LV feeder clustering

Alexander Hoogsteyn

August 2020

1 Introduction

Due to the rise of PV installations and heavy loads such as electrical-vehicle chargers in LV grids it is of interest to the local DSO to analyze the influence of these recent trends on his network. However, these LV networks often contain thousands of feeders and therefore it is of interest to obtain a subset of these feeders that are representative for the whole network. Results obtained for this subset can then be extrapolated to the whole network which simplifies the analysis to be carried out. Usually this is done by means of clustering algorithms such as Hierarchical clustering, K-means++, K-medoids++ or Gaussian mixture model as described in [12]. In this report the state of the art of clustering for LV feeders is firstly discussed. Secondly, the results of V. Rigoni et al. [12] are firstly reproduced for a data set of 160 LV feeders in a Spanish urban area. Thirdly, there is expanded upon the work of Rigoni et al. using state of the art clustering techniques. Finally, conclusions are drawn.

2 State of the art of LV feeder clustering

The K-means algorithm and its variants are the de facto standard for clustering of LV feeders. K-means aims to partition a data set in clusters by allocating each point to the nearest 'mean'. Then new 'means' are calculated for the clusters by finding the point with the minimal squared Euclidean distance to it's data points. This is repeated till convergence. K-means++ adapts the initialization procedure of K-means to enhance the computational efficiency. While K-means initializes the 'means' i.e. cluster centers randomly, K-means++ does this for the first cluster center only, thereafter it allocates the centers using a probabilistic distribution which is proportional to the squared distance from the point to the nearest existing cluster center. Thus improving the chance that the farthest patterns will have a center initialized close to it. K-means is used for obtaining representative feeders in a LV feeder network in [5], [2], [1] and [14]. Variants of K-means++ such as K-medoids++, which use data points as clustering centers instead of being able to use any point, are used in [7] and [4]. Another adaption to K-means is to use partial assignment of data points to clusters, this method is referred to as "fuzzy K-means". This can be useful since LV feeder data often shows no clear separation of clusters therefore it can make sense to allocate data points partly to multiple clusters. Because of this reason, such an approach is used in [11]. According to [9], clustering using K-means is still an ill-posed problem. It is hard for K-means to converge to a global minimum and it is questionable whether minimizing average squared Euclidean distance is the best approach (since it naturally tends towards circular clusters and fails to detect arbitrarily shaped clusters) Therefore, a more robust simple statistical approach is often taken such as in [8] were data stratification is used. In data stratification the data is separated based on a predetermined set of criteria.

Others recommend improvements to the K-means algorithm. In [10] an algorithm is described to increase the reproducibility of the outcome of the clusters. A. Maniar et al. combine different cluster validation indices (CH[1], DB[2], D[3] and SI[4]). In their algorithm the solutions are sorted according to all indices after a

[1] Calinski–Harahasz index
[2] Davies–Bouldin index
[3] Dunn's indices
[4] Silhouette index

number of repeats and the best solution is selected as the highest ranking common solution. According to their findings this will reduce the variance found in the solution especially for large data sets. These results were obtained for clustering load profiles i.e. for clustering certain consumer types, not for obtaining representative feeders for the network. In [3] clustering is performed for predicting PV hosting capacity. They improved the accuracy of K-means by weighting the features according with their correlation factor with PV hosting capacity (higher weight for more correlated features). This is different to the goal for finding reference feeders since in the latter you want to capture all the variety in the data set of LV feeders rather than clustering according to a single property as is the case in [3]. Having minimal correlated features will result in clusters that better capture the variety in a data set of feeders.

Clustering is well researched field of machine learning. Its challenges are well described in [9] which although it dates from 2010, is still quite relevant today. The author, Anil K. Jain, describes that a recent trend in the field is to use clustering ensembles. The idea is to combine different clustering algorithms, cluster sizes or different data to obtain superior clusters. the relevance of this to LV clustering is explored in section 5.

