Analysis of correlation relation between pollution and petroleum consumption using algorithms (spectral clustering and K-mean)

by

Ashraf Rafa Mousoud Mohamed

Master degree, Electrical and Computer Engineering, 2018

Submitted to the Graduate Faculty of

Science in partial fulfillment

of the requirements for the degree of

Master of Electrical and Computer Engineering

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

Doç. Dr. Oğuz BAYAT

________________________________ ________________________________

Co-Supervisor Supervisor

Examining Committee Members (first name belongs to the chairperson of the jury and the second name belongs to supervisor)

Doç. Dr. Oğuz BAYAT (Committee Member) _______________________

Yrd. Doç. Dr. Çağatay AYDIN (Committee Member) _______________________

Yrd. Doç. Dr. Adil Deniz DURU (Committee Member) _______________________

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

____

Yrd. Doç. Dr. Çagatay Aydin

Head of Department

____

Approval of [Institution] ____/____/____

Doç. Dr. Oğuz BAYAT

Director

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Ashraf Rafa Mousoud Mohammed

[Signature]

DEDICATION

To my father and mother

And to my wife thank you for helping me

And everyone who encouraged me to go to the sea of knowledge

And thanks to all those who support me

ACKNOWLEDGEMENTS

Thank you may be a few to you: Doç. Dr***. Oğuz BAYAT*** AND .Dr. ***Hüseyin Afşer***. To support me and help me to complete my research.

We also thank the members of the Committee:

- Yrd. Doç. Dr. ***Çağatay AYDIN***
- Yrd. Doç. Dr. ***Adil Deniz DURU***

Thank you to all university professors I learned from them at the university.

I must express my gratitude to my friends and colleagues for any support me.

ABSTRACT

Analysis of correlation relation between pollution and petroleum consumption using algorithms (spectral clustering & K-mean)

[Author’s Mohammed, Ashraf],

M.S./ Electrical and Computer Engineering, Istanbul Altınbaş University

Supervisor: Doç. Dr. Oğuz BAYAT

Co-/

Date: 8 / 2017

Clustering can be considered the most significant unsupervised learning methods that reveal similar behaviors (sets) on large sets of data.

Clustering is the process of organizing objects into groups that are similar in some way to their members.

Spectral clustering data points as nodes of a connected graph and clusters are found by partitioning this graph, based on its spectral decomposition, into sub graphs.

K - means assembly: Divide objects into k groups so that some measurements are minimized relative to the middle points of the clusters.

In this thesis, we have applied both of algorithms (spectral clustering and k-mean) on data from the life process and these data are statistical data. The field of data is the quantities of pollution and petroleum consumption. The spectral clustering algorithm was applied several times and k-mean was applied several times and the best result was compared and the best classification was also obtained.

Keywords: Clustering ,Spectral Clustering ,K-mean ,pollution, petroleum.

Table of content

LIST OF ABBREVIATIONS

Avr	average of values.
Str	standard deviation.
S.C.A	Spectral clustering algorithm.
EEA	European Economic Association.

1. INTRODUCTION

Constant advance in science and technology makes collection of data and storage much easier and very inexpensive than ever before. This led to form large data sets in science and government Industry, which should be processed or sorted for useful information.

For example, if we consider the results generated by a search engine for a specific query, the user has to clear through the long lists and find them the solution is required. But this task can be very difficult for the user if there is Millions of WebPages have been shown as a workaround for some queries. Etc. and so on Partition techniques can be very useful in a closely related compilation.

Specific query solutions and display results in the form to set so that even documents that have no relationship to each other can be uninstalled without.

That behind compiling any set of data is to find the authentic Structure in the data, interpret this structure as a set of groups, Data objects within each group must appear very high of the similarity known as the similarity within the mass, while the similarity.

1.1 CLUSTERING

Clustering is a Data collection is a way to make similar objects in groups Thus, same objects are in the same group and different objects are also in different groups Data clustering is considered as an unsupervised learning technique in which objects are grouped in unknown predefined clusters. On the contrary classification is a supervised education in the objects are set to predefined classes clusters.

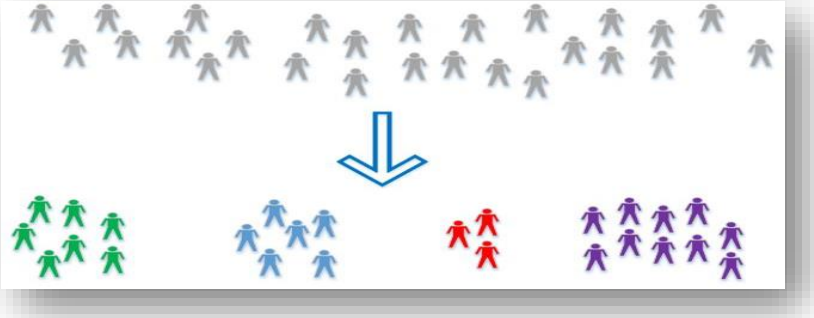


Figure. 1.1 clustering

1.2 Basic Concepts Of Clustering

The problem of data clustering can be formulated as follows: given a dataset **D** that contains n objects x1, x2, x_n (data points, records, instances, patterns, observations, items) and each data point is in a d-dimensional space, i.e. each the data points has A dimensions (attribute, feature, variable, and component that is Can be expresses in the form of a matrix as allow.

$$\mathbf{A}= \begin{bmatrix} \boldsymbol{x1.1} & \cdots & \boldsymbol{x1.n} \\ \vdots & \ddots & \vdots \\ \boldsymbol{xn,1} & \cdots & \boldsymbol{xn.n} \end{bmatrix}$$

Data clustering is based on the similarity or dissimilarity (distance) measures between data points. Hence, these measures make the cluster analysis meaningful [28]. The high quality of aggregation is obtained in strong similarity in mass and weak similarity between groups as shown in Figure 1.2. In addition, when we use the dissimilarity (distance) concept, the latter sentence becomes: the high quality of clustering is to obtain low intra-cluster dissimilarity.

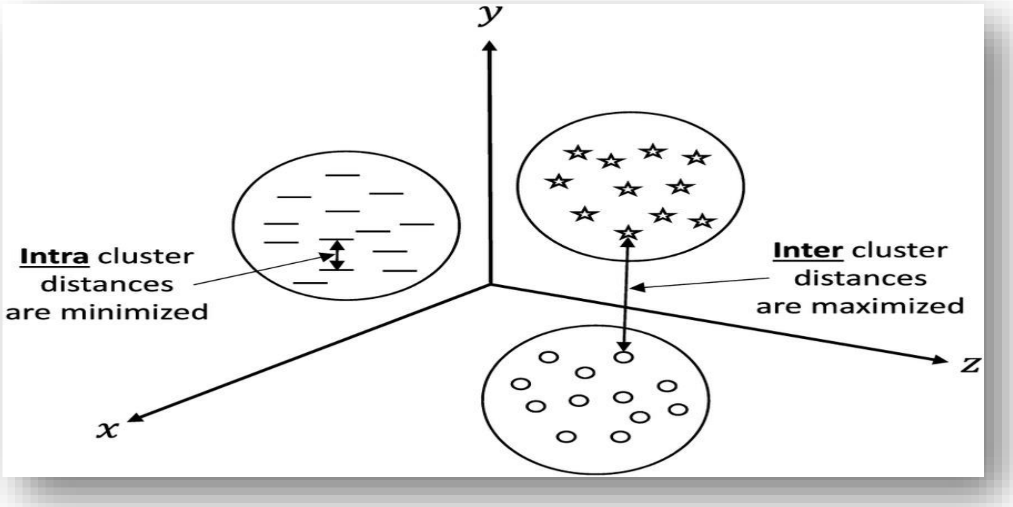

y
Intra cluster distances are minimized
Inter cluster distances are maximized
x
z

Figure 1.2 . Inter-clustering and Intra-clustering similarities of clusters.

1.3 Type Of Clustering

We can conduct groups in a number of different ways. Each assembly technique produces several different types of cluster.

Some take input parameters from the user like number clusters to be formed etc., but some decide on the type and amount of data given. The main developments have been the introduction to density based and grid based clustering methods. We can classify cluster algorithms into five different types.

