Jerzy TCHÓRZEWSKI[1]
Tomasz KANIA[2]

1 Siedlce University of Natural Sciences and Humanities
Faculty of Exact and Natural Sciences
Institute of Computer Science
ul. 3 Maja 54, 08-110 Siedlce, Poland

2 Siedlce University of Natural Sciences and Humanities
Faculty of Exact and Natural Sciences
Institute of Computer Science, GENBIT Student Branch
ul. 3 Maja 54, 08-110 Siedlce, Poland

Cluster analysis on the example of work data of the National Power System. Part 1. Comparative study of methods and conditions

DOI: 10.34739/si.2019.23.02

Abstract. The paper presents the results of the research on the comparative study of the methods of cluster analysis and conditions, which was carried out from the point of view of their use, on the example of data concerning the operation of the National Power System. Two algorithms were used for the clustering analysis, i.e. the Ward algorithm and the algorithm of self-organizing two-dimensional maps. Cluster analysis was preceded by a review of hierarchical and non-hierarchical methods of data analysis and a description of the prepared experiment. The obtained results were interpreted. The work consists of two parts published under the same main title with different subtitles. This part 1 presents the results of the conducted review of selected methods of cluster analysis and the research conditions resulting from the adopted data on the operation of the National Power System. Part 2 presents the cluster analysis process and selected research results and their discussion.

Keywords. Cluster analysis, National Electric Power System, MATLAB and Simulink Environment, Ward's algorithm, Self-Organizing Maps.

1. Introduction

At present, the operation of the National Power System (NPS) is associated with the need to conduct a very complex data analysis with the use of large sets of measurement data. The analysis consists in extracting data broken down into groups or classes in such a way that the objects of the same groups or classes are similar to each other, and at the same time clearly differ from elements from other groups or classes [1, 8, 12-14, 15- 18, 27].

The data for the experiments were downloaded from the website of Polskie Sieci Elektroenergetyczne S.A. (PSE) [19, 29], which include, among others: date and time of measurement, national electricity demand [MW], national balance of parallel cross-border electricity exchange (ee), national balance of non-parallel interconnection ee [MWh] and other size.

The values of the downloaded data set is a good basis for cluster analysis using hierarchical and non-hierarchical methods in the MATLAB and Simulink environment [30]. Appropriate cluster analysis was preceded by a data analysis in order to identify and assess various operating states of the tested PEPS system [8, 12-13, 24].

There are many definitions of the concept of Data Mining, one of them is as follows: "Data mining is the process of discovering interesting patterns and knowledge from big data sets" [12]. Data analysis is one of the steps in the process of knowledge discovery and its projection in big databases (Big Data) [24]. Contemporary methods of data analysis are mainly based on the use of computers to search for regularities and interdependencies in large databases, the size of which reaches even several dozen terabytes.

Extracting knowledge from the database is possible thanks to computer analysis combined with interactive cooperation of an expert [24]. Appropriate models and methods are used in the process of acquiring knowledge, and they are used to search for regularities in large data sets. The process includes on searching for patterns, regularities between data, or other relationships and relationships between data stored in large knowledge bases (Knowledge Base), and therefore the first question is the degree of accessibility to large databases [5, 14, 27].

The main purpose of data mining is to search for and describe the existing knowledge contained in data sets as well as to present and interpret the detected regularities. In turn, the process of describing the detected regularities boils down to providing patterns that define the data so that they are readable for the recipient of the information [2, 6, 28], and for obvious reasons the data analysis process requires appropriate preparation, and then cluster analysis . In

this study, two methods were selected, which were described in more detail and then applied in the cluster analysis tasks on the example of data related to the PEPS operation [19, 29].

2. Data analysis methods

2.1. Data mining techniques

In practical applications, depending on the demand, various methods or different techniques using Artificial Neural Networks are used, including neural modeling methods and methods of knowledge exploration [3, 15, 19, 37]. There are appropriate types of knowledge exploration methods from the point of view of, among others:

- classification - decision trees, Bayesian classification, Artificial Neural Networks (ANNs), etc.,
- cluster analysis - graphical cluster analysis, k-means method, Kohonen ANN, mixed models, etc.,
- regression - linear regression methods, non-linear regression methods, etc.,
- association rules - production methods, binary methods, quantitative and qualitative methods, one-dimensional and multi-dimensional methods, one-level and multi-level methods, methods of temporal exploration and methods of frequent data mining, etc. [3, 10, 12-13, 17, 27].

