Energy Proceedings
Vol 41, 2024
ISSN 2004-2965

Systematic Comparison of Imputation Models for Automatized Gap Filling on Electrical Load Data of Compressor Composites in the Industrial Sector

Anna Harman [1,2*], Lukas Baur [1,2], Alexander Sauer [1,2]

1 Fraunhofer Institute for Manufacturing Engineering and Automation IPA, Stuttgart, Germany

2 Institute for Energy Efficiency in Production (EEP), University of Stuttgart, Stuttgart, Germany
(*Corresponding Author: anna.harman@ipa.fraunhofer.de)

ABSTRACT

With increasing digitization for constructing intelligent energy systems, automated data processing is moving more and more into focus. Gaps in the recorded data pose a central problem for further processing instances. This work systematically investigates which methods are suitable for the imputation of data gaps of different sizes. It tackles the imputation performance's influence on overlying applications, such as load forecasting and total energy determination. The presented method is applied to four datasets of compressors of industrial. Based on these Use Case's evaluation results, recommendations for action are derived. Gap sizes should be considered when choosing an imputation method to minimize imputation error. For load forecasting, the prediction error correlates with the imputation error in certain missingness scenarios. Energy consumption analysis on the imputed data yields good results due to a balanced ratio of over- and undershooting of the imputation error.

Keywords: data imputation, automation, intelligent energy system, industrial energy system, data pre-processing, load forecasting

NONMENCLATURE

Abbreviations	
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error

1. INTRODUCTION

According to the German Federal Ministry for Economic Affairs and Energy study, which examined three scenarios to determine the techno-economic impact of different paths to decarbonizing the energy system and making Germany greenhouse gas neutral by 2050, the increased use of renewable electricity and energy efficiency play a key role [1]. Energy load data plays an essential role in various aspects such as energy load forecasting, efficiency improvements, and determining energetic transparency in order to be able to determine carbon emissions aggregated in the factory or product-specific. However, gaps in energy load data pose significant challenges in these tasks. Appropriate data preprocessing steps, including the application of reliable imputation methods are necessary. The effective handling of gaps in energy load data ensures reliable and complete datasets that are fundamental for efficient load management in modern industrial energy systems, contributing to increased energy flexibility and the transition to renewable energy sources.

This study investigates the research question of how the proportion and distribution of missing values and the chosen imputation method influence the quality of subsequent use cases that build on the imputed data.

1.1 Related Work

The performance of load management mainly depends on the data quality. However, missing data in energy time series is not uncommon [2]. One way to handle missing data is imputation. Sridevi et al. [3] propose an autoregressive-model-based missing value imputation, particularly effective when a time point is mainly or entirely missing. Kanda et al. [4] inserted data from similar meters in the missing periods before implementing a probabilistic load forecast in the GEFCom2017 final match competition. Lee et al. [5] introduced an imputation method using both univariate and multivariate imputation techniques combining spline interpolation and Expectation Maximization based on maximum likelihood estimation. Ryu et al. [6] applied

# This is a paper for 15th International Conference on Applied Energy (ICAE2023), Dec. 3-7, 2023, Doha, Qatar.

a denoising autoencoder for imputation on smart meter data with different missingness patterns, like random, block-wise, with different configurations and predefined missing scenarios. Khan et al. [7] presented a hybrid energy-forecasting model based on three machine-learning algorithms. Weber et al. [8] described a copy-paste imputation method in which gaps in smart meter data were eliminated by copying and scaling values from days with the lowest dissimilarity according to difference distance metrics. To conclude, related work has examined possible solutions for handling missing data in time series, particularly focusing on electric load data, involving both new and existing imputation methods.

1.2 Contributions

This paper investigates the effect of the proportion and distribution of missing values and the chosen imputation method on the usability of the dataset for further analysis. Therefore, the study introduces a simple method for artificially creating, imputing, and evaluating gaps in predefined scenarios.

