IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 March 2023, accepted 20 March 2023, date of publication 27 March 2023, date of current version 3 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3262188

RESEARCH ARTICLE

Data Imputation Techniques Applied to the Smart Grids Environment

JONAS FERNANDO SCHREIBER[1], AIRAM SAUSEN[1], MAURICIO DE CAMPOS[1], PAULO SÉRGIO SAUSEN[1], (Member, IEEE), AND MARCO THOMÉ DA SILVA FERREIRA FILHO[2]

[1]Department of Exact Sciences and Engineering, Regional University of Northwestern Rio Grande do Sul (UNIJUÍ), Ijui 98700-000, Brazil
[2]Department of Underground Networks, State Electric Power Distribution Company (CEEE-D), Porto Alegre 91350-180, Brazil

Corresponding author: Paulo Sérgio Sausen (sausen@unijui.edu.br)

This work was supported in part by the State Electric Power Distribution Company (CEEE-D) through Research and Development Program contract number 5000004772.

ABSTRACT The electricity sector has added plenty of new technologies in recent years. Smart Grids are characterized by the use of monitoring and communication technologies almost in whole system. The application and use of such new technologies triggers a significant growth in the data number, increasing the amount of errors and missing data, thus hindering the analysis. In this context, this paper performs the modeling, implementation, validation and comparative analysis of four data imputation techniques: K-Nearest Neighbor, Median Imputation, Last Observation Carried Forward, and Makima. The aim is to verify if they could be applied to the electric segment - more specifically to the Smart Grids environment. The database used in the research is obtained from the electricity utility CEEE and its underground substations, located in southern Brazil. Following this, five simulation scenarios are created and one data set is removed, based on pre-established criteria. Finally, the techniques are applied and the new database is compared with the original one. From the simulation results, the technique which presented the best results is Makima, it is validated as robust to be applied in the Smart Grids environment, especially in electrical data missing from an electric power substation.

INDEX TERMS Electric power system, smart grid, big data, data imputation.

I. INTRODUCTION

The Brazilian Electric Power System (EPS) has gone through changes, especially when it comes to the incorporation of new technologies in the electric grid, once it was developed in the twentieth century and, currently, it must meet new energy consumption needs and demands for security and quality. EPS is segmented into four subsystems: generation, transmission, distribution and consumers (loads), in monitored systems like this, the traditional processing of information from the generation and distribution of energy does not occur through simple operations of data analysis from the substation such as consultation, statistics, and modification [1]. More robust rules are required in those cases, so that it is possible to transform data into information and make decisions. This process is far from trivial, and only by correctly handling these data it is possible to transform a traditional grid into a Smart Grid.

Smart Grids are defined by the Electric Power Research Institute (EPRI) as the superimposition of a unified communication and control system on the existing power distribution system, where sensors are installed in the utility's equipment to monitor and verify its operating environment [1]. The ability to integrate data acquired by the sensors is necessary to solve problems and experience the benefits of this integration, aiming to analyse, monitor, process and obtain near real-time responses. Therefore, Smart Grids and smart meters open up unprecedented business opportunities for utility companies. On the other hand, huge challenges arise in regard to handling and managing large volumes of data, since traditional analysis techniques cannot process such amount effectively. The use of solutions derived from the concept of Big Data

The associate editor coordinating the review of this manuscript and approving it for publication was Akin Tascikaraoglu.

Analytics [2], [3], [4] becomes necessary, since it works with the application of advanced techniques specific to large data sets.

Many problems can occur when large amount of data is generated and handled. Using the extended monitoring system in conjunction with different transmission media, connecting varied sources that send many distributed data, can occasionally corrupt a piece of data from multiple record by different equipment. In addition, data can be corrupted, either by errors in the acquisition system or in the transmission system, or also because they are incomplete due to failures to communicate one or more system magnitudes. In summary, data gaps are likely to occur throughout the power generation and distribution process, resulting in inconsistency in the generation of information from this data. Data can be verified, processed and corrected at the edge of the network, i.e., at power substations, using a decentralized approach as applied in [5] and [6], or even in a centralized way through processing the data on the server, the centralized approach is the one used in this paper.

The search for improvement in data processing in the Smart Grids environment has been the subject of several research projects in recent years. However, as the number of information increases, an increase in record failures also occurs. Thus, the focus of this research is directed towards the treatment of these failures through data imputation techniques, so that existing records are not lost or eliminated due to missing data. In this sense, the choice of an adequate data imputation technique is important, since it brings benefits to society as a whole, providing quality energy to consumers, avoiding financial losses caused by the interruption of energy supply in various sectors of the economy. Rubin [7] discusses the types of failures and how the loss of records interferes with data analysis. Zhou et al. [8] use the based estimation method Last Observation Carried Forward (LOCF) to estimate the lost value and reconstruct the sensing dataset, considering temporal characteristics of sensing data in Wireless Sensor Networks (IWSNs).

Studies in several areas compare existing imputation techniques. Chang and Ge [9] present a comparison of ten data imputation techniques using Bayesian Principle Component Analysis (BPCA), they are applied to the problem of missing data in traffic flow. Majidpour et al. [10] perform a comparison between five techniques: Constant (zero), Mean, Median, Maximum Likelihood, and Multiple Imputation applied to compensate for missing values in Electric Vehicle (EV) charging, both Constant (zero) and Median techniques presented the best results. Pazhoohesh et al. [11] perform a comparative study of eight techniques for imputation of missing values in building sensor data, that is, Monte Carlo Markov Chain (MCMC), Hmisc aregImpute, K-Nearest Neighbours (KNN), Simple Mean, Expectation-Maximization, Random Value, Regression and Stochastic Regression, the authors got to the conclusion that one needs to identify the percentage of missing data before selecting the appropriate imputation technique in order to achieve the best result.

