Representative Residential LV Feeders: A Case Study for the North West of England

Valentin Rigoni, Luis F. Ochoa, *Senior Member, IEEE*, Gianfranco Chicco, *Senior Member, IEEE*, Alejandro Navarro-Espinosa, *Student Member, IEEE*, and Tuba Gozel, *Member, IEEE*

***Abstract*—The adoption of residential-scale low carbon technologies, such as photovoltaic panels or electric vehicles, is expected to significantly increase in the near future. Therefore, it is important for distribution network operators (DNOs) to understand the impacts that these technologies may have, particularly, on low voltage (LV) networks. The challenge, however, is that these LV networks are large in number and diverse in characteristics. In this work, four clustering algorithms (hierarchical clustering, k-medoids++, improved k-means++, and Gaussian Mixture Model—GMM), are applied to a set of 232 residential LV feeders from the North West of England to obtain representative feeders. Moreover, time-series monitoring data, presence of residential-scale generation, and detailed customer classification are considered in the analysis. Multiple validity indices are used to identify the most suitable algorithm. The improved k-means++ and GMM showed the best performances resulting in eleven representative feeders with prominent characteristics such as number and type of customers, total cable length, neutral current, and presence of generation. Crucially, the results from studies performed on these feeders can then be extrapolated to those they represent, simplifying the analyses to be carried out by DNOs. This is demonstrated with a hosting capacity assessment of photovoltaic panels in LV feeders.**

***Index Terms*—Clustering techniques, low voltage, representative feeders, taxonomy.**

I. Introduction

Low carbon technologies (LCT), from renewable generation to new types of customer-side loads, are incentivized by governments around the world to try to reduce their CO_2 emissions. In the U.K., a 34% reduction with respect to the values from 1990 is expected to be achieved by 2020. This will be in part accomplished by increasing distributed generation capacity (e.g., wind power, photovoltaic systems), and electrifying the transport (e.g., electric vehicles) and heating (e.g., electric heat pumps) sectors [1]. The expectation is for these technologies to be adopted mostly by householders as they currently represent approximately 29% of the U.K. energy consumption [2].

High penetrations of residential-scale LCT are likely to have technical impacts on the very circuits they will be connected to, i.e., the low voltage (LV) feeders [3]. Hence, it is imperative for distribution network operators (DNOs) to understand these impacts so preventive actions or long-term solutions can be deployed in a timely manner. This, however, poses a significant challenge as these LV feeders are large in number and diverse in characteristics. With approximately 1.3 million LV feeders in the U.K., detailed analyses of individual feeders would require enormous efforts from the DNOs.

The adoption of representative LV feeders opens a window for DNOs to better understand the characteristics of the whole population, and also to simplify any study needed to be carried out on it. Each representative feeder can link its characteristics and behaviors to the population it represents. Therefore, the results from any study performed on them can be as well extrapolated reducing the complexity of multiple analyses.

Sets of representative feeders have been obtained in the past for the USA, Australia, and the U.K. [4]–[9]. In [4], the Pacific Northwest National Laboratory (PNNL) provides a set of 24 medium voltage (MV) representative feeders across the USA considering 35 attributes to characterize each feeder. Nevertheless, the mathematical justification of the number of clusters relies on their compactness without taking into account their dissimilarity. A similar approach is used in [5] for Australian MV feeders where the features considered are limited to network parameters. A more complex index is considered but with a relatively low performance.

Representative LV and MV feeders for Western Australia were produced in [6] by combining cluster and discriminant analysis (DA). A mathematical methodology was proposed in order to significantly reduce the number of features used for the analysis. Hierarchical clustering was applied and the optimal solution found by analyzing three statistical parameters. The characterization of customers was limited to 4 attributes that considered number of residential and non-residential customers and their corresponding nominal capacities. In [7], the same authors proposed a set of 9 representative MV feeders (22 kV) for Western Australia by limiting the analysis on 6 experience-based features and applying the same statistical parameters.

In [8], a different clustering algorithm, k-means, resulted in 12 representative MV feeders for the California area (USA). After reducing the dimension of the original data, fifteen fea-

Manuscript received June 27, 2014; revised October 27, 2014 and January 25, 2015; accepted January 26, 2015. Date of publication February 27, 2015; date of current version December 18, 2015. This work was supported in part by Electricity North West Limited, U.K., through the Ofgem's Low Carbon Networks Fund Tier 1 Project "LV Network Solutions", 2011-2014. Paper no. TPWRS-00870-2014.

V. Rigoni and G. Chicco are with the Energy Department, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy (e-mail: valentinrigoni@yahoo.com.ar; gianfranco.chicco@polito.it).

L. F. Ochoa, A. Navarro-Espinosa, and T. Gozel are with the School of Electrical and Electronic Engineering, The University of Manchester, Manchester M13 9PL, U.K. (e-mail: luis_ochoa@ieee.org; alejandro.navarroespinosa@manchester.ac.uk; tgozel@gtu.edu.tr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TPWRS.2015.2403252

tures were used to cope with the complexity of the problem. Only one criterion, cubic clustering, was used to determine the optimal number of clusters.

In the U.K., MV networks (11 kV and 6.6 kV) were clustered in [9] using a "decision-tree" approach. Representative networks were constructed using the average characteristics of the feeders belonging to each cluster. No index was used to quantify the quality of the clusters.

This paper presents a clustering-based methodology that identifies a set of representative LV feeders from the North West of England. The strength of the analysis relies on two main aspects. First, four clustering algorithms (hierarchical clustering, k-medoids++, improved k-means++, and Gaussian Mixture Model) are adopted and compared using multiple validity indices. Second, feeders are characterized using 19 features including time-series monitoring data (10-min resolution), presence of residential-scale generation, network parameters, and multi-customer classification. Both aspects allow ensuring high quality clusters with an adequate compromise between diversity and number of representative LV feeders. Furthermore, the validity and application of the obtained representative feeders is demonstrated with a hosting capacity assessment of photovoltaic (PV) panels in LV feeders.

This paper is structured as follows: Section II describes the initial data processing. Section III proposes a set of 4 clustering algorithms and a set of validity indices to determine the optimal cluster structure. Section IV presents the results of the clustering analyses while Section V provides details of the final 11 representative feeders along with a quantitative characterization. Section VI presents an application of the representative feeders by assessing their PV hosting capacity. Finally, conclusions are drawn in Section VII.

II. Feeder Data Processing

A set of 383 feeders with network and monitoring data was created from 141 LV distribution substations (11 kV/0.4 kV and 6.6 kV/0.4 kV) from Electricity North West Limited (ENWL). After a validation and cleansing process (see below) this initial number of feeders was reduced to 232. A set of 19 normalized features was used to characterize each feeder, obtaining the definitive data set for a clustering process.

A. Data Cleansing and Validation of Feeders

For clustering algorithms to perform adequately any sort of outliers need to be eliminated. An initial data cleansing process was applied to remove feeders with uncommon characteristics. This includes issues with the monitoring data (e.g., voltages below 0.9 p.u. at the busbar) or isolated network characteristics (e.g., only two feeders with micro combined heat and power).

Due to reconfiguration practices, it is challenging for DNOs to keep updated records of type and number of customers per feeder. Consequently, it is essential to validate the feeder data with the corresponding monitoring data. This validation consists in checking the correspondence of the aggregated active power values and the "known" customer composition of the feeder. To do this, the daily and peak (5 pm to 8 pm) energy consumption of each feeder was compared with an estimation based on the known type and number of customers. Diversified profiles based on the Elexon profile classes adopted in the U.K. [10] were aggregated accordingly per feeder. For those feeders with PV panels, the corresponding generation profiles were created considering registered capacities and solar irradiance values from [11].

Although the diversified profiles provide a realistic representation of demand per customer type, this is accurate only when aggregated in large numbers—not the case for the studied feeders (mostly ranging from 5 to 180 customers). In addition, the adopted PV profiles do not necessarily represent the actual generation behavior of the analyzed days. Consequently, these potential inaccuracies should be taken into account when defining the acceptable error, i.e., the boundary that delimits if a feeder is valid or not.

An acceptable error of 60% was adopted according to the industrial report presented in [12]. This error was set for both daily and peak energy consumption per feeder. In [12], this value was obtained by statistically analyzing the mismatches with monitoring data of substations with verified topologies.

After the data cleansing and validation of feeders the initial data set of 383 feeders was reduced to 232.

B. Selection of Features

A compromise has to be found in order to take into account diversity and the identification of truly representative feeders. Here, the selection and simplification of the features used for the description of the 232 feeders was made considering those that are relevant to describe the findings from the analyses (e.g., LCT impact analyses) that will be performed on them. For instance, the total path impedance can help understanding the electrical strength of a feeder.

Following the approach in [4], the monitoring related features were processed using mean values and associated standard deviations. This allowed reducing three-phase and time-series data into representative values, thus resulting in a smaller number of features. This is important in order to maintain a balance between network and monitoring data related features.

In addition to the features that provide a general picture of the characteristics of a feeder (e.g., number of customers, customer type, active and reactive power, etc.), parameters such as impedance and neutral current were also considered as important. The selection of the features required a trial and error approach by which the set of features with the best performance and relevance was selected. The final features used for the description of the feeders are presented in Table I.

