Agglomerative Hierarchical Clustering Applied to Medium Voltage Feeder Hosting Capacity Estimation

Seddik Yassine Abdelouadoud, Sébastien Vallet, Robin Girard

Agglomerative Hierarchical Clustering Applied to Medium Voltage Feeder Hosting Capacity Estimation

Seddik Yassine Abdelouadoud
Centre for Processes, Renewable Energy and Energy Systems (PERSEE)
Mines Paris - PSL University
Sophia Antipolis, France
yassine.abdelouadoud@minesparis.psl.eu

Robin Girard
Centre for Processes, Renewable Energy and Energy Systems (PERSEE)
Mines Paris - PSL University
Sophia Antipolis, France
robin.girard@minesparis.psl.eu

Sébastien Vallet
Roseau Technologies
Grenoble, France
sebastien.vallet@roseautechnologies.com

***Abstract*—In this paper, we will assess the relevance of applying hierarchical agglomerative clustering algorithms on medium voltage feeder descriptive parameters (average line impedance, capacity, total length, etc.) in order to select representative feeders for long term planning studies. To achieve this, we start by creating a dataset of medium voltage feeders (bus and line geometries and characteristics) by combining domain knowledge with open datasets of distribution network layouts, district level electricity consumptions and building footprints. We then use this dataset to calculate descriptive attributes for each feeder as well as their maximal load hosting capacity and the associated type of constraint. Afterwards we perform a statistical analysis of the descriptive attributes in order to select the most relevant to use as inputs of the clustering algorithms. Finally, we apply a hierarchical agglomerative clustering algorithm with varying number of clusters, assess the quality of the results using internal and external validation and evaluate the ability of the medoids of each cluster to represent the behavior of the corresponding feeders.**

***Index Terms*—medium voltage feeders; hosting capacity; clustering**

I. Introduction

To achieve their long term goals of decarbonisation, many countries are planning to replace their use of fossil fuels by electric alternatives (e.g. EVs and heat pumps) along with the development of low-carbon electricity generation, including distributed renewable energies. From the standpoint of the electricity distribution network, this will entail shifts in loading conditions and power flows (including the possible appearance of reverse power flows). The extent to which these shifts will result in changes in operational and investment costs will depend on the electrical and topological characteristics of each medium and low voltage network as well as the geographical distribution of the associated loads and generations and their future evolution. As a consequence, evaluating the costs and benefits of a decarbonisation strategy (including the deployment of Smart Grid technologies such as distributed storage or VAR control) would, in theory, require the simulation of projected load flows for every medium and low voltage network. As this would be impractical for large distribution networks comprising thousands of medium voltage networks and hundreds of thousand of low voltage networks, significant research interest has been devoted to identifying "representative" or "typical" medium or low voltage feeders. The purpose of this effort is to allow complex network studies to be performed on a limited subset of feeders, from which results for the complete set can be extrapolated. The first example of such a study is [1], in which 12 representative medium voltage feeders are selected using a K-Means clustering algorithm. An external validation is performed by comparing the results of network studies (voltage drop, total losses, power factor, etc.) performed on the entire system with those extrapolated from the representative feeders. The second example is [2] which used an eigenvalue approach to clustering to select representative feeders based on their impedance and active and reactive power flows. After the authors of [3] (summarized in [4]) produced a "taxonomy of 24 prototypical feeder models" for the U.S. Department of Energy, a renewed interest in the subject led to the publication of several studies following a similar pattern such as [5], [6] or [7] :

1) Computation of feeder parameters
2) Elimination of outliers or erroneous feeders
3) Statistical study of feeder parameters (e.g. analysis of correlation and principal components)
4) Selection of parameters to use as inputs of the clustering algorithm
5) Selection and design of the clustering algorithm, including the selection of the number of clusters
6) Interpretation and/or validation of the results

The authors in [8] and [9] provide a complete review of such studies, in particular by categorizing them according to the type of clustering algorithm and the scales of both the system studied and the representative networks selected. In this article, we will adopt a similar approach and focus in particular on examining to what extent the networks selected

979-8-3503-9678-2/23/.00

can be used to extrapolate results on hosting capacity for the complete system. To achieve this, we will introduce quality measures of the clustering results related to the ability of the representative feeders to predict both the type of constraint and the hosting capacity of the feeders in their cluster.