- ✓ Partitioning methods
- ✓ Hierarchical methods.
- ✓ Model-based methods.
- ✓ Density based methods.
- ✓ Grid based methods.

1.3.1 Partitioning Methods.

The easiest and most important version of cluster analysis is the division that organizes group objectives in many exclusive groups or clusters. To keep the problem specification concise, we can assume that the number of clusters is given as background knowledge. This parameter is the first point of division methods.

The data set comes out of n objects formally and k is the number of groups to form a partitioning algorithm that arranges objects in k k_n sections where. cluster is configured to improve an objective partitioning criterion such as a difference function based on the distance so that the objects are within a block, So the objects include the group “identical”. If a database containing n data objects is given, then a partitioning

method constructs k clusters of the data where k<=n and k is that the input parameter provided by the user. That is, it classifies the data into groups which should satisfy the following conditions: (1) each group must contain at least one data object and (2) each data object ought to belong to just one group. The second demand becomes easy in fuzzy k mean agglomeration at intervals that one object square measure typically resembled by two or lots of teams. With k as the given number of partitions to be MADE the partitioning technique creates an initial partition methodology creates an initial partition method MADES an initial partition. Then more number of iterations are followed in which objects are moved from one group to other making sure that in-cluster similarity is more than similarity with Objects in another cluster.

Popular Partitioning methods k means and k-means the most well-known and usually used partitioning methods are k-means proposed by (Mac Queen 1967) and k-means proposed by (Kaufman and Rousseau 1987).

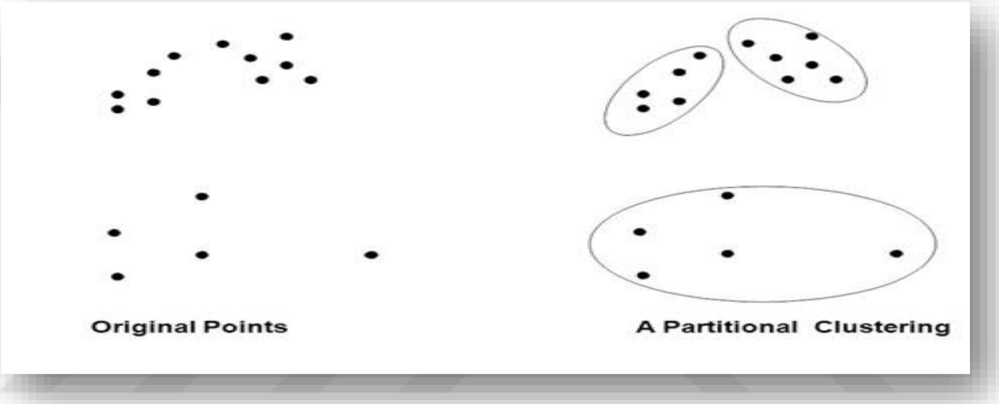


Figure 1.3. Original Points and **A partitional clustering**

1.3.2 Hierarchical Methods

The set of given data objects are partitioned in form of a tree like structure or nested clusters in hierarchical clustering. The hierarchical They were classified into two kinds.

- **Agglomerative**
- **Divisive**

In agglomerative method also known as bottom-up approach, Forms of all objects separate group. It successively merges the groups close tone another by checking the similarity function, until all the groups are merged into one, that's until the top most level of hierarchy is reached or until a termination condition holds. In divisive clustering also known a stop-down approach, initially all the objects are grouped into a single cluster which can also be called as parent. The repetition process is performed each time the object is divided into group until the object becomes in one group or it may be alone.

1.3.2.1 Agglomerative Method

This method begins by treating each object as an individual cluster and then proceeds by groping two nearest clusters. The distance between any two clusters m and n is defined by a metric $D_{m,n}$. Metrics can be single-link, complete-link and group average etc. A general class of metrics was given by Lance and Williams [1]. If Dk, ij be the distance between cluster k and the union of cluster i and cluster j, then: $Dk, ij = \_iDk, i + \_jDk, j + \beta Di, j + \gamma|Dk, i - Dk, j|$ The agglomerative method is as follows:

Consider each object to be an atomic cluster. The $(n\,x\,n)$ distancematrix represents the distance between all possible pairs of clusters. Find the smallest element in the matrix. This corresponds to the pair of clusters that are most similar. Merge these two clusters, say and n, together. Measure the distances between the newly formed cluster and the other remaining clusters using a distance function. Delete the rowan column of m and overwrite row and column of cluster n with the new values. If the current number of clusters is more than k then go to step 2. otherwise stop. The merging process can continue until all the objects are in one cluster. The advantages of hierarchical methods are that It is therefore easy to implement computationally. They can tackle larger datasets than the k-means method and we can run the algorithm without providing the input k (the number of clusters to be formed). The drawbacks of agglomerative method are:

The algorithm has $O(n3)$ time complexity. Even though the order of the distance matrix decreases with each iteration, the cost of Step2 on iteration k is $O\ ((n - k)2)$, and we are guaranteed $(n - k)$ iterationsbefore we get to k;The clusters produced are heavily dependent on the metric Di, j. Different metrics can produce different clusters. For instance, the complete-link metric tends to produce spherical clusters, whereas the single-link metric produces elongated clusters [1].

1.3.2.2 Divisive Method

The contrast procedure of agglomerative clustering is the divisive method. Initially all the data objects are considered in one cluster. Then for each object the degree of irrelevance is measured and the most irrelevant data object is split from the main cluster and a new cluster is formed with only that data object in it. The highest degree of irrelevance of an object corresponds to the one that is most distant from all other objects in that cluster. Let the average distance between object i and then cluster Cj be defined as [1]: Di, Cj =The most irrelevant object splits off and forms a new cluster. This is equivalent to splitting the cluster with the bigger diameter. The process will continue until it perfect the specific termination requirement or the required crusts above a certain threshold distance. These methods face the difficulty of making a right decision of splitting at a high level. The algorithm for divisive method is [1]:

- Choose the cluster that has the most remote pair of objects. Thesis the cluster with the largest diameter.
- In this cluster, we find the object with an average distance greater than other objects.
  Remove the object from the cluster, allowing it to form a new atomic cluster.
- For object h in the cluster being split, calculate the average distance between it and the current cluster; and the average distance between the object and the new cluster. If the distance to the new cluster is less than the distance between it and the16-current cluster, move the object h to the new cluster. Loop over althea objects in the cluster.
- If no objects can be moved, but the current number of clusters is greater than k, go to step1. Otherwise stop. The drawbacks of divisive method are:
- The time complexity of algorithm is O(n3), O(n2) on the step1 of the algorithm for each iteration. Moreover, there are expensive calculations that may take place in step3 of the algorithm.
- In step 3 the group averages between an object and the new and existing clusters need to be recalculated after an object is moved.
  This will be costly in terms of number of calculations and the amount of storage required.
- The method only searches one of the $N(n, k)$ possible partitions.

In hierarchical clustering once a split or merge is done, it cannot be undone. This fact acts as both key to success and drawback for hierarchical clustering. The firmness of hierarchical method leads to less computational cost without a combinatorial number of choices but the main problem with it is invalid decisions cannot be corrected.

Hierarchical clustering methods are simple but encounter problems at making critical decisions for selection of correct merge are split points. Such a decision is critical because once a group of objects is merged or split, the process at the next step will work on the newly generated clusters. It will ne'er undo what was done antecedently nor perform object swapping between clusters therefore merge or split if not done wise might end in caliber clusters. Thus, merge or split if not done wise may result in low quality clusters. These ways have scaling downside since the choice of merge or split must examine and measure a decent variety of objects or clusters Hierarchical clustering can be improved Processing By integration this technique with alternative cluster techniques for multiple part cluster. One such technique known as birch 1st partitions objects hierarchically use tree structures then applies different cluster techniques to provide refined clusters.

