Evaluating the Impact of Data Imputation on Knowledge Discovery in Smart Grid Data

Leonardo Minelli*
leonardo.minelli@sou.unijui.edu.br
UNIJUÍ
Ijuí, RS, Brazil

Paulo Sérgio Sausen†
sausen@univali.br
Univali
Ijuí, RS, Brazil

Airam Sausen‡
airam@unijui.edu.br
UNIJUÍ
Ijuí, RS, Brazil

Rene Reinaldo Emmel Junior‡
rene.junior@sou.unijui.edu.br
UNIJUÍ
Ijuí, RS, Brazil

Abstract

Missing data are common in smart grid environments, especially in underground substations where operational constraints lead to incomplete time-series measurements. Because the Knowledge Discovery in Databases (KDD) process depends on consistent datasets, imputation is essential to preserve analytical reliability. This study evaluates how data reconstruction affects a previously validated hybrid KDD framework applied to real substation measurements. Missingness levels of 5%, 10%, 20%, and 30% were simulated under a Missing Completely at Random (MCAR) mechanism, and the Modified Akima Interpolation Method (MAKIMA) was used to restore the affected series. The reconstructed datasets were then processed through EM clustering and Apriori association rule mining and compared with the original data. Error metrics (MAE, RMSE, R^2) showed high reconstruction accuracy, with R^2 above 0.999. Clustering deviations remained below 0.5%, and association rules retained their structure with minimal changes. The findings indicate that the KDD framework remains stable with up to 30% MCAR missingness, demonstrating that MAKIMA-based imputation preserves both statistical properties and the consistency of the extracted knowledge in smart grid datasets.

CCS Concepts: • **Information systems → Data mining**.

Keywords: Smart grids, Data imputation, Knowledge discovery

ACM Reference Format:
Leonardo Minelli, Paulo Sérgio Sausen, Airam Sausen, and Rene Reinaldo Emmel Junior. 2018. Evaluating the Impact of Data Imputation on Knowledge Discovery in Smart Grid Data. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

In addition to their accelerated generation, data have become critical assets in modern systems, supporting applications ranging from natural language processing and healthcare to autonomous vehicles and industrial automation. Artificial Intelligence (AI) and data-driven analytics increasingly influence not only decision-making but also the efficiency and resilience of complex infrastructures. However, even the most sophisticated algorithms and software systems can fail when the data supporting them are incomplete, inconsistent, or poorly prepared [20].

In smart grids, which rely on massive streams of operational and environmental data for real-time control and predictive maintenance, the integrity and completeness of datasets are essential. The absence of information caused by hardware malfunctions, sensor outages, network interruptions, or maintenance procedures can severely impair analytical processes and hinder automated decision-making [3]. This challenge is even more critical in underground power substations, where environmental constraints, limited accessibility, and complex interconnections increase both the probability and the impact of data loss.

Data imputation is a key preprocessing step for addressing missing values in databases. It aims to reconstruct absent or

*Writing - original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis and Conceptualization
†Writing - review and editing, Supervision, Methodology and Conceptualization
‡Writing - review and editing, Supervision

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

incomplete information while preserving statistical properties and temporal coherence, ensuring that subsequent analyses, including clustering and association rule mining, remain reliable. Improper handling of missing data can distort model training, bias parameter estimation, and compromise the interpretation of discovered knowledge [7, 16].

The imputation process presents several challenges: accurately identifying the missing-data mechanism (MCAR, MAR, MNAR), selecting suitable algorithms, and validating the quality of the reconstructed data. Temporal datasets from smart grids exhibit non-linear behaviors and abrupt variations, requiring techniques capable of preserving local patterns and smooth transitions. The choice of imputation technique depends on the type of data, the distribution of missing values, and the specific context of the problem. Moreover, assessing the quality of the imputed data is essential for avoiding analytical distortions [11]. Traditional linear interpolation may fail to capture these characteristics, motivating the use of advanced methods such as the Modified Akima Interpolation Method (MAKIMA), which provides smoother and more realistic estimations [17].

From a knowledge discovery perspective, imputation is not merely a preprocessing task but a determining factor in the accuracy and stability of the entire KDD process. Errors or distortions introduced during imputation can propagate through the clustering and association stages, ultimately altering the extracted insights. Despite its importance, few studies have quantitatively assessed how imputation affects the consistency of knowledge extracted from smart grid data, especially in underground substations.