II. Dataset creation

Obtaining a large enough, homogeneous and coherent network dataset (including network topology, line characteristics and load distributions) is the first step in the selection of representative networks. In this article, we will achieve this by relying on open datasets made available in France under a 2016 law governing the release of data produced by public entities. As these datasets do not contain all the information we need, we will combine them with domain knowledge to fill in the gaps. While the datasets are available for all of mainland France, we will focus in this study on Auvergne-Rhône-Alpes, a large region with a diverse mix of climate conditions and population densities.

A. Data sources

We will make use of three open datasets to create the network dataset :

1) Distribution network layout. This dataset contains the geographic positions of HV/MV and MV/LV substations as well the geometries of the MV and LV lines. Overhead and underground lines are differentiated but no other characteristics are known. Currently the dataset is available only for the parts of the network managed by the french DSO ENEDIS (95% of the country size).
2) Local energy consumption. This dataset contains the annual energy consumptions (electricity and utility gas) as well as number of customer by sector (residential, industry, service sector, agriculture) and by IRIS[1].
3) BD TOPO®, produced by IGN. This dataset contains, among other things, the footprints, sector (residential, industry, service sector, agriculture) and number of floors of every building

B. Network topology creation and line characteristics allocation

To establish the topology (i.e. how the lines and buses are connected), the first step is to create the set of feeders supplied by a given HV/MV substation by selecting the lines closest to the substation. Then the feeders are grown by connecting subsequent lines according to the proximity of their extremities. If a line can be connected to more than one feeder, it is connected to the shortest one. Buses are created at the points of connection between two lines. If a MV/LV substation is close to a bus, it is merged with it.

To allocate the line characteristics, two different procedures are applied for overhead and underground lines. For overhead lines, the backbone of the feeder is identified by iteratively adding to the backbone the line with the most downstream MV/LV substations, starting from the HV/MV substations. The procedure is stopped when there are no lines with more than 10 downstream MV/LV substations. Then the characteristics of 148 mm² cables are allocated to lines that are part of the backbone, while the other lines receive the characteristics of 54 mm² cables. For the underground lines, the feeders are separated into two categories : "dedicated" feeders supplying less than 5 MV/LV substations, for which the characteristics of 240 mm² cable are allocated, and other feeders for which the characteristics of 150 mm² cable are allocated.

This algorithm has been designed and implemented by Roseau Technologies.

C. Bus annual and peak load allocation

To allocate the annual and peak loads to each bus, we use the local energy consumption and BD TOPO®datasets in the following manner :

1) Determine the buildings inside each IRIS by performing a spatial join between building footprints and IRIS borders using the "within" relationship. For the building footprints that intersect several IRIS, assign them to the IRIS for which the intersection has the largest area.
2) For each building, calculate the floor area by multiplying the footprint area by the number of floors. If it is unknown, assume that there is one floor.
3) For each IRIS and each sector, allocate the annual load and number of customers to each building of the same IRIS and sector proportionally to its floor area
4) Define the building connected directly to the MV network as those for which the annual load by customer is above a given threshold. The threshold is set to 70 MWh/customer so that the number of MV customers fits publicly available country level statistics on MV customers.
5) Define a bus as supplying a MV customer when a MV/LV substation is present but no public LV network is connected to it.
6) Perform a spatial join between LV-supplied building footprints and LV buses using the "closest" relationship and aggregate the annual loads and number of customers if a bus is the closest one to several buildings. If there are multiples buses that are the closest ones for a given building, divide the annual loads and number of customers by the number of buses and allocate the result to them. Do the same for MV-supplied buildings and MV customer buses.
7) For each MV/LV substations with LV feeders connected to it, aggregate the annual loads and numbers of customers of the LV buses it supplies.
8) Compute the peak load by dividing the annual loads by a factor of 3900 hours for LV customers and 5500 hours for MV customers. These factors have been computed using country level open data on 30-minute resolution load data.

