Hierarchical Agglomerative Clustering of Selected Hungarian Medium Voltage Distribution Networks

Attila Sandor Kazsoki[1,2], Balint Hartmann[2]

[1] Department of Electric Power Engineering, Budapest University of Technology and Economics, Egry József u. 18, 1111 Budapest, Hungary

[2] Department of Environmental Physics, Centre for Energy Research, KFKI Campus, Konkoly-Thege Miklós u. 29-33, 1121 Budapest, Hungary;

kazsoki.attila@vet.bme.hu, hartmann.balint@energia.mta.hu

Abstract: Nowadays the increase of photovoltaic penetration and simultaneously, the decentralization of electricity system, poses a number of challenges for distribution system developers and operators. The spread of high output power photovoltaic power plant connections demands the development of a network infrastructure. The analysis of development directions can be done with software simulation, for which network models are needed, which can characterize real networks well. To create such reference networks, knowing existing topologies, hierarchical agglomerative clustering can be a solution. When the parameters of the clusters are specified well, their software implementation can be done. In this study, a possible clustering process of selected Hungarian medium voltage overhead networks (including the determination of the optimal cluster number too), and the formulated network clusters are presented. The clustering of twenty selected 22 kV medium voltage networks was done using hierarchical agglomerative clustering. Then the optimal cluster number was determined. Based on Davis-Bouldin and Silhouette criterions, this cluster number was four. Two of the four generated clusters are single clusters, containing only one feeder. The size and looping of the characterized sample networks are well observable. In this paper a method has been created to generate medium voltage distribution network models, which can be used to simulate the effects of the growing photovoltaic penetration in the Hungarian distribution network.

Keywords: distribution network; network clustering; hierarchical agglomerative clustering

1 Introduction

Nowadays, in both domestic and international energy market trends, photovoltaic penetration quickly increases. Photovoltaic systems, considering their output power, covering the entire power plant range (from household size small to size

small and over that). A large amount of this (approx. 333 MWp, which is 56%) is household size small power plants. Due to a change of the renewable support system at the end of 2016, the number of "applications for licenses for the installation" of photovoltaic power plants with 500 kVA output power increased significantly. Based on the photovoltaic market predictions, in the next decade, the number of small power plants is going to increase, which will increasingly decentralize the structure of the electricity infrastructure (firstly at the distribution voltage level) [1] [2]. In Table 1, the increasing tendency of the built-in photovoltaic capacity is shown.

Table 1
Cumulative photovoltaic capacity development in Hungary [1] [2]

	2015	**2016**	**2017**	**2018**
Built-in household size small power plant capacity [MWp]	129	164	241	324
Built-in small power plant capacity [MWp]	16	27	61	approx. 240
Total built in photovoltaic capacity [MWp]	172	235	344	665

In order for the low, and medium voltage networks to approximate the smart grid structure, electricity infrastructure development is necessary. Modeling of these distribution networks (here on medium voltage level) is essential to determine development directions and to answer emerging questions. For these simulations, the software implementation of medium voltage networks is recommended. Because in the examined area, there is a significant number of varied topology medium voltage networks, their software implementation and running simulations is a powerful time and resource consuming task. It is recommended to formulate reference networks with which the real feeders can well be described. Such reference networks can be created by clustering real networks. These distribution network models can be approximated more precisely than the mathematical models described in the literature [5]. Thus, real decision situations can be handled by the generated reference networks. In this paper a method has been created to generate Hungarian medium voltage distribution network models.

In Section 2 the used hierarchical agglomerative clustering method is presented, with which the numerous medium voltage networks are decreased to a manageable number of clusters. In Section 3.1, the examined distribution feeders are presented. In Section 3.2, the clustering method and principal component analysis are used, in Section 3.3 the determination process of the optimal cluster number are presented. Section 4 d the results of clustering and the generated clusters are presented.

2 Data Analysis Techniques

Data mining techniques can be used to get significant information from the examined database. [2].By reviewing a number of studies (Table 2) in which some kin fog these techniques (typically clustering methods) were used for grouping low or medium voltage electricity network, it can be said that classification, K-means (and K-medoids) clustering and hierarchical clustering are the most frequently used methods. [3]

Table 2
Methods of network analyzing techniques

Name of method	Referenced in literature	Number of groups	Number of analyzed networks
Classification	[3] [4] [5] [6] [7] [8] [9] [10] [11] [30]	Small number (approx. 3–5)	some 100– some 1000
Dimension reduction (SOM)	[3] [4] [12] [13]	Medium number (approx. 8–9)	some 100
Agglomerative clustering	[3] [4] [10] [14] [16] [26]	Large number (approx. 10–25)	some 100–10000+
Partitional clustering	[3] [4] [10] [17] [26]	–	some 100
K-means clustering	[3] [4] [14] [17] [19] [20] [22] [23] [24] [25] [26] [27] [29]	Variable number (approx. 2–12)	some 10–10000+
K-medoids clustering	[3] [4] [17] [23]	Medium number (approx. 8–9)	some 1000

The description of the mentioned methods are not presented in this paper, the clustering processes, the advantages and disadvantages of them can be found in another review paper of the authors [3].