In the electricity sector Khan et al. [12] use a simple average either to impute missing values or to solve data imbalance problem in Electricity Theft Detection (ETD) applied to the smart meter environment. Similarly, Weber et al. [13] came up with a technique called Copy-Paste Imputation (CPI) for energy time series, it checks one or several consecutive missing values and fills in missing values by copying blocks from similar days. For data from smart meters, Peppanen et al. [14] optimize an imputation technique based on Optimally Weighted Average (OWA). Razavi-Far et al [15] develop a new technique for imputation of missing energy data based on correlation-connected clusters. It not only considers local correlation between energy network measurements in estimating missing data, but also handles high-dimensional data and tolerates high missing rates. Zhang et al. [16] propose a new technique for imputation of solar data. They also performed modifications to the unsupervised learning algorithm Generative Adversarial Network (GAN), and they made a comparison with other machine learning techniques, the authors highlight that solar GAN has a great potential to facilitate the forecasting of photovoltaic generation.

In the Smart Grids scenario there is a growing need to apply some procedure or technique that makes it possible to detect and correct the absence of data. By analyzing the related works, the great majority of them uses synthetic data from the most diverse areas, for simulations and evaluation of data imputation algorithms. In the case of the electricity sector, the few existing works that use real data are restricted to applications linked to consumer profiles, such as smart meters, solar systems and electric vehicles. Different from the related works, the contribution of this paper is related to the administration and management of an utility company inserted in the Smart Grid environment, and by using real data from an underground electric power substation of this utility. In the Smart Grids segment, it is noteworthy that there are no studies that clearly and objectively present which and how these techniques can be applied, not even if it is possible to perform their application directly to the problem, or even if they can be adapted.

In this context, this paper performs the modeling, implementation, validation and comparative analysis of four data imputation techniques: K-Nearest Neighbor (KNN) [17], Median Imputation [18], Last Observation Carried Forward (LOCF) [8], and Makima [19], [20] (i.e., Akima's modified algorithm) aiming to verify if they can be applied to the electric segment, more specifically to the Smart Grids environment. The techniques KNN and Makima are chose because they are used in different areas of knowledge in the case of missing data, Imputation by Median and LOCF are chosen because they are the most widely used in monitoring systems for their simplicity of implementation. Following the line of originality, many techniques analysed in the literature are based on synthetic data, focusing only on the evaluation

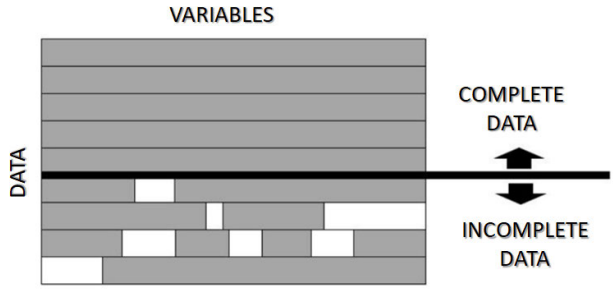


FIGURE 1. Illustration of a database with missing values.

of the technique and not on the scenario in which they are applied. Respecting the importance of working with real data and scenarios, the database used in this paper is obtained from a set of underground substations of the electricity utility CEEE, located in southern Brazil. In the sequence, five simulation scenarios are created and a dataset is removed from each scenario, following pre-established criteria. Finally, the techniques are applied and the new database is compared with the original one. They are implemented in MATLAB®, version R2019a (student license) and evaluated after eight runs and from the calculation of the relative mean error.

II. DATA IMPUTATION

With the rapid development of Information Technology (IT), the Internet of Things (IoT), social networking, e-commerce, and Smart Grids, the amount of data is growing and being accumulated at an unprecedented rate. However, the emergence of incomplete records grows at the same pace, degrading the quality and usability of those databases, turning the research field in data imputation - especially in the electricity sector - a fruitful area for studies and research.

In the power sector, most monitoring systems use data only to generate alarms in certain situations and events. However, with the application of new technologies and the increase in the number of sensors, data is being stored and, above all, analyzed, aiming to transform traditional electrical grids into intelligent environments. In order to perform an adequate analysis of this large volume of data, the computational complexity and the correct sizing of the algorithms must be analyzed, along with the problem of missing values in these bases.

Database failures may occur for a number of reasons. Often data is not recorded due to either an outage, or a sensor failure (i.e., instrumentation). On other occasions the value that is stored is far from the expected range, or not relevant, making it an invalid value. In these scenarios, the reported value is not the real value and it is considered invalid and labeled as a missing value, as shown in Figure 1.

Studies on incomplete data can be found as early as 1962, in which Sebestyen [21] proposes a solution based on chance probability. In the seventies, Rubin [22] presents the multiple imputation method, widely used to solve missing data problems, Chang and Ge [9] define the proportion p of missing data in relation to the total data set, given by:

$$p = \frac{P}{T} \tag{1}$$

where: P is the number of missing datas, and T is the total number of data in the sample.

A. NON-RESPONSE MECHANISMS

The absence of data, also known as Missing Variables (MVs), should not only be considered a problem, but also a way on how to interpret results. The presence of MVs happens frequently in various fields of knowledge and, based on the classification system created by Rubin in [23], a matrix D fromcollected data is considered. This matrix has R lines(i) representing the register and A the columns (j) representing attributes, so that $d_i = (d_{i1}, \ldots, d_{iA})$, where d_{ij} is the value from attribute j for the register i. Aware of this information, matrix D is divided in two sets of data:

$$D = \{D_{existent}, D_{nonexistent}\} \tag{2}$$

where: $D_{existente}$ are the non-missing data and $D_{inexistente}$ are the missing data. For each matrix D exists a matrix F of the same dimension that identifies the missing data, in which $f_{ij} = 1$ if d_{ij} exists, and $f_{ij} = 0$ if d_{ij} does not exist.

