IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

A PCA-based consistency and sensitivity approach for assessing linkage methods in voltage sag studies

Fabricio Alves de Almeida[1,2], Luiz Gustavo de Mello[1], Estevão Luiz Romão[1], Guilherme Ferreira Gomes[3], José Henrique de Freitas Gomes[1], Anderson Paulo de Paiva[1], Jacques Miranda Filho[4], and Pedro Paulo Balestrassi[1]

[1]Institute of Industrial Engineering and Management, Federal University of Itajubá, Itajubá, Brazil
[2]Departament of Economics, Faculty of Economic Sciences Southern Minas, Itajubá, Brazil
[3]Mechanical Engineering Institute, Federal University of Itajubá, Itajubá, Brazil
[4]IFES Federal Institute of Espírito Santo, Vitória, Brazil

Corresponding author: Estevão L. Romão (e-mail: estevaoromao@unifei.edu.br).

This work was supported in part by CAPES, and in part by CNPq.

ABSTRACT In the light of Brazilian energy regulatory context, cluster strategies are required to classify groups of substations for voltage sag purposes. Tuning cluster algorithms is not a trivial task, due the fact that these methods are sensitive to small errors. Therefore, this study proposes a new methodology based on principal components analysis (PCA), attribute agreement and analysis of covariance to verify the level of consistency and sensitivity of the linkage methods in the cluster formation for voltage sag studies. In order to prove this methodology, real data from power quality indices of distribution substations are used. Four distinct scenarios with disturbances are evaluated. PCA is applied for dimensionality reduction of the data. Then, grouping is performed for eight different linkage methods and agreement analysis is applied. Ward method was the only one that presented 100% consistency in all scenarios, being as the most robust method while k-means showed consistency of 94.11%, with inversion of the clusters. However, when evaluating their groupings, it was found that k-means was unable to adequately separate the groups for this data set. Finally, the proposed methodology is adequate for choose cluster methods for extensive data and it can be extended to applications in different areas.

INDEX TERMS Substation cluster, voltage sag, principal components analysis, linkage methods, attribute agreement analysis.

I. INTRODUCTION

Quality improvements are widely studied in several power quality (PQ) sectors, where the quality of generation and distribution significantly influences industrial sectors [1]. Among the variables researched in PQ distribution, the voltage sag is characterized as a metric of great importance in these studies [2], as it directly influences losses in industrial processes with sensitive loads. From this, it is possible to verify that several studies focused on PQ, investigate the phenomenon of voltage sag applying different strategies, in which, we can highlight: the use of evolutionary algorithm to optimize the allocations of PQ monitors in distribution systems [3]; use of battery energy storage systems in the investigation of voltage sag and voltage deviation problems in distribution networks [4]; a new approach to asses equipment trip using fuzzy probabilities and possibility distribution in order to mitigate voltage sag [5]; simulations of different strategies to identify voltage sag sources [6]; the use of non-hierarchical linkage method of k-means for PQ event recognition [7]; and the use of convolutional neural networks with weighted k-nearest neighbor classifier for identification of voltage sag events [8]; a methodology which can be applied as a voltage sag mitigation solution to distribution of utility in a group of customers installing a dynamic voltage restorer [9]. These and other studies infer the importance of using modern strategies to investigate the phenomenon of voltage sag for the power quality distribution.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3088436, IEEE Access

Voltage sag studies are also applied to classify PQ based on the number of incidences and the influence of other variables in this phenomenon, where regulatory agencies map substations to assess the quality in power distribution. Among the studies, we can highlight the research by Miranda Filho *et al.* [10], in which he presented a proposal that combines the use of the principal components analysis (PCA) strategy and the Ward linkage method, creating substation groups to evaluate power quality based on voltage sag. Studies, such as this one, are needed in view of the new Brazilian context regulation and PQ control. Considering the variance-covariance structure of the data (common in large datasets), the authors used the PCA to model correlated data from its variance-covariance structure. In addition, this strategy aims to reduce data dimensionality and promoting a combination of non-correlated response vectors [11]. These combinations explain the original variables in a lean and appropriate way. In addition, the reduction of dimensionality promotes a decrease in computational effort, favoring the analysis of extensive data. The use of PCA is present in several studies that analyze PQ and also in applications in different segments, such as [12]–[17].

Another multivariate technique widely used for applications in the electrical sector is known as cluster analysis. This technique is characterized as a data mining strategy [18], which is especially applied in studies related to the electricity sector through the use of linkage methods [1], [10]. These methods are characterized as techniques for estimating patterns and clusters, in which they are widely used in the literature, such as: Jasiński et al. [19] used the k-means algorithm to analyze long-term PQ data in the mining industry; López *et al.* [20] uses the non-hierarchical k-means method merged with Hopfield's autonomous recurrent neural network to classify electricity utility customers (industrial, residential and administrative); Vinothkumar and Selvam [21], who applied a hierarchical clustering algorithm in the development of a new method for grid integration points identification of distributed generator units; The use of clustering is also investigated in Pinel [22], but in a context of designing energy system of zero emission neighborhood using. For this purpose, the authors used two distinct clustering methods, k-means and k-medoids, were evaluated and k-means presented better results in this specific application; We can also comment on the study by Ferreira *et al.* [23], in which they proposed a new method for clustering and pattern recognition of multivariate time series. The algorithm, which extracts the main features of the series, was applied in a real context in Brazil. These studies highlight the importance and wide applicability of cluster analysis aimed at the electricity sector and in PQ studies.

When exploring the cluster strategy in the literature, Fávero [24] highlight that this approach can be divided into two distinct groups, being the methods of hierarchical clusters (such as Average, Centroid, Complete, McQuitty, Median, Single and Ward) and the non-hierarchical methods (such as k-means). Usually, the choice of these methods is made in an arbitrary way, where the authors choose a method to work. However, it is possible to find sources of variations and errors when formulating clusters, since these techniques are sensitive to outliers [25]. In addition, Pinel [22] states that the best clustering method depends on the data set and application. In this sense, the configuration of the techniques used to generate clusters must be examined in a detailed and careful manner, evaluating the sensitivity and the consistency of their groupings. Johnson and Wichern [25] affirm that it is a good measure to apply several clusters methods allied to small perturbations (small errors) to the unit of data, in order to verify if there are inversions in the formation of clusters and to analyze the variability and agreement of the methods for a particular case.

In this sense, the most suitable technique for analyzing variability within and between systems is the repeatability and reproducibility (GR&R) study [26], [27]. This strategy is the most appropriate technique to evaluate the principles of variation of a measurement system [28]. This allows the numerical identification of the standard errors of highest significance. These variation principles are measured to verify that the variation presented in the measurement system is less than the true variation of the process [29], [30]. For the substation clusters evaluation, it is more appropriate to apply a GR&R by attributes, also known as Attribute Agreement Analysis (AAA), since the clusters are classified through memberships. There are some studies in the literature that apply the AAA technique to assess study variability in different applications, such as [31], [32]. It is known that attribute agreement analysis can evaluate the quality of a cluster method based on its degree of agreement in front of an established confidence interval.

Based on the studies previously analyzed, it is possible to verify that several studies that use the cluster approach in the electric sector do not present an adequate analysis for the choice of linkage methods, neglecting their instability in different scenarios. Thus, our study proposes a strategy that combines exploratory multivariate techniques and quality control statistics to analyze the consistency and sensitivity of the linkage methods in voltage sag studies. For that, we used data from distribution substations with similar features for voltage sags applied in Brazil. As suggested by Johnson and Wichern [25], 1.5% perturbations were applied to the original data for four replicates, whereas the data characteristics (variance-covariance structure) were preserved. Given the correlated nature of the data, the PCA strategy will be used and the scores of the principal components will be stored. After that, the clustering was performed for eight different linkage approaches (Average, Centroid, Complete, McQuitty, Median, Single, Ward and k-means) and storing their respective memberships. Then, the attribute agreement study was performed to verify the variability in hierarchical and non-hierarchical methods. In addition, the degree of sensitivity of the results were evaluated based on the main voltage sag variables. To the authors best knowledge, there's no study that compares different linkage methods, assessing the level of agreement and sensitivity analysis in cluster formation for voltage sag

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3088436, IEEE Access

studies. The proposed method stands out as an important alternative for the evaluation of substation classifications, given the new Brazilian context of power quality regulation.

This paper is organized as follows: A theoretical background is presented in section 2, describing all the techniques used in this study. Section 3 presents the materials and methods used. Section 4 describes the application for the data set from distribution substations in southeastern Brazil, detailing the design, steps and the results. Finally, section 5 presents the study's conclusions.

II. THEORETICAL BACKGROUND

A. PRINCIPAL COMPONENTS ANALYSIS

PCA is a multivariate analysis technique used to find a combination of uncorrelated variables that adequately explains the original variables [33]. Considering the random vector $\boldsymbol{X^T}=[X_1, X_2, \ldots, X_p]$ that has the covariance matrix $\boldsymbol{\Sigma}$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_p \geq 0$. Then, the linear combinations can be described as in Eq. (1).

$$\begin{aligned} Y_1 &= a_1^T X = a_{11}X_1 + a_{21}X_2 + \ldots + a_{p1}X_p \\ Y_2 &= a_2^T X = a_{12}X_1 + a_{22}X_2 + \ldots + a_{p2}X_p \\ &\bullet\bullet\bullet \\ Y_p &= a_p^T X = a_{1p}X_1 + a_{2p}X_2 + \ldots + a_{pp}X_p \end{aligned} \quad (1)$$

According to Johnson and Wichern [25], these linear combinations can replace the original variables by reducing the dimensionality of the problem. If Y_i is the i^{th} principal component, then (Eqs. (2) and (3)).

$$Var(Y_i) = a_i^T \sum a_i = e_i^T \sum e_i; \qquad \forall i = 1, 2, \ldots, p \quad (2)$$

$$CoVar(Y_i, Y_k) = a_i^T \sum a_k = e_i^T \sum e_k; \qquad \forall i, k = 1, 2, \ldots, p \quad (3)$$

The principal components are those uncorrelated linear combinations $Y_1, Y_2, \ldots, Y_p$ whose variances in Eq. (2) are as large as possible. The first principal component is the linear combination with maximum variance. That is, it maximizes $Var(Y_1) = a_1^T \sum a_1$. It is clear that $Var(Y_1) = a_1^T \sum a_1$ can be increased by multiplying any a_1 by some constant. To eliminate this indeterminacy, it is convenient to restrict attention to coefficient vectors of unit length. Therefore, the first principal component in Eq. (4),

$$\mathbf{PC}_1 = \begin{cases} Maximize: \ Var(Y_1) \\ Subject\ to: \ a_1^T a_1 = 1 \end{cases} \quad (4)$$

Similarly (for Eq. (5)),

$$\mathbf{PC}_2 = \begin{cases} Maximize: \ Var(Y_2) \\ Subject\ to: \ a_2^T a_2 = 1 \\ \qquad\qquad CoVar(Y_1, Y_2) = 0 \end{cases} \quad (5)$$

B. HIERARCHICAL CUSTER ANALYSIS

Hierarchical Cluster Analysis (HCA) consists in agglomeration techniques whose purpose is to group objects with a certain similarity level. These grouping methods start with single objects, which means that initially the number of clusters is the same as the total number of objects. Next, similar objects form groups, which are merged according to their similarities. Inasmuch as the similarity is reduced, the subgroups tend to form a single cluster [25]. In this context, different linkage methods are presented.

