# Analysing Typology of European Energy Structure by Using Clustering Algorithms : Proceedings of the 12th International Conference on Informatics in Economy Education, Research & Business Technologies, 25-28 April 2013

Furtuna, Titus Felix; Reveiu, Adriana; Kanala, Roman

# ANALYSING TYPOLOGY OF EUROPEAN ENERGY STRUCTURE BY USING CLUSTERING ALGORITHMS

**Titus Felix FURTUNĂ**
The Bucharest University of Economic Studies
titus@ase.ro
**Adriana REVEIU**
The Bucharest University of Economic Studies
reveiua@ase.ro
**Roman KANALA**
Université de Genève
roman.kanala@unige.ch

**Abstract.** *The paper proposes to identify the energetic profile for the European countries by using of the nine indicators concerning the energetic sector. The European countries were grouped using the clustering and the feature extraction techniques. The analysis results were visualized on maps in a geographic information system (GIS). The used algorithms for grouping are bisecting KMeans and Non Negative Matrix Factorization. These are provided by the Java Data Mining API framework in the Oracle implementation. The countries groups are viewed on maps that are built in ArcGIS by ESRI. The proposed IT solution in this paper mixed the various solutions in a unitary manner and contained of the specialized elements for programming Java applications. It's about programming data mining applications using Oracle Data Mining Java API and the programming elements through ArcObjects SDK for Java. This two frameworks (Data Mining Java API and ArcObjects SDK) are used because of the Java background and on the other hand, Oracle offers the necessary support for the spatial database and the Data Mining engine.*

**Keywords**: Cluster Analysis, Feature Extraction, GIS, Symbologies
**JEL classification:**
C38 - Classification Methods; Cluster Analysis; Factor Models
C8 - Data Collection and Data Estimation Methodology; Computer Programs

## 1. Introduction

In 2003 the European Commission launched the Intelligent Energy-Europe (IEE) [6] in order to encourage the development of sustainable activities in energy. The program is part of a larger complex of actions undertaken to create a smart energy future for all of us. EU aims to increase energy efficiency and use of renewable energy sources. For 2020 are set some targets such as a 20% reduction in emissions of greenhouse gases, improving energy efficiency by 20%, increasing to 20% of the share of renewable sources in the energy consumption and so on.

Achieving the perspective objectives will inevitably lead to the implementation of the similar energy policy in the EU countries. Therefore structural differences among EU countries in the field of energy will gradually fade. Currently there are countries that are much closer to the proposed target than the others. This study highlights the structural differences and similarities that currently existing among the EU countries.

Energy indicators used in analyze include energy dependency, energy intensity, primary energy production, final total energy consumption, installed capacity of electricity generation plants, renewable energy sources, energy efficiency, electricity prices in households and electricity prices in industry.

## 2. Unsupervised classification algorithms

The algorithmic support for unsupervised classification is ensured by the Oracle Data Mining Java API framework [7,8]. Oracle Data Mining Java API is build over the Java Data Mining API [3] being an implementation of vendor type. The usage functions are: Clustering and Feature Extraction with KMeans algorithms on clustering side and Non Negative Matrix Factorization for Feature Extraction.

### 2.1 Feature extraction

Feature extraction is a process of information reduction by building new attributes most significant in terms of information. New attributes are linear combinations of the original attributes. Type feature extraction algorithms are used in data compression, pattern recognition and generally in supervised learning approaches. The algorithm implemented in Oracle Data Mining is Non-negative Matrix Factorization (NMF) [4]. By combining the attributes NMF generates meaningful patterns in the dataset description. According NMF, input matrix, X, can be approximated by a product of two lower rank matrices, *W* and *H*. NMF uses an iterative procedure to change the initial values of matrices W and H so that the product to get close of X. The procedure ends when it achieved a margin of error or the specified number of iterations.

***Problem definition***. Given a nonnegative matrix $X \in R^{m \times n}$, where *m* is the number of cases and *n* is the number of features and a positive integer number $k<\{m, n\}$ representing the number of essential features (patterns), to find the positive defined matrix $W \in R^{m \times k}$ and $H \in R^{k \times n}$ so that to minimize the function: $f(W,H) = \frac{1}{2}\|X - W \cdot H\|^2$.

The $W \cdot H$ product is the non-negative factorization of the *X* matrix. The method extracts the essential features of data as column vectors in the *W* matrix, features that can be used in classification. These vectors are expression of patterns by which can be identified items from the studied collection. By not allowing of the negative values in the *W* and *H* matrix, NMF creates non-subtractive combinations of features to form a pattern (typically).

