Hindawi
International Transactions on Electrical Energy Systems
Volume 2022, Article ID 8546588, 10 pages
https://doi.org/10.1155/2022/8546588

Research Article

Preprocessing Approach for Power Transformer Maintenance Data Mining Based on *k*-Nearest Neighbor Completion and Principal Component Analysis

Moïse Manyol,[1] **Samuel Eke**,[1] **Alphonse J. M. Massoma**,[1] **Alain Biboum**,[2] **and Ruben Mouangue**[1]

[1] *Energy, Materials, Modeling, and Methods Research Laboratory (LE3M), Polytechnic Higher National School, University of Douala, 2701, Pk. 17 Logbessou, Douala, Cameroon*
[2] *Mechanical and Industrial Engineering Department, National Advanced School of Engineering of Yaounde, University of Yaounde I, 33100 Melen1, Yaounde, Cameroon*

Correspondence should be addressed to Moïse Manyol; moisemany@yahoo.fr

Received 7 August 2022; Revised 13 September 2022; Accepted 15 September 2022; Published 3 October 2022

Academic Editor: Ibrahim Mahariq

The accuracy of a knowledge extraction algorithm in a large database depends on the quality of the data preprocessing and the methods used. The massive amounts of data that we collect every day are putting storage capacity at a premium. In reality, many databases are characterized by attributes with outliers, redundant, and even more missing values. Missing data and outliers are ubiquitous in our databases, and imputation techniques will help us mitigate their influence. To solve this problem, as well as the problem of data size, this paper proposes a data preprocessing approach based on the *k*-nearest neighbor (KNN) completion for imputation of missing data and principal component analysis (PCA) for processing redundant data, thus reducing the data size by generating a significant quality sample after imputation of missing and outlier data. A rigorous comparison is made between our approach and two others. The dissolved gas data from Rio Tinto Alcan's transformer T0001 were imputed by KNN, where *k* equals 5. For 6 imputed gases, the average percentage error is about 2%, 17.5% after average imputation, and 23.65% after multiple imputations. For data compression, 2 axes were selected based on the elbow rule and the Kaiser threshold.

1. Introduction

The large size of today's databases poses the problem of archiving and preprocessing raw data. The data may be missing, aberrant, or redundant. The application of data analysis algorithms on such data complicates the learning process and affects the performance and reliability of the model [1]. Data preprocessing is undoubtedly crucial in the process of knowledge discovery from these voluminous databases. Indeed, it allows for improving the quality of the data submitted later to the data mining algorithms. As far as outliers are concerned, it is good to keep them original and mysterious in the raw data if possible. In other words, the reason for removing outliers should come from outside the dataset only when you already know the originals [2]. Ignoring missing data can lead to a loss of precision and strong biases in the analysis models. Missing data are represented by the so-called missing values matrix, the form of which depends on the type of missing data. Generally, we distinguish between MCARs, MARs, and MNARs. The MCAR is the missing data in a completely random way if the probability of absence is the same for all the observations. This probability depends only on external parameters independent of this variable. MARs are data that are not missing completely at random if the probability of absence is related to one or more other observed variables. MNARs are nonrandomly missing data if the probability of absence depends on the variable in question. In summary, these fall into three categories and are extensively detailed in [3]. A presentation and a review have been made on the different assumptions and techniques for processing these data [4, 5]. In 2009, the authors of the work [4] developed a modern method called Multiple Imputation by Chained Equations (MICE) based on a Monte Carlo Markov Chain algorithm for imputing missing values. To eliminate missing values, noise, and redundant attributes, and also to reduce their size by generating representative and quality samples. The work [6] presents a method based on the calculation of the empirical copula of the original sample. The performance of this work was only compared to the one whose imputation is based on the mean. To classify data from the gases below [7], preprocess was performed in 2014 using the *k*-nearest neighbors (KNN). This shows that classification without preprocessing the data first is less accurate. Grzymala–Busse processed the unknown values of variable attributes by setting up new decision tables with known attributes instead of the original table, which had unknown and inconsistent attributes in their work [8]. The rough set theory is developed for learning inconsistent rules..

This paper proposes a preprocessing method that combines the *k*-nearest neighbors and the principal component analysis (PCA). The *k*-nearest neighbor classifier is generally used in supervised classification because of its simplicity and robustness. To make it even more robust, i.e., less sensitive to data variations as shown in this work, references [9, 10] have respectively developed a *k*-nearest neighbor classifier based on the generalized mean distance. The main objective is to overcome the sensitivity of the *k*-neighborhood size and improve the KNN-based classification's performance. In this proposed approach, the nested generalized mean distance computed by the multi-generalized mean distances in each class is designed for KNN-based classification decisions. The proposed method is shown to be suitable for pattern detection, on the one hand. On the other hand, a new representation method based on the *k*-nearest neighbor centroid coefficient is developed, which also aims to further improve the classification performance and reduce the sensitivity of the method to the size of the *k*-neighborhood, especially in cases of small sample size. In this work, the *k*-nearest neighbor-based classifier is used in the framework of data completion, and some works nowadays have used it in the framework of data imputation [11, 12].

The *k*-nearest neighbor method used for the imputation of missing data is based on the optimization of the choice of the *k* parameter, which relies on mechanical or discontinuous mean fitting techniques. For a *k*-nearest neighbor mean equal to 5 and for a Euclidean distance between two observations, 6 missing values are imputed. To observe the variables H_2 (hydrogen) and CH_4 (methane), which are low molecular weight gases, a weighting of N_2 (nitrogen), CO_2 (carbon dioxide), and CO (carbon monoxide) is recommended.