1) SINGLE LINKAGE METHOD

Single linkage is a hierarchical method where groups are formed by merging individual entities considering the largest similarity, i.e., a smaller separation [25]. Let A and B be two different clusters and $\mathbf{y}_i$ and $\mathbf{y}_j$ be the observations vectors for A and B, respectively. Considering n observations (i = 1, 2, …, n and j = 1, 2, …, n), the purpose is to minimize the distance between them as in Eq. (6).

$$D_{single}(A,B) = \min\left\{d\left(\mathbf{y}_i, \mathbf{y}_j\right)\right\} \quad (6)$$

2) COMPLETE LINKAGE METHOD

The complete linkage method is also known as the farthest neighbor method. Differently from the single linkage, it seeks to maximize the distance of two objects between clusters [34]. Again, let A and B be two different clusters, the complete linkage method separates the most distant objects [35], as shown in Eq. (7).

$$D_{complete}(A,B) = \max\left\{d\left(\mathbf{y}_i, \mathbf{y}_j\right)\right\} \quad (7)$$

3) AVERAGE LINKAGE METHOD

The average linkage method defines the distance between two clusters as the average distance between all pairs of objects, where each item belongs to a specific cluster [25], [35]. According to Rencher [35], the distance between two clusters can be expressed as shown in Eq. (8), where n_A and n_B, represent the number of objects in cluster A and B, respectively.

$$D_{average}(A,B) = \frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d\left(\mathbf{y}_i, \mathbf{y}_j\right) \quad (8)$$

4) CENTROID LINKAGE METHOD

The centroid of a cluster is defined as its center of mass and the distance between the clusters' centroids defines the similarity between them [36]. Therefore, let A and B be two different clusters, the distance between them depends on the Euclidean distance between their centroids, which means, the average vectors $\bar{\mathbf{y}}_A$ and $\bar{\mathbf{y}}_B$, respectively. This formulation is indicated in Eqs. (9) and (10) presenting the weighted average, which calculates the centroid of the new cluster AB.

$$D_{centroid}(A,B) = d(\bar{\mathbf{y}}_A, \bar{\mathbf{y}}_B) = \sqrt{(\bar{\mathbf{y}}_A - \bar{\mathbf{y}}_B)^T (\bar{\mathbf{y}}_A - \bar{\mathbf{y}}_B)} \quad (9)$$

$$\bar{\mathbf{y}}_{AB} = \frac{n_A \bar{\mathbf{y}}_A + n_B \bar{\mathbf{y}}_B}{n_A + n_B} \quad (10)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3088436, IEEE Access

where $\bar{\mathbf{y}}_A = \sum_{i=1}^{n_A} \mathbf{y}_i / n_A$ and $\bar{\mathbf{y}}_B = \sum_{i=1}^{n_B} \mathbf{y}_i / n_B$.

5) MEDIAN LINKAGE METHOD

The median linkage method calculates the median distance between the elements of different groups and Eq. (11) shows how the distance matrix is obtained. The variables D_{mj}, D_{kj}, D_{lj} and D_{kl} are defined as the distances between the clusters m and j; k and j; l and j; and k and l, respectively. m represents the merged group consisting of the clusters k and l, with $m = (k, i)$.

$$D_{median_{mj}} = \frac{D_{kj} + D_{lj}}{2} - \frac{D_{kl}}{4} \tag{11}$$

6) MCQUITTY LINKAGE METHOD

In the McQuitty linkage, the distance between a cluster AB and a given cluster C is calculated in Eq. (12) [37].

$$D_{mcquitty_{(AB-C)}} = \frac{D_{AC} + D_{BC}}{2} \tag{12}$$

7) WARD LINKAGE METHOD

Ward linkage method merges two distinct clusters to minimize the loss of information, which is described as an increase in the error sum of squares criterion (ESS). Grouping the clusters into a specific group of n variables, ESS can be described as shown in Eq. (13) [38].

$$ESS_i = \sum_{j=1}^{n_i} (X_{ij} - \overline{X_i})' (X_{ij} - \overline{X_i}) \tag{13}$$

where $\overline{X_i}$ is the mean of the objects and X_{ij} is the multivariate measurement associated with the j^{th} object. A deeper explanation about this method can be found in Ward [38].

C. NON-HIERARCHICAL CUSTER ANALYSIS

Non-hierarchical cluster techniques aim to group the items into a certain number (k) of clusters. An important consideration for these methods is that the number of final groups must be determined before starting the clustering procedure [39]. Furthermore, these techniques may be applied in situations where the collection of data are considerably large, since there is no need to calculate distance matrices nor to store basic data during the computer run. One of the most popular technique in this context is the k-means method [25].

The algorithm described by the k-means method assigns to each item of a cluster a nearest mean [10]. According to Johnson and Wichern [25], the simplest version of this procedure consists on the three main steps listed below:

--Initially, we partition the items in k distinct clusters and we calculate the coordinates of the clusters' centroids.

--Next, we assign an item to the cluster whose centroid is the nearest. For that, we compute the Euclidian distance. It is necessary to recalculate the centroids of the clusters that receive and loose an item.

--Finally, we perform the second step until there are no more reassignments to be done.

C. ATTRIBUTE AGREEMENT ANALYSIS

Attribute agreement analysis is a recommended strategy to analyze the variability presented by discrete variables. It is a statistical strategy used to verify if the appraisers present consistency among themselves and among the standards previously known. This technique allows reducing or eliminating the subjectivity in the analysis, e.g., in the classification of substations clusters. In order to create this study, it is necessary to define the number of samples, appraisers and replicates to be analyzed. After that, specific hypothesis tests are used to define the variability/agreement as Kappa statistics and Kendall coefficients. These statistics describe the agreement level existing among different classifications calculated from Fleiss strategy [40], [41].

Kappa index shows the ratio between the proportions that the appraisers agree and the maximum proportions that they could agree. Eq. (14) indicates the agreement degree of the classifications performed by several appraisers analyzing the same responses.

$$K = \left(\frac{P_o - P_e}{1 - P_e} \right) = \frac{\left[\frac{1}{N_k n_k (n_k - 1)} \left(\sum_{i=1}^{k} \sum_{j=1}^{k} x_{ij} - N_k n_k \right) \right] - \sum_{j=1}^{l} p_j^2}{\left(1 - \sum_{j=1}^{l} p_j^2 \right)} \tag{14}$$

Where P_o is the observed agreement mean proportion; P_e is the expected agreement mean proportion. p_j^2 is the expected agreement proportion for each category; N_k and n_k represent the number of evaluated items and the number of appraisers, respectively; k is the number of categories of the adopted scale. Finally, x_{ij} represents the number of appraisers that classified the i^{th} item as belonging to the j^{th} category.

Analyzing the Kendall coefficient in Eq. (15), it is possible to verify that it is more adequate to perform ordinal data analysis, being used as classifications, in [43], or in Likert scale as in [44]. Hence, Eq. (15) measures the agreement level between the appraisers, for both appraisers between and within.

$$W = \frac{12 \sum_{i=1}^{n} R_i^2 - 3p^2 n(n+1)^2}{p^2 (n^3 - n) - p \left(\sum_{k=1}^{m} (t_k^3 - t_k) \right)} \tag{15}$$

where n is the number of items, R_i^2 is the sum of squares for the classification sums R_i, p refers to the number of appraisers, and t_k is the number of tied classifications in each one of the m groups of ties.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3088436, IEEE Access

Table 1 summarizes the agreement indexes explored in this paper, and Table 2 shows the acceptability levels of agreement according to [42], [46].

TABLE I
AGREEMENT INDEXES

AGREEMENT INDEX	RANGE	INTERPRETATION
Kappa statistic (*K*)	-1 to 1	*K = 1* means a perfect agreement
		K = 0 means that agreement is the same that expected by chance
		K = -1 means that agreement is less than expected by chance
Kendall's coefficient of concordance (*W*)	0 to 1	*W = 1* indicates a perfect association
		W = 0 indicates that there is no association

TABLE II
ACCEPTABILITY LEVELS OF AGREEMENT

ACCEPTABILITY	KAPPA STATISTIC VALUE (K)	KENDALL'S COEFFICIENTS VALUES (*W* AND τ)
Poor	$K < 0.40$	$W < 0.30$
Good	$0.75 < K < 0.90$	$0.70 < W < 0.90$
Excellent	$K > 0.90$	$W > 0.90$

III. MATERIALS AND METHODS

A. POWER QUALITY INDICATORS OF SUBSTATIONS OCATED IN SOUTHEASTERN BRAZIL

A real example of network modeling and fault simulations in transmission and distribution levels is performed aiming to validate the proposal of the present paper. We considered an electricity distribution system consisting of 17 substations whose total area is about 41,241 km^2, around 90% of the state where they are located. All these substations are located in southeastern Brazil and can be viewed, geographically, in Fig. 1. The data, also available in [10], were obtained in a 30 month duration research and development project managed by EDP ES Distribution Utility, an electricity distribution company in partnership with the Federal University of Itajubá (UNIFEI).