Since the KDD process depends on complete and consistent data to ensure valid knowledge extraction, the presence of missing values can compromise analytical integrity and the reliability of discovered patterns [5]. In this context, the present work evaluates the impact of data imputation on the KDD process applied to underground smart grid substations, focusing on how the MAKIMA method can restore missing values while preserving the integrity of the derived knowledge.

The main contributions of this study can be summarized as follows: (i) the simulation of controlled missing data using a MCAR mechanism at absence levels ranging from 5% to 30%, enabling a systematic evaluation of the imputation impact; (ii) the application of the MAKIMA to reconstruct missing operational measurements while preserving temporal coherence and local behavioral patterns; (iii) the reapplication of the previously developed KDD framework—including EM clustering and Apriori association rule mining—to both the original and the imputed datasets for comparative analysis; and (iv) the quantitative assessment of the robustness and consistency of the KDD process after imputation, using metrics that evaluate data accuracy and the stability of the knowledge extracted.

Through this approach, the study improves the understanding of how imputation strategies affect the reliability of knowledge discovery in smart grids, providing insights for more resilient data-driven systems and predictive models in critical urban infrastructure.

This paper is organized as follows. Section 2 reviews related work on data imputation and the KDD process in smart grids. Section 3 describes the methodology used for data selection, preprocessing, imputation, and knowledge discovery. Section 4 presents the experimental results and discusses their implications for data and knowledge integrity. Finally, Section 5 summarizes the main conclusions and outlines directions for future research.

2 Related works

This section presents an overview of the most relevant studies in the literature, establishing the technical context and foundation for the methodology adopted in this work.

2.1 The KDD process

The KDD process provides a systematic pathway from raw data to actionable knowledge, comprising the stages of selection, preprocessing, transformation, data mining, and evaluation. It has become a cornerstone for extracting meaningful patterns from large and heterogeneous datasets [5]. In the context of smart grids, the KDD process enables the identification of operational patterns, fault conditions, and opportunities for predictive maintenance through the integration of data mining and machine learning algorithms. Several studies have demonstrated its applicability in power systems, highlighting its capacity to support decision-making and enhance the reliability of critical infrastructure [8, 10].

2.2 Missing data and imputation techniques in smart grid

Missing data represent a recurrent challenge in smart grid monitoring systems, commonly arising from sensor malfunctions, communication failures, or temporary shutdowns during maintenance activities. Such gaps can distort statistical analyses, compromise anomaly detection, and impair the predictive accuracy of models used for operational planning. Therefore, properly addressing missing data is essential for ensuring the consistency and interpretability of energy analytics [3, 18]. Several studies have shown that the effects of missingness extend beyond data completeness, influencing the entire analytical workflow, particularly when integrated with processes such as KDD and machine learning [20].

Several imputation strategies have been investigated for smart grid time-series, including statistical methods, machine learning approaches, and hybrid interpolation techniques [2, 9]. Recent studies highlight the effectiveness of piecewise cubic Hermite-based approaches, such as the MAKIMA,

which preserves smoothness and local behavioral patterns in complex temporal datasets [17].

3 Methods and materials

This research is an original and applied study aimed at generating new knowledge to enhance both the understanding and practical application of data analytics in smart grid environments. Applied research seeks to produce solutions for real-world problems by integrating theoretical frameworks with practical experimentation [4, 6]. The study combines exploratory and explanatory characteristics. It is exploratory because it investigates the behavior of data imputation within the KDD process, a topic that remains underexplored in the context of underground smart grids. It is explanatory because it analyzes causal relationships between the presence of missing data, the imputation process, and the resulting quality of the knowledge discovered. The research adopts experimental technical procedures [19], as it involves the controlled manipulation of datasets and the evaluation of outcomes using quantitative performance measures.

3.1 Data collection

The dataset used in this study consists of real operational measurements obtained from underground power substations operated by CEEE Equatorial in Porto Alegre, Brazil. These substations belong to the RW-4 group, specifically units 11 and 35, which represent distinct operational conditions within the same underground distribution network. The organizational structure of the monitored substations is illustrated in Fig. 1.

