Jerzy TCHÓRZEWSKI[1]
ORCID: 0000-0003-2198-7185

Bartłomiej LONGOTA[2]
ORCID: 0000-0002-4011-3056

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

Analysis of data quoted on the Day-Ahead Market of TGE S.A. using Statistics and Machine Learning Toolbox

DOI: 10.34739/si.2022.27.03

Abstract. The publication contains the results of research in the field of cluster analysis carried out using data quoted on the Day-Ahead Market of TGE S.A. Two methods were used in the analysis, one hierarchical known as the Ward's method, and the other non-hierarchical - the k-means method. Many interesting research results have been obtained, which are illustrated, among others, in in the form of dendrograms, silhouette graphs and graphs in the form of clusters. Data on the volume and the volume-weighted average price of electricity were examined for various types of quotations: fixing 1, fixing 2 and continuous quotations. The research was carried out in the MATLAB and Simulink environments using a library called Machine and Statistics Learning Toolbox. Selected test results were interpreted.

Keywords. Artificial Neural Network, cluster analysis, Day-Ahead Market, k-means method, Matlab and Simulink environment, Statistics and Machine Learning Toolbox, Ward's method.

1. Introduction

Data analysis is a very important process carried out in the field of operation and development of enterprises, state institutions, and even states and international communities in order to learn their current state and search for new states of development. For these reasons, in every company or institution that wants to develop, there is a position of a business analyst filled by specialists in the field of data analysis. A business analyst prepares data for processing, prepares and interprets statistics and tries to use them to modernize a company or institution in order to obtain, for example, higher benefits in the future and to ensure a competitive advantage, or one of the products on the market using artificial intelligence methods [1, 10, 14, 20-21].

At the moment, we are talking not only about data analysis, but more and more often we are talking about information analysis and even knowledge analysis. This means that larger and larger data sets or more and more complex data warehouses are subject to analysis over time. Their analysis requires an increasing number of treatments and activities related to the preparation of data, information and knowledge for further processing. The purpose of data preprocessing is to check the correctness and completeness of the data, to select characteristic cases among the data and to divide them to facilitate the intended study, as well as to pre-process the data. Correct data analysis can only be performed when the data is correct and well prepared for further processing. This involves the use of newer and newer methods of data cleaning, data transformation and the selection of variables and special cases for analysis [20].

After this type of initial processing of data, it is possible to analyze large data sets, which requires the preparation of appropriate methods for analyzing large data sets, information and even knowledge, which makes the analysis an increasingly complex process. There are basically three methods of data analysis, which are generally carried out using machine learning methods, namely: classification, regression and clustering, also known as grouping. From the point of view of the data used, classification and regression are somewhat similar processes, while clustering is significantly different from them, as shown in Table 1 [3-5, 9-10, 20-21].

Table 1. Summary of significant features in terms of similarity and differences in machine learning methods (classification, regression and clustering). Source: Own study based on [3-5, 9-10, 20-21]

Item	**Classification**	**Regression**	**Grouping (clustering)**
Method name	**Classification method**	**Regression method**	**Grouping method**
The essence of the method	Classifying the input data as one of the class labels of the output variable	Predicting the value of the output variable based on the values of the input variables	Grouping (combining) points into clusters
Results	Obtaining the classes of the analysed quantities	Obtaining the value of the output variable based on the model and the values of the input variables	Obtaining clusters of similar objects

Types of data used	training and testing	training, testing and validation	training
Data label	Labelled data	Data in input-output pairs	Unlabelled data
Stages	Teaching and testing	Teaching, testing and validation	Learning
Complexity	Medium	High	Low

Clustering methods include: cluster analysis methods. Cluster analysis is generally understood as a method of grouping n - objects, described by a vector of p - features, into k non-empty, disjoint and possibly homogeneous groups called clusters [3-6, 9-10, 20-21]. It is said that elements or objects that lie in a certain cluster should have common features of similarity to each other or different features of dissimilarity (dissimilarity) [3, 20-21]. Cluster analysis is supposed to lead to the search for groups of objects in the examined data set with significant similarity to each other, which can be interpreted in a logical and sensible way [14-15, 21]. Cluster analysis methods fall into two main categories, i.e. hierarchical methods and non-hierarchical methods.

Hierarchical methods are characterized by the formation of clusters as a result of the occurrence of a certain hierarchy of data in order to create a cluster structure by observing similarities between them, common features or regularities, etc. In cluster analysis using hierarchical methods, it is initially not necessary to know the number of clusters created [3-6, 9-10, 13]. The main hierarchical methods are: agglomeration methods operating on the principle of combining the resulting clusters and deglomeration methods operating on the principle of dividing the resulting clusters. In order to compare objects present in clusters, the measure of dissimilarity is examined, which is necessary during the cluster analysis process. This measure can be defined as a semi-metric or a simple metric in case it is not necessary to satisfy the triangle condition. Generally speaking, it is the distance between given objects such that the greater it is, the more dissimilar the elements are [21].

On the other hand, objects with a relatively small distance, whose position is sufficiently distant from other objects, form a separate cluster. The measure of distance is defined by a function defined on pairs of objects that takes values of real numbers. There are different types of measures of dissimilarity of objects (or in another way of their similarity), such as Euclidean distance, city distance, Chebyshev distance, power distance, cosine distance, Bergman distance, Minkowski distance, Mahalanobis distance, and many others described in detail among others at work [21].

The most commonly used method is the Euclidean distance, and the other measures of dissimilarity/similarity of objects are rarely used. Also, cluster binding methods are often used in the analysis, which have unique properties and are very easy to apply. The following are distinguished here: methods: single linkage, complete linkage, unweighted pair-group average, weighted pair-group centroid, minimum Ward variance (so-called Ward Algorithm) and a number of other methods [3-6, 10-11, 20-22], where the Ward's method is different from the other methods because the distance between clusters is determined by minimizing the sum of squared deviations within the appropriate group.

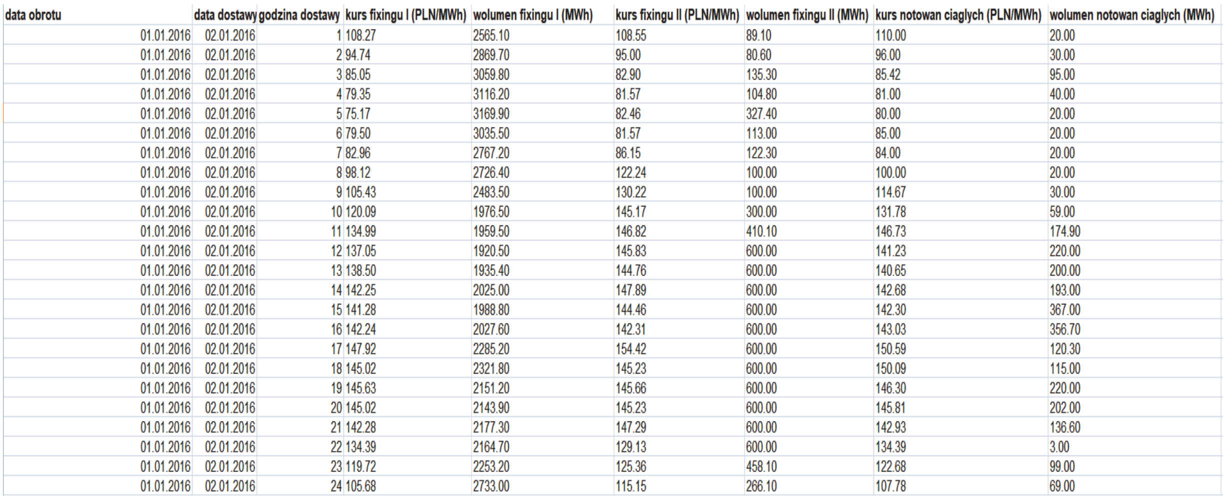
Non-hierarchical methods include, among others: the k-means method, the k-medoid method and artificial self-organizing neural networks (SOM for Self Organizing Maps), in which information about the number of clusters should be known at the beginning of the analysis. Non-hierarchical methods, in contrast to hierarchical methods, do not combine clusters into one cluster, but on the contrary, they create them over time into an increasing number of clusters until the assumed number is obtained. The most commonly used method is the k-means method, which consists in creating k clusters of objects that are as different from each other as possible. In this method, at each stage of the algorithm, the centers of gravity of the clusters (so-called centroids) are determined, with the next values assigned to the closest centers of gravity, which is repeated until the initial number of clusters is obtained [9, 15, 21].

2. Research experiment design

2.1. Research experiment design

The data used in the conducted research experiments relate to the Day-Ahead Market (DAM) system, which is a subsystem of the Polish Power Exchange (TGEE) system, which is in turn a subsystem of Towarowa Giełda Energii S.A. (TGE S.A.) [5, 8, 19]. These data are quoted in hourly contracts on the Day-Ahead Market and cover a period of four years, i.e. from 01/01/2016 to 12/11/2020. They include the volume of electricity and the volume-weighted average price of electricity (the so-called exchange rate) each hour of the day, with quotations relating to fixing 1, fixing 2 and continuous quotations. The structure and fragment of data used to conduct research experiments is presented in the form of Print Screen in Fig. 1.

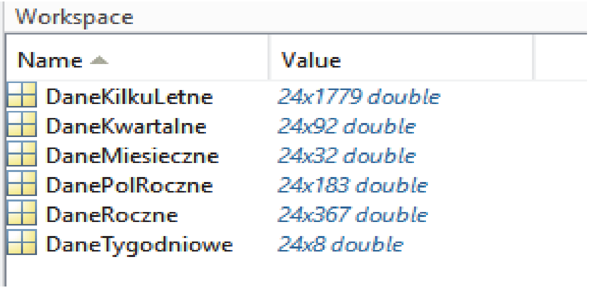
The number of records from the above-mentioned 4 years assumed in the research is 42,672. A significant number of records in the case of the price and volume for continuous trading had incorrectly recorded or missing values, therefore, average values from neighboring records, both for the price and volume, were inserted in these places . Due to the relatively large amount of data, they were divided into appropriate time periods to show the difference between them depending on a larger amount of data, e.g. fixing rate 1 was distinguished for data concerning: week, month, quarter, half-year and year and for the entire period of the analyzed data, i.e. for 4 years, as shown in Fig. 2.



data obrotu	data dostawy	godzina dostawy	kurs fixingu I (PLN/MWh)	wolumen fixingu I (MWh)	kurs fixingu II (PLN/MWh)	wolumen fixingu II (MWh)	kurs notowan ciaglych (PLN/MWh)	wolumen notowan ciaglych (MWh)
01.01.2016	02.01.2016	1	108.27	2565.10	108.55	89.10	110.00	20.00
01.01.2016	02.01.2016	2	94.74	2869.70	95.00	80.60	96.00	30.00
01.01.2016	02.01.2016	3	85.05	3059.80	82.90	135.30	85.42	95.00
01.01.2016	02.01.2016	4	79.35	3116.20	81.57	104.80	81.00	40.00
01.01.2016	02.01.2016	5	75.17	3169.90	82.46	327.40	80.00	20.00
01.01.2016	02.01.2016	6	79.50	3035.50	81.57	113.00	85.00	20.00
01.01.2016	02.01.2016	7	82.96	2767.20	86.15	122.30	84.00	20.00
01.01.2016	02.01.2016	8	98.12	2726.40	122.24	100.00	100.00	20.00
01.01.2016	02.01.2016	9	105.43	2483.50	130.22	100.00	114.67	30.00
01.01.2016	02.01.2016	10	120.09	1976.50	145.17	300.00	131.78	59.00
01.01.2016	02.01.2016	11	134.99	1959.50	146.82	410.10	146.73	174.90
01.01.2016	02.01.2016	12	137.05	1920.50	145.83	600.00	141.23	220.00
01.01.2016	02.01.2016	13	138.50	1935.40	144.76	600.00	140.65	200.00
01.01.2016	02.01.2016	14	142.25	2025.00	147.89	600.00	142.68	193.00
01.01.2016	02.01.2016	15	141.28	1988.80	144.46	600.00	142.30	367.00
01.01.2016	02.01.2016	16	142.24	2027.60	142.31	600.00	143.03	356.70
01.01.2016	02.01.2016	17	147.92	2285.20	154.42	600.00	150.59	120.30
01.01.2016	02.01.2016	18	145.02	2321.80	145.23	600.00	150.09	115.00
01.01.2016	02.01.2016	19	145.63	2151.20	145.66	600.00	146.30	220.00
01.01.2016	02.01.2016	20	145.02	2143.90	145.23	600.00	145.81	202.00
01.01.2016	02.01.2016	21	142.28	2177.30	147.29	600.00	142.93	136.60
01.01.2016	02.01.2016	22	134.39	2164.70	129.13	600.00	134.39	3.00
01.01.2016	02.01.2016	23	119.72	2253.20	125.36	458.10	122.68	99.00
01.01.2016	02.01.2016	24	105.68	2733.00	115.15	266.10	107.78	69.00

Figure 1. Data quoted on the Day-Ahead Market of TGE S.A. for the first 24 hours of the analyzed period. Source: Own study [5, 19].