In this context, voltage sags are caused by occurrences of lightning and short-circuit, since most of the overhead lines and feeders are not covered. The voltage-rated feeders, the length of the distribution lines, and the fault statistics (faults/100 km/year) that were used in the short-circuit simulations causing the analyzed sags can be viewed in Table 3, where VL, L and FR represent the voltage level, the length and the failure rate, respectively. The power quality monitors recorded the events that were collected in the secondary of power transformers. It was possible to obtain a distribution line equivalent to 13.8kV from the total long-term rates (greater than or equal to 3 minutes) and short duration of the statistics, in which medium voltage failure rate was used. We can also verify the existence of three-phase type with a lower incidence, whereas a single phase has a higher incidence, among the occurrences. Finally, just as in [10], we considered a normal distribution to estimate the transmission and sub transmission systems whose mean (μ) was equal to 5 Ω and whose standard deviation (σ) was equal to 1 Ω. We also worked with a uniform distribution ranging from zero to one for the distribution network, where a maximum value of 30 Ω was assigned for 1LG faults, 30 Ω for 2LG faults and 10 Ω for 3LG faults. Further details about the substation buses being analyzed are available in [10]. For each substation considered in the present study, 32 design characteristics and power qualities were considered. The main quality variable is the TNE, which is obtained through simulation, whereas the monitored number of events (MNE) is obtained by monitoring. Tables from 4 to 7 show the values of the 32 variables used in this study, and for more details on the data collection, readers should consult [10]. PQ measurements were collected over one year so that it was possible to cover different seasonality that has influence on the Distribution Network performance, such as rainfall, winds, etc. In addition, 30 Schweitzer Engineering Laboratories PQ meters were used to acquire these data by model SEL 734. The nomenclature of all variables used can be found in Appendix A, in Table A.1. We highlight that in this project we used software such as *Minitab18*® and *R Studio*® for statistical purposes and other developments.

TABLE III
LENGTH OF DISTRIBUTION LINES AND FAULT STATISTICS USED IN SHORT-CIRCUIT SIMULATIONS

VL [*kV*]	L [*km*]	FR	1LG [%]	2LG [%]	LL [%]	LLL [%]
138	2125.5	2.33	75	13	10	2
69	1033	6.34	58	25	11	2
34.5	619	43.13	70	15	10	5
13.8	22.750	MVFR	78	10	9	3

B. SUBSTATION CLUSTERING METHOD BASED ON MULTIVARIATE TECHNIQUE

Characterized as one of the most worrying PQ variables in sensitive industrial loads, voltage sags are widely analyzed in studies that make use of hierarchical and non-hierarchical methods of clustering applied to this topic (as highlighted in section 1). Thus, it appears that several authors analyze one, or few, linkage methods without checking the degree of agreement. In other words, without checking if the method is stable or even robust for the data set under analysis. In their study, Miranda Filho *et al.* [10] briefly discuss the behavior of a few hierarchical methods and the non-hierarchical methods, k-means, for voltage sag studies. However, the authors do not analyze, or demonstrate, the behavior of all methods. In addition, Johnson and Wichern [25] infer that there are sources of errors and variations when analyzing the clusters from these strategies and that, when discussing cluster methods, one should verify the behavior of the methods in the face of minimal disturbance scenarios. Thus, this work proposes a methodology, of a multivariate character, aiming to find the best choice among the linkage methods for voltage sag studies of PQ distributions data. Therefore, it is possible to perform a sensitivity and variability check on the formation and degree of agreement from small disturbances for each cluster. Based on these

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3088436, IEEE Access

analyzes, confidence intervals can be generated. The proposed methodology is illustrated in Fig. 1.

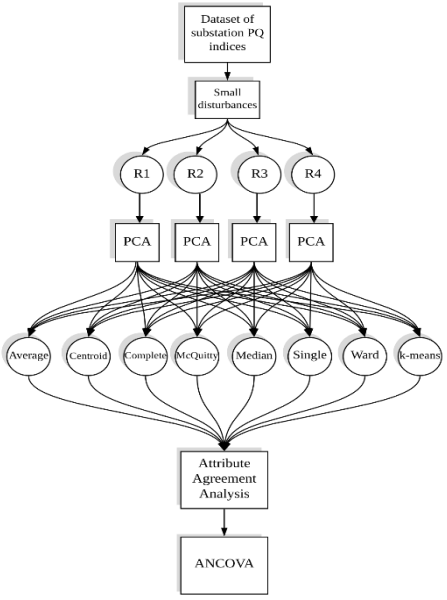


FIGURE 1. Flowchart of the proposed methodology.

--*Step 1*: From the main variables to carry out a voltage sag study, different scenarios should be generated by applying small disturbances to the data set (in the range of 1.5%). Based on this, four distinct replicas should be generated at random. In this case study, real data from the Brazilian Power Distribution Company described in Tables 4 to 7 will be used, containing a total of 544 data. It is important to emphasize that the replicates must maintain a significant variance-covariance structure, according to the original data set;

--*Step 2*: After performing the four sets of random replicas with minor disturbances, the PCA multivariate technique should be applied, storing the components scores, for the following purposes: i) reduce the dimensionality of the data (minimizing computational effort); ii) generate scores for non-correlated components that adequately represent the variance-covariance structure of the data. The use of this strategy is necessary, since analyzing such data with a univariate technique, may present inadequate results [27];

--*Step 3*: In view of the PCA application and the components scores of the different scenarios, clusters should be performed using each of the linkage methods listed above, hierarchical and non-hierarchical, namely: Average, Centroid, Complete, McQuitty, Median, Single, Ward and k-means, respectively. For each method, the memberships must be stored, totaling eight response vectors from different voltage sag scenarios;

--*Step 4*: Based on the previous analysis, it is possible to verify the behavior and robustness of the clusters in each linkage method. For this, the strategy called Attribute Agreement Analysis will be used (or GR&R by attributes). Along with the other techniques, the analysis by attributes allows to verify the consistency of the groupings, even in scenario with small disturbances. Furthermore, the method allows to verify, analytically, which linkage methods performed better, with lower incidence of cluster inversion;

--*Step 5*: Finally, it is possible to check the sensitivity of the results for the voltage sag data set. For this, the confidence interval of the clusters must be calculated and evaluated from the analysis of covariance (ANCOVA). It is possible to infer the best linkage method for a given data set.

TABLE IV
POWER QUALITY INDICES OF SUBSTATION (PART I)

SUBSTATION	HVFR[a]	TNE	NEMV	MVFR	MNE	LNE	ANE	UNE
Aracruz	2.33	210.8953	179.60	235	79	47	61.22	82
Baixo Guandu	2.33	131.93174	105.70	126	72	28	43.69	64
Barra Sahy	2.33	196.8775	174.70	404	77	53	76.7	104
Ecoporanga	2.33	198.81387	163.27	90	216	94	117.38	142
Itarana	2.33	322.34765	293.25	192	172	92	114.3	132
Jaguaré	2.33	249.6614	214.14	206	83	58	88.81	109
João Neiva	2.33	426.1293	394.61	212	144	88	114.02	134
Juncado	2.33	498.07767	466.78	184	269	138	171.59	214
Linhares A	2.33	543.97146	513.89	279	165	139	169.45	197
Linhares C	2.33	292.0779	261.28	474	149	73	96.79	128
Montanha	2.33	113.1813	77.98	100	303	125	159.5	183
Nova Venecia	2.33	51.20909	13.84	174	118	95	123.79	156
Paulista	2.33	149.87619	114.02	97	176	71	93.52	187
Pinheiros	2.33	314.99604	280.13	148	244	107	134.94	154
Santa Tereza	2.33	289.72731	259.65	203	314	86	109.26	129
São Francisco	2.33	164.39265	129.33	159	184	135	164.92	200
Suiça	2.33	177.98019	153.63	83	261	38	56.83	76

[a]Constant variable (not simulated)

SOURCE: MIRANDA E FILHO ET AL. [8]

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3088436, IEEE Access

TABLE V
POWER QUALITY INDICES OF SUBSTATION (PART II)

SUBSTATION	FKVAr	SAIFI1	SAIFI2	STIFI	FL	AREA	EMVVA	EVAHV
Aracruz	6900	5.567	6.531	1130	365.22	555.414	76.42	1343.24
Baixo Guandu	5700	5.070	7.392	1136	885.65	1080.046	83.89	1125.83
Barra Sahy	4800	9.592	12.531	1014	164.74	255.178	43.24	951.75
Ecoporanga	4500	6.163	9.481	941	1086.75	1499.906	181.41	1525.50
Itarana	12000	10.329	11.885	2672	1685.29	1303.753	152.74	1248.75
Jaguaré	7500	6.490	7.317	1561	863.08	1254.651	103.95	1524.50
João Neiva	6600	10.575	12.225	2029	903.17	950.368	186.14	1352.75
Juncado	9000	14.668	16.040	1298	625.00	749.424	253.68	1343.25
Linhares A	24300	10.538	11.677	3117	1800.09	2057.083	184.19	1290.89
Linhares C	9300	10.538	11.677	2814	323.23	537.861	55.12	1321.75
Montanha	6900	10.253	11.166	891	853.20	1242.832	77.98	1511.00
Nova Venecia	10500	8.695	11.118	2392	1892.82	1335.288	7.95	1603.90
Paulista	3600	10.062	12.781	193	244.11	205.801	117.54	1539.00
Pinheiros	9900	10.253	11.166	1477	1173.21	1307.187	189.28	1496.50
Santa Tereza	4500	13.109	15.398	1899	1136.70	789.833	127.91	1290.75
São Francisco	9000	8.135	10.796	2006	1441.63	977.427	81.34	1504.70
Suiça	1200	13.109	15.398	521	536.59	231.392	185.09	1045.25

SOURCE: MIRANDA E FILHO ET AL. [8]

IV. CONSISTENCY AND SENSITIVITY ANALYSIS BASED ON PCA FOR ASSESSING LINKAGE METHODS IN VOLTAGE SAG STUDIES

In view of the real data of design features and power quality indices from distribution substations conducted by EDP ES Distribution Utility, we will apply the proposed method to find the best and most robust linkage method for generating clusters. Considering the total number of data (32 variables of 17 substations), initially the data will be disturbed in the range of 1.5%, as suggested by Johnson and Wichern [25]. It is important to note that the replicas with disturbances were randomly generated. In addition, a significant degree of data structure was maintained. In this way, a total of 2,176 data were generated and, due to the large extension of the sets, they are available in the supplementary material.