The main path distance is equal to the path connecting the farthest customer to the head of the feeder. The average path distance is the mean value of each customer's path to the head of the feeder. The total path impedance is calculated as the total sum of the impedances of each customer's path to the head of the feeder.

The "PV-supplied demand" is calculated as the ratio between the total declared PV capacity and the typical aggregated demand of the feeder during midday (i.e., when maximum PV generation is expected). The latter is quantified using the values from the corresponding Elexon profile classes given that the

TABLE I
ADOPTED FEATURES

Feature	Description
1	Total number of customers (PC1 to PC8)
2	No. of Domestic Economy 7 Two Rate customers (PC2)
3	No. of Non-Domestic Unrestricted customers (PC3)
4	No. of Non-Domestic Unrestricted (PC3) and Economy 7 Two Rate (PC4) customers
5	No. of Non-Domestic Maximum Demand customers (PC5 to PC8)
6	Total conductor length [m]
7	Main path distance [m]
8	Average path impedance [ohms]
9	Total path impedance [ohms]
10	Daily mean neutral current [A]
11	Mean 3ϕ daily active power [kW]
12	Daily mean standard deviation of 3ϕ active power [kW]
13	Daily mean standard deviation of 1ϕ active power [kW]
14	Mean 3ϕ daily reactive power [kvar]
15	Power Factor (PF)
16	No. of PV installations
17	PV-supplied demand
18	PV penetration level (n° customers/PV installations)
19	Mean PV installation capacity (kW/unit)

TABLE II
ELEXON PROFILES

ENWL Description	**Profile Classes Code**
Domestic Unrestricted	PC1
Domestic Economy 7 Two Rate	PC2
Non-Domestic Unrestricted	PC3
Non-Domestic Economy 7 Two Rate	PC4
Non-Domestic Economy 7 Off Peak	PC4
Non-Domestic Maximum Demand	PC5, PC6, PC7, PC8

monitoring data is already affected by the local PV production. This feature is to some extent a proxy of the actual auto-consumption of the feeders with PV and hence useful to understand the potential impacts.

The considered customer classification corresponds to Elexon profiles classes (PC) [10]. These profile classes are a categorization used throughout the U.K. for electricity settlement and tariff purposes. The 8 profile classes, shown in Table II, were defined by Elexon by applying the concept of load profiling and represent the consumption of all customers below 100 kW of maximum demand.

Domestic Economy 7 Two Rate customers, different from Domestic Unrestricted customers, have electric space heating. The rest are non-domestic customers with varied characteristics and behavior during the day. A high granularity in customer's classification was considered. This is important given that the impacts of LCT will depend on the interactions with traditional loads. For instance, PC3 and PC4 have both a peak close to 3.5 kW but temporarily located in different times of the day (noon and night, respectively).

The number of Domestic Unrestricted customers (domestic customers without electric space heating) was not considered as a feature. This is because they represent 91% of the customers in the total set of feeders. Together with the total number of customers they would have had an excessive weight in front of the remaining features.

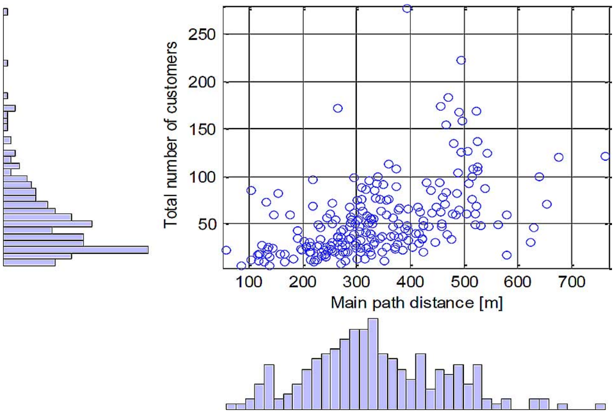


Fig. 1. Scatter plot with marginal histograms of features 1 and 7.

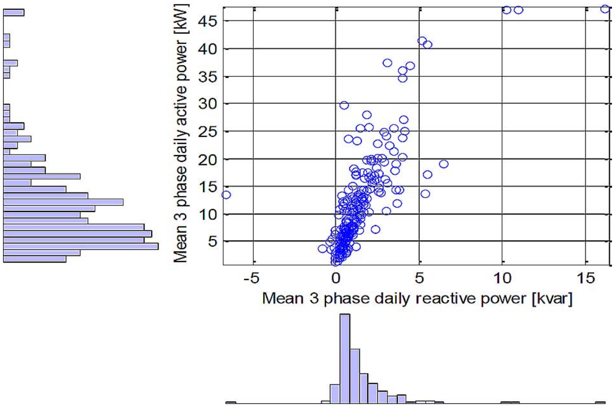


Fig. 2. Scatter plot with marginal histograms of features 11 and 14.

C. Properties of the Adopted Features

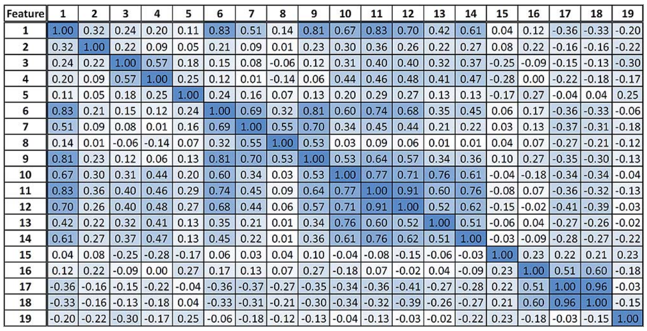
An analysis of the statistical distribution and correlation of the features used for the characterization of the data set was performed. This analysis provides a better understanding of the properties associated to the population under analysis.

The diverse nature of the adopted features resulted in different probability distributions. In most cases, distributions were found to be very or slightly skewed. For instance, features associated with the number of customers per profile class presented the most skewed distributions while others related to cable characteristics (features 6 to 9) and mean PV installation capacity were found to be closer to normal distributions. For example, Fig. 1 shows the histograms of the total number of customers and main path distance together with the corresponding scatter plot.

In the case of monitoring data, features were found to have skewed distributions as shown in Fig. 2. Features related to the penetration of PV panels presented skewed distributions except for the "Mean PV installation capacity" which was clearly normal.

On the other hand, Fig. 3 presents the correlation matrix $\mathbf{R}$ of the data set. The entries of the matrix $\mathbf{R}$ are related to those of the covariance matrix $\mathbf{C}$ as in (1):

$$R_{ij} = \frac{C_{ij}}{\sqrt{C_{ii}C_{jj}}}. \tag{1}$$



Feature	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	1.00	0.32	0.24	0.20	0.11	0.83	0.51	0.14	0.81	0.67	0.83	0.70	0.42	0.61	0.04	0.12	-0.36	-0.33	-0.20
2	0.32	1.00	0.22	0.09	0.05	0.21	0.09	0.01	0.23	0.30	0.36	0.26	0.22	0.27	0.08	0.22	-0.16	-0.16	-0.22
3	0.24	0.22	1.00	0.57	0.18	0.15	0.08	-0.06	0.12	0.31	0.40	0.40	0.32	0.37	-0.25	-0.09	-0.15	-0.13	-0.30
4	0.20	0.09	0.57	1.00	0.25	0.12	0.01	-0.14	0.06	0.44	0.46	0.48	0.41	0.47	-0.28	0.00	-0.22	-0.18	-0.17
5	0.11	0.05	0.18	0.25	1.00	0.24	0.16	0.07	0.13	0.20	0.29	0.27	0.13	0.13	-0.17	0.27	-0.04	0.04	0.25
6	0.83	0.21	0.15	0.12	0.24	1.00	0.69	0.32	0.81	0.60	0.74	0.68	0.35	0.45	0.06	0.17	-0.36	-0.33	-0.06
7	0.51	0.09	0.08	0.01	0.16	0.69	1.00	0.55	0.70	0.34	0.45	0.44	0.21	0.22	0.03	0.13	-0.37	-0.31	-0.18
8	0.14	0.01	-0.06	-0.14	0.07	0.32	0.55	1.00	0.53	0.03	0.09	0.06	0.01	0.01	0.04	0.07	-0.27	-0.21	-0.12
9	0.81	0.23	0.12	0.06	0.13	0.81	0.70	0.53	1.00	0.53	0.64	0.57	0.34	0.36	0.10	0.27	-0.35	-0.30	-0.13
10	0.67	0.30	0.31	0.44	0.20	0.60	0.34	0.03	0.53	1.00	0.77	0.71	0.76	0.61	-0.04	-0.18	-0.34	-0.34	-0.04
11	0.83	0.36	0.40	0.46	0.29	0.74	0.45	0.09	0.64	0.77	1.00	0.91	0.60	0.76	-0.08	0.07	-0.36	-0.32	-0.13
12	0.70	0.26	0.40	0.48	0.27	0.68	0.44	0.06	0.57	0.71	0.91	1.00	0.52	0.62	-0.15	-0.02	-0.41	-0.39	-0.03
13	0.42	0.22	0.32	0.41	0.13	0.35	0.21	0.01	0.34	0.76	0.60	0.52	1.00	0.51	-0.06	0.04	-0.27	-0.26	-0.02
14	0.61	0.27	0.37	0.47	0.13	0.45	0.22	0.01	0.36	0.61	0.76	0.62	0.51	1.00	-0.03	-0.09	-0.28	-0.27	-0.22
15	0.04	0.08	-0.25	-0.28	-0.17	0.06	0.03	0.04	0.10	-0.04	-0.08	-0.15	-0.06	-0.03	1.00	0.23	0.22	0.21	0.23
16	0.12	0.22	-0.09	0.00	0.27	0.17	0.13	0.07	0.27	-0.18	0.07	-0.02	0.04	-0.09	0.23	1.00	0.51	0.60	-0.18
17	-0.36	-0.16	-0.15	-0.22	-0.04	-0.36	-0.37	-0.27	-0.35	-0.34	-0.36	-0.41	-0.27	-0.28	0.22	0.51	1.00	0.96	-0.03
18	-0.33	-0.16	-0.13	-0.18	0.04	-0.33	-0.31	-0.21	-0.30	-0.34	-0.32	-0.39	-0.26	-0.27	0.21	0.60	0.96	1.00	-0.15
19	-0.20	-0.22	-0.30	-0.17	0.25	-0.06	-0.18	-0.12	-0.13	-0.04	-0.13	-0.03	-0.02	-0.22	0.23	-0.18	-0.03	-0.15	1.00

Fig. 3. Features' correlation matrix $\mathbf{R}$.

