Electric Power Systems Research 206 (2022) 107778

Contents lists available at ScienceDirect

Electric Power Systems Research

journal homepage: www.elsevier.com/locate/epsr

Combining machine learning techniques with Kappa–Kendall indexes for robust hard-cluster assessment in substation pattern recognition

Fabricio Alves de Almeida [a,*], Estevão Luiz Romão [b], Guilherme Ferreira Gomes [c], José Henrique de Freitas Gomes [b], Anderson Paulo de Paiva [b], Jacques Miranda Filho [d], Pedro Paulo Balestrassi [a,b]

[a] Institute of Electrical Systems and Energy, Federal University of Itajubá, Brazil
[b] Institute of Industrial Engineering and Management, Federal University of Itajubá, Brazil
[c] Mechanical Engineering Institute, Federal University of Itajubá, Brazil
[d] IFES Federal Institute of Espírito Santo, Vitória, Brazil

ARTICLE INFO

Keywords:
Machine learning
Kappa–Kendall
Rotated factor analysis
Voltage sag
Power quality substation
Pattern recognition

ABSTRACT

This study proposes a method that combines different machine learning and lean six sigma techniques to calibrate cluster analysis through linkage methods. The power quality indexes of substations in Brazil, which are of interest to regulatory agencies, are used. The method uses the random forest mixed with rotated factor analysis to filter, minimize, and improve the interpretation of latent information. Variability scenarios are created using the Monte Carlo simulation to assess the stability of the cluster analysis using the design of experiments and the Kappa–Kendall indexes. The Ward method shows a better consistency in all scenarios and a better discriminatory power among the clusters. The optimal result is used to predict different scenarios with high levels of variability (5, 10, and 15%) by comparing the behaviors of different supervised machine learning techniques for classification. The results show that the k-nearest neighbors, support vector classifier, and logistic regression approaches can accurately predict, even in scenarios with high variability in the dataset.

1. Introduction

Advanced statistical techniques have been widely investigated in power quality (PQ) studies [1], promoting the development of new technologies and supporting decision making [2]. The advent of computation has provided the creation and improvement of mathematical/machine learning approaches in strategic sectors, such as energy generation and distribution, thus, impacting the industrial sector significantly [1]. Voltage sag is a significant characteristic of PQ distribution [3] and is a variable caused by the short-duration voltage variation. Additionally, this variable economically impacts the production processes because industrial processes have sensitive loads. Several existing studies investigated the voltage sag phenomenon [4–8].

As an object of study, regulatory agencies consider voltage sag and other characteristics to classify PQ substations based on the number of voltage sag events. Some studies have used exploratory techniques combined with cluster analysis (CA) to group substations based on the PQ [1,9]. These studies were based on the regulatory agencies' need to assess and control the PQ. In these studies, specific techniques were applied in view of the multivariate characteristics of the data.

Multivariate techniques were used to analyze a dataset with multiple correlated characteristics [10]. Factor analysis (FA), an exploratory strategy, is one the most robust techniques. This technique transforms several variables into a few common factors, thereby reducing the data dimensionality. This technique also allows rotation of factor loads, simplifying the load matrix and creating an easy-to-interpret structure [11]. Another multivariate strategy widely used in the electricity sector is the CA. This technique creates clusters based on the level of similarity using techniques such as the hierarchical and non-hierarchical linkage methods. Both approaches can recognize patterns of observations based on the available characteristics. These strategies were applied in several studies on PQ [12–16], highlighting their importance.

The most used and significant linkage methods are: k-means, Ward, single, average, complete, median, centroid, and McQuitty. However, many authors arbitrarily employ linkage methods obtaining unsatisfactory results because linkage methods, which are sensitive to outliers,

* Corresponding author.
E-mail address: fabricio-almeida@unifei.edu.br (F.A. Almeida).

https://doi.org/10.1016/j.epsr.2022.107778
Received 29 August 2020; Received in revised form 23 November 2021; Accepted 4 January 2022

may present inversions in the ideal formation of clusters [17]. According to Pinel [15], the ideal method selection depends on the characteristics of the data and its application. Therefore, the best linkage method may vary depending on the dataset structure and usage.

Thus, this study proposes a method that combines different machine learning techniques (MLTs), indicators of variability, and agreement analysis (Kappa–Kendall) to assist decision making in selecting the robust linkage method to be applied based on the dataset. The investigation in this study is conducted using a real dataset, which includes several PQ characteristics impacting voltage sag studies. Initially, this study uses replicas with small to moderate perturbations (3%), creating different scenarios through the Monte Carlo simulations. Using replicas is justified by the study conducted by Johnson and Wichern [17], in which the authors claimed that using small perturbations in the set is a good measure while evaluating the behavior of cluster methods. As a strategy to treat the data, MLTs, such as FA, CA (considering eight different linkage methods), and random forest regression (RFR), are used to initially analyze the most significant characteristics. Additionally, other MLTs are used to predict results confirming the stability of the best method. In this study, design of experiments (DOE) is applied to create the experimental matrix, which is evaluated using the Kappa and Kendall indicators. This technique allows selecting methods with better consistency and stability based on international criteria [18]. This approach is widely used in the lean six sigma methodology. Finally, the proposed approach presents a strategy to identify the best linkage method and, consequently, the ideal discrimination between different observations for a given set of data (in this context, in the formation of substation hard clusters based on voltage sag studies).

2. Theoretical background

2.1. Machine learning techniques

2.1.1. Random forest

Random forest is based on the combination of many trees, where the final decision is obtained considering the decisions of several individual trees [19]. This method can be applied to regression problems (using RFR) or classification problems (using a random forest classifier (RFC)).

As the number of trees increases, the limiting value of the generalization error is obtained; however, there is no overfitting, which is explained by the law of large numbers [19]. However, a small number of estimators makes the model unable to perceive the relationships between the input variables and investigated response.

2.1.2. Factor analysis

FA is an exploratory multivariate technique that minimizes the repetition of information in observed variables, using of a smaller number of latent variables [17]. The factor model can be expressed as Eq. (1), where **Y** is the vector of the observable random variables, **μ** is the vector of population means, **L** is the factor loading matrix, **F** is the random vector of latent variables, and **ε** is the error vector or specific factors.

$$\mathbf{Y}_{(p\times 1)} - \boldsymbol{\mu}_{(p\times 1)} = \mathbf{L}_{(p\times m)}\mathbf{F}_{(m\times 1)} + \boldsymbol{\varepsilon}_{(p\times 1)} \tag{1}$$

Using this technique depends on the adequacy of the data to be investigated [20]. Thus, strategies such as the Bartlett sphericity test (BST) and Kaiser–Meyer–Olkin (KMO) measure of the sampling adequacy index should be applied. To use the BST, the data must follow a normal multivariate distribution, i.e., the null hypothesis that the correlation matrix is equal to the identity matrix is not rejected $\left(\chi^2 > \chi^2_{\alpha;[p(p-1)/2]}\right)$ [21]. Meanwhile, the KMO index analyzes the proportion of common variance for the original variables in Eq. (2), where *rij* and *qij* represent the sample correlation matrices **R** and anti-image **Q**, respectively. This indicator returns values between 0 and 1, where a KMO index ≥ 0.5 is desirable [20,22].

$$KMO = \frac{\sum_{i\neq j} r_{ij}^2}{\sum_{i\neq j} r_{ij}^2 + \sum_{i\neq j} q_{ij}^2} \tag{2}$$

For adequate data, the parameters of the FA can be estimated. The estimation methods include the principal components (PC) and maximum likelihood (ML) methods. The PC approach does not require a specific probability distribution (such as ML), in addition to producing factor scores with independent vectors (null correlation) through lexicographical optimization. This approach estimates the factor loadings and specific variances from the spectral decomposition of the sample correlation matrix **R**. In addition to the estimation method, the number of factors to be extracted is defined, which uses the Kaiser criterion that considers the amount related to eigenvalues ≥ 1 and a percentage of total explanation of the variance of at least 80% [17,23].

Before extracting the factor scores, the FA technique rotates the axes to approximate the factors of the factor loadings. This creates more simplified and easily-explained models based on the principle of parsimony established by Thurstone [24] for FA. The *varimax* method is one of the most used approaches with best performances as it uses the **T** orthogonal matrix to maximize the value of Eq. (3). The relationship between the i^{th} communality and factor loading under rotation is represented by $\tilde{l}_{ij}^{\circ} = l_{ij}^{\circ}/\sqrt{h_i^2}$.

$$varimax = \frac{1}{p}\sum_{j=1}^{m}\left[\sum_{i=1}^{p}\tilde{l}_{ij}^{4} - \left(\sum_{i=1}^{p}\tilde{l}_{ij}^{2}\right)^2 \Big/ p\right] \tag{3}$$

After simplification, the rotation of the factor scores is obtained from Eq. (4), where **F** represents the matrix of the latent variables, **Z** represents the standardized matrix of the response variables, and **L°** is the matrix of rotated factor loads obtained through the *varimax* method.

$$\mathbf{F} = \mathbf{Z}\left[\mathbf{L}^{\circ}\left(\mathbf{L}^{\circ T}\mathbf{L}^{\circ}\right)^{-1}\right] \tag{4}$$

2.1.3. Cluster analysis

CA uses a strategy that divides and classifies elements of a given sample (or population) into groups of similar elements [25]. Thus, homogeneous information tends to be classified in the same group and, in a complementary way, the formed groupings are heterogeneous. According to Mingoti [25], this technique has wide applicability and is used in several fields, such as psychology, economics, ecology, and geochemistry. The CA is also remarkable in research and data mining applications, relating this strategy to other computational tools for searching and identifying patterns in datasets [1,15].

In addition to distance measurements and techniques, the CA uses link metrics to calculate and group the elements based on mathematical equations. These methods are sensitive to outlier data [17] and can be divided into two categories: hierarchical and non-hierarchical [26]. The hierarchical methods are often used in exploratory analyses by performing groupings based on similarity characteristics between elements. Table 1 presents the main hierarchical linkage methods (examples of different groups *A, B*, and *C,* and their elements X_i, X_k and X_m, respectively; n_A and n_B represent the number of elements for groups *A* and *B*, respectively, whereas $\overline{X}$ indicates the mean of the groups) and their descriptions and mathematical modeling.

In contrast, the non-hierarchical methods directly identify the best partition of the set based on the number of clusters. This technique has two requirements: internal similarity between the elements and separation of clusters [25]. The most widespread strategy for this method is the k-means method. This technique performs an iterative process, assigning a closest average to each item in a cluster, as follows [17]: First, the elements are partitioned into k clusters, defining the coordinates of the centroids. From the Euclidean distance, the element to be assigned in each cluster is defined, based on the nearest centroid. Thus, for each element change in a cluster, the group centroid is

Table
Hierarchical linkage methods.