Considering the correlated nature of the data structure, in step 2 it is necessary to apply the multivariate strategy of principal component analysis. For the extraction of component scores, the Kaiser criterion was considered, where the percentage of explanation must be greater than or equal to 80% [27], [47]. Fig. 2 presents the statistical explanation for replica 1 (R1) of each component, eigenvalue and percentage of explanation. The first seven principal components (PC) have explanation values greater than or equal to 92.9% for the replicas R1, R2, R3 and R4, respectively. Thus, the choice of seven components is ideal to represent the original variables with adequate statistical validation. Table 8 presents the component scores for R1.

TABLE VI
POWER QUALITY INDICES OF SUBSTATION (PART III)

SUBSTATION	FKVAr	SAIFI1	SAIFI2	STIFI	FL	AREA	EMVVA	EVAHV
Aracruz	6900	5.567	6.531	1130	365.22	555.414	76.42	1343.24
Baixo Guandu	5700	5.070	7.392	1136	885.65	1080.046	83.89	1125.83
Barra Sahy	4800	9.592	12.531	1014	164.74	255.178	43.24	951.75
Ecoporanga	4500	6.163	9.481	941	1086.75	1499.906	181.41	1525.50
Itarana	12000	10.329	11.885	2672	1685.29	1303.753	152.74	1248.75
Jaguaré	7500	6.490	7.317	1561	863.08	1254.651	103.95	1524.50
João Neiva	6600	10.575	12.225	2029	903.17	950.368	186.14	1352.75
Juncado	9000	14.668	16.040	1298	625.00	749.424	253.68	1343.25
Linhares A	24300	10.538	11.677	3117	1800.09	2057.083	184.19	1290.89
Linhares C	9300	10.538	11.677	2814	323.23	537.861	55.12	1321.75
Montanha	6900	10.253	11.166	891	853.20	1242.832	77.98	1511.00
Nova Venecia	10500	8.695	11.118	2392	1892.82	1335.288	7.95	1603.90
Paulista	3600	10.062	12.781	193	244.11	205.801	117.54	1539.00
Pinheiros	9900	10.253	11.166	1477	1173.21	1307.187	189.28	1496.50
Santa Tereza	4500	13.109	15.398	1899	1136.70	789.833	127.91	1290.75
São Francisco	9000	8.135	10.796	2006	1441.63	977.427	81.34	1504.70
Suiça	1200	13.109	15.398	521	536.59	231.392	185.09	1045.25

SOURCE: MIRANDA E FILHO ET AL. [8]

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3088436, IEEE Access

TABLE VII
POWER QUALITY INDICES OF SUBSTATION (PART IV)

SUBSTATION	R+	X+	Xo	ZBASE	Zohm	Zpu	MVASC	BMVAr
Aracruz	6900	5.567	6.531	1130	365.22	555.414	76.42	1343.24
Baixo Guandu	5700	5.070	7.392	1136	885.65	1080.046	83.89	1125.83
Barra Sahy	4800	9.592	12.531	1014	164.74	255.178	43.24	951.75
Ecoporanga	4500	6.163	9.481	941	1086.75	1499.906	181.41	1525.50
Itarana	12000	10.329	11.885	2672	1685.29	1303.753	152.74	1248.75
Jaguaré	7500	6.490	7.317	1561	863.08	1254.651	103.95	1524.50
João Neiva	6600	10.575	12.225	2029	903.17	950.368	186.14	1352.75
Juncado	9000	14.668	16.040	1298	625.00	749.424	253.68	1343.25
Linhares A	24300	10.538	11.677	3117	1800.09	2057.083	184.19	1290.89
Linhares C	9300	10.538	11.677	2814	323.23	537.861	55.12	1321.75
Montanha	6900	10.253	11.166	891	853.20	1242.832	77.98	1511.00
Nova Venecia	10500	8.695	11.118	2392	1892.82	1335.288	7.95	1603.90
Paulista	3600	10.062	12.781	193	244.11	205.801	117.54	1539.00
Pinheiros	9900	10.253	11.166	1477	1173.21	1307.187	189.28	1496.50
Santa Tereza	4500	13.109	15.398	1899	1136.70	789.833	127.91	1290.75
São Francisco	9000	8.135	10.796	2006	1441.63	977.427	81.34	1504.70
Suiça	1200	13.109	15.398	521	536.59	231.392	185.09	1045.25

SOURCE: MIRANDA E FILHO ET AL. [8]

From the non-correlated and dimensionless scores (Table 8), it is possible to plan and apply hierarchical and non-hierarchical clustering methods. The ideal number of clusters was defined by the group categorization rule, the quantity (k_c) being defined by $k_c = 1+3.32log(N)$, where N is the number of objects. Thus, the number of clusters for the study was defined as 5. Regarding the hierarchical methods, the seven most used linkage methods were considered: Ward, Average, Centroid, Complete, McQuitty, Median and Single. All hierarchical applications used the Euclidean distance, in which the square root of the sum of the square differences is calculated.

In addition, the non-hierarchical method of k-means was also applied. For each of these applications, memberships are stored.

Regarding the results, it is possible to check the behavior of the clusters in the replicas, analyzing the consistency and finding the best method. Given the information stored, we can assess the variability of the clusters from the attribute agreement analysis (*step 4*). The planning of the concordance analysis for this study can be defined as the 17 substations (number of samples), 8 linkage methods (as number of appraisers) and the 4 disturbance scenarios, with different data sets (number of replicates). Thus, we have 544 data combinations for variables containing design features and power quality indices. The memberships for this experimental matrix are described in Tables 9 and 10.

Applying this strategy with a 95% confidence interval (CI), we can analyze the degree of precision in which the methods group the substations for PQ through the Fleiss' Kappa statistics and Kendall's Coefficient of Concordance (detailed in section 2.4). Initially evaluating the agreement within appraisers (or repeatability), it is possible to verify through the Table 11 that the only method that showed 100% consistency for all clusters was the Ward method. This linkage method did not show cluster inversion in any of the four scenarios with disturbance. This level of agreement can be validated through Tables 10 and 11, where all clusters (and the Overall assessment) had a Kappa and Kendall index equal to 1, inferring an excellent level of agreement according to the AIAG criteria (Table 2). Then, the non-hierarchical method (k-means) showed the second-best behavior, with 94.12% agreement, with a confidence interval between 71.31% and 99.98%. Evaluating the Kappa statistic, it appears that the k-means method presented an overall agreement of 95.81%, obtaining an inversion of clusters between clusters 2 and 3 for the Linhares C substation (with original TNE of 292 - faults on medium voltage and high voltage). Thus, from Kendall's Coefficient of Concordance, it appears that k-means presented a value equal to 0.9984, which is an excellent level of agreement.

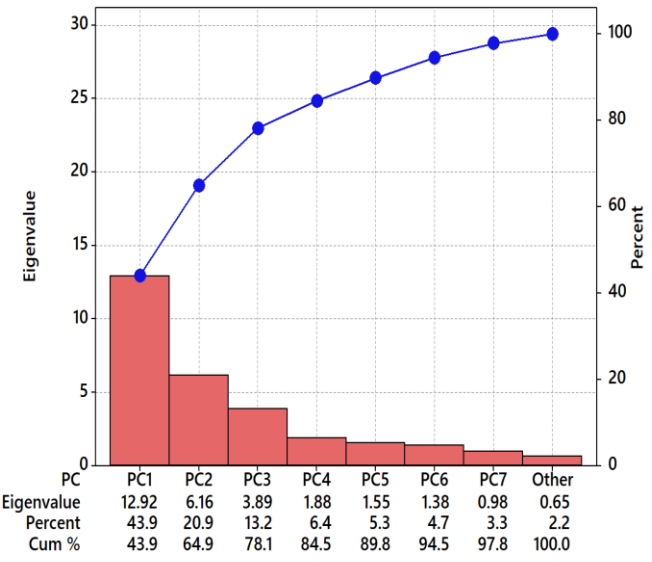


FIGURE 2. **Pareto chart and number of principal components for R1**

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3088436, IEEE Access

TABLE VIII
PRINCIPAL COMPONENT SCORES FOR R1

PC1	PC2	PC3	PC4	PC5	PC6	PC7
-2.0517	-3.1187	-1.0727	1.1744	1.4562	-0.8041	0.3894
2.3045	-3.9988	-0.9081	1.2495	-0.5337	-1.3051	0.0667
2.7694	-3.5133	2.1727	-1.6681	0.7160	0.4507	2.1251
-6.4086	0.2667	-2.2395	2.2523	0.8561	0.5533	0.6340
1.9546	1.5594	0.5521	0.3833	-1.2904	-1.8531	-0.1178
3.9800	-1.2948	-2.0907	0.8286	0.6073	-0.2134	-1.6660
2.9093	0.2906	1.9104	2.4158	-1.0711	3.1196	-0.6129
-2.0797	3.8062	2.6560	-0.4301	2.4489	0.5436	-0.0778
4.3743	4.3973	0.6552	1.4933	0.1112	-1.0118	1.5910
4.8217	-0.8104	0.6612	-1.3263	2.0056	-0.1697	-1.4840
-4.6583	1.6130	-1.6229	-1.1980	-0.7258	-0.0859	-0.2351
3.5950	0.3832	-3.4473	-1.8142	-1.4573	1.5951	0.9624
-3.7211	-1.4335	-0.2222	-1.2687	0.0499	1.0049	-0.6805
-0.2164	2.2281	-0.2329	-0.1688	-0.9198	-0.5652	-0.4360
-1.6639	0.9285	2.0550	-0.9802	-1.5239	-0.8240	-0.9882
-1.3938	1.6169	-2.1300	-1.0342	0.6074	-0.1364	0.3416
-4.5151	-2.9204	3.3039	0.0912	-1.3364	-0.2984	0.1879

Considering the methods that showed a good, or even adequate, behavior in view of the acceptance criteria established by AIAG [42], we can verify another hierarchical method, the Centroid. This method showed an 82.35% agreement in view of the disturbance scenarios, providing total consistency only for clusters 4 and 5. When evaluating the Fleiss 'Kappa statistics and Kendall's Coefficient of Concordance (Table 12 and 13), it appears that the method had an overall value of 0.7801 and 0.9054, respectively. Such results infer that, according to the Kappa criterion, the method presents a good quality of agreement (but not excellent), while the Kendall's coefficient infers a high level of agreement. Evaluating in detail, we have that Clusters 1 had a consistency of 83.65%, with inversions linked to the Barra Sahy substation, whereas clusters 2 and 3 presented inversions for the Ecoporanga and Juncado substations (46.85% consistency for both clusters).