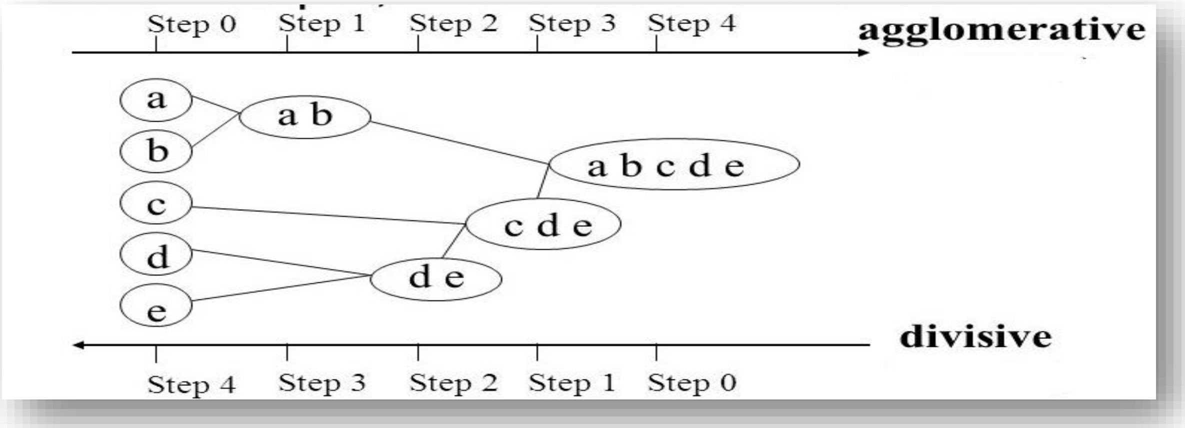


Figure 1.4. Agglomerative and Divisive

1.3.3 Model-based methods

Each component is described by a density function and has an associated probability or “weight” in the mixture. In precept, we will able to} adopt any probability model for the parts however usually we'll assume that parts are p variety normal distributions. (This does not necessarily mean things are easy: inference in tractable, however).

1.3.4 Density Based Methods

Density-based clustering ways are supported a local cluster criterion. Clusters area unit assumed as regions within the data area during which the objects area unit dense and therefore the clusters area unit separated by regions of low object density. These regions have a spot shape and the data points inside a cluster may be spot distributed.

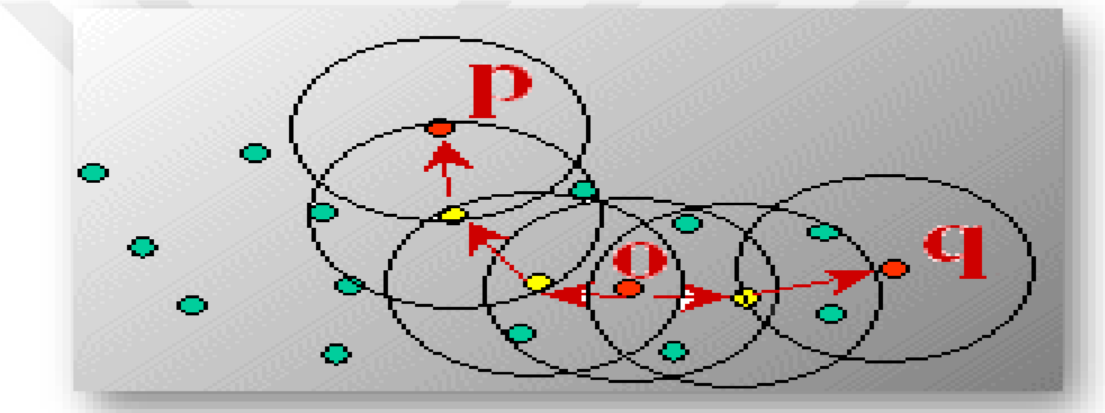


Figure 1.5. Density-Based clustering

1.3.5 Grid Based Methods

A grid based mostly clustering technique takes inside the thing space and quantizes it into affinity range of cells forming a grid structure. Then the method performs all the operations on it grid structure. The main advantage of this methodology is its quick interval that is freelance of the quantity of objects, and dependent solely on the quantity of cells in each dimension within the quantized area. Grid-based ways use one uniform grid mesh to partition the whole downside domain into the cells and the data objects settled within a cell area unit represented by the callusing a set of statistical attributes from the objects. cluster is performed on the grid cells, instead of info itself. Since the scale of the grid is far but the quantity of data objects, the process speed will be significantly improved.

1.4. Importance of Clustering

Data clustering is one of the main tasks of data mining [1] and pattern recognition [2]. Moreover, it can be used in many applications such as:

1. Data compression [3].

2. Image analysis [5].

3. Bioinformatics [6].

4. Academics [9].

5. Search engines [79].

6. Wireless sensor networks.

7. Intrusion detection [81].

8. Business planning [82].

9. Add another application to use the monastery These are social and statistical studies of real life data Such as the rate of crime and unemployment in a statistical study as well as pollution and its relationship to oil consumption, pollution and the production of oil. This is the focus of our study on this proposal using dye in the classification of countries in terms of pollution, oil consumption, pollution and oil production.

2. DATA CLUSTERING

data clustering may be a crucial obstacle with many applications in

- Machine learning.
- computer vision.
- Signal method.

the item of clustering is to divide a dataset into natural groups such as:

Points among constant cluster area unit similar. Points incurious groups area unit dissimilar to each different. Clustering methods can be: Hierarchical: Single Link, Complete Link, etc. Partitioned or flat: k-means, Gaussian Mixture, Mode Seeking, Graph partitioning.

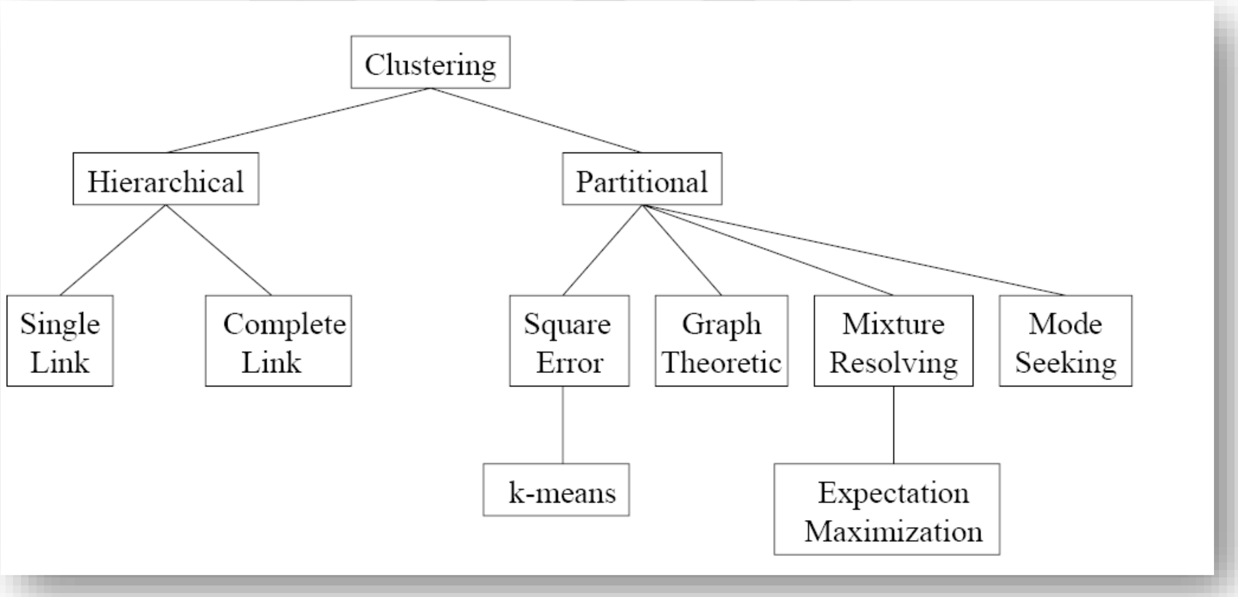


Figure 2.1. classification of clustering algorithms

2.1 Clustering Algorithms

There are many types of cluster algorithms, we will focus on only two types, which are used in thesis. K- means cluster and spectral cluster. But the study was applied to spectral cluster.