The correlation coefficient R_{ij} varies from -1 to $+1$. The lower limit indicates perfect negative correlation, 0 indicates no correlation and the upper limit indicates perfect positive correlation. The color of the cells in Fig. 3 tends to a darker blue with high correlations (positive or negative). Note that the calculation of the coefficient for features 16 to 19 only considered those feeders with PV panels (null values would falsely alter the correlation).

The total number of customers as well as some features related to cable characteristics (6, 7, and 9) and monitoring data (10 to 14) present the highest correlations with other features. The highest correlation coefficient (0.96) was found for features 17 (PV-supplied demand) and 18 (PV penetration level). This is due to the fact that most feeders are purely domestic with similar mean declared capacity per PV installation. However, the presence of both features is important as they allow identifying cases where significant penetration levels are not relevant in terms of the PV-supplied demand, e.g., many houses but with very small PV panels.

Correlations can as well be seen with the scatter plots in Figs. 1 and 2. Features 11 and 14 are more correlated than 1 and 7 as also quantified by the correlation matrix.

Features with correlations too close to unity should not be considered at the same time if they do not allow identifying particular cases in the data set (as explained before). This is to avoid redundancy and bias. Therefore, it can be concluded that the features adopted in this work are adequate and each of them provide relevant characteristics of the feeders.

D. Normalization of Data

Mathematically, the data set $\boldsymbol{X}$ is composed of M patterns defined as vectors $\mathbf{x}_m$, each of which contains H features:

$$\boldsymbol{X} = \{\mathbf{x}_m = \{x_{mh}\}, m = 1, \ldots, M; h = 1, \ldots, H\}. \quad (2)$$

A normalization process was applied to make the selected features to vary in a comparable range. This is required given that the clustering algorithms use the Euclidean distance to determine similarities among samples, and therefore these distances have to be comparable to give appropriate importance (or weight) to each feature [13]. For this purpose, the max-min normalization formula (4) was used to transform the original data set $\boldsymbol{X}$ into the normalized data set

$$\boldsymbol{L} = \{\mathbf{l}_m = \{l_{mh}\}, m = 1, \ldots, M; h = 1, \ldots, H\} \quad (3)$$

whose entries, by denoting with $\min_{m=1,\ldots,M}\{x_{mh}\}$ and $\max_{m=1,\ldots,M}\{x_{mh}\}$ the minimum and maximum values of the h^{th} feature, are

$$l_{mh} = \frac{x_{mh} - \min_{m=1,\ldots,M}\{x_{mh}\}}{\max_{m=1,\ldots,M}\{x_{mh}\} - \min_{m=1,\ldots,M}\{x_{mh}\}}. \quad (4)$$

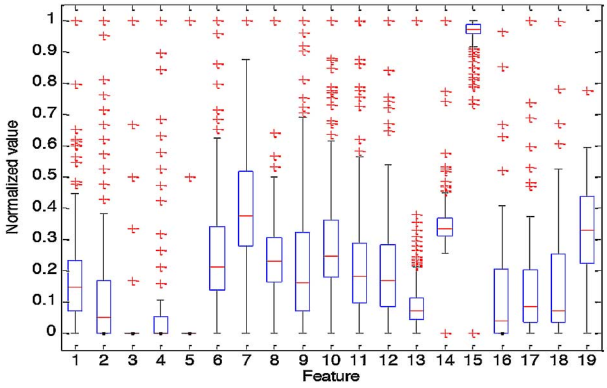


Fig. 4. Normalized features using the min-max formula.

The resulting normalized values for the $H = 19$ features are presented in Fig. 4 as boxplots. These boxplots reveal median values for each feature (medium horizontal bar), the inter-quartile range (rectangle, 50% of a normal distribution probability), the typical extremes (slim black lines, 99.3%) and outlier values (small crosses). Fig. 4 shows that the normalized features maintain the probabilistic distribution they had before the normalization. Features with skewed probabilistic distributions keep concentrating values on the extremes of their range (now limited between 0 and 1) and those with a normal distribution present the inter-quartile range closer to the center of the range.

It is important to highlight that in the context of the clustering process the outliers of those features with skewed probabilistic distributions (farther from mean values) will be penalized more than those from normal distributions. This is because the concept of Euclidean distance is used to determine similarities among feeders.

III. Clustering Process

For data grouping, the possible approaches include clustering analysis (an unsupervised learning technique useful when no a priori information is known about the groups) and the discriminant analysis (a supervised learning technique useful when some information on the groups is known). For the problem under analysis, there is no group structure defined a priori that can be used for training a supervised learning-based method. As such, clustering analysis is applied.

In clustering analysis, the representative patterns of a data set are obtained by grouping similar samples of data according to the features that characterize them. The result is a set of K clusters (groups) each one defined by a centroid, which in turn corresponds to (or leads to) a representative pattern [14]. The ideal clusters are those that are compact (i.e., patterns within the cluster are close to each other) and also distant from each other (i.e., clusters are clearly dissimilar). In this process, the cluster of each pattern is initially unknown and the corresponding assignation is based on a given rule (e.g., based on the Euclidean distance).

The clustering process results in K clusters. Each cluster $k = 1, \ldots, K$ can be defined as the disjoint set $\boldsymbol{L}_k$ with $\cup_{k=1}^{K} \boldsymbol{L}_k = \boldsymbol{L}$ and $\cap_{k=1}^{K} \boldsymbol{L}_k = \varnothing$. Finally, the resulting set of centroids $\boldsymbol{R}$ is composed of K patterns (vectors) $\mathbf{r}_k$ each one associated with its corresponding cluster:

$$\boldsymbol{R} = \{\mathbf{r}_k = \{r_{kh}\}, k = 1, \ldots, K; h = 1, \ldots, H\}. \tag{5}$$

Within the studied LV feeders, it was found that those with PV panels resulted in significant reduction of active power demand, power factors much closer to unity, etc. In some cases, these changes were subtle in comparison to feeders without PV. Consequently, to have a better, independent representation of these particular cases, it was decided to divide the set of 232 feeders into two macro-categories: with (156) and without (76) PV panels. Each macro-category will be treated separately, i.e., forming independent sets $\mathbf{L}$.

A. *Clustering Algorithms*

According to [14] and [15], data clustering algorithms can be generally divided into partitional and hierarchical algorithms. Further details in the categorization may indicate distribution-based and density-based approaches to clustering. This work presents and applies a variety of clustering algorithms, including two different partitional clustering algorithms (improved k-means++ and k-medoids++), one hierarchical algorithm, and one distribution-based algorithm (Gaussian Mixture Model).

1) Improved k-means++*:* k-means groups the set of M patterns (i.e., feeders) into the desired number of clusters K by minimizing the total Sum of Square Distance (*SSE*) between the data and the associated clusters, as shown in the following equation:

$$SSE = \sum_{\mathbf{r}_k \in \boldsymbol{R}} \sum_{\mathbf{l}_m \in \boldsymbol{L}_k} [d(\mathbf{r}_k, \mathbf{l}_m)]^2 \tag{6}$$

where $d(\mathbf{r}_k, \mathbf{l}_m)$ is the Euclidean distance between the pattern $\mathbf{l}_m$ and centroid $\mathbf{r}_k$. The algorithm starts by arbitrarily initializing (locating) the set of K centroids to then associate the closer elements so as to minimize the *SSE*. Once all the elements have been assigned, the algorithm recalculates the new centroids as the average of the elements belonging to each cluster. The whole procedure is repeated until there is no further variation. In this algorithm, the centroids are H-dimensional points (i.e., according to the number of features) calculated as an average, therefore, they are not necessarily an element of the population. Thus, the representative feeder has to be defined as the closest element to the centroid.

The k-means++ clustering algorithm adopts a more effective, distance-based initialization by randomly allocating only the first centroid. The subsequent centroids are defined based on the remaining patterns not closely associated with the previous centroids [16]. Therefore, the farthest patterns will have a higher probability to become centroids. This probability is calculated using the distance of the remaining patterns to the closest previously defined centroids.

Given the stochastic nature of both k-means and k-means++, each of their initializations is likely to have a new location of the first centroids resulting in different solutions. Therefore, typically, the *SSE* is used again to find the best result among a number of runs. This index, however, as it can be seen in (6), considers only the compactness of each cluster without taking into account the dissimilarity among them.