For PCA based on the weighted strategy, robustness will be improved by mitigating the statistical impact of outliers through reduced weighting [13]. Similarly, it will solve the normalization problem and the increasing difficulties of archiving after a judicious choice of the number of axes to be retained by using the kink theory and the Kaiser threshold. From the works [7, 14] it appears that the PCA associated with a classifier (ANN) is more accurate than the support vector machine (SVM) associated with the KNN classifier. It then becomes judicious to join a classifier (KNN) to an exploratory technique on the data (PCA). After a reminder of the KNN algorithm and the PCA algorithm in Section 2, the proposed preprocessing approach is presented in Section 3. Finally, the results of our method applied to a dissolved gas database are compared to other methods presented in Section 4, followed by a discussion.

2. Materials and Methods

The approach proposed here is to present the imputation techniques by completion, in which the data will be submitted and to consider outliers.

2.1. Imputation Methods. The most common imputation techniques are presented here in a nonexhaustive way, namely: stationary, linear combination, *k*-nearest neighbor, NIPALS, and multiple completions, to mention only those. A dataset consists of p quantitative or qualitative variables ($Y = (y_{ij}) = (Y_1, \ldots, Y_p)$) observed on a sample of n individuals; M denotes the matrix indicating the missing values by $M = m_{ij} = 1_{\{y_{ij}, \text{mis}\}}$.

2.1.1. Stationary and Linear Combination Completion. There are several possible stationary imputations. The most common attribute value fitting (CMCF) [15] or simply the last known observation carried forward (LOCF) is given as follows:

$$\left(y_{ij}\right)_{\text{mis}} = y_{i^* j^*} = \left\{y_{i^* j} | m_{i^* j} = 0, j < j^*\right\}, \quad (1)$$

where $(y_{ij})_{\text{mis}}$ represents the missing data.

This method may seem too naive, but it is often used to lay the foundation for a comparison between imputation methods. Another common technique is to replace all missing values with a linear combination of observations. The case of imputation by the mean is given as follows:

$$\left(y_{ij}\right)_{\text{mis}} = y_{i^* j^*} = \overline{Y_{j^*}}. \quad (2)$$

Or by the median, as follows:

$$\left(y_{ij}\right)_{\text{mis}} = y_{i^* j^*} = \widetilde{Y_{j^*}}. \quad (3)$$

But this case is generalized to any weighted linear combination of observations. Instead of using all the available values, it is possible to restrict oneself to methods that select the most influential values by local aggregation or regression or even by combining different aspects.

2.1.2. Completion by the Nearest Neighbor Method. The *k*-nearest neighbors (KNN) imputation consists of running the following algorithm that models and predicts the missing data. First, the choice of the parameter k ($1 \leq k \leq n$), calculate the metric distances $d(y_{i^*}, y_i)$, $i = 1, \ldots, n$ retain the k observations $y_{(i1)}, \ldots, y_{(ik)}$, for which these distances are smaller; and finally, assigning to the missing values the arithmetic mean of the values of the k neighbors, such that is given as follows:

$$\left(y_{ij}\right)_{\text{mis}} = y_{i^*j^*} = \frac{1}{k}\left(y_{i_1}, \ldots, y_{i_k}\right). \tag{4}$$

Or y_{i^*} an observation with q missing values.

2.1.3. NIPALS Completion Algorithm. The NIPALS (Nonlinear Iterative Partial Least Squares) algorithm is an iterative method for estimating PLS (Partial Least Square) regression. This algorithm can be adapted to impute missing data. Or $Y = (Y_1, \ldots Y_p)$ such as $\forall i \in 1, \ldots, p$, mathematical expectation $E(Y_i) = 0$ (each column of the matrix is centered). The expansion of Y in terms of principal components and principal factors is given by following equation:

$$Y = \sum_{h=1}^{q} \varepsilon_h u_h, \tag{5}$$

where $q = \dim L_2(X)$; $\{\xi_h\}_{h=1,\ldots,q}$ are the principal components and $\{u_h\}_{h=1,\ldots,q}$ the principal vectors of the PCA of Y. For each variable of Y_i, there is the equation given as follows:

$$Y_i = \sum_{h=1}^{q} \varepsilon_h u_h(i), \tag{6}$$

where $u_h(i)$ represents the slope of the linear regression of Y_i on the component ξ_h. The development of this algorithm is contained in [16].

2.1.4. Multiple Imputation. Multiple imputations retain the virtues of single imputation and correct its main shortcomings. As its name suggests, it consists of imputing missing values several times to combine the results to reduce the error due to imputation [17]. The multiple imputation procedure consists of two phases: the imputation phase and the statistical analysis phase. These two phases use two different models: the imputation model and the analysis model. Once the imputations have been performed, the statistician can perform any type of analysis, according to the standard procedures for the analysis of complete datasets [4]. In 2011, work [18] developed a multiple imputation program called Amelia II. The model is based on a normality assumption:

$Y \sim N_k(\mu, \Sigma)$, or Y has a multivariate normal distribution with a mean vector μ and a covariance matrix Σ. Regarding multiple imputations, we are concerned with the parameters of the complete data, $\theta = (\mu, \Sigma)$. Under the assumption that the data are MAR, we define Y_{obs} as the observed data and Y_{mis} as the missing data, such that $Y = \{Y_{\text{obs}}, Y_{\text{mis}}\}$. The missing data mechanism is characterized by the conditional distribution of M knowing Y given by $p(M|Y)$.

$$p(M|Y) = p(M|Y_{\text{obs}}). \tag{7}$$

Likelihood $p(Y_{\text{obs}}|\theta)$ is then, written as

$$p(Y_{\text{obs}}, M|\theta) = p(M|Y_{\text{obs}})\,p(Y_{\text{obs}}|\theta). \tag{8}$$

Since we are only interested in the inference of the parameters of the complete data, the likelihood can be written in the following form:

$$L(\theta|Y_{\text{obs}}) \infty p(Y_{\text{obs}}|\theta). \tag{9}$$

Now, using the iterative property of the expectation

$$p(Y_{\text{obs}}|\theta) = \int p(Y|\theta)\mathrm{d}Y_{\text{mis}}, \tag{10}$$

the obtained a posteriori law is as follows:

$$p(\theta|Y_{\text{obs}}) \infty p(Y_{\text{obs}}|\theta) = \int p(Y|\theta)\mathrm{d}Y_{\text{mis}}. \tag{11}$$