***Numerical solutions***. There was given different solutions to solve the problem: Patero and Tapper (1994), Lee și Seung (2001), Berry-Plemmons (2004). W represents the matrix of essential feature. *H* is the matrix of factors. The column vectors of the *W* matrix are nor orthogonal. Among essential features can be exist an informational redundancy in comparison with the principal components from the principal components analysis.

The link among *X*, *W* and *H* is $X = W \cdot H$.

If we multiply this relation at the left side with $(W^T W)^{-1} W^T$ obtain: $H = \frac{W^T X}{W^T W}$, where, $W^T$ is the transpose of the *W* matrix. If we multiply the same relation at the right side with $H^T (H \cdot H^T)^{-1}$, obtain: W = $\frac{X \cdot H^T}{H \cdot H^T}$. Taking into consideration these relations a possible implementation [4] could be formalized in this way:

Procedure NMF(*X*,*W*,*H*,*m*,*n*,*k*,*maxIter*)
W = random(m,k)
H = random(k,n)
for i=1,maxIter

$$h_{ij} = h_{ij}\frac{\left(W^T X\right)_{ij}}{\left(W^T WH\right)_{ij}}, i = 1,k \quad j = 1,n$$
$$w_{li} = w_{li}\frac{\left(XH^T\right)_{li}}{\left(WHH^T\right)_{li}}, l = 1,m \quad i = 1,k$$

endfor
end

### 2.2 Bisecting KMeans

Oracle implements an enhanced version of Bisecting K-Means algorithm, which is closer to the hierarchical algorithms class [2]. K-Means Oracle has the following features:
- It is a hierarchic model of the top-down type that consists in the clusters progressive splitting according to the K-Means classical algorithm [5] and the progressive improving solution;
- At each step only one cluster is splitting (node cluster) according to the following criterion: the cluster variance and the number of entities from node;

It provides a probabilistic score of the elements belonging to a cluster;
- The algorithm computes for each cluster the following indicators: confidence level, support level, the centroid, the rule, the parent cluster, the ancestors and various statistics. The following algorithm is the bisecting version of the general K-means algorithm.

```
Procedure BisectingKMeans(X,n,m,k)
X - input table, n, m - numbers of instances and attributes, k - desired number of clusters
List[] L // List of the clusters
call Init(L,n) // The root cluster is the entire collectivity
for i=1,k-1
        call Select(X,n,m,L;M) // Selecting the cluster according of splitting principle
        call InitCenters(X,m,n,M;gl, gr)
        do // Using of KMeans classic algorithm for identifying children
                call Divide(X,n,m,M,gl,gr;ML,MR)
                call Centers(X,n,m,ML,MR;cl,cr)
                call Distance(gl,gr,cl,cr;dist)
        while dist>eps // Adding children
        call Add(L,ML)
        call Add(L,MR)
endfor
call Write(L)
end
```

### 3. Energy structure typology

The selected indicators present national data for the 27 EU Member States, the EFTA and candidate countries [6]. Data are generally available for the 2010. The main data sources are: reporting under Regulation (EC) No 1099/2008 of the European Parliament and of the Council on Energy Statistics and Directive 2008/92/EC concerning transparency of gas and electricity prices. Indicators cover aspects of production, consumption, prices, consumption and production structure and energy efficiency. Data are grouped in the Oracle database in three tables: Energy, Feg, Iceg. Indicators contain in the Energy table are the following: ED - Energy dependency - Energy dependency shows the extent to which a country relies upon imports in order to meet its energy needs. It is measured in percents; EI - Energy intensity -

Energy intensity gives an indication of the effectiveness with which energy is being used to produce added value. It is measured in kilograms of oil equivalent - kgoe; PEP - Primary energy production, by fuel, per capita (kilotons of oil equivalent to one million habitants); FEC - Final energy consumption per capita (megatons to one million habitants); ICEG - Installed capacity of electricity generation plants per capita (MW to one million habitants); RES - Share of renewable energy in gross final energy consumption (%); EE - Energy efficiency - Gross inland consumption per capita (toe per capita); EPH, EPI - Electricity prices in households and industry (EUR/100 KWh).
In the Feg table is the structure of final energy consumption, by sector (industry, transport, residential, services) expressed as a percentage. In the Iceg table is the structure of installed capacity of electricity generation plants, by type (thermal, nuclear, hydro and other) also expressed as a percentage.
Figure 1 presents a division into 5 groups of countries by KMeans algorithm. We can observe a territorial distribution groups with a cluster of countries in Eastern Europe, a cluster with the western countries, a cluster of countries in southern Europe (Italy, Cyprus, Malta), a cluster with Norway and a cluster with Sweden, Finland and Luxembourg. In Table 1 the centroid values for each indicator are presented. For instance the first cluster contains countries having decrease energy efficiency, with a low production yield and lower per capita consumption. The map is completed with information concerning the final energy consumption on types of activities.