Linkage method	Equation	Description
Average	$d(A,B) = \sum_{l\in A}\sum_{k\in B}(1/n_A n_B)\, d(X_l,X_k)$	Consider the distance between two clusters as the average distance between all pairs of objects [21]
Centroid	$d(A,B) = (\overline{X}_l - \overline{X}_k)^T(\overline{X}_l - \overline{X}_k)$	Sets distance based on mean vectors (known as centroids)
Complete	$d(A,B) = max\{d(X_l,X_k)\}$	Known as the "farthest neighbor method", it considers the elements with the least similarity to create the groups
McQuitty	$d(A,B-C) = \frac{d(\overline{X}_l,\overline{X}_k) + d(\overline{X}_k,\overline{X}_m)}{2}$	Similar to the Average method, this strategy uses the weighted average to define the clusters
Median	$d(A,B-C) = \frac{d(X_l,X_m) + d(X_k,X_m)}{2} - \frac{d(X_l,X_k)}{4}$	It uses the distance matrix to calculate the median distance of the elements
Single	$d(A,B) = min\{d(X_l,X_k)\}$	It seeks to define the similarity by the two most similar elements
Ward	$d(A,B) = [n_A n_B/n_A + n_B](\overline{X}_{l.} - \overline{X}_{k.})^T(\overline{X}_{l.} - \overline{X}_{k.})$	It joins two clusters to minimize the loss of information, improving the error sum-of-squares (ESS) criterion

recalculated, repeating this step until it is not necessary to perform other reassignments. Fig. 1 illustrates the iterative process of the cluster formation.

2.1.4. Other classification techniques

2.1.4.1. Artificial neural networks. Artificial neural networks (ANNs) are formed by neurons (processing units) grouped into input, hidden, and output layers [27]. The neurons from the same layer have no connection but are connected to all the neurons in the adjacent layers through synaptic weights.

The layers of a network can be arranged in a feedback or feedforward manner; in the former, a neuron can be visited more than once. Typically, a backpropagation algorithm is used to adjust weights [28].

2.1.4.2. Support vector machine classification. The support vector machine classifier (SVC) algorithm separates initially non-separable data linearly using kernel functions [29]. The most important algorithm parameters are the kernel type, regularization (C), and gamma value.

The hyperplane is built using a kernel function, which can be a linear, sigmoidal, polynomial, or radial basis function [30]. The value of C penalizes each misclassification made by the model, while the gamma determines the influence of the training observations on the decision boundary.

2.1.4.3. k-nearest neighbors' algorithm. k-nearest neighbors (kNN) is a non-parametric method based on instances [31]. This strategy stores the training set, which is used for each new forecast. Thus, the predominant class among the k nearest neighbors is identified and assigned to the new data.

While determining the nearest neighbors, distances such as Euclidean, Minkowski, and Manhattan are used [32]. It is essential to standardize the predictor variables such that the order of magnitude does not erroneously influence the decisions for the closest neighbors.

2.1.4.4. Multinomial logistic regression. The logistic regression technique is similar to linear regression; however, in the latter, continuous variables are considered, and in the former, the response variable represents a class. Thus, it is necessary to estimate the value of $P[c \mid \mathbf{X}]$, i.e., the probability that vector $\mathbf{X}$ belongs to class c. When there are more than two categories to which an observation belongs with no hierarchical order, multinomial logistic regression (MLR) can be used, as exemplified by Hosmer *et al.* [33].

2.1.3. Kappa–Kendall indexes

The Kappa and Kendall indexes are used to analyze the variability between discrete variables to determine whether appraisers have a good performance. Thus, the Kappa index, calculated using Eq. (5) [34], indicates the ratio between the following values: the proportion in which the appraisers agree and maximum proportion in which they can agree.

$$K = (P_o - P/1 - P) = \frac{\left[1\Big/N_k n_k(n_k-1)\left(\sum_{i=1}^{k}\sum_{j=1}^{k}x_{ij} - N_k n_k\right)\right] - \sum_{j=1}^{l}p_j^2}{\left(1 - \sum_{j=1}^{l}p_j^2\right)} \tag{5}$$

In Eq. (5), the numbers of appraised items and appraisers are N_k and n_k, respectively, where k is the number of categories in the adopted scale; P_o and P_e represent the mean proportions of observed and expected agreements, respectively; X_{ij} represents the number of appraisers that classify a certain item i to class j.

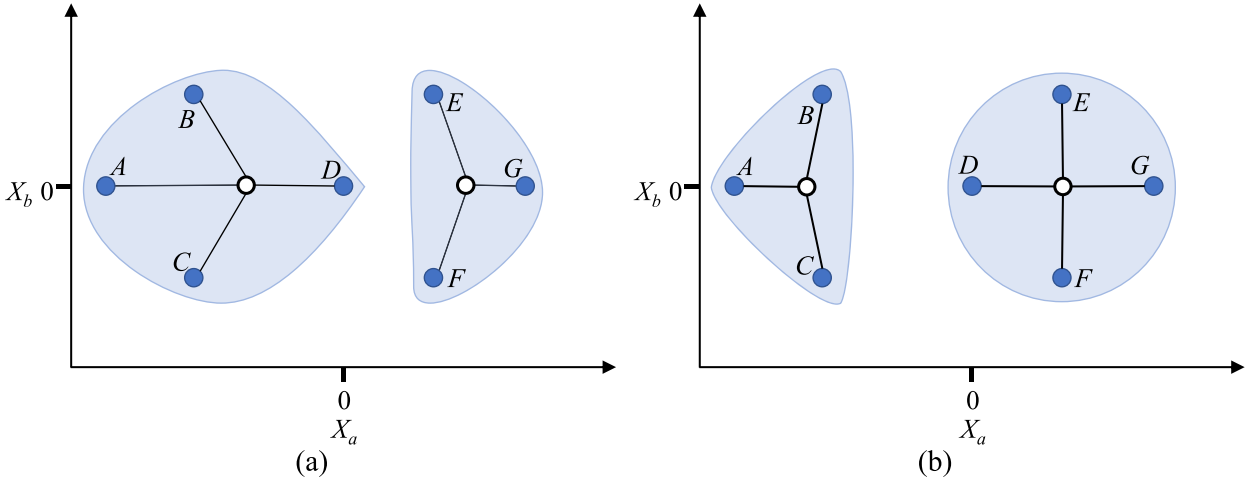

Fig. 1. Behavior of the k-means cluster: (a) initial situation; (b) final situation.

The Kendall's coefficient measures the level of agreement between and within the appraisers. This can be obtained using Eq. (6), where R^2 is the sum of squares for classification sums R_i, n is the number of items, t_k represents the number of ranks tied in each group of ties, and p represents the number of appraisers. It is noteworthy that m indicates the total number of groups of ties.

$$W = 12\sum_{i=1}^{n} R_i^2 - 3p^2 n(n+1)^2 \Big/ p^2\left(n^3 - n\right) - p\left(\sum_{k=1}^{m}\left(t_k^3 - t_k\right)\right) \tag{6}$$

According to AIAG criteria [18], the Kappa index values (K) range between −1 and 1, where a value of 1 indicates a perfect agreement, and 0 indicates that the agreement is the same as that expected by chance. A value of −1, however, indicates an agreement lower than that expected by chance. Furthermore, according to Hinkle [35], the Kendall index (W) varies between 0 and 1, where 0 and 1 indicate no association and perfect association, respectively.

3. PQ characteristics: Brazilian case study

A real example of network modeling and fault simulation is investigated in this study. Seventeen substations, located in the state of Espírito Santo in southeastern Brazil, are considered. Fig. 2 depicts the geographical locations of these substations serving 90% of the municipalities in that state, covering an area of approximately 41,214 km^2. The data used (from [9]) were collected over 30 months in a research project developed by the Federal University of Itajubá and EDP ES Distribution Utility.

Bare conductors in overhead lines and feeders facilitate the occurrence of short circuits. During the simulations, the distribution line length, voltage-rated feeders, and fault statistics, i.e., the number of faults within 100 km per year, are considered.

Based on the database, 32 characteristics related to the design and PQ of each substation are considered, and the total number of sag events per year (TNE) is considered as the main variable related to the PQ. The values associated with the TNE variable and number of events are obtained through simulations and monitoring, respectively. Measurements related to the PQ are acquired considering a period exceeding one year to cover different seasonal events influencing the distribution network performance. The data are obtained using 30 PQ meters of model SEL 734 from the Schweitzer Engineering Laboratories.

Events collected from the secondary power transformers are recorded using the PQ monitors. Considering the 13.8 kV distribution lines, the medium-voltage failure rate is obtained by summing the long-and short-duration failure rates from the statistics. It is noteworthy that the existence of the three-phase type can also be identified, even though it presents a lower incidence, compared with a single-phase type.

Table A.1 in Appendix A lists the nomenclature of all the variables considered in this study. Further details regarding data collection and substation buses can be found in [9].