2.2. Outlier Management. Outliers can come from two possibilities. Either they come from errors, or they have a history behind them. In principle, outliers should be very rare; otherwise, the experiment/investigation to generate the dataset will be inherently flawed. Defining an outlier is tricky. Outliers may be legitimate because they are part of the long tail of the population. For example, a team working on predicting a financial crisis determines that a financial crisis occurs in one out of every 1000 simulations. Of course, the result is not an outlier that should be discarded. The reason for removing outliers only comes into play when you know the original data for them. For example, if the heart rate data is strangely fast and you know that there is a problem with the medical equipment, you can remove the bad data. Rejecting mysterious outliers is risky for downstream tasks. For example, some regression tasks are sensitive to extreme values. It takes more experiments to decide whether the outliers exist for a reason. In such cases, do not remove or correct outliers in the data preprocessing steps [2].

3. A Proposed Approach to Data Preprocessing

The algorithm in Figure 1 is a maintenance data preprocessing technique that combines missing and redundant data management with data size reduction techniques (compression). The interest of such a work is the combination of its advantages in speed, robustness, and archiving.

The first operation consists of having a table of maintenance data for the power transformers. For this purpose, we have a database from Rio Tinto Alcan of Canada, in which we will exploit the GD (dissolved gas) data of equipment T0001. Due to this database being complete, we will simulate missing data and submit this incomplete table to our algorithm. For the imputation of missing data, we chose the k-nearest neighbor algorithm (KNN). This imputation technique requires the choice of the parameter k by optimization of a criterion. The missing values are imputed, taking into account the class of data to which they belong. Moreover, the notion of distance between observations must be chosen with care. We proceed essentially by learning Euclidian, Mahalanobis, or Minkowski distance metrics to evaluate the similarity between classes. The notion of learning metrics is a recent field in machine learning. The work presented in 2002 in the article [19] is considered a

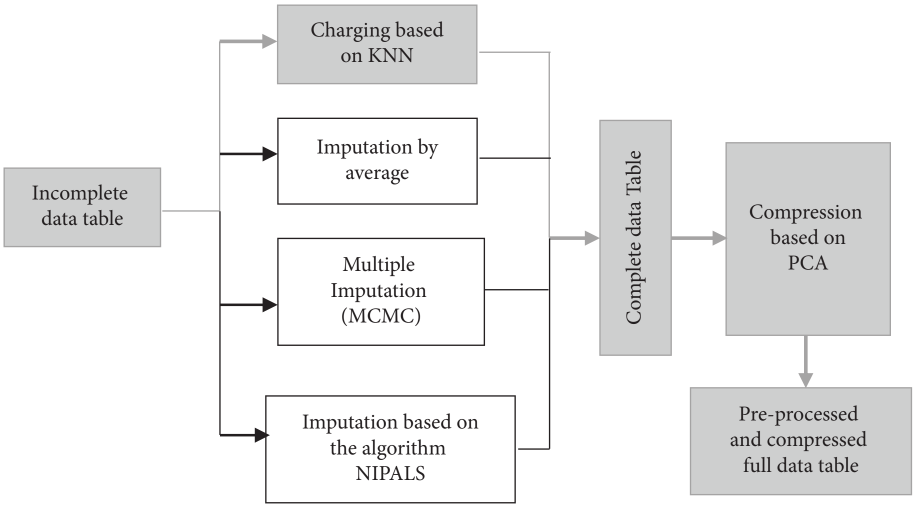
Incomplete data table
Charging based on KNN
Imputation by average
Multiple Imputation (MCMC)
Imputation based on the algorithm NIPALS
Complete data Table
Compression based on PCA
Pre-processed and compressed full data table

Figure 1: Proposed preprocessing algorithm.

pioneer. The goal of metric learning is to answer a recurrent need for comparison functions. Indeed, in many learning algorithms, the notion of metric plays a fundamental role. This is, for example, the case in the k-nearest-neighbor algorithm [20]. Indeed, this classification algorithm proposes associating the majority class of the k closest points with a point of unknown class. This proximity is defined by a metric. The imputed value will thus be the sum of the k-nearest values belonging to the same class.

3.1. Distances of Similarity or Dissimilarity. Euclidean distance. The Euclidean distance is probably the most used distance. For two vectors, x and x', it is noted as follows:

$$d_2(x, x') = x - x'_2 = \sqrt{\sum_i (x_i - x'_i)^2} = \sqrt{(x - x')^T (x - x')}. \tag{12}$$

Minkowski distance. It is defined, for any $p \geq 1$ and for two vectors x and x', by the following equation:

$$d_p(x, x') = x - x'_p = \left(\sum_i |x_i - x'_i|^p \right)^{1/p}. \tag{13}$$

Thus, for $p = 1$, we get the Manhattan distance, for $p = +\infty$, we have the Chebyshev distance, and for $p = 2$, the Euclidean distance is found.

Distance from Mahalanobis is defined as follows:

$$d_M(x, x') = \sqrt{(x - x')^T M (x - x')}. \tag{14}$$

This distance is parameterized. Indeed, depending on the matrix M chosen, the result obtained changes. Thus, if we put $M = \mathrm{I}$, where I is the identity matrix, we find the Euclidean distance.

