18th International Conference on Renewable Energies and Power Quality (ICREPQ'20)
Granada (Spain), 1st to 2nd April 2020
Renewable Energy and Power Quality Journal (RE&PQJ)
ISSN 2172-038 X, Volume No.18, June 2020

Methodology for the formulation of medium voltage representative networks in three DSO areas

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

[1] Department of Electric Power Engineering
Budapest University of Technology and Economics
Egry Jozsef Str. 18. V1 III-IV. Floor, 1111 Budapest (Hungary)

[2] Department of Environmental Physics
Centre for Energy Research
KFKI Campus, Konkoly-Thege Miklos Str. 29–33., 1121 Budapest (Hungary)

Phone number: +36-30/8729134, e-mail: kazsoki.attila@vet.bme.hu
Phone number: +36-20/4825310, e-mail: hartmann.balint@energia.mta.hu

Abstract

In the 2020s, in line with international trends, a rapidly increasing photovoltaic penetration is expected in Hungary. These decentralized electricity sources are typically connected to the low and medium voltage distribution network, affecting their operation. The spread of different output power connection demands the development of the network infrastructure. The development directions and opportunities can be examined with software simulation, so the formulation of reference networks is required. In order to get valid reference networks, clustering of real feeders can be a solution. When the clusters and the values of the variables that describe the clusters are defined, the networks can be implemented in simulation software. In this paper, the clustering process of Hungarian distribution networks, the determination of the optimal cluster number, and the reference networks are all presented, on which the effects of the growing photovoltaic penetration in the Hungarian medium voltage distribution network system can be simulated.

Key words

Distribution network; Network clustering; Hierarchical agglomerative clustering; Determination of the optimal cluster number; Representative networks

1. Introduction

In Hungary, in line with international trends, photovoltaic penetration quickly increases. Photovoltaic systems, considering their output power, belong to the household size small power plant and small power plant range. Due to the change of the renewable support system in 2016, the number of applications for licenses for the installation of small photovoltaic power plants increased significantly. As a result, the regulator in Hungary gave permission to the construction and installation of approximately 800 MWp photovoltaic systems, for the whole sector, in 2019. In the next years, the number of small power plants with 500 kWp output power is expected to increase. As a result of these investments, the hierarchical, centralized structure of the electricity network will be increasingly decentralized, typically at the low and medium voltage level [1].

In order that the distribution networks approximate the smart grid structure, their development is necessary. To determine the directions (and opportunities) of electricity network development and to answer the emerging questions, it is necessary to model these distribution networks. To perform simulation, the software implementation of networks is recommended. [2–4].

Since there is a significant number of various topology medium voltage networks in Hungary, their software implementation and a large range of simulations is a powerful time and resource absorbing exercise. It is recommended to construct reference networks with which the real system can be described well. Such reference networks can be created by the clustering real networks. These distribution network models can be approximated more precisely than in the mathematical models used in the literature. Thus, real decision situations can be handled by the generated reference networks.

2. Data analysis techniques

Data mining techniques use numerous statistics-derived algorithms. Compared to statistics, it is generally presumed that a large set of relevant data is available, but the method of analysis used to get valuable information is unpredictable. One of the biggest questions of data mining is which methods can be used on large data sets and how these tools are used.

After reviewing studies in which electricity network topologies are grouped, it can be stated that for classification, k-means (and k-medoids) clustering and hierarchical clustering are the most frequently used techniques [2–4]. In this study, hierarchical agglomerative clustering is used for the formulation of medium voltage representative networks.