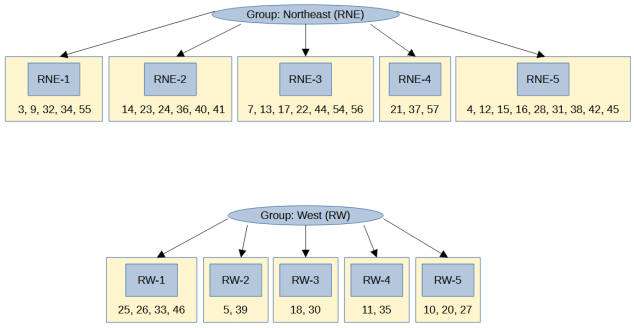

Figure 1. Organization of groups, subgroups, and substations.

The dataset includes electrical and environmental parameters such as primary and secondary voltage, current, ambient temperature, transformer temperature, and humidity, all recorded at regular time intervals by intelligent monitoring devices installed in the substations. These data were acquired through the utility's supervisory system and stored in tabular format for subsequent analytical processing.

Table 1. Monthly data volume and quality indicators for the RW-4 subgroup (2018–2019).

Period	Data volume	Cov. (%)	LDA	Seg.	Anom.
01/2018	405,301	78%	○	○	○
02/2018	237,570	46%	●	○	○
03/2018	22,195	4%	●	●	-
04/2018	143,366	28%	●	●	-
05/2018	415,751	80%	○	○	○
06/2018	387,481	75%	○	○	●
07/2018	160,763	31%	●	●	-
08/2018	175,930	34%	●	●	-
09/2018	270,304	52%	○	●	-
10/2018	127,512	25%	●	○	○
11/2018	235,871	45%	●	○	○
12/2018	66,537	13%	●	○	○
01/2019	100,062	19%	●	○	●
02/2019	90,205	17%	●	●	-
03/2019	166,774	32%	●	○	●
04/2019	85,146	16%	●	●	-
05/2019	41,997	8%	●	○	●
06/2019	139,088	27%	●	○	●
07/2019	148,194	29%	●	●	-
08/2019	129,409	25%	●	●	-
09/2019	258,773	50%	○	○	●
10/2019	311,082	60%	○	●	-
11/2019	399,736	77%	○	○	●
12/2019	416,777	80%	○	○	○

Note. ● Yes ; ○ No ; - Inconclusive.

Table 1 presents the data volume for each period, along with quality indicators that support the selection of the analyzed months. The table also evaluates whether the data exhibit significant segmentation, defined as the presence of large temporal gaps that may hinder or complicate pattern recognition and anomaly detection. Following visual inspection of data distribution, identification of recurring patterns, and detection of potential anomalies, the most consistent datasets were selected for further analysis.

In Table 1, symbols are used to simplify interpretation: the circle (○) indicates that the property is absent, the bullet (●) indicates that the property is present, and a blank entry (-) is used when the dataset exhibits an unfavorable profile for visual analysis, typically due to a combination of low data availability and segmentation that renders it inconclusive for anomaly detection.

In this context, months with a data volume below 50% of the expected 259,200 records (assuming a 30-day average collection period) are classified as having low data availability (LDA) due to insufficient coverage (Cov. (%)). Additionally, datasets are classified as segmented (Seg.) when they contain more than three interruptions within a month, potentially indicating irregularities in data collection that may affect the effectiveness of the mining process. The anomalous column

(Anom.) refers to months in which visual inspection revealed irregular patterns in the monitored parameters, including sudden and inconsistent variations, atypical behaviors deviating from expected operational conditions, and values that diverge from historical trends.

Given that the proposed analysis requires complete, consistent, and temporally continuous data to ensure methodological reliability, the months of January 2018, May 2018, and December 2019 were selected. These datasets present the highest coverage rates and minimal segmentation among all available periods, providing a robust foundation for the subsequent stages of data preprocessing, imputation, and knowledge discovery. These selected months were then used as input for the preprocessing and simulation steps described in the following sections.

3.2 Data preprocessing

The preprocessing stage aimed to ensure that the dataset was properly structured, consistent, and suitable for the subsequent experimental procedures. This phase involved organizing, cleaning, and preparing the data for the simulation of missing values and their later imputation. The datasets were imported and processed using the R software environment, which provides flexibility for handling large time-series datasets and performing controlled manipulations such as random record removal and numerical interpolation.