2.1.1 K-means cluster Algorithms

Some objects are separated into groups such as K cluster. Some groups are reduced to centroids of the clusters is minimized.

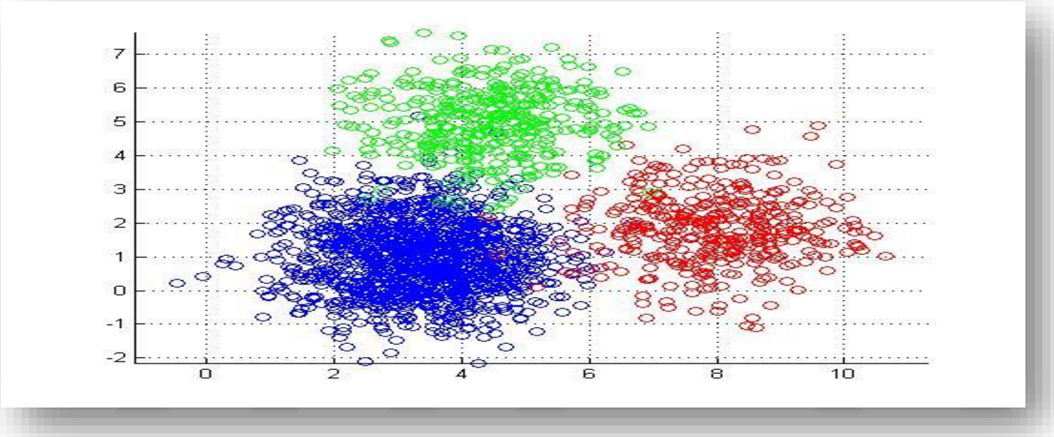


Figure 2.2. k-mean Clustering

2.1.1.1 Pros

- It's easier to understand.
- comparatively active.
- We get a better result when they are separate. So that each group is separated by a special center.
- You only need to calculate distances between the center only until its center is configured. The distance between the inner points is not measured.
- The following figure illustrates the concentration of data and the composition of the class.

2.1.1.2 Cons

- The education algorithm order apriority specialization of the number of cluster centers.
- Euclidean, space menstruation can unequally be weighing underlying factors.

- The education, algorithm supply the local optima of the square mistake function.

- The learning algorithm provides the native optima of the square error function.
- Random configuration of the cluster center may not lead us to a useful result.
- The application is only -used when finding and defining the average any failure in the class data Unable to handle extreme data the algorithm failed to set nonlinear data.

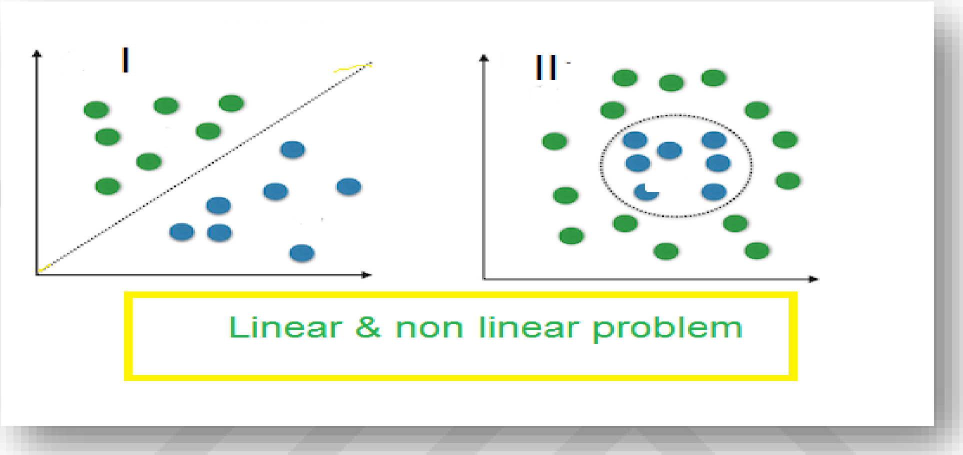


Figure 2.3 linear and nonlinear

2.1.2 Spectral Clustering Algorithms

It is considered one of the best and most important algorithms in the extraction of data, and useful in the search for a large amount of data, and this is through the process of linking the analysis of these data and methods of artificial intelligence to become the best efficiency in the search process the Clustering is a process of data from similar clustering, a branch of data mining. The aggregation algorithm divides data sets into several groups, where the similarities between points within a group are greater than the similarities between two points within the two different communities. The idea of simple data collection in nature and very close to the human in his way of thinking where we whenever we deal with a large amount of data tend to a huge amount of data to summarize a few groups or categories, to facilitate the analysis process. Not only used to organize and classify the data, but used in data compression and build the sample system algorithms to collect data widely. If data sets are found, it is possible to build

a problem model based on those groups. The clustering, in data extraction, can detect clusters and identify interesting distributions in basic data. It is illustrated by this figure (1). The data points are also nodes of connected.

It studies the relationship between the data points themselves and then distributes them to groups where data in the same group must be connected and coherent.

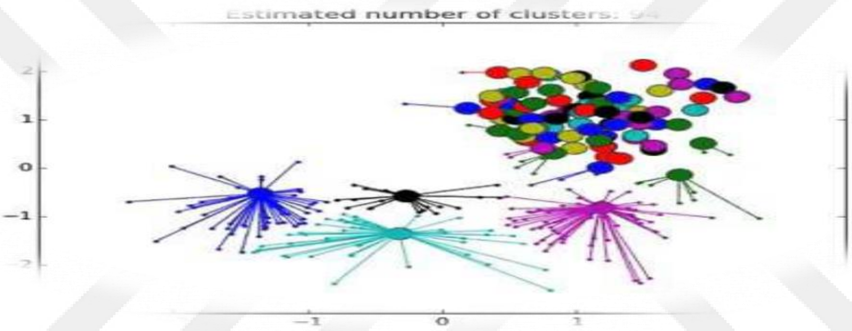


Figure 2.4. spectral clustering

2.1.2.1 Cons

- ❖ Strong assumptions cannot be obtained from cluster statistics.
- ❖ It features easy implementation.
- ❖ Reasonable clusters results.

2.1.2.2 Pros

- ❖ Some of the times are sensitive to the selection of parameters.
- ❖ When large data is made, it costs a lot to calculate.
- ❖ The data set that can be able to execute is not common and in real life.

2.2 ALGORITHMS (k-means and spectral)

2.2.1 K-mean clustering algorithm

These centers should be placed in a boring manner because of a different location causing a different result. Therefore, the best option is to put them as far as possible apart from each other. The next step is to take each point belonging to a particular data set and connect it to the nearest center.

To apply the algorithm correctly, it should be positioned as far as possible apart from each other. The next step is to take each point belonging to a particular data set and connect it to the nearest center.

When no point is outstanding, the primary step is completed and also the group is aged early. At this time, we'd like to calculate the new C-centroid because the Paris center of the clusters ensuing from the previous step. when we've these new centroid k. a new binding must be done between an equivalent data set points and the closest new center. A loop has been generated. As a result of this loop we could notice that the k centers amendment their location step by step till no additional changes are done or in alternative words centers don't move to any extent further. Finally, this algorithmic rule aims at minimizing AN objective function know as square error function given by

$$\boldsymbol{J}\,(\boldsymbol{V}) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{c}} \sum_{\boldsymbol{j=1}}^{\boldsymbol{ci}} \left(\|x_i - v_j||\|\right)^{\mathbf{2}}$$

Where:

$||x_i - v_j||$' Is the Euclidean space between x_i and v_j 'c_i' is the number of data points in $I$the cluster. 'c' is the number of cluster centers.

2.2.1.1 Algorithms steps for k-means clustering

Let $X = \{x1, x2, x3, \dots \dots . . . , xn\}$ be the set of data points and $V = \{v1, v2, \dots \dots . , vc\}$ be the set of centers.

1. Random selection of 'C' cluster centers.

2. Calculate the distance between every data point and cluster centers.
3. Let the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
4. Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
5. Repeat the new cluster center account using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

where, 'c_i' represents the number of data points in it cluster.