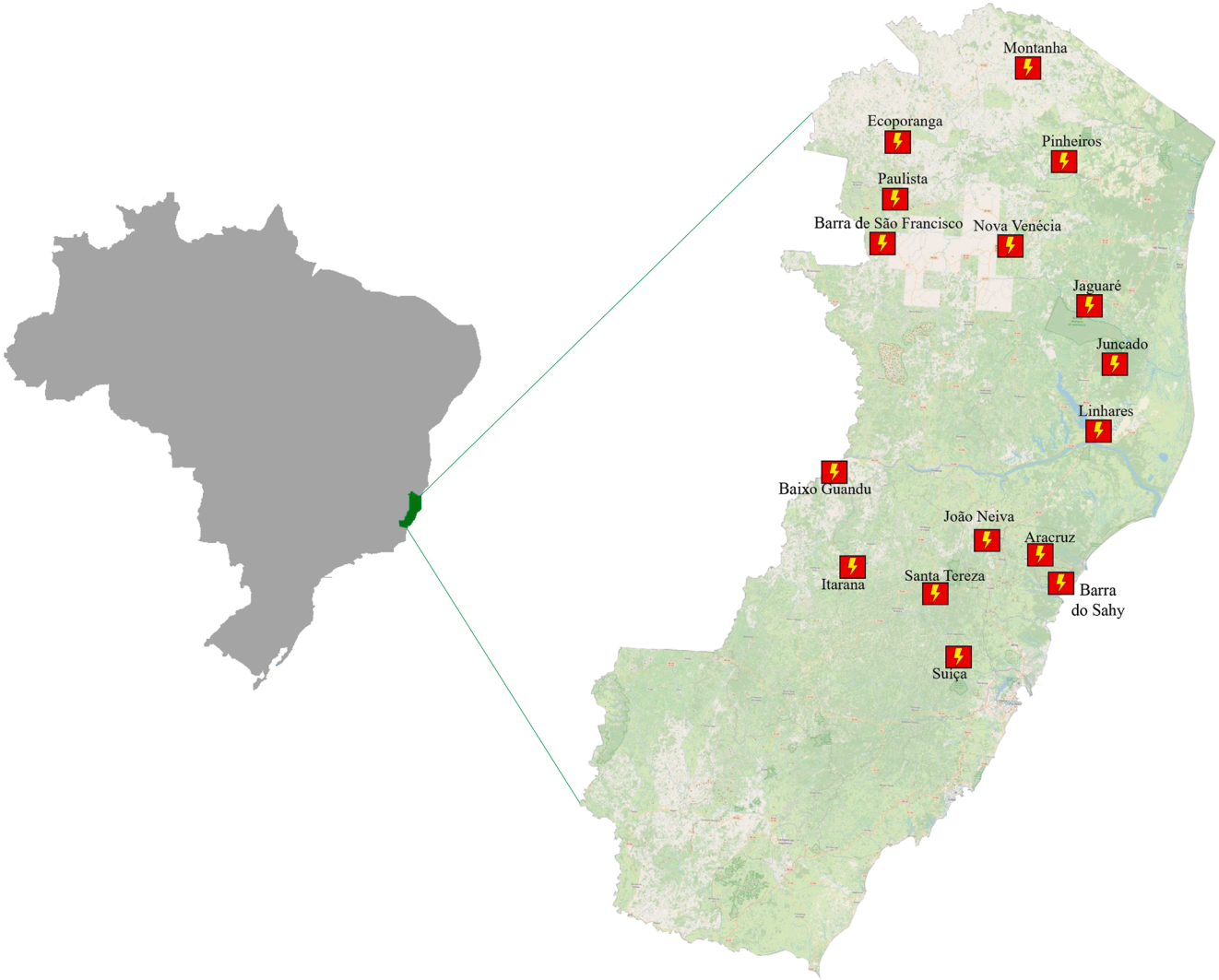

Fig. 2. Geographic location of substations (State of Espírito Santo – Brazil).

4. Robust hard-clustering assessment through machine learning approaches

Technological advent has favored the application and improvement of MLT for various purposes, in which the CA is widely used in several sectors, mainly in the energy sector. For CA calibration, linkage methods are characterized as the main parameters directly impacting the classification results. Many studies use one or several linkage methods in different applications but do not consider sufficient criteria to support the selection of these methods. Furthermore, Johnson and Wichern [17] inferred that linkage techniques can present cluster inversions in the presence of outliers. Thus, it is necessary to verify the behavior of the data under different variability conditions. This study proposes a methodology that combines different MLTs to assist in analysis and decision-making while selecting a robust linkage method, considering latent variables and other characteristics of the dataset. For this study, the dataset that includes the substation PQ characteristics are investigated (as described in Section 3 and [9]). The software *Minitab®, R Studio®* and *algorithms* in *Python* language are used for the analyses. The pseudocode of the proposed method is presented in Table A.2 of the appendix.

Step 1: From the original data, the structure of the set must initially be verified. Knowing the characteristics of the FA exploratory strategy, the number of observations must be greater than the number of variables. Thus, if necessary, an initial filtering must be performed using the RFR strategy. This technique identifies the most important variables for the set based on the response of interest (in this case, the TNE). Filtering promotes prior creation of a lean dataset and reduction of the data dimensionality.

Step 2: Identifying the most important variables, scenarios with small perturbations should be created, as suggested by Johnson and Wichern [17]. These scenarios should have their variations applied to a small standard deviation in the range of 3%. The Monte Carlo simulation creates the replicas. Each of the replicas represents the mean value of 10, 000 simulations created using different disturbance scenarios. For this case study, real data from the Brazilian power distribution company is used to develop these replicas.

Step 3: The next step involves the application of the FA technique subject to adequacy tests, such as the BST and KMO index. If the data structure is adequate, the FA can be applied. Otherwise, another exploratory strategy must be used.

Step 4: As the data are adequate for the FA application, the number of factors to be used in the study is determined using the Kaiser criterion. Additionally, factors presenting an explanation percentage of minimum 80% of the data can be considered. Subsequently, the FA must be applied to extract the factor scores. The rotation of factorial loads is fixed in the *varimax* method to meet the principle of parsimony [24,36]. This process is repeated for each replica.

Step 5: Considering the rotated factor scores, it is necessary to apply the CA (using different linkage methods) to assess the formation of memberships. This planning can be considered as a specific type of DOE, which is a multilevel full factorial design with two factors and four replicas: an 8-level factor (linkage methods), another factor with 17 levels (number of observations/substations), and the replicas representing the different scenarios with 3% disturbances.

To apply the CA, the number of clusters to be considered must be determined initially, which can vary for each objective. This can be achieved through the categorization using the Sturges rule [37], where the ideal value (k_c) is defined by the equation $k_c = 1 + 3.322log(\zeta)$, where ζ is the number of objects in the study (in this case, the number of substations). Thereafter, the CA is applied considering each hierarchical (single, centroid, complete, average, median, McQuitty, and Ward) and non-hierarchical (*k-means*) linkage method, repeating the application for each replica and storing the classification indicated through its respective membership.

Step 6: After extracting and storing the memberships of each linkage method (for each replica), the Kappa statistic and Kendall concordance coefficient can be calculated. These indicators assess the variability within and between linkage methods in categorical results, which are the characteristics of the hard-cluster approaches. These indexes indicate the linkage methods (after performing all the treatments) that have the best consistency of results under small disturbances. Therefore, the most robust method for a given dataset can be identified by checking the best Kappa–Kendall results ($\geq$0.9) based on the AIAG [18] classification criteria.

Step 7: Identifying the most robust methods, the quality of the results must be verified by comparing the classification results with the response of interest, indicating confidence intervals to assess the separability and non-overlapping of the groups. For this, the analysis of covariance (ANCOVA) approach is used to make an adjustment to the analysis of variance (ANOVA), explaining the main variable in terms of a concomitant variable impacting the analysis [38].

Step 8: Finally, after determining the most consistent method with better cluster discrimination, the robustness in predicting different results is confirmed by comparing some machine learning methods (such as *kNN algorithm, SVC, ANN, RFC*, and *MLR*).

Initially, the training is conducted using the original data as the "*input*" (in this study, the substation data from [9]) and the classification resulting from the best method as the "*output.*" Subsequently, the Monte Carlo simulation is performed again to create scenarios with higher levels of disturbance: 5, 10, and 15% using the original data. The sets of scenarios with disturbances are used as test sets for the MLT to predict the robust results obtained through the proposed method and determine the best technique for forecasting these scenarios.

5. Application of the proposed method

5.1. RFR-FA-based approach to substation clustering

Based on the substations' PQ data, the variables are initially analyzed to define the characteristics that best explain the TNE (main variable). This analysis is necessary because the dataset presents a larger number of variables in relation to the observations. In Step 1, the RFR strategy is applied considering 100 estimators for the method. The maximum number of features is 5, because the parameter is assumed to have a value of $\log_2 N$ (N is the number of input variables for the problem).

The random state is a parameter that can vary the precision of the RFR as it controls the randomness of the bootstrapping of the samples used during the construction of the trees. Thus, the sampling of the features is considered when the method seeks the best division in each node. This parameter ranges from 0 to 49, storing the R^2 score values for each of them. Therefore, the model producing the highest R^2 score (93.84%) is used, indicating that the model can adequately adjust to the past data. Fig. 3(a) illustrates all 50 cumulative executions, and Fig. 3(b) shows the behavior of the RFR with the best R^2 score.

Based on the algorithm results, the variables exhibiting the highest impact on the amount of voltage sags are selected, with respect to the constraint of the FA technique, which includes: NEMV, EMVVA, SAIFI1, LNE, UNE, ANE, MVRF, STIFI, Xo, FKVAr, EVAHV, MAXA, SAIFI2, MAXS, and 3LG. The results show that the NEMV and EMVVA variables are the most significant variables, representing the number of events at medium voltage and the vulnerability area at medium voltage, respectively. It is noteworthy that these variables are relevant in voltage sag studies [1,9]. Therefore, 16 variables are selected (considering the TNE), providing an initial reduction of 48.38% in the data dimensionality. Fig. 4 illustrates the variance–covariance structure of the filtered variables, indicating a multivariate behavior between the PQ characteristics.

To verify the behavior and robustness of the cluster methods, replicas must initially be created with small perturbations, as reported by Johnson and Wichern [17]. Thus, a disturbance of 3% is applied to the data set, creating four scenarios with random values (R_1, R_2, R_3, and R_4).

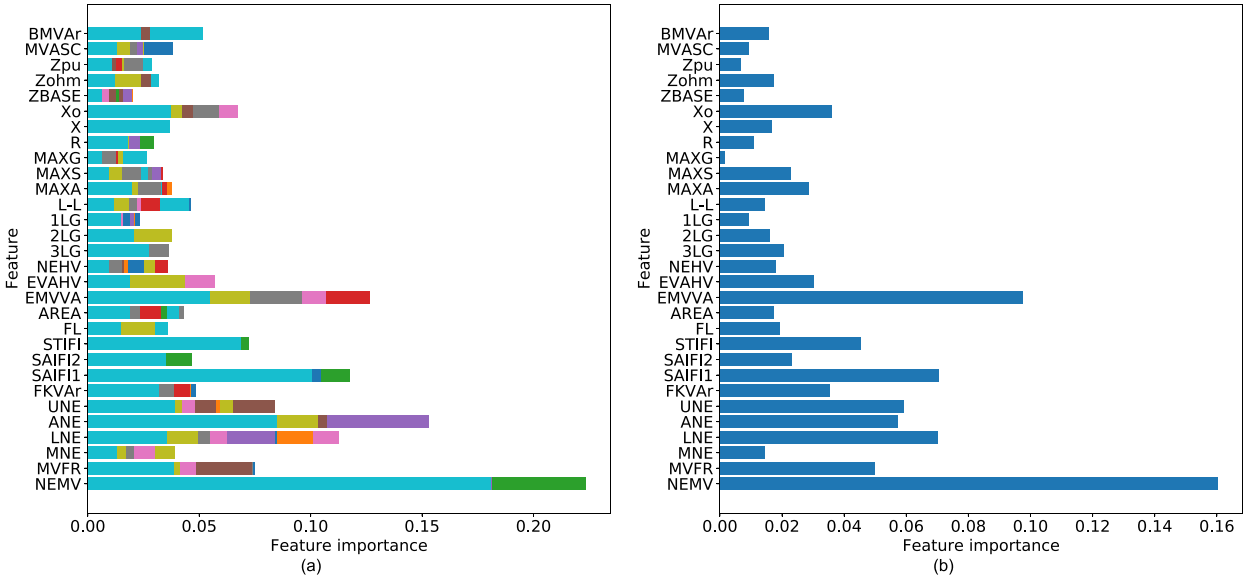

Fig. 3. Feature importance graph for (a) 50 RFR models and (b) best RFR model (TNE as investigate response).