3 Method

The clustering is performed using several features that characterize the feeder. The features were normalized such that their minimum and maximal value are 0 and 1 respectively. This will ensure that the different features with incomparable units are equally accounted for while preserving the probabilistic distribution within features which are skewed. Thirteen feeders did not contain any customers at all and were therefore excluded from the data set. For each remaining feeder in the data set the following information is gathered and used as features:

Number of customers For each feeder the total number of customers on that feeder was calculated

Average active energy consumption This data was obtained by aggregating measurements from smart meters in the network. The total annual energy consumption of each customer was calculated. Then the average energy consumption per customer is calculated for each feeder. This is preferred over the total consumption in one feeder because the former is less correlated with the number of customers. For customers for which this data was not available, an estimation was made based on the average consumption of all customers that have the same type of connection (1 or 3 phase).

Average reactive energy consumption idem as above

Main path length For each feeder the longest path was calculated from a customer to the head of the feeder, this is referred to as the main path length

Average path impedance For each feeder the impedance from each customer to the head of the feeder was calculated. The Average path impedance is then taken.

Four different clustering techniques were performed: Hierarchical clustering, K-means++, K-medoids++, and Gaussian mixture model. Due to the statistical nature of the last Three they were performed 1000 times and the best results are kept according to their average silhouette score. A points silhouette score is a measure of how similar that point is to its cluster. Scores range from 1 to -1 with 1 indicating high similarity. Solutions are then compared by comparing the average silhouette score of all points in the data set. Alternatively the global silhouette score could be used which takes the average silhouette score within each cluster and then averages over all clusters. The global silhouette score is also evaluated but not used for validation. Validation is done using average silhouette score as in [12]. Since it is hard to predict a priori which number of clusters will yield the best separation the experiment was repeated for cluster sizes ranging from 2 to 25.

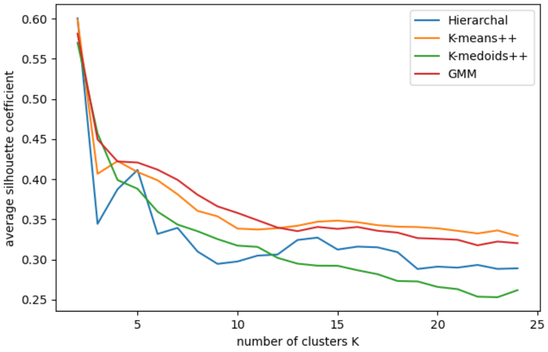

Figure 1: Comparison of AvgS for different number of clusters

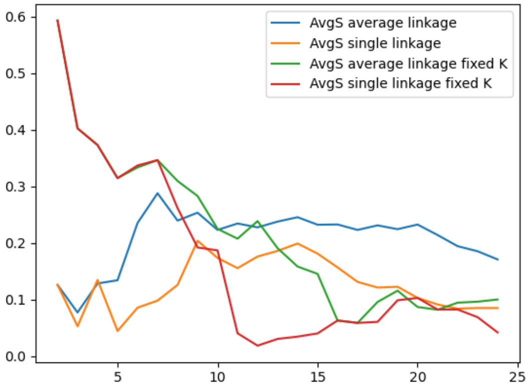

Figure 2: AvgS for the considered cluster ensembles

4 Results

The best results for the four algorithms considered are shown in Figure 1. The best results are found using K-means++ and GMM. From this result the best value for the number of clusters K was chosen. In general it is very hard to find an optimum for K since it is dependent on what metric you deem relevant for your application. The number of clusters was selected to be 5 due to their relatively high average silhouette score. The above average scores for K=2 and 3 were ignored since at such low cluster count the clusters would contain close to no details, thus making it impossible to choose representative feeders from them.[12]

The feeders are plotted in Figure 21 and 22 according to their number of customers and average consumption and colored according to which cluster they were assigned by the algorithm used (GMM and K-means++ respectively). Both algorithms seem to identify a cluster with above average consumption per customer thus indicating high penetration of heavy electric loads such as electric heating or vehicle charging. Another interesting observation is that these feeders all have few customers connected to them. From inspecting Figure 6 can be seen that The K-means++ algorithm divides the remaining feeders then into 4 clusters as follows: High number of customers, medium number of customers, low number of customers with shorter connections and low number of customers with long connections. The Gaussian mixture model shows a solution that is quite similar as can be seen in Figure 5.