To improve the final set of clusters obtained by the k-means++ algorithm, here it is proposed the use of the *Global Silhouette Coefficient* (*GS*) [17]. This coefficient quantifies the clusters' compactness and similarity, and therefore the solution is expected to be of higher quality.

2) k-medoids++*:* The k-medoids++ algorithm is strongly related to k-means but with the particular difference that the initial and final centroids are elements of the data set. It also adopts the same initialization process of k-means++.

3) Hierarchical Clustering: An agglomerative hierarchical clustering algorithm was adopted. This algorithm starts by considering that each pattern (feeder) of the data set is its own cluster. Then, it merges pairs of clusters based on their distances until all patterns have been grouped together in one unique cluster [14]. In this work, the distances were calculated adopting the Ward's Variance Method (WVM) [18] by which the distance between a pair of clusters is equal to the potential increase in *SSE* if merged.

4) Gaussian Mixture Model: In the distribution-based approach, the data set is modelled by using a mixture of parametric distributions. Using a combination of Gaussian distributions it is possible to formulate the so-called GMM, in which the Expectation Maximization algorithm is used to estimate the GMM parameters [15], [19]. One of the interesting aspects of this algorithm is the possibility of taking into account the correlations among the entries of the data set, represented by the $M \times M$ covariance matrix. In order to avoid having an ill-conditioned covariance matrix, a non-negative regularization value can be added to its diagonal so as to make the matrix positive-definite.

B. *Cluster Assessment*

None of the clustering algorithms presented above (or found in the literature) provide the optimal number of clusters. This has to be determined by assessing the quality of the clusters resulting from different values of K. To perform this cluster assessment in a comprehensive way, four indices are adopted in this work. In addition, these indices will also be used to assess the relative performance of the four clustering algorithms. The first two of the adopted indices were presented in [20]. The formulation of these indices has been adapted in such a way that higher values imply better performance. The values for each indicator will depend on the clustering algorithm and number K of clusters.

They are based on aspects related to the Euclidean distance, structure compactness, distance among different clusters, etc. The set of different distances needed for their calculation are presented in Appendix A

1) Variance Ratio Criterion (VRC)*:* This index not only considers clusters' compactness; it takes into account its relation with the dispersion of patterns within the total population (set $\boldsymbol{L}$):

$$VRC = M\left(1 + \frac{W}{K-1}\right)^{-1}\left(1 - \frac{W}{M-K}\right) \tag{7}$$

$$W = \sum_{k=1}^{K}(n_k - 1)\left(1 - \frac{n_k[\hat{d}(\boldsymbol{L}_k)]^2}{M[\hat{d}(\mathbf{X})]^2}\right). \tag{8}$$

2) Similarity Matrix Indicator (SMI)*:* This index depends on the distance between centroids. It is penalized by the presence of clusters with centroids close to each other:

$$SMI = \left\{ \max_{i>j} \left[\left(1 - \frac{1}{\ln[d(\mathbf{r}_i, \mathbf{r}_j)]} \right)^{-1} \right] \right\}^{-1}. \quad (9)$$

Two more indices were considered. They are based on the silhouette width index (10) which represents how strongly related is a feeder to the cluster it has been associated to [17]:

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}, \text{ silhouette width index for the } i\text{th feeder} \quad (10)$$

$$\text{with } -1 \le s_i \le 1$$

where a_i is the mean distance between the ith feeder and other feeders of the same cluster, and b_i is the minimum among the mean distances between the ith feeder and the feeders of the other clusters (taken cluster by cluster). In practice, b_i represents the dissimilarity existing between the ith feeder and the other cluster (not containing the ith feeder) whose feeders are most similar to the ith feeder. When s_i is close to 1 (i.e., a_i is much smaller than b_i) it means that the ith feeder was assigned to an appropriate cluster. If a cluster contains only one feeder, s_i is conventionally set to zero.

3) Global Silhouette Coefficient (GS)*:* This index provides an idea about the global quality of clusters:

$$GS = \frac{1}{K} \sum_{k=1}^{k} S_k, \text{ global coefficient} \quad (11)$$

$$S_k = \frac{1}{n_k} \sum_{\mathbf{l}_i \in \boldsymbol{L}_k} s_i, \text{ local coefficient.} \quad (12)$$

4) Average Silhouette Coefficient ($AvgSC$)*:* The average value of the silhouette width index (*AvgSC*) for the whole population was also used as an index. Higher values are related to a more adequate allocation of feeders to the corresponding clusters:

$$AvgSC = \frac{1}{M} \sum_{\mathbf{l}_i \in \boldsymbol{L}} s_i. \quad (13)$$

The above four indices provide a quantification of different aspects required to assess the quality of clusters. If multiple indices tend to converge to a specific number K of clusters, then the most robust solution has been found.

IV. Clustering Results

This section presents the clustering process for each macro-category (i.e., without and with PV panels). The process to determine the best clustering algorithm as well as the number of clusters is described and discussed. A cluster inspection is then carried out to obtain truly representative clusters. Finally, the main statistical characteristics of these clusters are presented and discussed.

Due to the heuristic nature of the k-means++ and k-medoids++ they were executed 1000 times. The GMM results depend on the regularization value and on the seed for random number extraction. Hence, the GMM was also executed 1000 times for different regularization values chosen in the range from 0.001 to 0.1 (identified after an initial parametric analysis as a suitable range, i.e., providing relatively good values for the *AvgSC* index for the method and data set used). The values presented correspond to the best solution provided by each algorithm from the parametric analysis.

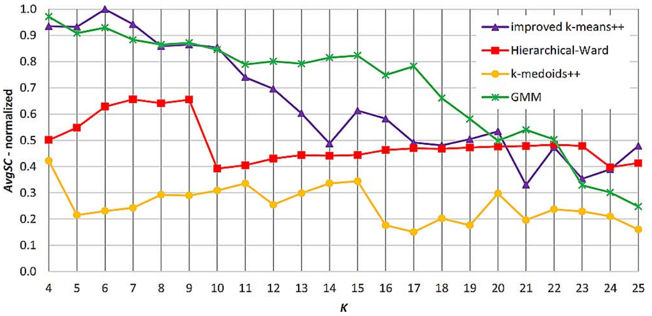


Fig. 5. Normalized *AvgSC* indicator for $4 \le K \le 25$.

A. Macro-Category Without PV Panels

A set of 156 feeders characterized by the features described in Table I was separately clustered. Features 16 to 19 were not considered given the absence of PV panels.

1) Algorithm Selection and Determination of K*:* The macro-category without PV panels was analyzed applying the four clustering algorithms. The results were compared using the four indices previously presented. All indices were calculated without considering clusters comprised of only one feeder as they tend to falsely increase their values making any comparison unrealistic. Indeed, a single-feeder cluster is mathematically considered by the indices as "compact". However, in the context of the production of representative feeders, it becomes an outlier.

For example, Fig. 5 shows the values for the *AvgSC* index for the four clustering algorithms varying the number K of clusters. For visualization purposes, *AvgSC* values have been normalized with respect to the results from the improved k-means++. It can be seen that due to the diverse nature of the four algorithms they result in different values of the index. Therefore, the optimal number of clusters will differ from algorithm to algorithm.

The optimal number of clusters for the 156 feeders has to be found by considering each clustering algorithm and all the indices. Figs. 6 and 7 show the values of the four indices, varying the number K of clusters, for the improved k-means++ and the GMM, respectively. The *VRC* index is not directly suitable to determine the optimal number of clusters as the values obtained for different K cannot be compared. Therefore, the formulation presented in [21] was considered:

$$wVRC_k = (VRC_k - VRC_{k-1}) - (VRC_{k+1} - VRC_k). \quad (14)$$

It is important to mention that this formulation is used to help identifying the optimal K but it is not suitable for assessing the quality of clusters and/or compare the results from different algorithms.

The selection of the optimal number of clusters is not always straightforward. The optimal K needs to come from a compromise between a high performance of the different indices and

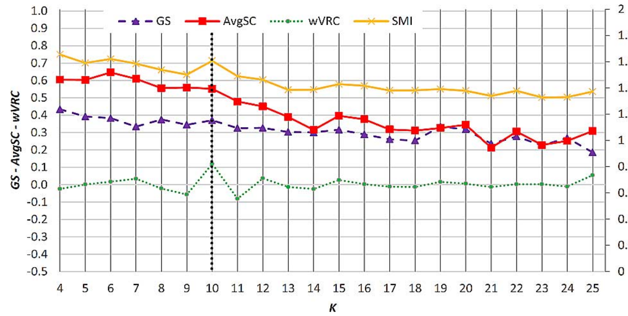


Fig. 6. Improved k-means++ indices' values for $4 \leq K \leq 25$.

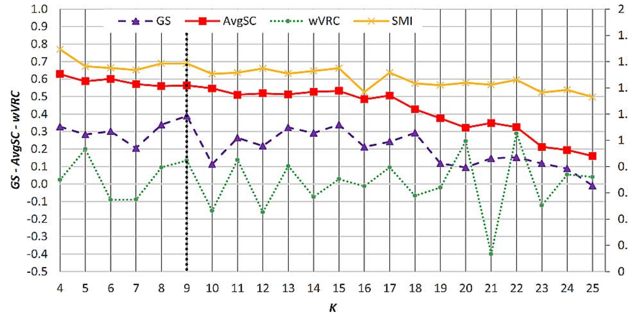


Fig. 7. GMM indices' values for $4 \leq K \leq 25$.