Concepts and research methodology, especially in the field of conducting discoveries, were published, among others in works [2, 16-17], the results of a comparative study of methods of conducting discoveries at work [18], and the preliminary results of research in this area were published, among others, in at work [15]. It is worth adding that in the field of searching for models of the Day-Ahead Market system of TGE S.A., especially neural models, significant research results have been published, e.g. in works [11-12].



Workspace

Name	Value
DaneKilkuLetne	24x1779 double
DaneKwartalne	24x92 double
DaneMiesieczne	24x32 double
DanePolRoczne	24x183 double
DaneRoczne	24x367 double
DaneTygodniowe	24x8 double

Figure 2. An example of data used in the analysis, taking into account different periods. Designations: DaneKilkuLetnie - Several Years Data (here: four years), DaneKwartalne - Quarterly Data, DaneMiesieczne - Monthly Data, DanePolRoczne - Half Year Data, DaneRoczne - Yearly Data, DaneTygodniowe - Weekly Data. Source: Own elaboration in MATLAB and Simulink environment [5-6].

On the other hand, non-hierarchical methods include, among others: the k-means method, the k-medoid method and self-organizing artificial neural networks (SOM), which have been exhaustively described, e.g. in the book by S. Wierzchoń and M. Kłopotek entitled Algorithms of cluster analysis [26]. In these methods, initial information about the expected number of clusters is very important. This means that the non-hierarchical method, unlike hierarchical methods, does not combine clusters into one cluster, but on the contrary, separates them to obtain the desired number of groups (clusters) as a result [21]. Among these methods, an important method is the k-means method, which consists in creating k-clusters of objects that differ as much as possible from each other.

Thus, the first part of this method is for the researcher to specify the number of clusters into which the data collected in the set are to be divided. The algorithm then assigns some specific data points as initial cluster centers, which can be chosen depending on the method adopted. Then, for each data, the nearest cluster center is assigned, after which the resulting new cluster is checked and a new center of gravity (centroid) is determined. The algorithm repeats these steps until the user reaches the specified number of clusters. Next, an error function is determined, which may cause the algorithm to stop due to the lack of significant changes in the value of this function.

2.2. Data analysis project

In the field of data analysis, the MATLAB and Simulink environments provide a library of m-files called Statistics and Machine Learning Toolbox (SMLT), which allows the use of many dedicated methods related to cluster analysis [5-7]. Exploratory data analysis concerns e.g. grouping, matching probability distributions, generating random numbers needed for simulations, and testing given hypotheses, etc. [6].

Regression and classification algorithms support the extraction of knowledge from data, and support the construction of predictive models along with interaction using properly prepared applications, or in a programmatic way. For these reasons, SMLT is a willingly used application in the process of data regression, classification and clustering, especially in the search for various specific features between data and subgroups. With the use of clusters, i.e. clusters of data created by appropriate methods, it is possible to e.g. note whether the analyzed data have similar values, or on the contrary - there are cases of large differences between certain data and other data, including dominant data. In this regard, it is also usually checked how individual values are recorded on particular days and how algorithms behave when dividing objects into clusters under the pressure of increasing the number of target clusters, e.g. for the k-means method. In the case of clustering using the Ward's method, there is no need for a similar check, because in this case the aim is to obtain the smallest number of groups, i.e. one group in the end [10, 13, 21].

The data analysis experiment carried out in this study aims to demonstrate clustering using two types of analysis methods, namely the hierarchical method and the non-hierarchical method. The data used in the experiment are, therefore, data on the volume and price of electricity quoted on the TGE S.A. Day-Ahead Market. In the hierarchical method, these data were analyzed using the agglomeration method using the Ward algorithm, while in the case of the non-hierarchical method, the analysis was carried out using the k-means method [10, 21].

In the program implementation, solutions presented by the creators of the MATLAB environment were used, which were exhaustively explained in the description of the Machine and Statistics Learning Toolbox, e.g. for the cluster analysis process using both described methods, i.e. the Ward's method and the k-means method [10, 13, 21]. In addition, in this respect, the differences between the test results for different time intervals were also shown, with the obtained results presented on the appropriate graphs so as to facilitate their interpretation. The experiment in this regard used m-files built using specially programmed methods supporting the process of cluster analysis, such as classifying functions and evaluation functions [5-6] of the type: [5-6]:

kmeans() - a method supporting the grouping of data into appropriate clusters, which is to create data clusters that match each other, and depending on the "k" variable specified in the parameter, any number of final clusters can be obtained;

pdist() – a method that aims to determine the distance between pairs of observations, which by default returns the Euclidean distance, and when there is a distance parameter, then various other distance measures can be used, such as cosine distance or Minkowski distance;

silhouette() – a method that allows to create a graph of the results based on data aggregated using different distance metrics;

linkage() – a method for creating an agglomeration hierarchical cluster tree;

dendrogram() - a method used to generate a dendrogram plot of a cluster tree, which consists of multiple lines connecting data points, where the height represents the distance between two connected values.

In the library of the MATLAB and Simulink environment entitled Statistics and Machine Learning Toolbox has many different datasets may be clustered using the k-means method and the hierarchical method, where the k-means function is used for clustering using an algorithm that assigns specific elements to clusters in such a way that, for example, the sum of the distances of each object from its center of gravity is as small as possible.

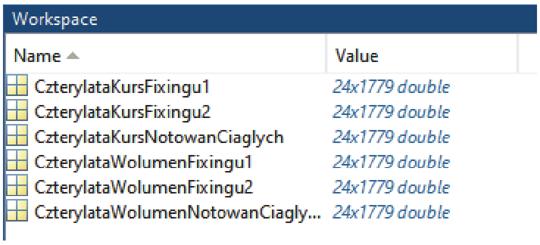
3. Implementation of selected methods

3.1. Data preparation

The data for the purposes of the experiment were transposed in such a way as to be correctly read, with the data quoted in fixing 1, fixing 2 and continuous trading separated from each other in order to clearly distinguish each of the measures. Then, the pre-prepared data in MS Excel were imported to Workspace of MATLAB and Simulink [5-6, 15].

3.2. K-means data analysis

After proper preparation of data, broken down into appropriate data types as shown in Fig. 3, they were introduced to the Workspace of the MATLAB and Simulink environments in the form of properly prepared matrices for the price and volume of fixing 1, fixing 2 and continuous trading conducted in the course of quotations on the stock exchange at particular hours of the day. It is worth noting that the data set includes hourly quotations of transactions concerning the volume of supplied and sold electricity and the obtained volume-weighted average electricity.



Workspace

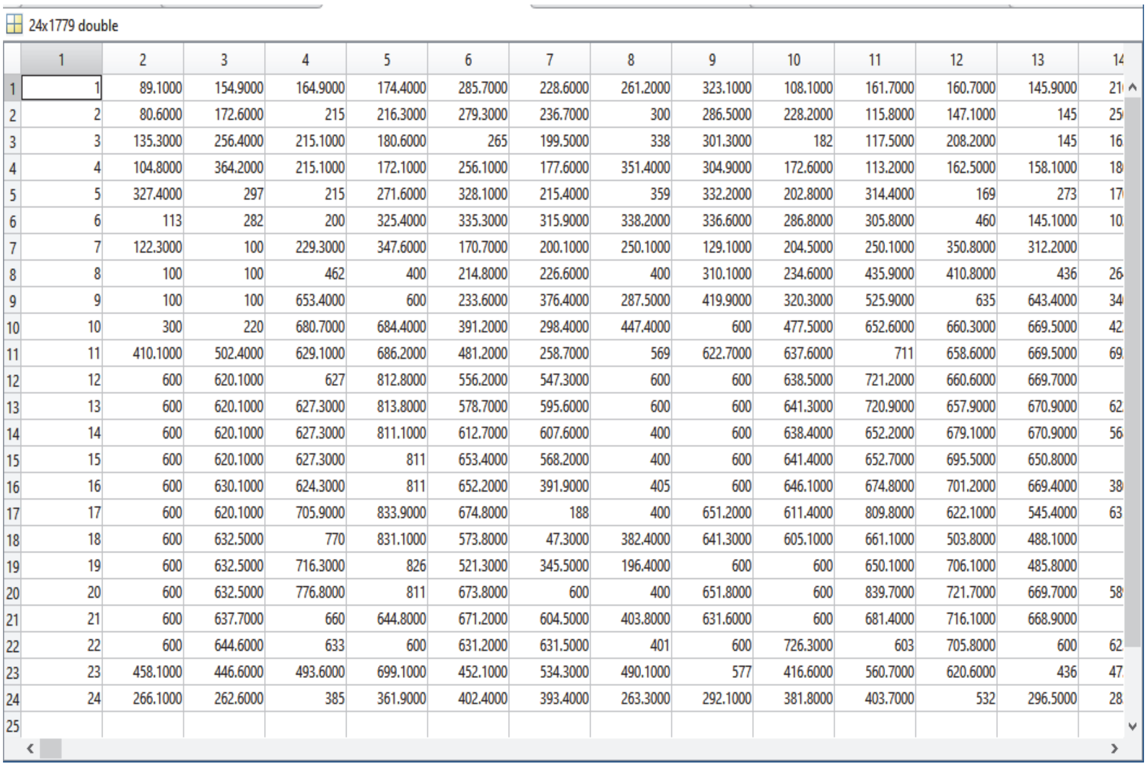
Name ▲	Value
CzterylataKursFixingu1	24x1779 double
CzterylataKursFixingu2	24x1779 double
CzterylataKursNotowanCiaglych	24x1779 double
CzterylataWolumenFixingu1	24x1779 double
CzterylataWolumenFixingu2	24x1779 double
CzterylataWolumenNotowanCiagly...	24x1779 double

Figure 3. MATLAB and Simulink Workspace content. Designations: CzterylataKursFixingu1 – Four Years Fixing1 Rate, CzterylataKursFixingu2 – Four Years Fixing2 Rate, CzterylataKursNotowańCiagłych – Four Years Continous Trading Rate, CzterylataWolumenFixingu1 – Four Years Fixing Volume1, CzterylataWolumenFixingu2 – FourYearsFixingVolume2, CzterylataWolumenNotowańCiaglych - Four Years Continuous Trading Volume. Source: Own elaboration in MATLAB and Simulink [5-7].