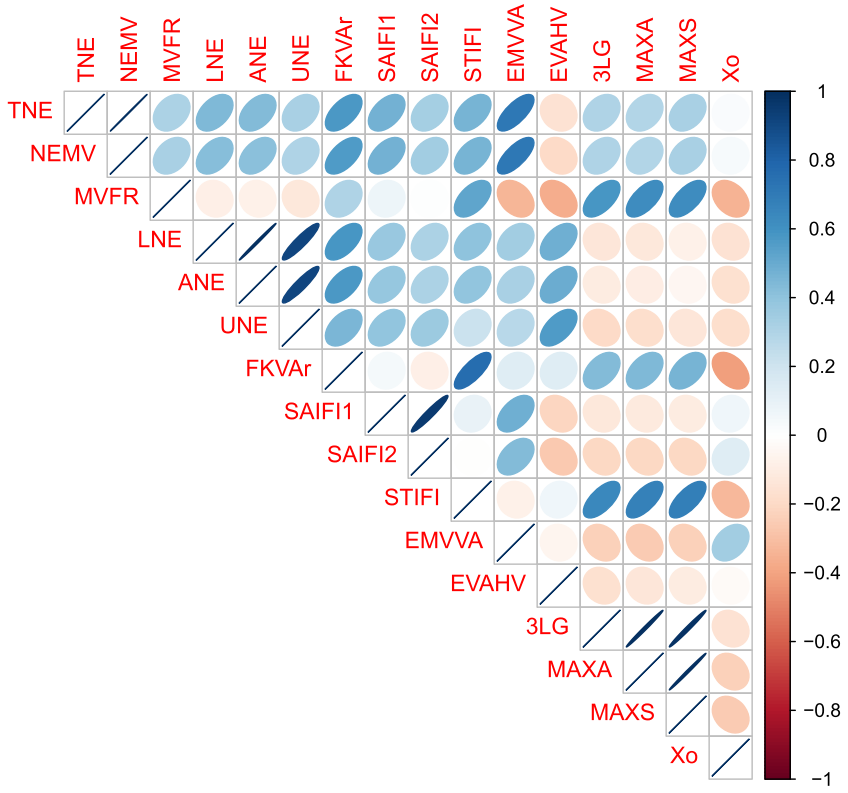

Fig. 4. Correlation results from of main variables.

Each scenario represents the average value of 10,000 simulations created using the Monte Carlo approach. The 3% replica sets (disturbed scenarios) are provided in the supplementary material.

Before applying the FA, the adequacy tests must be performed for the filtered data (from the RFR strategy). The Mardia test indicates that the data do not represent a normal multivariate distribution; therefore, the Bartlett's sphericity test should not be used (because this test is subject to that specific distribution). In the KMO test, it is observed that the indicator shows a value of 0.5, indicating that the data are adequate for the FA application.

For this application, the factor numbers must be defined initially. Assessing the percentage of factor contributions, 5 factors are verified to be ideal for the analysis. This behavior is repeated in both scenarios, with values >90% of the data explanation. Therefore, the substations' PQ variables can be represented using 5 factor score vectors, presenting a new dimensionality reduction of 68.75% (with a total reduction of 83.87% of the original dataset). This data structure minimization requires less computational effort. Thereafter, the FA is applied using the rotation of the factorial loads by the *varimax* method, which presents load separability, with a communality level of 0.9196 for R_1. Table 2

Table 2
Rotated factor loads and communalities for R_1.

Variable	FA_1	FA_2	FA_3	FA_4	FA_5	Communality
MAXA	0.976	0.083	−0.032	−0.087	−0.064	0.9731
3LG	0.975	0.103	−0.057	−0.1	0.003	0.9742
MAXS	0.974	0.045	−0.06	−0.096	−0.075	0.9684
STIFI	0.738	−0.341	−0.233	−0.004	−0.28	0.7938
MVFR	0.642	0.23	−0.058	0.185	−0.484	0.7369
ANE	−0.007	−0.912	−0.254	0.223	−0.108	0.957
LNE	−0.03	−0.906	−0.283	0.213	−0.116	0.9607
UNE	−0.09	−0.902	−0.132	0.283	−0.08	0.9259
EVAHV	−0.086	−0.814	0.283	−0.265	0.219	0.8684
FKVAr	0.459	−0.52	−0.456	−0.16	−0.398	0.8723
NEMV	0.276	−0.12	−0.916	0.216	−0.052	0.9785
TNE	0.275	−0.15	−0.912	0.208	−0.045	0.9756
EMVVA	−0.304	−0.124	−0.839	0.19	0.302	0.9385
SAIFI2	−0.13	−0.148	−0.166	0.951	0.057	0.9741
SAIFI1	−0.052	−0.182	−0.285	0.916	−0.005	0.957
Xo	−0.161	0.131	−0.116	0.062	0.893	0.8586
Variance	4.3255	3.713	3.0107	2.2064	1.4574	14.7131
% Var	0.27	0.232	0.188	0.138	0.091	0.9196

presents the rotated loads and communality for R_1, indicating the best behavior based on the principle of parsimony. Similar behavior is observed for the other replicas. The extracted factor scores for R_1 are presented in Table 3. Owing to the extent of the data, information from the other scenarios (R_2, R_3, and R_4) are provided in the supplementary material.

The factor scores (from each replica) represent the original variables, indicating the dimensionless and independent values. For the CA application, it is necessary to define the number of categories for the substations to be classified using the Sturges rule (Step 5). Considering the number of substations investigated (observations), $k_c = 1 + 3.322log(17) = 5.087 \cong 5$ is the number of categories. Based on this, the CA must be applied to the 8 different linkage methods, storing the memberships created for each method in their respective replicas. It should be emphasized that there is no need to standardize the variables, as the factor scores are dimensionless. Furthermore, the Mahalanobis distance is not used because the variables are independent (in this way, the Euclidean distance is fixed as a dissimilarity metric). Table 4 presents the classifications created for each linkage method and their respective replica. Fig. 5 illustrates the dendrograms created using the hierarchical methods for replica R_1.

The measurement system analysis can be implemented as a multilevel full factorial design by considering substations and linkage methods as multilevel factors of a DOE (with 4 replicas) and performing the experimental analysis with membership values. Analyzing the experimental design, there is an adequate model fit with R^2 and R^2_{adj} values of 81.62 and 75.54%, respectively. Considering a confidence interval (CI) of 95% in the ANOVA, it can be observed that a statistically significant difference exists between the substations (p-value = 0.000) as these substations have distinct PQ characteristics (indicating that the experimental design is adequate). Analyzing the cluster method, a statistically significant difference is found to exist between the linkage methods, presenting a high F value (of 73.74) and, consequently, a p-*value* = 0.000. This result indicates that some linkage methods show different behaviors through the replications. Therefore, the methods present different classifications with cluster inversions under the perturbation scenarios. This indicates the need to assess the variability and consistency of linkage methods using adequate statistics.

Table 3
Rotated factor scores for R_1.

Variable	FA_1	FA_2	FA_3	FA_4	FA_5
TNE	1.9690	−8.9360	4.3490	−3.6240	7.5260
NEMV	−2.0010	9.1100	−5.0820	3.5180	−7.6230
MVFR	0.0910	0.1090	0.0350	0.1520	−0.2930
LNE	−0.0120	−0.2370	−0.0050	0.0160	−0.0290
ANE	0.0030	−0.2430	0.0160	0.0310	−0.0130
UNE	0.0040	−0.2500	0.0800	0.0890	0.0030
FKVAr	0.0000	−0.1040	−0.1770	−0.1960	−0.2300
SAIFI1	0.0490	0.0130	0.0890	0.4730	0.0230
SAIFI2	0.0560	0.0100	0.1460	0.5180	0.0650
STIFI	0.1570	−0.0880	−0.0070	−0.0070	−0.0500
EMVVA	−0.1040	0.0390	−0.3650	−0.1050	0.1520
EVAHV	0.0000	0.0000	0.0000	0.0000	0.0000
3LG	0.2960	−0.0090	0.0610	0.0320	0.2240
MAXA	0.2860	−0.0100	0.0730	0.0420	0.1690
MAXS	0.2790	−0.0190	0.0600	0.0280	0.1600
Xo	0.1550	−0.0290	−0.0080	0.0400	0.7390

Based on the previous design, the agreement and Kappa–Kendall indicators are calculated considering a CI of 95%. In this analysis, the Ward method homogeneously classifies all 17 substations in the four scenarios with disturbances, showing 100% agreement in the cluster formation. The Median and k-means methods show a degree of agreement of 82.35% in both approaches, with a CI ranging between 56.57 and 96.2%. The other methods show values below 60%, indicating a low consistency. Fig. 6 illustrates the behavior of the agreement values of each method and the respective CI for an α value of 0.05.

In addition to the consistency of the methods, the most suitable statistic to assess the level of agreement is the Fleiss' Kappa Statistics. Table 5 presents the Kappa values and other necessary statistics. The results verify that the Ward method shows a better degree of agreement ($K = 1.00$), indicating no inversions in the classification of substations in the four different scenarios. This verifies the robustness of this approach as the scenarios present a 3% disturbance. The k-means method shows the second-best behavior ($K = 0.882$), presenting inversions only for clusters 4 and 5. Furthermore, the inversions occur between substations s12, s16, and s7, the latter being one of the substations with the highest incidences of TNE (nominal voltage of 138 kV). The median method obtains a generalized K value of 0.780, indicating instability in the classification of clusters 1, 2, and 3, with individual K values of 0.83, 0.46, and 0.46, respectively. Median values show inversions in the formation of clusters 2 and 3, indicating a difficult classification under small perturbations for this data set. Substations *s4, s7*, and *s8* present inversions, the last two substations showing the highest number of voltage variation incidences causing the sag events. The other linkage methods show generalized Kappa values below the criteria recommended by the AIAG [18], indicating a low quality of consistency in the results.

To complement the linkage concordance assessment, it is necessary to analyze the Kendall's agreement coefficient, which presents the degree of agreement between the linkage methods and calculates the W indicator for this purpose. Table 6 presents the results required to assess this indicator. Based on the results and established criteria [18], the Ward method presents an excellent value ($W = 1.00$). The k-means and Median methods also present excellent values (W values of 0.964 and 0.905, respectively). The McQuitty, Centroid, and Single methods show acceptable Kendall's coefficient of concordance only for this analysis, with W values of 0.821, 0.772, and 0.720, respectively. For this indicator, the other methods present inadequate values ($W < 0.7$).