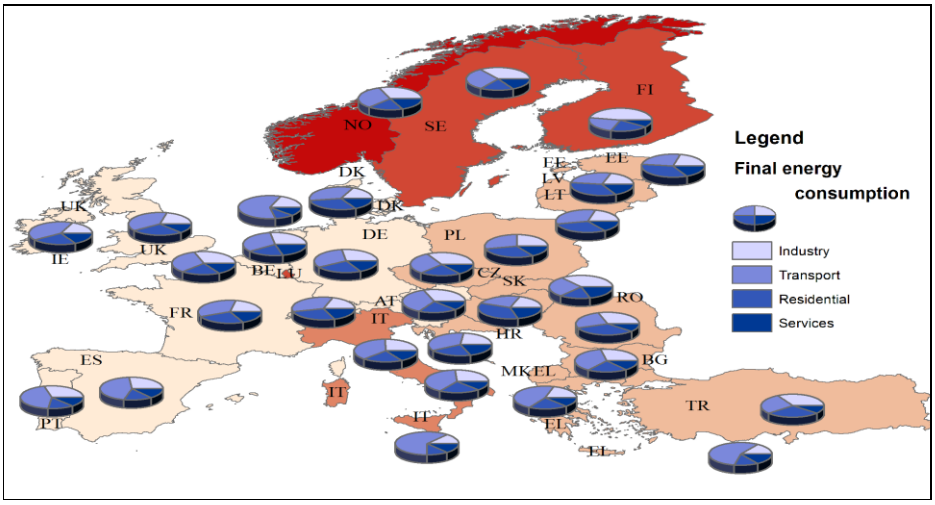


*Figure 1. Unsupervised classification by Oracle KMeans*

*Table 1. Centroid values for clusters*

| Attribute | Centroid Values | | | | |
|---|---|---|---|---|---|
| | 1 (East) | 2 (Norway) | 3(SE, FI, LU) | 4 (South) | 5 (West) |
| RES | 16.0031 | 58.53 | 27.6333 | 5.1 | 14.0558 |
| PEP | 1369.1207 | 24837.39 | 2309.2283 | 199.3437 | 1864.825 |
| ICEG | 1140.7825 | 5307.3 | 3446.4423 | 1639.4233 | 1945.46 |
| FEG | 1.6146 | 3.9997 | 5.108 | 1.8089 | 2.6334 |
| EPI | 9.5894 | 9.1 | 8.6 | 18.3167 | 10.5551 |
| EPH | 12.9649 | 18.7 | 16.9 | 20.6333 | 19.6619 |
| EI | 457.4615 | 158 | 182.6667 | 176 | 156.6667 |
| EE | 2.5246 | 6.9 | 7.1133 | 2.83 | 3.8175 |
| ED | 47.0154 | -276.89 | 60.4667 | 95.1667 | 52.475 |

Figure 2 presents a distribution on groups according with the Non Negative Matrix Factorization algorithm. There are identified and highlighted on the map four patterns. The map is completed with information concerning the types of plants used in the production of electricity.