On the other hand, Fig. 4 shows in detail the data on the fixing 2 volume, where it can be noticed, among others, that the first column contains individual delivery times, while the subsequent columns contain the existing values of the fixing 2 volume for subsequent days on which electricity was supplied. The matrix has 24 rows x 1 779 columns, which makes it easy to pre-analyze individual delivery days.

The method responsible for non-hierarchical grouping used in the research experiment is the k-means method, which is used to assign objects to appropriate clusters so that the sum of the distances from each object to its center, i.e. from the centroid, is as small as possible, where in the case under consideration, the default distance measure used by the k-means method to initialize the center of a given cluster is the square Euclidean distance.

A very important way of visualizing the obtained results is the silhouette graph, which allows to plot the course of particular clusters for the set, in this case for the input data matrix, taking into account the assignment of clusters to each observation point [5-6].



24x1779 double

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	1	89.1000	154.9000	164.9000	174.4000	285.7000	228.6000	261.2000	323.1000	108.1000	161.7000	160.7000	145.9000	21
2	2	80.6000	172.6000	215	216.3000	279.3000	236.7000	300	286.5000	228.2000	115.8000	147.1000	145	25
3	3	135.3000	256.4000	215.1000	180.6000	265	199.5000	338	301.3000	182	117.5000	208.2000	145	16
4	4	104.8000	364.2000	215.1000	172.1000	256.1000	177.6000	351.4000	304.9000	172.6000	113.2000	162.5000	158.1000	18
5	5	327.4000	297	215	271.6000	328.1000	215.4000	359	332.2000	202.8000	314.4000	169	273	17
6	6	113	282	200	325.4000	335.3000	315.9000	338.2000	336.6000	286.8000	305.8000	460	145.1000	10
7	7	122.3000	100	229.3000	347.6000	170.7000	200.1000	250.1000	129.1000	204.5000	250.1000	350.8000	312.2000	
8	8	100	100	462	400	214.8000	226.6000	400	310.1000	234.6000	435.9000	410.8000	436	26
9	9	100	100	653.4000	600	233.6000	376.4000	287.5000	419.9000	320.3000	525.9000	635	643.4000	34
10	10	300	220	680.7000	684.4000	391.2000	298.4000	447.4000	600	477.5000	652.6000	660.3000	669.5000	42
11	11	410.1000	502.4000	629.1000	686.2000	481.2000	258.7000	569	622.7000	637.6000	711	658.6000	669.5000	69
12	12	600	620.1000	627	812.8000	556.2000	547.3000	600	600	638.5000	721.2000	660.6000	669.7000	
13	13	600	620.1000	627.3000	813.8000	578.7000	595.6000	600	600	641.3000	720.9000	657.9000	670.9000	62
14	14	600	620.1000	627.3000	811.1000	612.7000	607.6000	400	600	638.4000	652.2000	679.1000	670.9000	56
15	15	600	620.1000	627.3000	811	653.4000	568.2000	400	600	641.4000	652.7000	695.5000	650.8000	
16	16	600	630.1000	624.3000	811	652.2000	391.9000	405	600	646.1000	674.8000	701.2000	669.4000	38
17	17	600	620.1000	705.9000	833.9000	674.8000	188	400	651.2000	611.4000	809.8000	622.1000	545.4000	63
18	18	600	632.5000	770	831.1000	573.8000	47.3000	382.4000	641.3000	605.1000	661.1000	503.8000	488.1000	
19	19	600	632.5000	716.3000	826	521.3000	345.5000	196.4000	600	600	650.1000	706.1000	485.8000	
20	20	600	632.5000	776.8000	811	673.8000	600	400	651.8000	600	839.7000	721.7000	669.7000	58
21	21	600	637.7000	660	644.8000	671.2000	604.5000	403.8000	631.6000	600	681.4000	716.1000	668.9000	
22	22	600	644.6000	633	600	631.2000	631.5000	401	600	726.3000	603	705.8000	600	62
23	23	458.1000	446.6000	493.6000	699.1000	452.1000	534.3000	490.1000	577	416.6000	560.7000	620.6000	436	47
24	24	266.1000	262.6000	385	361.9000	402.4000	393.4000	263.3000	292.1000	381.8000	403.7000	532	296.5000	28
25														

Figure 4. Matrix contents CzterylataWolumenFixingu2. Designations: column 1 – delivery hours, next columns – fixing volume2 values for subsequent days. Source: Own study [5-6]

It is worth noting that due to the fact that the electricity volume values for individual hours of the day and for individual days and types of quotations were of the same order, and similarly, the volume-weighted average electricity prices for individual hours of the day and for individual days and types of quotations were of the same order, so the data was not normalized. The silhouette graph allows to find out if the clusters are well separated from each other and shows how close to each point in one cluster is a point in neighboring groups. It is worth noting here that an appropriate measure of distance was given here, i.e. the square Euclidean distance (sqeuclidean) was used, which is shown in Listing 1, and the obtained result is shown in Fig. 5. clusters defined in the k-means method, i.e. with the number of clusters: 3, 5, 7 and cluster 9.

```
[cidx2,cmeans2] = kmeans(CzterylataWolumenFixingu2,5,'dist','sqeuclidean');
[silh2,h] = silhouette(CzterylataWolumenFixingu2,cidx2,'sqeuclidean');
```

Listing 1. An example of calling the k-means method on data related to the Day-Ahead Market system. Source: Own elaboration in MATLAB and Simulink [5-7]

It is worth noting that the higher the silhouette value (maximum it can be 1), the better the points in clusters are separated from neighboring groups. However, this is not always an unambiguous situation. Table 2 shows that not all clusters are sufficiently well separated. In order to conduct an in-depth analysis of the data, they were prepared for analysis in such a way (Figure 5) that the columns concern, respectively, starting from the left: delivery time, fixing price 1, fixing volume 1, fixing price 2, fixing volume 2, continuous trading price and volume continuous trading. Due to the relatively large number of analyzed data and their richness, an attempt was made to show the results of the analysis on the example of the first week in such a way as to better show the assignment of objects to the appropriate clusters. The presented m-file code written in Matlab shows the process of creating a three-dimensional graph based on a set of input data, which is a matrix containing information covering all data for the first week.

Due to the fact that the values of individual prices and volumes for fixing 1, fixing 2 and continuous trading are similar and the amount of information in the set is quite large, exceeding 40 thousand. Therefore, to increase the transparency of the obtained results, they are shown on the example of the first week. As a result of the analysis, a chart is obtained with data on the values of individual columns according to specific delivery times.

The objects were appropriately divided using the k-means method, the operation of which is shown in Listing 2. They belong to a specific cluster and have been highlighted in the graph in such a way that they differ from other data belonging to a different cluster. In order to present the results in three-dimensional form using the plot3() m-file, the following columns 1, 2 and 3 representing the delivery time, the fixing volume 1 and the fixing volume 2 were assigned respectively to the successive x, y and z axes. The results in the form of graphs showing the elements grouped to the appropriate clusters were prepared using a 3D graph, in which each axis is responsible for the given values. Objects representing the values of the fixing 1 and fixing 2 volumes, respectively: week, quarter, half-year, year and 4 years, are presented in Table 3.

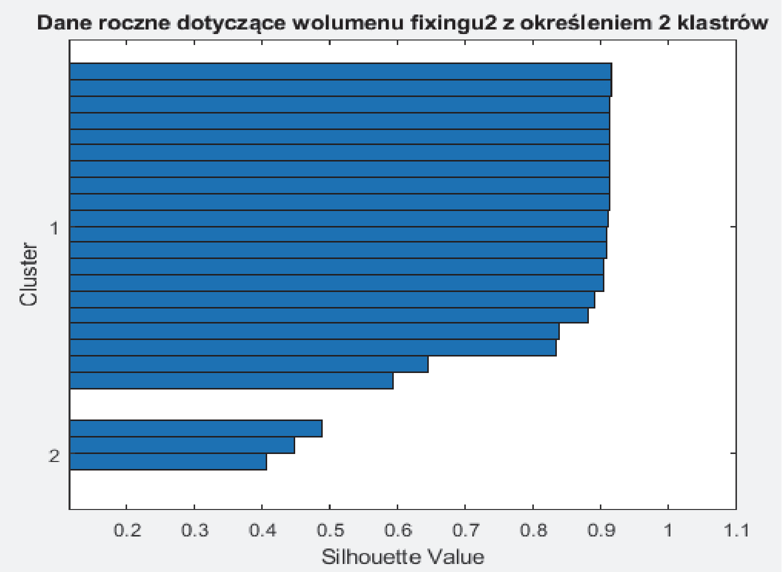
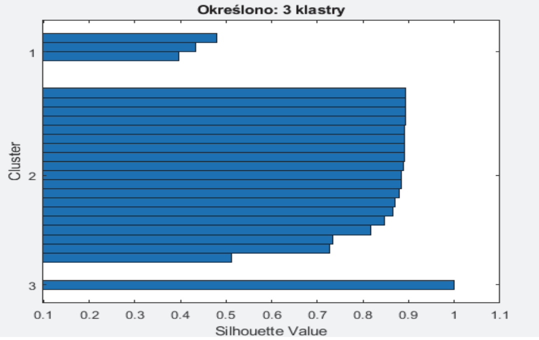
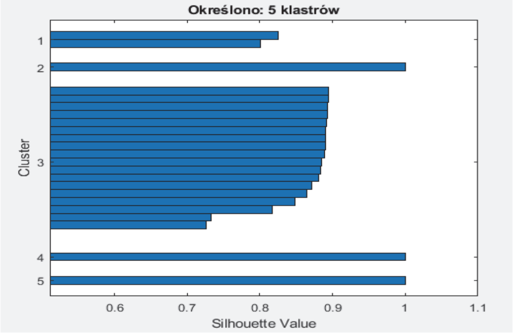
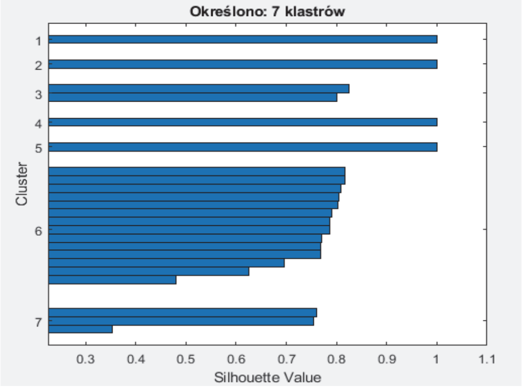
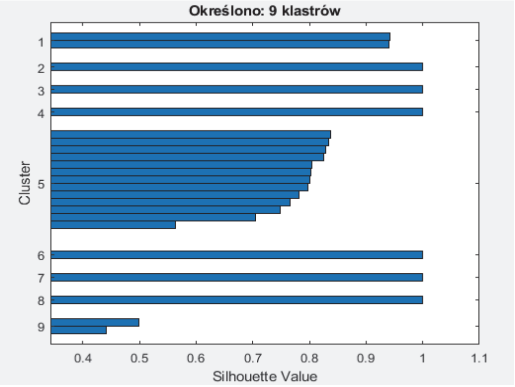
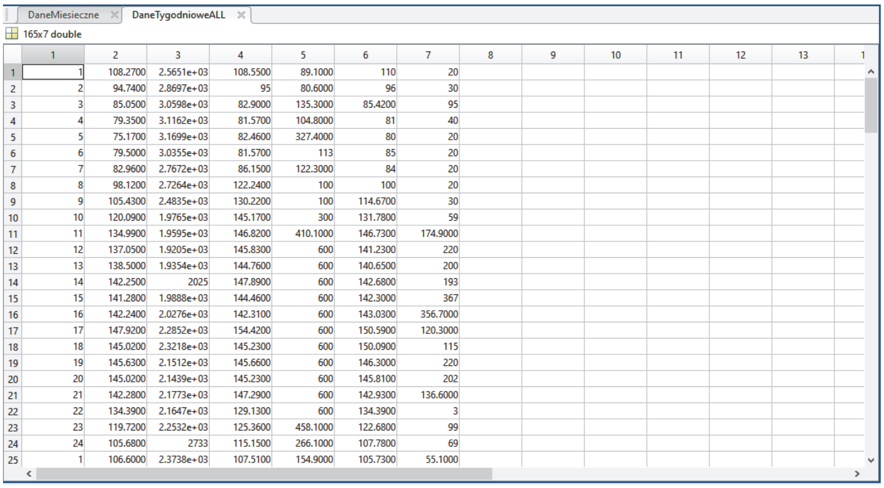