Regarding the Single method, it appears that it did not show total consistency in any of the five clusters, with inversions in the substations: Ecoporanga, João Neiva, Juncado and Linhares A. However, its overall agreement percentage is 64.71%. The other methods analyzed (Average, Complete, McQuitty and Median) obtained a degree of agreement below 50%, showing great instability, with inversions in several clusters and between replicates. Fig. 3 depicts the degree of agreement and confidence intervals (95%) for all linkage methods used in the study. Further details on the results of the Kappa and Kendall metrics are available in Tables 12 and 13. The assessment agreement between appraisers (or reproducibility) shows an agreement of 5.88%, with Kappa and Kendall values of 0.2442 and 0.3934, respectively. This result implies that the linkage methods are not in agreement, validating the need to find a method that is robust for each type of data set.

TABLE IX
MEMBERSHIPS OF THE CLUSTERS FORMED BY THE LINKAGE METHODS (PART I)

SAMPLE	AVERAGE				CENTROID				COMPLETE				MCQUITTY			
	REPLICA				REPLICA				REPLICA				REPLICA			
	R1	R2	R3	R4	R1	R2	R3	R4	R1	R2	R3	R4	R1	R2	R3	R4
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	2	2	2	1	1	1	1	2	2	2	2	1	2	2	2
3	1	2	2	2	2	1	1	1	2	2	2	2	1	2	2	2
4	2	1	1	1	3	2	2	2	3	1	1	1	2	1	1	1
5	3	1	1	1	1	1	1	1	4	3	3	3	3	3	3	3
6	1	2	2	2	1	1	1	1	5	2	2	2	1	2	2	2
7	1	3	3	3	1	1	1	1	5	4	4	4	4	4	4	4
8	3	1	1	1	1	3	3	3	4	3	3	3	3	3	3	3
9	4	4	4	4	4	4	4	4	5	4	4	4	4	5	5	5
10	1	2	2	2	1	1	1	1	5	2	2	2	1	2	2	2
11	3	1	1	1	1	1	1	1	3	1	1	1	3	1	1	1
12	1	2	2	2	1	1	1	1	5	2	2	2	1	2	2	2
13	3	1	1	1	1	1	1	1	1	1	1	1	3	1	1	1
14	3	1	1	1	1	1	1	1	4	3	3	3	3	3	3	3
15	3	1	1	1	1	1	1	1	4	3	3	3	3	3	3	3
16	3	1	1	1	1	1	1	1	3	1	1	1	3	1	1	1
17	5	5	5	5	5	5	5	5	1	5	5	5	5	1	1	1

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3088436, IEEE Access

TABLE X
MEMBERSHIPS OF THE CLUSTERS FORMED BY THE LINKAGE METHODS (PART II)

SAMPLE	MEDIAN				SINGLE				WARD				K-MEANS			
	REPLICA				REPLICA				REPLICA				REPLICA			
	R1	R2	R3	R4	R1	R2	R3	R4	R1	R2	R3	R4	R1	R2	R3	R4
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	2	2	2	1	1	1	1	2	2	2	2	2	2	2	2
3	1	2	2	2	2	2	1	2	2	2	2	2	3	3	3	3
4	2	1	1	1	1	1	1	1	3	3	3	3	4	4	4	4
5	1	1	1	1	1	1	1	1	4	4	4	4	5	5	5	5
6	1	2	2	2	1	1	1	1	2	2	2	2	2	2	2	2
7	1	3	3	3	3	3	2	3	5	5	5	5	5	5	5	5
8	3	4	4	4	4	4	3	4	4	4	4	4	5	5	5	5
9	4	5	5	5	1	1	4	1	5	5	5	5	5	5	5	5
10	1	2	2	2	1	1	1	1	2	2	2	2	3	2	2	2
11	1	1	1	1	1	1	1	1	3	3	3	3	4	4	4	4
12	1	2	2	2	5	1	1	1	2	2	2	2	5	5	5	5
13	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
14	1	1	1	1	1	1	1	1	4	4	4	4	5	5	5	5
15	1	1	1	1	1	1	1	1	4	4	4	4	5	5	5	5
16	1	1	1	1	1	1	1	1	3	3	3	3	5	5	5	5
17	5	1	1	1	1	5	5	5	1	1	1	1	1	1	1	1

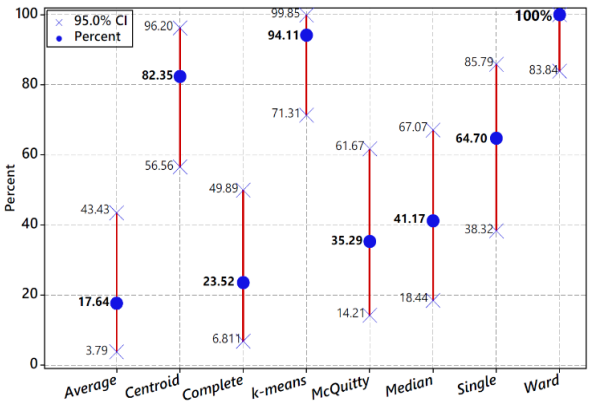


FIGURE 3. Degree of agreement for linkage methods in disturbed scenarios

Considering the method that showed consistency in all disturbance scenarios, we can apply it to the original data, using the PCA technique. Fig. 4 shows the groupings generated by the Ward method from the PC scores. When evaluating these clusters, it is possible to verify that Clusters 1, 2 and 3 present a lower occurrence of voltage sag, compared to the other clusters. The first cluster is formed by Suíça substation (with a similarity of 52.85%) and by Paulista and Aracruz substations, both showing similarity of 64.46%. The second cluster consists of five substations, where we can highlight Linhares C and Jaguaré, which have the highest level of similarity (69.4%). It is important to note that the five substations in Cluster 2 have identical primary voltage values (138 kV). Cluster 3 is formed by three substations (Ecoporanga, São Francisco and Montanha), with similarity levels greater than 52% and equal capacity transformers and primary voltages for the substations (15/20MVA and 69kV, respectively).

When analyzing clusters 4 and 5, we found that both represent the groups with the highest incidence of voltage sag. In detail, we verified that cluster 4 infers the grouping of the Pinheiros and Itarana substations (77.3% similarity). These substations have equivalent characteristics, such as the power and voltages transformers capacities (25/33/41 MVA; 138–69 kV and; 15/20 MVA; 69–13.8 kV, respectively). Such characteristics stand out for their importance for checking sags and voltage regulation [8]. In addition, Juncado and Santa Teresa substations (both with primary voltages of 69kV), are part of Cluster 4. Finally, we can see that cluster 5 performed the grouping of nearby substations, Linhares A and João Neiva, presenting 13.8/138kV transformers. It is important to note that the Ward linkage method, in addition to presenting 100% agreement between the disturbance scenarios, also showed full agreement with the values generated by the original data.

In order to better evaluate the classification of clusters, it is possible to determine the mean values of total number of sag events (TNE), for each group and the respective confidence intervals for the results. For this, analysis of variance (ANOVA) is usually used. However, given the multivariate nature of the data, it is more appropriate to consider the variance-covariance structure of the data. Such assessments will be carried out by analysis of covariance, which considers a concomitant variable. From the original data, it is possible to verify that the variable “number of events in medium voltage” (NEMV) has a higher level of significance for the TNE. However, this variable has a linear dependence for TNE (since this variable is part of the calculation to find TNE per year). Thus, the variable “equivalent medium voltage vulnerability area” (EMVVA), indicating the short circuits that cause sags in the substation busbars (represented in kilometers), is the most significant and does not have linear dependence. ANCOVA is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3088436, IEEE Access

performed considering EMVVA as a concomitant variable. Fig. 5 depicts the confidence intervals for clusters formed by the Ward method, where it is possible to easily verify the division of clusters, with significant partitions. Clusters 4 and 5 have a high incidence of voltage sag events, with average values above 300 sag events per year while clusters 1, 2 and 3 have a low incidence, with average values below 200 sag events.

TABLE XI
ASSESSMENT AGREEMENT WITHIN APPRAISERS

APPRAISER	# INSPECTED	# MATCHED	%	95% CI
Average	17	3	17.65	(3.80, 43.43)
Centroid	17	14	82.35	(56.57, 96.20)
Complete	17	4	23.53	(6.81, 49.90)
k-means	17	16	94.12	(71.31, 99.85)
McQuitty	17	6	35.29	(14.21, 61.67)
Median	17	7	41.18	(18.44, 67.08)
Single	17	11	64.71	(38.33, 85.79)
Ward	17	17	100	(83.84, 100.00)

#Matched: Appraiser agrees with him/herself across trials.