The missing data, or non-response mechanism, is considered through the conditional distribution of F in relation to D ($P(F|D)$), and it can be:

1) Missing Completely at Random (MCAR): missing data occurs at random and is not related to any variable, that is, the probability of a not recorded observation does not depend on any other observation in the matrix D, it is given by:

$$P(F|D) = P(F) \tag{3}$$

implying that the probability of occurrence of missing data is the same for all cases, and the cause leading to the occurrence of missing data is a random event.

2) Missing at Random (MAR): missing data has a predictable loss pattern from other variables, that is, the missing data depends only on the recorded information and is correlated with the variable with missing data, that is:

$$P(F|D) = P(F|D_{existent}) \tag{4}$$

if missing data does not depend on values $D_{inexistente}$ and only the values $D_{existente}$, then missing data is caused by some observed variable, which is available for analysis and correlated with the variable that has missing data.

3) Missing Not at Random (NMAR): the most difficult type of data to deal with in an analysis is related to unobserved values that are higher or lower than the sample standard, i.e. the probability of missing data varies in unknown ratios. It happens when F depends

on what data is missing from the matrix $D_{inexistente}$, may also depend on the existing data $D_{existente}$, that is:

$$P(F|D) \neq P(F|D_{existent}). \quad (5)$$

B. TREATMENT TECHNIQUES FOR MISSING VARIABLES

Several techniques can be used for the treatment of MVs in the database; simpler techniques that perform record deletion, and techniques that insert records where data is missing. Directly deleting a record where data is missing is a quick and easy practice, but it leads to data loss, since the existing information is simply ignored. If the absence of records occurs, the best to do is to elect one data imputation technique, where the incomplete records are filled in, thus avoiding loss of information and the amount of data to be analysed.

The Incomplete Case Deletion (ICD) is an example of deletion technique of record affected by missing data [24], in which every record that contains a missing variable is removed, this technique has a simple implementation, but also a high potential for information loss. The Pairwise Deletion Option [25] excludes only samples with missing data in the variables required for analysis are excluded, it also causes loss of information available in the eliminated data and itis simple to be implemented.

There are several techniques for substitution of MVs. The most important ones are listed below. The Average Imputation technique [26] allows an missing value to be replaced by the average of values present in the variable of interest, generating reasonable and fast results. A variation of this technique is Median Imputation [18], this measure of central tendency requires the data to be ordered, so the missing value is replaced by the median, which is the value that divides the data set into equal parts, thus offering good performance. Another commonly used technique is the LOCF [8] which consists in identifying the missing record and replacing it with the non-missing value prior to the missing record, proving to be a quick and easy-to-implement technique. For specific cases, imputation by Zero [18] can be employed, in this technique the missing data are replaced by a constant - in this case zero - and it is often used in situations which either the variable is binary, or zero value is plausible.

Another existing imputation option is the Hot Deck replacement [27] in this technique the replacement of a missing value is performed from a similar value in the current data set. A variation of Hot Deck is Cold Deck replacement [16], this technique differs from the previous for the fact that it uses a value from an external origin dataset to the analysed set. It is an easy technique to implement, although it may not work well when there is a large amount of missing data.

There are also more elaborate techniques, such as Regression Imputation [23] which replaces missing data with predicted values from a regression model. In other words, it imputes missing data based on other variables in the data set. As an example, the KNN technique [17] chooses k neighbors based on some distance measure, and its average is used as an imputation estimate. Another existing technique is Expectation Maximization (EM) [28] it seeks to estimate the parameters of the joint distribution from data, such as the mean vector and the covariance matrix, resulting in punctual estimates of these vectors. And there is also the Cubic Interpolation technique developed by Akima [19] with low computational cost and oscillation, it presents more accurate values close to the data. A modified version of it named Makima is available in the Matlab computational tool [20] directly preventing oscillation when compared to the original technique.

Among the techniques mentioned above, four of them are selected to perform a comparative study and to define which is the most suitable to the data imputation problem related to the electricity sector. Regarding the reasons for choosing these four techniques, the factors taken into consideration are the number of citations in works in related areas, the ease of implementation, use of computer resources, use in monitoring systems, and performance in relation to assertiveness. For a better understanding, the four techniques previously selected will be detailed below.

The first technique chosen is KNN [17], which basically consists of returning the data after replacing the values *NaN* (“Not a Number”) by matching values from the nearest neighbouring column. If the matching value is also *NaN*, the next neighbour column is used. This algorithm calculates the Euclidean distance between observed columns using only the rows that do not have *NaN*, values, so the analyzed data set must have at least one row with complete and error/absence free values.

The second technique chosen is Median Imputation [10], [18], [29], it is technique widely referenced in the literature and easy to implement. This algorithm uses the median calculation to obtain the missing data from the division in two equal parts of a window with k defined records.

The third technique chosen is LOCF [8]. The application of this technique consists of examining the database, locating the missing value *NaN*, and then replacing it with the most recent non-missing value before it. The choice of this technique followed the same criteria previously mentioned and it is widely used in the literature and simple to implement. If the directly preceding position also contains a missing value, the search continues recursively until a valid value is found. If two or more *NaN* records in a row are missing, the LOCF technique continues searching the previous field until a valid record is found. This technique has a disadvantage, it presents execution failure when there is an absence in the first register. This happens due to the nonexistence of a previous register to be used in the replacement, and in this specific case it is necessary to implement an error/exception treatment in the computational implementation of this technique.

