Energy cluster analysis based on consumption data in different weather condition

Dimitrios Th. Kazolis¹*, Panagiotis G. Kogias¹, and Nikolaos I. Roumeliotis²

¹International Hellenic University, Department of Physics, Agios Loukas, 65403, Kavala, Greece
²HENDO Public’s Electricity Company, Kavala, Greece

Abstract. The main aim of this effort is the discovery of knowledge from data, concerning consumption of electric energy, during the year 2022, based on unattended learning methods. These data were collected from the Public Electricity Company of Kavala and the methods used are, at first the Factor analysis and second the K-means clustering algorithm. The overhead methodologies are realized by the use of Statistica Data Miner software.

1 Introduction

The purpose of all Electricity Companies is to provide stable and sufficient energy to all of their customers at any time of the year. A decisive factor in achieving this is not only the production of energy, but the prediction of future consumption and the identification of the factors that affect it. This has as result to constructive energy modeling and modifications that are essential, since nowadays; there is an increase in the energy needs of man. For this reason, there has been a growing interest in use of various techniques for extraction of knowledge from databases related to energy consumption. [1, 2, 3].

The pursuit of this knowledge is, to reveal and create conclusions, which are useful for the energy consumption process [4], [5]. Thus, there are articles aiming to the extraction of knowledge, for example in the energy consumption of buildings [6], for the characterization of electricity customers [7] or for the Analysis to Power Grid System [8]. The production of the electricity from renewable energy sources as wind farms and photovoltaic systems is also appropriate to be study [9, 10].

Generally, many data mining techniques have been implemented in the development of predicting and solving the various problems of energy consumption. [11, 12].

As a continuation of previous research, the current effort aims to contribute to the development of the specific scientific field. So, the rationale behind this work is, to exploit information that may arise from large databases.

This procedure applies in practice all key stages of the extracting knowledge process from large databases [13, 14].

Thus, a main file was created, which had 17521 lines and 20 columns. Initially, these data were changed to a numeric form. Then, they fed the process of factor analysis. This resulted in three factors and after that, the total sample led the clustering algorithm which divided it

* Corresponding author: author@email.org

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).
into 60 groups. Those groups had a logical relevance in their categorization and proved the validity of the procedures used.

As a conclusion this article proves that, the used methods, can cooperate together perfectly in constructing understandable relationships, during the process of knowledge discovery.

2 Data and methods

2.1 Data selection

Initially, data from eight energy distribution points in the city of Kavala were recorded, with a half an hour frequency, for the entire year 2022. So, an initial table was created with 17521 lines. Then, this table combined with more variables such as, the month, the time, the days of the week, the minimum and the maximum temperature of the day, the average temperature of the month, the average humidity, the wind, the rain height, the average pressure, the sunny and rainy days and the sun intensity. [15]

As a result of the above, a final table with a total of 17521 lines and 32 columns was constructed.

Although many formats could be imported directly into Statistica, it was considered appropriate to change the format of data as many of them had a text formatting. So they were converted into Byte and Numeric type, for the better operation of the clustering algorithm.

2.2 Factor analysis

As mentioned, the result of the above actions was to create a table of 17521 lines and 32 columns. This makes a total of 543151 entries, a fact which creates great difficulty in the identification of relationships and in addition, has a negative effect on the clustering algorithm to be used. So, a multivariate exploratory technique as factor analysis was used in order to reduce the number of variables available [14]. The default minimum eigenvalues was left to be 1, and by the use of the Scree-plot and checking the Kaiser criterion, the number of the chosen factors was three. This is shown in Figure 1.

![Scree plot](image)

**Fig. 1.** Scree plot.

Furthermore a part of the correlation Matrix is depicted in the following Figure 2.
For a factor rotation method, the Varimax raw method was used, and in the final stage of the process the loads of the factors were created. Those characterize the point that every factor interprets every variable. If a load is greater than 0.3 then it can be considered significant, but the larger it is, the more the factor interpret the corresponding variable.

So, the interpretation of the first factor correlates the time with the consumptions. As it can be observed here, there is an abnormality at the energy distribution point P230. This is because initially, the distribution point P230 was used to distribute electric energy to the industrial and not the residential area of the city of Kavala.

The second factor interprets the change of temperature and moreover, the effect of the sun on the electricity consumptions.

Finally, the third factor interprets the variables connected with the days of the week. Here, a decline can be observed during Sundays and on the other hand, during Thursdays and Fridays the consumptions are increasing.