TABLE XII
RESULTS FOR FLEISS' KAPPA STATISTICS WITHIN APPRAISERS

Appraiser	Response	Kappa	SE Kappa	Z	P(vs>0)	Appraiser	Response	Kappa	SE Kappa	Z	P(vs>0)
Average	1	0.176	0.099	1.782	0.037	*McQuitty*	1	0.356	0.099	3.596	0.000
	2	0.510	0.099	5.147	0.000		2	0.510	0.099	5.147	0.000
	3	0.062	0.099	0.627	0.265		3	0.781	0.099	7.887	0.000
	4	1.000	0.099	10.100	0.000		4	0.784	0.099	7.919	0.000
	5	1.000	0.099	10.100	0.000		5	0.469	0.099	4.734	0.000
	Overall	0.382	0.060	6.378	0.000		*Overall*	0.559	0.057	9.770	0.000
Centroid	1	0.837	0.099	8.449	0.000	*Median*	1	0.514	0.099	5.194	0.000
	2	0.469	0.099	4.734	0.000		2	0.510	0.099	5.147	0.000
	3	0.469	0.099	4.734	0.000		3	0.469	0.099	4.734	0.000
	4	1.000	0.099	10.100	0.000		4	0.469	0.099	4.734	0.000
	5	1.000	0.099	10.100	0.000		5	0.469	0.099	4.734	0.000
	Overall	0.780	0.061	12.809	0.000		*Overall*	0.500	0.065	7.749	0.000
Complete	1	0.698	0.099	7.047	0.000	*Single*	1	0.673	0.099	6.798	0.000
	2	0.765	0.099	7.723	0.000		2	0.469	0.099	4.734	0.000
	3	0.401	0.099	4.053	0.000		3	0.469	0.099	4.734	0.000
	4	0.297	0.099	2.995	0.001		4	0.469	0.099	4.734	0.000
	5	0.150	0.099	1.515	0.065		5	0.469	0.099	4.734	0.000
	Overall	0.512	0.051	9.993	0.000		*Overall*	0.560	0.061	9.200	0.000
k-means	1	1.000	0.099	10.100	0.000	*Ward*	1	1.000	0.099	10.100	0.000
	2	0.892	0.099	9.004	0.000		2	1.000	0.099	10.100	0.000
	3	0.784	0.099	7.919	0.000		3	1.000	0.099	10.100	0.000
	4	1.000	0.099	10.100	0.000		4	1.000	0.099	10.100	0.000
	5	1.000	0.099	10.100	0.000		5	1.000	0.099	10.100	0.000
	Overall	0.958	0.056	17.260	0.000		*Overall*	1.000	0.051	19.542	0.000

For comparative purposes, confidence intervals were performed using linkage methods that presented consistency greater than 80%: k-means and Centroid. From the confidence intervals by the k-means method (Fig. 6), it was found that it performed clusters with close means values and with long confidence intervals, showing the lack of precision in the estimation for voltage sag analysis. Furthermore, it is not possible to perform an adequate separation of the clusters, making their classification unfeasible.

TABLE XIII
KENDALL'S COEFFICIENT OF CONCORDANCE WITHIN APPRAISERS

Appraiser	Coef	Chi - sq	DF	P-value
Average	0.56406	36.0996	16	0.0028
Centroid	0.90542	57.9469	16	0.0000
Complete	0.77393	49.5313	16	0.0000
k-means	0.99845	63.9005	16	0.0000
McQuitty	0.72691	46.522	16	0.0001
Median	0.73145	46.8129	16	0.0001
Single	0.7246	46.3717	16	0.0001
Ward	1.0000	64.0000	16	0.0000

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3088436, IEEE Access

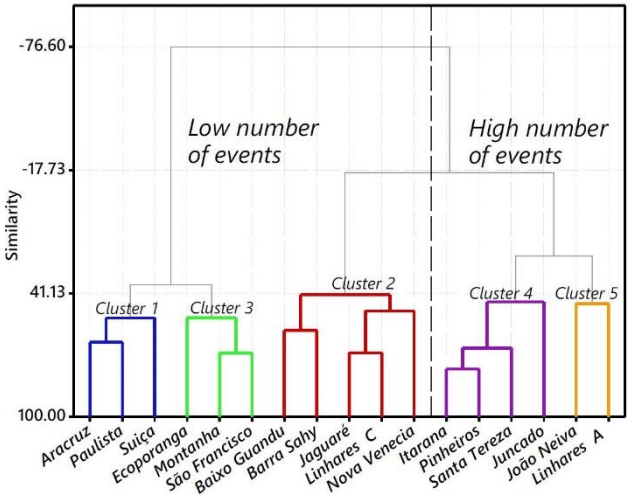


FIGURE 4. Ward method dendrogram for substations

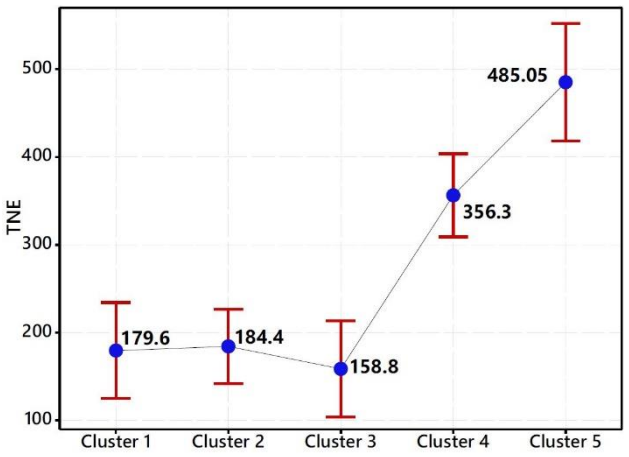


FIGURE 5. 95% interval plot of voltage sags for Ward linkage method

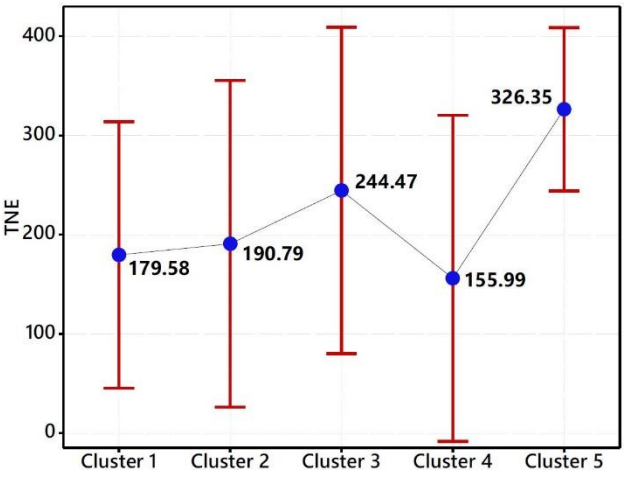
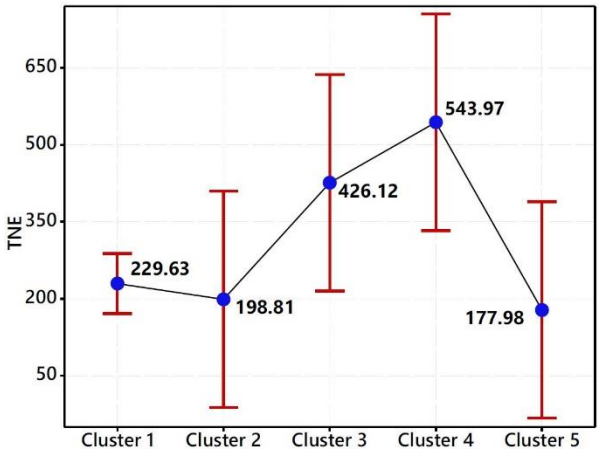


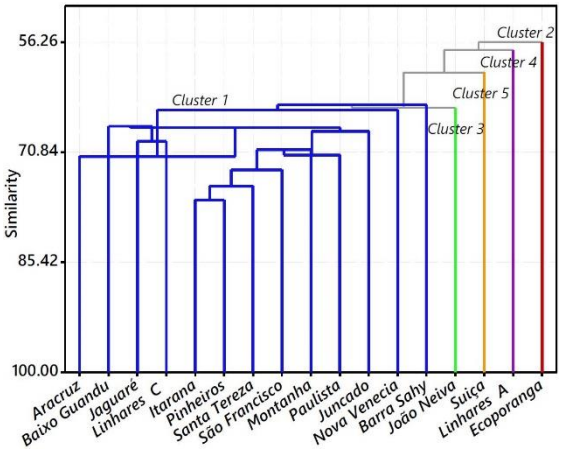
FIGURE 6. 95% interval plot of voltage sags for k-means linkage method

Fig. 7(a) shows the intervals for the Centroid method. In this graph it can be seen that the groups have different mean values. However, the confidence intervals are long, showing overlaps. These results infer that it is not possible to say that the mean is different. When analyzing the separation of clusters, it is possible to verify that the Centroid method added most of the substations in just one cluster (the same performance was verified in the Single method). Such behavior cannot be justified, making its analysis inadequate and unsatisfactory. The dendrogram of the Centroid method illustrates this behavior, as shown in Fig. 7(b).

The verification and the evaluation of these results demonstrates the importance to analyze the sensitivity of the results and not just agreement, as suggested by Johnson and Wichern [25]. Ward linkage method proved to be the most suitable for analyzing PQ data when classifying these substations to provide a diagnosis for concessionaires and regulatory agents.



(a)



(b)

FIGURE 7. (a) 95% Interval plot and (b) dendrogram of voltage sags for Centroid method.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3088436, IEEE Access

V. CONCLUSIONS

This work sought to present a methodology capable of finding the best linkage method from the scores of principal components. From the application of this methodology, the following conclusions can be inferred:

--The proposed approach was able to provide a viable and adequate alternative to verify the consistency of linkage methods in small disturbance scenarios;

--Performing the application in PQ data for substations, the Ward linkage method showed 100% consistency (for that specific data set), demonstrating to be a robust alternative when analyzing this data set.

--The non-hierarchical method k-means did not show absolute agreement in this case. When analyzing the classification criteria, k-means proved to be an adequate method with an excellent degree of agreement (Kappa and Kendall). However, when checking the clusters and their CIs, it was found that k-means had overlapping, making it difficult to separate groups with high and low incidence of voltage sag phenomena. The other methods were not able to present an adequate degree of agreement through the evaluated scenarios;

--When applying the Ward method to the original data, it confirmed the consistency of the oscillating scenarios. In addition, from ANCOVA it was possible to estimate the CI for each cluster, considering TNE. With this result, it was possible to verify that the Ward method provided a better grouping for this dataset, allowing to find two different levels of sag event incidence, promoting a good discriminatory level and favoring decision-making.

Finally, it can be inferred that the proposed methodology allows analyzing the behavior of the linkage methods, being able to be applied in different correlated data sets. In addition to the concordance results, the sensitivity analysis of the clusters allowed to confirm the robustness and behavior of the clusters through the confidence intervals. As future suggestions, this methodology can be applied using different data sets and different multivariate methods, such as the analysis of rotated factors scores.