3.2. Compression Based on PCA. The principal component analysis, beyond being a descriptive technique of data analysis, is an extremely powerful tool of compression and synthesis of information, very useful when one is in the presence of an important quantity of quantitative data to be processed and interpreted. PCA consists of synthesizing the number of observed variables, in other words, it attempts to summarize the information contained in the data table into a reduced set of linear combinations of the initial variables, taking care to minimize the loss of information due to this reduction [21, 22]. These new synthetic variables, called "principal components or factors or macrocharacteristics" have the following properties:

The principal components, as noted $(C^1, C^2, \ldots, C^q)$, are linear combinations of the initial variables

$$\begin{gathered} \left(X^2, X^2, X^p\right)\colon C^j = a_1 X^1 + a_2 X^2 + \cdots a_p X^p, \\ \text{for any } j = 1, q \text{ with } q \leq p. \end{gathered} \tag{15}$$

They are uncorrelated (the linear correlation coefficients of the components taken two by two are zero), which avoids the redundancy of the already summarized information. The first component carries or summarizes more information than the second, which carries more than the third, and so on, so that by limiting ourselves to the first 2 or 3 components, we have a good summary of the information contained in the data [23]. The mathematical tools used are those of linear algebra and matrix calculation. The correlation matrix is diagonalized, and the eigenvectors of this matrix define the new variables sought: these are the principal components. We can show that the principal components thus defined, verify well the sought properties: uncorrelated between them, of decreasing variance, and linear combinations of the starting variables. This last property allows us to construct graphs representing the individuals as well as the variables in the space defined by the components [23]. In the article [14], PCA is used to improve the preprocessing of data.

4. Results and Discussion

This part is presented in two steps: the results and discussions after imputation and after compression.

Table 1: RIO Tinto Alcan of Canada database of dissolved gases from transformer T0001.

Date	H_2	CH_4	C_2H_2	C_2H_4	C_2H_6	CO	CO_2	O_2	N_2
25/01/1985	85,00	40,00	0,00	55,00	52,00	1290,00	11500,00	6380,00	72600,00
23/10/1986	89,00	41,00	0,00	44,00	48,00	1403,00	15092,00	8203,00	78927,00
15/12/1986	90,00	35,00	0,00	37,00	48,00	1310,00	12000,00	10600,00	79600,00
08/06/1987	80,00	40,00	0,00	46,00	47,00	1160,00	10400,00	13700,00	82500,00
06/09/1988	80,00	30,00	2,00	34,00	42,00	1200,00	10600,00	8710,00	70500,00
17/05/1989	85,00	40,00	2,00	39,00	50,00	1190,00	10800,00	3800,00	68500,00
17/10/1989	90,00	40,00	2,00	36,00	48,00	1210,00	11200,00	7130,00	64200,00
30/07/1990	80,00	30,00	2,00	31,00	39,00	1080,00	9680,00	9360,00	73800,00
16/04/1991	90,00	30,00	2,00	22,00	44,00	1150,00	10400,00	7580,00	65300,00
20/04/1992	85,00	37,00	0,00	32,00	46,00	1262,00	11344,00	1092,00	67199,00
14/08/1992	94,00	57,00	0,00	34,00	49,00	1391,00	11694,00	3422,00	97388,00
02/11/1992	86,00	33,00	0,00	34,00	45,00	1317,00	11218,00	1331,00	63320,00
29/04/1993	73,00	29,00	0,30	30,00	43,00	1172,00	10802,00	1574,00	58747,00
02/08/1993	72,00	33,00	0,00	29,00	40,00	1158,00	10906,00	4266,00	62380,00
22/12/1993	74,00	35,00	0,00	27,00	44,00	1278,00	10899,00	3140,00	74873,00
01/05/1995	34,00	5,30	0,00	2,70	2,70	242,00	1181,00	19996,00	64920,00
10/05/1995	34,00	5,30	0,00	2,70	2,70	242,00	1181,00	19996,00	64920,00
29/10/1997	107,00	24,00	0,00	19,00	14,00	1153,00	5626,00	454,00	66308,00
31/05/1999	101,00	26,00	0,30	23,00	16,00	1192,00	5870,00	4709,00	65210,00
10/05/2000	89,00	26,00	0,00	23,00	18,00	1177,00	6476,00	2595,00	77360,00
10/05/2000	89,00	26,00	0,00	23,00	18,00	1177,00	6476,00	2595,00	77360,00
15/05/2001	132,00	44,00	0,00	39,00	33,00	1929,00	12520,00	5494,00	105155,00
13/11/2001	94,00	30,00	0,00	30,00	22,00	1349,00	7976,00	875,00	64907,00
06/05/2002	70,00	25,00	0,00	26,00	19,00	1171,00	7416,00	3512,00	63358,00
14/10/2003	73,00	28,00	0,00	32,00	21,00	1167,00	6446,00	575,00	60260,00
13/05/2004	58,00	23,00	0,00	26,00	20,00	978,00	6999,00	403,00	59725,00
14/05/2005	66,00	25,00	1,10	29,00	21,00	1220,00	8354,00	1067,00	64411,00
10/05/2006	69,00	26,00	0,00	33,00	23,00	1186,00	8099,00	2561,00	67380,00
17/05/2007	68,00	25,00	0,00	35,00	24,00	1211,00	9018,00	6181,00	64829,00
22/05/2008	69,00	24,00	0,00	38,00	23,00	1184,00	8205,00	2816,00	64983,00
17/06/2009	84,00	29,00	0,00	44,00	29,00	1293,00	10276,00	9234,00	77994,00

Table 2: Simulated missing values of transformer T0001.

Dates	H_2	CH_4	C_2H_4	CO	CO_2	N_2
06/09/1988		30,00				
17/10/1989				1210,00		
30/07/1990	80,0					
16/04/1991						65300,00
10/05/1995						64920,00
31/05/1999					5870,00	
10/05/2000	89,0		23,00			

4.1. Imputation of Data. The table of data that we submitted to the test is one of the dissolved gases taken from the transformer equipment (T0001) and presented in the Table 1. Afterward, we will simulate missing values (Table 2), and then submit the data table to the data preprocessing algorithms. Table 3 presents the statistical data before any imputation.

The table of data containing missing values is presented in Table 2 along with several data completion methods in turn. The table 4 shows the results of the different completion techniques.

Table 4 presents the statistical data after imputation by KNN and for a k-neighbor average equal to 5, by the mean and multiple. The results of Table 5 show that the imputation by KNN allows finding in the majority of the exact values.