TABLE III
Cluster Assessment (Macro-Category Without PV Panels)

Index	GMM	Improved k-means++	Hierarchical	k-medoids++
wVRC	0.13	0.15	0.02	0.01
SMI	1.58	1.58	1.46	1.36
GS	0.38	0.37	0.33	0.26
AvgSC	0.56	0.55	0.42	0.19
Optimal K	9	10	9	8

an adequate number of clusters. For instance, even if most indices present high values for low values of K there would not be any practical interest in obtaining a very low number of clusters as the level of detail for each one would be limited. Therefore, the evolution of the indices has to be studied in order to obtain a reasonable number of clusters able to properly characterize the data set under study. Because of this and the clear local convergence of all indices at $K = 10$ for the improved k-means++ and $K = 9$ for the GMM, it can be concluded that these values can be adopted as the optimal number of clusters for each algorithm. The same process was repeated for the hierarchical and k-medoids++ algorithm. The optimal number of clusters for each algorithm and the corresponding values of the indices are presented in Table III.

The first important observation is that the four algorithms converge to a very similar number of clusters. The second observation is that the indices found with the improved k-means++ and the GMM algorithms exhibit the best performances (i.e., highest values) compared with the other 2 algorithms for each optimal K.

Given that the indices performance is very similar between the GMM and the improved k-means++, the selection of the best solution requires further analysis. By excluding clusters consisting of only one feeder with uncommon characteristics (within the sample) the GMM algorithm resulted in 7 final clusters (two clusters were excluded) while the improved k-means++ algorithms resulted in 8 final clusters (two clusters were excluded). The improved k-means++ algorithm allowed identifying an extra cluster formed by feeders with high penetration of electric heating (% of customers with electric heating). As this characteristic can be useful for further studies and considering that the indices indicate a very similar quality between both cluster structures, the improved k-means++ for $K = 10$ was adopted as the best solution.

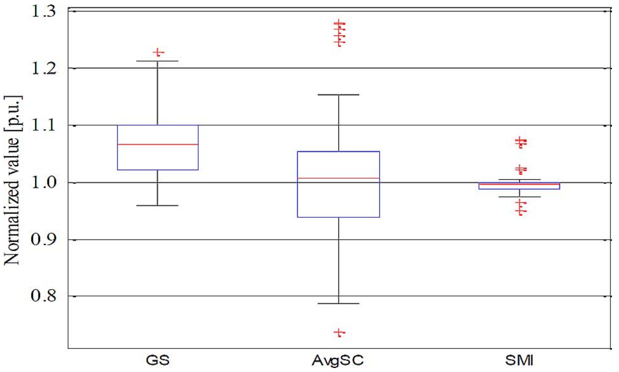


Fig. 8. Improved k-means++ for $K = 10$ versus hierarchical clustering for $K = 9$.

It is important to highlight that the resulting values from the indices *VRC* and *SMI* are strongly dependent on the characteristics of the data set. Therefore, two different feeder clustering studies cannot be directly compared. Only the *GS* and *AvgSC* indices are able to provide this direct comparison. This is due to the fact that they assess the relative adequacy of a feeder to different clusters. Taking this into account, the obtained *AvgSC* of 0.55 suggests a considerably better performance than the 0.22 found in [5].

In order to statistically compare the performance of the improved k-means++ algorithm with the hierarchical algorithm, the results from 100 runs of the improved k-means++ (each run considering 1000 initializations of the centroids) for $K = 10$ are presented in Fig. 8 as boxplots. To facilitate the comparison, values have been normalized with respect to the results from the hierarchical algorithm in Table III as in (15):

$$index_i^{(e)} = \frac{index_i^{(e)^*}}{index_{i_{hierarchical}}} \tag{15}$$

where $index_i^{(e)^*}$ is the non-normalized value of index i for the eth execution and $index_{i_{hierarchical}}$ the value obtained with hierarchical clustering for index i.

It is clear that the improved k-means++ can statistically lead to better results than hierarchical clustering. For instance, in the case of the *GS* index, 91% of the cumulative curve was above 1 per unit.

A detailed comparison of the performance of the improved k-means++ algorithm and the original versions (i.e., k-means++ and k-means) is presented in Appendix B.

2) Final Set of Clusters Without PV Panels: As previously mentioned, the solution from the improved k-means++ for $K = 10$ was adopted. Two of the 10 clusters found consisted of only one feeder with uncommon characteristics (within the sample). These two cases were excluded.

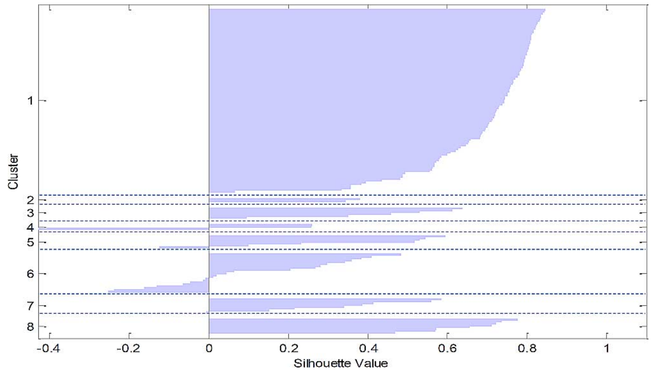


Fig. 9. Silhouette plot for improved k-means++ for $K = 10$ (Macro-Category without PV panels).

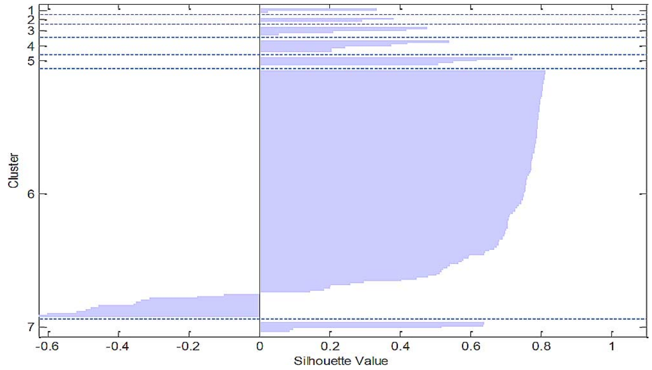


Fig. 10. Silhouette plot for GMM for $K = 9$ (Macro-Category without PV panels).

Fig. 9 presents the final set of 8 clusters in a silhouette plot. On it, feeders within the same cluster are joined together forming a figure similar to a silhouette. The width (vertical axis) of each independent silhouette is related to the number of feeders in the same cluster. The corresponding height (horizontal axis) is equal to the silhouette width index for each feeder. The negative values correspond to feeders that present similar characteristics to other clusters. It can be seen that the silhouette plot mostly presents positive values which means that most of the feeders have been adequately allocated to clusters. This figure also shows that most feeders have been allocated to cluster 1.

Fig. 10 presents the silhouette plot of the final set of 7 clusters from the GMM algorithm. They present the same characteristics obtained with the improved k-means++. Particularly, cluster 6 that consists of small pure-domestic feeders (approximately 30 houses) can be represented by joining clusters 1 (small pure-domestic feeders with small penetration of electric heating) and 8 (small pure-domestic feeders with high penetration of electric heating) from the improved k-means++ algorithm for $K = 10$ (Fig. 9).

The last step towards determining the final set of clusters requires the inspection of the features characterizing them by considering engineering aspects. This is necessary because the values could be found for different combinations of customer types/number and their aggregated demand behavior. However, in this case, it was found that the 8 clusters can indeed be considered as a statistical representation of the analyzed population.

TABLE IV
CLUSTER ASSESSMENT (MACRO-CATEGORY WITH PV PANELS)

Index	GMM	Improved k-means++	Hierarchical	k-medoids++
wVRC	0.09	0.12	0.07	0.01
SMI	1.65	1.51	1.44	1.41
GS	0.49	0.47	0.41	0.31
AvgSC	0.53	0.41	0.34	0.22
Optimal K	10	9	13	17

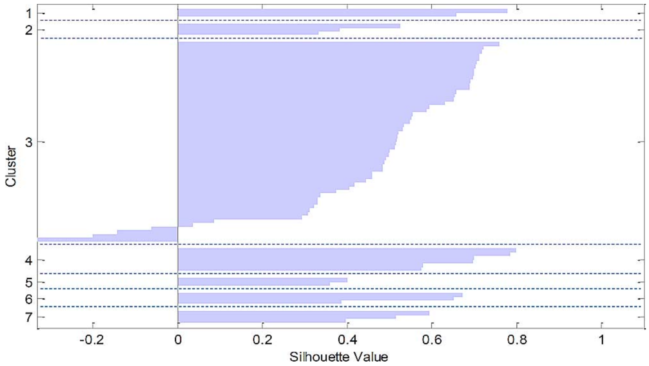


Fig. 11. Silhouette plot for GMM for K=8 (Macro-Category with PV panels).

B. Macro-Category With PV Panels

A set of 76 feeders with declared PV panels characterized by the features described in Table I was separately clustered.