[1]Sub-municipal territorial division created for the purpose of statistics publication

D. Medium to low voltage transformer nominal power allocation

As we aim to compute the load hosting capacity of MV feeders, it is necessary to estimate the nominal power of MV/LV transformers as this can constitute a limit to the electricity supplied by a given feeder. To achieve this, we rely on domain knowledge of the nominal powers of standard transformers (between 50 kVA and 1000 kVA) and the associated share in the existing stock.

To assign a nominal power to each MV/LV substations substations supplying LV customers, we sort them by increasing order of peak load and associate them with increasing transformer nominal power while respecting the share of nominal powers in the existing stock. When the peak load is above 1000 kW, the first multiple of 1000 kW above the peak load is selected. The same procedure is applied to substations supplying MV customers.

E. Descriptive attributes

Using the procedure described above, we obtain a set of 4551 MV feeders supplied by 391 HV/MV substations, for which the topology, geometry, line characteristics and MV/LV substation nominal powers are known. From this, we compute a set of 24 descriptive parameters taken mainly from [3] and [8], with particular care taken to include parameters relevant to hosting capacity estimation.

F. Elimination of erroneous feeders

Due to the exhaustivity of the source datasets, special purpose MV feeders (e.g. backup feeders) are present in the resulting dataset. After analyzing a sample of feeders, we have decided to remove the feeders with a total line length below 50 m or a total nominal power below 250 kVA. This corresponds to 395 feeders, or 8.7 % of the sample, but only 0.53 % of the total line length and 0.39 % of the total nominal power.

G. Load hosting capacity

The load hosting capacity of each feeder is computed by performing a sequence of power flows[2] with increasing apparent power supplied to the MV/LV substations. At each step, the apparent power supplied to every substation is increased by 0.5 % until it reaches 110 % of the nominal power. Then, the compliance with voltage and thermal line rating constraints is evaluated for each step. If a constraint is violated at step N, the hosting capacity is set to the total active load supplied at step $N-1$ and the type of constraint ("voltage" or "line rating") is attributed to the feeder. If the constraints are violated in none of the steps, the hosting capacity is set to the total active load supplied at the last step and the type of constraint "transformer rating" is attributed to the feeder. The voltage compliance check is performed for minimal voltages ranging from 0.95 p.u. to 1.02 p.u. in order to evaluate the sensitivity to voltage quality requirements.

[2]BFS using a three-phase balanced, radial, line impedance model as described in [10]. The source voltage is set to 1.03 p.u.

III. Clustering

A. Variable selection

Using correlated variables as clustering inputs can lead to bias, redundancy and create a false impression of cluster separation. To prevent this, we will select a subset of "poorly-correlated" variables, using a methodology heavily inspired by [8]. Concretely, we calculate the Spearman correlation coefficient for each pair of variables and apply a hierarchical clustering algorithm using the distance defined in Eq. 1.

$$Dist(x,y) = 1 - |Corr^{spearman}(x,y)| \tag{1}$$

The rationale for such a procedure is to identify clusters of variables among which the correlation (negative or positive) is high. Fig. 1 displays the evolution of the distance between merged clusters as a function of the number of clusters, with an "elbow" suggesting the presence of 6 clusters.

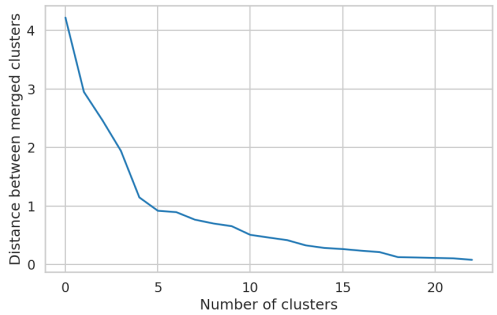

Fig. 1: Distance between merged clusters

For each of the 6 clusters, we choose the variable with the highest relative standard deviation as its representative :