The fourth technique chosen is Makima [20] based on the Akima technique [19]. The Makima algorithm for one-dimensional interpolation performs cubic interpolation to produce piecewise polynomials with continuous first-order

derivatives. This algorithm avoids excessive local undulations. If

$$\delta_i = \frac{v_{i+1} - v_i}{x_{i+1} - x_i} \tag{6}$$

is the slope on interval $[x_i, x_{i+1})$ the piecewise cubic interpolation gives a cubic polynomial that can interpolate values v_i and v_{i+1} at nodes x_i and x_{i+1}, then the value of the derivative d_i at the sample point x_i is a weighted average of nearby slopes, given by:

$$d_i = \frac{w_i}{w_1 + w_2}\delta_{i-1} + \frac{w_2}{w_1 + w_2}\delta_i. \tag{7}$$

The weights used in the Makima algorithm are:

$$\begin{aligned} w_1 &= |\delta_{i+1} - \delta_i| + \frac{|\delta_{i+1} + \delta_i|}{2}, \\ w_2 &= |\delta_{i-1} - \delta_{i-2}| + \frac{|\delta_{i-1} + \delta_{i-2}|}{2}. \end{aligned} \tag{8}$$

The Makima algorithm gives priority to the side closest to the horizontal, which is more intuitive and avoids overshoot. In particular, whenever there are three or more consecutive collinear points, the algorithm connects them with a straight line and thus avoids an overshoot [20].

III. STUDY OF CASE

A real database[1] is used in order to carry out this research. It is obtained from sensors installed in a set of underground substations at the electricity utility CEEE, located in southern Brazil. Initially the entire database is analysed, and then a period of 2 years is defined (i.e., 2018/2019). This period is chosen with the objective of obtaining a significant number of substations that reflect the energy consumption pattern of two complete annual cycles, which makes it possible to reduce any variation in energy consumption and consequently in the variation of data.

From the set of 42 substations, the unit chosen for the implementation of data imputation techniques is the one with the highest number of samples as the shortest time interval between samples and no interruptions or data loss. The sampling rate of the monitoring system is one package every 10 seconds. However, due to several problems, either in the acquisition system or even in the transmission system, it is not possible to maintain this sampling frequency for a period that could be considered satisfactory.

Then, the reading periodicity is increased by one packet every 60 seconds, and of the 42 substations only the substation named in this paper as *id*17 managed to have a complete sequence of error-free data for 14,427 uninterrupted minutes, which is equivalent to a 10-day window of complete records, i.e., 09/06/2019 to 19/06/2019. This case study, along with the applied methodology, proves the difficulty of obtaining an integrated and complete set of data in inhospitable environments such as an underground electric power substation.

[1]The database can be made available upon request by email to the authors.

TABLE 1. Measurements used in this research.

Measurements
Secondary Voltage A
Secondary Voltage B
Secondary Voltage C
Primary Current A
Primary Current B
Primary Current C
Secondary Current A
Secondary Current B
Secondary Current C
Room Temperature
Transformer Temperature Range

Table 1 presents the measurements that form the database used in this research.

Once the database is obtained, the methodology chosen to perform the comparison among the four imputation techniques aforementioned is the one which inserts errors/faults randomly into the original database, generating a new database with errors [30]. From this methodology it is possible to measure the effectiveness of mechanisms to correctly complete missing data. The insertion of errors in the original base follows an increasing percentage of errors that generate different scenarios to simulate and analyse the behaviour and the efficiency of the analysed methods from the average of runs, by comparing differences between the estimated/imputed data with the original base and without errors.

The Relative Error (RE) is used to compare the chosen techniques, that is:

$$E_r = \frac{|x_i - \overline{x}|}{\overline{x}}. \tag{9}$$

where: E_r is the relative error, x_i is the imputed value, and $\overline{x}$ is the actual value.

IV. RESULTS AND DISCUSSIONS

This section presents the simulation results considering the four techniques: Median Imputation, LOCF, KNN, and Makima. From the substation database *id*17, five scenarios are created with 14,427 records of 11 analog magnitudes. The first three 1%, 3% and 5% of records are respectively removed from the original database. With the aim of evaluating the effectiveness of the techniques in extreme situations, in the fourth scenario the Percentage of Missing Records (PMR) is increased and can reach up to 99% of missing data. The fifth scenario aims to analyse which technique has better performance for imputation of missing data when applied to the primary currents that suffer significant variation. Afterwards, the data imputation techniques are applied, then the new database is compared with the original one, and the techniques are evaluated from eight runs and the calculation of the relative average error. The five scenarios, application of the techniques, simulation results and discussions are presented below.

A. FIRST SCENARIO

From the complete database 1% of data is randomly removed, i.e. 1,586 records. This random removal consists of replacing the existing value with the content *NaN*. An auxiliary base of the same size is created to store the values removed from the complete base, to facilitate comparison of results.

Among the techniques presented above, KNN is the first one evaluated. To perform the simulation, the function *knnimpute* of the MATLAB software is used. The generated base is shielded for later analysis, after imputation of the missing values.

The KNN technique is run eight times and then the Mean Relative Error for evaluation (ER_{med}) is calculated. At the end a scan is performed on the simulated base and the imputed value is compared with the original value of the initial complete base. For the simulation results of the KNN technique in this scenario, it is found the ER_{med} of 10.61%.

The second technique evaluated is Median Imputation. The function *fillmissing* of MATLAB is used to fill the registers with value *NaN* by a value calculated from an argument given by the *fillmissing*. In the case of the median, the argument used by the function is the *movmedian* with a calculation window of 10 records. The Median Imputation technique is run eight times and the ER_{med} found is of 1.56%.