Through the systematic comparison of the quality of use cases built on four real-world electric load time series datasets imputed with six easily implementable imputation methods this study provides valuable guidance for selecting appropriate imputation methods, considering both the missingness scenario and the purpose of the analysis.

The rest of this paper is structured as follows. The gap creation, imputation, and evaluation methodology are introduced briefly in Section 2 and supplemented with concrete dataset description and implementation details in Section 3. Section 4 summarizes the resulting findings, followed by a substantive discussion in chapter 5. The paper ends with a short summary in Section 6.

2. METHODOLOGY

To systematically investigate the imputation methods' performance and the impact of different types of gaps, the following three-step approach, summarized in Figure 1, is presented: First, gaps are systematically generated on a complete dataset, which are then repaired by the models. The resulting time series are evaluated in three scenarios to assess performance.

2.1 Gap Generation

Gaps with varying frequency and in different sizes were introduced randomly into the complete datasets.

2.2 Gap Filling

To address the gaps created in the previous steps, six distinct imputation methods were employed in this step.

2.3 Evaluation

This module evaluates the methods by comparing the imputed time series to the original, gap-free version. Initially, the proximity between the imputed load curves to the actual values was evaluated solely on the created gaps. This comparison focused on how well the imputed values capture the actual values. Additionally, two practical scenarios were examined: first, the comparison of load forecasting performance, and second, the contrast in the total energy consumption calculated using imputed versus the actual data.

3. USE CASES

3.1 Used Datasets

Four electricity load datasets recorded at companies from the manufacturing sector are used for evaluation, which will be enumerated 1-4 from now. The profiles are recorded from a smart meter of air pressure generators. The temporal resolution was 1 minute.

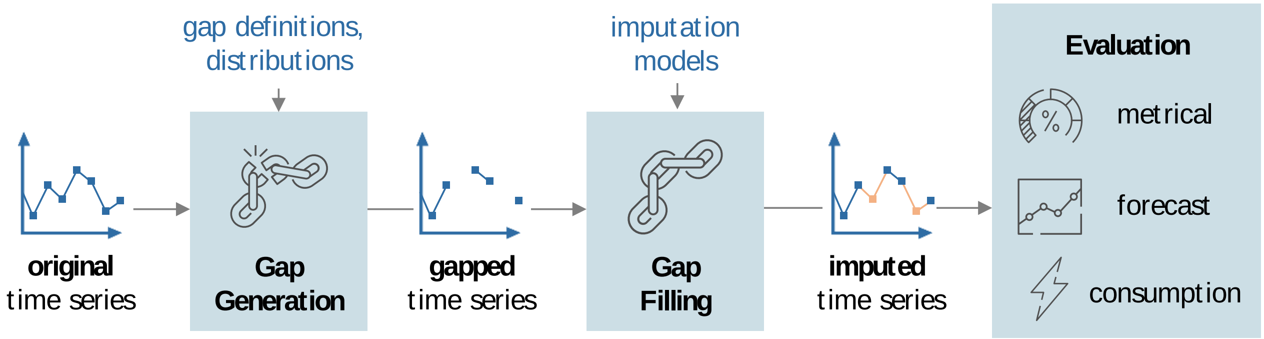

Fig. 1. The overall methodology

The datasets exhibit temporal characteristics. For instance, the energy consumption scaled between 0 and 1 in dataset 4 is displayed in Figure 2. Notably, energy consumption on weekends and during nighttime shows a substantial reduction compared to workdays, indicating reduced operations or downtime during these periods. Between January and March, the decrease in energy consumption at night is less prominent, while during other periods, it nearly reaches zero, suggesting variations in operational patterns through time.

3.2 Implementation

Our primary goal in this study was to assess the impact of the imputation algorithm on several aspects: the proximity of imputed values to the actual values and the performance of subsequent scenarios on the imputed data. To compare the results with complete data, first, we selected four datasets with a possibly long interval without or with minimal proportion of missing values. For all four selected datasets we defined an identical, 206 days long time interval, where the proportion of missing values were below 0.03%. These missing values were imputed with the values from a week before.