The number of the examined feeders is relatively small (20), hierarchical agglomerative clustering can be used.

2.1 Hierarchical Agglomerative Clustering

“In hierarchical clustering, clusters are determined with the relative distance (Euclidean distance) between the examined data points. The main concept is that a selected item is more tied to a closer data point that to a farther one.” [3] [15]

At the beginning of this process, all the data points (n) are considered as a single cluster. At each step of the algorithm, all data points are moved to a larger cluster. The clustering algorithms stop when all the n points are in the same cluster. As the graphical representation of the clustering, a tree-structure (dendrogram) can be used, which can be cut off at any level. At this level, the leave elements of the tree represent the clusters [2] [3] [4] [14].

The advantage of the algorithm is that "it corrects the distance errors between the local minimum and the center of the clusters" [3] [4]. Besides the positive attributions, there are many negative ones too. The greatest one is the irrevocability of cluster merges. In one step if two clusters are combined, they cannot be divided again later, since the new cluster is used in the future steps of the algorithm. These steps are critical because incomplete mergers give incorrect results (clusters) [3] [4] [14]

3 Clustering Method

3.1 Input Network Data

In this publication, 20 selected Hungarian medium voltage, 22 kV overhead distribution feeders which can be found in the same distribution system operator area, but at four different locations were examined. At the selection of the feeders, the most important criteria were to be able to physically accommodate (approx. 500 kVA) photovoltaic small power plants (output power approx. at least 500 kVA, area is at least 1 ha). Half of the examined networks are located in rural areas and the other half of them are located in suburban settings.

Here, the examined networks are handled as graphs. These graphs can be characterized by specific mathematical variables, such as:

- Total node number
- Average node degree
- Clustering coefficient (CC)
- Characteristic path length (CPL)

The average node degree can be defined with Eq. 1. [4]

$$average\ node\ degree = \frac{2 * E}{N} \tag{1}$$

"where E is the number of edges, N is the number of nodes of the graph". [4]

The clustering coefficient can be defined with Eq. 2. [4]

$$CC = \frac{1}{n} * \sum_{i=1}^{n} \frac{2|\{e_{j,k}: v_j, v_k \in N_i, e_{j,k} \in E\}}{k_i * (k_i - 1)} \tag{2}$$

"where $e_{j,k}$ is edge between vertex v_j with v_k; N_i is the set of immediately connected neighboring vertices for a vertex v_i; k_i is the element number of N_i and n is the size of the graph". [4]

The CPL is interpreted as the impedance values of the lines (feeders). It can be defined with Eq. 3. [4]

$$CPL = \frac{1}{n * (n-1)} * \sum_{i \neq j}^{k} d(v_i, v_j) \tag{3}$$

"where *n* is the size of the graph, and *d* is the distance between any two nodes of the graph". [4]

The values of these parameters can be found in Table 3. To calculation they, the built-in functions of MATLAB R2018b were used.

Table 3
The parameters of the examined feeders

Network identifier	Node number	Average node degree	CC	CPL
N1	350	2.0057	0.7230	27.0648
N2	100	2.0000	0.7273	12.4315
N3	153	1.9869	0.7349	18.0999
N4	193	1.9896	0.7375	25.2386
N5	32	1.9375	0.7646	6.1734
N6	835	2.0216	0.7225	48.4478
N7	82	2.0000	0.7224	11.5414
N8	228	2.0175	0.7294	31.2079
N9	244	2.0164	0.7279	22.4480
N10	243	1.9918	0.7373	22.2443
N11	125	1.9840	0.7347	20.6679
N12	180	1.9889	0.7378	21.1089
N13	59	1.9661	0.7480	10.1473
N14	153	1.9869	0.7383	19.4555
N15	140	1.9857	0.7438	16.7857
N16	491	2.0000	0.7238	27.9195
N17	175	1.9886	0.7290	18.8393
N18	89	1.9775	0.7367	15.4949
N19	166	1.9880	0.7301	21.4499
N20	64	1.9688	0.7370	10.8889

Based on the parameters, it can be said that the feeders have a varied size (node number) and topology.

According to the confidentiality agreement signed between the Centre for Energy Research and the Distribution System Operator (DSO), the authors are not allowed to publish raw data.

3.2 Principal Component Analysis

In this article, hierarchical agglomerative clustering is used.

The most frequently used variables to describe clusters are the size of the network (number of nodes), the degree distribution of feeders (average node degree), the clustering coefficient and the characteristic path length of the feeders. The values of the parameters can be seen in Table 3.

The network analysis is a procedure, in which often more than two variables are taken into account. The handling of a large dataset of multiple variables as a compact unit is a tough assignment. It is recommended to decrease the number of variables, without losing information. A solution can be for this reduction is the principal component analysis (PCA). Using PCA the nature of the array can be written with fewer mathematical parameters (factors) that contain most of the original information. Another task is to describe the nature of correlation between the original variables with the principal components [15] [18].

In this case, the feeders are characterized by four variables. Treating them as a unit is not easy, it is recommended to complete the principal component analysis. To get the values of the principal components, MINITAB 18.0, a statistical software was used. [4] The description of the algorithm used in the MINITAB software can be found at [36].