There is also a general division of data mining methods into methods using Artificial Neural Networks, which can be taught under supervision, learned without supervision or learned with a critic (then the last layer plays the role of a teacher) [7, 9, 12-13, 16, 27].

The ANN methods learned with the teacher include: Logistic Regression, Perceptron ANN, Decision Trees, Decision Rules, etc. [8, 9, 12-13].

On the other hand, ANNs learned without a teacher are: Self-Organizing Artificial Neural Networks, Cluster Analysis, Clustering, association rules, etc. [1-2, 7, 21-22, 27]

Among the methods of cluster analysis, there are especially divisible algorithms, agglomerative algorithms, Partitioned Clustering algorithms, data-based analysis algorithms (Incremental Clustering), etc. [5,9-10,18,23,27].

It is worth noting that, if necessary, exploratory techniques can be combined with each other to create specialized hybrid systems. The choice of the cluster analysis method should be

analyzed in terms of advantages and disadvantages, which always translates into the end result in the form of a reliable result of the cluster analysis [27], while the primary goal of the cluster analysis is to divide the set of objects (into not necessarily disjoint) homogeneous groups taking into account the mutual similarity or dissimilarity of the compared objects.

2.2. Discovering associations

One of the most popular data mining techniques to find interesting relationships, patterns, or correlations is the associative network method. The end result of the association discovery process is a set of association rules that represent the found relationships between data. Initially, the association discovery problem was considered in the context of the so-called "Market Basket Analysis". The classic problem of shopping cart analysis is to analyze data from cash registers and describing purchases made by customers in a supermarket.

The purpose of this analysis is to find natural patterns in consumer behavior by analyzing the products that are most often bought together by the supermarket customers (i.e. identifying the groups of products that customers most often place in their shopping carts). The identified patterns of customer behavior can then be used to develop promotional campaigns, new organization of supermarket shelves, to develop a concept of a catalog of offered products, etc. [12-13]. Market Basket Analysis is used wherever customers collect a specific set of things, products or services. It is not necessarily a commercial application, because the basket may be both a set of weather events that occurred in a certain period, as well as a basket of disease symptoms used in medicine, or the behavior of customers on the stock exchange [6].

2.3. Classification

In the clustering method, specific objects are assigned to specific defined classes based on a set of attribute values that characterize the relevant objects. The main purpose of this method is to predict the value of a selected decision attribute by assigning tested objects to similar classes, eg based on descriptors, ie a set of values characterizing these objects [5-6, 14].

Objects are assigned according to the decision attribute key to fixed classes with objects with the same decision attribute value.

The advantages of this approach include: obtaining homogeneous objects of analysis, from which it is much easier to distinguish repeat factors, the possibility of discovering a previously unknown data structure that has been analyzed, and accelerating the time of data mining by reducing a large number of objects to a few basic categories [9, 12-13, 27]. Data classification

is a two-stage process, i.e. the first stage related to building the model to describe a predefined set of classes or concepts, and the second stage in which the model is used to classify the data.

Moreover, the classification is divided into two categories [4, 6, 27]: the model classification, which can be used when the structure of the object category is fully known, and the model-free classification, which consists in recognizing object structures based on data analysis by extracting on the basis of the specified features of similar clusters of these objects.

Due to the existence of many different data sets with separate purposes, classification methods can take various forms that are characterized by one or more characteristics such as: accuracy, errors within the training data set, scalability, interpretability or noise resistance, and many others.

2.4. Grouping

Grouping generally means searching for sets of objects obtained in the course of training ANN without supervision. You can then group records, cases, or cases into classes of similar features, where the group is a set of items that are similar to each other and dissimilar to items in other groups. Grouping differs from classification in that there is no target variable when grouped. The grouping task does not attempt to classify, evaluate, or predict the values of the target variable. Instead, the clustering algorithm tries to divide the entire data set into relatively compatible subgroups or groups, where the similarity of records within groups is maximized and the similarity with records outside groups is minimized [8-9].