1) Algorithm Selection and Determination of K: The macro-category with PV panels was clustered following the same methodology applied for feeders without PV panels. The four clustering algorithms were applied and the results were again compared using the four indices. The optimal number of clusters for each algorithm and the corresponding values for indices are presented in Table IV.

Most indices found with the GMM algorithm have the best performances (i.e., highest values). Consequently, the most robust clustering approach is, in this case, to adopt the GMM algorithm considering $K = 10$.

2) Final Set of Clusters With PV Panels: Three of the ten clusters found in the previous subsection consisted of only one feeder with uncommon characteristics and, hence, were excluded. The silhouette plot of the final seven clusters is presented in Fig. 11.

After the inspection of the features characterizing the seven clusters, only three clusters (1, 2 and 4; total 11 feeders) were considered as representative clusters with PV panels. Clusters 3, 5, 6, 7, and two of the three excluded clusters (with only one feeder) presented small PV penetrations (less than 5% of the customers) or very low PV-supplied demand resulting in monitoring data similar to that found in feeders without PV panels. Therefore, it can be argued that the features related to the presence of PV panels become irrelevant in those cases. Thus, given the similarity of the monitoring data-related features and the fact that topological features are not influenced by the presence of PV panels, these 64 feeders can be re-allocated to the macro-category without PV panels in order to enrich the sample.

TABLE V
CLUSTER ASSESSMENT (UPDATED MACRO-CATEGORY WITHOUT PV PANELS)

Index	GMM	Improved *k*-means++	Hierarchical	*k*-medoids++
wVRC	0.02	0.16	0.02	0.02
SMI	1.50	1.48	1.41	1.39
GS	0.27	0.35	0.28	0.22
AvgSC	0.55	0.57	0.37	0.22
Optimal K	10	10	9	9

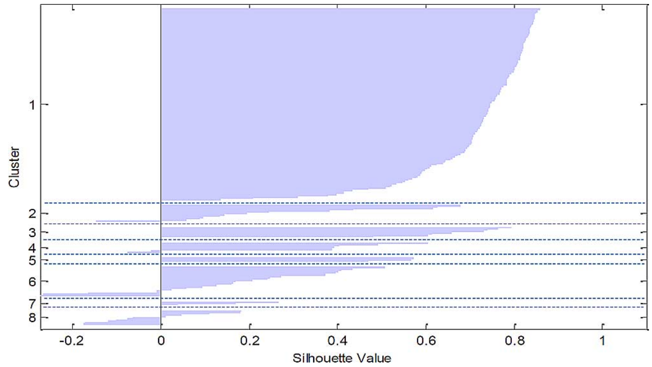


Fig. 12. Silhouette plot for improved k-means++ for $K = 10$ (updated Macro-Category without PV panels).

C. Updated Macro-Category Without PV Panels

After the re-allocation of the 64 feeders excluded from the previous subsection, the process described in Section IV-A was repeated with the new set of 220 feeders. The new cluster assessment is presented in Table V.

The improved k-means++ with $K = 10$ resulted in the best performance for most indices. Two of the ten clusters found had only one feeder (with uncommon characteristics) and, hence, were excluded. The silhouette plot of the final eight clusters is presented in Fig. 12.

As expected, due to the monitoring data and topological similarities of the re-allocated feeders, the new clusters were not significantly different from those obtained previously (the methodology produced again a total of 8 clusters). For instance, the behavior and characteristics of feeders in the new cluster 1 are closely related to those in the previous cluster 1. However, the new data set (41% larger) resulted in a different distribution of feeders within clusters.

D. Statistical Characteristics of the Cluster

Figs. 13–17 present the boxplots corresponding to some of the features of the final set of clusters: 8 without PV (clusters 1 to 8) and 3 with PV (clusters 9 to 11). For all the cases it is possible to identify significant differences between clusters, demonstrating an adequate grouping. By contrasting these boxplots, it is also possible to visually investigate the interdependencies among features.

Figs. 13 and 14 present the total number of customers and the daily mean three-phase active power. The active power shows, as presented in Section II, a strong correlation with the number of customers. However, it can also be seen that for a similar number of customers (e.g., clusters 3 and 4) the median of the active power is different. This is due to the customer type composition that, for cluster 4, includes a significant number of non-domestic customers.

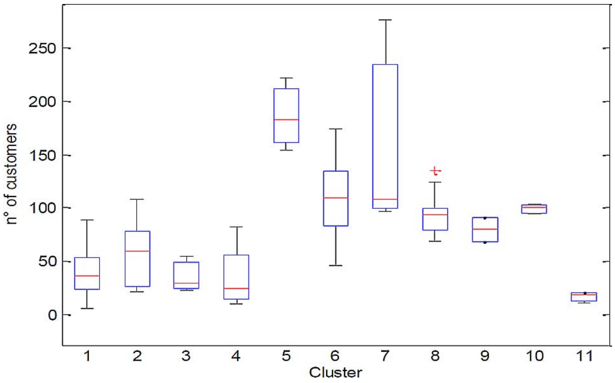


Fig. 13. Total number of customers.

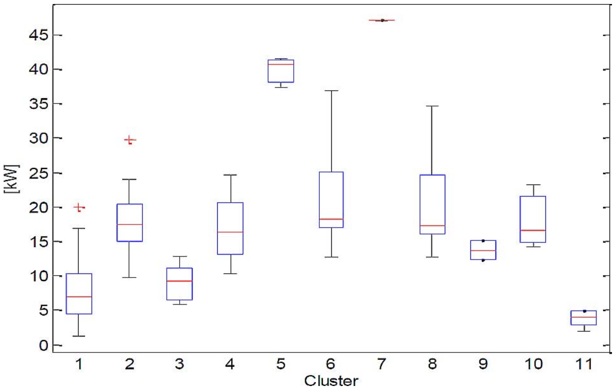


Fig. 14. Mean 3ϕ daily active power.

Fig. 15 shows the neutral current (*In*) among different clusters. Average *In* values were not insignificant due to the unbalanced nature of residential LV feeders in the U.K. (mostly single-phase connected customers). The neutral currents were found to depend on the number and type of customers. For instance, although clusters 1 and 4 have a similar number of customers, *In* values differ significantly. This can be attributed to the high penetration (61%) of non-domestic customers in cluster 4. Interestingly, the clusters with PV panels (clusters 9 to 11) resulted in the lowest neutral currents.

Fig. 16 shows the power factor (PF) values. The average PF was found, with rare exceptions, to exceed 0.95 (traditionally considered for U.K. LV network design). In addition, clusters with PV panels had the highest PF values. The corresponding PV penetrations are shown in Fig. 17.

V. REPRESENTATIVE FEEDERS

For each of the 11 clusters a representative feeder (closest feeder to the centroid) was obtained. The set of representative feeders is listed in Table VI including some of the features used for their characterization. The corresponding distribution, i.e., how representative is each feeder, is also presented.

The type of customer field shows the percentages of the main type of customers per profile class. The remainder to reach the total composition corresponds to PC2 or PC3-PC4 customers if not specified.

TABLE VI
REPRESENTATIVE FEEDERS

k	No. customers	Type of customers (share)	Total cable length [m]	Total Path Imp. [ohm]	Mean 3ϕ daily active power [kW]	In [A]	PV penetration	PV-supplied demand	Distribution
1	31	PC1 / PC2 (90 / 10)	935	2.26	5.75	16.05	N/A	N/A	65.1%
2	38	PC1 / PC3-PC8 (79 / 19)	1591	2.33	16.34	45.54	N/A	N/A	5.7%
3	31	PC1 / PC2 (61 / 39)	561	1.11	8.58	22.80	N/A	N/A	3.5%
4	23	PC1 / PC3-PC4 (39 / 61)	764	1.08	10.28	17.63	N/A	N/A	3.5%
5	222	Mainly PC1 (99)	4589	15.38	37.31	59.43	N/A	N/A	1.3%
6	126	Mainly PC1 (94)	2450	9.29	17.68	36.14	N/A	N/A	10%
7	97	PC1 / PC3-PC4 (70 / 12)	1617	1.31	47.05	53.47	N/A	N/A	1.3%
8	100	PC1 / PC3-PC4 (93 / 5)	1406	7.11	17.00	47.65	N/A	N/A	4.8%
9	91	PC1 / PC3-PC8 (92 / 3)	2708	5.30	15.08	16.57	19%	0.92	0.9%
10	100	PC1 / PC2 (93 / 7)	1912	7.90	14.24	1.59	25%	1.71	1.3%
11	18	Only PC1 (100)	593	0.63	4.77	19.25	39%	3.38	2.6%

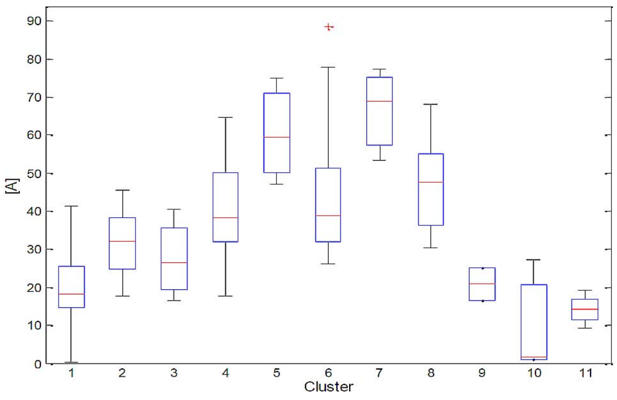


Fig. 15. Neutral current (*In*).