It can be concluded that similar to the findings of [12] GMM and K-means++ deliver the best results according to the average silhouette coefficient. It should be mentioned however that GMM is computationally heavier. Also, for $K = 5$ Hierarchical clustering obtained similar results to the former 2 algorithms although only being executed once. The best number of clusters was found to be lower however. This could be due to lower variety in the data set. The test data considered here has low PV penetration for example. The silhouette scores of the data points are visualized in Figure 7 and 8. The average and global silhouette coefficient are also summarized in table 1 where they are compared against the results found by Rigoni et al. for a LV network in North-west England. The results found here are considerable worse than those found by Rigoni et al. indicating that contains less distinct clusters or that additional features are needed to make a better distinction.

	Spain		North-West England	
	K-means++	GMM	K-means++	GMM
Average silhouette	0.35	0.32	0.55	0.56
Global silhouette	0.33	0.32	0.37	0.38

Table 1: Comparison of silhouette coefficients found by Rigoni et al. [12]

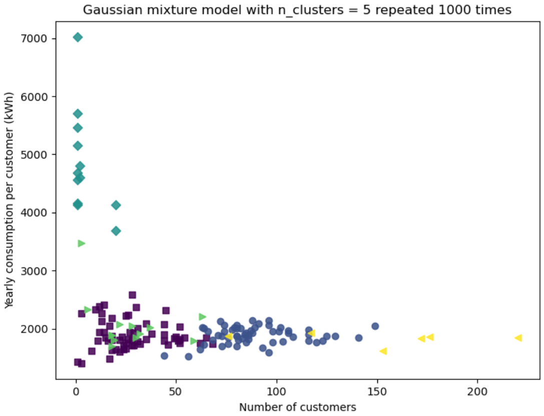

Figure 3: GMM clusters in function of average consumption for K=5

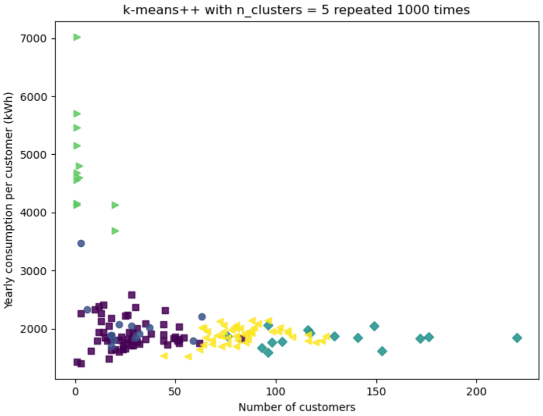

Figure 4: K-means++ clusters in function of average consumption K=5

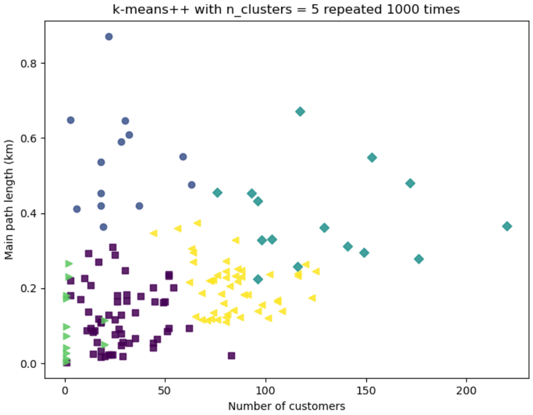

Figure 5: GMM for clusters in function of main path length K=5

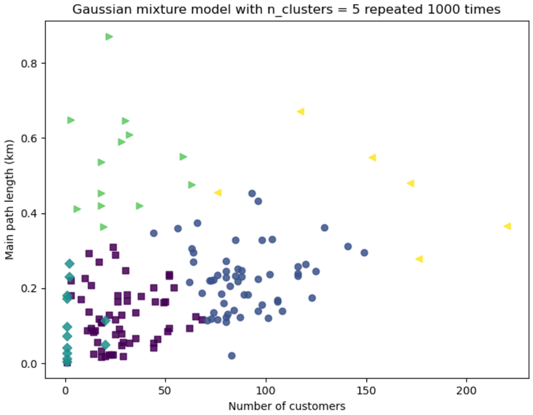

Figure 6: K-means++ clusters in function of main path length for K=5

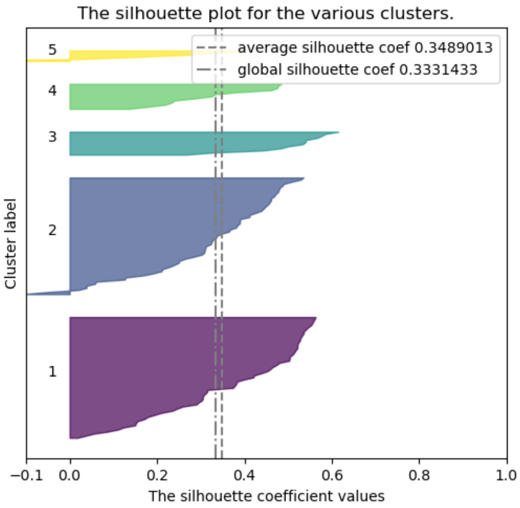

Figure 7: GMM Silhouette for K=5

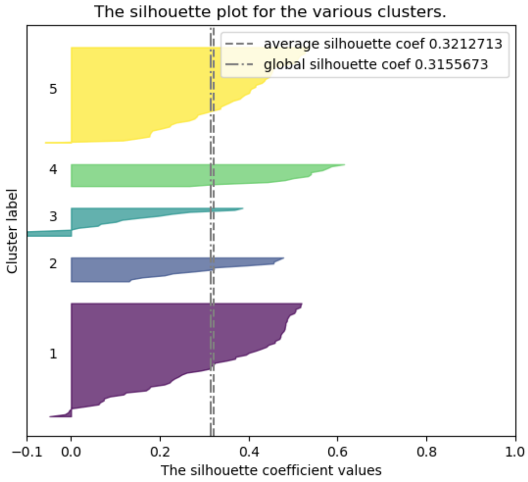

Figure 8: K-means++ Silhouette for K=5

5 Improvements