2.1 *Hierarchical agglomerative clustering*

In hierarchical clustering, clusters are determined with the relative Euclidean distance between the examined data points. The main concept is that a selected item is more tied to a closer data point that a further one. The name of the clustering method refers to the process of data processing. Within the group of hierarchical clustering algorithms, depending on the direction of the clustering process, two different algorithms can be identified: (i) agglomerative clustering and (ii) divisive clustering [2–4].
Hierarchical agglomerative clustering is a federation-based algorithm. At the beginning of the process, all the data points (n) are considered as a single cluster. In the next step, each of them is moved to a larger cluster. The shutdown condition of the algorithm is that all the elements are in the same (root) cluster. Agglomerative clustering is looking for new clusters based on the previously generated ones. A dendrogram can be used as the graphical representation of the results. The tree-structured dendrogram can be cut off at any level, allowing each cluster to be examined at the cut-off level. By determining the optimal number of clusters, it is possible to determine the level of the tree where it should be interrupted. As a result of this interruption, the required clusters are already available. The step number of clustering algorithms is relatively high; it is around the size of the input data array to the third power ($\sim n^3$). In the case of large input data sets, this algorithm may be slow [2–4].
The advantage of this algorithm is that it corrects the distance errors between the local minimum and the centre of the clusters. Besides this advantage, there are many disadvantages too. The greatest one is the irrevocability of decisions. If two clusters are merged in one step, they cannot be divided again later, since the new cluster is used in the next step of the algorithm. The merging steps are considered critical, because incomplete mergers result in incorrect clustering [2–4].
In this paper, 1769 selected Hungarian medium voltage distribution feeders are examined and clustered with hierarchical agglomerative clustering (one of the most frequently used). The examined networks can be found in three different distribution system operator areas.

3. Clustering method

In this study, hierarchical agglomerative clustering is used. In MINITAB 18.0 [5], the agglomerative clustering method is based on the complete linkage method (also called the furthest neighbour method). In this method 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 [5].

3.1 *Input network data*

The examined networks are handled as graphs. They are characterized by mathematical and electrical parameters, such as

v1 - number of switches,
v2 - number of transformers,
v3 - average impedance of feeders [Ω],
v4 - total node number,
v5 - average node degree,
v6 - characteristic impedance [Ω],
v7 - (impedance) diameter of the feeders [Ω],
v8 - betweenness centrality [Ω].

Since the range of these parameters is fairly different, their normalization had to be done. On these modified data series, correlation analysis also has been done.

3.2 *Principal component analysis*

The network analysis is a procedure in which more than two variables are taken into account. To examine multiple variables on a large data array and to handle the dataset as a compact unit is complicated. In this case, it is recommended to decrease the number of variables, without losing information [6][7].
A possible way to reduce the number of variables is the principal component analysis (PCA). The task of the PCA is to describe the data array with fewer factors than the number of variables so that the factors contain most of the original information [6][7].
Another target is to describe the nature of the correlation between the numerous variables with the factors. Since in this study the number of variables is 8 and treating them as a unit is a difficult task, it is recommended to carry out. To get the optimal number and the values of the main components, statistical software (MINITAB 18.0) was used [6]. In this software at the first step of the PCA, the optimal number of the main components was determined. The cumulative proportion of the first 3 eigenvalues is 0.95. In this case, the loss of information is approx. 5%. This means that the feeders can be characterized well with the first 3 principal components. The scree-plot of the main components can be seen in Fig. 1, on which the "Elbow point" can also be observed [6][7].

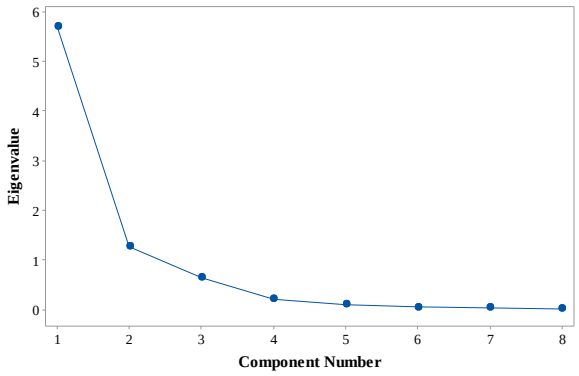

Fig. 1: The scree plot for the eigenvalue of principal components for the feeders

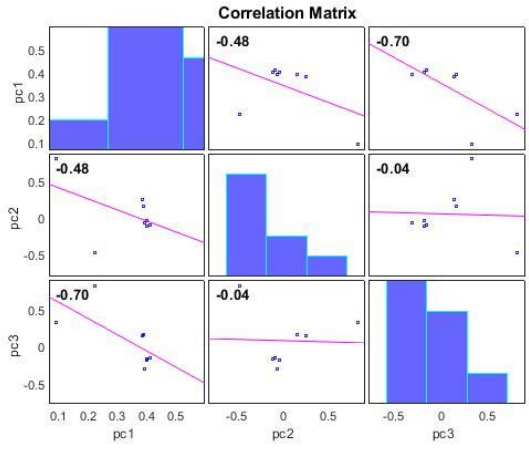

Fig. 2: The graphical representation the correlation analysis

The correlation between the PCA components can be seen in Fig. 2. The R- and P-values of the PCA correlation matrix are given in Table 1.