During preprocessing, non-numerical attributes (e.g., identification codes and timestamps) were retained for traceability but excluded from analytical computations. Outliers and duplicated records were removed to avoid distorting the statistical relationships among the variables. The resulting dataset maintained temporal continuity and uniform sampling intervals, preserving the operational characteristics in smart grid data [14] required for consistent and reliable knowledge discovery analysis.

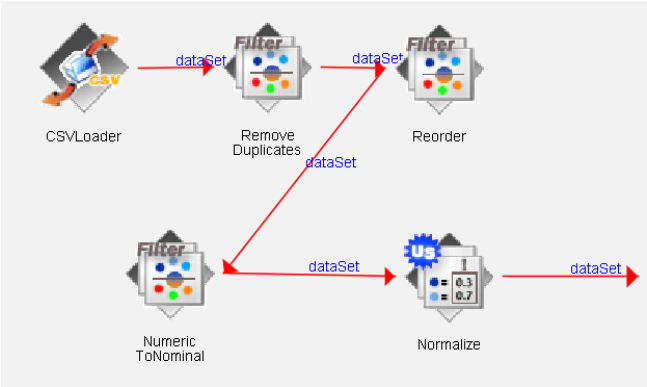

Figure 2. Preprocessing workflow.

The preprocessing workflow comprises a sequence of structured operations designed to prepare the datasets for analysis. Initially, irrelevant attributes were removed to retain only the variables essential for the mining tasks. Missing values were then introduced under a controlled MCAR mechanism [15], followed by their reconstruction using the MAKIMA interpolation method [1, 12]. Subsequently, as shown in Fig. 2, data in CSV format is ingested via the *CSV Loader*. Furthermore, the *Remove Duplicates* operator eliminates redundant records to ensure balanced weighting during mining. Attributes are then reorganized via the *Reorder* block to ease visual analysis, while the *NumericToNominal* operator ensures that numerical attributes are converted to nominal types as required by specific techniques. Finally, all numerical attributes were normalized using the *Normalize* block to ensure consistent scaling. The resulting datasets were converted into ARFF format for integration with the data mining framework. This preprocessing workflow ensured data consistency and full compatibility with the KDD framework adopted for the subsequent analyses.

3.2.1 Generating missing data. To evaluate the performance of the imputation techniques, missing values were artificially introduced into the original datasets under a MCAR mechanism. This approach ensures that the probability of data removal is independent of both observed and unobserved values, thereby isolating the imputation performance from potential biases in data generation [11].

In this study, the missing data mechanism was applied to the primary current attribute, which plays a central role in the operational characterization of underground substations. As demonstrated in previous analyses of the RW-4 subgroup, variations in primary current were decisive for identifying clusters, classifying operating states, and generating association rules.

Controlled levels of missingness were simulated at 5%, 10%, 20%, and 30%, representing typical and extreme conditions encountered in real-world smart grid environments. The removal process was applied independently to each variable, preserving the multivariate structure and temporal alignment among attributes. This strategy enabled a fair comparison between the imputed datasets and the original complete data during the validation phase.

Table 2. Data volume removed and imputed for each month.

Period	Data volume	5%	10%	20%	30%
01/2018	405,301	20,265	40,530	81,060	121,590
05/2018	415,751	20,788	41,575	83,150	124,725
12/2019	416,777	20,839	41,678	83,356	125,034

Table 2 presents the resulting data volumes after the introduction of the MCAR mechanism at different missingness levels. These values represent the number of records to be imputed in each dataset following the random removal of samples.

In the resulting datasets, the removed records were replaced with the question mark symbol (?), which explicitly denotes missing values in the data files. This placeholder ensures compatibility with subsequent imputation procedures and facilitates the identification of missing entries during the preprocessing and validation stages.

3.2.2 Data imputation. After generating the incomplete datasets, the missing values were reconstructed using numerical interpolation methods implemented in R. Among the various techniques available for time-series imputation, this study focuses on the MAKIMA interpolation method, selected for its ability to preserve local trends and prevent overshooting in irregularly spaced data [1, 13], as well as for its proven performance in smart grid applications [17].

The imputation process was independently applied to each monthly dataset, which contained the combined measurements from both underground substations considered in this study. This approach ensured that the temporal and operational patterns intrinsic to each period were preserved during reconstruction, avoiding cross-month interference and maintaining consistency with the original data collection cycles. For each month and missingness level (5%, 10%, 20%, and 30%), the MAKIMA method was applied to estimate the missing values in the primary current of phase B, generating a separate imputed dataset for each case.