The LOCF technique is the third one evaluated. For the implementation, the function *fillmissing* of MATLAB software is used by passing the argument *previous*. The implemented algorithm replaces the value *NaN* with the first existing valid previous record. From eight runs performed on the database with 1% of missing data the ER_{med} of 1.64% is obtained.

Makima technique is the last one evaluated. For this simulation the function *fillmissing* of the MATLAB software is also used, inserting the argument *makima*. From the results of the simulations of eight runs on the database the ER_{med} of 1.38% is found. This technique got the best performance among the four ones evaluated.

B. SECOND SCENARIO

Using the same methodology presented in the first scenario, another database is created, now with 3% of the removed records, called the second scenario. From the complete base, 4,759 records are replaced by the content *NaN*. The same methodology adopted in the first scenario for the simulations and evaluation of the four techniques is adopted in the second scenario.

After performing the simulations, the KNN technique presents an ER_{med} de 16.94%, the Median Imputation technique presents an ER_{med} de 1.60%, while for LOCF technique a ER_{med} of 1.66% is found, and lastly Makima technique has a ER_{med} of 1.41%. For the second scenario, the Makima technique also performed the best.

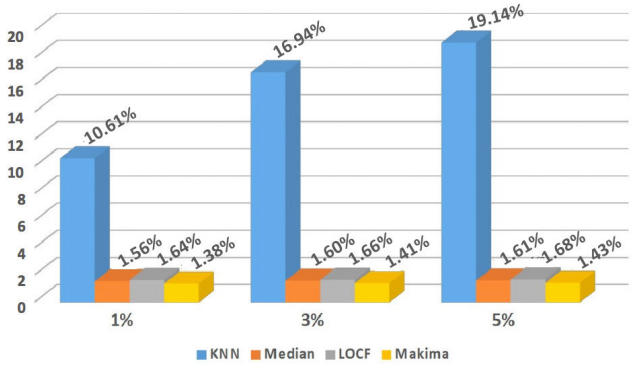


FIGURE 2. Comparative analysis of the ER_{med} of the four data imputation techniques.

TABLE 2. ER_{med} for 8 techniques execution for each scenario.

Technique	First scenario (1%)		Second scenario (3%)		Third scenario (5%)	
	ER_{med}	SD	ER_{med}	SD	ER_{med}	SD
KNN	10.61%	2.5157	16.94%	2.9277	19.14%	2.2530
Mediana	1.56%	0.0414	1.60%	0.0307	1.61%	0.0275
Previous	1.64%	0.0389	1.66%	0.0223	1.68%	0.0247
Makima	1.38%	0.0378	1.41%	0.0193	1.43%	0.0297

C. THIRD SCENARIO

The third scenario extends the removal to 5% of the original database records, i.e. 7.925 *NaN* contents are inserted. The same methodology adopted in the scenario 1 and 2 to carry out the simulations and evaluate the four techniques is adopted in the third scenario. After performing the simulations, the KNN technique presents an ER_{med} de 19.14%, the Median Imputation technique presents an ER_{med} of 1.61%, the LOCF technique has a ER_{med} of 1.68%, and again the Makima technique obtained the best result with a ER_{med} of 1.43%.

Figure 2 presents a comparison between the four data imputation techniques. The superiority of Makima technique is observed in relation to the others, followed by the Median Imputation, LOCF, and with minor results, the KNN technique.

In order to facilitate the comparison between the four techniques evaluated, Table 2 presents the ER_{med} results for the three scenarios in summarised form and the respective Standard Deviation (SD).

D. FOURTH SCENARIO

The effectiveness of the methods in extreme situations is evaluated in the fourth scenario, here the PMR is incremented and can reach up to a limit of 99% of missing data. To do so, from the original database 10% of records are incrementally and randomly removed and are replaced by *NaN* until the point

TABLE 3. PMR versus ER_{med} for the KNN technique.

PMR	10%	20%	30%	40%
ER_{med}	30.63%	52.39%	78.85%	108.99%

TABLE 4. PMR versus ER_{med} for the Median Imputation technique.

PMR	10%	20%	30%
ER_{med}	1.62%	1.66%	-

TABLE 5. PMR versus versus ER_{med} for the LOCF technique.

PMR	10%	20%
ER_{med}	1.73%	-

when the technique can no longer perform the imputation of missing data in a satisfactory way.

The KNN technique is evaluated first, it reaches a threshold of 30% of missing data, as seen in Table 3. When 40% of data are removed the technique achieves an ER_{med} superior to 100%. In this case, the percentage of 40% is considered the limit for the application of this technique, since the ER_{med} represents an imputed value more than double the original value. It can also be seen that as the PMR is incremented, the ER_{med} also increases.

The Median Imputation technique is evaluated by the same methodology. Table 4 shows that this technique reaches a limit of 20% of missing data, because when a PMR of 30% is applied an execution error occurs and the technique cannot correctly estimate the missing values. This is because the Median Imputation technique cannot calculate the missing value because the averaging method finds a *NaN* inside the calculation window, which makes the simulation unfeasible and it is therefore terminated.

The next technique analysed is LOCF. Table 5 shows that this technique does not perform well, as it reaches its limit at 10% of missing data, when a PMR of 20% is applied there is an execution error and the technique is no longer able to correctly estimate the missing values. The LOCF technique uses the immediately preceding record to replace the missing one, which causes the failure. If there are two or more *NaN* records in a sequence, the algorithm searches the previous one recursively until it finds the first valid value available. This technique does not work if the first record is a *NaN* value. Since random removals are carried out, when 20% of the records are removed, a data that is in the first position is also removed, making it impossible to proceed with the tests.