Based on the scree plot of the main components (seen in Figure 1) and using the "Elbow" criterion, the optimal number of principal components can be determined, which is equal to 2. This means that the examined feeders can be described with the first and second principal components [4].

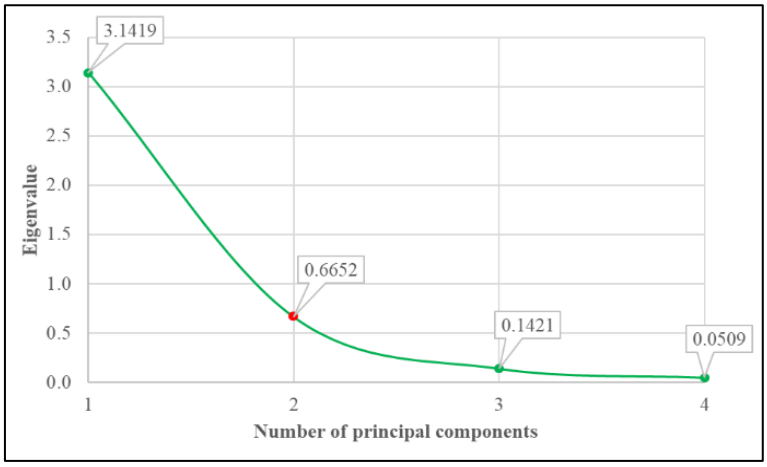

Figure 1
The scree plot for the eigenvalue of principal components for the feeders
(the Elbow point is marked with a red dot)

The numeric values of the principal components for the original parameters and for the examined networks can be seen in Tables 4 and 5, respectively [4]. These values can be used in the clustering process, using a hierarchical agglomerative clustering algorithm.

Table 4
Values of the principal components for the variables

	PCA 1st component	**PCA 2nd component**
Node number	0.49287	-0.53713
Average node degree	0.51853	0.37171
CC	-0.47021	-0.62420
CPL (impedance)	0.51682	-0.42860

Table 5
Values of the principal components for the feeders

Network identifier	**The eigenvalue of the 1st principal component**	**The eigenvalue of the 2nd principal component**	**The value of the 1st principal component**	**The value of the 2nd principal component**
N1	1.711	0.262	187.192	-199.301
N2	-0.136	1.301	56.407	-58.752
N3	-0.379	0.164	85.448	-89.659
N4	-0.079	-0.395	108.853	-114.204
N5	-4.092	-1.722	19.608	-19.591
N6	4.669	-1.817	437.294	-468.968
N7	-0.003	1.703	47.078	-48.699
N8	1.637	0.266	129.207	-135.547
N9	1.230	0.700	132.565	-140.386
N10	0.112	-0.351	131.950	-139.777
N11	-0.381	0.082	72.974	-75.721
N12	-0.221	-0.198	100.311	-105.452
N13	-2.253	-0.408	34.991	-35.776
N14	-0.466	-0.115	86.147	-90.242
N15	-0.938	-0.314	78.357	-82.119
N16	1.949	-0.351	257.125	-275.406
N17	0.041	0.462	96.677	-101.788
N18	-1.037	0.177	52.553	-54.171
N19	0.091	0.283	93.590	-98.074
N20	-1.612	0.277	37.846	-38.772

3.3 Optimal Cluster Number

At the first step of the agglomerative clustering, the cluster number is decided. The determination of the optimal cluster number is based on the simultaneous application of Davies-Bouldin (*DB*) validity criteria and Silhouette (*Si*) validity criteria [4] [21] [28]. For the determination of *DB* and *Si* values the built-in functions of MATLAB R2018b academic version is used.

3.3.1 Determination of the Optimal Cluster Number

Empirically, in the case of a small number of data sets (around some 10 to 100), the optimal cluster number is between 2 and 5. It coincides with what was described in [28]. Therefore, the minimum number of the clusters can be determined with Eq. 4, and the maximum number of clusters can be determined with Eq. 5.

$$M_{min} = 1 + 1 = 2 \tag{4}$$

"where M_{min} is the minimal number of the clusters" [28].

$$M_{max} = \left\lceil \sqrt{N/2} \right\rceil + 1 = \left\lceil \sqrt{20/2} \right\rceil + 1 = 5 \tag{5}$$

"where M_{max} is the maximal number of the clusters, N is the number of examined data points" [28].

$$M_{opt} = [M_{min}; M_{max}] \tag{6}$$

In this paper, the optimal cluster number has been investigated in the range, defined in Eq. 6 (cluster number is 2, 3, 4 or 5), their values are calculated with the simultaneous application of Davies-Bouldin and Silhouette criterions.

3.3.2 Davies-Bouldin Criterion

"The Davies-Bouldin evaluation is an object consisting of sample data, clustering data, and Davies-Bouldin criterion values used to evaluate the optimal number of clusters. This criterion is based on a ratio of within- and between-cluster distances." [4] [31] The Davies-Bouldin index can be defined with Eq. 7 [4] [28] [31] [32].

$$DB = \frac{1}{k} * \sum_{i=1}^{k} max_{j \neq i} \{D_{i,j}\} \tag{7}$$

"where $D_{i,j}$ is the within-to-between cluster distance ratio for the i^{th} and j^{th} clusters" [28]. The mathematical description of this distance can be seen in Eq. 8 [24] [25] [28].

$$D_{i,j} = \frac{(\overline{d}_i + \overline{d}_j)}{d_{i,j}} \tag{8}$$

"where $\overline{d}_i$ is the average distance between each point *i* and the centroid of the i^{th} cluster, $\overline{d}_j$ is the average distance between each point and the centroid of the j^{th} cluster, $d_{i,j}$ is the Euclidean distance between the centroids of the i^{th} and j^{th} clusters" [24] [25] [28].