This imputation strategy also supports scalability for multistation analysis, as each file represents a self-contained operational snapshot. Moreover, by keeping the substation data integrated within each monthly file, the imputation captures the shared load behavior and mutual influence between substations, which is relevant for detecting correlated anomalies in subsequent analyses.

3.3 Data mining

The data mining process was executed on both the complete and imputed datasets to evaluate whether the imputation preserved the knowledge patterns identified by the KDD framework. The analysis was conducted using a computational workflow composed of sequential and interdependent stages designed to ensure methodological consistency, as illustrated in Fig. 3.

The framework begins with data loading and validation, followed by transformation and preparation steps that standardize and organize the information for processing. In the subsequent stage, data mining algorithms are applied to extract patterns, correlations, and structural relationships within the operational data. The final phase involves the evaluation and representation of the discovered knowledge, allowing a direct comparison between the results obtained from the original and imputed datasets. This design enables the assessment of whether the KDD process maintains equivalent analytical performance and interpretative validity when applied to reconstructed data.

For the clustering stage, the analysis focused on the primary current of phase B, which was the attribute most affected by the imputation process. The substation identifier was retained to distinguish operational contexts between sites. Other attributes (e.g., voltage and temperature) remained available in the dataset and were used only for visualization and interpretive purposes. This controlled attribute selection enabled a focused evaluation of how the imputation influenced the data structure and the resulting knowledge patterns.

In the association rule mining stage, the Apriori algorithm was applied to both the original and the imputed datasets to evaluate whether the imputation process preserved the underlying relationships among operational variables. The rules were extracted using identical parameters and support/confidence thresholds, enabling a direct comparison of the discovered patterns.

4 Results

The following results summarize the comparative outcomes obtained from applying the KDD process to both the complete and the imputed datasets. To assess the impact of data imputation on the KDD workflow, the results derived from the imputed datasets were directly compared with those obtained from the original complete data. The analysis focused on the preservation of clustering structures and association rules. Clustering performance was quantitatively evaluated using the Silhouette Coefficient, which measures intra-cluster cohesion and inter-cluster separation. The association rules were analyzed in terms of their overlap ratio and the stability of support and confidence values. This procedure enabled a concise yet comprehensive validation of the effectiveness of the MAKIMA-based imputation within the KDD framework.

4.1 Imputation results

To quantitatively assess the accuracy of the MAKIMA interpolation, the imputed datasets were compared with the original measurements using three complementary metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). These metrics respectively evaluate the average deviation, the penalization of larger errors, and the linear agreement between the reconstructed and original values. Table 3 summarizes the results for all months and missingness levels considered in this study.

The quantitative results in Table 3 confirm the high accuracy and stability of the MAKIMA interpolation method. Across all months and missingness levels, the MAE remained below 0.018, while the RMSE values were consistently below 0.053. The coefficient of determination exceeded 0.999 in every case, indicating an almost perfect linear agreement between the imputed and original datasets. These findings

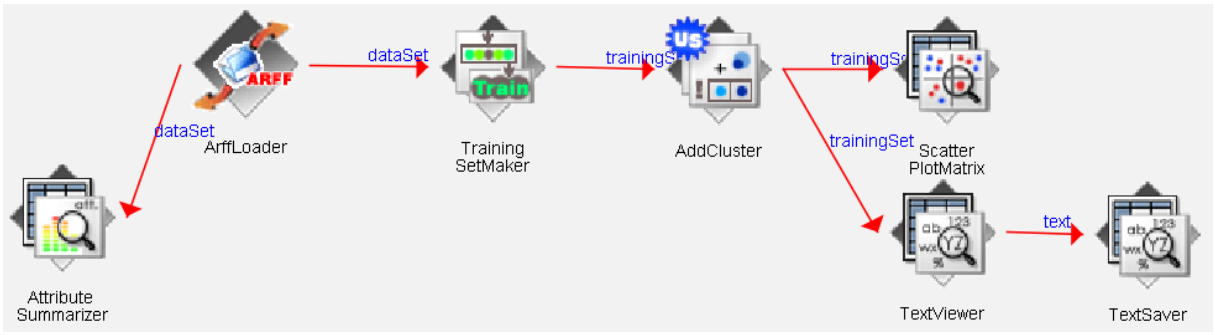

Figure 3. Data mining workflow.