- Overhead line mean length (in km)
- Maximal line thermal capacity (in kVA)
- Total line length divided by total nominal power (in km/kVA)
- Total nominal power of MV/LV substations (in kVA)
- kWΩ (see [8])
- Impedance closeness

B. Algorithm description

While K-means clustering has drawbacks in the context of representative feeder identification [3], it has been the most widely used for this purpose to date as it is more computationally efficient than the alternatives on large datasets. For this study, we have elected to use an agglomerative hierarchical clustering (AHC) with a Ward linkage and euclidean distance by taking advantage of the efficient implementation provided by [11]. Input variables are standardized by removing the mean and scaling to unit variance. The number of clusters considered varies from 10 to 4150 by increment of 10. For each cluster, the medoid (i.e. the feeder whose average distance with the other cluster member is the smallest) is used as representative.

[3]Difficulty in identifying non-globular, unevenly sized clusters

C. Internal validation

Internal validation of the clustering results is often used to assist in selecting a suitable number of clusters. It relies on quality measurements that depend only on the input variables. For AHC using Ward linkage, a straightforward option is to use the distance between the two clusters merged at any given step. In that case, the elbow method is applied to select the number of clusters. Another widely used quality measurement is the silhouette score, whose value should be maximised. Fig. 2 presents the two quality measurements for clusters counts varying between 2 and 100. Using the distance between merge clusters, the optimal number of clusters seems to lie between 20 and 30, while using the silhouette score it seems to lie between 10 and 15.

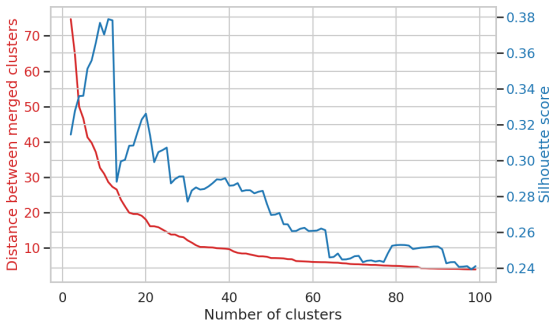

Fig. 2: Clustering internal quality measurements

D. External validation

External validation consists in confronting the clustering results to an external measurement of quality, i.e. one not relying on the variables used in the clustering process. In general, this is performed after the internal validation, in order to verify that the number of clusters chosen is suitable for the purpose of the study. In our case, we want to ensure that the representative feeders are able to accurately estimate both the hosting capacities and the types of constraint of the feeders from their cluster. To illustrate this, Fig. 3 contains a scatter plot of actual feeder hosting capacity (minimal voltage set to 0.98) as a function of their representative feeder hosting capacity in the case of 30 clusters.

While this number of clusters was deemed suitable according to internal measurements, we can observe that the correlation between actual and predicted values is poor and that some clusters seem to contain feeders with different types of constraints.

By building on the works in [8] and [12], we propose two external quality measures that can be computed for each number of clusters :

- The coefficient of determination of the predicted hosting capacity relative to the actual one, which aims at measuring the ability of the representative feeders to predict the hosting capacity of their cluster members.
- The share of feeders with a type of constraint identical to their representative feeder, which aims at measuring the ability of the representative feeders to predict the type of constraint of their cluster members.

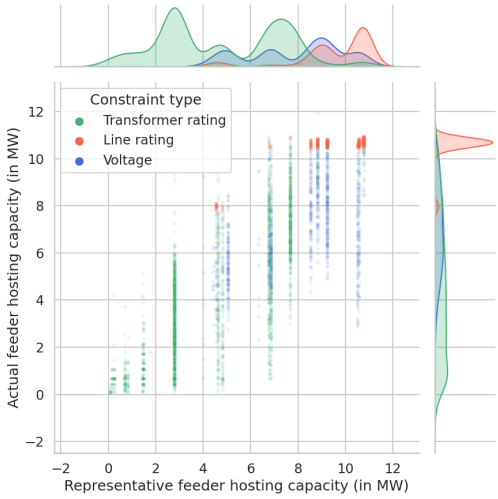

Fig. 3: Actual vs predicted hosting capacities (30 clusters)

As an example, in the case presented in Fig. 3, the coefficient of determination is 0.66, while the share of feeders with a correct type of constraint is 81.4 %. While the specific level of accuracy needed is purpose-dependent, it is safe to say that such a low level would not be suitable for most of the applications of representative that have been envisioned. Fig. 4 and Fig. 5 show the evolution of the two external quality measurements for numbers of clusters ranging from 10 to 1000, and for the range of minimal voltages considered.