There is the worst-case for cluster *i* when $D_{i,j}$ has a global maximum at within-to-between cluster ratio. The optimal cluster number can be identified when the Davies-Bouldin index has a global minimum [4] [24] [25] [28].

The objective function of the optimization problem based on Davies-Bouldin validity index is defined with Eq. 9.

$$M_{opt} = \min_{m\in[M_{min};M_{max}]} DB_m \tag{9}$$

"where M_{opt} is the optimal number of the clusters, *m* is the number of clusters" [28].

3.3.3 Silhouette Criterion

"The value of the Silhouette criterion is a metric of how similar is the examined point to the other points in the same cluster, compared to points in other clusters." [4] [33] The Silhouette value (*Si*) for the point *i*, can be defined with Eq. 10 [4] [28] [33].

$$S_i = \frac{(b_i + a_i)}{\max\{a_j, b_i\}} \tag{10}$$

"where a_i is the average distance from point *i* to the other points of the cluster, b_i is the minimum average distance from point *i* to the points in another cluster" [4] [28] [33].

The value of the *Si* can be in the range from *-1* to *+1*. If it is closer to *+1*, point *i* is well-matched to its own, and poorly-matched to the other clusters. The optimal cluster number is then when the Silhouette index has a global maximum [4] [28] [33].

The objective function of the optimization problem based on Davies-Bouldin validity index is defined with Eq. 11.

$$M_{opt} = \max_{i=m\in[M_{min};M_{max}]} S_i \tag{11}$$

"where M_{opt} is the optimal number of the clusters, *m* is the number of clusters"[28].

The results of the two methods described above can be seen in Figure 2. The values of the validity indexes for each cluster number are depicted in Figure 2. Based on these, the optimal cluster number is 4 [4] [28] [33].

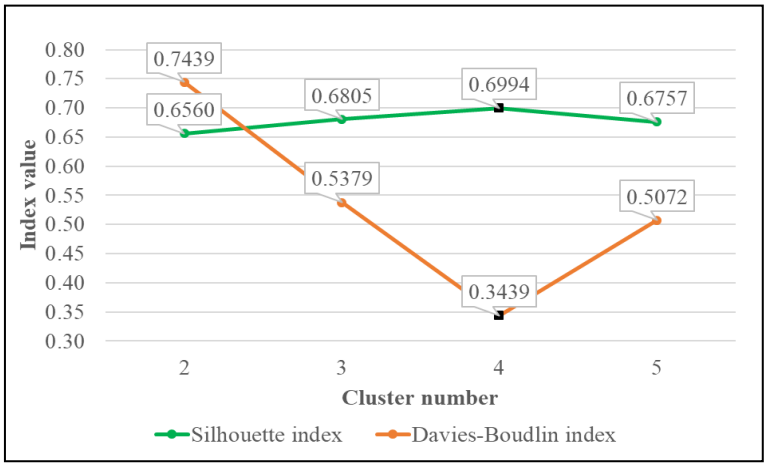

Figure 2
The values of the Davies-Bouldin and Silhouette evaluations in the case of 2, 3, 4, and 5 clusters calculated with MATLAB R2018b built-in functions

The clustering algorithm was run 20 times to avoid local minima. The result of clustering was always the same. The clustering algorithm was convergent.

4 Results

In this publication, from the previously presented 20 medium voltage networks described above, by using principal component analysis and hierarchical agglomerative clustering algorithm, 4 network clusters were created. Clustering was done using the tutorial version of MINITAB 18.0 statistical software.

In MINITAB 18.0, the agglomerative clustering method is based on the complete linkage method (also called furthest neighbor method), in which "the distance between two clusters is the maximum distance between an observation (feeder or data point) in one cluster and an observation (feeder or data point) in the other cluster" [37]. The complete distance is calculated with Eq. 10 [37].

$$d_{m,j} = max\{d_{k,j}; d_{l,j}\} \tag{10}$$

where $d_{m,j}$ is the distance between clusters m and j; m is a merged cluster that consists of clusters k and l, with $m = (k,j)$; $d_{k,j}$ distance between clusters k and j; $d_{l,j}$ distance between clusters l and j [37].

The graphical representation of the clustering (dendrogram), can be seen in Fig. 3. In this figure, the clusters are colored respectively.