Table 1: The R- and P-values of the PCA correlation matrix

	PC1	PC2	PC3
R-values of the principal components			
PC1	1.0000	-0.4769	-0.6957
PC2	-	1.0000	-0.0351
PC3	-	-	1.0000
P-values of the principal components			
PC1	1.0000	0.2321	0.0553
PC2	-	1.0000	0.9343
PC3	-	-	1.0000

In this matrix, the P-value is the Pearson correlation coefficient, which is used to examine the strength and direction of the linear relationship between two continuous variables. If the P-value is less than the significance level (0.05), the correlation is significant [6].

3.3 Determination of the optimal cluster number

In the first step of agglomerative clustering, the cluster number is decided. The range in which the optimal cluster number is searched can be calculated with the number of data points. Therefore, the minimum number of clusters can be determined with Eq. 1, and the maximum number of clusters can be determined with Eq. 2.

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

where M_{min} is the minimal number of the clusters.

$$M_{max} = \left\lfloor \sqrt{N/2} \right\rfloor + 1 = \left\lfloor \sqrt{1769/2} \right\rfloor + 1 = 30 \tag{2}$$

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

$$M_{opt} = [M_{min}; M_{max}] \tag{3}$$

In this paper, the optimal cluster number was investigated in the range defined by Eq. 3, their values are calculated with the simultaneous application of the Calinski-Harabasz, the Davies-Bouldin, the Silhouette, and the Gap validity indexes.

3.3.1 Calinski-Harabasz criterion

The Calinski-Harabasz (*CH*) index is defined by Eq. 4.

$$CH = \frac{SS_B}{SS_W} \times \frac{N-k}{k-1} \tag{4}$$

where SS_B is the overall between-cluster variance, SS_W is the overall within-cluster variance, k is the number of clusters, and N is the number of observations.
SS_B is defined by Eq. 5.

$$SS_B = \sum_{i=1}^{k} n_i * \|m_i - m\|^2 \tag{5}$$

where k is the number of clusters, n_i is the number of observations in the i^{th} cluster, m_i is the centroid of the i^{th} cluster, m is the mean of the sample data, and $\|m_i\text{-}m\|$ is the Euclidean distance between the two vectors.

SS_W is defined as Eq. 6.

$$SS_W = \sum_{i=1}^{k} \sum_{x \in c_i} \|x - m_i\|^2 \tag{6}$$

where k is the number of clusters, x is a data point, c_i is the i^{th} cluster, m_i is the centroid of the i^{th} cluster, and $\|x\text{-}m_i\|$ is the Euclidean distance between the two vectors.
The optimal cluster number can be identified when the *CH* index has a global maximum. The objective function of the optimization problem based on *CH* validity index is defined with Eq. 7 [7][8].

$$M_{opt} = \max_{m \in [M_{min}; M_{max}]} CH_m \tag{7}$$

where M_{opt} is the optimal number of clusters, m is the number of clusters.

3.3.2 Davies-Bouldin criterion

The Davies-Bouldin (*DB*) evaluation is an object consisting of sample data, clustering data, and *DB* criterion values used to evaluate the optimal number of clusters. This criterion is based on a ratio of within- and between-cluster distances. The *DB* index can be defined with Eq. 8.

$$DB = \frac{1}{k} * \sum_{i=1}^{k} max_{j \neq i} \{D_{i,j}\} \tag{8}$$

where $D_{i,j}$ is the within-to-between cluster distance ratio for the i^{th} and j^{th} clusters. The mathematical description of this distance can be seen in Eq. 9 [6][8].

$$D_{i,j} = \frac{(\overline{d}_i + \overline{d}_j)}{d_{i,j}} \tag{9}$$

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. The worst-case for cluster i appears when $D_{i,j}$ has a global maximum at within-to-between cluster ratio. The optimal cluster number can be identified when the *DB* index has a global minimum. The objective function of the optimization problem based on DB validity index is defined by Eq. 10 [7][9].

$$M_{opt} = \min_{m \in [M_{min}; M_{max}]} DB_m \tag{10}$$

where M_{opt} is the optimal number of clusters, m is the number of clusters.