5.2. Sensitivity analysis

Based on the results presented in Section 5.1, it is necessary to compare the classification results with the TNE values from the substations. At this stage, the linkage methods presenting the best simultaneous results for the Kappa–Kendall evaluation are considered ($K \geq 0.75$ and $W \geq 0.9$). For this analysis, the original dataset is used to perform the entire procedure (RFR-FA and CA), which defines the clusters to be investigated. Table 7 presents the FA scores rotated for the original data and the memberships of the more stable linkage methods (Ward, k-means, and Median). Fig. 7 illustrates the dendrograms for the hierarchical methods (k-means does not present this graphic illustration). The figure shows that the Ward method presents a good

Table 4
Membership values of each linkage method from all disturbed scenarios.

S	Linkage	Membership				S	Linkage	Membership				S	Linkage	Membership			
		R_1	R_2	R_3	R_4			R_1	R_2	R_3	R_4			R_1	R_2	R_3	R_4
Aracruz (s1)	Average	1	1	1	1	*João Neiva (s7)*	Average	4	5	2	4	*Paulista (s13)*	Average	3	3	1	5
	Centroid	1	1	1	1		Centroid	2	2	2	2		Centroid	1	1	1	1
	Complete	1	1	1	1		Complete	4	3	3	4		Complete	2	5	2	5
	McQuitty	1	1	1	1		McQuitty	3	4	2	4		McQuitty	1	1	1	1
	Median	1	1	1	1		Median	3	2	2	2		Median	1	1	1	1
	Single	1	1	1	1		Single	2	2	2	3		Single	1	1	1	1
	Ward	1	1	1	1		Ward	5	5	5	5		Ward	3	3	3	3
	k-means	1	1	1	1		k-means	5	4	5	5		k-means	4	4	4	4
Baixo Guandu (s2)	Average	1	1	1	1	*Juncado (s8)*	Average	3	4	1	3	*Pinheiros (s14)*	Average	3	4	1	3
	Centroid	1	1	1	1		Centroid	3	1	1	3		Centroid	1	1	1	1
	Complete	1	1	1	1		Complete	3	4	4	3		Complete	3	1	1	3
	McQuitty	1	1	1	1		McQuitty	4	3	3	3		McQuitty	1	3	1	3
	Median	1	1	1	1		Median	1	3	3	3		Median	1	1	1	1
	Single	1	1	1	1		Single	3	3	3	1		Single	1	1	1	1
	Ward	1	1	1	1		Ward	4	4	4	4		Ward	4	4	4	4
	k-means	2	2	2	2		k-means	5	5	5	5		k-means	5	5	5	5
Barra do Sahy (s3)	Average	2	2	1	2	*Linhares A (s9)*	Average	3	4	3	3	*Santa Tereza (s15)*	Average	2	2	1	2
	Centroid	1	1	1	1		Centroid	4	3	3	4		Centroid	1	1	1	1
	Complete	1	2	2	2		Complete	3	4	4	3		Complete	1	5	2	5
	McQuitty	2	2	1	2		McQuitty	4	3	3	3		McQuitty	2	2	1	2
	Median	1	1	1	1		Median	4	4	4	4		Median	1	1	1	1
	Single	1	1	1	2		Single	4	4	4	1		Single	1	1	1	1
	Ward	2	2	2	2		Ward	4	4	4	4		Ward	2	2	2	2
	k-means	3	3	3	3		k-means	5	5	5	5		k-means	3	3	3	3
Ecoporanga (s4)	Average	3	3	1	1	*Linhares C (s10)*	Average	2	2	1	2	*São Francisco (s16)*	Average	3	3	1	5
	Centroid	1	1	1	1		Centroid	1	1	1	1		Centroid	1	1	1	1
	Complete	2	1	1	1		Complete	1	2	2	2		Complete	2	2	1	2
	McQuitty	1	1	1	1		McQuitty	2	2	1	2		McQuitty	1	1	1	1
	Median	2	1	1	1		Median	1	1	1	1		Median	1	1	1	1
	Single	1	1	1	1		Single	1	1	1	2		Single	1	1	1	1
	Ward	3	3	3	3		Ward	2	2	2	2		Ward	3	3	3	3
	k-means	4	4	4	4		k-means	3	3	3	3		k-means	4	5	4	4
Itarana (s5)	Average	3	4	1	3	*Montanha (s11)*	Average	3	3	1	5	*Suíça (s17)*	Average	2	2	5	2
	Centroid	1	1	1	1		Centroid	1	1	1	1		Centroid	1	5	5	5
	Complete	3	1	1	3		Complete	2	2	1	2		Complete	1	5	2	5
	McQuitty	1	3	1	3		McQuitty	1	1	1	1		McQuitty	2	2	5	2
	Median	1	1	1	1		Median	1	1	1	1		Median	1	1	1	1
	Single	1	1	1	1		Single	1	1	1	1		Single	1	1	1	5
	Ward	4	4	4	4		Ward	3	3	3	3		Ward	2	2	2	2
	k-means	5	5	5	5		k-means	4	4	4	4		k-means	3	3	3	3
Jaguaré (s6)	Average	1	1	1	1	*Nova Venécia (s12)*	Average	5	3	4	5	Intentionally left blank					
	Centroid	1	1	1	1		Centroid	5	4	4	1						
	Complete	1	1	1	1		Complete	5	2	5	2						
	McQuitty	1	1	1	1		McQuitty	5	5	4	5						
	Median	1	1	1	1		Median	5	5	5	5						
	Single	1	1	1	1		Single	5	5	5	4						
	Ward	1	1	1	1		Ward	3	3	3	3						
	k-means	2	2	2	2		k-means	4	5	4	4						

separability between the groups (Fig. 7(a)), with high levels of similarity between them. However, the Median method shows difficulties in the classifications, associating 76.47% of the substations in only one cluster (Fig. 7(b)). This indicates the limitation of this method to correctly classify the information in this dataset. It is also verified that the other clusters have only one substation for each classification, where substations *s7* (cluster 2), *s8* (cluster 3), and *s9* (cluster 4) have the highest TNE values (426.129, 498.078, and 543.971, respectively). Substation *s12* (in cluster 5) has the lowest TNE value of 51.209.

From the classifications, the concomitant variable is defined based on the results of the RFR for the application of the ANCOVA. The NEMV variable exhibits the largest impact; however, it presents a linear dependence on the TNE (Fig. 4, Pearson coefficient of 1.00). Thus, the EMVVA is defined as the next most important variable. The selection of this variable is practically justified as it considers the vulnerability area (in km^2) for the occurrence of sags in the substation busbars (caused by short circuits).

Thus, the ANCOVA is performed for the three selected linkage methods. Fig. 8 shows that the Ward method (Fig. 8(a)) presents a good discrimination between the groups, where the substations with higher incidences (*s5, s7, s8, s9*, and *s14*) can be differentiated from the substations with the best PQ level adequately. Additionally, the analysis allows for narrow and precise intervals, considering a 95% confidence level. Analyzing the second method with the best consistency result, the k-means method (Fig. 8(b)) presents discrimination only for cluster 5, indicating an overlap of the CIs for the other groups. This hinders the separability of groups and consequently favors the confounding interpretation analysis. Finally, the Median method (Fig. 8(c)) presents a narrow CI for cluster 1 (because the method classifies 76.47% of the observations in the same group). However, it presents a CI amplitude identical to those of the other clusters; because it has only one substation in each grouping, the pooled standard deviation is used to calculate the intervals.

From the results of the cluster formation and CI analysis performed through the ANCOVA, it is observed that the Ward linkage method presents a better behavior and discrimination in the clusters of substations for this dataset. This result is reinforced through the analysis of the Kappa–Kendall indicators, in which only the Ward method presents a classification called *excellent* (K and $W > 0.9$), according to the international criteria for agreement analysis [18].

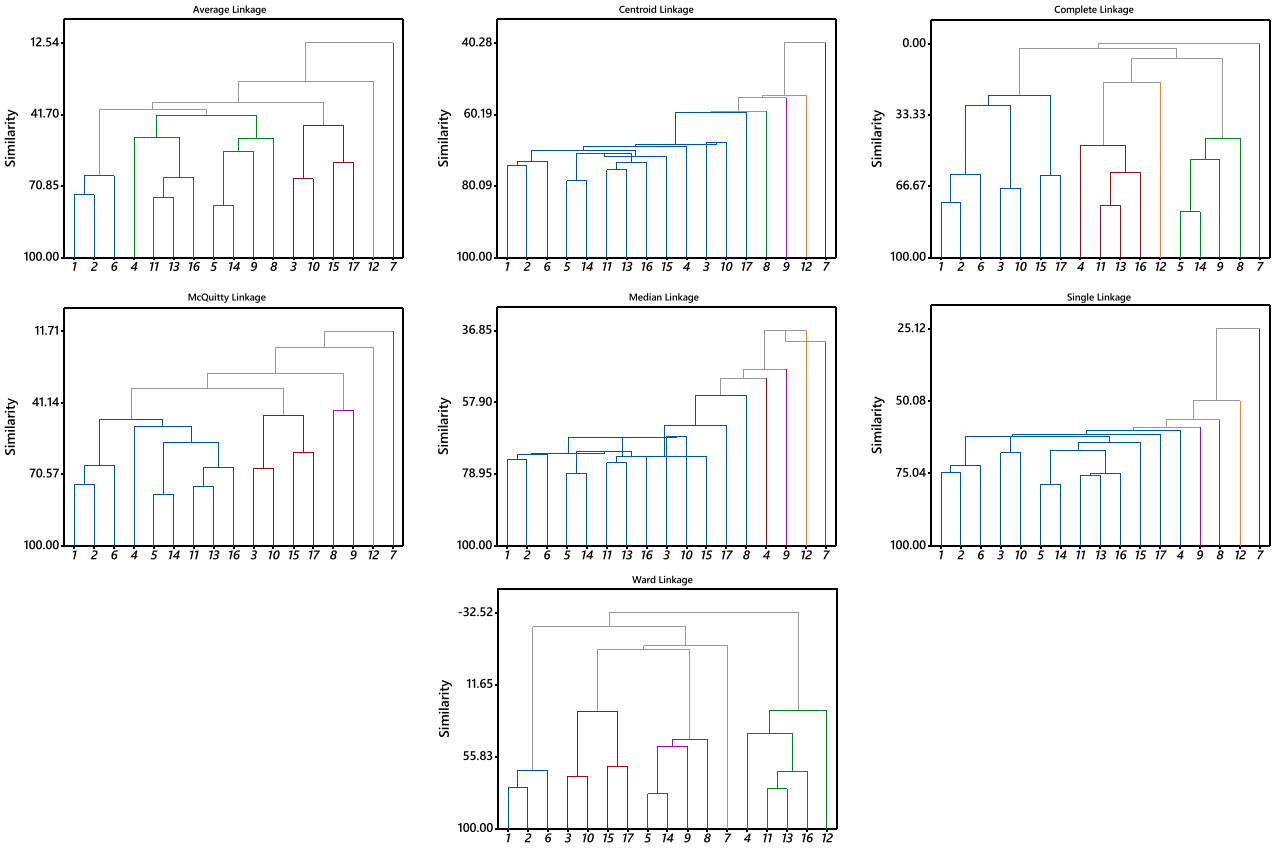

Fig. 5. Dendrograms of hierarchical linkage methods for R1.