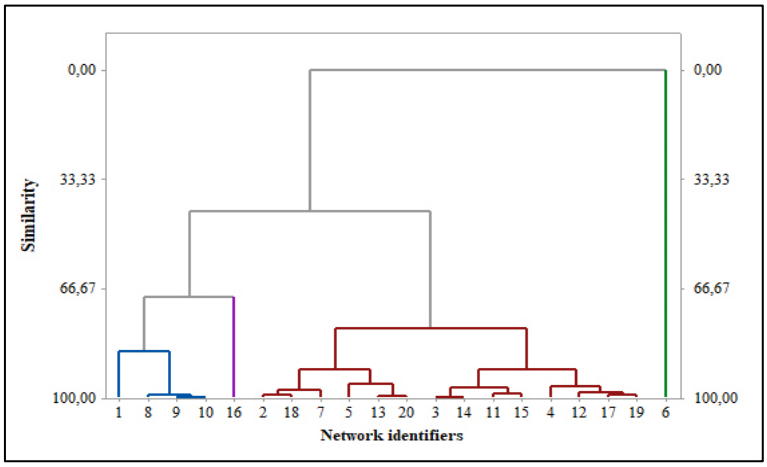

Figure 3
Dendrogram for the clustered feeders

For the graphical representation of the eigenvalues of principal components for the clustered feeders see Figure 4.

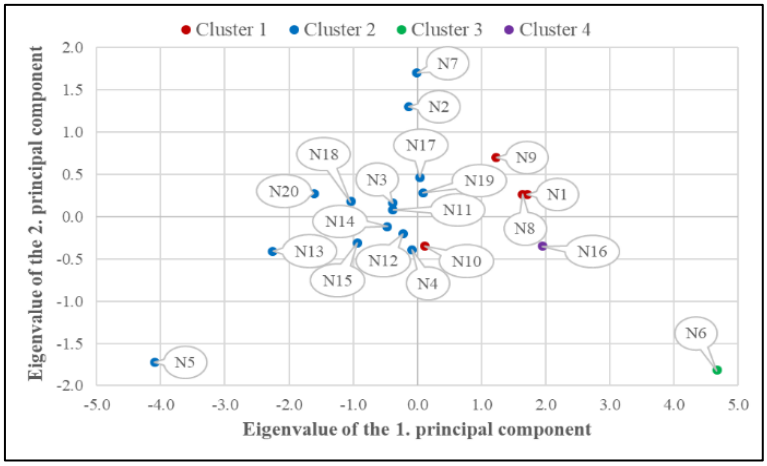

Figure 4
The score plot for the eigenvalues of principal components for the clustered feeders

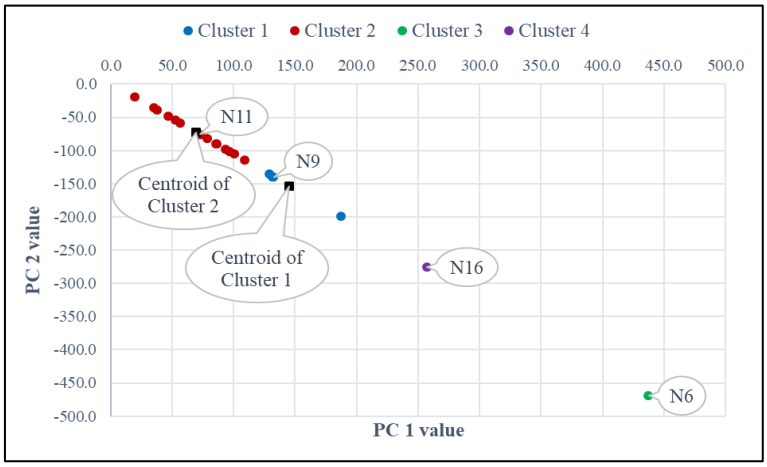

Figure 5
The score plot for the values of principal components for the clustered feeders

In Figure 5, the centroids of the non-single element clusters are marked with black square markers, and the markers of real networks closest to the centroids are labeled too. The values of the centroids (see Table 6), were determined as the average of the data points in a cluster with them. The centroids of the one element clusters are the feeders, included in each cluster.

Table 6
Calculated centroids of clusters

Variable	Cluster1	Cluster2	Cluster3	Cluster4
PCA 1st component	145.229	69.346	437.294	257.125
PCA 2nd component	-153.753	-72.359	-468.968	-275.406

The final partition of clustering and the various distance metrics of the clusters can be seen in Table 7.

Table 7
Final partition of clustering

	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster 1	4	5134.500	30.969	61.933
Cluster 2	14	21507.800	34.968	72.514
Cluster 3	1	0.000	0.000	0.000
Cluster 4	1	0.000	0.000	0.000

For the average numeric values of the variables, see Table 8.

Table 8
The average value of parameters in the four clusters

	Node number	Average node degree	CC	CPL
Cluster 1	266.250	2.008	0.729	25.741
Cluster 2	122.214	1.982	0.737	16.309
Cluster 3	835.000	2.022	0.723	48.448
Cluster 4	491.000	2.000	0.724	27.919

In order to illustrate the characteristics of the typical network topology of clusters, the representation of sample networks, which are the closest to the previously defined centroids, are presented. These networks with their identifier are shown in Figure 5, and their topology can be seen in Figures 6-9.

In Cluster 1, there are 4 weakly looped networks. While the element number of the cluster is not too large (4) and the feeders are fairly similar, the sum of squares of distances within the cluster is approximately the quarter of the same value in Cluster 2. The distances from the centroids are in the same range for Clusters 1 and 2, so it can be said that these clusters are compact. The graphical representation of feeder N9 is shown in Figure 6.