6. Recalculate the distance between every data point and new obtained cluster centers.
7. If no data point was reassigned then stop, otherwise repeat from step 3).

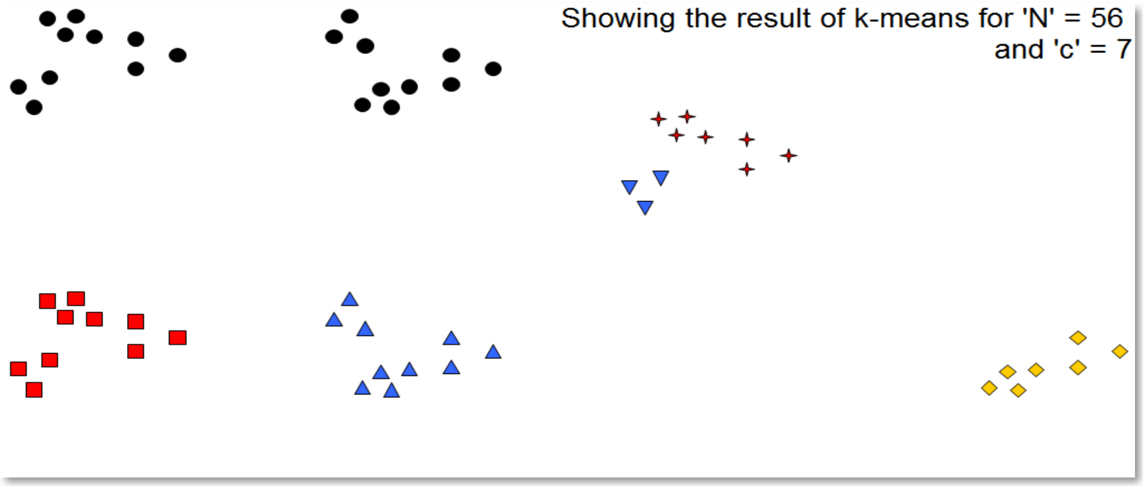


Figure 2.5. Showing the result of k-means for *'N'* = 56 and *'c'* = 7

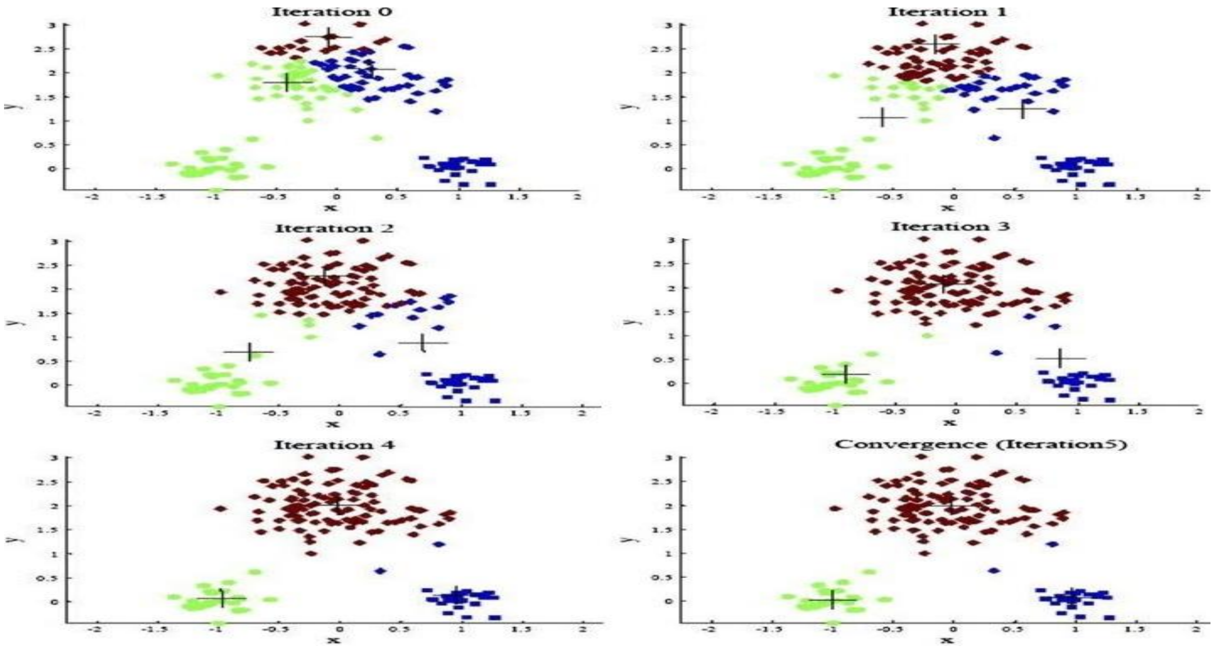


Figure 2.6. k-mean Clustering 1-5

2.2.2 Algorithms steps Spectral Clustering

supposed data from, read it from outer file to the matrix.

$$\mathbf{A}= \begin{array}{cc} X_1 & y_1 \\ X_2 & y_2 \\ . & . \\ . & . \\ . & . \\ X_n & y_n \end{array}$$

Find affinity matrix $A\ [n, n]$ from matrix $A[n, 2]$ $\boldsymbol{Affinity} = \dfrac{1}{e^{dis/2\sigma^2}}$

where d is the distance between every Element in and others in the same matrix A by the low

$$\mathrm{dis}\ (\mathrm{p}_i\ , \mathrm{p}_j) = \sqrt{(\mathrm{x}_i - \mathrm{x}_j)^2 + (\mathrm{y}_i - \mathrm{y}_j)^2} \quad \textbf{\textit{Where}} \quad \mathrm{p}_i = (\mathrm{x}_i, \mathrm{y}_i)\ \&, \mathrm{p}_j = (\mathrm{x}_j, \mathrm{y}_j).$$

Considering **K** is a parameter express a number of a suitable, vectors will be select from eigenvector in common, similar, to number of clusters k+1 to apply neigvec.

The scaling parameter $\boldsymbol{\sigma^2}$ has a control over how fast this approximation falls with the distance between p_i , p_jTwo Local Scaling is a measure of two similar points. This provides an intuitive way to determine possible values for σ. The chosen of σ is generally done manually. propose, choosing σ automatically by running their clustering algorithm frequently for a number of values of σ and selecting the one which supply lower distorted clusters.

We note. In addition, the set of values to be tested must still be determined manually. Furthermore, when input data includes groups with different local statistics, there may note a single value σ that works well for all data. The significant effect σ on the cluster. When the data contains multiple scales, even using optimization σ fails to provide a good set, the local scaling effect. The affinities across clusters are now safely, lower than the affinities within any single cluster. To set the value of σ, we calculate it for finding the Affinity of each two points i and j First we must analysis $\sigma 2\ to\ \sigma_i,\ \sigma_j \sigma_i = \boldsymbol{dis}(s_i - s_k)$.

We consider$s$ k is **K'th** neighbor of$\boldsymbol{si}$points. The choosing of K is not dependent or independent of the scale and is the dimension function of data from the embedding area. However, in most experiments (both on synthetic data and on images) one value of K = 3 was used, which yields good results even for high-dimensional data$\boldsymbol{\sigma_j}$= $\boldsymbol{dis(s_i - s_k)}$.

1	Affinity	…	Affinity
Affinity	1	…	Affinity
Affinity	Affinity	…	Affinity
…	…	…	…
	Affinity (p_n, p_2)	…	1

Figure 2.1 Affinity Matrix

computation the degree matrix the grade of matrix Degree is the same as that of matrix Affinity Matrix taking into consideration. Calculate the sum of each row of matrix Affinity in a position equal to row and column Degree matrix.

$$\begin{cases} \text{Degree } [\boldsymbol{i}, \boldsymbol{j}] = \sum_{j=1}^{n} \text{Degree } [\, i, j \,] \text{ If } i = j & \text{if } i=j \\ \text{Degree } [\boldsymbol{i}, \boldsymbol{j}] = 0 & \text{if } i \neq j \end{cases}$$

Consider n is a grade matrix with all elements of zeros except the main diameter.