The decomposition of mineral oil at low temperatures produces relatively large quantities of low molecular-weight gases such as hydrogen (H_2) and methane (CH_4). To better observe the evolution of these two quantities after different imputations presented in Table 5, a weighting of nitrogen (N_2), carbon dioxide (CO_2), and carbon monoxide (CO) is made.

Figure 2 shows the evolution of the exact values of the variables before any simulation of missing data (in blue) and the variations of these data after their imputation created by simulation. It can be seen from this figure that the evolution of the data imputed by KNN (in orange) is closer to that of the exact values. The imputation by KNN is robust because its standard deviation and its mean are less sensitive to the variations of the data. Example: H_2 and N_2 before imputation have, respectively, a mean of 80.034–71206.679 and a standard deviation of 19.320–10819.652. After imputation by KNN, they have 80.323–71286.032 and a standard deviation of 18.734–10399.787. The table of completed data after imputation by KNN is presented in Table 6.

The error in absolute value after the different imputations is given by

$$|\text{Error}| = \frac{y_i - y}{y_i}, \tag{16}$$

where y_i represents the imputed value and y is the exact value.

Table 3: Statistical data before any imputation.

Variable	Obs	Obs. with MV	Obs. without MV	Mean	Standard deviation
H_2	31	2	29	80,034	19,320
CH_4	31	1	30	30,387	10,246
C_2H_4	31	1	30	31,080	10,918
CO	31	1	30	1174,400	298,059
CO_2	31	1	30	9159,467	3057,737
O_2	31	0	31	5591,968	5126,700
N_2	31	2	29	71206,679	10819,652

Table 4: Statistical data after imputations.

		Imputation by KNN ($k = 5$)				Imputation by mean		Imputation multiple	
Variable	Obs	Obs. with MV	Obs. without MV	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
H_2	31	0	31	80,323	18,734	80,034	18,665	79,064	19,062
CH_4	31	0	31	30,374	10,074	30,387	10,074	30,291	10,088
C_2H_4	31	0	31	30,819	10,832	31,080	10,734	30,772	10,870
CO	31	0	31	1174,903	293,062	1174,400	293,049	1184,598	298,499
CO_2	31	0	31	9045,484	3072,597	9159,467	3006,342	9173,076	3007,297
O_2	31	0	31	5591,968	5126,700	5591,968	5126,70	5591,968	5126,700
N_2	31	0	31	71286,032	10399,787	71206,679	10264,423	71291,446	10547,466

*MV: missing value.

Table 5: Evolution of imputed values.

	Exact values	I-KNN	I-mean	I-multiple
H_2	80	80	80,034	68,664
CH_4	30	30	30,387	27,411
C_2H_4	23	23	31,08	21,542
CO	1210	1190	1174	1490,549
CO_2	5870	5626	9159,467	9581,353
N_2	64920	64920	71206,679	82476,217

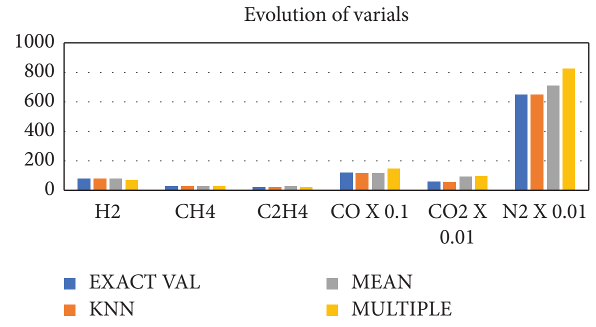

Figure 2: Evolution of imputed variables.

Figure 3 shows us that the error committed by imputing the 6 gases by the k-nearest neighbors (KNN) has an average error percentage of 2%, 17.5% by the average, and 23.65% by multiple imputations.

4.2. Data Compression Based on PCA. One of the applications of principal component analysis is compression. PCA consists of synthesizing the number of observed variables, i.e., summarizing the information contained in the data table, into a reduced set of linear combinations of the initial variables, taking care to minimize the loss of information. The table of data imputed by the k-nearest neighbors is subjected to compression by PCA, and the results are obtained from the Anaconda (Python) and XLStat software.

Table 7 and Figure 4 show that the first eigenvalue λ is 5.054 and represents 56.153% of the variability (inertia I). This means that if we represent the data on two axes, then we will always have 71.299% of the total variability preserved. Each eigenvalue corresponds to a factor. Each factor is a linear combination of the starting variables. In principal component analysis, the problem is to be able to determine the dimension of the optimum representation space. It is a question of preserving all the stable and important characteristics of the data studied while ignoring the unstable and meaningless axes [24].

4.2.1. Number of Axes to Retain. In practice, the only criteria applicable to the choice of the number of axes are empirical, the best known of which is that of Kaiser: in reduced centered data, we retain the principal components corresponding to eigenvalues greater than 1, which means that we are only interested in those components that «contribute» more than the initial variables [25]. We also use the broken sticks test and the kink rule, which consist of detecting the existence of a kink in the eigenvalue diagram. Figure 5 shows the kink thus formed, and Table 8 shows the results of the different tests.

For the Kaiser threshold:

Table 6: Data completed after imputation by KNN.