Figure 5. Silhouette chart of fixing 2 volume broken down into 2 clusters. Source: Own elaboration in the MATLAB and Simulink environment using SMLT [5-7]

Table 2. Silhouette graph for the number of clusters: 3, 5, 7 and 9. Source: Own elaboration in the MATLAB and Simulink environment using SMLT [5-7]

Wykres sylwetkowy	Wykres sylwetkowy
	
	

Differences in creating clusters by increasing the number of clusters to 3 are shown in Table 4. Data grouping in Fig. 6 and Fig. 7 shows the data corresponding to the data from the first week with the difference that 4 and 5 clusters were used.



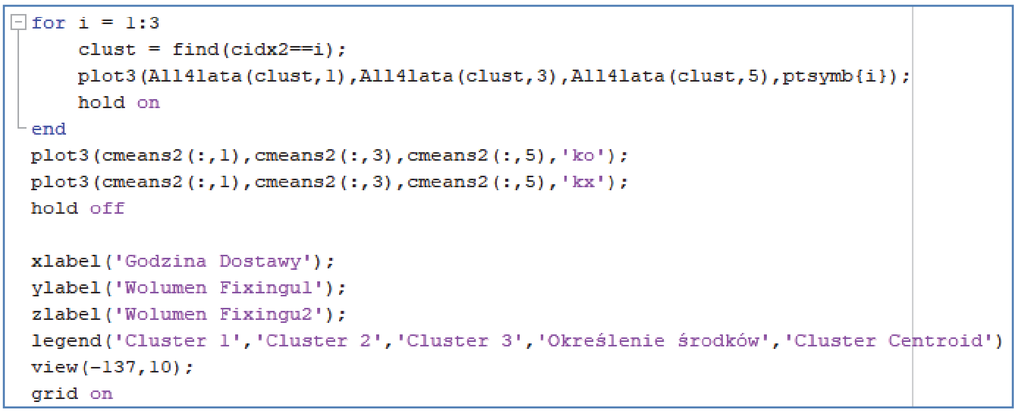
165x7 double

	1	2	3	4	5	6	7
1	1	108.2700	2.5651e+03	108.5500	89.1000	110	20
2	2	94.7400	2.8697e+03	95	80.6000	96	30
3	3	85.0500	3.0598e+03	82.9000	135.3000	85.4200	95
4	4	79.3500	3.1162e+03	81.5700	104.8000	81	40
5	5	75.1700	3.1699e+03	82.4600	327.4000	80	20
6	6	79.5000	3.0355e+03	81.5700	113	85	20
7	7	82.9600	2.7672e+03	86.1500	122.3000	84	20
8	8	98.1200	2.7264e+03	122.2400	100	100	20
9	9	105.4300	2.4835e+03	130.2200	100	114.6700	30
10	10	120.0900	1.9765e+03	145.1700	300	131.7800	59
11	11	134.9900	1.9595e+03	146.8200	410.1000	146.7300	174.9000
12	12	137.0500	1.9205e+03	145.8300	600	141.2300	220
13	13	138.5000	1.9354e+03	144.7600	600	140.6500	200
14	14	142.2500	2025	147.8900	600	142.6800	193
15	15	141.2800	1.9888e+03	144.4600	600	142.3000	367
16	16	142.2400	2.0276e+03	142.3100	600	143.0300	356.7000
17	17	147.9200	2.2852e+03	154.4200	600	150.5900	120.3000
18	18	145.0200	2.3218e+03	145.2300	600	150.0900	115
19	19	145.6300	2.1512e+03	145.6600	600	146.3000	220
20	20	145.0200	2.1439e+03	145.2300	600	145.8100	202
21	21	142.2800	2.1773e+03	147.2900	600	142.9300	136.6000
22	22	134.3900	2.1647e+03	129.1300	600	134.3900	3
23	23	119.7200	2.2532e+03	125.3600	458.1000	122.6800	99
24	24	105.6800	2733	115.1500	266.1000	107.7800	69
25	1	106.6000	2.3738e+03	107.5100	154.9000	105.7300	55.1000

Figure 6. Weekly data on all volume and volume-weighted electricity (ee) values quoted in particular hours of the day (fixing 1, fixing 2, continuous trading). Symbols: column 1 – delivery time, column 2 – fixing price 1, column 3 – fixing price 1, column 4 – fixing price 2, column 5 – fixing price 2, column 6 – continuous trading price, column 7 – trading volume continuous. Source: Own elaboration in the MATLAB and Simulink environment using SMLT [5-7]

During the experiment conducted using the non-hierarchical k-means method, objects were grouped into groups called clusters, obtaining many different graphs showing the forms of data divided into a specific number of clusters, as well as graphs showing clusters of objects (Table 3). In the silhouette chart of the annual volume values of electricity (ee) fixing 2 divided into 2 clusters, it can be seen that the silhouette values in the first cluster are correct, because they take values close to 1, which is the maximum value that can be obtained. The second cluster takes values close to 0.4 and 0.5, which indicates poor separation of objects from other groups in this cluster.

The following charts in Table 3 show how the data in a specific data set are matched to individual groups in terms of the number of clusters. It is clearly visible that with the increase in the number of clusters, the data in individual groups are characterized by the correct silhouette.



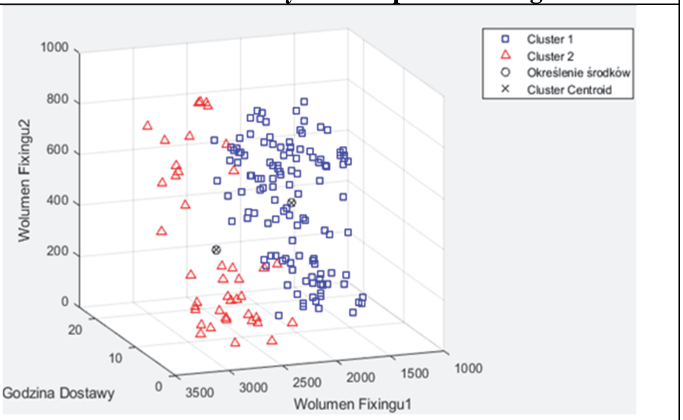
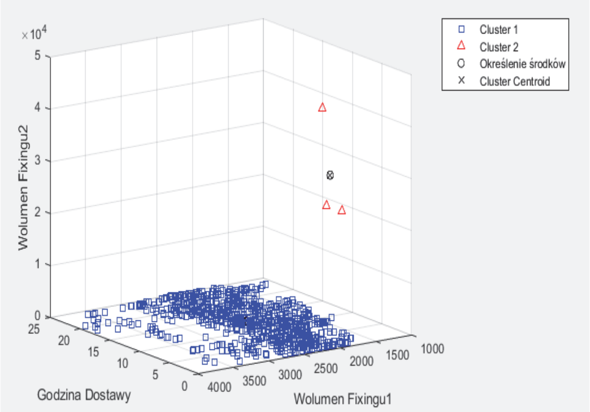
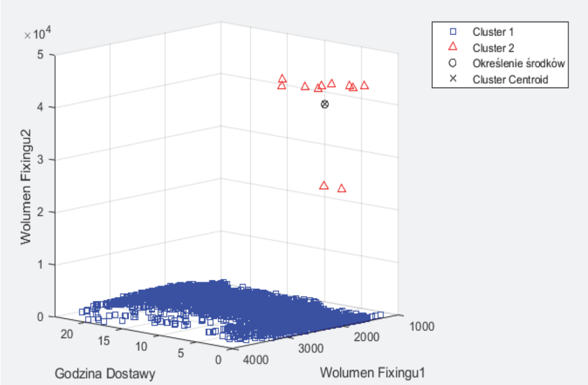
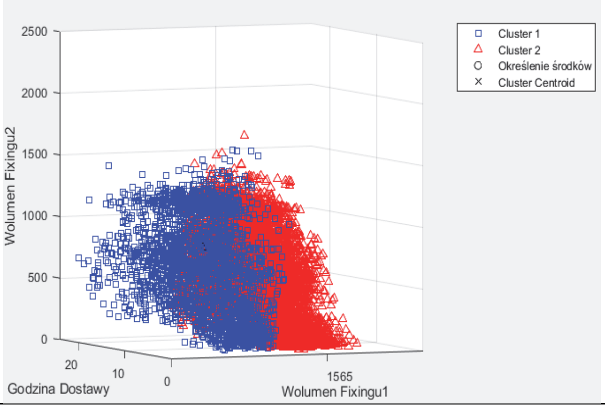
```
for i = 1:3
    clust = find(cidx2==i);
    plot3(All4lata(clust,1),All4lata(clust,3),All4lata(clust,5),ptsymb{i});
    hold on
end
plot3(cmeans2(:,1),cmeans2(:,3),cmeans2(:,5),'ko');
plot3(cmeans2(:,1),cmeans2(:,3),cmeans2(:,5),'kx');
hold off

xlabel('Godzina Dostawy');
ylabel('Wolumen Fixingu1');
zlabel('Wolumen Fixingu2');
legend('Cluster 1','Cluster 2','Cluster 3','Określenie środków','Cluster Centroid')
view(-137,10);
grid on
```

Listing 2. Creating a three-dimensional chart using the m-file plot3() for fixing volume 1 and fixing volume 2. Source: Own study in Matlab [5-7]

However, not all the obtained clusters have the same amount of data, and the clusters in which the silhouette value is greater than 0.8 can be considered sufficiently separated from the others. With the number of clusters equal to 3 (Table 4), you can see that cluster number 2 is the largest in size, and with the number of clusters equal to 5, another cluster, cluster number 3, has the largest number of features. The division into groups is similar also for the number of clusters equal to 7 and 9.

In the next stage of the research, data sets were used, which were divided into specific periods of time: a week, a quarter, a half-year, a year and 4 years. The ee fixing 1 volume and the ee fixing 2 volume were selected as the observed data, juxtaposing them (Table 3 and Table 4). Starting from the initial one, we see how the algorithm divided the objects between the two resulting clusters.