3.3.3 Silhouette criterion

The value of the Silhouette criterion is a metric of how similar the examined point to the other points in the same cluster is, compared to points in other clusters. The Silhouette value (S) for the point i, can be defined by Eq. 11.

$$S_i = \frac{(b_i + a_i)}{\max\{a_j, b_i\}} \tag{11}$$

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. The optimal

cluster number is then when S has a global maximum. The objective function of the optimization problem based on the S_i validity index is defined by Eq. 12 [7][10][11].

$$M_{opt} = \max_{i=m\in[M_{min};M_{max}]} S_i \qquad (12)$$

where M_{opt} is the optimal number of clusters, m is the number of clusters.

3.3.4 Gap criterion

Gap criterion is a graphical approach to cluster evaluation that involves plotting an error measurement versus several proposed numbers of clusters, locating the "elbow" which occurs at the highest decrease in error measurement. The gap value (G) is defined by Eq. 13 [7][12].

$$G_n(k) = E_n^*\{\log(W_k)\} - \log(W_k) \qquad (13)$$

where n is the sample size, k is the number of clusters being evaluated, and W_k is the pooled within-cluster dispersion measurement.

$$W_k = \sum_{r=1}^{k} \frac{1}{2 * n_r} * D_r \qquad (14)$$

where n_r is the number of data points in cluster r, and D_r is the sum of the pairwise distances for all points in cluster r, $E_n^*\{log(W_k)\}$ is determined by Monte Carlo sampling from a reference distribution, and $log(W_k)$ is determined from the sample data [12]. The gap value defined for clustering solutions contains one cluster, used with a distance metric. The optimal cluster number arises when the local or global gap value is the largest, within a tolerance range. The objective function of the optimization problem based on G_i validity index is defined by Eq. 15 [7][12].

$$M_{opt} = \max_{m\in[M_{min};M_{max}]} G_m \qquad (15)$$

where M_{opt} is the optimal number of clusters, m is the number of clusters.
The determination of these index values is based on the built-in functions of MATLAB R2019b. The results of the four methods described above can be seen in Table 2.

Table 2: The optimal cluster number determined with different validity indexes

Validity index	K-means	Hierarchical agglomerative
CH	5	7
DB	6*	7*
S	6*	6*
G	5*	8*

**Point with an obvious change*

Based on Eq. 3 the range where the optimal cluster is examined is quiet wide. Based on [7] the optimal cluster number can be found where the value of validity indexes has a global minimum or maximum, or in the case of the validity curves, where a point with an obvious change can be found. The optimal cluster number is determined with the simultaneous application of the most frequently used methods, which are k-means and hierarchical agglomerative clustering. Based on the results of the clustering (see in Table 2), the optimal cluster number is set to 6.

The clustering algorithm was run 25 times to avoid local minima. The result of clustering was always the same. The clustering algorithm was convergent.

4 Results

The data processing method presented above is suitable for clustering networks that cover a larger area (three DSO are in Hungary), developing network topologies specific to the examined area. The dendrogram, as the graphical representation of the clustering, can be seen in Fig. 3, in which the clusters are marked with different colours.