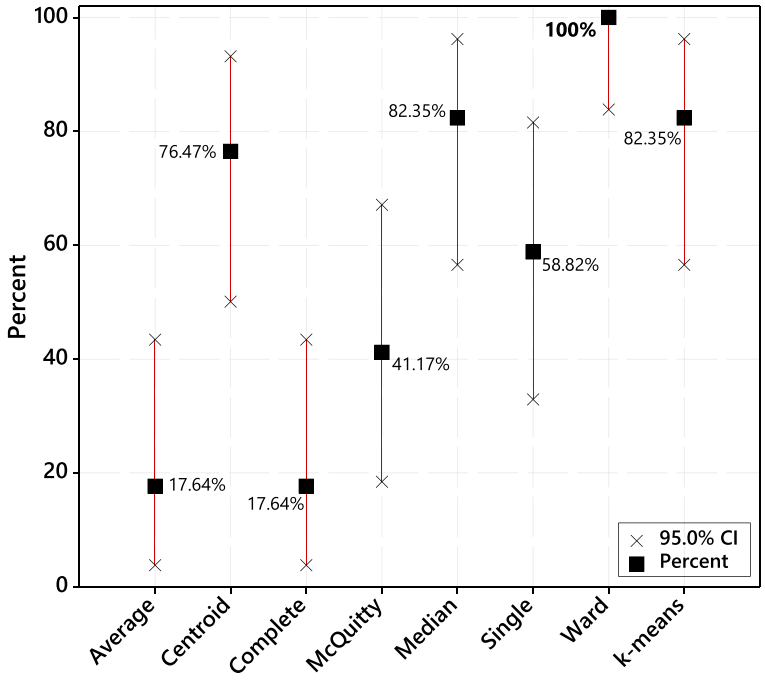

Fig. 6. Agreement chart for the linkage methods.

5.3. Prediction analysis through machine learning techniques

After identifying the linkage method with larger robustness and discrimination power (Ward method) in *Step 8*, the proposed method is made to perform a search to predict scenarios with other disturbance levels (5, 10, and 15%) for verification. The behavior of the results is observed in scenarios with significant levels of variability. For this purpose, five supervised machine learning classifier methods are

Table 5
Fleiss' Kappa statistics for substation clusters.

Linkage	*Cluster*	*Kappa*	*P(vs > 0)*	*Linkage*	*Cluster*	*Kappa*	*P(vs > 0)*
Average	1	0.3211	0.001	*Median*	1	0.8365	0.000
	2	0.5245	0.000		2	0.4688	0.000
	3	0.1437	0.073		3	0.4688	0.000
	4	−0.0086	0.535		4	1.0000	0.000
	5	−0.0086	0.535		5	1.0000	0.000
	Overall	**0.2367**	**0.000**		**Overall**	**0.7802**	**0.000**
Centroid	1	0.7276	0.000	*Single*	1	0.6078	0.000
	2	1.0000	0.000		2	0.3524	0.000
	3	0.2917	0.002		3	0.4688	0.000
	4	0.2917	0.002		4	0.4688	0.000
	5	0.4688	0.000		5	0.4688	0.000
	Overall	**0.6092**	**0.000**		**Overall**	**0.5118**	**0.000**
Complete	1	0.3885	0.000	*Ward*	1	1.0000	0.000
	2	0.2940	0.002		2	1.0000	0.000
	3	0.2184	0.014		3	1.0000	0.000
	4	0.2688	0.003		4	1.0000	0.000
	5	0.2444	0.007		5	1.0000	0.000
	Overall	**0.3011**	**0.000**		**Overall**	**1.0000**	**0.000**
McQuitty	1	0.6664	0.000	*k-means*	1	1.0000	0.000
	2	0.5245	0.000		2	1.0000	0.000
	3	0.3854	0.000		3	1.0000	0.000
	4	0.0646	0.257		4	0.7809	0.000
	5	0.4688	0.000		5	0.7933	0.000
	Overall	**0.4976**	**0.000**		**Overall**	**0.8830**	**0.000**

Bold:overall result of the Kappa indicator.

Table 6
Kendall's coefficient of concordance for substation clusters.

Linkage	*Coef.*	*Chi - Sq*	*DF*	*P*
Average	0.6670	42.6902	16	0.0003
Centroid	0.7721	49.4159	16	0.0000
Complete	0.5943	38.0369	16	0.0015
McQuitty	0.8214	52.5672	16	0.0000
Median	0.9054	57.9469	16	0.0000
Single	0.7203	46.0987	16	0.0001
Ward	1.0000	64.0000	16	0.0000
k-means	0.9644	61.7218	16	0.0000

Table 7
Rotated factor scores from the original data set.

Variable	F1	F2	F3	F4	F5
TNE	2.643	−10.032	−5.026	−4.313	9.753
NEMV	−2.678	10.208	5.768	4.199	−9.857
MVFR	0.098	0.101	−0.041	0.151	−0.286
LNE	−0.011	−0.243	−0.007	0.025	−0.028
ANE	0.005	−0.252	−0.029	0.041	−0.009
UNE	0.006	−0.264	−0.096	0.096	0.008
FKVAr	−0.013	−0.074	0.200	−0.197	−0.257
SAIFI1	0.053	−0.013	−0.104	0.464	0.036
SAIFI2	0.056	−0.014	−0.159	0.508	0.070
STIFI	0.160	−0.091	−0.006	0.004	−0.049
EMVVA	−0.109	0.053	0.382	−0.108	0.149
EVAHV	0.000	0.000	0.000	0.000	0.000
3LG	0.293	−0.015	−0.066	0.026	0.218
MAXA	0.285	−0.017	−0.080	0.038	0.168
MAXS	0.280	−0.026	−0.068	0.024	0.163
Xo	0.144	−0.033	0.006	0.037	0.711

selected: kNN, SVC, ANN, RFC, and MLR.

Initially, the complete original dataset (in [9]) is used as the "*input*" for training with 32 variables for 17 observations, and the classification vectors performed by the proposed method (RFR-FA) is used as the "*output*" together with the best performance linkage method (Ward), described in Section 5.2. The behavior of the data with the cluster classification is shown in Fig. 9. This figure shows the main variable (TNE) and the four variables that have the largest impact in this study, considering the feature importance results.

To create more complex scenarios, the Monte Carlo simulation is again used to create three data with 5, 10, and 15% disturbances. The mean value of 10,000 simulations is considered for each scenario. Subsequently, these new scenarios are used to verify the accuracy of the MLT for correct classification, based on supervised learning, created by the proposed model. To generate the results, the algorithms are executed 50 times, varying the random state value from 0 to 49. Therefore, the demonstrated precisions deal with the mean value of 50 executions (with the exception of the kNN algorithm, because it does not have this randomness parameter). Table 8 presents the level of accuracy of each algorithm and their performance evaluations for each scenario.

The results show that the MLT can learn the behavior of the input data, even with complex disturbance levels. Only two methods do not show absolute accuracy (RFC and ANN), but they yield good results. Furthermore, the MLR method, despite being a simpler strategy, presents satisfactory results and better performance than those of other more sophisticated and difficult-to-implement methods, such as ANN and RFC. It is also noteworthy that the kNN, ANN, and SVC methods show significantly improved accuracy when the input data are standardized, which do not occur for the RFC and MLR methods.

To clarify the application and performance of each method, a brief discussion on each MLT used is presented below, describing their parameterization:

- *RFC*: In this study, 100 estimators are considered, and the maximum number of features is 5. Other values are tested, such as a larger number of trees and features; however, they do not show improvements or more accurate results;
- *ANN*: The multilayer perceptron classifier is used for confirmation experiments. Only a hidden layer with 10 neurons is considered, in addition to an input layer composed of 31 neurons and an output layer with only one processing unit. The learning rate is 0.30. The Levenberg–Marquardt optimization is used for the parameterization of weights and bias, and the goal train parameter is 0.001;
- *k-NN algorithm*: The Euclidean distance (most widespread metric) is used to test different numbers of neighbors: $k = 1$, 3, 5, and 7. The best accuracy obtained is $k = 1$. Therefore, an accuracy of 100% is obtained for all test sets, considering 5, 10, and 15% disturbances in the original data;

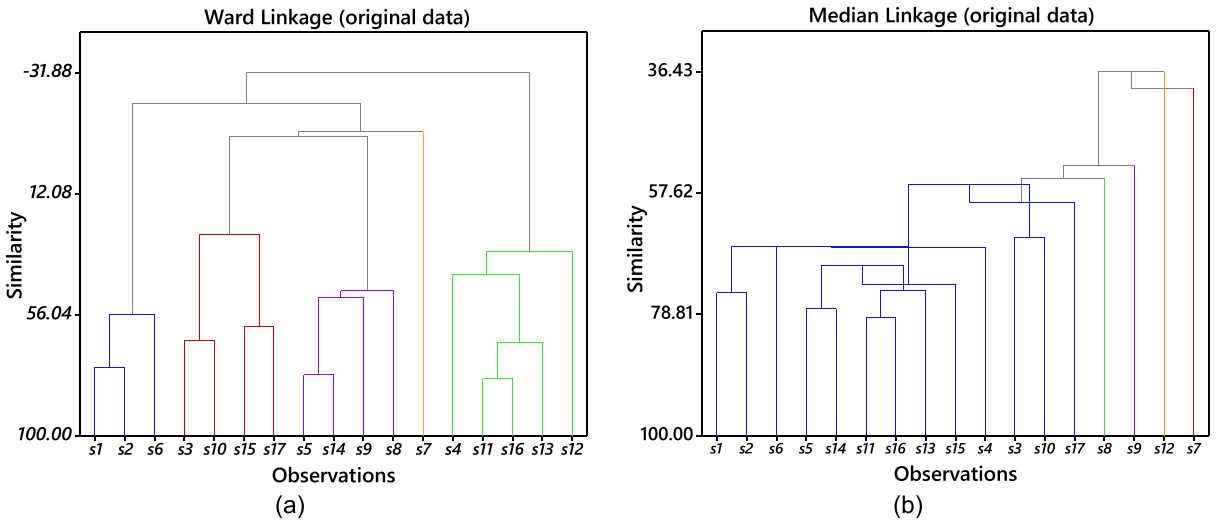

Fig. 7. Original data dendrogram from (a) Ward and (b) Median method.