Next, it is investigated whether using cluster ensemble techniques can improve the results. In this technique different clustering results are combined. Different results can be obtained by using different algorithms, data, initializations or cluster sizes. These results are then combined using a consensus function such as CSPA[5] to form a consensus matrix. Then an algorithm has to be chosen to extract the final clusters from the consensus matrix.

In this work, four variants were considered. Different clusters are obtained by performing K-means++ (i) using 1000 different initializations and keeping K fixed and (ii) using 1000 different initializations while varying K. The final clusters were extracted using (a) average linkage and (b) single linkage as proposed in [6]. average and single linkage are two variants of hierarchical clustering algorithms. Thus four variants were considered (ia, iia, ib, iib). The consensus matrix C is obtained using CSPA. CSPA uses a voting system to determine which samples are more correlated to each other. The entry's of C are determined as in (1) where n_{ij} is the number of times the i'th and j'th element are partitioned in the same cluster. N is the total number of partitions, which is 1000 in this case. Thus essentially, CSPA performs a non-linear transformation on the original features, obtaining an alternative representation of the features (denoted as consensus matrix), which then can be used to extract the final clusters. CSPA is rather simple but has a computational and storage complexity of $O(n^2)$. Since a small data set is considered here (160 feeders), this is no problem but for bigger data sets alternative algorithms such as HGPA[6] and MCLA[7] exist[13].

$$C(i,j) = \frac{n_{ij}}{N} \tag{1}$$

A comparison of the average silhouette coefficient of the four considered variants is shown in Figure 2. For lower cluster numbers a fixed number of cluster works best. But for higher cluster numbers the results get worse. Average linkage outperforms single linkage in almost all cases. The results for both variants for average linkage and K=7 score best according to average silhouette score. Their resulting clusters are shown in Figure 9, 10, 11, 13 and 14.