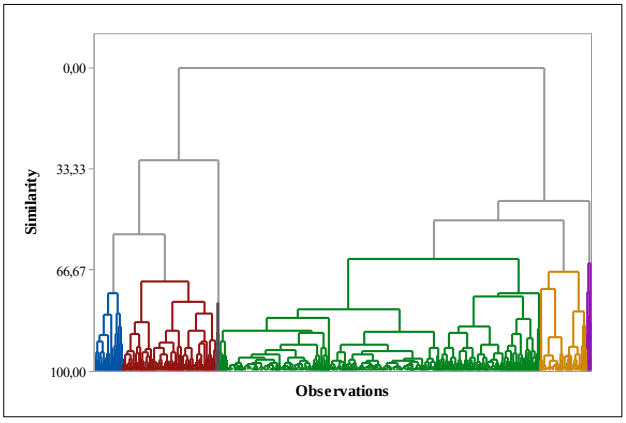

Fig. 3: Dendrogram: the result of the clustering - Complete linkage method with Euclidean distance

Ideally, the clusters are described with the centroid of the clusters, but in most cases this is not a real data point. Then (in this paper too) the clusters can be described with the closest (the smallest distance) feeder to the centroid. The centroids and the nearest networks to the centroids can be seen in Fig. 4–6, marked with red and black cross markers, respectively.

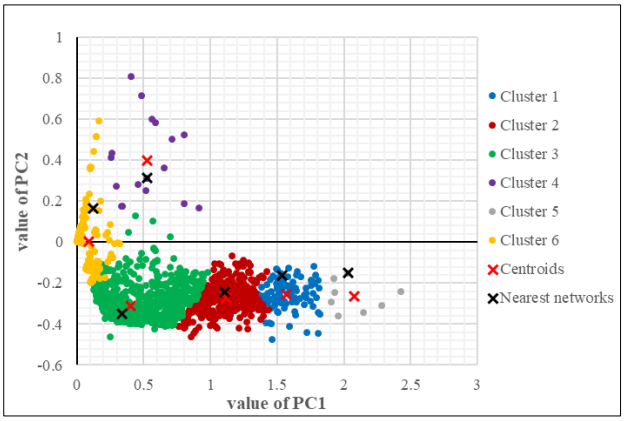

Fig. 4: The result of the clustering - score plot of the 1st and 2nd principal component of the feeders

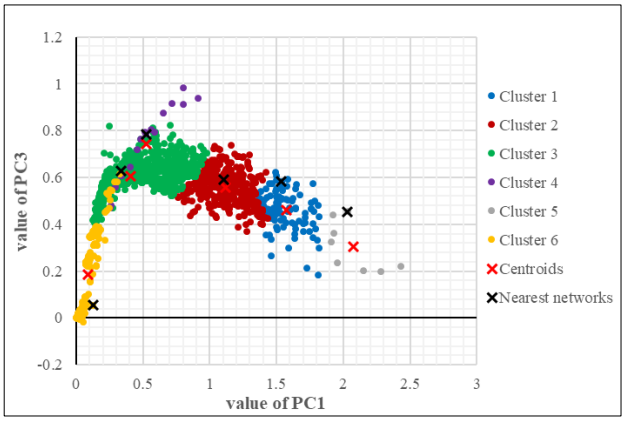

Fig. 5: The result of the clustering - score plot of the 1st and 3rd principal component of the feeders

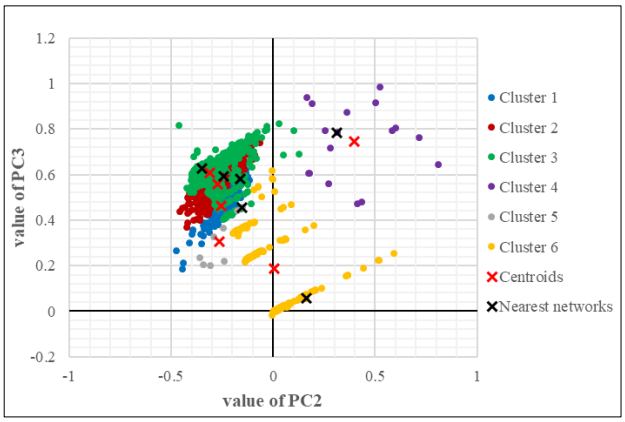

Fig. 6: The result of the clustering - score plot of the 2nd and 3rd principal component of the feeders

The numerical result of the clustering can be seen in Table 3. Based on the parameters, it can be said that the examined networks have a varied size (node number) and topology. The largest cluster is Cluster 3, with 1145 feeders. This cluster covers approx. 65% of the examined area. Cluster 4 and 5 are small clusters with 17 and 8 networks, respectively. The other three clusters (Cluster 1, 2 and 6) are medium size ones.