APPENDIX

HVFR*	High Voltage Failure Rate (faults per 100 km line per year)	NEHV	Number of Events High Voltage (faults on high voltage lines)
TNE	Total number of Sag events per year (faults on MV and HV)	3LG	Three phase to ground short-circuit current (on the busbar)
NEMV	Number of Events in Medium Voltage (faults on MV);	2LG	Double phase to ground short-circuit current (on the busbar)
MVFR	Medium Voltage Failure Rate (faults per 100 km feederper year)	1LG	Single phase to ground short-circuit current (on the busbar)
MNE	Monitored Number of Events (one year 13.8 kV bus bar)	L-L	Phase to phase short-circuit current
LNE	Lower Number of Events (100 Scenarios simulated)	MAXA	Maximum asymmetric short circuit Current (on the busbar)
ANE	Average Number of Events (100 Scenarios simulated)	MAXS	Maximum symmetric short circuit Current (on the busbar)
UNE	Upper Number of Events (100 Scenarios simulated)	MAXG	Maximum symmetric short circuit Current to Ground (on the bus bar)
FKVAr	Shunt Capacitor KVAr Installed on the Feeders	R+	Positive Sequence Resistance (on the bus bar)
SAIFI1	System Average Interruption Frequency Index (without critical day)	X+	Positive Sequence Reactance (on the bus bar)
SAIFI2	System Average Interruption Frequency Index (with critical day)	Xo	Zero Sequence Reactance (on the bus bar)
STIFI	System Total Interruption Frequency Index (number of events)	ZBASE	Base Impedance ohm (on the bus bar)
FL	Feeders Length (km)	Zohm	Equivalent Impedance ohm (on the bus bar)
AREA	Cluster Area (square km) according to ANEEL [30]	Zpu	Per Unit Impedance (on the bus bar)
EMVVA	Equivalent Medium Voltage Vulnerability Area (km), within which the short-circuits cause sags on the substation bus	EVAHV	Equivalent High Voltage Vulnerability Area (km), within which the short-circuits cause sags on the substation bus
BMVAr	Shunt Capacitor MVAr installed on the bus bar	MVASC	Short Circuit Power MVA (1000/Zpu on the bus bar)

REFERENCES

[1] F. A. de Almeida, J. Miranda Filho, L. F. Amorim, J. H. de F. Gomes, and A. P. de Paiva, "Enhancement of discriminatory power by ellipsoidal functions for substation clustering in voltage sag studies," *Electr. Power Syst. Res.*, vol. 185, p. 106368, 2020, doi: https://doi.org/10.1016/j.epsr.2020.106368.

[2] Y. Han, Y. Feng, P. Yang, L. Xu, Y. Xu, and F. Blaabjerg, "Cause, Classification of Voltage Sag, and Voltage Sag Emulators and Applications: A Comprehensive Overview," *IEEE Access*, vol. 8, pp. 1922–1934, 2020, doi: 10.1109/ACCESS.2019.2958965.

[3] H. M. G. C. Branco, M. Oleskovicz, D. V Coury, and A. C. B. Delbem, "Multiobjective optimization for power quality monitoring allocation considering voltage sags in distribution systems," *Int. J. Electr. Power Energy Syst.*, vol. 97, pp. 1–10, 2018, doi: https://doi.org/10.1016/j.ijepes.2017.10.011.

[4] H. M. A. Ahmed, A. S. A. Awad, M. H. Ahmed, and M. M. A. Salama, "Mitigating voltage-sag and voltage-deviation problems in distribution networks using battery energy storage systems," *Electr. Power Syst. Res.*, vol. 184, p. 106294, 2020, doi: https://doi.org/10.1016/j.epsr.2020.106294.

[5] C. Behera, G. H. Reddy, P. Chakrapani, A. K. Goswami, C. P. Gupta, and G. K. Singh, "Assessment of Equipment Trip Probability Due to Voltage Sags Based on Fuzzy Possibility Distribution Function," *IEEE Access*, vol. 6, pp. 76889–76899, 2018, doi: 10.1109/ACCESS.2018.2884562.

[6] M. H. Moradi and Y. Mohammadi, "Voltage sag source location: A review with introduction of a new method," *Int. J. Electr. Power Energy Syst.*, vol. 43, no. 1, pp. 29–39, 2012, doi: https://doi.org/10.1016/j.ijepes.2012.04.041.

[7] H. Erişti, Ö. Yıldırım, B. Erişti, and Y. Demir, "Optimal feature selection for classification of the power quality events using wavelet transform and least squares support vector machines," *Int. J. Electr. Power Energy Syst.*, vol. 49, pp. 95–103, 2013, doi: https://doi.org/10.1016/j.ijepes.2012.12.018.

[8] H. Sun, H. Yi, G. Yang, F. Zhuo, and A. Hu, "Voltage Sag Source Identification Based on Few-Shot Learning," *IEEE Access*, vol. 7, pp. 164398–164406, 2019, doi: 10.1109/ACCESS.2019.2953226.

[9] S. Majumder, A. P. Agalgaonkar, S. A. Khaparde, P. Ciufo, S. Perera, and S. V. Kulkarni, "Allocation of Common-Pool Resources in an Unmonitored Open System," *IEEE Trans. Power Syst.*, vol. 34, no. 5, pp. 3912–3920, 2019, doi: 10.1109/TPWRS.2019.2901389.

[10] J. Miranda *et al.*, "A PCA-based approach for substation clustering for voltage sag studies in the Brazilian new energy context," *Electr. Power Syst. Res.*, vol. 136, pp. 31–42, 2016, doi: 10.1016/j.epsr.2016.02.012.

[11] F. A. de Almeida, A. C. O. Santos, A. P. de Paiva, G. F. Gomes,

and J. H. de F. Gomes, "Multivariate Taguchi loss function optimization based on principal components analysis and normal boundary intersection," *Eng. Comput.*, 2020, doi: 10.1007/s00366-020-01122-8.

[12] R. Machlev, D. Tolkachov, Y. Levron, and Y. Beck, "Dimension reduction for NILM classification based on principle component analysis," *Electr. Power Syst. Res.*, vol. 187, p. 106459, 2020, doi: https://doi.org/10.1016/j.epsr.2020.106459.

[13] G. Belinato, F. A. de Almeida, A. P. de Paiva, J. H. de Freitas Gomes, P. P. Balestrassi, and P. A. R. C. Rosa, "A multivariate normal boundary intersection PCA-based approach to reduce dimensionality in optimization problems for LBM process," *Eng. Comput.*, vol. 35, no. 4, pp. 1533–1544, 2019, doi: 10.1007/s00366-018-0678-3.

[14] B. Song, X. Zhou, S. Tan, H. Shi, B. Zhao, and M. Wang, "Process Monitoring via Key Principal Components and Local Information Based Weights," *IEEE Access*, vol. 7, pp. 15357–15366, 2019.

[15] F. A. Almeida *et al.*, "Measurement data from bobbins of Partially Oriented Yarns: Univariate and multivariate aspects," *Data Br.*, vol. 27, p. 104637, Dec. 2019, doi: 10.1016/j.dib.2019.104637.

[16] A. Bosisio, A. Berizzi, D.-D. Le, F. Bassi, and G. Giannuzzi, "Improving DTR assessment by means of PCA applied to wind data," *Electr. Power Syst. Res.*, vol. 172, pp. 193–200, 2019, doi: https://doi.org/10.1016/j.epsr.2019.02.028.

[17] B. Song, X. Zhou, H. Shi, and Y. Tao, "Performance-Indicator-Oriented Concurrent Subspace Process Monitoring Method," *IEEE Trans. Ind. Electron.*, vol. 66, no. 7, pp. 5535–5545, 2019, doi: 10.1109/TIE.2018.2868316.

[18] N. M. Puggina Bianchesi, E. L. Romao, M. F. B. P. Lopes, P. P. Balestrassi, and A. P. De Paiva, "A design of experiments comparative study on clustering methods," *IEEE Access*, vol. 7, pp. 167726–167738, 2019, doi: 10.1109/ACCESS.2019.2953528.

[19] M. Jasiński, T. Sikorski, and 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.*, vol. 166, pp. 52–60, 2019, doi: https://doi.org/10.1016/j.epsr.2018.09.020.

[20] J. J. López, J. A. Aguado, F. Martín, F. Muñoz, A. Rodríguez, and J. E. Ruiz, "Hopfield–K-Means clustering algorithm: A proposal for the segmentation of electricity customers," *Electr. Power Syst. Res.*, vol. 81, no. 2, pp. 716–724, 2011, doi: https://doi.org/10.1016/j.epsr.2010.10.036.

[21] K. Vinothkumar and M. P. Selvan, "Hierarchical Agglomerative Clustering Algorithm method for distributed generation planning," *Int. J. Electr. Power Energy Syst.*, vol. 56, pp. 259–269, 2014, doi: 10.1016/j.ijepes.2013.11.021.

[22] D. Pinel, "Clustering methods assessment for investment in zero emission neighborhoods' energy system," *Int. J. Electr. Power Energy Syst.*, vol. 121, p. 106088, 2020, doi: https://doi.org/10.1016/j.ijepes.2020.106088.

[23] A. M. S. Ferreira, C. H. De Oliveira Fontes, C. A. M. T. Cavalcante, and J. E. S. Marambio, "Pattern recognition as a tool to support decision making in the management of the electric sector. Part II: A new method based on clustering of multivariate time series," *Int. J. Electr. Power Energy Syst.*, vol. 67, pp. 613–626, 2015, doi: 10.1016/j.ijepes.2014.12.001.

[24] L. P. Fávero, *Análise de Dados: Técnicas Multivariadas Exploratórias com SPSS e STATA*. Elsevier, Grupo Gen, 2015.

[25] D. Johnson, R.A., Wichern, *Applied Multivariate Statistical Analysis*, 6th ed. New Jersey: Prentice-Hall, 2007.

[26] W. H. Woodall and C. M. Borror, "Some Relationships between Gage R&RCriteria," *Qual. Reliab. Eng. Int.*, vol. 24, 2008, doi: 10.1002/qre.870.

[27] F. A. Almeida, R. R. Leite, G. F. Gomes, J. H. de F. Gomes, and A. P. de Paiva, "Multivariate data quality assessment based on rotated factor scores and confidence ellipsoids," *Decis. Support Syst.*, vol. 129, p. 113173, Oct. 2020, doi: 10.1016/j.dss.2019.113173.

[28] F. A. De Almeida, S. C. Streitenberger, A. F. Torres, A. P. De Paiva, and J. H. D. F. Gomes, "A Gage Study Through the Weighting of Latent Variables Under Orthogonal Rotation," *IEEE Access*, vol. 8, pp. 183557–183570, 2020, doi: 10.1109/ACCESS.2020.3019031.