Conclusion

5 clusters of representative feeders for a spanish urban area were found using GMM and K-means++. Similar to the findings in [12] these algorithms obtained the best results. The best number of clusters was found to be lower in case of the Spanish urban area. This could be due to the low penetration of PV in this area. The experiment with ensemble clustering techniques led to equally good or in some cases better results then for simply repeating K-means++. This quite an interesting result, this means that in some cases combining 1000 results of K-means++ delivers superior results then choosing the best run out of 1000.

[5] Cluster-based similarity partitioning algorithm
[6] Hyper-graph partitioning algorithm
[7] Meta-clustering algorithm

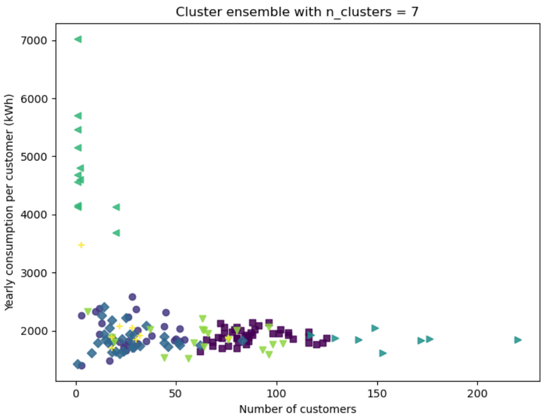

Figure 9: AL with fixed K clusters in function of active energy consumption for K=7

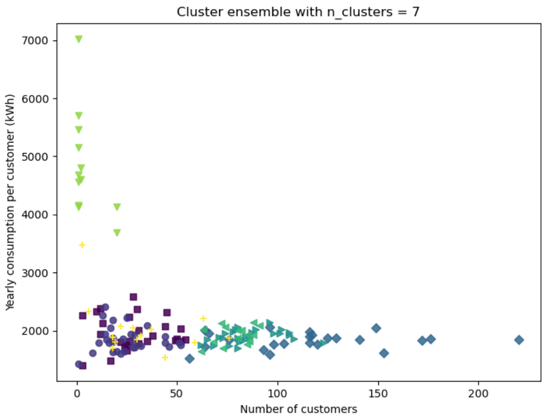

Figure 10: AL with varying K clusters in function of active energy consumption for K=7

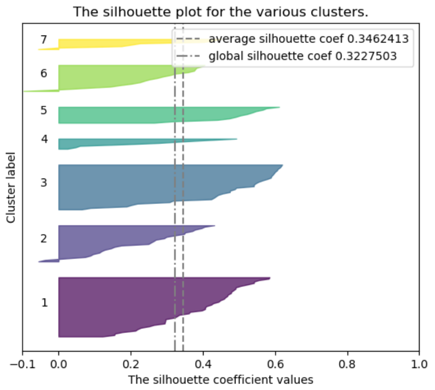

Figure 11: AL with fixed K clusters in function of active energy consumption for K=7

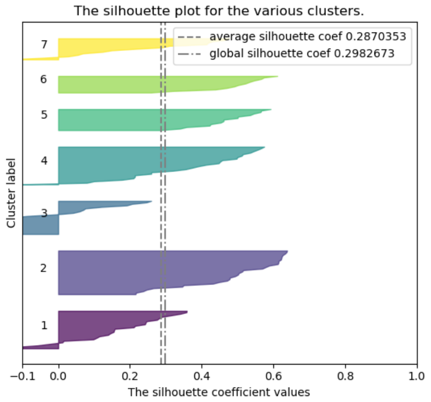

Figure 12: AL with varying K clusters in function of active energy consumption for K=7