For gap creation, we introduced gaps into complete datasets through nine different missingness scenarios. These scenarios comprise datasets with varying percentages of missing values (1%, 5%, and 10%) and gaps of different sizes: small (1-10 minutes), middle (10-60 minutes), large (60 minutes to 1 day) and mixed (1 minute to 1 day). The length of the gaps was assigned randomly within the specified ranges, and the gaps were randomly distributed across the datasets. For more reliable results, the entire process was repeated 5 times, permuting the gaps differently for each.

For gap filling, the gaps created in the previous step were imputed using six different imputation methods. One of these methods fills the gaps with constant values, namely with the last value (Padded Last) before the gap. Additionally, we used linear interpolation, Kalman smoothing, and imputation through moving averages with the imputeTS R package [9]. Furthermore, two methods for imputing values with similar characteristics based on calendar features, such as weekday and business holidays, were utilized. Once, we imputed measurements from the corresponding time window in the previous week, capturing reoccurring patterns present in the data (Last Week). In the other method (KNN), we performed imputation based on values measured in the k nearest neighbors for each hour with missing values. We searched for the three hours from the past data with the smallest Euclidean distance [10] calculated on the added calendar features and the total energy consumption in the previous six hours. The missing values were then imputed as the average in the three nearest hours.

For the metrical evaluation, i.e., to assess the proximity of the imputed values to the actual values, the mean absolute error (MAE) was calculated for the imputed gaps.

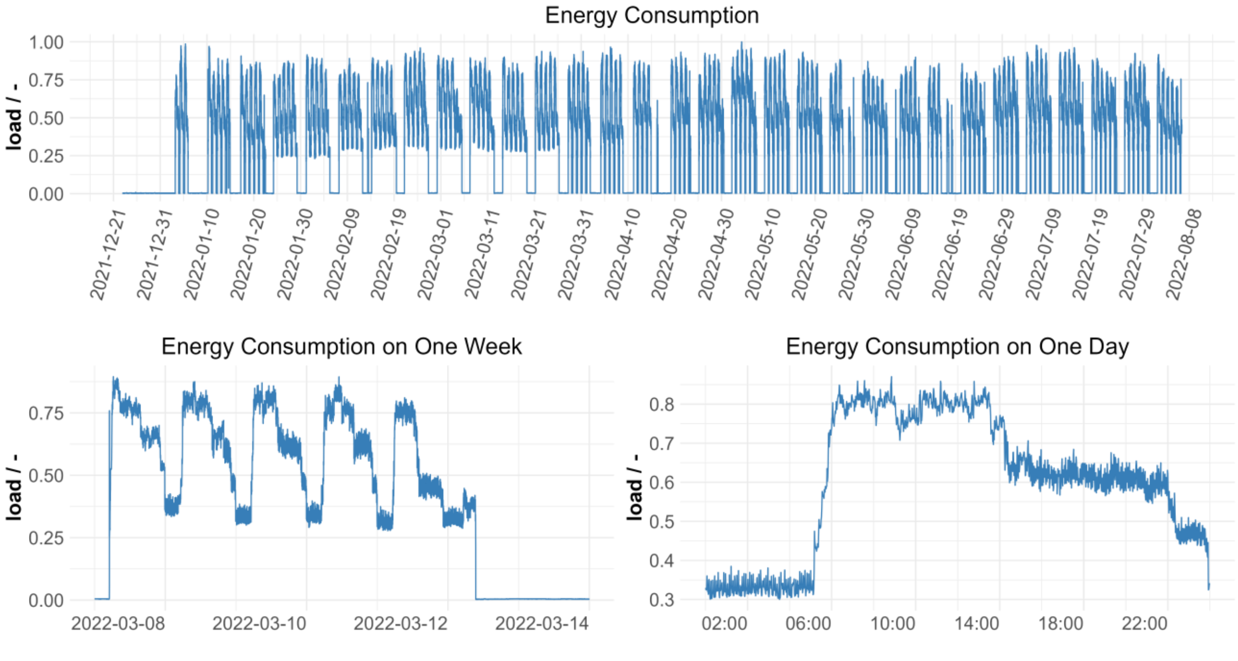

Fig. 2. Energy consumption overall, daily and weekly. The load was normalized to [0,1]