Table 3. Error metrics for MAKIMA-imputed datasets.

Period	MCAR	MAE	RMSE	R^2
01/2018	5	0.002514	0.019369	0.999933
	10	0.005079	0.027344	0.999866
	20	0.010382	0.039566	0.999719
	30	0.015756	0.048780	0.999573
05/2018	5	0.002822	0.020462	0.999909
	10	0.005679	0.029115	0.999817
	20	0.011651	0.042156	0.999616
	30	0.017885	0.052768	0.999398
12/2019	5	0.002435	0.018407	0.999952
	10	0.004921	0.026216	0.999903
	20	0.010029	0.037682	0.999800
	30	0.015308	0.047174	0.999687

demonstrate that the MAKIMA method can accurately reconstruct missing primary current measurements, even when up to 30% of the data are absent, without compromising the numerical integrity of the signals used in subsequent data mining and knowledge discovery stages.

4.2 Clustering results

For December 2019, Fig. 4 illustrates the clustering results obtained from the original dataset and Fig. 5 from the dataset imputed with 30% MCAR missingness. In both cases, the clusters were formed based on the primary current of phase B over time, with colors representing distinct operational groups. The visual patterns exhibit a high degree of structural similarity between the two datasets, confirming that the imputation process did not distort the temporal dynamics or the relative boundaries between clusters.

The imputed data exhibit smooth transitions in regions originally affected by missing values, while the overall cluster configuration remains nearly identical to that of the complete dataset. Both figures display two predominant operational regimes, characterized by distinct current amplitude ranges that are preserved after interpolation. This visual consistency reinforces the quantitative findings summarized in Table 4, demonstrating that the MAKIMA interpolation successfully

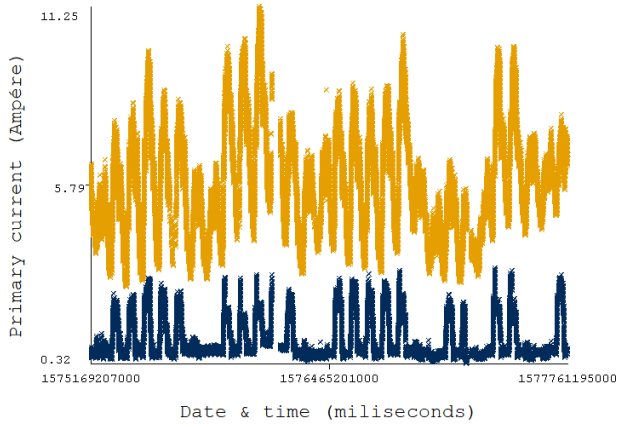

Figure 4. December 2019 clustering - original data.

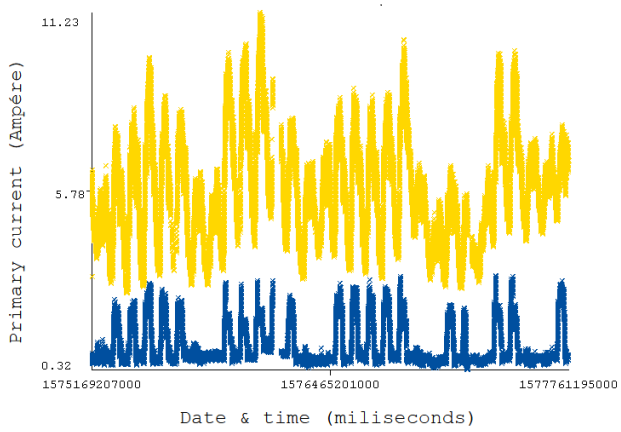

Figure 5. December 2019 clustering - 30% MCAR and MAKIMA.

reconstructs missing values without compromising the underlying data structure or the coherence of the knowledge recovered during the clustering stage.

Table 4 summarizes the average silhouette coefficients obtained for both the original and imputed datasets. Across all evaluated months and missingness levels, the differences

Table 4. Silhouette coefficient of clustering for original and imputed data.

