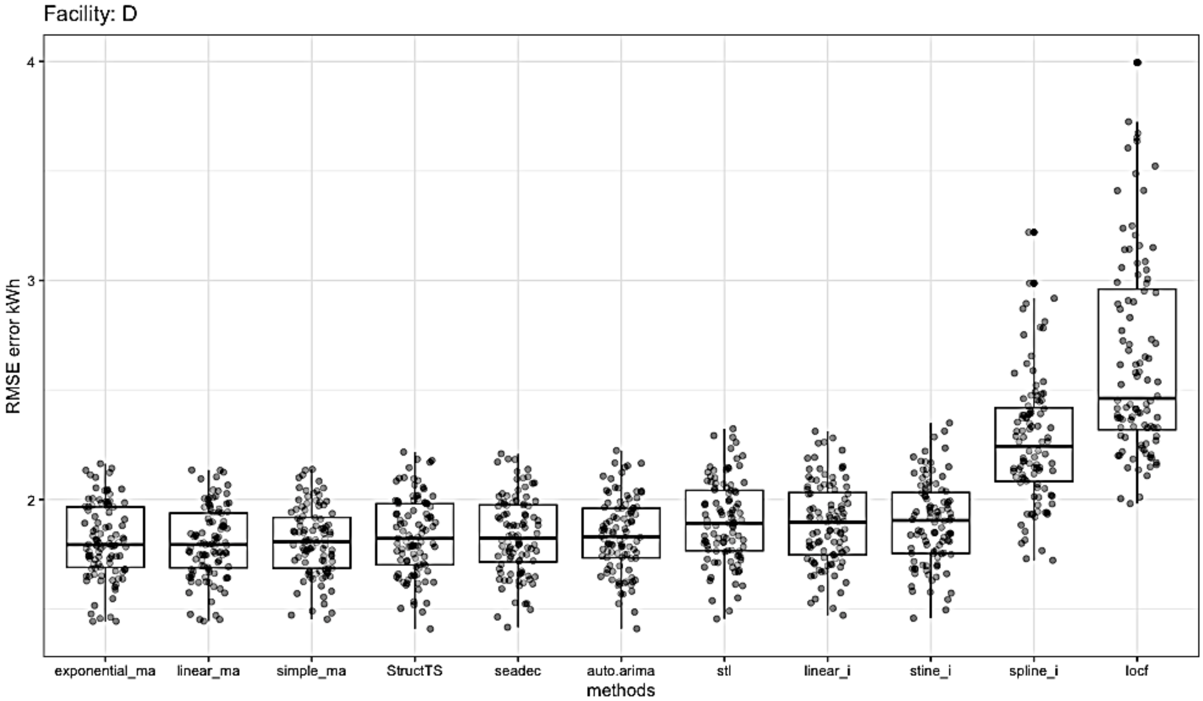

Figure 5. Distributions of RMSE error values for different imputation methods for Facility D.

Source: own research.