Table 3. Objects representing the values of fixing 1 and fixing 2 volumes for: week, quarter, half-year, year and 4 years, respectively. Designations: Godzina dostawy - Delivery time, Source: Own elaboration in the MATLAB and Simulink environment using SMLT [5-7]

Description	The result of data analysis in a spatial arrangement
Weekly data including fixing1 and fixing2 volume	
Quarterly data including fixing1 and fixing2 volumes	
. Semi-annual data including fixing1 and fixing2 volumes	
Annual data including fixing1 and fixing2 volumes	

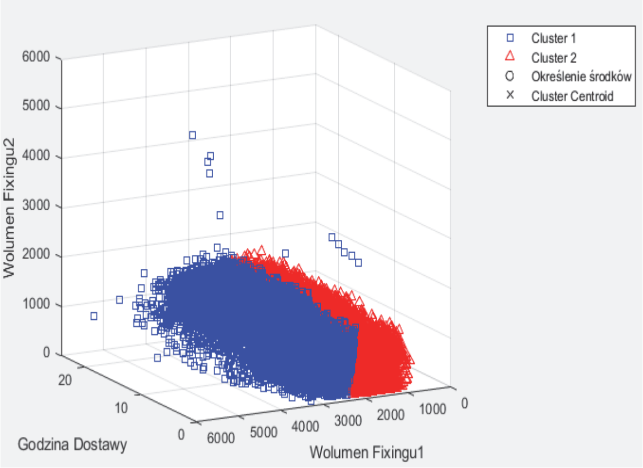
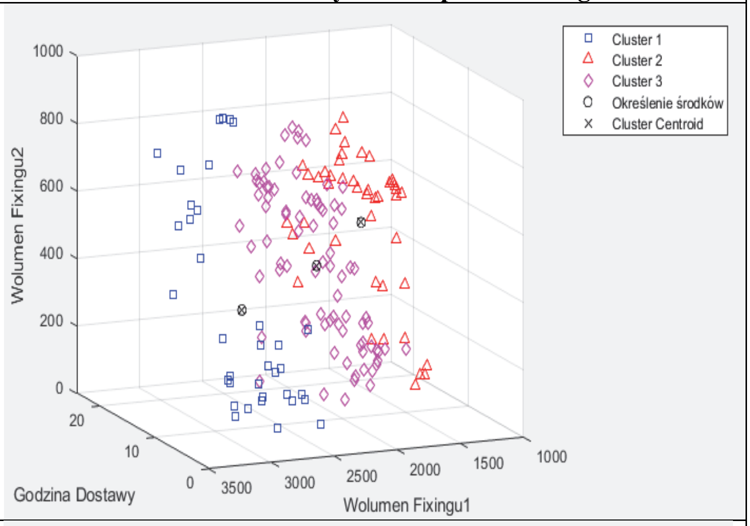
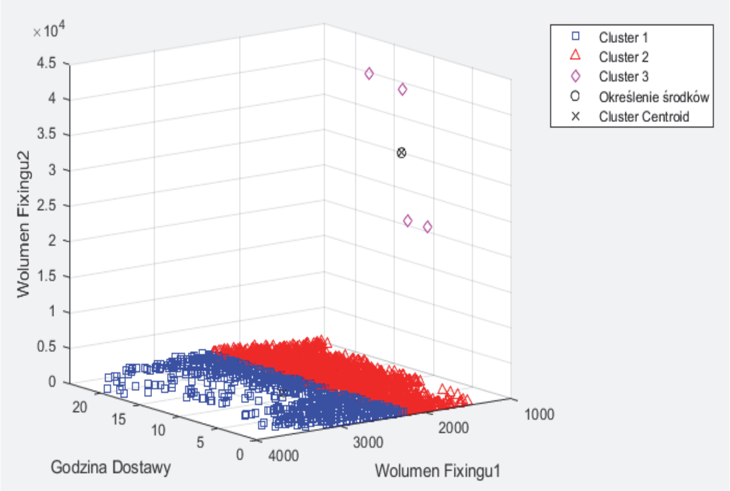
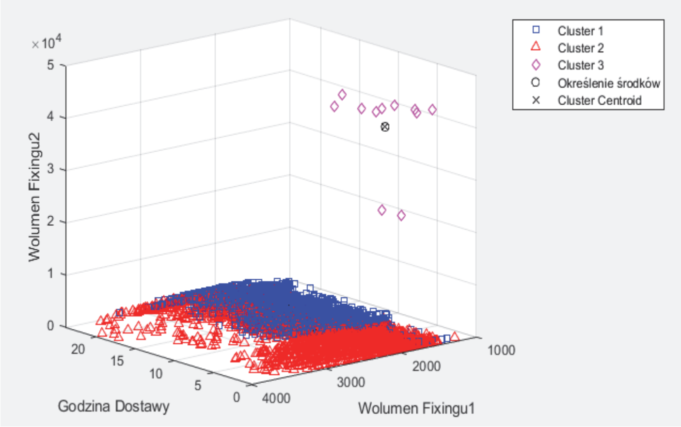
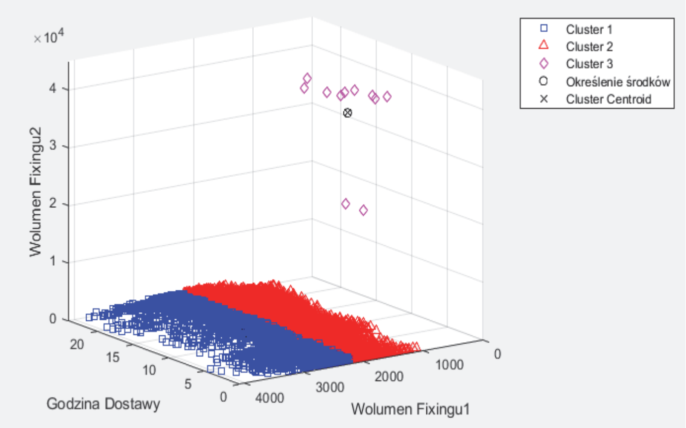
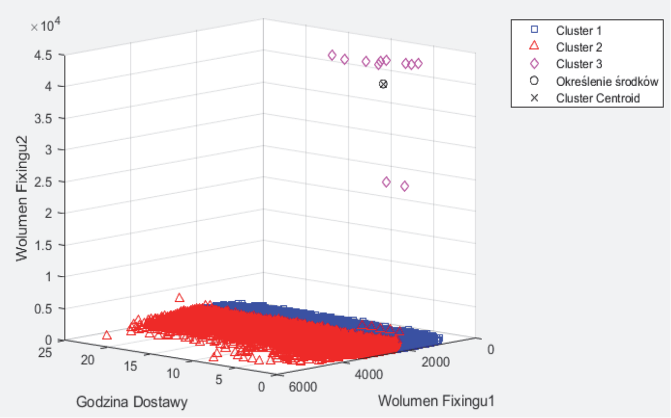
4-year data covering fixing1 and fixing2 volumes	

Table 4. Differences in the formation of clusters by increasing the number of clusters to 3 for the value of the fixing 1 and fixing 2 volumes of the week, quarter, half-year, year and 4 years, respectively. Designations: Godzina dostawy – Delivery time, Source: Own elaboration in the MATLAB and Simulink environment using SMLT [5-7]

Description	The result of data analysis in a spatial arrangement
Weekly data including fixing1 and fixing2 volume with 3 clusters	
Quarterly data including the volume of fixing1 and fixing2 with the number of clusters equal to 3	

Semi-annual data including fixing1 and fixing2 volumes with the number of clusters equal to 3	
Annual data including the volume of fixing1 and fixing2 with the number of clusters equal to 3	
Data from 4 years covering the volume of fixing1 and fixing2 with the number of clusters equal to 3	

Two centers of each of the clusters were determined and marked accordingly on the graph. The sets are adequately separated from each other, while some values are quite close to those of the neighboring cluster. The following charts show the results for the quarter. The algorithm generated a situation where the most distant objects had their cluster, hence all values at the bottom of the graph were assigned to one and the same group. It can be concluded that in this case the groups are very well separated from each other and no elements from different clusters are adjacent. A similar situation also occurred for the six-month period, with more data, but no significant difference in the appearance of the clusters, the clusters are still far apart and none of the objects is close to neighboring groups.

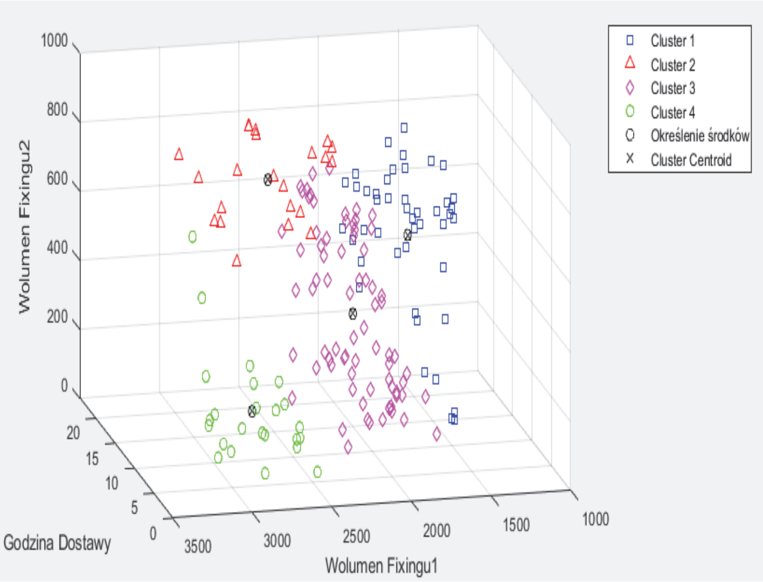


Figure 7. Weekly data including the volume of ee fixing1 and fixing2 with the number of clusters equal to 4. Designations: Godzina dostawy - Delivery time, Source: Own study in the MATLAB and Simulink environment using SMLT [5-7]

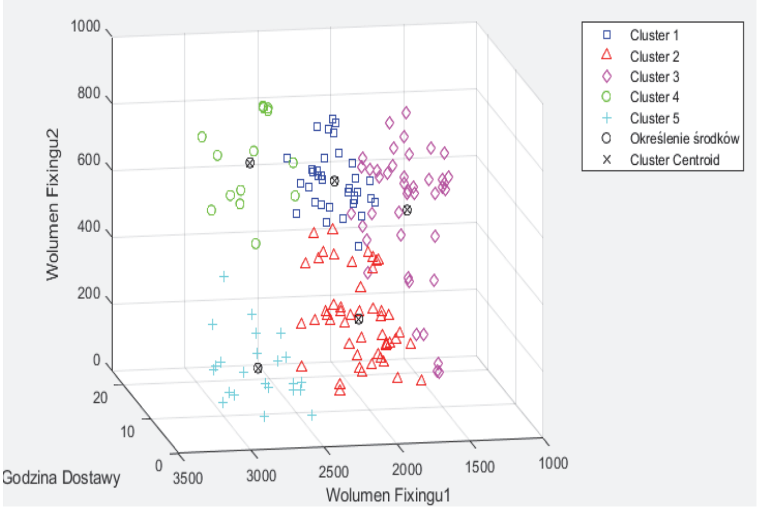


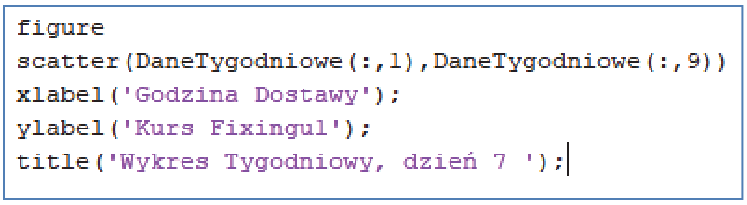
Figure 8. Weekly data including fixing 1 and fixing 2 ee volume with 5 clusters. Desigantions: Godzina dostawy – Delivery time, Source: Own elaboration in the MATLAB and Simulink environment using SMLT [5-7].