After the factor analysis, the 32 columns of the original file were replaced by 8 contributing to the reduction of data.

The above file is now ready to be led to the process of data mining by the use of generalized cluster analysis.

### 2.3 Cluster process

The main purpose of this step is to find clusters with common features, which can be interpreted by a human.

For this purpose, the Generalized k-Means Cluster Analysis is used, as, it can easily handle categorical variables. In these, the maximum frequency category grows into the center of the corresponding cluster, and all distances can be zero and one.

In this process, before the algorithm is applied, the number of clusters and the number of iterations is defined by the user. Here, the number of clusters is set to be 60 and the number of repetitions, defining group centers, is chosen to be 600. The selection of this big number was made just in order to avoid interrupting the process before it is properly completed.

Moreover, for the initial cluster centers the k observations were selected randomly and the Manhattan method is defined. This method determines the distances of non-uniformity.
and is the average of the modification among the elements. This is done in order to prevent the analysis from being influenced from variables with different gradient.

After all the above manipulations, a final record was created. Here, every specific hour of every month was grouped into a cluster and also was presented its distance from the cluster’s center. The histogram in Figure 3 is showing the number of members of each cluster.

![Histogram of final classification](image)

**Fig. 3.** Histogram of final classification.

At the present the Cluster Analysis process has been finalized. As a continuation, the examination of the similarity of each clusters’ member characteristics must validate the knowledge discovery effort.

However, it should be emphasized that one of the characteristics of unsupervised learning and the K-Means algorithm is that it gives similar but not exactly the same results whenever the process occurs from the opening. And this is because the selection of the initial K centers is random.

### 3 Validity control

As mentioned above, the number of clusters was set to be 60, a number which can help with the interpretation of results. So, in the specific execution, some of the clusters will be examined in order to justify the reliability of this knowledge discovery process.

To begin with, cluster 2, is selected. As it can be seen in the following Figure 4, this cluster includes days of month May except Sundays. This group consists of high energy demand hours, from 8-9 o’clock in the morning until late in the evening.
Next the cluster 38, in Figure 5, was selected. Here are days of February but the group consists of low energy demand hours, from midnight up to early in the morning.

The following, cluster 43 in Figure 6, has the same characteristics as the previous cluster, with low energy demand hours, from midnight up to early in the morning. The clustering algorithm differentiated them because of the month August.
Fig. 6. The number 43 cluster.

A similarity can be observed in the following cluster 57, in Figure 7. As with the previous groups this consists of the low energy demand hours, from midnight up to early in the morning of November days. Furthermore, the low consumption can be observed here until midday during Sundays.

Fig. 7. The number 43 cluster.

Similarly the cluster 9 in Figure 8, consists of days of April. There is a low energy demand period from midnight up to 7-7:30 o’clock in the morning. Moreover during Sundays, this low energy demand period, reaches the early afternoon hours.
Fig. 8. The number 9 cluster.

Similar examination can be performed in all of the created clusters. In general it can be noted that members of each cluster, have many common features, while on the other hand, the clusters have several differences among them, so they can be separated by the algorithm.

In general, it can be observed that, all of the clusters contain time periods with similarities in the energy consumption.

4 Conclusions

The conclusions from all of the above procedure can be derived from each step of it.

Initially it can be concluded that, a very effective statistical method for multivariate exploration, is factor analysis. With it, the large dimensional data, which must be processed, can be transformed into a relative smaller group of variables by replacing many of them, with factors. As a result the difficulties which could be raised from the very large dimensional data can be erased and moreover, the created factors include the essential meaning of the replaced variables.

It can be inferred that, in such cases, Factor Analysis is considered to be necessary, as it is a reliable and effective method.

Moreover, the above procedure has also included the K-Means clustering algorithm. This is a very active method of classification. It can easily handle categorical variables especially in cases without learning examples.

To conclude with, this paper is one more proof that these two methods can cooperate together perfectly. They are following their theoretical background in a predictable way and they can be applied for further knowledge discovery research.

So, for further research, one could investigate the same data model with the method of V-fold cross validation, or to discover the effectiveness of different data mining algorithms in this common data model.

References

3. Y. Zhang, J. Liu, Sensors, 22 (10), 3704 (2022)
6. A. Abdelaziz, V. Santos, M. Dias, Energies, 14 (22), 7810 (2021)
13. W. Liu, J. Zhao, D. Wang, WIREs Data mining and Knowledge Discovery, 11(4), e1406 (2021)