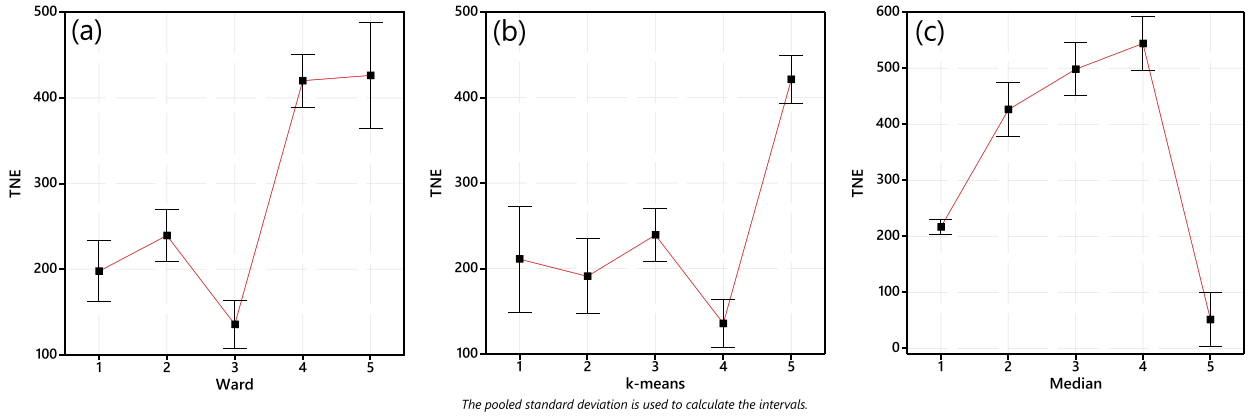

Fig. 8. 95% confidence interval plot for (a) Ward, (b) k-means and (c) median.

- *SVC*: To perform the parameterization of this algorithm, an exhaustive search over specified parameter values for an estimator is performed, such as C (0.1, 1, 10, and 100), gamma (1, 0.1, 0.01, and 0.001), and finally the selected kernel (RBF, polynomial, and sigmoidal). After performing the search, the following values are obtained: C = 0.1, gamma = 1, and kernel = polynomial;
- *MLR*: In this application, a maximum iteration value of 100 is considered, with a solver-type limited-memory Broyden–-Fletcher–Goldfarb–Shanno optimization algorithm and a tolerance of 10^{-4}.

6. Conclusions

Based on the proposed method and application with real data from PQ distributions in Brazil, the following conclusions can be drawn:

- The methodology proves to be suitable for planning and assessing different linkage methods, favoring the selection of a robust method in disturbance scenarios;
- The mixed application of the RFR and FA techniques provide filtering and minimizing the data dimension by 83.87%, favoring the reduction of computational effort by creating independent vector scores capable of adequately representing the dataset. Additionally, the Monte Carlo simulation creates the disturbance scenarios adequately, without significantly de-characterizing the variance–covariance structure of the data;
- The use of the multilevel factorial DOE strategy allows the creation of an appropriate experimental design for the study and application of the Kappa–Kendall strategy. The DOE promotes the prior analysis of the relationship between the linkage methods and observations (substations)
- The assessment of the Kappa–Kendall indicators reveal that the Ward linkage method presents a better behavior for this dataset, with K and W values equal to 1000. The non-hierarchical k-means method presents acceptable results ($K = 0.882$; $W = 0.964$), but with difficulties in classifying clusters 4 and 5, presenting inversions in memberships. Similarly, the Median approach presents acceptable Kappa indices ($K = 0.780$), presenting difficulty in classifying the three substations (*s4, s7*, and *s8*). This approach also presents acceptable values for Kendall's coefficient of agreement ($W = 0.905$)
- The sensitivity analysis reveals that the Ward method has a good level of separability through the ANCOVA analysis, discriminating substations with narrow and precise CIs. The k-means method does not exhibit the same performance, showing overlapping CIs from clusters 1 to 4. The Median method presents acceptable Kappa-Kendall indexes but does not perform an adequate grouping, allocating 13 substations in a single cluster. The Median method also classifies other substations into single clusters. Therefore, the Ward method exhibits the best behavior in scenarios with variability, presenting a high level of robustness and better discrimination between the substation clusters in this dataset;

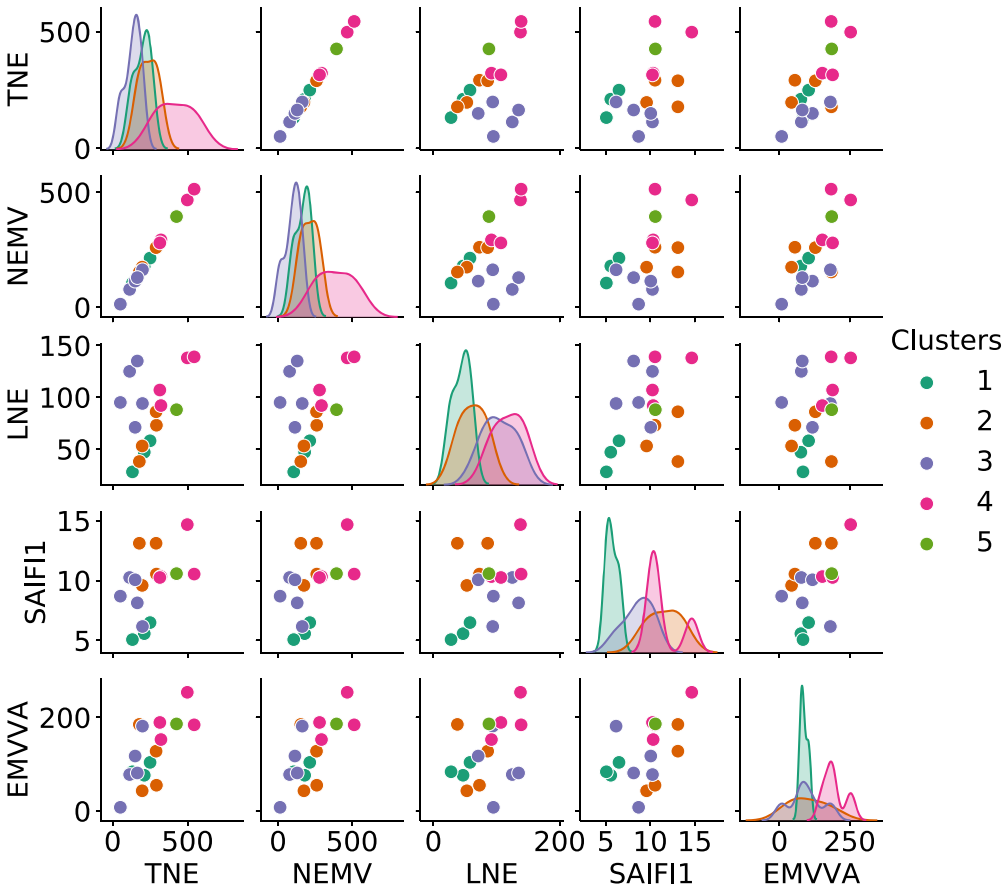

Fig. 9. Behavior of RFR-FA-Ward clusters.

Table 8
Accuracy of the algorithms on the test set.

Machine learning algorithms	Simulation of different scenarios			Overall
	5% disturbances	*10% disturbances*	*15% disturbances*	
RFC	99.88%	95.76%	90.47%	95.37%
kNN	100.00%	100.00%	100.00%	**100%**
ANN	81.18%	76.70%	58.23%	72.04%
SVC	100.00%	100.00%	100.00%	**100%**
MRL	100.00%	100.00%	100.00%	**100%**

- Using other MLTs to predict different scenarios with higher disturbance levels (5, 10, and 15%), it is verified that all analyzed supervised learning approaches present satisfactory values. However, some strategies are better (kNN, SVC, and MLR). It is also noteworthy that the MLR presents better results than those of other more complex techniques, such as ANN and RFC.

Based on the obtained results, the proposed method proves to be adequate for evaluating and defining a method with larger robustness in different voltage sag scenarios. The method also promotes an optimal interpretation of the data structure with a dimensionality reduction. The results for this set favor the creation of predictive models capable of interpreting the entire treatment performed by the techniques imposed in the method. As future proposals, new hard and soft cluster methods can be investigated and applied to datasets from other sectors.

Table A.1
Nomenclature of the variables used in the study.

HVFR	*High voltage failure rate*	*NEHV*	*Number of events high voltage*
TNE	*Total number of sag events per year*	*3LG*	*Three phase to ground short-circuit current*
NEMV	*Number of events in medium voltage*	*2LG*	*Double phase to ground short-circuit current*
MVFR	*Medium voltage failure rate*	*1LG*	*Single phase to ground short-circuit current*
MNE	*Monitored number of events*	*L-L*	*Phase to phase short-circuit current*
LNE	*Lower number of events*	*MAXA*	*Maximum asymmetric short circuit Current*
ANE	*Average number of events*	*MAXS*	*Maximum symmetric short circuit Current*
UNE	*Upper number of events*	*MAXG*	*Maximum symmetric short circuit Current to Ground*
FKVAr	*Shunt capacitor KVAr installed on the feeders*	*R+*	*Positive sequence resistance*
SAIFI1	*System average interruption frequency index (without critical day)*	*X+*	*Positive sequence reactance*
SAIFI2	*System average interruption frequency index (with critical day)*	*Xo*	*Zero sequence reactance*
STIFI	*System total interruption frequency index*	*ZBASE*	*Base impedance ohm*
FL	*Feeders length (km)*	*Zohm*	*Equivalent impedance ohm*
AREA	*Cluster area (km2)*	*Zpu*	*Per unit impedance*
EMVVA	*Equivalent medium voltage vulnerability area (km)*	*EVAHV*	*Equivalent high voltage vulnerability area (km)*
BMVAr	*Shunt capacitor MVar installed on the bus bar*	*MVASC*	*Short circuit power MVA (1000/Zpu on the bus bar)*

Table A.2
Pseudocode of the proposed method.