However, in the case of the whole year, you can see how the k-means method split the data, placing the randomized centroids not as far apart as you might expect. For these reasons, the scale of the y-axis has been changed to improve the visibility of clusters. In the next step, the number of clusters was increased to learn about the behavior of the algorithm and how objects would be divided between clusters. At that time, there were 3 clusters, the centers of which were located at a slight distance, but sufficient for an unambiguous division, with the values of the ee fixing 2 volume amounting to about a thousand, while the values on the X axis amount to as much as 3.5 thousand electricity volume values, which results from the fact that that they concern a period of one week from the entire set containing information from the entire period under examination, i.e. from 4 years.

3.3. Ward's data analysis

As a result of data analysis using the Ward‘s method, among others, in the form of a chart, the value of the ee fixing 1 exchange rate in each of the 24 hours of a specific day of electricity delivery. It has been established that the day presented is each of the last days of the period that have been prepared in advance. In table 5 the following days were taken into account for the relevant periods used in the data analysis: week period - 7th day, month period - 31st day, quarterly period - 91st day, half-year period - 182nd day, yearly period - 366th day, 4-year period – 1 778 days.

The M-file used in this analysis is shown in Listing 3, while the "scatter" function was used in the visualization of the results in order to best reflect the numerical values present in the data set. The X-axis shows the delivery times of the day, while the Y-axis shows the value of the fixing 1 rate during these hours. As input data, the data of the first week was used, with an indication of the hour of deliveries on the last day of this period. This made it possible to check at what times on a given day the value of the fixing 1 exchange rate was the highest and lowest.

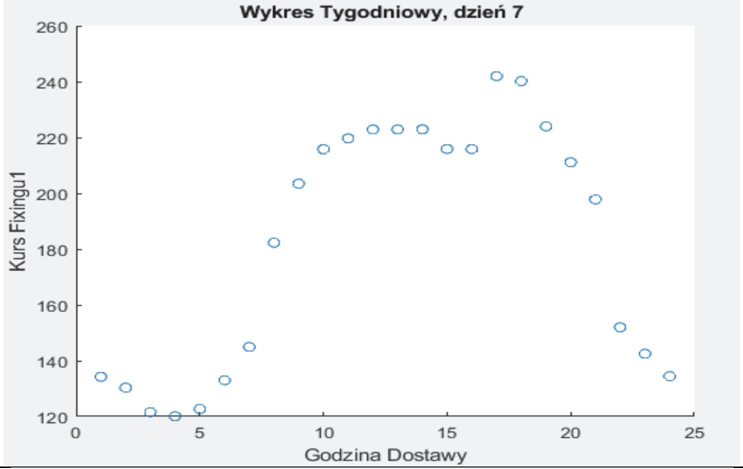
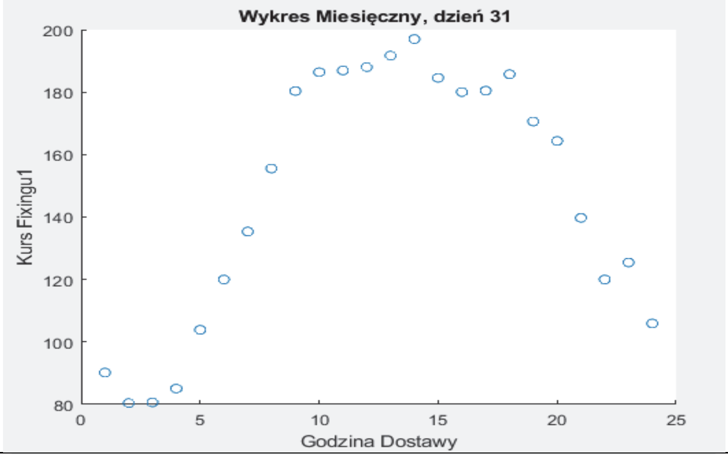
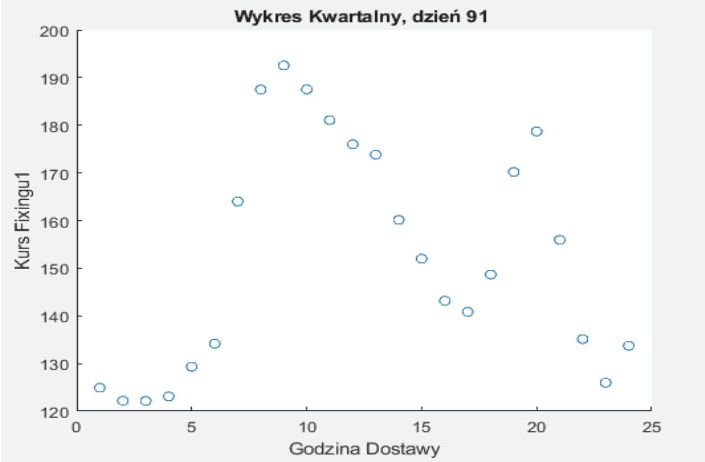


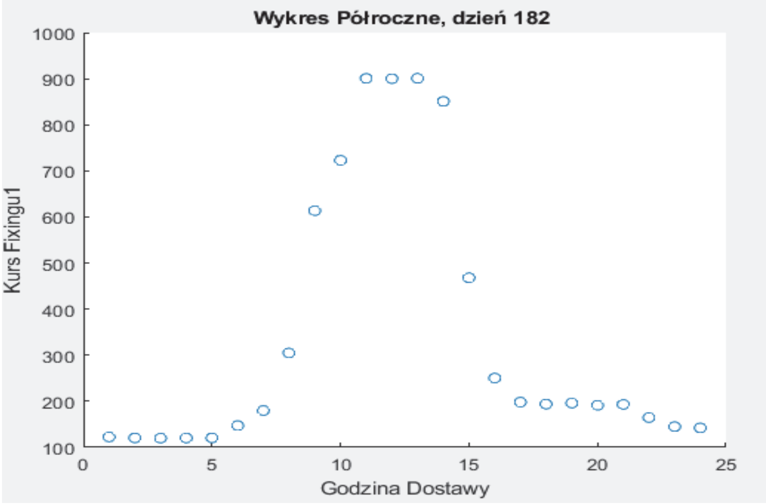
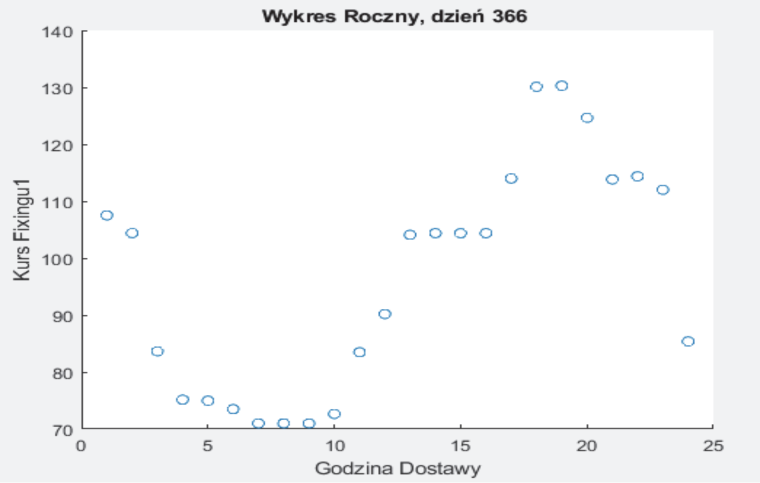
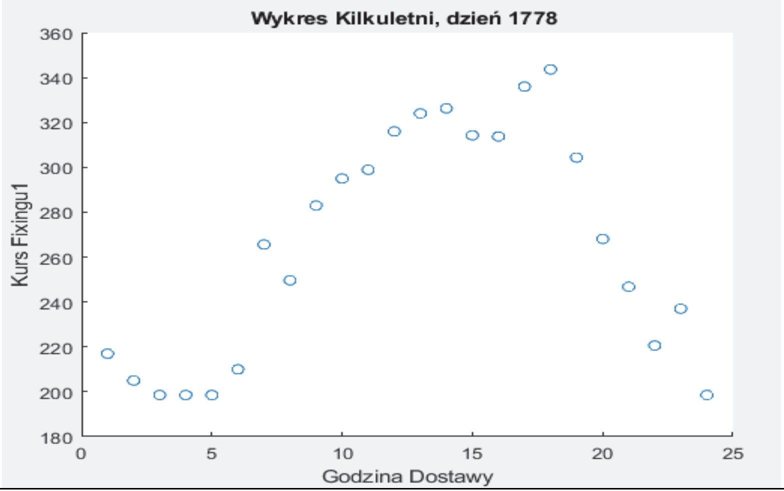
```
figure
scatter(DaneTygodniowe(:,1),DaneTygodniowe(:,9))
xlabel('Godzina Dostawy');
ylabel('Kurs Fixingu1');
title('Wykres Tygodniowy, dzień 7 ');
```

Listing 3. M-file for creating a scatter plot. Designations: scatter – chart type, weekly data. Source: Own elaboration in MATLAB and Simulink environment with the use of SMLT [5-7]

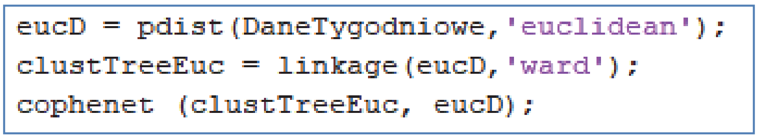
Subsequently, hierarchical clustering was performed using the Ward’s method. Initially, the Euclidean distance between different pairs of observations was determined using the pdist() function, which takes a data set and a user-specified distance metric as parameters. Then, a cluster tree called clustTreeEuc was created using a special function available in the MATLAB and Simulink SMLT library, in which the input data matrix was given as the first parameter, and the method for determining the distance between clusters as the second parameter. In addition, a cophenetic correlation coefficient was introduced to determine to what extent the resulting cluster tree faithfully reflects the differences between observations (Listing 4). The linkage() function shown here creates a cluster tree based on predetermined distances between pairs of objects. The result is a matrix called tree with 3 columns. The first two are related to nodes, which in the case under consideration are delivery times, and the last column contains the value of the distance between these nodes.