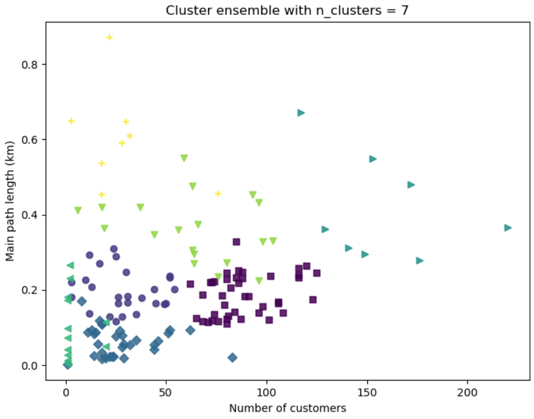

Figure 13: AL with fixed K clusters in function of main path length for K=7

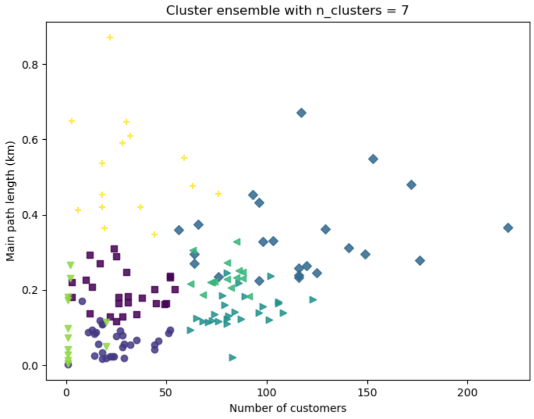

Figure 14: AL with varying K clusters in function of main path length for K=7

Appendix: Additional plots

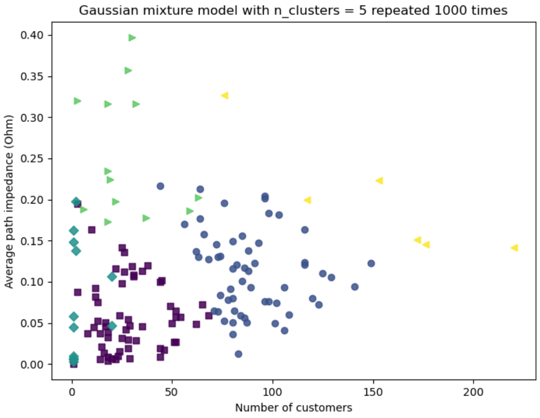

Figure 15: GMM clusters in function of average path impedance for K=5

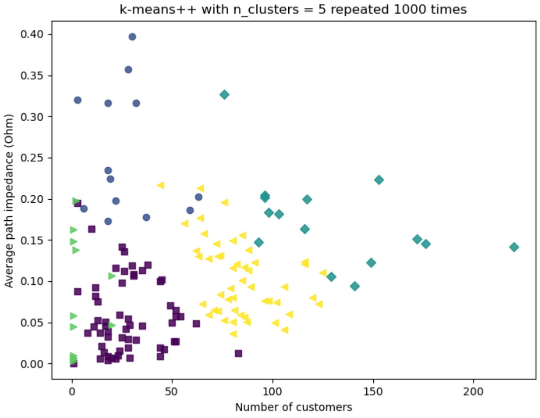

Figure 16: K-means++ clusters in function of average path impedance for K=5

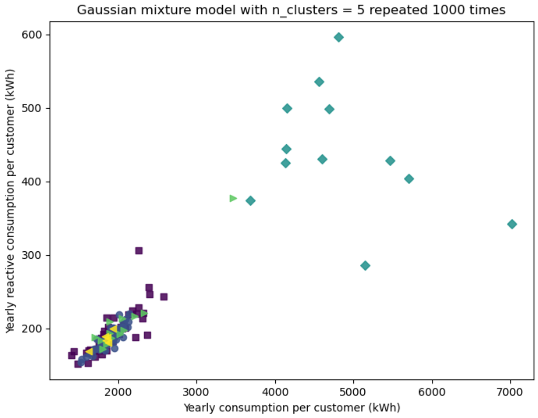