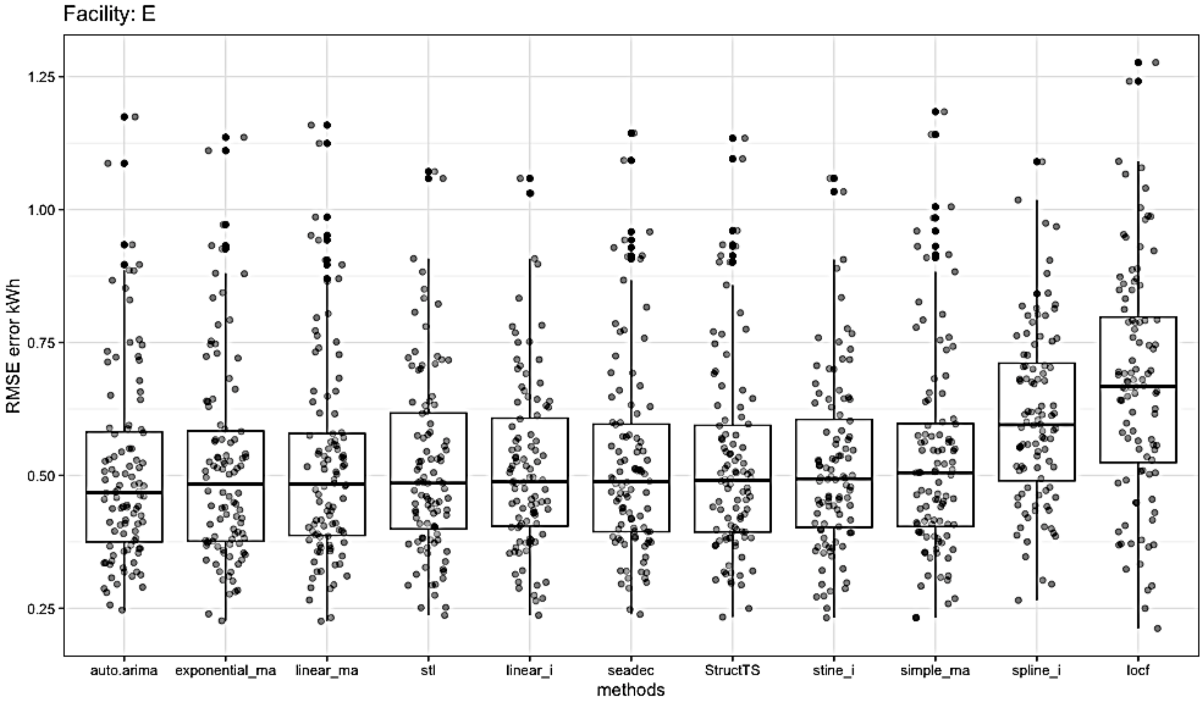

Figure 6. Distributions of RMSE error values for different imputation methods for Facility E.

Source: own research.

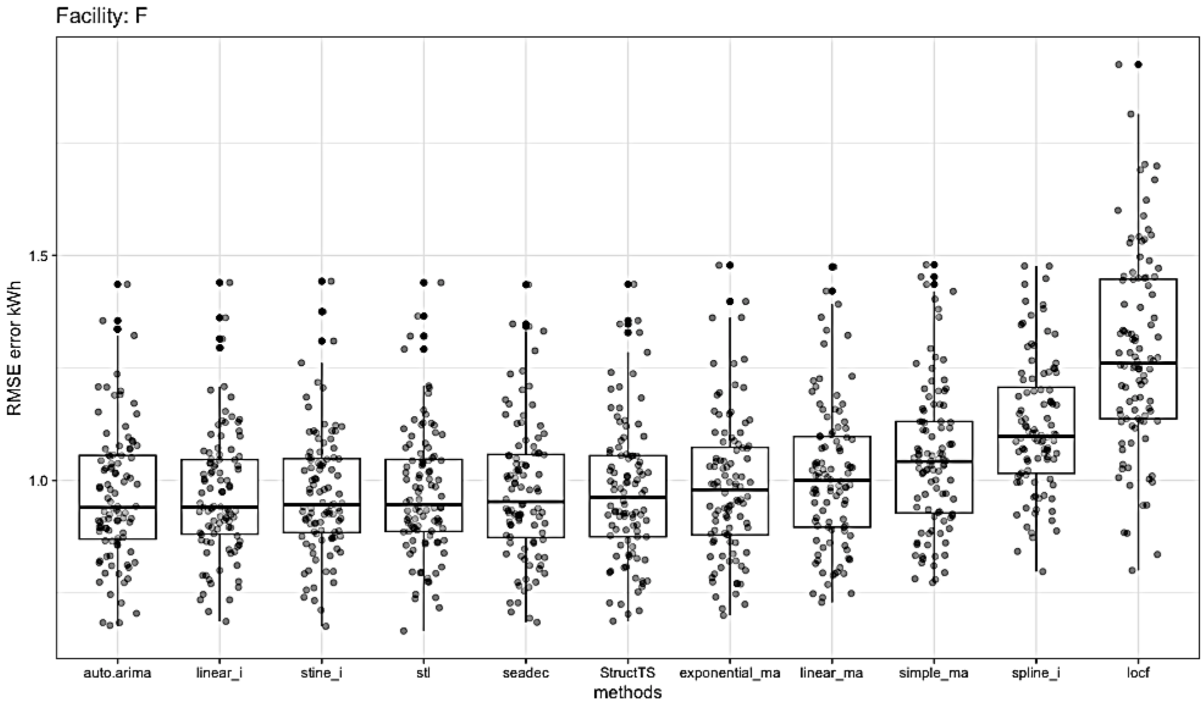

Figure 7. Distributions of RMSE error values for different imputation methods for Facility F.