Table 5. Charts of the fixing 1 exchange rate for the last day of each analyzed data analysis period. Designations: Wykres tygodniowy - Weekly chart, Wykres miesięczny - Monthly chart, Wykres kwartalny - Quarterly chart, Wykres pólroczny - Half-yearly chart, Wykres kilkuletni - Several years chart, Godzina Dostawy – Delivery time, dzień - day. Source: Own elaboration in MATLAB and Simulink environment with the use of SMLT [5-7]

Description	Data analysis results
Graph of the fixing rate1 for day 7	
Graph of the fixing rate1 for day 31	
Graph of the fixing rate1 for day 91	

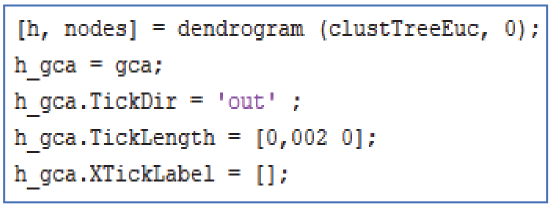
Graph of the fixing rate1 for day 182	
Graph of the fixing rate1 for day 366	
Graph of the fixing rate1 for day 1 778	

Thus, the dendrogram shows the hierarchy of the resulting clusters for the fixing 1 rate (Fig. 9), where the previously created cluster tree was used as the input data for the dendrogram. The X-axis shows the hours and the Y-axis shows the distances between different pairs of objects. Listing 5 shows an example of creating a dendrogram.



```
eucD = pdist(DaneTygodniowe,'euclidean');
clustTreeEuc = linkage(eucD,'ward');
cophenet (clustTreeEuc, eucD);
```

Listing 4. Create a cluster tree using Ward's method for the Euclidean distance. Source: Own elaboration in MATLAB and Simulink environment with the use of SMLT [5-7].



```
[h, nodes] = dendrogram (clustTreeEuc, 0);
h_gca = gca;
h_gca.TickDir = 'out' ;
h_gca.TickLength = [0,002 0];
h_gca.XTickLabel = [];
```

Listing 5. Visualization of the dendrogram with the appropriate parameters. Source: Own elaboration in MATLAB and Simulink environment with the use of SMLT [5-7].

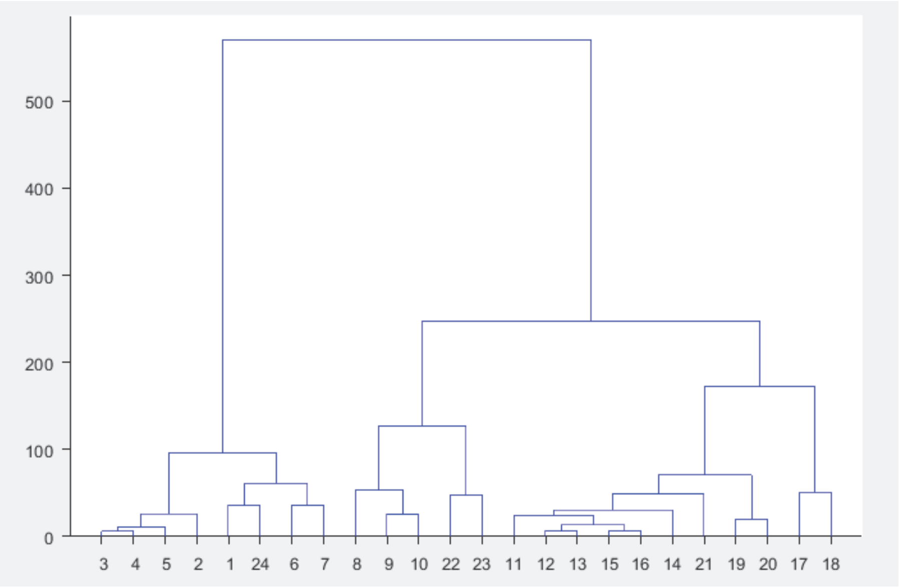
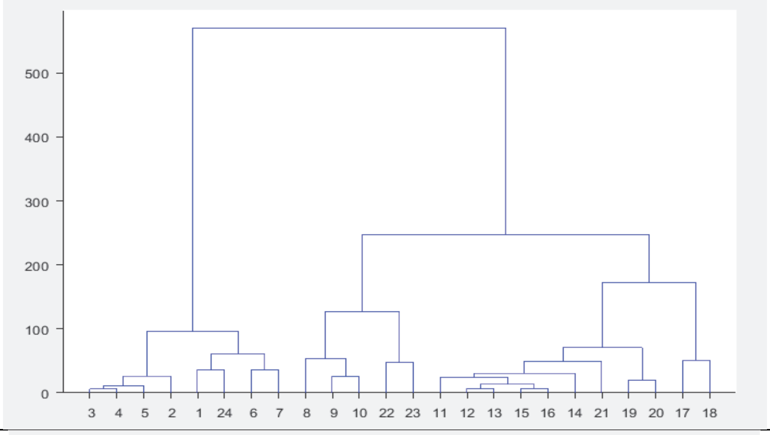
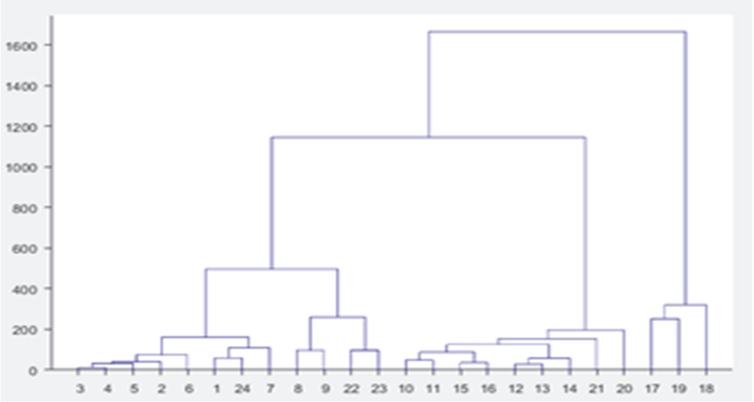
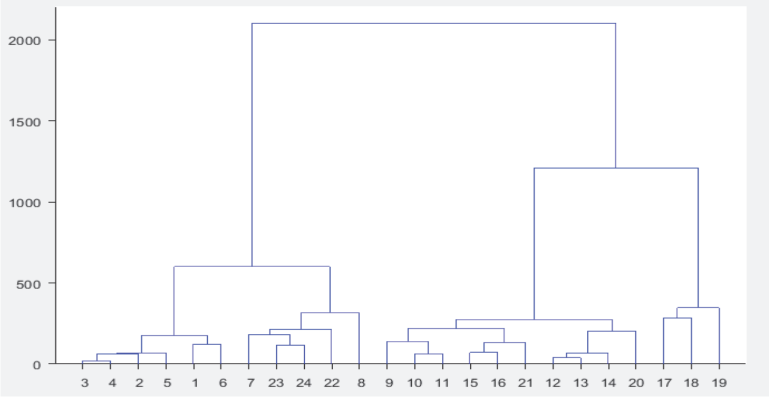
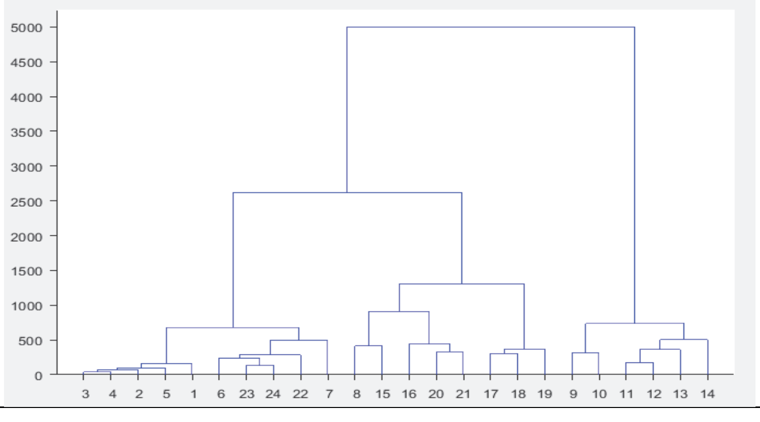
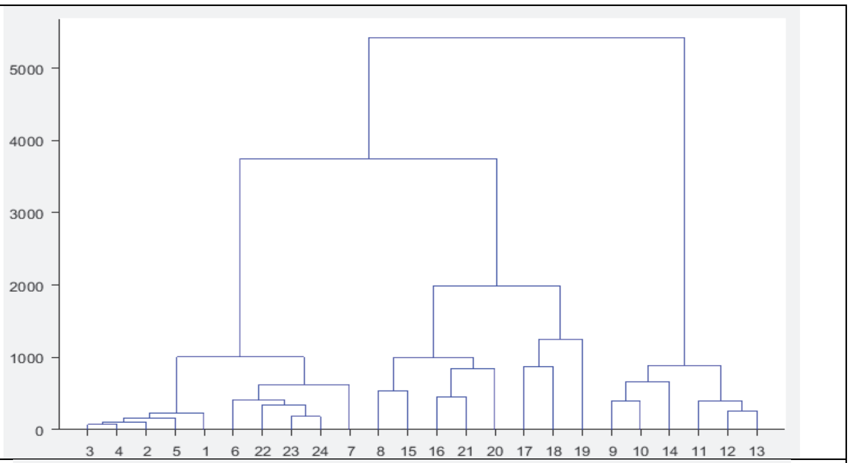
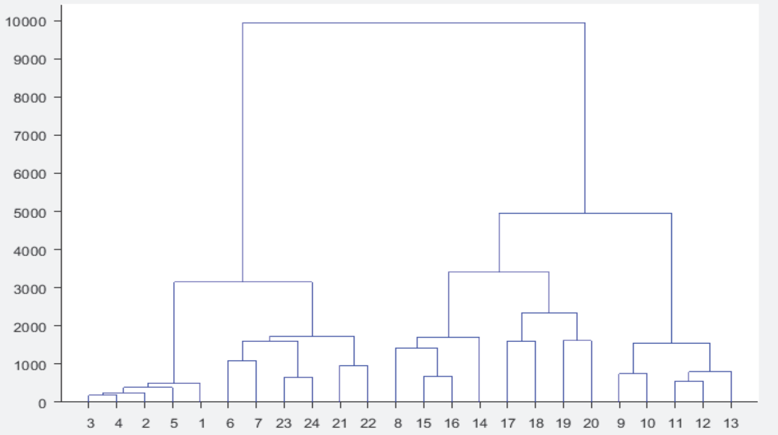


Figure 9. Dendrogram showing the hierarchy of clusters for the first week of the study period. Markings: Values on the Y axis - distances between pairs of objects, X axis - hours of energy delivery. Source: Own elaboration in MATLAB and Simulink environment with the use of SMLT [5-7]

The hierarchy of clusters of all time periods that were created in this experiment is presented in Table 6 in the appropriate order for the period of: week, month, quarter, half-year, year and 4 years. The dendrograms show many U-shaped lines that connect the points. The height of each U indicates the distance between two connected data points. Thanks to such results, it is easy to determine between which pairs there is the smallest distance, and between which the distance is the greatest.

Table 6. Dendrograms as cluster trees for various lengths of the analyzed data. Source: Own elaboration in MATLAB and Simulink environment with the use of SMLT [5-7].