Period	S.C. original	MCAR	S.C. imputed	Δ (%)
01/2018	0.7008	5	0.7012	+ 0.06
		10	0.6994	- 0.20
		20	0.7020	+ 0.17
		30	0.6992	- 0.23
05/2018	0.8515	5	0.8506	- 0.11
		10	0.8506	- 0.11
		20	0.8509	- 0.07
		30	0.8515	0.00
12/2019	0.7241	5	0.7249	+ 0.11
		10	0.7240	- 0.01
		20	0.7265	+ 0.33
		30	0.7271	+ 0.41

Note. S.C. = Silhouette coefficient.

remained below 0.5%, indicating a high degree of clustering stability after the MAKIMA interpolation. These results suggest that the imputation preserved the intrinsic structure of the data, maintaining the coherence and separability of operational patterns within the substations. In practical terms, the reconstructed datasets reproduce the same behavioral groupings as the complete data, demonstrating that the proposed imputation process does not distort the knowledge generated by the KDD framework, even under higher levels of missingness.

4.3 Association rules results

The association rule mining process was performed using the Apriori algorithm to identify co-occurring attribute patterns that characterize the operational behavior of the substations. Rules were generated for both the original and the imputed datasets under different levels of missingness to evaluate the stability of the discovered knowledge after data reconstruction. Table 5 summarizes the number of rules generated, as well as the variations in support and confidence between the original and imputed datasets.

The results reveal that the MAKIMA interpolation preserved the rule structure across all missingness levels, with no variation in the total number of generated rules and only negligible differences in support and confidence values. These findings indicate that the imputation process maintained the statistical consistency of the attribute relationships, ensuring that the KDD framework yields equivalent knowledge from both complete and reconstructed data. Consequently, the Apriori-based analysis confirms the robustness of the proposed approach for handling missing data in smart grid datasets.

4.4 Discussion

The analysis of the imputation accuracy metrics reveals a controlled and predictable degradation pattern. The MAE and RMSE values increased linearly with the MCAR level, while the coefficient of determination (R^2) remained consistently above 0.999, indicating a strong preservation of the temporal dynamics in the reconstructed time series. Methodologically, the negligible deviations in the silhouette coefficient (below 0.5%) demonstrate that the MAKIMA interpolation does not significantly alter the density or the separation of the operational clusters identified by the KDD framework.

Table 5. Rules generated from original and imputed datasets.

Period	Rules	MCAR	Δ Conf. (%)	Δ Supp. (%)
01/2018	1000	5	0	0.529
		10	0	0.556
		20	0	0.526
		30	0	0.515
05/2018	846	5	0.070	0.656
		10	0.082	0.654
		20	0.076	0.672
		30	0.077	0.673
12/2019	1000	5	0	0.001
		10	0	0.001
		20	0	0.002
		30	0	0.002

The association rule analysis reinforces this structural stability, as both the number and the logical composition of the rules remained unchanged across all imputed datasets. The minimal fluctuations in support and confidence metrics (below 1%) indicate that the logical relationships among attributes were preserved. This behavior suggests that the MAKIMA method maintains the semantic consistency of the knowledge extracted through the Apriori algorithm, allowing the same diagnostic patterns to be identified despite data gaps.

From an operational standpoint, these findings reflect the framework's ability to deal with data interruptions typical of underground substations, such as communication issues or equipment failures. The stability of the extracted patterns suggests that the MAKIMA interpolation can be integrated into knowledge discovery pipelines to maintain the interpretability and reliability of analytical outcomes. The results indicate a threshold of up to 30% MCAR where the reconstruction effectively supports data-driven decision-making for monitoring and maintenance in underground power distribution systems without distorting the underlying operational reality.

5 Conclusion

This study evaluated the reliability of the hybrid KDD framework for underground substations when applied to datasets reconstructed through the MAKIMA interpolation method. The experimental design, covering three representative months

and missingness levels up to 30%, enabled a systematic validation of the imputation effect on clustering and association rule mining results.

The results confirmed that the proposed approach preserves both the statistical and structural properties of the original datasets, maintaining a high-fidelity reconstruction of the primary current signal. The analysis of the silhouette coefficient and association rules demonstrated that cluster cohesion and logical structures remain virtually unchanged, confirming the stability of the knowledge extracted after imputation.