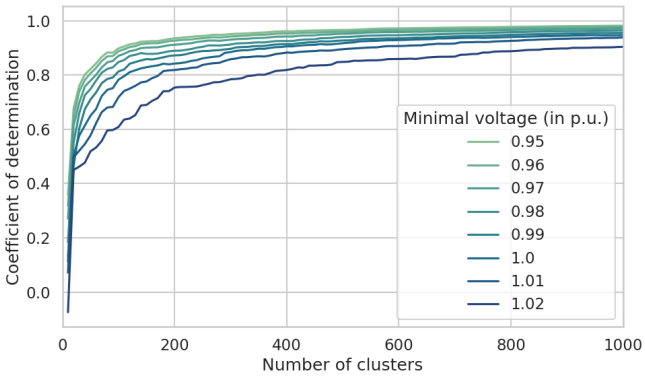

Fig. 4: Quality measurement of hosting capacity prediction

E. Result interpretation

In a broad sense, we have managed to reproduce both the results in [12], where large dispersions of hosting capacities inside clusters were observed when using numbers of clusters in the range of 10 to 30, and the results of [8], where relatively low levels of constraint type purity were observed when using numbers of clusters below 10. However, it should be noted that the quality measurements used in those studies (boxplot-based

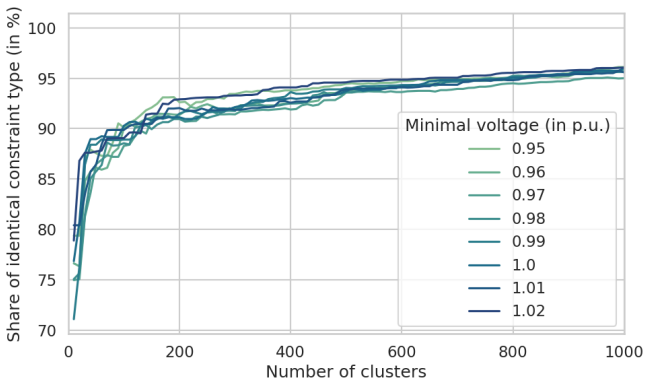

Fig. 5: Quality measurement of constraint type prediction

in [12] and cluster purity in [8]) only aim at measuring the level of homogeneity inside clusters, while the measurements introduced here aim at measuring the quality of the prediction provided by the representative feeders. As a consequence, these can be used to select a number of clusters (and thus representative networks) to use in a particular study as a function of the level of accuracy expected for the study.

We can also observe that both quality measurements increase rapidly up to around 200 clusters, after which the incremental gains accrued by increasing the number of clusters become marginal. Thus we can consider that, absent specific accuracy requirements, the choice of 200 clusters represents a good compromise between number of clusters and quality (see Fig. 6 for a scatter plot with 200 clusters). We can remark that this is one to two orders of magnitude higher than what is generally considered an optimal choice of number of clusters (see Table 2 in [9], in which 10-25 is considered a "large number" of clusters).

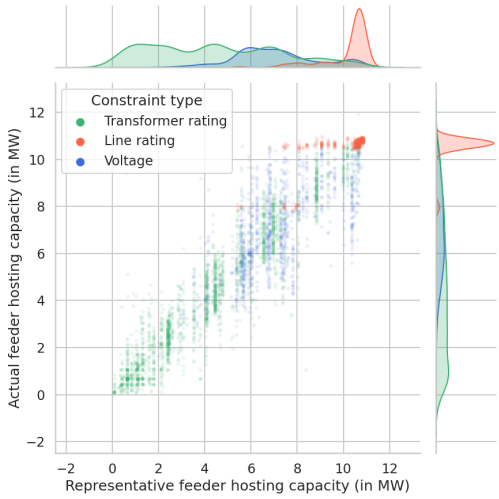

Fig. 6: Actual vs predicted hosting capacities (200 clusters)