Finally, the Makima technique is the last one evaluated. This technique presented an excellent result, since it reached a percentage of 99% of missing values managing to estimate a value to replace the removed record with low ER_{med}. Table 6 presents the results found for the Makima technique.

TABLE 6. PMR versus versus ER_{med} for the Makima technique.

PMR	ER_{med}
10%	1.44%
20%	1.48%
30%	1.51%
40%	1.56%
50%	1.65%
60%	1.71%
70%	1.84%
80%	2.05%
90%	2.28%
99%	5.22%
100%	-

The technique uses the argument *makima* based on the existing records in the base to perform the calculation and subsequent replacement of the *NaN* records by the values obtained. Thereafter, the technique responds well until it reaches 99% of missing data. Considering the results for the ER_{med} presented in Table 6 an elevation of this error can be oserved when the percentage of missing data exceeds 70%. From the simulation results of all the evaluated techniques for the fourth scenario, it can be clearly seen that the Makima technique obtained the best performance.

On the other hand, Median Imputatio and LOCF techniques are discarded for not being able to present a stable imputation sequence, since these techniques cannot determine a valid value to be imputed and they depend on the position of the missing records. The KNN technique is able to impute values for the removed data, however, when it reaches a percentage of 40% of record removal, it inserts values that resulted in a difference of over 100% from the original database values. Makima is therefore the one to achieve the best results in the fourth scenario.

E. FIFTH SCENARIO

The fifth scenario is defined in order to confirm if Makima really has the best performance in the imputation of missing data applied to the electric sector - more specifically in Smart Grids. In this scenario, only the three primary currents of the system are kept to evaluate the techniques performance. These measurements are chosen because they suffer significant variation when compared to other system measurements

In order to simplify and potentiate the tests considering the three distinct phases of the current (i.e., phase A, B and C), the removal of the registers is performed only in phase B. Following the same methodology as the three first scenarios,

TABLE 7. Result of data imputation for the primary current in phase B.

Technique	Removal of 1%		Removal of 3%		Removal of 5%	
	ER_{med}	SD	ER_{med}	SD	ER_{med}	SD
KNN	6.27%	0.2754	6.28%	0.2716	6.20%	0.1076
Median	3.04%	0.1684	3.11%	0.1369	3.14%	0.0652
LOCF	3.21%	0.4032	3.18%	0.1356	3.30%	0.0886
Makima	2.67%	0.2625	2.81%	0.1416	2.80%	0.1029

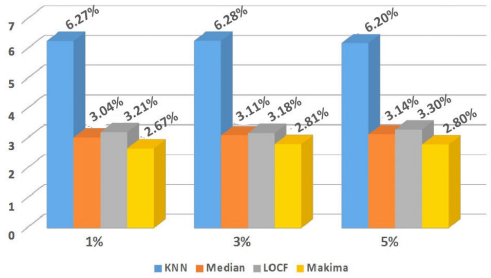


FIGURE 3. Comparison of ERM between mechanisms.

TABLE 8. PMR versus versus ER_{med} for the Makima technique.

PMR	ER_{med}
10%	2.78%
20%	2.84%
30%	2.98%
40%	3.00%
50%	3.13%
60%	3.28%
70%	3.54%
80%	3.80%
90%	4.16%
99%	6.92%
100%	-

the data is randomly removed with percentages of 1%, 3% and 5% from the total of records.

Considering the random removal of 1% of data from the B-phase primary stream, 144 records are removed; for 3% of the data, 432 records are removed; and with 5% of the data, 721 records are replaced by the *NaN* content. Auxiliary bases of the same size are created to store values removed from the main base, to facilitate the comparison with the simulated values. To evaluate all techniques, the ER_{med} is used; for this scenario eight repetitions are also performed. Table 7 presents the simulation results.

Figure 3 presents the ER_{med} for the four techniques evaluated in the fifth scenario, except for the KNN technique, all others increased their ER_{med} rates when compared to the results obtained in scenarios 1, 2 and 3. On the other hand, the Makima technique achieved once again the best performance. By repeating the methodology used in the fourth scenario, data removal percentage for the Makima technique is incrementally and randomly extended, starting again at 10%. This technique again presented an excellent result, since it reached a percentage of 99% of missing values managing to estimate a value to replace the removed record with ER_{med} relatively low. Table 8 shows the percentages of ER_{med},a small variation in the error rate is observed even significantly increasing the percentage of missing data. Therefore, the Makima technique is validated as robust for application in Smart Grids environment, especially in missing electrical data of an electric power substation.

V. CONCLUSION

This paper performs the implementation, validation and comparative analysis of four data imputation techniques: K-Nearest Neighbor, Median Imputation, Last Observation Carried Forward, and Makima to be applied in the electric segment, more specifically in the Smart Grids environment. From a real database of an underground substation of the electricity utility CEEE, located in southern Brazil, five test scenarios are created and data is removed according to different pre-established criteria, even extreme data missing situations are tested. In sequence, data imputation techniques are applied, and the new database is compared with the original one.

In the first three scenarios 1%, 3% and 5% of records are removed from the original database, and in all simulation results the Makima technique showed the lowest ER_{med} with 1.38%, 1.41% and 1.43% respectively. In the fourth scenario, extreme situations are evaluated and the percentage of missing records is increased, reaching up to 99% of the data, considering the application and analysis of the four techniques, the only one that managed to complete the database with up to 99% of missing data is Makima with an ER_{med} of at most 5.22%. The fifth scenario aims to analyze which technique has the best performance for imputation of missing data when applied to the currents of primary phase B, and again 1%, 3% and 5% of records are removed from the original database, in all simulations results the Makima technique presented the lowest ER_{med} with 2.67%, 2.81% and 2.80% respectively. The fifth scenario is also evaluated for extreme conditions, and the Makima technique is again the only one that was able to complete the database with up to 99% of missing data with a ER_{med} of 6.92% at most.