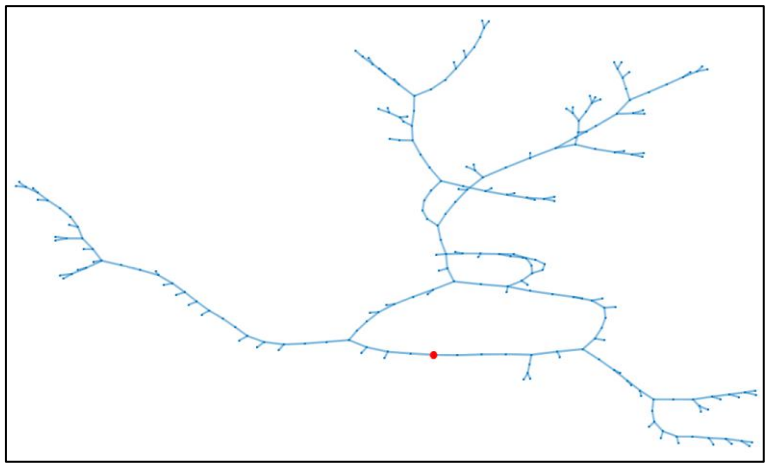

Figure 6
The topological representation of feeder N9 in Cluster 1

Topology N9 is a medium-sized, weakly looped medium voltage (22 kV) overhead network, located in a suburban area. In Figure 6, the HV/MV (132 kV/22 kV) substation is marked with red.

Cluster 2 is the highest element number cluster with 14 feeders. The networks in Cluster 2 are small and medium size and have throughout radial topology, located in a rural area. For the graphical representation of feeder N11, see Figure 7.

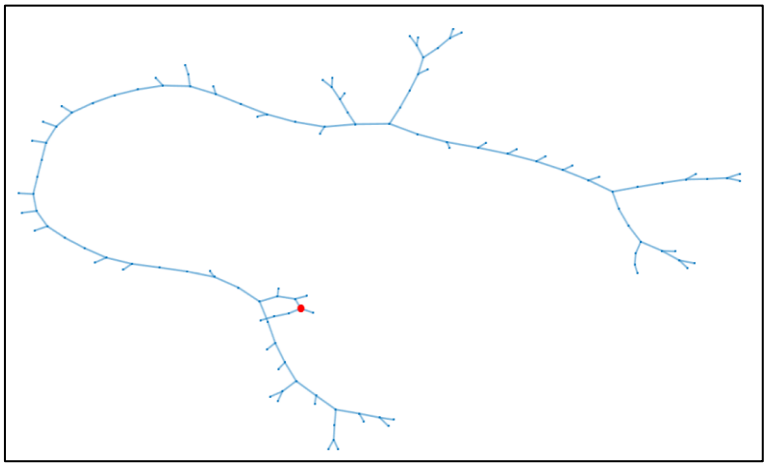

Figure 7
The topological representation of feeder N11 in Cluster 2

Topology N11 is a radial network, placed in the rural area. From the four reference networks, this is the smallest one (smallest node number and CPL value). In Figure 7, the HV/MV (132kV/22kV) substation is marked with cyan.

Clusters 3 and 4 are single clusters. The networks in these clusters are fairly large, and the CCs are the biggest ones too. These networks are located in a suburban area (Cluster 3). These clusters cannot be as relevant as Clusters 1 and 2, because the element number is only 1. The graphical representation of feeder N6 can be seen in Figure 8.