Pseudocode: Robust hard-cluster assessment using MLT and Kappa-Kendall Indexes
Input: Dataset that will be applied to cluster analysis
Output: Best linkage method to use

$n \leftarrow$ number of variables *//Power Quality characteristics*
$N \leftarrow$ number of observations *//Substations*
$x_{11}, x_{12}, \ldots, x_{ni} \leftarrow$ original data for each observation *//PQ indexes for substation* ($i = 1, 2, \ldots, N$)
$V_1 \leftarrow$ main variable *//Total number of sag events per year*
if $n \geq N$
Apply RFR (x_{ni}) *//Random forest regressor*
$rfr \leftarrow 1$
$w_1, w_2, \ldots, w_k \leftarrow k$ variables with higher level of feature importance for V_1
end if
//Define disturbance percentage to be applied to the original data
$disturb \leftarrow$ 3% disturbance level
Apply Monte Carlo simulation
//Create at least 4 different replicas
if correlation (w_k variables for each replica)
Apply BST or KMO
if $X\_Sqr \leq Chi\_Sqr$ or KMO ≥ 5
$Adequation \leftarrow 1$
end if
Apply factor analysis *//For each replica*
Choose number of factor *//Kaiser criterion*
Use varimax rotation
Extract factor scores
end if
Apply linkage method ($L_1, L_2, \ldots, L_i$) *//Single, Average, Centroid, Complete, Median, McQuitty, Ward and k-means*
Store memberships *//Each cluster classification for observation/substation*
Define the multilevel DOE
Analyze the experimental matrix
Apply Kappa-Kendall indexes
Define linkage methods more consistently *// Kappa (K) and Kendall (W)*
if $K > 0.9$ and $W > 0.9$
Return $L_i \leftarrow$ **Best linkage method**
end if
Apply ANCOVA (x) for L_i
Plot CI 95%

//Confirmation of the best linkage method to predict scenarios with a high level of variation
Apply Monte Carlo simulation *// Scenarios with 5%, 10% and 15% disturbance*
Create 3 new scenarios
Apply supervised machine learning techniques for classification *//kNN, ANN, SVC, RFC, MLR*
if *accuracy* > 0.8
Return ← Machine learning technique with better accuracy
end if

CRediT author statement

Fabrício Alves de Almeida: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing - Review & Editing. **Estevão Luiz Romão:** Investigation, Software, Writing - Review & Editing. **Guilherme Ferreira Gomes:** Writing - Review & Editing. **José Henrique de Freitas Gomes:** Writing - Review & Editing. **Anderson Paulo de Paiva:** Writing - Review & Editing, Methodology. **Jacques Miranda Filho:** Writing - Review & Editing. **Pedro Paulo Balestrassi:** Supervision, Writing - Review & Editing.

Declaration of Competing Interest

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

Acknowledgments

The authors would like to express their gratitude to Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) agency under process number 150117/2021–3 and to Fundação de Amparo a Pesquisa do Estado de Minas Gerais (FAPEMIG) under project APQ-00385–18.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.epsr.2022.107778.

Appendix A

Table A.1. Nomenclature of the variables used in the study
Table A.2. Pseudocode of the proposed method

References

[1] F.A. de Almeida, J. Miranda Filho, L.F. Amorim, J.H. de F. Gomes, A.P. de Paiva, Enhancement of discriminatory power by ellipsoidal functions for substation clustering in voltage sag studies, Electr. Power Syst. Res. 185 (2020), 106368, https://doi.org/10.1016/j.epsr.2020.106368.
[2] C.T. Ragsdale, Spreadsheet Modeling and Decision Analysis: A Practical Introduction to Business Analytics, 7th ed., Cengage Learning, 2014.
[3] Y. Han, Y. Feng, P. Yang, L. Xu, Y. Xu, F. Blaabjerg, Cause, classification of voltage sag, and voltage sag emulators and applications: a comprehensive overview, IEEE Access 8 (2020) 1922–1934, https://doi.org/10.1109/ACCESS.2019.2958965.
[4] H.M.A. Ahmed, A.S.A. Awad, M.H. Ahmed, M.M.A. Salama, Mitigating voltage-sag and voltage-deviation problems in distribution networks using battery energy storage systems, Electr. Power Syst. Res. 184 (2020), 106294, https://doi.org/10.1016/j.epsr.2020.106294.
[5] M. Amini, A. Jalilian, Modelling and improvement of open-UPQC performance in voltage sag compensation by contribution of shunt units, Electr. Power Syst. Res. 187 (2020), 106506, https://doi.org/10.1016/j.epsr.2020.106506.
[6] M.H. Sadeghi, A. Dastfan, Y. Damchi, Optimal coordination of directional overcurrent relays in distribution systems with DGs and FCLs considering voltage sag energy index, Electr. Power Syst. Res. 191 (2021), 106884, https://doi.org/10.1016/j.epsr.2020.106884.
[7] Y. Mohammadi, R.C. Leborgne, A new approach for voltage sag source relative location in active distribution systems with the presence of inverter-based distributed generations, Electr. Power Syst. Res. 182 (2020), 106222, https://doi.org/10.1016/j.epsr.2020.106222.
[8] M.V. Costa, J.M.C. Filho, R.C. Leborgne, N.B. Pereira, A novel methodology for determining the voltage sag Impact Factor, Electr. Power Syst. Res. 174 (2019), 105865, https://doi.org/10.1016/j.epsr.2019.105865.
[9] J. Miranda, J. Maria, D.C. Filho, A. Paulo, P. Vitor, G. De Souza, S. Tomasin, A PCA-based approach for substation clustering for voltage sag studies in the Brazilian new energy context, Electr. Power Syst. Res. 136 (2016) 31–42, https://doi.org/10.1016/j.epsr.2016.02.012.
[10] F.A. de Almeida, A.C.O. Santos, A.P. de Paiva, G.F. Gomes, J.H. de, F. Gomes, Multivariate Taguchi loss function optimization based on principal components analysis and normal boundary intersection, Eng. Comput. (2020), https://doi.org/10.1007/s00366-020-01122-8.
[11] A.B. Costello, J.W. Osborne, Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis, Pract. Assess. Res. Eval. (2005) 1–9, papers2://publication/uuid/256DEBAD-ECC4-4DC4-99FD-E67E4D567AEC.
[12] C. Ge, R.A.D. Oliveira, I.Y.H. Gu, M.H.J. Bollen, Unsupervised deep learning and analysis of harmonic variation patterns using big data from multiple locations, Electr. Power Syst. Res. 194 (2021), 107042, https://doi.org/10.1016/j.epsr.2021.107042.
[13] M. Jasiński, T. Sikorski, K. Borkowski, Clustering as a tool to support the assessment of power quality in electrical power networks with distributed generation in the mining industry, Electr. Power Syst. Res. 166 (2019) 52–60, https://doi.org/10.1016/j.epsr.2018.09.020.
[14] J.J. López, J.A. Aguado, F. Martín, F. Muñoz, A. Rodríguez, J.E. Ruiz, Hopfield–K-means clustering algorithm: a proposal for the segmentation of electricity customers, Electr. Power Syst. Res. 81 (2011) 716–724, https://doi.org/10.1016/j.epsr.2010.10.036.
[15] D. Pinel, Clustering methods assessment for investment in zero emission neighborhoods' energy system, Int. J. Electr. Power Energy Syst. 121 (2020), 106088, https://doi.org/10.1016/j.ijepes.2020.106088.
[16] J. Mora-Flórez, J. Cormane-Angarita, G Ordóñez-Plata, k-means algorithm and mixture distributions for locating faults in power systems, Electr. Power Syst. Res. 79 (2009) 714–721, https://doi.org/10.1016/j.epsr.2008.10.011.
[17] D. Johnson, R.A. Wichern, Applied Multivariate Statistical Analysis, 6th ed., Prentice-Hall, New Jersey, 2007.
[18] AIAG, Measurement Systems Analysis: Reference Manual, 4th ed., Automotive Industry Action Group, Detroid, MI, USA, 2010.
[19] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32, https://doi.org/10.1023/A:1010933404324.
[20] F.A. De Almeida, S.C. Streitenberger, A.F. Torres, A.P. De Paiva, J.H.D.F. Gomes, A gage study through the weighting of latent variables under orthogonal rotation,

IEEE Access 8 (2020) 183557–183570, https://doi.org/10.1109/ACCESS.2020.3019031.
[21] Alvin C. Rencher, Methods of Multivariate Analysis, 2nd ed., John Wiley & Sons, Inc., New York, 2002 https://doi.org/10.1002/0471271357.
[22] F.A. Almeida, R.R. Leite, G.F. Gomes, J.H. de F. Gomes, A.P. de Paiva, Multivariate data quality assessment based on rotated factor scores and confidence ellipsoids, Decis. Support Syst. 129 (2020), 113173, https://doi.org/10.1016/j.dss.2019.113173.
[23] H.F. Kaiser, The varimax criterion for analytic rotation in factor analysis, Psychometrika 23 (1958) 187–200, 10.1007/BF02289233.
[24] L.L. Thurstone, Multiple-factor Analysis, Chicago Press, Chicago Univ, 1947, p. 535.
[25] S.A. Mingoti, Análise de dados através de métodos de estatística multivariada, Editora UFMG (2005).
[26] S. Sharma , Applied Multivariate Techniques, John Wiley & Sons, Inc., 1996.
[27] R.F. Ribeiro Junior, F.A. de Almeida, G.F. Gomes, Fault classification in three-phase motors based on vibration signal analysis and artificial neural networks, Neural Comput. Appl. (2020), https://doi.org/10.1007/s00521-020-04868-w.
[28] H. Chen, X. Liu, Z. Jia, Z. Liu, K. Shi, K. Cai, A combination strategy of random forest and back propagation network for variable selection in spectral calibration, Chemom. Intell. Lab. Syst. 182 (2018) 101–108, https://doi.org/10.1016/j.chemolab.2018.09.002.
[29] L.D. Simões, H.J.D. Costa, M.N.O. Aires, R.P. Medeiros, F.B. Costa, A.S. Bretas, A power transformer differential protection based on support vector machine and wavelet transform, Electr. Power Syst. Res. 197 (2021), 107297, https://doi.org/10.1016/j.epsr.2021.107297.
[30] I. Mitiche, G. Morison, A. Nesbitt, M. Hughes-Narborough, B.G. Stewart, P. Boreham, Classification of EMI discharge sources using time–frequency features and multi-class support vector machine, Electr. Power Syst. Res. 163 (2018) 261–269, https://doi.org/10.1016/j.epsr.2018.06.016.
[31] S. Zhang, D. Cheng, Z. Deng, M. Zong, X. Deng, A novel kNN algorithm with data-driven k parameter computation, Pattern Recognit. Lett. 109 (2018) 44–54, https://doi.org/10.1016/j.patrec.2017.09.036.
[32] S. Zhang, Nearest neighbor selection for iteratively kNN imputation, J. Syst. Softw. 85 (2012) 2541–2552, https://doi.org/10.1016/j.jss.2012.05.073.
[33] D.W. Hosmer, S. Lemeshow, R.X. Sturdivant, Applied Logistic Regression, Third Edit, Wiley, New Jersey, 2013, https://doi.org/10.1002/9781118548387.
[34] J.L. Fleiss, B. Levin, M.C. Paik, Statistical Methods for Rates and Proportions, Third Edit, John Wiley & Sons, Inc ., 2003. 10.1002/0471445428.
[35] D.E. Hinkle, W. Wiersma, S.G. Jurs, Applied Statistics for the Behavioral Sciences, Houghton Mifflin, Boston, 2002.
[36] M.W. Browne, An overview of analytic rotation in exploratory factor analysis, Multivariate Behav. Res. 36 (2001) 111–150, 10.1207/S15327906MBR3601_05.
[37] H.A. Sturges, The choice of a class interval, J. Am. Stat. Assoc. 21 (1926) 65–66, https://doi.org/10.1080/01621459.1926.10502161.
[38] D.C. Montgomery, Design and Analysis of Experiments, 9th ed., John Wiley & Sons, Inc., New York, 2017.