Sum 1st row	0	0		0
0	Sum 2st row	0		0
0	0	Sum 3st row		0
0	0	0	0	Sum n st row

Figure 2.2 .Degree Matrix

Find an affinity matrix for Laplacian normalization through these low levels and is used to find each element in the N.Lap matrix Normalized L.N= AFFINIT[I,J] $/ \sqrt{\text{dgree}[i, i]} * \sqrt{\text{dgree}[j, j]}$

N.laplacian[1.1]	N.laplacian[1.2]	N.laplacian[1.3]	…	N.laplacian
N.laplacian[2.1]	N.laplacian[2.2]	N.laplacian[2.3]	…	N.laplacian
N.laplacian[2.3]	N.laplacian[3.2]	N.laplacian[3.3]	…	N.laplacian
…	…	…	…	…
N.laplacian[1.n]	N.laplacian[n.2]	N.laplacian[n.3]	…	N.laplacian

Figure 2.3. Normalized Laplacian Matrix

- **Decomposition(** linear algebra)

E.vectors [1.1]	E.vectors [1.2]	E.vectors [1.3]	…	E.vectors [1.4]
E.vectors [2.1]	E.vectors [2.2]	E.vectors [2.3]	…	E.vectors [2.4]
E.vectors [3.1]	E.vectors [3.2]	E.vectors [3.3]	…	E.vectors[3.4]
…	…	…	…	…
E.vectors [n.1]	E.vectors [n.2]	E.vectors [n.3]	…	Eigenvectors [n.n]

Figure 2.4. Eigen values, Eigenvectors

Vn	Vn-1	Vn-2	Vn-3
v (1,n)	v (1,n-1)	v (1,n-2)	v (1,n-3)
v (2,n)	v (2, n-1)	v (2, n-1)	v (2, n-3)
…	…	…	…
v (n,n)	v (n, n-1)	v (n, n-2)	v (n, n-3)

Figure 2.5. Eigenvalues, Eigenvectors

λ_1
λ_2
.
λ_n

v(1,1)	v(1,2)	v(1,3)	v(1,4)
v(2,1)	v(2,2)	v(2,3)	v(2,4)
v(n,1)	v(n,2)	v(n,3)	v(n,4)

Figure 2.6. Rearrange Neigenvec

$\sqrt{\mathrm{v}(1,1)^2 + v(1.2)^2 + v(1.3)^2}$	v(1,2)/round	v(1,3)/round	v(1,4)/round
$\sqrt{\mathrm{v}(2,1)^{\mathbf{2}} + \boldsymbol{v}(\mathbf{2.2})^2 + \boldsymbol{v}(\mathbf{2.3})^2}$		v(2,2)/ round	v(2,4)/ round
$\sqrt{\mathrm{v}(\mathrm{n},1)^{\mathbf{2}}} + \boldsymbol{v}(\boldsymbol{n}.\mathbf{2})^2$		v(n,2)/ round	v(n,4)/ round

Figure 2.7 .Normalized Matrix

3- METHODLOGY

Overview: This search is described by applying this algorithm. All data is real, which is one of the advantages of this study. Real and realistic data were obtained from the European Economic Association website. a world-class statistical data center. In this study, we will focus on the pollution data from the European statistical site EEA and its correlation relation with oil consumption in 31 EU countries from 2006 to 2014.Throughout this study, we will identify the countries most affected by pollution and petroleum consumption by applying the algorithm (**Spectral**, **K-means**)and phases of implementation.

3.1 General stages of implementation.

- ***First phase.***

Obtaining data from the warehouse which mostly unworkable. After the data are obtained, it is processed and prepared using a number of data mining techniques.

- ***Second phase.***

The spectral and k-means algorithm are implemented more than once on the algorithms to obtain a good result.

- **Fourth phase**.

Results: We obtain a graphical pattern showing the distribution of the cluster in the two algorithms(spectral and k-means).

- **Final phase.**

The results are checked in terms of data classification and as to whether there is a correlation relation.

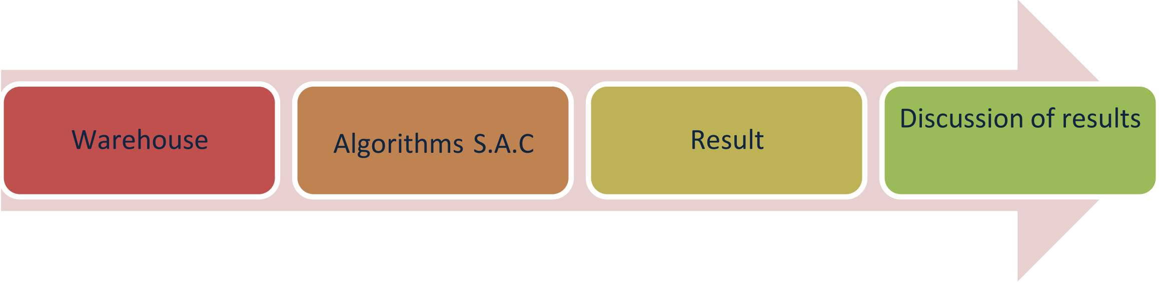


Figure 3.1 .phases Of implementation

3.2 Preprocessing Data Phases.

The following figure, in general, illustrates a number of data mining techniques.

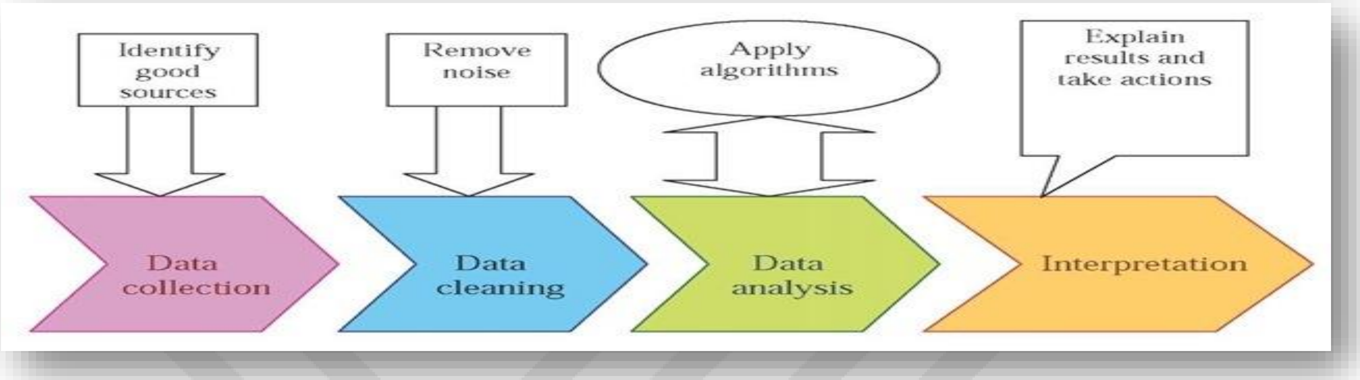


Figure 3.2 .phases preprocessing data.

3.2.1 Data Selection

The data to be analyzed is specified to arrive at a logical classification.

3.2.2 Data Clean

Exclude incomplete data and data not compatible with the remaining data.