For the forecast evaluation, the following additional features were added: periodic calendar and time-series information (month, day, day of week, relative day minute, week of year, quarter), holiday information (holidays, bridging days, weekend flags), and lagged values (one hour, one day, one week). After train/test splitting and normalization, four one-hour-ahead (60 steps) models (namely Linear and Lasso Regression, Random Forest, and Decision Tree Regressor) were trained in a k-fold (k=5) cross-validation setting. The prediction quality was compared through the MAPE metric, which is well-interpretable and commonly used.

Also, we measured the proportion of overshoots in error and the proportion of total energy consumption to the actual energy consumption (consumption evaluation). These two additional metrics were employed to determine how well the imputed data aligns with the actual energy consumption patterns.

For each dataset, the calculations for the three evaluations were performed with the described four missingness proportions and three gap sizes for the five gap permutations, resulting in 240 cases, each of which tested all six imputation models. To quantify uncertainty, for all cases and each evaluation metric, the 95% confidence interval for the mean was calculated.

4. EXPERIMENTAL RESULTS

In this section, the empirically determined results of the three evaluation modules are addressed.

The results of the metrical evaluation, the MAE comparison of the imputation methods applied to the different missingness scenarios are presented in Figure 3. Among the applied methods, Kalman smoothing consistently yielded the smallest MAE between real and imputed values in almost all cases. For Dataset 4, imputation with values from the corresponding time window of the previous week and KNN on large and mixed-sized gaps performed better, indicating stronger temporal patterns in the dataset. Across all four datasets, large and mixed-sized gaps resulted in a higher MAE with a larger variance between the permutations of the gaps, implying greater challenges in the imputation in these missingness scenarios.

The calculated confidence intervals for the mean of the MAE revealed that the performance varies across the datasets. Notably, all imputation methods performed better on dataset 2 and 4 and worse on dataset 1 and 3. The latter two not only exhibit higher MAE means on the 240 cases, but also wider confidence intervals.

To assess the forecast evaluation, the prediction errors were plotted based on the imputation errors of the datasets on which they were trained. Also, the correlations were determined. Figure 4 shows the results for dataset 2 as an example. The plots of the other datasets are qualitatively similar and, therefore not included.

Based on the evaluations of the four datasets, the following can be stated: With increasing gap size, the