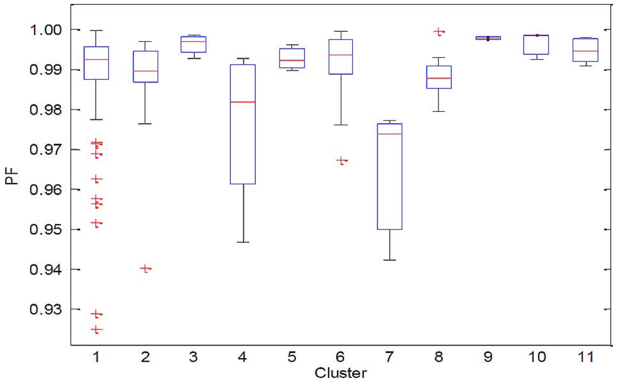


Fig. 16. Power factor (PF).

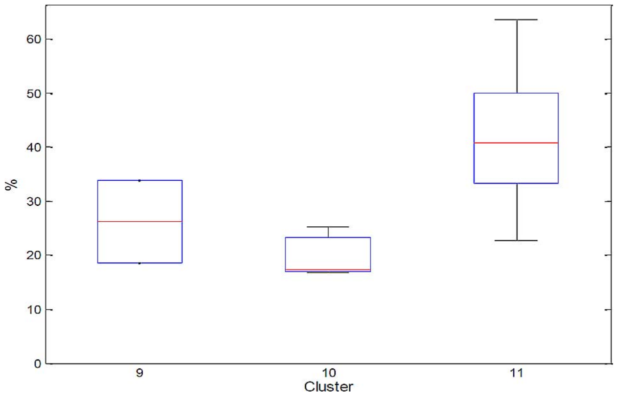


Fig. 17. PV penetration level.

It is important to highlight that each one of the representative feeders is clearly distinguishable from the rest (with respect to their features). For example, representative feeders 1, 5, and 6 (adding up to approx. 76% of the population) correspond to pure domestic feeders with prevalence of PC1 customers but with different monitoring data and topological characteristics.

As shown by the boxplots, the feeders belonging to the same cluster are not exactly equal. However, the representative feeders cover the overall characteristics of each cluster. Consequently, the results from any study performed on them can be extrapolated to the corresponding group, thus reducing the complexity of multiple analyses.

The level of precision of a sample (also called sampling error) can be calculated by using a simplified formula (16)

$$e = \sqrt{\frac{(\frac{N}{n} - 1)}{N}} \tag{16}$$

where n is the sample size, N the population size, e the sampling error, and the confidence level is equal to 95% [22]. Therefore, given that ENWL has approx. 186 890 feeders (33 425 LV substations), the adopted sample of 232 has a statistical accuracy of 93.4% (6.6% of sampling error) with a 95% of confidence.

VI. APPLICATION OF REPRESENTATIVE LV FEEDERS: PV HOSTING CAPACITY ASSESSMENT

This section presents a deterministic analysis in which the capacity of the representative LV feeders to host PV panels is assessed. This is quantified considering voltage rise and thermal issues. Results are presented in a way that allows comparing the findings from analyzing the whole population of feeders (232) and the representative ones.

A. Overview

To assess the PV hosting capacity of LV feeders, time-series power flows are required in order to capture the correlation (or lack of it) between generation and demand. For this, demand profiles are first allocated to all customers according to their Elexon profile class. Then, for a given PV penetration level, customers are allocated generation profiles. Finally, voltage and thermal issues are identified from the resulting day-long power flow analysis.

B. Demand and Generation Profiles

To represent the behavior of customers, the Elexon profile classes presented in Section II are considered (PC1 to PC8). The corresponding demand profiles are obtained by averaging the weekdays in August 2012 (summer).

In the case of PV panels, every installation is considered to have an installed capacity of 3 kWp (aligned to the average found in the 232 feeders). For the solar irradiance, the sunniest day of August 2012 (highest daily mean solar irradiance) is used to create a 5-min resolution profile (necessary to avoid underestimating voltage issues [3]).

C. Methodology

Given a feeder with N customers and a penetration level of $P\%$ (% of customers with PV panels), the general procedure consists of the following steps:

1) Allocation of load profiles to the N customers.
2) Allocation of generation profiles to $P\%$ of the N customers (selected from those farthest to the transformer progressively to those closest to the transformer).
3) Three-phase power flow analysis.
4) Identification of voltage and/or thermal issues.

The above procedure is repeated from 10% to 100% penetration levels with 10% increments.

D. Identification of Technical Issues

Voltage problems are identified by checking compliance of customer connection points with the British Standard BS EN50160 [23]. Considering a nominal line-to-neutral voltage of 230 V, during normal conditions, voltages must be between 0.94 and 1.1 p.u. for at least 95% of data measured in a week (10-min average rms values), and never outside 0.85 and 1.1 p.u. This has been adapted for one day.

Thermal problems are identified by calculating the utilization level at the head of the feeder considering the hourly maximum current and the corresponding ampacity. This current is calculated by averaging the results from the 5-min power flow simulations within each hour.

E. Results

Once all the feeders have been analyzed, the results can be summarized by cluster (only considering the first eight clusters without PV). The boxplots in Figs. 18 and 19 show the hosting capacities for each cluster (i.e., maximum penetration level for which customers do not have technical problems). The figures also show the hosting capacity of each representative feeder (black stars).

Given that the representative feeders describe the overall characteristics of each cluster, their results are expected to be close to the median of each boxplot. This is verified in most cases. In addition, despite the distance to the corresponding median, the representative feeders do show consistency within the clusters in terms of the impacts from PV panels. Clusters where all or most of their feeders have hosting capacity of 100%, have also representative feeders without technical problems. Similarly, those showing lower hosting capacities have representative feeders that do present technical problems with PV panels.

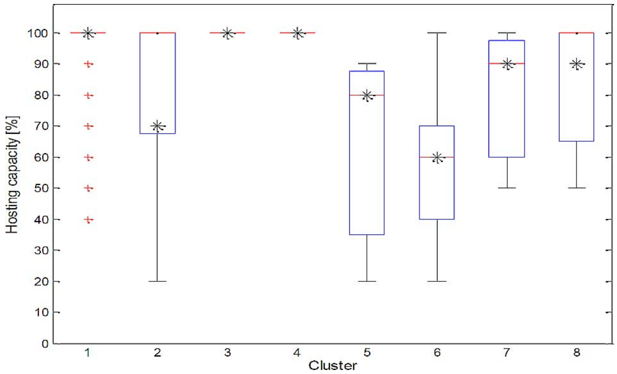


Fig. 18. Hosting capacity due to voltage problems. The stars represent the hosting capacity of the representative feeders.

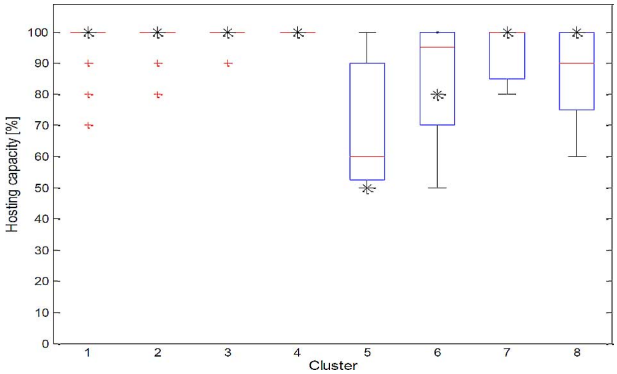


Fig. 19. Hosting capacity due to thermal problems. The stars represent the hosting capacity of the representative feeders.

In the context of LV feeders with low carbon technologies, the results demonstrate that the representative feeders allow identifying types of feeders that are likely to present (or not) problems at a particular penetration level. This is of great interest to DNOs as it can provide useful information to determine corrective or preventive actions. For instance, representative feeders 1, 3, and 4, corresponding to approximately 72% of the analyzed population, do not present thermal or voltage problems for any penetration level. Consequently, it could be said that the LV feeders they represent are robust in terms of PV panels and therefore the DNO can confidently allow high penetration levels of this technology. The findings can be further refined if other more sophisticated approaches (e.g., [3]) are considered.

VII. Conclusions

This work presented a methodology to obtain a set of representative feeders from a sample of 232 monitored residential LV feeders from the North West of England. A final set of 11 clusters and their representative feeders were obtained with a high statistical accuracy.

Multiple validity indices were used to ensure the highest quality of the clusters found by four algorithms. The improved k-means++ and the GMM had the best performances compared with any other previously considered algorithm for similar studies. In addition, the results obtained from the improved k-means++ and GMM were very similar. As such, both methods can be considered as viable alternatives to perform the clustering of LV networks. The methodology developed in this paper is general and can be applied to different sets of feeders within other regions or countries.

It is important to mention that the comparison of the four clustering techniques considered in this work is limited to the studied data set, i.e., other data sets can result in different performances. However, the performance assessment carried out in this work shows that the arbitrary selection of an algorithm can lead to clustering results of lesser quality.

The use of time-series monitoring data, the consideration of PV panels, and the detailed customer classification were also unique aspects of this study. The presence of PV panels required the separation of the original data set into two macro-categories. By adopting monitoring data and detailed customer classification it was possible to cater for the real diversity and behavior of LV feeders. Consequently, these features, in addition to the typical topological ones, were important to enrich the characterization of LV feeders and hence produce truly representative feeders.