3.2.3 Data Set 1

Data Set 1 includes the quantities of emissions of pollution in tons in 31 countries in the European Union for the 2006 to 2014 period (in Belgium, Bulgaria, Czech Republic, Denmark, Germany, Estonia, Ireland, Greece, Spain, France, Croatia, Italy, Cyprus, Latvia, Lithuania, Luxembourg Hungary, Malta, Netherlands, Austria, Poland, Portugal, Romania, Slovenia, Slovakia, Finland Sweden, United Kingdom, Norway and Turkey). This group shows the amount of pollution and the amount of oil consumption in tons.

geo\time	2006	…	2008	…	2012	2013	2014
Belgium	…	…	2,561,412	…	47,309	44,702	
Bulgaria	…	…	571,696	…	329,980	195,832	188,937
…	…	…		…			
Sweden	…	…	30,501	…	28,339	26,802	23,973
Norway	…	…	20,027	…	17,300	16,664	16,633
United Kingdom	…	…	491,595	…	439,147	385,982	307,638
Turkey	…	…	2,561,412	…	2,715,911	1,939,104	2,147,499

Table. 3.1 .Data set 1

3.2.4 Data Set *2*

Quantities of oil consumption per ton. And 31 countries - in the European Union and for the period 2006 to 2014 **States are**: (Belgium, Bulgaria, Czech Republic, Denmark, Germany, Estonia, Ireland, Greece, Spain, France, Croatia, Italy, Cyprus, Latvia, Lithuania, Luxembourg Hungary, Malta, Netherlands, Austria, Poland, Portugal, Romania, Slovenia, Slovakia, Finland Sweden, United Kingdom, Norway .and Turkey) from 2006 to 2014. In this group shows the amount of oil, and the amount of oil consumption by tons.

geo\time	2006	…	2008	…	2012	2013	2014
Belgium	…	…	24,600	…			23,123
Bulgaria	…	…	4,858	…			4,078
…	…	…		…	…	…	
Sweden	…	…	13,967	…			12,040
Norway	…	…	12,380	…			10,640
United Kingdom	…	…	75,214	…			66,263
Turkey	…	…	30,608	…			34,447

Table. 3.2 .Data set 2

3.2.5 Data Transformation

At this stage, the data is converted into custom templates to suit the search and retrieval procedures. Through the compilation process, in our study, the years are converted into columns and placed in a sequential order from 2006 to 2014. with 31 countries in the European Union, maintaining the sequence. The following figure shows pollution data.

Countries	YEARS	POLLUTION
Belgium	2006	133,599
Bulgaria	2006	765,131
…	…	…
Belgium	2007	124,092
Bulgaria	2007	820,614
Belgium	2010	60,615
Norway	2014	16,633
Turkey	2014	2,147,499

Figure. 3.3 .Data transformation

The following figure shows oil consumption data for these countries and for the same years starting from **Belgium** to **Turkey** between **2006** and **2014,** as follows.

Countries	Years	OIL
Belgium	2006	23,619
Bulgaria	2006	5,126
…	…	…
Belgium	*2007*	22,932
Bulgaria	2007	4,954
..	…	…
Belgium	2010	24,363
Bulgaria	2010	4,046
…	…	…
Norway	2014	10,640
Turkey	2014	34,447

Figure. 3.4 .Data transformation

3.2.6 Data Integration

At this stage, similar and relevant data are collected from multiple data sources and combined. Data were collected on petroleum consumption and pollution emissions while maintaining the order of countries and the sequence of years (from 2006 to 2014, which are relevant to our study).

COUNTRY	YEAR	Pollution	oil
Belgium	2006	133,599	23,619
Bulgaria	2006	765,131	5,126
…	…	…	…
Germany	2007	459,942	110,214
Estonia	2007	88,034	1,203
Portugal	2009	61,071	12,526
Romania	2009	444,827	9,023
Slovenia	2009	10,820	2,531
…	…		
Lithuania	2010	21,120	2,537
Luxembourg	2010	1,756	2,819
…	…	…	…
Portugal	2011	48,065	11,250
Romania	2011	320,090	8,963
Sweden	2012	28,339	12,998
United Kingdom	2012	439,147	67,636
…	…	..	…
Norway	2014	16,633	10,640
Turkey	2014	2,147,499	34,447

figure. 3.5 .data integration

3.3 Data Mining

Combine the values associated with the study.With the exception of year and country and maintain sequence to indicate data- Oil consumption - tons of pollution. As in the following figure So that we get a good pattern. Be in three columns - Sequence - Petroleum - Pollution. While maintaining the previous sequence of data. To extract a knowledge-based pattern of petroleum-pollution.

NO	Pollution	Oil	Cluster
1	133,599	1,364.6	...
2	459,942	23,728.1	...
...		...	...
3	21,120	1,064.6	...
270	2,147,499	12,080.5	...

Figure 3.6 . Data Patterns

This is the final form of the modification of the previous data.

No	Pollution	Oil
1	133,599	1,364.6
2	459,942	23,728.1
...	...	...
3	21,120	1,064.6
270	2,147,499	12,080.5

Figure3.7 . Dataset (output dataset)

3.4 Implantation

After completion of data mining and data filtering, the spectral algorithm is ready to be implemented. The implementation phase of the algorithm is as follows:

1- Mathematical steps to calculate the value of matrix A (a square matrix is created from the original data). Also at this stage, the matrix of Affinity is calculated, in which we find the values of the distance between the points).Matrix D is also calculated. The diagonal of matrix A is used to find an N.L matrix.

2- Steps of Linear Algebra. >In this phase, the eigenvalue value is calculated. The matrix is also calculated by finding vector v.Apply the k-means algorithm to the U matrix.

3- Details of the implementation phase mathematical steps.

3.4.1 Load Dataset

Data is represented in the figure of a two-dimensional matrix. In our study, the matrix was arranged in**2** columns and in **270** rows, which are the data of pollution emission points and oil consumption [2 × 270].

No	pollution	Oil
1	133,599	23,619
2	765,131	5,126
...	...	...
270	2,147,499	34,447

Figure3.8 . Dataset

3.4.2 Running the Algorithm.

- ***Preprocessing.***
- ***Decomposing.***
- ***Running k-means algorithm.***

To run the algorithm, there are three operations which execute sequentially, as follows:

3.4.2.1 Preprocessing

1) Calculate Affinity matrix:

Find affinity matrix A $[270{,}270]$ from matrix $\mathbf{A}$ $[270{,}2]$

Affinity = $\mathbf{1/e\ dis/2\sigma 2}$

where d is the distance between every element in and others in the same.

Matrix A by the low

$$\boldsymbol{dis\ p1, p2 = \sqrt{}x1 - x2)2\ +\ (y1 - y2)2}$$

Where $p1\ =\ (x1, y1)$ **&** $p2\ =\ (x2, y2)$The trial and error value were calculated to obtain a satisfactory result.

The value was k=3.

2) Calculate the Row Matrix

Calculate the sum of each row of matrix A in a position equal to the rows and columns of

Matrix **D**.

$$D[\,i,j\,] \;= \textstyle\sum D[\,i,j\,]\, nj = 1 \quad If \;\; i \;=\; j$$

$$D[\,i,j\,] \;=\; Zero \qquad\qquad if \;\; i \;\neq\; j$$

where $n = 270$is the matrix degree

All other elements in matrix D will be equal to zero.

3) Calculation to normalize the Laplacian affinity matrix

$$L = \frac{affinty\,\{i.j\}}{\sqrt{d\{i,j\}} * \surd d\{j,j\}}$$

3) Calculation is to normalize Laplacian affinity matrix

$$L = \frac{affinty\,\{i.j\}}{\sqrt{d\{i,j\}} * \surd d\{j,j\}}$$

3.4.2.2 Decomposing.

We used linear algebra techniques to find both V and the eigen value **λ (* V = λ * v)**,where v is the eigenvector of N.lap corresponding to λ. We have 270 values of eigenvectors and eigenvale, so for each λ, there is V (λ1. V1, λ2 ... V2, ..., λ270. V270). From the eigenvectors, we choose (k + 1) = 2 similar to k from any group. We select the largest integers according to k which correspond to the same number of the largest eigenvalues because k = 3 equals 4.

We construct a normalization matrix [U] from the Nig matrix obtained [270,2] to find each element of this matrix using the code.

$$U\{I,J\} = \frac{v(i,j)}{\surd v(i,1)2 + v(i,2)2 + v(i,3)2}.$$

3.4.2.3 Running k-means Algorithm.

The K-means algorithm is run with the [U] matrix. Then, the data will be broken down into groups according to spectral clustering.

3.4.2.4 Results Analysis

Data is represented in the two dimensions. All cluster data are colored to distinguish one from the others in order to determine whether these results are acceptable. The results cannot be accepted in one way or another in the parameter change models for the k and σ values, or only one of them. For meaningful results, the properties of each group must be distinguished between the groups that support the end outputs of spectral clustering algorithm implementation.