Grouping, otherwise known as data clustering or data clustering, is therefore a concept in the field of data mining and machine learning, which derives from a broader concept, which is patternless classification [1, 16, 26]. In this approach, cluster analysis is the so-called Unsupervised Learning, a method of grouping elements into relatively homogeneous classes [27].

The basis of grouping in most algorithms is the similarity between elements - expressed by the similarity function (metric). Grouping can also solve genre problems of discovering structure in data and making generalizations. Grouping consists in distinguishing groups (classes, subsets). The following grouping objectives are distinguished: obtaining homogeneous research subjects, reducing a large number of data to several basic categories, reducing the workload and time of carrying out the analysis, discovering an unknown structure of the analyzed data or comparing multi-feature objects [5, 14, 27].

3. Cluster analysis algorithms

Cluster analysis is a set of multidimensional statistical analysis methods that divide objects into homogeneous subsets in the studied population. Cluster analysis methods are used when the hypotheses are not known in advance.

Finding clusters of objects is based on the elements (variables) characterizing the tested objects, which makes it very important to select the appropriate variables that are used to distinguish compact groups of objects.

Cluster analysis is susceptible to different cases, not matching typical objects, therefore, before starting the analysis, outliers and variables with a slight degree of differentiation in the context of the compared objects should be removed [3, 5-6, 10, 25-27, 30].

Thanks to cluster analysis, you can also treat group separation as an introduction to further multi-dimensional analyzes, or reduce a large data set to average values for individual groups.

The basic role in cluster analysis is played by distance measures, i.e. the dissimilarity functions of a pair of objects that describe the degree of similarity of objects. The smaller the distance between the objects, the more the objects are similar to each other. Elements that are separated by a short distance are combined into clusters. The distance function is calculated on the basis of qualitative, ordinal or quantitative variables that characterize the given objects, the result of the calculation is the distance value expressed as a non-negative number from the set of real numbers.

The most popular distance functions include [5, 13, 14, 27]: Minkowski distance, Chebyshev distance, Euclidean distance, Euclidean distance squared, city distance, Bregman divergence, Mahalanobis distance, cosine distance, power distance, and many other distance measures.

3.1. Hierarchical methods of cluster analysis

Depending on the defined goal, data evaluation can be carried out in various ways. Over the years, many different cluster analysis algorithms have been developed, although due to the grouping approach, two basic methods of data division emerge [5-6, 14, 37]: hierarchical cluster analysis methods and non-hierarchical cluster analysis methods, while the non-hierarchical cluster analysis methods , also known as combinatorial methods, are based on the mapping of objects to created clusters, and hierarchical cluster analysis methods use the possibilities of

creating a classification hierarchy, that is, creating a hierarchy of classes for a specified number of observations.

The methods of hierarchical cluster analysis belong to the traditional methods of cluster analysis and consist in successively combining or dividing the observations [27], which in turn leads to obtaining a tree (dendrogram). Cluster analysis is mainly aimed at obtaining homogeneous clusters of data. As with almost any data mining technique, there are many different choices to classify your data breakdown into clusters.

The most frequently used criteria in hierarchical cluster analysis are the similarity of the objects assigned to a given group or the dissimilarity of elements of one group from the other cluster groups.

With regard to hierarchical methods, there are two approaches to data sharing. The first is the use of the agglomeration technique, while in the second case, the fragmentation technique is used, where the analysis begins with one cluster, which is divided successively according to the increasing degree of intragroup similarity.

Agglomeration techniques are based on the initial indication of the division of objects, in which each of them forms a separate group for classification, where all the given objects belong to one group, taking into account which clusters were combined in subsequent iterations. In the case of fragmentation techniques, during the first step, each element belongs to only one group, after which the group is divided into subgroups until a classification with a certain number of groups is obtained. The maximum of groups in this method is the number of individual elements.