Table 3: Final partition of clustering

	Number of observations	Within cluster sum of squares	Av. distance from centroid	Max. distance from centroid
Cluster 1	99	2.6775	0.1467	0.4131
Cluster 2	336	11.2049	0.1636	0.3836
Cluster 3	1145	53.9177	0.1866	0.6058
Cluster 4	17	1.6827	0.2947	0.4927
Cluster 5	8	0.3772	0.2070	0.3630
Cluster 6	164	8.5243	0.2110	0.5964

The values of the variables characterizing the reference networks in the clusters can be seen in Table 4.

Table 4: The value of parameters of reference networks

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
v1	67.00	61.00	3.00	0.00	106.00	0.00
v2	60.00	40.00	0.00	13.00	76.00	14.00
v3	0.28	0.28	1.59	0.07	0.32	0.58
v4	150.00	130.00	7.00	26.00	223.00	2.00
v5	1.99	1.98	1.71	1.92	1.99	1.00
v6	8.14	4.99	4.88	0.35	9.78	0.58
v7	39.29k	22.94k	19.5k	6.89k	48.85k	3.42k
v8	1451.77	712.20	5.00	100.00	2535.76	0.00

The graphical representation of the reference networks representing the clusters can be seen in Fig. 7–12.

Network *NC1* is a medium size, 22 kV, mainly overhead (96.6% overhead lines) medium voltage network, placed in a rural area.

Network *NC2* in Cluster 2 is similar to NC1 (99.2% overhead lines), the difference between the networks comes from nature and the size of the supplied are. In NC2 there are more branches covering a larger area, the graph of the network is more centralized.

Network *NC3* is a medium sized, 22 kV, underground (100%) medium voltage network, situated in an urban area. The network can be powered by switching on either of the endpoint interruptions.

Network *NC4* in Cluster 4 is similar to NC3 (in NC4 80% overhead lines), the difference between the two feeders come from the size and the nature of the supplied area. With this kind of network (typically) large consumers are supplied.

Network *NC5* is a large, 22 kV, mainly overhead (94.1% overhead lines) medium voltage network, placed in a suburban area. In the graph, there are two branches where a loop (ring) can be created by switching on an interruption.

Network *NC6* in Cluster 6 is a 3.4 km long 22 kV underground cable with which large consumers are supplied.

5 Conclusion

In this paper, a network clustering method on Hungarian medium voltage feeders has been presented, which is suitable for processing a larger amount of the data array. Based on the international literature on the creation of reference networks, PCA and agglomerative hierarchical clustering were used together. In the first step, the dimension of the examination space was decreased from 8 to 3 (using PCA), and the feeders were clustered in this 3D PCA component space. As the first step of the clustering, the optimal cluster number is described using the Calinski-Harabasz, Davies-Bouldin, Silhouette and Gap criterions. The optimal cluster number was found to be 6. The results of the clustering are presented in Section 4. Because of the variety of topologies, diverse clusters have been created.

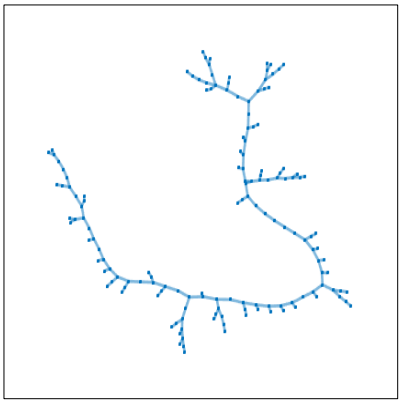

Fig. 7: The topological representation of Cluster 1 (NC1)

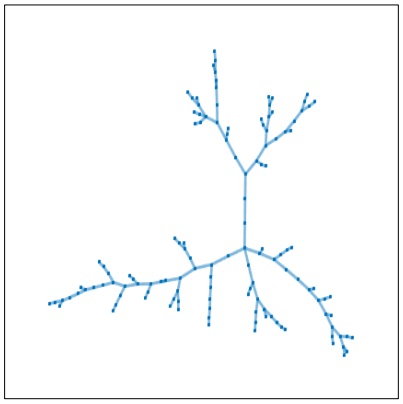

Fig. 8: The topological representation of Cluster 2 (NC2)