Figure 17: GMM clusters in function of reactive energy consumption for K=5

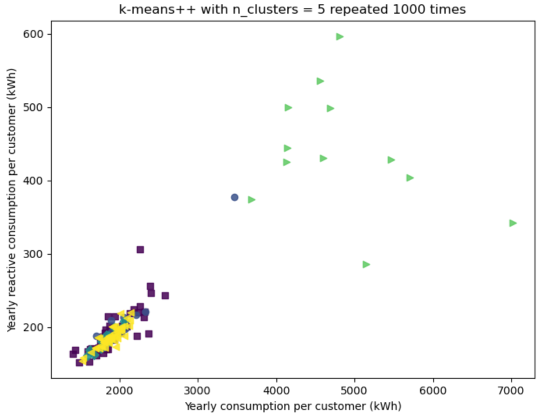

Figure 18: K-means++ clusters in function of reactive energy consumption for K=5

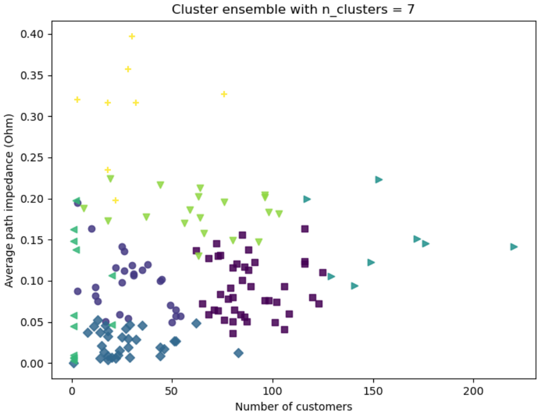

Figure 19: AL with fixed K clusters in function of average path impedance for K=7

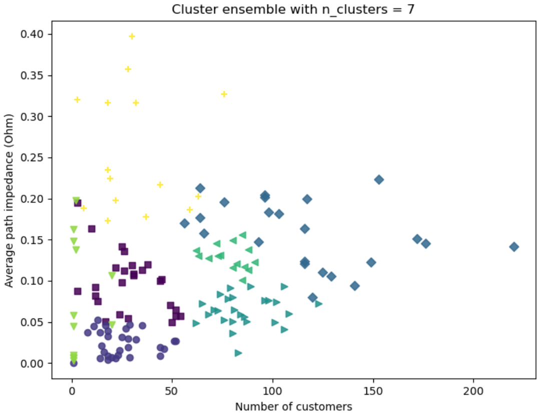

Figure 20: AL with varying K clusters in function of average path impedance for K=7

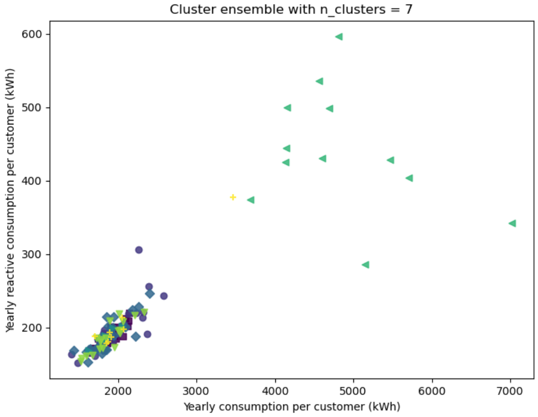

Figure 21: AL with fixed K clusters in function of reactive energy consumption for K=7