Aglomeration methods of cluster analysis

Among the hierarchical methods, agglomeration techniques are the most frequently used. With these techniques, the objects initially become separate clusters, then objects that are separated by the shortest distance are combined into a new cluster until one cluster is obtained. A very important and, at the same time, difficult issue is the determination of the distance between the resulting clusters, which are formed from connected objects, it is the so-called determination of the binding principle [3, 9, 12-13, 37].

There are many different binding principles. However, they only differ in the method of calculating the distance between groups of objects. Classically, agglomeration techniques are characterized by the fact that after each step of classification, the number of classes is reduced by one, and the reduction in the number of classes is related to the merging of existing classes.

An important feature here is the starting point as n single-element classes and the existence of n-1 classification steps, after which one class consists of all the elements of the set.

The most frequently used agglomeration techniques include [12-13]:

Single Linkage method, also called Nearest Neighbor method, which is based on the minimum distance between elements of two separate groups, i.e. the distance between two clusters is assumed to be defined by the distance closest to located neighbors (being objects of two different collections);

Complete Linkage method, also known as the Furthest Neighbor method, in which the principle of operation is similar to the single linkage method with the difference that the distance between clusters is expressed as the distance between the farthest neighbors, i.e. between the farthest neighbors positioned elements of sets, and this method works well when objects form separate clusters, while in the case of clusters arranged in an elongated form with chain characteristics, this method does not bring favorable results;

Unweighted Pair-Group Method Using Arithmetic Averages, which was developed in order to limit the impact of strongly divergent, i.e. extreme values, on the criterion of group merging. In the mean linkage method, the distance is calculated based on the arithmetic mean of the distance of all elements belonging to different clusters. The resulting cluster groups tend to have very low variability within one group. The method is applicable both when clusters take the form of elongated clusters and in the case of objects clustered in separated groups;

Weighted Pair-Group Method Using Arithmetic Averages, which is similar to the standard method of average connections, which was developed by taking into account the size of the clusters, i.e. the number of objects in them. This method is used when there is a potentially significant discrepancy between the size of the analyzed sets;

Center of gravity method, also known as the Centroid Method or Unweighted Pair-Group of Using the Centroid Average Methods, in which the distance between a pair of clusters is represented as the value of the segment connecting their centers of gravity calculated from the averaged values of the features of the elements within the cluster. The average values are called centroid. This type of method is very similar to the method of determining the center of gravity on a map using geographic coordinates;

The Median Clustering method, also known as the Weighted Group Centroid method, is similar to the method of mean connections with the difference, however, that possible differences between the size of the clusters are taken into account, thanks to which, in the case

of clusters with highly differentiated cardinality, the distance between these clusters will be better reflected;

Ward's Error Sum of Squares method, also known as the Increase in Sum of Squares method, uses analysis of variance to estimate the distance between clusters. This method aims at minimizing the sum of squares of deviations in clusters and uses the Error Sum of Squares (ESS) as a measure of diversity with the search for the minimum diversity of attribute values, which constitute the segmentation criteria for the subsequent steps of the algorithm for creating consecutive clusters. This method works very well in practice, which means that the clusters created are very homogeneous on the one hand, and it tends to create clusters of similar and small sizes on the other hand.

3.2. Non-hierarchical methods of cluster analysis

Algorithms within the methods of non-hierarchical cluster analysis are reduced to the search for the best division of the analyzed data set by stepwise optimization, i.e. improving the quality of clustering, obtained in the subsequent stages of data processing of the input set. The starting division is usually done randomly. Non-hierarchical methods rely on assigning n objects to a predetermined number of clusters k. Regardless of the value of the parameter k, the algorithm does not rely on previously determined clusters separately for each value.

The number k is a key input parameter that is passed as a set point or random value. It determines the number of clusters obtained as a result of the algorithm's operation, therefore it is very important to choose an adequate criterion function by which the quality of the grouping will be assessed. The disadvantage of the non-hierarchical cluster analysis method is its greed, so only the local optimum is obtained as a result. Moreover, there is no guarantee that the global optimum will be achieved at all, but the great advantage is the ease of implementation of this method and the low computational complexity [4, 25, 37].