These findings validate the robustness of the KDD framework under data incompleteness scenarios typical of underground power systems. The MAKIMA method proved capable of reconstructing reliable operational profiles, ensuring that the discovered knowledge remains equivalent to that obtained from complete datasets. This establishes a strong basis for applying the framework to real-world smart grid environments, where analytical integrity must be maintained despite sensor or communication failures. Future work will extend this validation to higher levels of missingness and evaluate the performance limits of the MAKIMA approach in larger-scale or multistation datasets.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

References

[1] Hiroshi Akima. 1970. A new method of interpolation and smooth curve fitting based on local procedures. *Journal of the ACM (JACM)* 17, 4 (1970), 589–602.

[2] Gustavo EAPA Batista and Maria Carolina Monard. 2003. An analysis of four missing data treatment methods for supervised learning. *Applied artificial intelligence* 17, 5-6 (2003), 519–533.

[3] Xinyu Chen, Zhaocheng He, and Lijun Sun. 2019. A Bayesian tensor decomposition approach for spatiotemporal traffic data imputation. *Transportation research part C: emerging technologies* 98 (2019), 73–84.

[4] John W Creswell and J David Creswell. 2017. *Research design: Qualitative, quantitative, and mixed methods approaches.* Sage publications, Thousand Oaks, CA.

[5] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. The KDD process for extracting useful knowledge from volumes of data. *Commun. ACM* 39, 11 (1996), 27–34.

[6] Antonio Carlos Gil. 2008. *Métodos e técnicas de pesquisa social.* 6. ed. Ediitora Atlas SA, São Paulo.

[7] John W Graham. 2012. *Missing data: Analysis and design.* Springer Science & Business Media, New York.

[8] Eklas Hossain, Imtiaj Khan, Fuad Un-Noor, Sarder Shazali Sikander, and Md Samiul Haque Sunny. 2019. Application of big data and machine learning in smart grid, and associated security concerns: A review. *Ieee Access* 7 (2019), 13960–13988.

[9] Heikki Junninen, Harri Niska, Kari Tuppurainen, Juhani Ruuskanen, and Mikko Kolehmainen. 2004. Methods for imputation of missing values in air quality data sets. *Atmospheric environment* 38, 18 (2004), 2895–2907.

[10] Mahdi Khodayar, Guangyi Liu, Jianhui Wang, and Mohammad E Khodayar. 2020. Deep learning in power systems research: A review. *CSEE Journal of Power and Energy Systems* 7, 2 (2020), 209–220.

[11] Roderick JA Little and Donald B Rubin. 2019. *Statistical analysis with missing data.* Vol. 793. John Wiley & Sons, Hoboken, NJ.

[12] MathWorks. 2020. Modified Akima Interpolation (`makima`). Available at: https://www.mathworks.com/help/matlab/ref/makima.html. Acessed on: Sep 01, 2023.

[13] Mathworks. 2023. Modified Akima piecewise cubic Hermite interpolation - MATLAB makima. Available at: https://www.mathworks.com/help/matlab/ref/makima.html. Acessed on: Sep 01, 2023.

[14] Leonardo Minelli, Jonas Fernando Schreiber, Paulo Sérgio Sausen, Airam Teresa Zago Romcy Sausen, and Maurício de Campos. 2024. Smart Grids data characterization: a revision. *Cuadernos de Educación y Desarrollo* 16, 2 (2024), e3357–e3357.

[15] Donald B Rubin. 1976. Inference and missing data. *Biometrika* 63, 3 (1976), 581–592.

[16] Joseph L Schafer. 1997. *Analysis of incomplete multivariate data.* CRC press, Boca Raton, FL.

[17] Jonas Fernando Schreiber, Airam Sausen, Mauricio De Campos, Paulo Sérgio Sausen, and Marco Thomé Da Silva Ferreira Filho. 2023. Data imputation techniques applied to the smart grids environment. *IEEE Access* 11 (2023), 31931–31940.

[18] Yi Wang, Qixin Chen, Tao Hong, and Chongqing Kang. 2018. Review of smart meter data analytics: Applications, methodologies, and challenges. *IEEE Transactions on smart Grid* 10, 3 (2018), 3125–3148.

[19] Raul Sidnei Wazlawick. 2009. *Metodologia de pesquisa para ciência da computação.* Vol. 2. Elsevier Rio de Janeiro, Rio de Janeiro.

[20] Steven Euijong Whang, Yuji Roh, Hwanjun Song, and Jae-Gil Lee. 2023. Data collection and quality challenges in deep learning: A data-centric ai perspective. *The VLDB Journal* 32, 4 (2023), 791–813.

Received ...; revised ...; accepted ...