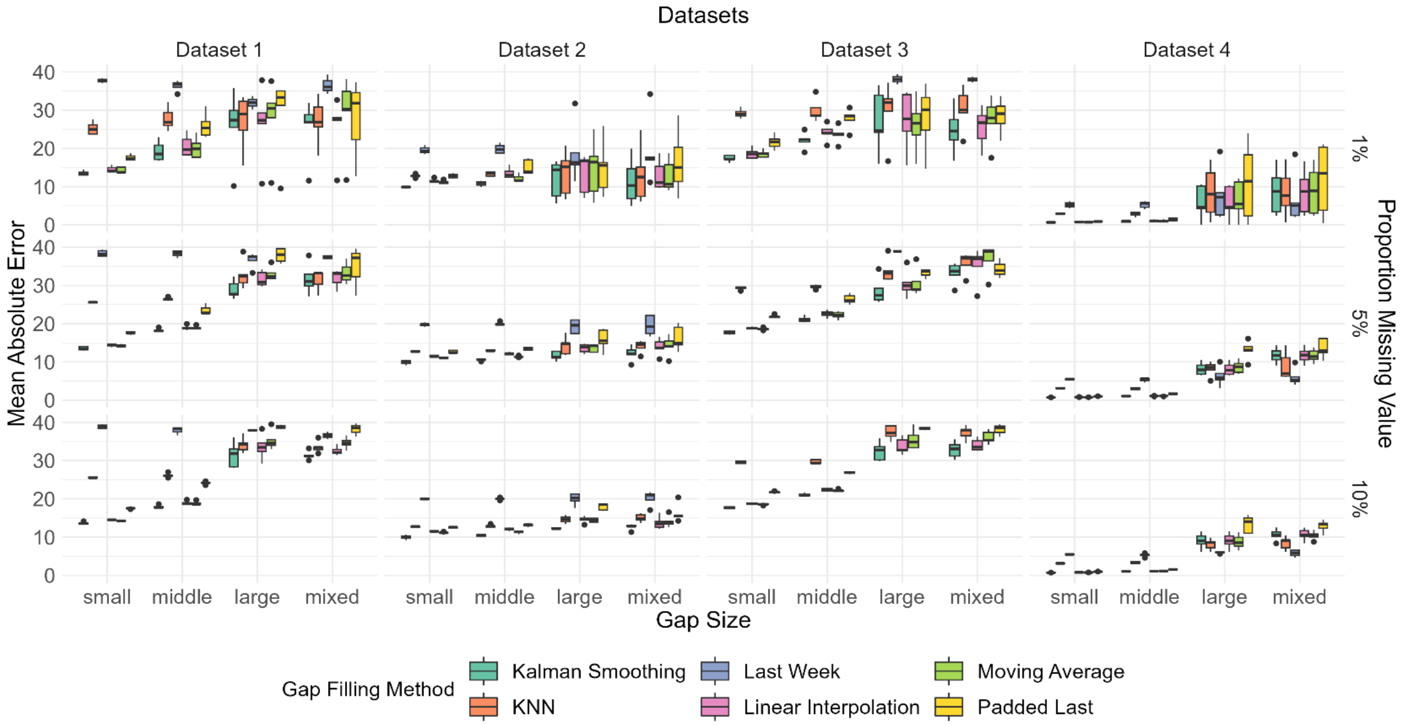

Fig. 3. Results of metric evaluation

variance of the prediction model performance increases. For small and medium gaps, the prediction error of the regression models correlates with the imputation error. In all other cases, no significant correlation can be detected. The results of mixed gaps are quantitatively dominated by the large gaps among them. Their distributions are very similar. Overall, the choice of model has a significantly higher impact on the prediction quality than the choice of imputation. For example, prediction with Lasso Regression on dataset 2 in most cases yielded to a MAPE lower than with other prediction models with a narrower confidence interval.

For energy consumption evaluation, we considered two aspects to answer the question of how imputation affects the total energy consumption, i.e., the integral: the proportion of over- and undershoots of the imputed values on the one hand, and the relative difference between the areas as a whole on the other.

The evaluation of the proportion of overshoots in MAE revealed that 77% imputation models achieved a proportion between 0.4 and 0.6, indicating that approximately half of the difference between the imputed and real values was positive, and the other half was negative. The total energy consumption calculated on the overall range of the imputed data compared to the actual values varied between 0.995 and 1.001, suggesting a close alignment with the real values. Accordingly, the mean of the proportion of total energy consumption was consistently near to 1 with a very narrow corresponding confidence interval.

5. DISCUSSION

In this study, we focused exclusively on the case where missing data occurred completely at random [11]. However, incorporating the missingness mechanism in the imputation design could greatly effect the analysis and an important scope for future work. The number of datasets and missingness scenarios used in this study was limited. Also, there are numerous ways to improve load prediction: Besides integrating additional external time series, the extension to additional model classes would be conceivable in future work.

6. CONCLUSIONS

Filling gaps in time series data is a significant step in data preprocessing. The question of which models are helpful for which types of gaps and the impact of imputation on subsequent tasks was investigated.