Therefore, the Makima technique is considered superior to the others, being robust and suitable to be applied in a Smart Grid environment, especially in missing data in the processing and monitoring systems of electric power underground substations. As future works, it is suggested both the application of the Makima data imputation technique in a

decentralized way, that is, directly in the substation's internal acquisition system, and also in the real monitoring system of the underground substations installed at the electricity utility CEEE, RS, Brazil.

REFERENCES

[1] *The Green Grid—Energy Savings and Carbon Emissions Reductions Enabled by a Smart Grid*, Electr. Power Res. Inst., Washington, DC, USA, 2007. Accessed on: Oct. 11, 2022. [Online]. Available: https://tinyurl.com/rbhrf3q

[2] D. Alahakoon and X. Yu, "Advanced analytics for harnessing the power of smart meter big data," in *Proc. IEEE Int. Workshop Inteligent Energy Syst. (IWIES)*, Vienna, Austria, Nov. 2013, pp. 40–45, doi: 10.1109/IWIES.2013.6698559.

[3] P. Russom. (2011). *Big Data Analytics. TDWI Best Practices Report 4th Quarter*. pp. 1–34, Accessed: Oct. 11, 2022. [Online]. Available: https://tinyurl.com/49tcz2j5

[4] T. Dargahi, H. Ahmadvand, M. N. Alraja, and C.-M. Yu, "Integration of blockchain with connected and autonomous vehicles: Vision and challenge," in *Proc. IEEE Int. Workshop Intell. Energy Systems (IWIES)*, Vienna, Austria, Dec. 2022, vol. 14, no. 1, pp. 1–10, doi: https://doi.org/10.1145/3460003.

[5] H. Ahmadvand, T. Dargahi, F. Foroutan, P. Okorie, and F. Esposito, "Big data processing at the edge with data skew aware resource allocation," in *Proc. IEEE Conf. Netw. Function Virtualization Softw. Defined Netw. (NFV-SDN)*, Heraklion, Greece, Nov. 2021, pp. 81–86, doi: 10.1109/NFV-SDN53031.2021.9665051.

[6] H. Ahmadvand and F. Foroutan, "DV-ARPA: Data variety aware resource provisioning for big data processing in accumulative applications," 2020, *arXiv:2008.04674*.

[7] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976, Accessed 11, Oct. 2022, doi: 10.2307/2335739.

[8] H. Zhou, K.-M. Yu, M.-G. Lee, and C.-C. Han, "The application of last observation carried forward method for missing data estimation in the context of industrial wireless sensor networks," in *Proc. IEEE Asia–Pacific Conf. Antennas Propag. (APCAP)*, Aug. 2018, pp. 1–2, doi: 10.1109/APCAP.2018.8538147.

[9] G. Chang and T. Ge, "Comparison of missing data imputation methods for traffic flow," in *Proc. Int. Conf. Transp., Mech., Electr. Eng. (TMEE)*, Dec. 2011, pp. 639–642, doi: 10.1109/TMEE.2011.6199284.

[10] M. Majidpour, P. Chu, R. Gadh, and H. R. Pota, "Incomplete data in smart grid: Treatment of missing values in electric vehicle charging data," in *Proc. Int. Conf. Connected Vehicles Expo (ICCVE)*, Nov. 2014, pp. 1041–1042, doi: 10.1109/ICCVE.2014.7297505.

[11] M. Pazhoohesh, Z. Pourmirza, and S. Walker, "A comparison of methods for missing data treatment in building sensor data," in *Proc. IEEE 7th Int. Conf. Smart Energy Grid Eng. (SEGE)*, Aug. 2019, pp. 255–259, doi: 10.1109/SEGE.2019.8859963.

[12] I. U. Khan, N. Javaid, C. J. Taylor, K. A. A. Gamage, and X. Ma, "Big data analytics for electricity theft detection in smart grids," in *Proc. IEEE Madrid PowerTech*, Jun. 2021, pp. 1–6, doi: 10.1109/PowerTech46648.2021.9495000.

[13] M. Weber, M. Turowski, H. K. Cakmak, R. Mikut, U. Kuhnapfel, and V. Hagenmeyer, "Data-driven copy-paste imputation for energy time series," *IEEE Trans. Smart Grid*, vol. 12, no. 6, pp. 5409–5419, Nov. 2021, doi: 10.1109/TSG.2021.3101831.

[14] J. Peppanen, X. Zhang, S. Grijalva, and M. J. Reno, "Handling bad or missing smart meter data through advanced data imputation," in *Proc. IEEE Power Energy Soc. Innov. Smart Grid Technol. Conf. (ISGT)*, Sep. 2016, pp. 1–5, doi: 10.1109/ISGT.2016.7781213.

[15] R. Razavi-Far, M. Farajzadeh-Zanjani, M. Saif, and S. Chakrabarti, "Correlation clustering imputation for diagnosing attacks and faults with missing power grid data," *IEEE Trans. Smart Grid*, vol. 11, no. 2, pp. 1453–1464, Mar. 2020, doi: 10.1109/TSG.2019.2938251.

[16] W. Zhang, Y. Luo, Y. Zhang, and D. Srinivasan, "SolarGAN: Multivariate solar data imputation using generative adversarial network," *IEEE Trans. Sustain. Energy*, vol. 12, no. 1, pp. 743–746, Jan. 2021, doi: 10.1109/TSTE.2020.3004751.