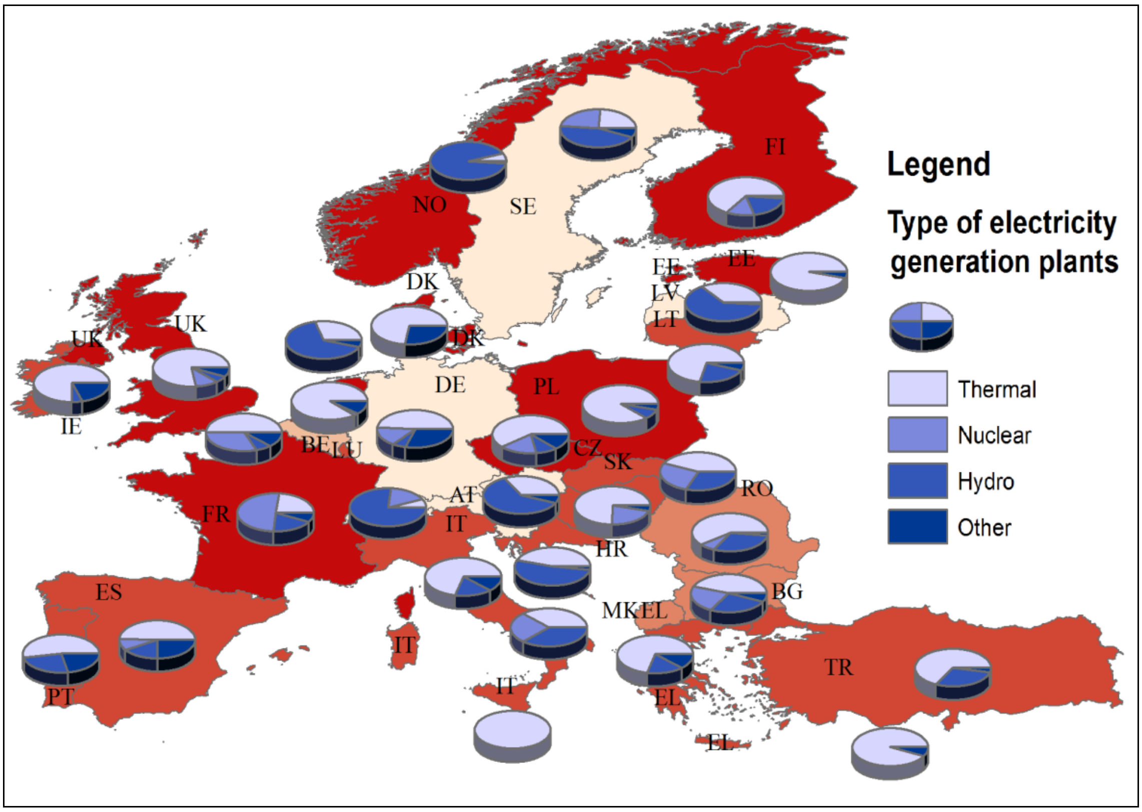


*Figure 2. Unsupervised classification by NMF*

To view information on the map we used a combination of symbologies[1]: filling the polygons with a specific color mixed with pie chart drawing on spatial data.

## 4. Conclusions

Constraints concern nonrenewable resources, environment, urbanization and population growth require diversification of activity in energy field, orientation toward alternative resource, clean and renewable, implement of better technology to improve energy efficiency. European countries are in different stages concerning the creating of an energy sector that to be suitable according the future demands. In this article the structural typologies are highlighted for European space, showing the differences among countries. Cluster analysis can show where and what to do to achieve the established objectives.

**Acknowledgment**

This work was supported by the Swiss Enlargement Contribution in the framework of the Romanian-Swiss Research Program (Grant IZERZ0_142217).

**References**

[1] Dârdală, M., Reveiu, A., *Soluţie software de reprezentare a datelor statistice în GIS, prin combinarea simbologiilor*, Revista Română de Interacţiune Om-Calculator, vol 4, nr 1, Editura MATRIX ROM, ISSN 1843-4460, Bucureşti, 2011

[2] Han, J., Kamber, M., *Data Mining: Concepts and Techniques*, Elsevier, 2006

[3] Hornick, M.,F., Marcade, E., Venkayala, S., *Java Data Mining. Strategy, Standard, and Practice*, Elsevier, Morgan Kaufmann Publishers, 2007

[4] Berry, M.W., Browne, M, Langville, A.N., Pauca, V.P, Plemmons, R. J., *Algorithms and applications for approximate nonnegative matrix factorization*, Computational Statistics & Data Analysis, Volume 52, Issue 1, 15 September 2007, Pages 155-173

[5] MacQueen, J. B., *Some Methods for classification and Analysis of Multivariate Observations*, Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, pp. 281–297, 1967

[6] Eurostat pocketbook, *Energy, transport and environment indicators*, 2012 Edition, http://ec.europa.eu/eurostat

[7] http://docs.oracle.com/cd/B28359_01/datamine.111/b28132/toc.htm

[8] http://www.oracle.com/pls/db111/portal.portal_db?selected=6&frame=#data_mining