Dates	H_2	CH_4	C_2H_2	C_2H_4	C_2H_6	CO	CO_2	O_2	N_2
25/01/1985	85,00	40,00	0,00	55,00	52,00	1290,00	11500,00	6380,00	72600,00
23/10/1986	89,00	41,00	0,00	44,00	48,00	1403,00	15092,00	8203,00	78927,00
15/12/1986	90,00	35,00	0,00	37,00	48,00	1310,00	12000,00	10600,00	79600,00
08/06/1987	80,00	40,00	0,00	46,00	47,00	1160,00	10400,00	13700,00	82500,00
06/09/1988	80,00	30,00	2,00	34,00	42,00	1200,00	10600,00	8710,00	70500,00
17/05/1989	85,00	40,00	2,00	39,00	50,00	1190,00	10800,00	3800,00	68500,00
17/10/1989	90,00	40,00	2,00	36,00	48,00	1190,00	11200,00	7130,00	64200,00
30/07/1990	80,00	30,00	2,00	31,00	39,00	1080,00	9680,00	9360,00	73800,00
16/04/1991	90,00	30,00	2,00	22,00	44,00	1150,00	10400,00	7580,00	73800,00
20/04/1992	85,00	37,00	0,00	32,00	46,00	1262,00	11344,00	1092,00	67199,00
14/08/1992	94,00	57,00	0,00	34,00	49,00	1391,00	11694,00	3422,00	97388,00
02/11/1992	86,00	33,00	0,00	34,00	45,00	1317,00	11218,00	1331,00	63320,00
29/04/1993	73,00	29,00	0,300	30,00	43,00	1172,00	10802,00	1574,00	58747,00
02/08/1993	72,00	33,00	0,00	29,00	40,00	1158,00	10906,00	4266,00	62380,00
22/12/1993	74,00	35,00	0,00	27,00	44,00	1278,00	10899,00	3140,00	74873,00
01/05/1995	34,00	5,30	0,00	2,700	2,700	242,00	1181,00	19996,00	64920,00
10/05/1995	34,00	5,30	0,00	2,700	2,700	242,00	1181,00	19996,00	64920,00
29/10/1997	107,00	24,00	0,00	19,00	14,00	1153,00	5626,00	454,00	66308,00
31/05/1999	101,00	26,00	0,300	23,00	16,00	1192,00	5626,00	4709,00	65210,00
10/05/2000	89,00	26,00	0,00	23,00	18,00	1177,00	6476,00	2595,00	77360,00
10/05/2000	89,00	26,00	0,00	23,00	18,00	1177,00	6476,00	2595,00	77360,00
15/05/2001	132,00	44,00	0,00	39,00	33,00	1929,00	12520,00	5494,00	105155,00
13/11/2001	94,00	30,00	0,00	30,00	22,00	1349,00	7976,00	875,00	64907,00
06/05/2002	70,00	25,00	0,00	26,00	19,00	1171,00	7416,00	3512,00	63358,00
14/10/2003	73,00	28,00	0,00	32,00	21,00	1167,00	6446,00	575,00	60260,00
13/05/2004	58,00	23,00	0,00	26,00	20,00	978,00	6999,00	403,00	59725,00
14/05/2005	66,00	25,00	1,10	29,00	21,00	1220,00	8354,00	1067,00	64411,00
10/05/2006	69,00	26,00	0,00	33,00	23,00	1186,00	8099,00	2561,00	67380,00
17/05/2007	68,00	25,00	0,00	35,00	24,00	1211,00	9018,00	6181,00	64829,00
22/05/2008	69,00	24,00	0,00	38,00	23,00	1184,00	8205,00	2816,00	64983,00
17/06/2009	84,00	29,00	0,00	44,00	29,00	1293,00	10276,00	9234,00	77994,00

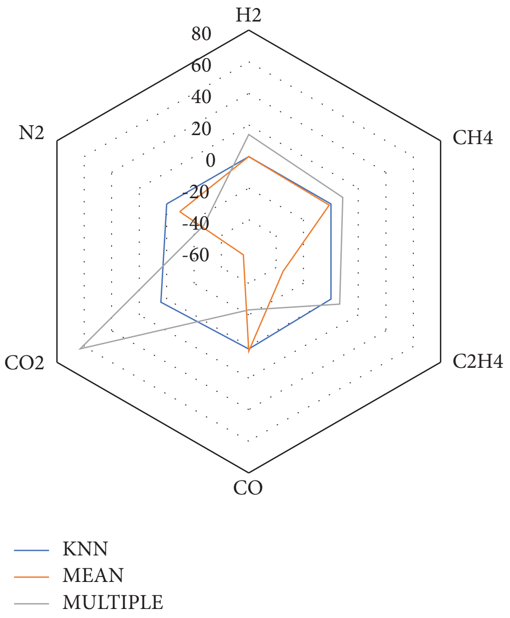

Figure 3: Change in error due to imputation.

$$\lambda > 1 + 2\sqrt{\frac{p-1}{n-1}} \longrightarrow \lambda > 1 + 2\sqrt{\frac{9-1}{31-1}} = 1,188. \quad (17)$$

For broken sticks: $b_k = \sum_{m=k}^{p} 1/m$ The component is validated if $\lambda_k > b_k$.

The calculation of the threshold of the eigenvalues is given by the expression (17) where λ is the eigenvalue, p is the number of variables, and n is the number of observations.

The use of the threshold (Kaiser-Saporta), and elbow (Cattell) rules limits the number of axes to two, while the broken sticks test (Frontier 1976) limits the number of axes to be retained to one. All of these approaches are consistent in that only one factor or axis appears to be sufficient in this study. For the sake of future interpretation and rotation of the axes, we have opted to maintain two axes as recommended by the elbow and Kaiser methods.

4.2.2. Observations. The study of the observations consists of examining their coordinates and especially their graphic representations. Figures 6 and 7 show the evolution of the two components retained according to the type of pre-treatment applied.

The reference variable here is SI (without imputation), which represents the component retained at the end of the principal component analysis without having first carried out an imputation.

To take just one example, on 15/05/2001, the components F1 without imputation (F1 SI), with imputation by the

Table 7: Eigenvalues and inertia.