Source: own research.

As can be seen from Figures 2-7, in most cases the RMSE error is smallest for the ARIMA method (auto_arima).

In order to select the best imputation method, ranks were used. As shown in Table 3, the lowest imputation error (RMSE) is generated by the 'auto.arima' method for 5 out of 6 facilities. Only for Facility D, the best imputation method was found to be 'exponential_ma'. As indicated in Table 1, in this case, the percentage of missing values was the highest (around 19%), which is approximately 4 times higher than in the other cases.

Table 3.
Ranks for facilities due to RMSE

method	facility						facility rank						SUM
	A	B	C	D	E	F	A	B	C	D	E	F	
auto.arima	0.90	1.05	1.06	1.83	0.47	0.94	1	1	1	6	1	1	11
exponential_ma	0.92	1.08	1.10	1.79	0.48	0.98	4	4	2	1	2	7	20
seadec	0.90	1.06	1.11	1.82	0.49	0.95	3	2	5	5	6	5	26
StructTS	0.90	1.06	1.11	1.82	0.49	0.96	2	3	4	4	7	6	26
linear_ma	0.92	1.10	1.12	1.79	0.48	1.00	5	5	8	2	3	8	31
linear_i	0.96	1.12	1.11	1.90	0.49	0.94	7	6	6	8	5	2	34
stl	0.97	1.13	1.11	1.89	0.49	0.95	8	8	3	7	4	4	34
simple_ma	0.94	1.12	1.16	1.81	0.51	1.04	6	7	9	3	9	9	43
stine_i	0.97	1.13	1.12	1.90	0.49	0.95	9	9	7	9	8	3	45
spline_i	1.13	1.37	1.32	2.24	0.60	1.10	10	10	10	10	10	10	60
locf	1.17	1.39	1.44	2.46	0.67	1.26	11	11	11	11	11	11	66

We are looking for a universal method of the imputation of missing values that can be used for the series of energy consumption from individual consumers. As pointed out by Moritz et al. (2015), one-dimensional time series is a particular challenge in the field of imputation research.

One-dimensional series of energy consumption in the context of missing data were previously analyzed for business customers data (Kowalska-Styczeń et al., 2022). The analysis of the structure of this data prompted the authors to choose three methods of the imputation of missing data: the calendar method, the imputation method by separating the phases of seasonal cycles and the imputation method using seasonality decomposition. In this article, also methods from "ImputeTS" package in R were used, such as "Seasonal decomposition with Kalman smoothing" (seadec), "Seasonally Splitted Missing Value Imputation" (na_seasplit), "Simple moving average" (simple_ma), "Linear moving average" (linear_ma) and "Exponential moving average" (exponential_ma).

The data analyzed in this article concerns individual customers and their structure is different than in the case of business customers (especially the distribution of missing data was very diverse). Therefore, a procedure based on the KSSA algorithm, which allows for simultaneous testing of many imputation methods for one-dimensional series, has been proposed. This may be interesting from the point of view of practitioners (e.g. energy trading companies). The proposed automated method of dealing with missing values may contribute to the improvement of electricity consumption forecasts.

5. Conclusion

The increasing demand for electricity necessitates the need for energy conservation, which includes the implementation of improved energy management systems, even for households. In this context, the key aspect is the application of methods that enhance the accuracy of individual consumption forecasts at specific points of energy consumption. Missing values in historical data pose a barrier to improving forecasting accuracy for individual consumption points. In this case, an estimate of the deficiencies in the historical time series has to be made and then the missing values should be replaced with these estimates, which is called missing data imputation or gap filling. As shown earlier, there are many methods and approaches for the imputation issue. However, it should be noted that univariate time series require an individual approach to data imputation problems as they do not contain additional attributes. Therefore, we propose utilizing a method based on the KSSA algorithm, which allows for testing multiple methods of missing values imputation. This can be a great convenience for practitioners interested in improving the quality of data and, consequently, improving electricity consumption forecasts.

For our analysis, we selected objects characterized by varying lengths of energy consumption series and different statistics regarding missing values and energy consumption. The results of our analysis demonstrate that the KSSA algorithm is a reliable tool for imputing missing values in electricity consumption series. For 5 out of 6 facilities, the 'auto.arima' method proved to be the best imputation method. In further research, we intend to analyze a larger number of objects with diverse characteristics to ultimately confirm the selection of this method as the best for imputing missing values in household electricity consumption series.