For this purpose, a method was introduced to systematically create gaps, impute them with different models and test them in three realistic evaluation settings. To this end, the impact of imputation on curve

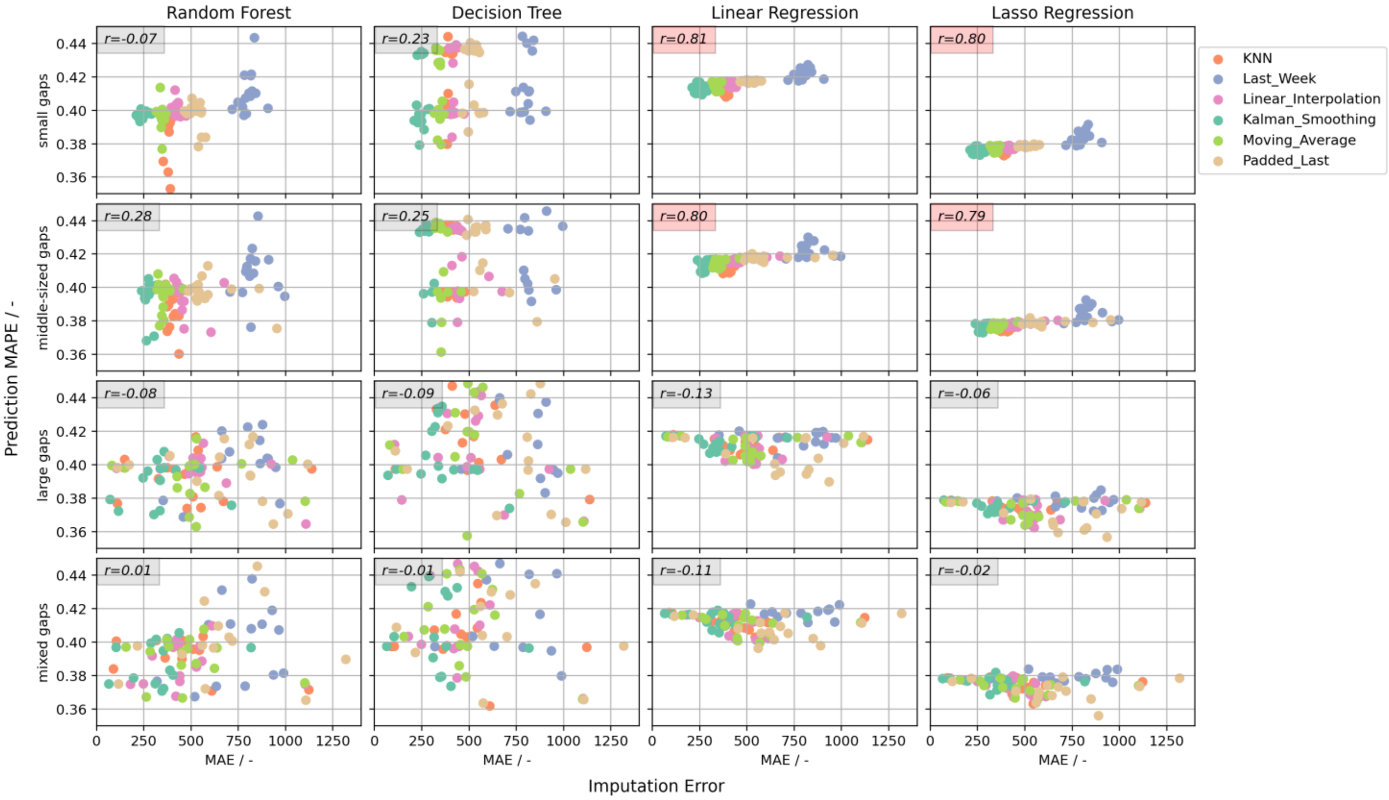

Fig. 4. Results of forecast evaluation

similarity (metric), the load forecasting task, and determining total energy consumption was investigated. The method was applied to four datasets from the industry. Based on the evaluations, the following conclusions emerge:

To minimize the mean imputation error, different models for different gap sizes can be recommended. However, the concrete choice is company load-specific, but overall, the MAE between the real and imputed data was often the lowest with Kalman smoothing. The larger the gaps, the more difficult the imputation. For load forecasting, the choice of the forecasting model has a significantly higher impact on the performance than the choice of the imputation method. Here, the imputation performance correlates with prediction performance for small (1-10 minutes) and medium gaps (10- 60 minutes), but not in all other cases. In the use case of energy consumption analysis, very good results can be obtained since over- and undershoots occur in a balanced ratio.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the financial support of the Invest BW project KIDruId by the Ministry of Economic Affairs, Labour and Tourism of Baden-Württemberg and the project Partnership-based development to detect compressed air leaks in machines and systems by SICK AG, as well as WRS Energie + Druckluft GmbH for providing the data, and the project's supervision by the project management organization VDI/VDE/IT.

DECLARATION OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. All authors read and approved the final manuscript.

REFERENCE

[1] Sensfuß, Frank; et al. (2021) "Langfristszenarien für die Transformation des Energiesystems in Deutschland 3." Potentiale der Windenergie auf See Datensatz 127.

[2] Peppanen, J; Reno, MJ.; Thakkar, M; Grijalva, S; Harley, RG. (2015): Leveraging AMI Data for Distribution System Model Calibration and Situational Awareness. In: IEEE Trans. Smart Grid 6 (4), S. 2050–2059. DOI: 10.1109/TSG.2014.2385636.

[3] Sridevi, S.; Rajaram, S.; Parthiban, C.; SibiArasan, S.; Swadhikar, C. (2011): Imputation for the analysis of missing values and prediction of time series data. In: 2011 International Conference on Recent Trends in Information Technology (ICRTIT). 2011 International Conference on Recent Trends in Information Technology (ICRTIT). Chennai, India, 03.06.2011 - 05.06.2011: IEEE, S. 1158–1163.

[4] Kanda, I; Veguillas, JQM. (2019): Data preprocessing and quantile regression for probabilistic load forecasting in the GEFCom2017 final match. In: International Journal of Forecasting 35 (4), S. 1460–1468. DOI: 10.1016/j.ijforecast.2019.02.005.

[5] Lee, B; Lee, H; Ahn, H. (2020): Improving Load Forecasting of Electric Vehicle Charging Stations Through Missing Data Imputation. In: Energies 13 (18), S. 4893. DOI: 10.3390/en13184893.

[6] Ryu, S; Kim, M; Kim, H. (2020): Denoising Autoencoder-Based Missing Value Imputation for Smart Meters. In: IEEE Access 8, S. 40656–40666. DOI: 10.1109/ACCESS.2020.2976500.

[7] Waqas Khan, P; Byun, YC; Lee, SJn; Park, N. (2020): Machine Learning Based Hybrid System for Imputation and Efficient Energy Demand Forecasting. In: Energies 13 (11), S. 2681. DOI: 10.3390/en13112681.

[8] Weber, M; Turowski, M; Cakmak, HK.; Mikut, R; Kuhnapfel, U; Hagenmeyer, V. (2021): Data-Driven Copy-Paste Imputation for Energy Time Series. In: IEEE Trans. Smart Grid 12 (6), S. 5409–5419. DOI: 10.1109/TSG.2021.3101831.

[9] 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 <https://doi.org/10.32614/RJ-2017-009>

[10] Dokmanic, I; Parhizkar, R; Ranieri, J; Vetterli, M. (2015): Euclidean Distance Matrices: Essential Theory, Algorithms and Applications. In: IEEE Signal Process. Mag. 32 (6), S. 12–30. DOI: 10.1109/MSP.2015.2398954.

[11] Baraldi, AN.; Enders, CK. (2010): An introduction to modern missing data analyses. In: Journal of school psychology 48 (1), S. 5–37. DOI: 10.1016/j.jsp.2009.10.001.