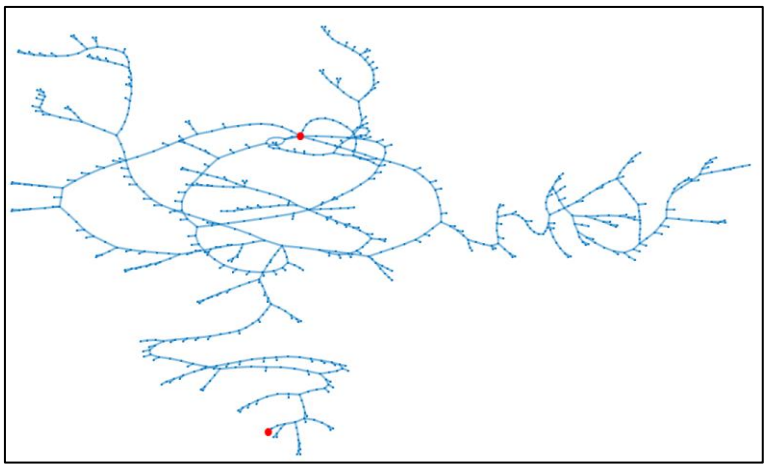

Figure 8
The topological representation of feeder N6 in Cluster 3

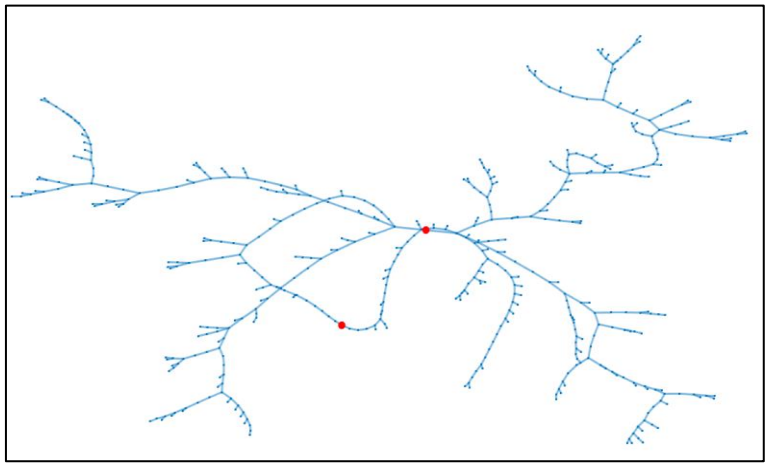

Figure 9
The topological representation of feeder N16 in Cluster 4

Topology N6 is a large, and heavily looped medium voltage network, located near the suburban area of the capital city of the county. Out of the four reference networks, this is the biggest one (highest node number and CPL value). In Fig. 8, the HV/MV (132 kV/22 kV) substations are marked with red (the two substations are different).

Topology N16 is similar to N9, but the node number is much higher. This topology is a large (not as large as N6), weakly looped medium voltage network, in a suburban area of a town. As in the case of N6 (Cluster 3), this topology also contains loops, but less. In Figure 9, the HV/MV (132 kV/22 kV) substations are marked with red (the two substations are different).

Conclusions

In this study, a network clustering method on Hungarian medium voltage distribution feeders has been displayed, which is suitable for the efficient processing of smaller or larger amounts of a data array. Found on the international studies published in the literature, at the formulation of network groups the agglomerative hierarchical clustering and PCA were used. With PCA, the number of the original variables space was reduced from four to two, and this two-dimensional component space was clustered. At the first step, as an input parameter of the clustering algorithm, the optimal number of the clusters was described using the Davies-Bouldin and Silhouette criterions. Both methods led to the same result, the optimal cluster number is 4. The results of the clustering were presented in Section 4. As the topologies of feeders are fairly varied, distinct clusters have been formulated. Two of the clusters (Clusters 3 and 4) are single clusters because in each of these clusters the is only one feeder.

The data processing and clustering method presented in this paper can be well used for clustering networks that cover a physically large area (eg. a country), formulating network topologies specific to the examined area.

Herein, a method has been created to generate Hungarian medium voltage distribution network models, which can be used to simulate the effects of the growing photovoltaic penetration within the Hungarian distribution network. In addition, these results can also help in modeling the voltage and power changing effects on these networks. On the reference networks the effects of growing electrical car numbers, the energy storage penetration and the opportunities for smart grid development can also be simulated.

Acknowledgment

The VEKOP-2.3.2-16-2016-00011 grant is supported by the European Structural and Investment Funds, financed jointly by the European Commission and the Hungarian Government.

References

[1] A. Whiteman, J. Esparrago, S. Rueda, S. Elsayed, I. Arkhipova, Renewable capacity statistics, International Renewable Energy Agency (IRENA) 2019

[2] Data of license exemptioned small power plants and household size small power plants between 2008 and 2017, Hungarian Energy and Public Utility Regulatory Authority, 2018

[3] A. S. Kazsoki, B. Hartmann, Data Analysis and Data Generation Techniques for Comparative Examination of Distribution Network Topologies, in: International Review of Electrical Engineering (IREE) Vol. 14, 2019: pp. 32-42

[4] A. S. Kazsoki, B. Hartmann, Typologization of medium voltage distribution networks using data mining techniques: A case study, 2019 7th International Youth Conference on Energy (IYCE), pp:1-8, 2019

[5] G. A. Pagani, From the Grid to the Smart Grid, Topologically, University of Groningen, 2014

[6] G. A. Pagani, M. Aiello, Towards Decentralization: A Topological Investigation of the Medium and Low Voltage Grids, 2011

[7] V. Lenz, Generation of Realistic Distribution GridTopologies Based on Spatial Load Maps, Swiss Federal Institute of Technology (ETH) Zurich, 2015

[8] P. Hines, S. Blumsack, E. C. Sanchez, C. Barrows, The Topological and Electrical Structure of Power Grids, in: 2010 43rd Hawaii Int. Conf. Syst. Sci., 2010: pp. 1-10

[9] Y. Wang, J. Zhao, F. Zhang, B. Lei, Study on structural vulnerabilities of power grids based on the electrical distance, in: IEEE PES Innov. Smart Grid Technol., 2012: pp. 1-5

[10] K. P. Schneider, YousuChen, D. P. Chassin, R. G. Pratt, D. W. Engel, S. E. Thompson, Modern Grid InitiativeDistribution Taxonomy Final Report, Pacific Northwest National Laboratory, 2008

[11] K. Vill, A. Rosin, Identification of Estonian weak low voltage grid topologies, in: 2017 IEEE Int. Conf. Environ. Electr. Eng. 2017 IEEE Ind. Commer. Power Syst. Eur. (EEEIC / I&CPS Eur., 2017: pp. 1-5