K-Means Method

The most popular, non-hierarchical combinatorial cluster analysis algorithm is the k-means algorithm, which is based on the minimized variability inside the resulting clusters, therefore the variability between clusters is automatically maximized. In other words, elements are moved from cluster to cluster until the within-group and between cluster variations are optimized.

The most important aspect is the variability optimization, where the objective function is the minimization of the trace of a certain intragroup covariance matrix, which is one of the

two decomposition matrices of the dispersion matrix (the second matrix is the intergroup covariance matrix) [12-13, 31, 37] this method creates k clusters, as diverse as possible.

As reported by the authors of the work [37], despite the over 50-year history of the k-means algorithm, there are relatively few works analyzing its properties, or works formulating its convergences. It is therefore a very simple and effective algorithm for finding groups in the database, and the most common measure of distance is the Euclidean distance, although other measures of object similarity are also used.

Grouping method around centroids

The method of grouping around medoids (called Partitioning Around Medoids) is based on a similar principle as the k-means method, where the centers of the clusters are the observations from the data set (called centroids or cluster centers - centers of gravity) [5, 12-13, 37] .

In this method, the set of potential clustering centers is therefore much smaller than in the k-means method, and often the results of operation are more stable.

The algorithm ends when the centroids no longer change. This method is computationally complex. In the case of a huge data set, its execution may be impossible, therefore the CLARA (Clustering Large Applications) method was created, which allows the use of the method of grouping around centroids on a large data set to divide the set of objects into segments with the merging of the obtained results into one division into clusters.

Kohonen Artificial Neural Networks

Kohonen neural networks belong to the Self Organizing Artificial Neural Networks (SOM), the aim of which is to transform multidimensional input signals into projections into one-, two- or more dimensions, which is associated with dimensional reduction and knowledge projection [7, 13, 17, 20-23].

Kohonen's neural networks belong to the group of Cell Artificial Neural Networks, and the most commonly used learning rules are the Winer Takes All (WTA) and Winer Takes Most (WTM) rules. The neighborhood radius plays an important role in teaching these Self-Organizing Artificial Neural Networks (SOMs), and the main goal is to classify the multidimensional inputs described by a large number of parameters so that the result is represented in a smaller number of dimensions [7, 20-23].

In practice, the analysis is usually limited to two dimensions, with the largest possible representation of the input vector structure. Thanks to this, these networks are used in the visualization of complex structures, and the data processed by them can constitute the

foundation for diagrams displayed on the screen. Another example of application are cases where it is important to reduce the size of the input data, which becomes possible due to the compression properties of the Kohonen's Self-Organizing Neural Network [7, 20-23].

Self-organizing Artificial Neural Networks are neural networks that learn without supervision (of the teacher), which means that only input signals are involved in teaching them, on the basis of which discoveries are made. The process of teaching Kohonen's Artificial Neural Networks is therefore independent and is based on the observations of data sent to the network, the internal structure of which and the logic hidden in the structure determine the final results of the classification, and thus the data clusters [7, 20-23, 37].

The SOM Artificial Neural Network consists of two layers [7, 17, 20-23, 37]: the input layer (input vectors) and the output layer (topological map). The SOM neural network is usually unidirectional, and each neuron is connected to all components of the N-dimensional input vector U. The weights of the connections of the neurons form the vector W^k. What is important in the structure of the SOM Artificial Neural Network is that each neuron of the input layer communicates with all neurons of the topological layer. However, neurons in the same layers cannot communicate with each other.

Maps can take many shapes and dimensions and are selected according to the purpose of the SOM network. For example, a topological map can be a one-dimensional, two-dimensional plane or a torus, as well as a three-dimensional plane, or any geometric figure, including spatial, etc.

The most commonly used topological map is a two-dimensional map. The analyzed data is placed on a user-defined map [12-13, 20-23, 37]. When learning a topological map, the notion of the neighborhood of neurons with the neighborhood radius already mentioned in this paper plays a very clear role.

The learning of the SOM network consists in the fact that during the administration of each training pattern at a given stage of SOM ANN learning, the winning neuron is selected, which becomes the neuron whose weight vector will be closest to the current input vector. The effect of teaching the winner will be his more and more binding to the next input signals that made him the winner. Such a neuron, thanks to training, will recognize a certain class of signals and thus becomes a specific detector of all signals similar to those that contributed to its winning position.