IV. Discussion and conclusion

In this study, we have applied an agglomerative hierarchical clustering algorithm to a set of more than 4000 MV feeders in order to extract representative feeders. We have then employed internal quality measurements that are generally used to determine the optimal number of clusters and have shown that this approach leads to a number of clusters that is too small to adequately represent the diversity of feeder behaviour in terms of load hosting capacity. This has led us to introduce two new external quality measurements related to the hosting capacity. These new measurements suggest that the appropriate number of clusters is one to two orders of magnitude higher than what is generally admitted. While this result is significant in itself, several avenues of inquiry remain open, in particular pertaining to its generalization to other clustering algorithms, feeder datasets (especially larger datasets) and study objectives (e.g. PV hosting capacity or Volt-Var control).

References

[1] H. Willis, H. Tram, and R. Powell. A Computerized, Cluster Based Method of Building Representative Models of Distribution Systems. *IEEE Transactions on Power Apparatus and Systems*, PAS-104(12):3469–3474, December 1985.

[2] S.S. Fouda. Eigenvalue approach clustering algorithm for building equivalent models of distribution systems. *IEE Proceedings - Generation, Transmission and Distribution*, 142(3):282, 1995.

[3] K. P. Schneider, Y. Chen, D. Engle, and D. Chassin. A Taxonomy of North American radial distribution feeders. In *2009 IEEE Power & Energy Society General Meeting*, pages 1–6, Calgary, Canada, July 2009. IEEE.

[4] Kevin P. Schneider, Yousu Chen, David P. Chassin, Robert G. Pratt, David W. Engel, and Sandra E. Thompson. Modern Grid Initiative Distribution Taxonomy Final Report. Technical Report PNNL-18035, 1040684, November 2008.

[5] Yingliang Li and P. Wolfs. Statistical identification of prototypical low voltage distribution feeders in Western Australia. In *2012 IEEE Power and Energy Society General Meeting*, pages 1–8, San Diego, CA, July 2012. IEEE.

[6] Robert J. Broderick and Joseph R. Williams. Clustering methodology for classifying distribution feeders. In *2013 IEEE 39th Photovoltaic Specialists Conference (PVSC)*, pages 1706–1710, Tampa, FL, USA, June 2013. IEEE.

[7] Valentin Rigoni, Luis F. Ochoa, Gianfranco Chicco, Alejandro Navarro-Espinosa, and Tuba Gozel. Representative Residential LV Feeders: A Case Study for the North West of England. *IEEE Transactions on Power Systems*, 31(1):348–360, January 2016.

[8] Benoît Bletterie, Serdar Kadam, and Herwig Renner. On the Classification of Low Voltage Feeders for Network Planning and Hosting Capacity Studies. *Energies*, 11(3):651, March 2018.

[9] Attila Sandor Kazsoki and Balint Hartmann. Hierarchical Agglomerative Clustering of Selected Hungarian Medium Voltage Distribution Networks. *Acta Polytechnica Hungarica*, 17(4):201–219, 2020.

[10] SY Abdelouadoud, Robin Girard, François-Pascal Neirac, and Thierry Guiot. Optimal power flow of a distribution system based on increasingly tight cutting planes added to a second order cone relaxation. *International Journal of Electrical Power & Energy Systems*, 69:9–17, 2015.

[11] Daniel Müllner. fastcluster: Fast hierarchical, agglomerative clustering routines for r and python. *Journal of Statistical Software*, 53:1–18, 2013.

[12] Robert J. Broderick, Karina Munoz-Ramos, and Matthew J. Reno. Accuracy of clustering as a method to group distribution feeders by PV hosting capacity. In *2016 IEEE/PES Transmission and Distribution Conference and Exposition (T&D)*, pages 1–5, Dallas, TX, May 2016. IEEE.