[29] F. A. Almeida *et al.*, "A Gage Study Applied in Shear Test to Identify Variation Causes from a Resistance Spot Welding Measurement System," *Stroj. Vestn. J. Mech. Eng.*, vol. 64, pp. 621–631, 2018, doi: 10.5545/sv-jme.2018.5235.

[30] F. A. de Almeida, J. H. de F. Gomes, G. F. Gomes, E. L. Romão, and P. P. Balestrassi, "Variation causes analysis attributed to different metrological instruments to verify the geometric characteristics of a spot welding process," *Soldag. e Insp.*, vol. 23, no. 4, pp. 485–504, 2018, doi: 10.1590/0104-9224/SI2304.05.

[31] C. Marques, N. Lopes, G. Santos, I. Delgado, and P. Delgado, "Improving operator evaluation skills for defect classi fi cation using training strategy supported by attribute agreement analysis," *Measurement*, vol. 119, no. January, pp. 129–141, 2018, doi: 10.1016/j.measurement.2018.01.034.

[32] P. Taylor, J. Lyu, and M. Chen, "Measurement of bivariate attributes using a novel statistical model Measurement of bivariate attributes using," no. September 2013, pp. 37–41, doi: 10.1080/02664760903030221.

[33] J. A. Velasco, H. Amaris, and M. Alonso, "Deep Learning loss model for large-scale low voltage smart grids," *Int. J. Electr. Power Energy Syst.*, vol. 121, p. 106054, 2020, doi: https://doi.org/10.1016/j.ijepes.2020.106054.

[34] R. C. Jain, A. K.; Dubes, *Algorithms for Clustering Data*, 1st ed. Pearson College Div, 1988.

[35] Alvin C. Rencher, *Methods of Multivariate Analysis*, 2nd ed. New York: John Wiley & Sons, Inc., 2002.

[36] H. C. Romesburg, *Cluster Analysis for Researchers*. Lulu Press, 2004.

[37] L. L. Mcquitty, "FOR DISCRETE AND CONTINUOUS DATA analysis (McQuitty , 1955). Equation ij," pp. 825–831, 1966.

[38] J. H. Ward, "Hierarchical Grouping to Optimize an Objective Function," *J. Am. Stat. Assoc.*, vol. 58, no. 301, pp. 236–244, Mar. 1963, doi: 10.1080/01621459.1963.10500845.

[39] T. Babnik, R. K. Aggarwal, and P. J. Moore, "Principal Component and Hierarchical Cluster Analyses as Applied to Transformer Partial Discharge Data With Particular Reference to Transformer Condition Monitoring," *IEEE Trans. Power Deliv.*, vol. 23, no. 4, pp. 2008–2016, 2008, doi: 10.1109/TPWRD.2008.919030.

[40] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychol. Bull.*, vol. 76, no. 5, pp. 378–382, 1971, doi: 10.1037/h0031619.

[41] J. L. Fleiss, B. Levin, and M. C. Paik, *Statistical Methods for Rates and Proportions*, Third Edit. John Wiley & Sons, Inc., 2003.

[42] AIAG, *Measurement systems analysis: reference manual*, 4th ed. Detroid, MI, USA: Automotive Industry Action Group, 2010.

[43] N. Gisev, J. S. Bell, and T. F. Chen, "Interrater agreement and interrater reliability: key concepts, approaches, and applications.," *Res. Social Adm. Pharm.*, vol. 9, no. 3, pp. 330–338, 2013, doi: 10.1016/j.sapharm.2012.04.004.

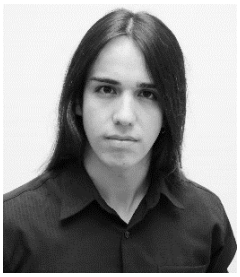
[44] G. H. Lewis and R. G. Johnson, "Kendall's Coefficient of Concordance for Sociometric Rankings with Self Excluded," *Sociometry*, vol. 34, no. 4, pp. 496–503, Aug. 1971, doi: 10.2307/2786195.

[45] A. Agresti, *Analysis of Ordinal Categorical Data*, 2nd ed. John Wiley & Sons, Inc., 2010.

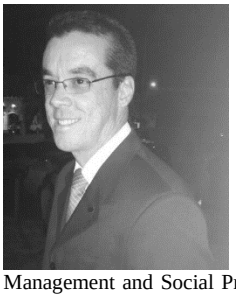
[46] D. E. Hinkle, W. Wiersma, and S. G. Jurs, *Applied Statistics for the Behavioral Sciences*. Boston: Houghton Mifflin, 2002.

[47] F. A. de Almeida, G. F. Gomes, J. H. D. Gaudêncio, J. H. de F. Gomes, and A. P. de Paiva, "A new multivariate approach based on weighted factor scores and confidence ellipses to precision evaluation of textured fiber bobbins measurement system," *Precis. Eng.*, vol. 60, pp. 520–534, Nov. 2019, doi: 10.1016/j.precisioneng.2019.09.010.

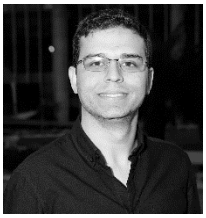
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3088436, IEEE Access



FABRÍCIO A. ALMEIDA received his B.S. degree in economics from the Faculty of Economic Sciences Southern Minas Gerais (2014). He received his M.S. degree in industrial engineering from the Federal University of Itajubá (2017) and he is currently pursuing a Ph.D. in industrial engineering at UNIFEI. Currently, he is Black Belt in Lean Six Sigma and a professor at the Faculty of Economic Sciences Southern Minas Gerais and a member of the research group of the Nucleus of Manufacturing Optimization and Innovation Technology. His research areas include multivariate statistical analysis, multiobjective optimization, quality engineering, industrial economics, and design of experiments. Metrics since 2018: 221 citations and *h9* factor.



LUIZ G. DE MELLO is a professor at the Federal Institute of Southern Minas Gerais (IFSULDEMINAS - since 2015) where he also works as Director of Educational Development at the Advanced Campus Carmo de Minas. He has the following academic trajectory: Bachelor of Business Administration (Faculty Santa Marta); Postgraduate in Instructional Design for Virtual Distance Learning (Federal University of Itajubá - UNIFEI); Postgraduate in People Management and Social Projects (UNIFEI); Postgraduate in Teaching in Professional and Technological Education (IFSULDEMINAS) and Master in Production Engineering (UNIFEI). He is a student of the Doctorate Degree program in Production Engineering (UNIFEI). He has been working as a professor since 2003 and taught at the following institutions: Green River Valley University (UNINCOR); President Antônio Carlos University (UNIPAC); Victor Hugo Faculty e São Lourenço Faculty. Developed research in the following areas: Multi-objective Optmization; Forecasting; Balanced Scorecard, Cronbach Alpha e Performance Measurement Systems.



ESTEVÃO L. ROMÃO received the M.Sc. degree in Industrial Engineering from Federal University of Itajubá in 2019, received B.S. degree in Industrial Engineering from Federal University of Viçosa, Minas Gerais, Brazil, in 2017. Currently pursuing the D.Sc. degree in Industrial Engineering at Federal University of Itajubá, Brazil. His research interests include Nonlinear Optimization, Times Series Forecasting, ANN and Statistics.

GUILHERME F. GOMES was born in Itajubá, Brazil in 1989. He received the B.S. and M.D in mechanical engineering from the Ecole Nationale d'Ingénieurs de Metz, France, in 2014 and the D.Sc. degree in mechanical engineering from Universidade Federal de Itajubá, Brazil, in 2017. Since 2017 he has been a Professor and researcher with the Mechanical Engineering Institute, Universidade Federal de Itajubá. He is the author of more than 75 articles and inventions. He is reviewer of more than 45 international journals and doctoral advisor. His research interests include: structural analysis, structural health monitoring, modal testing, applied artificial intelligence, optimization methods and composite structures: More information can be found on the website: https://guilherme.unifei.edu.br

JOSÉ H. F. GOMES is a Mechanical Industrial Engineer from the Federal University of Itajubá (2003-2007), with master's degree (2008-2010) and PhD (2010-2013) in Industrial Engineering from the same institution. Currently Associate Professor I of the Institute of Industrial and Management Engineering (IEPG) of the Federal University of Itajubá, in the undergraduate courses in Production Engineering and Administration, and postgraduate in Production Engineering. It acts in the research lines of modeling, analysis and optimization of production systems and manufacturing operations. His main research areas are: application and improvement of multi-objective methods, statistical methods, mathematical programming methods and materials management.

ANDERSON P. PAIVA is a professor of the Industrial Engineering and Management Institute at Federal University of Itajubá since 2003. He has 115 journal papers published on Scopus about nonlinear multiobjective optimization, multivariate statistics and Design of Experiments (DOE). He has achieved a H19 factor. He got his bachelor's degree in mechanical engineering in 1996, his master's degree in industrial engineering at Federal University of Itajubá (UNIFEI) in 2004 and his doctor's degree in mechanical engineering at Federal University of Itajubá (UNIFEI) in 2006.

JACQUES MIRANDA FILHO born in Rio de Janeiro, in 1954. Graduation in Electrical Engineering at Federal School Engineering of Itajubá in 1975. Msc Degree at Federal University of Itajubá in 2005 and PHD Degree at the same University in 2016. From 1976 to 2006 worked at a Distribution Utility, most of this period as Expansion Planning Engineer and Research and Development Projects Coordination. From 2006 until today working with R&D Projects and Professor of Electricity and Electrical Power Systems at Federal Institute of Espírito Santo State.

PEDRO PAULO BALESTRASSI received his his B.S. degree in electrical engineering from the Federal University of Espírito Santo and his M.S. degree in electrical engineering from the Federal University of Itajubá. He received his Ph.D. in industrial engineering from the Federal University of Santa Catarina through an exchange program of the Texas A&M University. He is a professor of the Industrial Engineering and Management Institute at the Federal University of Itajubá (UNIFEI). He has a research scholarship from the CNPq (level 1D). He is a visiting professor at the Polytechnical University of Catalunha (Barcelona Tech), the University of Tennessee in Knoxville, and the University of Texas in Austin. His research areas include quality engineering, statistics, design of experiments, time series and forecast, and artificial neural networks.