The network is trained by repeatedly introducing input vectors and changing the weights of output neurons. Depending on the learning rule (WTA or WTM), only one or a group of neurons can be modified.

In practice, typical ANN SOM learning takes place in two phases, in the first phase large learning coefficients and a long neighborhood radius are adopted, the WTM learning rule is then applied, and then the neighborhood radius is usually reduced and the WTA learning rule is applied [12- 13, 20-23, 31, 37].

4. Summary of method comparison results

Each of the methods has features that allow them to be used for specific types of data. The main difference between non-hierarchical and hierarchical algorithms is the need to specify the number of clusters in advance. However, differences may also be related to, among others: the size of the data set, the intuitiveness of the selection of features, computational complexity, the influence of data noise, initial parameters, distance between objects, etc.

Table 1. Summary of cluster analysis methods, non-hierarchical and hierarchical.
Source: Own elaboration based on works [5-6, 12-14, 20-23, 37]

	Non-Hierarchical		Hierarchical	
The name of method / *The distinctions*	**K-Means Method**	**Groupings around centroids**	**Method of mean connections**	**Ward's method**
Sensitivity to data noise and distant objects	centroids can become distorted by distant objects	the problem of distant objects does not disturb the shape of the centroids	the problem of distant objects has little effect on the shape of centroids	centroids can become distorted by distant objects
The influence of the initial parameters on the result	very clearly influenced the results	does not affect the results	no impact	no impact
Computational complexity	relatively small	moderate, in the case of large sets the algorithm cost is high	short	short
Efficiency	high	short	short	high only for small harvests
Big data support	relatively good	average	to a small extent	to a small extent

Methods of assessing the correctness of the analysis carried out. There are different ways to validate the data analysis methods developed. They can be divided into quantitative and qualitative methods. In the case of quantitative methods, the correctness is assessed on the basis of comparing the results generated by the method to the actual results (empirical verification or validation), and in the case of qualitative methods, substantive knowledge of the system analyzing the data is required.

This approach makes it possible to evaluate the results taking into account the type of phenomenon under study, hence it is called logical verification or verification. It is worth mentioning the following important data treatments [5-6, 12-13, 20-23, 37]:

Validation, including Simply Validation - checking the correctness of the model's operation or methods based on comparing the results from the test set to the actual results, while simple validation works by dividing the data into the training set and the training set test usually in the proportion of 2/3: 1/3 of the data sample, after which the results from the test set are assessed in comparison to the actual results,

Cross-Validation - it is used to verify the correctness of the method when the set consists of a small number of elements, the operation is based on multiple checks on different data sets, dividing the statistical set into smaller parts, and the next step is performing all possible analyzes on the selected population of the training set,

Verification (also known as logical verification) consists in checking the correctness of the operation of a method or model based on a substantive assessment, i.e. the logical correctness of intermediate and final results, as well as the implementation of the assumptions of the theory that was the foundation during model construction, etc.

Description of the conditions of the tested object and data source

Characteristics of the National Power System

The object under analysis is the National Power System (NPS), which was described taking into account the basic parameters related to its operation. The NPS covers the generation, transmission, distribution and collection of electricity by recipients who may also be prosumers in recent years. One of the subsystems of the Polish Electric Power System (PEPS) is the subsystem of the Highest Voltage Transmission Grid (polish: PSE) managed by Polskie Sieci Elektroenergetyczne S.A. (PSE S.A., PSE), which ensure the transmission of electricity from power plants to distribution networks.

The transmission of ee takes place with the support of power stations via power networks with a voltage of 220 kV, 400 kV and 750 kV. Electricity networks also use 110kV power grids [11, 19], which are high voltage power system networks used for transmission of ee at distances of no more than several dozen kilometers, and power grids with a voltage of 10-30 kV, i.e. medium-sized power grids voltages that are used in local distribution networks. Raising and lowering the voltage to 220/230V or 380/400V in order to enable the use of electricity for everyday devices requires the use of high voltage and high voltage system power stations and transformer stations.