[12] F. Dehghani, H. Nezami, M. Dehghani, M. Saremi, Distribution feeder classification based on self-organized maps (case study: Lorestan province, Iran), in: 2015 20th Conf. Electr. Power Distrib. Networks Conf., 2015: pp. 27-31

[13] Y. Li, P. Wolfs, Preliminary statistical study of low voltage distribution feeders under a representative HV network in Western Australia, in: AUPEC 2011, 2011: pp. 1-6

[14] A. Méffe, C. Oliveira, Classification techniques applied to electrical energy distribution systems, in: CIRED 2005 - 18th Int. Conf. Exhib. Electr. Distrib., 2005: pp. 1-5, Z. Ilonczai, Klaszter-analízis és alkalmazásai, Eötvös Loránd University, Budapest, 2014

[15] Z. Ilonczai, Cluster analysis and their applications (M.Sc. thesis), Eötvös Loránd University, Budapest, 2014

[16] Y. Li, P. Wolfs, Statistical identification of prototypical low voltage distribution feeders in Western Australia, in: 2012 IEEE Power Energy Soc. Gen. Meet., 2012: pp. 1-8

[17] R. J. Broderick, J. R. Williams, Clustering methodology for classifying distribution feeders, in: 2013 IEEE 39th Photovolt. Spec. Conf., 2013: pp. 1706-1710

[18] P. N. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining: Pearson New International Edition, Pearson Education Limited, 2013

[19] R. J. Broderick, K. Munoz-Ramos, M. J. Reno, Accuracy of clustering as a method to group distribution feeders by PV hosting capacity, in: 2016 IEEE/PES Transm. Distrib. Conf. Expo., 2016: pp. 1-5

[20] J. Watson, N. Watson, D. Santos-Martin, S. Lemon, A. Wood, A. Miller, Low Voltage Network Modelling, EEA Conf. Exhib. (2014) 15

[21] C. Gonzalez, J. Geuns, S. Weckx, T. Wijnhoven, P. Vingerhoets, T. De Rybel, J. Driesen, LV distribution network feeders in Belgium and power quality issues due to increasing PV penetration levels, in: 2012 3rd IEEE PES Innov. Smart Grid Technol. Eur. (ISGT Eur., 2012: pp.

[22] J. Dickert, M. Domagk, P. Schegner, Benchmark Low Voltage Distribution Networks Based on Cluster Analysis of Actual Grid Properties, 2013

[23] J. Cale, B. Palmintier, D. Narang, K. Carroll, Clustering distribution feeders in the Arizona Public Service territory, in: 2014 IEEE 40th Photovolt. Spec. Conf., 2014: pp. 2076-2081

[24] I. Borlea, R. Precup, F. Dragan, and A. Borlea, Centroid Update Approach to K-Means Clustering, Adv. Electr. Comput. Eng., Vol. 17, No. 4, pp. 3-10, 2017

[25] S. Chakraborty and S. Das, k − Means clustering with a new divergence-based distance metric: Convergence and performance analysis, Pattern Recognit. Lett., Vol. 100, pp. 67-73, 2017

[26] S. Zahra, M. A. Ghazanfar, A. Khalid, M. A. Azam, and U. Naeem, Novel Centroid Selection Approaches for KMeans-Clustering Based Recommender Systems, 2015

[27] R. Zall, M. R. Kangavari, On the Construction of Multi-Relational Classifier Based on Canonical Correlation Analysis, International Journal of Artificial Intelligence, Vol. 17, No. 2, pp. 23-43, 2019

[28] Q. Zhao, Cluster Validity in Clustering Methods, University of Eastern Finland, 2012

[29] I. Bonet, A. Escobar, A. Mesa-múnera, and F. Alzate, Clustering of Metagenomic Data by Combining Different Distance Functions, Acta Polytech. Hungarica, Vol. 14, No. 3, pp. 223-236, 2017

[30] A. Hamouda, Improvement of the Power Transmission of Distribution Feeders by Fixed Capacitor Banks, Acta Polytech. Hungarica, Vol. 4, No. 2, pp. 47-62, 2007

[31] MATLAB R2018b, Davies-Bouldin evaluation, (n.d.). https://www.mathworks.com/help/stats/clustering.evaluation.daviesbouldinevaluation-class.htm

[32] Davies, D. L., and D. W. Bouldin. A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. PAMI-1, No. 2, 1979, pp. 224-227

[33] MATLAB R2018b, Silhouette evaluation, (n.d.) https://www.mathworks. com/help/stats/clustering.evaluation.silhouetteevaluation-class.html

[34] Kaufman L. and P. J. Rouseeuw. Finding Groups in Data: An Introduction to Cluster Analysis. Hoboken, NJ: John Wiley & Sons, Inc., 1990

[35] Rouseeuw, P. J., Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics. Vol. 20, No. 1, 1987, pp. 53-65

[36] MINITAB 18.0, Principal Component Analyzis, (n.d.). https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/multivariate/how-to/principal-components/methods-and-formulas/methods-and-formulas/

[37] MINITAB 18.0, Linkage clustering methods, (n.d.). https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/multivariate/how-to/cluster-observations/methods-and-formulas/linkage-methods/