	F1	F2	F3	F4	F5	F6	F7	F8	F9
λ	5,054	1,363	1,141	0,799	0,295	0,173	0,126	0,033	0,014
Inertia (%)	56,153	15,146	12,683	8,878	3,281	1,927	1,405	0,372	0,156
% Cumule	56,153	71,298	83,981	92,859	96,140	98,067	99,472	99,844	100,000

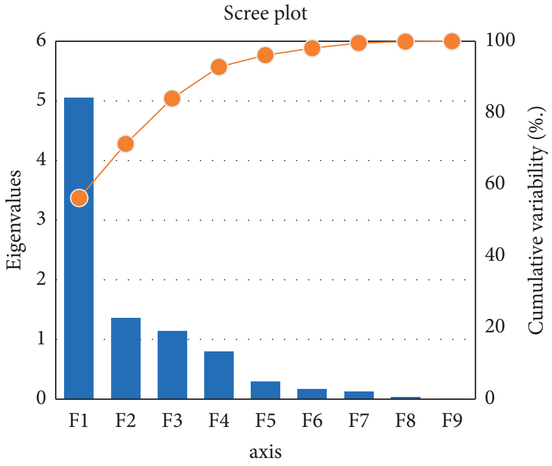

Figure 4: Eigenvalue distribution on the axes.

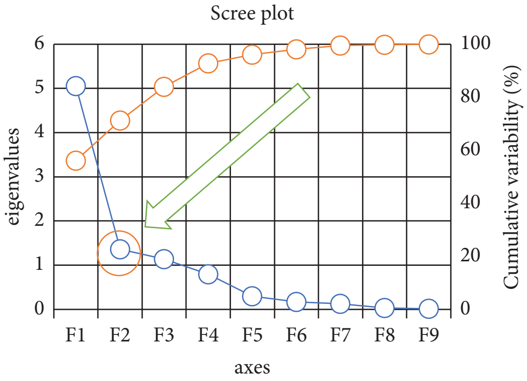

Figure 5: Observation of the middle elbow.

Table 8: Boundary values of the broken sticks test and the Kaiser threshold.

Axes	F1	F2	F3	F4	F5	F6	F7	F8	F9
Eigenvalues	5,054	1,363	1,141	0,799	0,295	0,173	0,126	0,033	0,014
b_k	2,828	1,828	1,328	0,995	0,745	0,545	0,379	0,236	0,111
Kaiser threshold	1,188	1,188	1,188	1,188	1,188	1,188	1,188	1,188	1,188

k-nearest neighbors + principal component analysis (F1 Iknnacp), with imputation by the mean + principal component analysis (F1 Imoyacp), and with multiple imputation + principal component analysis (F1 Imacp) have the values 4.142, 4.213, 4.167, and 3.917, respectively. The results show that, for component 1, the preprocessed data are faithful to the reference with a few precisions.

For the same example, on 15/05/2001, the F2 components without imputation (F2 SI), with imputation by the k-nearest neighbors + principal component analysis (F2 Iknnacp), with imputation by the mean + principal component analysis (F2 Imoyacp), and with multiple imputation + - principal component analysis (F2 IMacp) have, respectively, values −1.723, −1.822, −1.257, and 1.406. These results show that, for component 2, the preprocessed data with KNN + PCA imputation are more faithful to the reference data (F2 SI) with a few precisions. The combination of these two elements (KNN + ACP) produces the best accuracy for

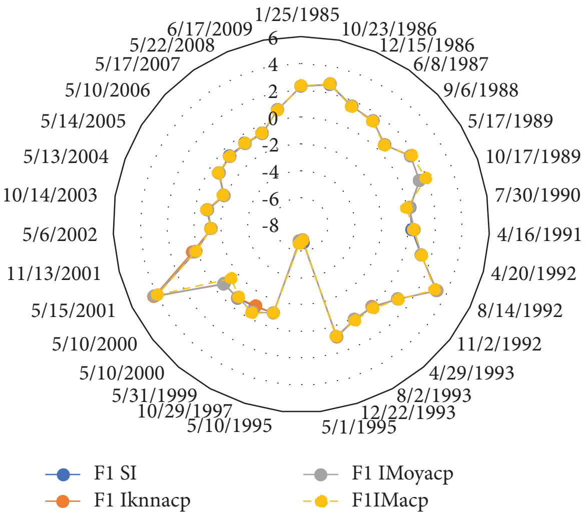

Figure 6: Evolution of component 1.

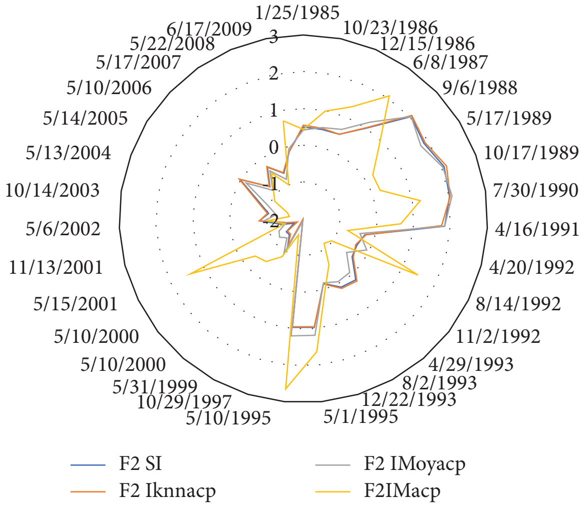

Figure 7: Evolution of component 2 [26].

the T0001 transformer DGA dataset of different sizes and different percentages of missing values.

5. Conclusion

In this paper, an approach for preprocessing power transformer maintenance data is proposed. This approach uses both KNN completion with the Euclidean metric to impute quantitative data and a PCA whose function here is the management of redundant values and especially the compression of a large amount of data. This preprocessing approach, i.e., imputation by KNN completion and PCA, was rigorously compared to two other approaches, such as imputation by the mean + PCA and multiple imputation + PCA. It is clear that for 6 missing values, the k-nearest neighbor imputation and for $k = 5$, the error committed is around 2%, while the multiple and the mean imputation have 23.65% and 17.5% errors, respectively. Similarly, to observe low molecular weight gases produced at low temperatures such as hydrogen (H_2) or methane (CH_4), the weighting of nitrogen (N_2), carbon dioxide (CO_2), and carbon monoxide (CO) is performed (Figure 2). The KNN + ACP preprocessing is robust because the standard deviation and mean obtained by KNN completion are less sensitive to data variations and present results close to reality on the one hand, and on the other hand, the amount of starting data is considerably reduced while keeping the originality of the starting data at the maximum. For 31 observations and 9 variables, the Kaiser threshold is 1.188, which allowed us, in this case, to retain 2 components based on the kink principle and the eigenvalue threshold. Experiments conducted using this proposed combination show significant performance, especially when the percentage of variables and missing values in the dataset would be high.