It is also worth mentioning the interconnections between the neighboring countries, thanks to which the export and import of ee are carried out depending on the current needs of individual countries. One of the results of the PEPS work is the forecast of the demand for electricity and

the value of imported or exported electricity, while the PEPS working conditions are subject to changes regulated by the energy law. Depending on the work schedule of electricity recipients, in particular large industrial recipients, the demand fluctuates throughout the day.

The value of demand for electricity, as well as power, is also influenced by weather conditions and other significant increased demands of industrial recipients.

Therefore, the NPS plays a very important role in the functioning of the state and large companies operating on its territory, therefore it must meet a number of economic and technical requirements in order to maintain the security, reliability and continuity of electricity supplies from electricity producers to recipients, both industrial and individual customers. [11, 19, 29].

Data set for cluster analysis

An important element of the cluster analysis is the appropriate preliminary analysis of the input data. Data on the operation of the National Power System were downloaded from the website of Polskie Sieci Elektroenergetyczne S.A., and, inter alia, [29]: measurement time, domestic power demand [MW], total CDGU generation [MWh], PI generation [MWh], IRZ generation [MWh], total nCDGU generation [MWh], national balance of parallel cross-border exchange [MWh], national balance of non-parallel cross-border exchange [MWh].

The collected data covers the period from 01/01/2014 to 31/12/2018, in an interval of one hour. In this way, a very large set of measurement data was obtained, but not all the columns of the table were used to construct the data sets that were intended for cluster analysis, among others, generations of PI and IRZ, as it turned out that these data do not exist in the entire scope of the period under examination. The structure of data collected from daily reports on the PEPS operation consists of the basic values described in Table 2.

Table 2. Basic quantities of the work of the National Polish Electric Power System. Source: [6, 19, 29]

Date	Houer	National power demand	Total CDGU generation	Total nCDGU generation	National balance of parallel interconnection	National balance of non-parallel interconnection
		[MW]	[MWh]	[MWh]	[MWh]	[MWh]
2014-08-01	1	15 811,738	12 265,100	3 405,600	49,013	108,463
2014-08-01	2	15 104,213	11 734,050	3 376,025	-51,775	62,088
…	…	…	…	…	…	…
2018-12-31	24	15 469,150	7 176,513	8 390,100	-164,675	90,113

5. Final remarks

The article compares hierarchical and non-hierarchical methods of cluster analysis from the point of view of using them to analyze the operation of the National Power System on the basis of selected numerical data.

It has been shown that two algorithms are particularly beneficial for the PEPS operation analysis, i.e. the Ward algorithm and the algorithm of self-organizing two-dimensional maps.

An initial analysis of the PEPS numerical data was performed. The work is continued in the article of the same main title in part 2 entitled: Analysis and research results.

References

[1] Cichosz P.: Systemy uczące się, WNT, Warszawa 2000.

[2] Reyes A. J. O., Garcia A. O., Mue Y. L.: System for Processing and Analysis of Information Using Clustering Technique, IEEE Latin America Transactions, IEEE Digital Library, Vol. 12, Issue 2/2014, pp. 364-371.

[3] Długosz M.: Materiały dydaktyczne do przedmiotu „Analiza danych pomiarowych. część IX - Analiza skupień", AGH, Kraków 2015, str. 2-6.

[4] Duraj A., Krawczyk A.: Dobór miar odległości w hierarchicznych aglomeracyjnych metodach wykrywania wyjątków, Przegląd Elektrotechniczny, R. 87 NR 12b, 2011.

[5] Jasiński M.: Zastosowanie analizy skupień oraz globalnego wskaźnika jakości energii do identyfikacji i oceny różnych stanów pracy elektroenergetycznych sieci górniczych w aspekcie jakości energii elektrycznej. Rozprawa doktorska pod kierunkiem prof. dr hab. inż. Tomasza Sikorskiego, Wydział Elektryczny PWr., Wrocław 2019.

[6] Kania T.: Analiza danych z wykorzystaniem analizy skupień na przykładzie Krajowego Systemu Elektroenergetycznego. Praca magisterska pod kierunkiem dr hab. inż. Jerzego Tchórzewskiego, porf. UPH w Siedlcach, Siedlce 2019.