Period length	Dendrogram as a cluster tree
week	
month	
quarter	
half year	

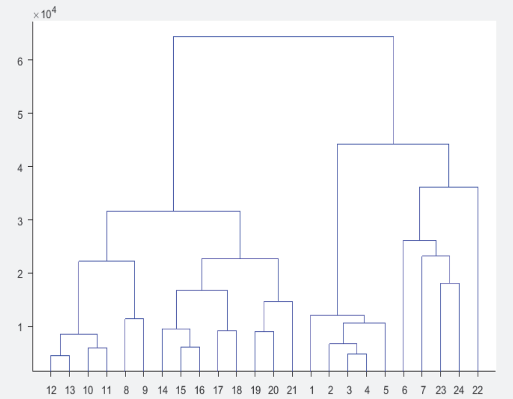
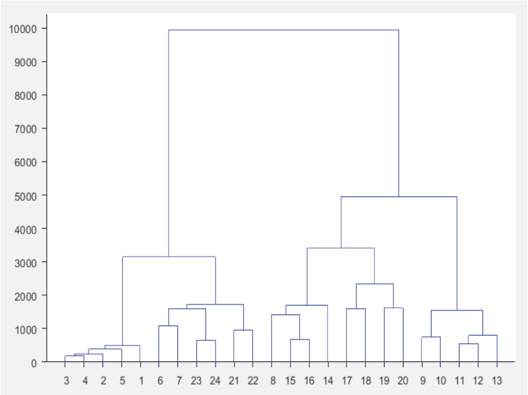
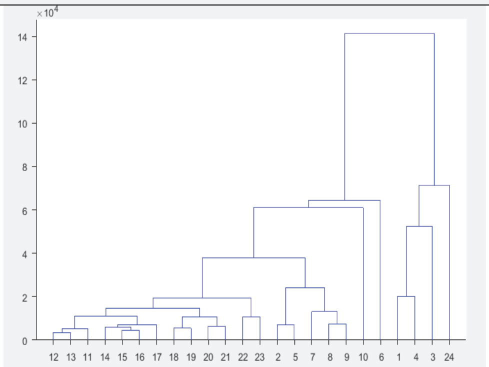
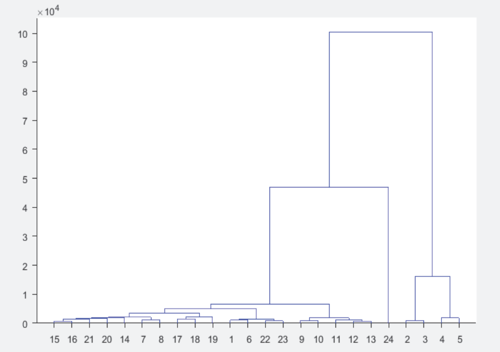
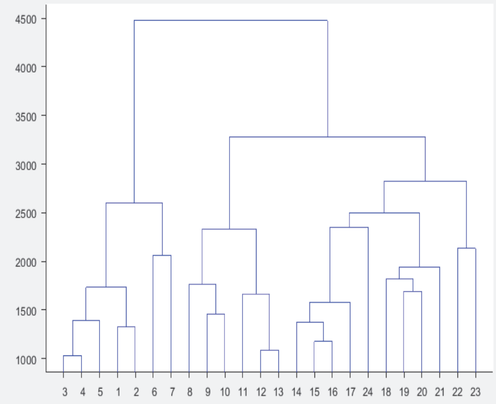
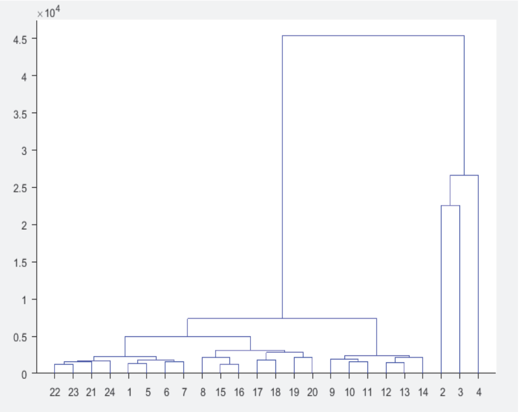
year	
4 years	

In addition, comparative studies were carried out, among others for a period of 4 years regarding fixing 1, fixing 2 and continuous quotations in terms of the exchange rate value and the volume of electricity (Fig. 4), the results of which are presented in table 7. Different distributions of dendrograms were recorded for individual hours, which was dictated by a change in the input data connecting pairs of objects.

The purpose of this experiment was to show data analysis using Ward's hierarchical method. Using different time periods, it was investigated, e.g. the value of the fixing rate1, including changes in the rate value in each hour on a specific day from a given period of time. It was noted, among others, that on days numbered: 7, 31, 91, 182, 366 and 1,778, the highest values were obtained at: 17, 14, 9, 11 and 1, 7 p.m. and 6 p.m., with the highest two-hour value on day 182, which was over PLN 900/MWh, and on other selected days the value was much lower, not exceeding the value of PLN 350/MWh.

Therefore, the analyzed statement shows that on day 182 the fixing 1 rate was definitely higher, especially between 9 and 14. At the next stage of the research, the course of the dendrograms obtained using the appropriate cluster tree was analyzed. Observation on dendrograms of the letter U with the height reflecting the distance between connected objects in specific hours indicated, among others, that the most similar objects occurred between 3 - 4, 12 - 13 and 15 - 16, and with the increase in the number of data in relation to the time period, the values that were most similar to each other were at similar times, with the largest U in the dendrograms related to the most distant objects, which were different with each time period. It is worth noting here that the range of values on the Y axis also increased its value, starting from a maximum of 500 and ending at 10 000.

Table 7. Dendrograms for various types of quotations conducted on the Day-Ahead Market of TGE S.A. Markings: X axis - time of a given distance value in terms of quotations of supplied and sold electricity, Y axis - distance between each pair of objects. Source: Own elaboration in MATLAB and Simulink environment with the use of SMLT [5-7].

X	Electricity volume	Electricity price
Fixing 1		
Fixing 2		
Continuous quotes		

It is also worth noting, among others, that the compilation of the hierarchy of clusters for electricity volumes and volume-weighted average electricity prices for the entire 4-year period made it possible to examine the difference between the fixing 1 and fixing 2 prices, as well as continuous quotations. In the first case, the objects were quite far from each other in relation to subsequent graphs, e.g. in the case of fixing 2 and continuous trading, these values were more similar to each other, hence the letters U were low and contained small distances, with most objects close to each other, but there were also elements distant from the cluster. In the case of comparing the volumes from the period of 4 years, a similar situation can be observed, because elements with a smaller distance occurred in the case of fixing 2, similarly to the comparison of electricity prices. It also turned out that the largest distances concerned the volume of continuous trading.

4. Conclusion

A cluster analysis was designed and carried out using data listed on the TGE S.A. Day-Ahead Market. Both the example of the hierarchical method (Ward's method) and the non-hierarchical method (k-means methods) were used, obtaining appropriate results in the field of data clustering.

The research used data on the volume and the volume-weighted average electricity price (price) quoted in the form of fixing 1, fixing 2 and continuous trading in the following hours of the day.

In the case of the Ward's method (hierarchical method), the fixing volume 2 had the most favorable values located close to each other, with the greatest distance between the continuous trading volume, which confirmed the fact that the values of the quoted electricity volume in this case differed significantly from each other.

However, in the case of the k-means method (non-hierarchical method), the 3D chart illustrated what clusters look like and how, for example, fixing volume 2 values were matched to the appropriate groups. Attention was also paid to the situation in which the algorithm increases the number of clusters and how individual silhouette graphs are shaped then.

It is also worth noting, among others, that an in-depth interpretation of the obtained research results is also possible, which has not been included in this work due to the need for further development of the article. In addition, such results would be of little interest to readers of a journal in the discipline of computer science, hence the authors intend to publish them in a journal in the discipline of automatics, electronics and electrical engineering.

References

1. Flasiński M.: Wstęp do sztucznej inteligencji (In Polish), In English: Introduction to Artificial Intelligence, PWN, Warszawa 2011, pages 332.

2. Kłopotek M., Tchórzewski J.: The concept of discoveries in evolving neural net, Advances in Soft Computing, IPI PAN, No. 17, Warszawa 2002, pp. 165-174.

3. Koronacki J., Ćwik J.: Statystyczne systemy uczące się (In Polish), In English: Statistical learning systems, EXIT, Warszawa 2005, pages 328.

4. Larose D.: Odkrywanie wiedzy z danych. Wprowadzenie do eksploracji danych. PWN. Warszawa 2006, pages 228.

5. Longota B.: Środowisko MATLAB-a oraz Statistics and Machine Learning Toolbox-a do analizy danych Rynku Dnia Następnego TGE S.A. (In Polish), In English: MATLAB environment and Statistics and Machine Learning Toolbox for the analysis of TGE S.A. Day-Ahead Market data, Master's thesis supervised by PhD Hab. Eng. Jerzy Tchórzewski, Univ. Prof., written at the Institute of Computer Science at the Faculty of Exact and Natural Sciences, UPH in Siedlce, Siedlce 2021, pages 91.

6. MathWorks, Statistics and Machine Learning Toolbox User's Guide, 2020, pages 7 984.

7. MATLAB & Simulink, MathWork, https://www.mathworks.com [access: 2021-04-16].

8. Mielczarski W.: Rynki energii elektrycznej. Wybrane aspekty techniczne i ekonomiczne (in Polish), In English: Electricity markets. Selected technical and economic aspects. ARE S.A., Warszawa 2000, pages 321.

9. Migdał-Najman K., Najman K.: Analiza porównawcza wybranych metod analizy skupień w grupowaniu jednostek o złożonej strukturze grupowej (In Polish), In English: A Comparative Analysis of Selected Methods of Cluster Analysis in the Grouping Units with a Complex Group Structure, Zarządzanie i Finanse, 2013, R. 11, No 3, part 2, pp. 179-194.

10. Osowski S., Metody i narzędzia eksploracji danych (In Polish), In English: Data mining methods and tools. Wyd. BTC, Legionowo 2017, pages 390.

11. Ruciński D.: The Influence of the Artificial Neural Network type on the quality of learning on the Day-Ahead Market model at Polish Electricity Exchange join-stock company, Studia Informatica. Systems and Information Technology, Vol. 1-2(23)2019,

pp. 77-94.

12. Ruciński D., The neural modelling in chosen task of Electric Power Stock Market. Studia Informatica. Systems and Information Technology. No. 21, Vol. 1 No. 21/2017, pp. 63-83.

13. Tadeusiewicz R., Szaleniec M.: Leksykon sieci neuronowych (In Polish), In English: Lexicon on Neural Networks, Wydawca Projekt Nauka, 2015, pages 134.

14. Tchórzewski J.: Metody sztucznej inteligencji i informatyki kwantowej w ujęciu teorii sterowania i systemów (In Polish), In English: Methods of artificial intelligence and quantum computing in terms of control theory and systems, Wydawnictwo Naukowe UPH w Siedlcach, Siedlce 2021, pages 343.

15. Tchórzewski J., Jezierski J.: Cluster analysis as a preliminary problem in neural modelling of the Polish Power Exchange, Information Systems in Management, Vol. 8, No. 1/2019, pp. 69-81.

16. Tchórzewski J., Kłopotek M.: A Case Study in Neural Network Evolution, Prace Naukowe Instytutu Podstaw Informatyki PAN, No. 943, IPI PAN, Warszawa 2002, pp. 1-12.

17. Tchórzewski J., Kłopotek M.: The Concept of Making Discoveries in Evolving Neural Net, Intelligent Information Systems 2002, Physica-Verlag HD, pp. 165-174.

18. Tchórzewski J., Kłopotek M., Kujawiak M.: Studium porównawcze metod prowadzenia odkryć (in Polish), In English: A comparative study of discovery methods. Studia Informatica. Systems and Information Technology. No. 1(4)2004, pp. 105-122.

19. TGE S.A., https://tge.pl [dostęp: 06.04.2021].

20. Trajer J., Janaszek-Mańkowska M., Mańkowski D. R., Komputerowa analiza danych w badaniach naukowych (Eng. Computer data analysis in scientific research), Wyd. SGGW, Warszawa 2016, 131 pages.

21. Wierzchoń S., Kłopotek M.: Algorithms of cluster analysis, Monograph Series Information Technologies: Research and Their Interdisciplinary Applications, No. 3, Institute of Computer Science, PAN, Warsaw 2015, 308 pages.

22. Zhan J., Matwin S., Chang L.: A Multi-Party Scheme for Privacy-Preserving Clustering Studia Informatica. Systems and Information Technology. No 1-2(7)2006, pp. 217-232.