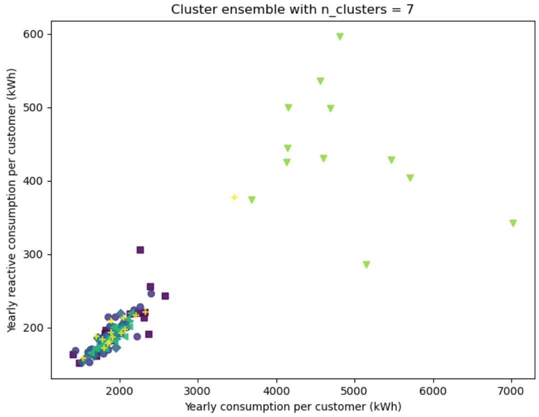

Figure 22: AL with varying K clusters in function of reactice energy consumption for K=7

References

[1] Benoit Bletterie, Serdar Kadam, and Herwig Renner. “On the Classification of Low Voltage Feeders for Network Planning and Hosting Capacity Studies”. In: *Energies* 11 (Mar. 2018), p. 651. DOI: 10.3390/en11030651.

[2] R. J. Broderick and J. R. Williams. “Clustering methodology for classifying distribution feeders”. In: *2013 IEEE 39th Photovoltaic Specialists Conference (PVSC)*. 2013, pp. 1706–1710.

[3] Robert Broderick, Karina Munoz-Ramos, and Matthew Reno. “Accuracy of Clustering as a Method to Group Distribution Feeders by PV Hosting Capacity”. In: May 2016. DOI: 10.1109/TDC.2016.7520083.

[4] J. Cale et al. “Clustering distribution feeders in the Arizona Public Service territory”. In: *2014 IEEE 40th Photovoltaic Specialist Conference (PVSC)*. 2014, pp. 2076–2081.

[5] Jörg Dickert, Max Domagk, and Peter Schegner. "Benchmark Low Voltage Distribution Networks Based on Cluster Analysis of Actual Grid Properties". In: June 2013. DOI: `10.1109/PTC.2013.6652250`.

[6] Ana Fred and Anil Jain. "Combining Multiple Clusterings Using Evidence Accumulation". In: *IEEE transactions on pattern analysis and machine intelligence* 27 (July 2005), pp. 835–50. DOI: `10.1109/TPAMI.2005.113`.

[7] Gunther Gust. "Analyse von Niederspannungsnetzen und Entwicklung von Referenznetzen". German. MA thesis. 2014. 100 pp. DOI: `10.5445/IR/1000045150`.

[8] Khairul Anwar Ibrahim, Mau Au, and Chin Gan. "Generic Characteristic of Medium Voltage Reference Network for the Malaysian Power Distribution System". In: May 2015. DOI: `10.1109/SCORED.2015.7449324`.

[9] Anil K. Jain. "Data clustering: 50 years beyond K-means". In: *Pattern Recognition Letters* 31.8 (2010). Award winning papers from the 19th International Conference on Pattern Recognition (ICPR), pp. 651–666. ISSN: 0167-8655. DOI: `https://doi.org/10.1016/j.patrec.2009.09.011`. URL: `http://www.sciencedirect.com/science/article/pii/S0167865509002323`.

[10] M. A. Maniar and A. R. Abhyankar. "Validity index based improvisation in reproducibility of load profiling outcome". In: *IET Smart Grid* 2.1 (2019), pp. 131–139.

[11] Michiel Nijhuis, Madeleine Gibescu, and Sjef Cobben. "Clustering of low voltage feeders form a network planning perspective". In: *CIRED 2015* (2015).

[12] Valentin Rigoni et al. "Representative Residential LV Feeders: A Case Study for the North West of England". In: ().

[13] Alexander Strehl and Joydeep Ghosh. "Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions". In: *Journal of Machine Learning Research* 3 (Jan. 2002), pp. 583–617. DOI: `10.1162/153244303321897735`.

[14] H. L. Willis, H. N. Tram, and R. W. Powell. "A Computerized, Cluster Based Method of Building Representative Models of Distribution Systems". In: *IEEE Transactions on Power Apparatus and Systems* PAS-104.12 (1985), pp. 3469–3474.