[7] Kohonen T.: Self-Organization of Very Large Document Collections: State of the Art, Helsinki University of Technology, Finland 2013.

[8] Koronacki J.: Statystyczne systemy uczące się, wydanie 2, EXIT, Warszawa 2008.

[9] Larose D.: Odkrywanie wiedzy z danych. Wprowadzenie do eksploracji danych, WN PWN, Warszawa 2006.

[10] Migdał-Najman K., Najman K.: Analiza porównawcza wybranych metod analizy skupień w grupowaniu jednostek o złożonej strukturze grupowej, CEON, Warszawa 2013.

[11] Mielczarski W.: Hannbook: Energy Systems&Markets. Part. 1 Structure and operation. Part 2. Technical aspects. Association of Polish Electrical Engineers, Division Łódź. Edition I., Łódź 2018.

[12] Morzy T.: Eksploracja danych, PWN, Warszawa 2007.

[13] Osowski S.: Metody i narzędzia eksploracji danych. Wyd. BTC. Legionowo 2017.

[14] Płoński P.: Zastosowanie wybranych metod przekształcenia i selekcji danych oraz konstrukcji cech w zadaniach klasyfikacji i klasteryzacji. Rozprawa doktorska pod kierunkiem prof. dr hab. inż. Krzysztofa Zaremby, Wydz. Elektroniki i Technik Informacyjnych PW, Warszawa 2016.

[15] Ruciński D.: The neural modelling in chosen task of Electric Power Stock Market, Studia Informatica. Systems and Information Technology. Vol. 1 No. 21/2017.

[16] Skorzybut M., Krzyśko M., Górecki T., Wołyński W.: Systemy uczące się. Rozpoznawanie wzorców, analiza skupień i redukcja wymiarowości. WNT. Warszawa 2009.

[17] Szeliga M.: Praktyczne uczenie maszynowe. PWN, Warszawa 2019.

[18] Tchórzewski J., Jezierski J.: Cluster Analysis as a Preliminary Problem Neural Modelling of the Polish Power Exchange, Information Systems in Management, Vol. 8/2019 (1), pp. 69-81.

[19] Tchórzewski J.: Rozwój system elektroenergetycznego w ujęciu teorii sterowania i systemów. OW PWr., Wrocław 2013.

[20] Tchórzewski J., Buziak R., Suszczyński P.: Model and Implementation of Self-Organising Neural Network for Searching Discovery in Databases. Studia Informatica. Systems and Information Technology. Vol. 1(5)/2005, pp. 35-47.

[21] Tchórzewski J., Kłopotek M., Kujawiak M.: Studium porównawcze metod prowadzenia odkryć. Studia Informatica. Systems and Information Technology. Vol. 1(4)/2004, pp. 105-122.

[22] Tchórzewski J., Kwiczak I.: Mapowanie informacji z baz danych za pomocą sieci neuronowych samoorganizujących się. Studia Informatica. Systems and Information Technology. Vol. 1(3)/2004, pp. 99-105.

[23] Tchórzewski J., Zarzycki I., Soćko M.: Poszukiwanie odkryć w rozwijającej się elektroenergetycznej sieci przesyłowej przy wykorzystaniu środowiska MATLAB i Simulink. Studia Informatica. Systems and Information Technology. Vol. 1(2)/2003, pp. 101-109.

[24] Trajer J., Janaszek-Mańkowska M., Mańkowski D. R.: Komputerowa analiza danych w badaniach naukowych, Wyd. SGGW, Warszawa 2016.

[25] Walesiak M., Dudek A.: ClusterSim package, R-Project, 2011.

[26] Witten I.H., Frank E.: Data Mining: Practical Machine Learning Tools and Techniques, IEEE, 2011.

[27] Wierzchoń S., Kłopotek M.: Algorytmy analizy skupień, PWN, Warszawa 2017.

[28] Zhang, E. A.: Graph degree linkage: Agglomerative clustering on a directed graph, 12th European Conference on Computer Vision, Florence Italy 2012.

Internet sources:

[29] https://www.pse.pl

[30] https://www.mathworks.com

[31] http://kognitywistyka.uwb.edu.pl