Acknowledgements

This paper was published as part of the statutory research ROZ 1: BK-274/ROZ1/2023 (13/010/BK_23/0072) at the Silesian University of Technology, Faculty of Organization and Management.

References

1. Aravkin, A., Burke, J.V., Ljung, L., Lozano, A., Pillonetto, G. (2017). Generalized Kalman smoothing: Modeling and algorithms. *Automatica, 86,* 63-86.
2. Benavides, I.F., Santacruz, M., Romero-Leiton, J.P., Barreto, C., Selvaraj, J.J. (2021). Assessing methods for multiple imputation of systematic missing data in arine fisheries time series with a new validation algorithm. *Aquaculture and Fisheries,* https://doi.org/10.1016/j.aaf.2021.12.013.
3. Bokde, N., Beck, M.W., Álvarez, F.M., Kulat, K. (2018). A novel imputation methodology for time series based on pattern sequence forecasting. *Pattern recognition letters, 116*, 88-96.
4. Chen, W., Zhou, K., Yang, S., Wu, C. (2017). Data quality of electricity consumption data in a smart grid environment. *Renew. Sust. Energy Rev., 75*, 98-105.
5. Cleveland, R.B., Cleveland, W.S., McRae, J.E., and Terpenning, I. (1990). STL: A seasonal-trend decomposition. *J. Off. Stat, 6(1),* 3-73.
6. Demirhan, H., Renwick, Z. (20180). Missing value imputation for short to mid-term horizontal solar irradiance data. *Appl. Energy, 225(9),* 98-1012.
7. Durbin, J., Koopman, S.J. (2001). *Time Series Analysis by State Space Methods*. Oxford University Press.

8. García-Laencina, P.J., Sancho-Gómez, J.L., Figueiras-Vidal, A.R. (2010). Pattern classification with missing data: a review. *Neural Computing and Applications, 19,* 263-282.
9. Golyandina, N., Korobeynikov, A. (2014). Basic singular spectrum analysis and forecasting with R. *Computational Statistics & Data Analysis, 71,* 934-954.
10. Hall, C.A., Meyer, W.W. (1976). Optimal error bounds for cubic spline interpolation. *Journal of Approximation Theory, 16(2),* 105-122.
11. Harvey, A.C. (1990) *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press.
12. Hyndman, R.J., Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of statistical software, 27,* 1-22.
13. Kim, T., Ko, W., Kim, J. (2019). Analysis and impact evaluation of missing data imputation in day-ahead PV generation forecasting. *Appl. Sci., 9,* 204.
14. Kowalska-Styczeń, A., Owczarek, T., Siwy, J., Sojda, A., Wolny, M. (2022). Analysis of Business Customers' Energy Consumption Data Registered by Trading Companies in Poland. *Energies, 15(14),* 5129.
15. Li, H., Li, M., Lin, X., He, F., Wang, Y. (2020). A spatiotemporal approach for traffic data imputation with complicated missing patterns. *Transportation Research, Part C, 119,* 102730.
16. Liu, C.-H., Tsai, C.-F., Sue, K.-L., Huang, M.-W. (2020). The Feature Selection Effect on Missing Value Imputation of Medical Datasets. *Appl. Sci., 10,* 2344.
17. Makonin, S. (2019). HUE: The hourly usage of energy dataset for buildings in British Columbia. *Data in brief, 23*.
18. Moritz, S., Bartz-Beielstein, T. (2017). "imputeTS: Time Series Missing Value Imputation in R. *" _The R Journal_, *9*(1),* 207-218. doi:10.32614/RJ-2017-009.
19. Moritz, S., Sardá, A., Bartz-Beielstein, T., Zaefferer, M., Stork, J. (2015). *Comparison of different Methods for Univariate Time Series Imputation in R,* arXiv:physics/1510.03924
20. Sefidian, A.M., Daneshpour, N. (2019). Missing value imputation using a novel grey based fuzzy c-means, mutual information based feature selection, and regression model. *Expert Systems with Applications, 115,* 68-94.
21. Stineman, R.W. (1980). A consistently well-behaved method of interpolation. *Creative Computing, 6(7),* 54-57.
22. Wang, M.-C., Tsai, C.-F., Lin, W.-C. (2021). Towards missing electric power data imputation for energy management systems. *Expert Syst. Appl., 174,* 14743.