Appendix

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

The authors thank American Journal Experts for their technical support in formatting this work. Also, the Rio Tinto Alcan in Canada for providing data through Pr Issouf FOFANA.

References

[1] S. Curtet, *Modules de prétraitement de données dans le cadre du Data Mining*, Haute Ecole d'Ingenierie et de Gestion du Canton de Vaud, Bains, Switzerland, 2006.

[2] R. Li, *Essential Statistics for Non-STEM Data Analysts*, 392 pages, Packt Publishing, Birmingham, UK, 2020, https://www.packtpub.com/product/essential-statistics-for-non-stem-data-analysts/9781838984847.

[3] R. J. Mislevy, R. J. A. Little, and D. B. Rubin, "Statistical analysis with missing data," *Journal of Educational Statistics*, vol. 16, no. 2, pp. 150–155, 1991.

[4] G. Cottrell, M. Cot, and J. Y. Mary, "[Multiple imputation of missing at random data: general points and presentation of a Monte-Carlo method]," *Revue d'Epidemiologie et de Sante Publique*, vol. 57, no. 5, pp. 361–372, 2009.

[5] S. A. Imtiaz and S. L. Shah, "Treatment of missing values in process data analysis," *Canadian Journal of Chemical Engineering*, vol. 86, no. 5, pp. 838–858, 2008.

[6] R. Houari, A. Bounceur, and T. Kechadi, "Nouvelle approche de prétraitement pour les fouilles de données numériques," in *Proceedings of the 2ième édition de la conférence nationale de l'informatique destinée aux étudiants de graduation et de post-graduation JEESI'12*, Oued-Smar, Algeria, April 2012.

[7] Z. B. Sahri and R. B. Yusof, "Support vector machine-based fault diagnosis of power transformer using k nearest-neighbor

imputed DGA dataset," *Journal of Computer and Communications*, vol. 02, no. 09, pp. 22–31, 2014.

[8] J. W. Grzymala-Busse, "On the unknown attribute values in learning from examples," in *Proceedings of the International Symposium on Methodologies for Intelligent Systems*, pp. 368–377, Charlotte, NC, USA, 1991.

[9] J. Gou, H. Ma, W. Ou, S. Zeng, Y. Rao, and H. Yang, "A generalized mean distance-based k-nearest neighbor classifier," *Expert Systems with Applications*, janv, vol. 115, pp. 356–372, 2019.

[10] J. Gou, L. Sun, L. Du et al., "A representation coefficient-based k-nearest centroid neighbor classifier," *Expert Systems with Applications*, vol. 194, Article ID 116529, 2022.

[11] D. M. P. Murti, U. Pujianto, A. P. Wibawa, and M. I. Akbar, "K-nearest neighbor (K-NN) based missing data imputation," in *Proceedings of the 2019 5th International Conference on Science in Information Technology (ICSITech)*, pp. 83–88, Yogyakarta, Indonesia, October 2019.

[12] F. Yang, J. Du, J. Lang et al., "Missing value estimation methods Research for arrhythmia classification using the modified kernel difference-weighted KNN algorithms," *BioMed Research International*, vol. 2020, Article ID 7141725, 9 pages, 2020.

[13] J. Zhu, Z. Ge, Z. Song, and F. Gao, "Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data," *Annual Reviews in Control*, vol. 46, pp. 107–133, 2018.

[14] A. J. C. Trappey, C. V. Trappey, L. Ma, and J. C. M. Chang, "Intelligent engineering asset management system for power transformer maintenance decision supports under various operating conditions," *Computers & Industrial Engineering*, vol. 84, pp. 3–11, 2015.

[15] J. W. Grzymala-Busse, W. J. Grzymala-Busse, and L. K. Goodwin, "Coping with missing attribute values based on closest fit in preterm birth data: a rough set approach," *Computational Intelligence*, vol. 17, no. 3, pp. 425–434, 2001.

[16] C. Preda and G. Saporta, "The nipals algorithm for missing functional data," *Romanian Journal of Pure and Applied Mathematics*, vol. 55, no. 4, pp. 315–326, 2010.

[17] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, Inc, Hoboken, NJ, USA, 1987.

[18] J. Honaker, G. King, and M. Blackwell, "Amelia II: a program for missing data," *Journal of Statistical Software*, vol. 45, no. 7, pp. 1–47, 2011.

[19] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell, "Distance metric learning, with application to clustering with side-information," in *Proceedings of the 15th International Conference on Neural Information Processing Systems*, pp. 521–528, MIT Press, Berkeley, CA, USA, January 2002.

[20] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[21] H. Zaki, M. Benlyas, F. Y. Zegzouti, and M. Bouachrine, "Méthodologie générale d'une étude ACP: généralités, concepts et exemples," *Revue Interdisciplinaire*, vol. 1, no. 1, pp. 1–8, 2016.

[22] E. Samuel, "Stratégie d'évaluation de l'etat des transformateurs: esquisse de solutions pour la gestion intégrée des transformateurs vieillissants," Universite de Lyon, Lyon, France, These de Doctorat, 2018.

[23] R. Abdesselam, *Analyse des Données Polycopié 1: Méthodes Factorielles*, Université Lumière Lyon, Lyon, France, 2013-2014.

[24] C. Baril, *Étude de la Stabilité de l'ACP par la Méthode du Bootstrap*, CIRAD-Forêt, Montpellier, 1993.

[25] G. Saporta and N. Niang, "Analyse en composantes principales," *Hermes*, pp. 19–42, 2003.

[26] Addinsoft, "XLSTAT statistical and data analysis solution," 2022, http://www.xlstat.com/fr.