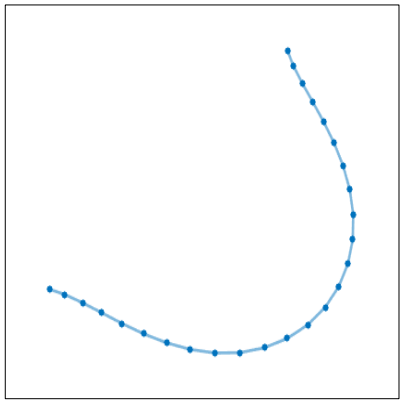

Fig. 9: The topological representation of Cluster 3 (NC3)

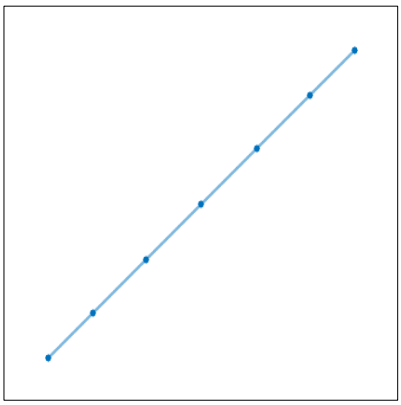

Fig. 10: The topological representation of Cluster 4 (NC4)

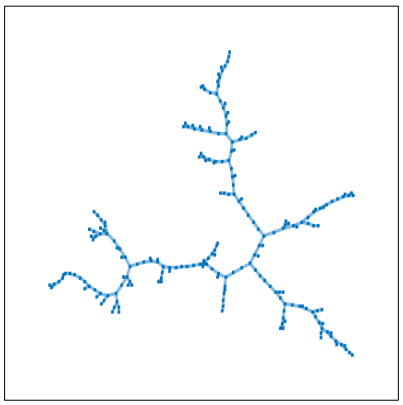

Fig. 11: The topological representation of Cluster 5 (NC5)

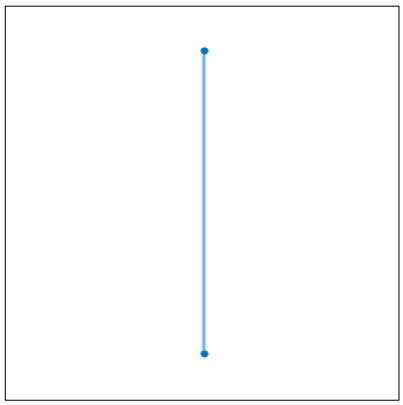

Fig. 12: The topological representation of Cluster 6 (NC6)

The data processing method presented in the present paper is suitable for clustering networks that cover a larger area (country), developing network topologies specific to the examined area.

The method created here to generate medium voltage distribution network models can be used to simulate the effects of the growing photovoltaic penetration in the Hungarian medium voltage distribution network system. The results can also help to model the voltage and power changing effects on these networks. The effects of the growing electrical car and energy storage penetration and the opportunities for smart grid development can also be simulated.

Acknowledgement

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] 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.

[2] 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.

[3] A.S. Kazsoki, B. Hartmann, Typologization of medium voltage distribution networks using data mining techniques, 7th International Youth Conference on Energy, 2019

[4] A.S. Kazsoki, B. Hartmann, Hierarchical Agglomerative Clustering of Selected Hungarian Medium Voltage Distribution Networks, Acta Polytechnica Hungarica, 2020

[5] 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/

[6] 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/

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

[8] Calinski, T., and J. Harabasz. A dendrite method for cluster analysis. Communications in Statistics. Vol. 3, No. 1, 1974, pp. 1–27.

[9] 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

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

[11] 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.

[12] Tibshirani, R., G. Walther, and T. Hastie. “Estimating the number of clusters in a data set via the gap statistic.” Journal of the Royal Statistical Society: Series B. Vol. 63, Part 2, 2001, pp. 411–423.