Finally, a deterministic PV hosting capacity assessment has been included primarily to allow interpreting the applicability of the representative LV feeders in the context of low carbon technologies. The findings show that it is possible for DNOs to reduce the complexity of assessing the individual feeders and demonstrate that the studies performed on a set of representative feeders can be extrapolated to the population they represent. The accuracy of this extrapolation is linked to the representativeness of the corresponding data set. The findings can be further refined if other more sophisticated approaches (e.g., [3]) are considered.

APPENDIX

A. Set of Distances for Validity Indices

Feeder-to-Feeder: Distance between two patterns $\mathbf{l}_i, \mathbf{l}_j$ of a cluster, each of them having H features:

$$d(\mathbf{l}_i,\mathbf{l}_j)=\sqrt{\frac{1}{H}\sum_{h=1}^{H}(l_{ih}-l_{jh})^2}. \tag{17}$$

Feeder-to-Cluster: Distance between a pattern $\mathbf{l}_i$ and the subset $\boldsymbol{L}_k$ it belongs to. This is calculated as the root mean square of the *feeder-to-feeder* distances between the pattern and each of the n_k feeders belonging to the cluster $\boldsymbol{L}_k$:

$$d(\mathbf{l}_i,\boldsymbol{L}_k)=\sqrt{\frac{1}{n_k}\sum_{\mathbf{l}_m\in\boldsymbol{L}_k}[d(\mathbf{l}_i,\mathbf{l}_m)]^2}\quad \mathbf{l}_i\in\boldsymbol{L}_k. \tag{18}$$

Intra-Class: Calculated by using the *feeder-to-cluster* distance for the n_k feeders of cluster $\boldsymbol{L}_k$:

$$\hat{d}(\boldsymbol{L}_k)=\sqrt{\frac{1}{2n_k}\sum_{\mathbf{l}_m\in\boldsymbol{L}_k}[d(\mathbf{l}_m,\boldsymbol{L}_k)]^2}. \tag{19}$$

The above formulations can be used by replacing the patterns with the centroids (number of features is the same). Likewise, the subset $\boldsymbol{L}_k$ can be replaced with the sets $\boldsymbol{X}$ or $\boldsymbol{R}$.

TABLE VII
IMPROVED k-means++ VERSUS k-means++ AND k-means

Index	Improved k-means++	k-means++	k-means
wVRC	0.15	0.00	0.00
SMI	1.58	1.45	1.44
GS	0.37	0.35	0.27
AvgSC	0.55	0.43	0.38
Optimal K	10	7	7

B. Performance of the Improved k-means++

Here, the performance of the proposed improved k-means++ algorithm is contrasted with the original versions (k-means and k-means++) using the macro-category from Section IV-A. The optimal K for each algorithm and the corresponding values for the indices are presented in Table VII.

As it can be clearly seen, the improved k-means++ resulted in the best performing solution for all the indices.

ACKNOWLEDGMENT

The authors would like to thank Electricity North West Limited, U.K. for providing the data for this work.

REFERENCES

[1] Low Carbon Innovation Coordination Group, Tech. Innovation Needs Assessment (TINA), Electricity Networks & Storage (EN&S) Summary Report, 2012.

[2] Department of Energy & Climate Change, "Domestic energy consumption in the UK between 1970 and 2012," in *Energy Consumption in the UK* ch. 3, 2013.

[3] A. Navarro, L. F. Ochoa, and D. Randles, "Monte Carlo-based assessment of PV impacts on real UK low voltage networks," in *Proc. IEEE Power and Energy Soc. General Meeting*, Jul. 2013.

[4] K. P. Schneider, Y. Chen, D. P. Chassin, and D. W. Engel, "A taxonomy of North American radial distribution feeders," in *Proc. IEEE Power & Energy Soc. General Meeting 2009 (PES '09)*, Jul. 2009.

[5] A. M. Berry, T. Moore, K. K. Ward, S. A. Lindsay, and K. Proctor, National Feeder Taxonomy—Describing a Representative Feeder Set for Australian Electricity Distribution Networks, Report for CSIRO, 2013.

[6] Y. Li and P. J. Wolfs, "Taxonomic description for Western Australian distribution medium-voltage and low-voltage feeders," *IET Gen., Transm., Distrib.*, vol. 8, pp. 104–113, Jan. 2014.

[7] Y. Li and P. J. Wolfs, "A statistical study on topological features of high voltage distribution networks in Western Australia," in *Proc. 20th Australasian Universities Power Eng. Conf. (AUPEC)*, Dec. 2010.

[8] R. J. Broderick and J. R. Williams, "Clustering methodology for classifying distribution feeders," in *Proc. IEEE 39th Photovoltaic Specialists Conf. (PVSC)*, Jun. 2013.

[9] V. Levi, G. Strbac, and R. Allan, "Assessment of performance-driven investment strategies of distribution systems using reference networks," *IEE Proc. Gen., Transm., Distrib.*, vol. 152, pp. 1–10, Jan. 2005.

[10] Load Profiles and their use in Electricity Settlement. Elexon, May 4, 2014 [Online]. Available: http://www.elexon.co.uk/wp-content/uploads/2013/11/load_profiles_v2.0_cgi.pdf

[11] Photovoltaic Geographical Information System (PVGIS). Joint Research Centre website, May 4, 2014 [Online]. Available: http://re.jrc.ec.europa.eu/pvgis/

[12] V. Rigoni and L. F. Ochoa, ENWL LV Network Solutions—Deliverable 3.7 Characterisation of LV Networks (Appendix J), Sep. 14, 2014 [Online]. Available: http://www.enwl.co.uk/lvns

[13] G. W. Milligan and M. C. Cooper, "A study of standardization of variables in cluster analysis," *J. Classification*, vol. 5, pp. 181–204, 1988.

[14] A. K. Jain and R. C. Dubes*, Algorithms for Clustering Data*. Englewood Cliffs, NJ, USA: Prentice Hall, 1988.

[15] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Comput. Surveys*, vol. 31, pp. 264–323, Sep. 1999.

[16] D. Arthur and S. Vassilvitskii, k-means + +: The Advantages of Careful Seeding, Stanford InfoLab, Tech. Rep. 2006-13, 2006.

[17] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *J. Computat. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987.

[18] C. Shalizi, *Distances between Clustering, Hierarchical Clustering*. Pittsburgh, PA, USA: Carnegie Mellon Univ.—Dept. Statist., 2009.

[19] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. B*, vol. 39, no. 1, pp. 1–38, 1977.

[20] G. Chicco, "Overview and performance assessment of the clustering methods for electrical load pattern grouping," *Energy*, vol. 42, pp. 68–80, Jun. 2012.

[21] E. Mooi and M. Sarstedt, "A concise guide to market research," in *Cluster Analysis*. New York, NY, USA: Springer, 2011, ch. 9.

[22] T. Yamane, *Statistics, an Introductory Analysis*, 2nd ed. New York, NY, USA: Harper and Row, 1967.

[23] British Standards Institution, BS EN 50160: Voltage characteristics of electricity supplied by public distribution systems, 2000.

Valentin Rigoni received the B.Eng. degree in mechanical and electrical engineering from UNC, Córdoba, Argentina, in 2014 and the M.Sc. degree in electrical engineering from Politecnico di Torino (PdT), Torino, Italy, in 2014.

He was a Visiting Researcher at The University of Manchester, U.K. His research interests include power system and distribution system analysis and integration of distributed energy resources.

Luis F. Ochoa (S'01–M'07–SM'12) received the B.Eng. degree from UNI, Lima, Peru, in 2000 and the M.Sc. and Ph.D. degrees from UNESP, Ilha Solteira, Brazil, in 2003 and 2006, respectively.

He is a Senior Lecturer in Smart Distribution Networks at The University of Manchester, U.K. His current research interests include network integration of distributed energy resources and future low-carbon distribution networks.

Gianfranco Chicco (M'98–SM'08) received the Ph.D. degree in electrotechnics engineering from Politecnico di Torino (PdT), Torino, Italy, in 1992.

Currently, he is a Professor of Electrical Energy Systems at PdT, Energy Department. His research interests include power system and distribution system analysis, energy efficiency, multi-generation, load management, artificial intelligence applications, and power quality.

Prof. Chicco is a member of AEIT.

Alejandro Navarro-Espinosa (S'08) received the Industrial Engineer and M.Sc. degrees from the Pontificia Universidad Catolica (PUC), Chile, in 2004 and 2007, respectively. In 2011, he also received the M.Sc. degree in power systems from The University of Manchester, U.K., where he is currently pursuing the Ph.D. degree in future low voltage distribution networks.

Previously, he was Technical Chief Executive at Systep and lecturer at Universidad de los Andes, Chile.

Tuba Gozel (S'00–M'14) received the B.Eng. degree from Selcuk University, Konya, Turkey, in 1994 and the M.Sc. and Ph.D. degrees from Gebze Institute of Technology, Kocaeli, Turkey, in 2002 and 2009, respectively.

She was a Research Associate at The University of Manchester, U.K. Currently, she is a Research Assistant at Gebze Technical University, Kocaeli, Turkey. Her research interests include power system analysis and distribution networks.