[17] P. Keerin, W. Kurutach, and T. Boongoen, "Cluster-based KNN missing value imputation for DNA microarray data," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2012, pp. 445–450, doi: 10.1109/ICSMC.2012.6377764.

[18] E. Patrick, K. McKnight, S. Souraya, and A. Figueredo, *Missing Data: A Gentle Introduction*. New York, NY, USA: Guilford Press, 2007.

[19] H. Akima, "A new method of interpolation and smooth curve fitting based on local procedures," *J. ACM*, vol. 17, no. 4, pp. 589–602, Oct. 1970, doi: 10.1145/321607.321609.

[20] MATLAB. (2022). *MathWorks*. Accessed: Oct. 11, 2022. [Online]. Available: https://https://www.mathworks.com/products/MATLAB.html

[21] G. S. Sebestyen, *Decision-Making Processes in Pattern Recognition*. New York, NY, USA: Macmillan, 1962.

[22] D. B. Rubin, "Multiple imputations in sample surveys: A phenomenological Bayesian approach to nonresponse," in *Proc. Surv. Res. Methods Sect., Amer. Stat. Assoc. (ASA)*, 1978, pp. 20–28.

[23] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*. Hoboken, NJ, USA: Wiley, 1987.

[24] J. Brand, S. Buuren, E. M. Mulligen, T. Timmers, and E. Gelsema, "Multiple imputation as a missing data machine," in *Proc. Annu. Symp. Comput. Appl. Med. Care*, 1994, pp. 6–303.

[25] P. D. Allison, "Multiple imputation: Basics," in *Missing Data*. Thousand Oaks, CA, USA: SAGE, 2002, pp. 28–50, doi: 10.4135/9781412985079.n5.

[26] M. Fichman and J. N. Cummings, "Multiple imputation for missing data: Making the most of what you know," *Organizational Res. Methods*, vol. 6, no. 3, pp. 282–308, 2003, doi: 10.1177/1094428103255532.

[27] K. Lakshminarayan, S. A. Harp, and T. Samad, "Imputation of missing data in industrial databases," *Appl. Intell.*, vol. 11, pp. 259–275, Nov. 1999, doi: 10.1023/A:1008334909089.

[28] G. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*. Hoboken, NJ, USA: Wiley, 2008.

[29] A. T. Sree Dhevi, "Imputing missing values using inverse distance weighted interpolation for time series data," in *Proc. 6th Int. Conf. Adv. Comput. (ICoAC)*, Dec. 2014, pp. 255–259, doi: 10.1109/ICoAC.2014.7229721.

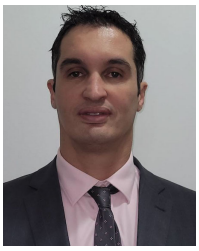
[30] T. Duy Le, R. Beuran, and Y. Tan, "Comparison of the most influential missing data imputation algorithms for healthcare," in *Proc. 10th Int. Conf. Knowl. Syst. Eng. (KSE)*, Nov. 2018, pp. 247–251, doi: 10.1109/KSE.2018.8573344.

JONAS FERNANDO SCHREIBER received the bachelor's degree in information systems and the master's degree in mathematical modeling from the Regional University of Northwestern Rio Grande do Sul (UNIJUI), Ijui, Brazil, in 2005 and 2013, respectively, where he is currently pursuing the Ph.D. degree in a mathematical modeling program. He has experience in the information technology (IT) area, with an emphasis on the software programming.

AIRAM SAUSEN received the degree in mathematics and the master's degree in mathematical modeling from the Regional University of Northwestern Rio Grande do Sul (UNIJUI), in 2002 and 2004, respectively, and the Ph.D. degree in electrical engineering from the Federal University of Campina Grande (UFCG), in 2009. Currently, she is an Adjunct Professor with UNIJUI and a Coordinator of the stricto sensu postgraduate program in mathematical modeling. She is also a Researcher of the Industrial Automation and Control Group (GAIC), and a Member of the NDE for a degree in mathematics. She has experience in the areas of applied mathematics, mathematical modeling, and simulation.

MAURICIO DE CAMPOS received the degree in electrical engineering from the Regional University of Northwestern Rio Grande do Sul (UNIJUI), Brazil, in 1997, the M.S. degree in electrical engineering from the Federal University of Santa Maria, Brazil, in 2000, and the Ph.D. degree in electrical engineering from the Federal University of Campina Grande, Brazil, in 2017. He is currently an Assistant Professor with UNIJUI. He has experience in electrical engineering with an emphasis on electronic and electrical process automation.

PAULO SÉRGIO SAUSEN (Member, IEEE) received the B.S. degree in computer science from the Regional University of Northwestern Rio Grande do Sul (UNIJUI), Ijui, Brazil, in 1993, and the M.Sc. degree in computer science from the Federal University of Paraiba (UFPB), Campina Grande, Brazil, in 1998, and the Ph.D. degree in electrical engineering from the Federal University of Campina Grande (UFCG), in 2008. He is currently an Associate Professor with UNIJUI. His current research interests include the design of algorithms and protocols for sensor networks and smart grids.



MARCO THOMÉ DA SILVA FERREIRA FILHO received the degree in civil engineering and the sensu lato postgraduate degree in public administration from the Federal University of Rio Grande do Sul, in 2006 and 2015, respectively. He is currently a Research and Development Manager and the Head of the services for underground networks. He coordinates the areas of maintenance, operation, and construction in underground power networks of State Electricity Distribution Company (CEEE-D)–Equatorial Group, Rio Grande do Sul.