No	Pollution	Oil	Clusters
1	133,599	23,619	2
2	765,131	5,126	3
270	2,147,499	34,447	5

Table 3.8 Outputs of Implementation S.C.A k=3,4,5,6

3.5 Result and Dissociation

3.5.1 Results of the Spectral Algorithm (*data set of oil and pollution*)

When the dataset (pollution, oil) was performed using MATLAB language and using the A.S.C. algorithm, we produced the following pattern: a drawing that shows several clusters such that each color is a different color. Each color represents a different classification.

The following figure illustrates this classification.

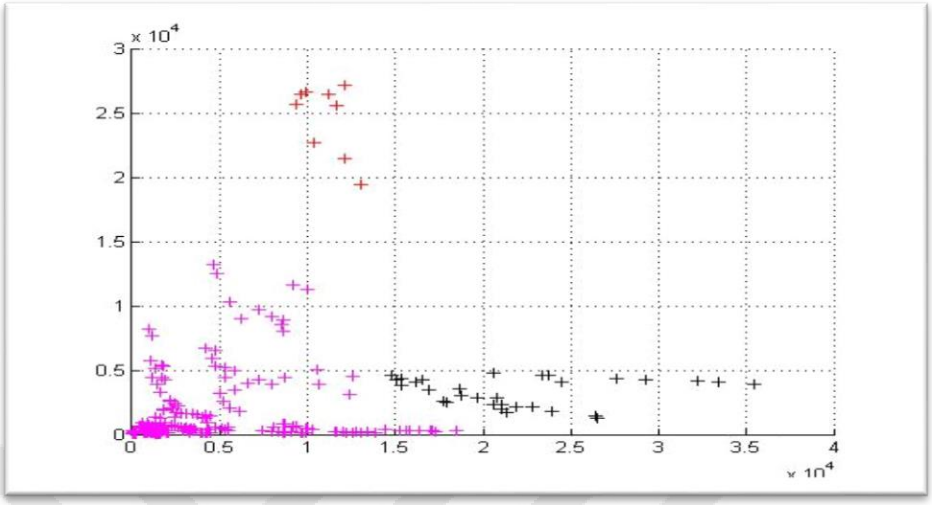


Figure 3.9. Represent Dataset Points in 2D

Application of the algorithm to data related to the production of oil energy and the amount of pollution combines the result of implementing the algorithm and the original data set.

For best results. Therefore, we implemented the program several times with the change K value as the cluster number each time, and then we studied these results for the best.

Point	Country	Year	Pollution	Oil	Cluster
1	Belgium	2006	1335.99	23619	5
2	Bulgaria	2006	7651.31	5126	4
3	Czech Republic	2006	2031.63	9867	3
4	Denmark	2006	299.58	8249	5
…	…	…	…	…	…
268	United Kingdom	2014	3076.38	66263	1
269	Norway	2014	166.33	10640	6
270	Turkey	2014	21474.99	34447	2

Figure 3.10 Outputs of Implementation with Dataset.

Result S. C.A with k=3

And through follow-up and observation When the algorithm is executed for the **k = 3** value we get a logical relationship. The higher the consumption of the average of the oil, the lower the average of the pollution. As shown in Figure 3.5,when **k = 3** is executed, the following was obtained. Figure 3.11, average of pollution, average of oil and average standard deviation. Graph 3 shows the number of clusters.

Note. We used standard deviation, which is a measure of the dispersion of a set of data from its mean. It is calculated as the square root of the variance by determining the variation between each data point relative to the mean. If the data points are further from the mean, there is higher deviation within the data.

No. Clusters	AVG.Pollution	AVG. Oil	Avg.str
Cluster**1**	1286.125	5630.6	4129.073
Cluster**2**	10204.45	7407.96	1772.304
Cluster**3**	17299.76	18375.74	10408.96

Figure 3.14 two Avg. Values (pollution and oil) for S.C.A (k=3)

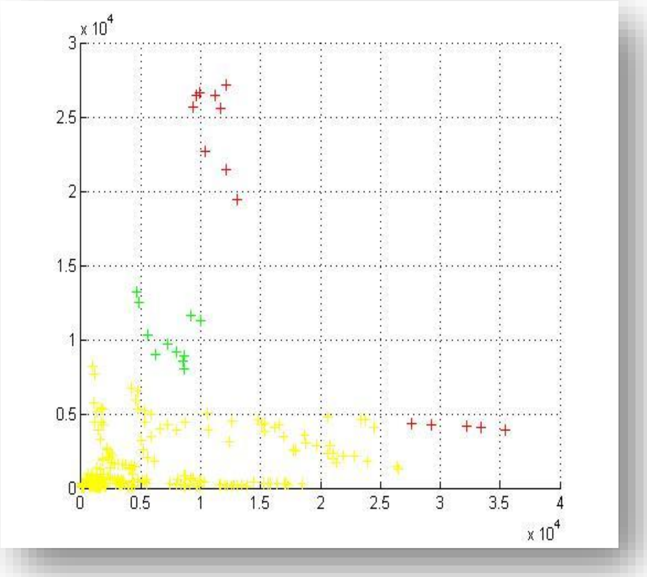


Figure 3.12. Representing S.C.A result with 3 Clusters

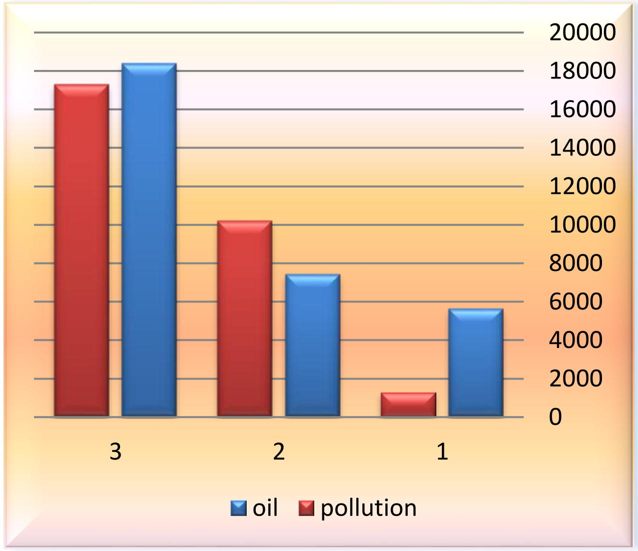


Figure 3.13. Graph result for S.C.A with k=3

Result S. C.A with k=4

The fourth cluster may not give a correlation relation. It gives only a classification. The remainder of the clusters gives a correlation relation and classification (clusters 1,2,3),as shown Figure 3.14.

No. Clusters	AVG.Pollution	AVG oil	Avg.Str
CLUSTER **1**	1285.3	5459	3997.923
CLUSTER **2**	8885.628	7872.9	1689.825
CLUSTER **3**	17299.76	18375.75	10030.33
CLUSTER**4**	1378.31	26441.4	71.945

Figure 3.14 two Avg. Values (pollution and oil) for S.C.A (K=4)

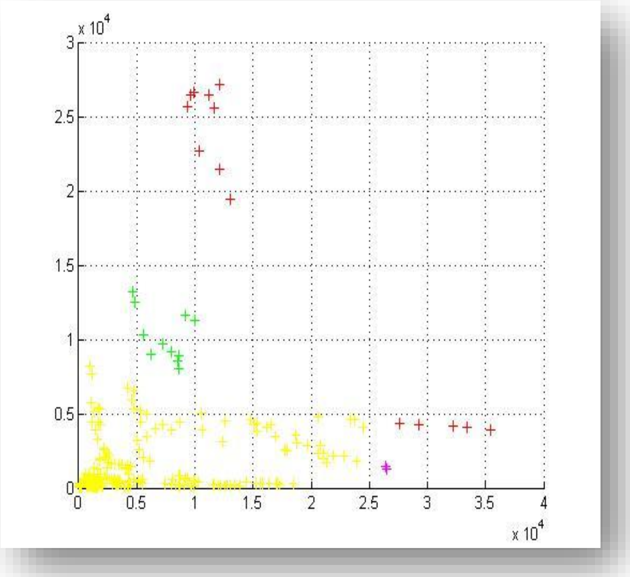


Figure 3.15 . Representing S.C.A result with 4 Clusters

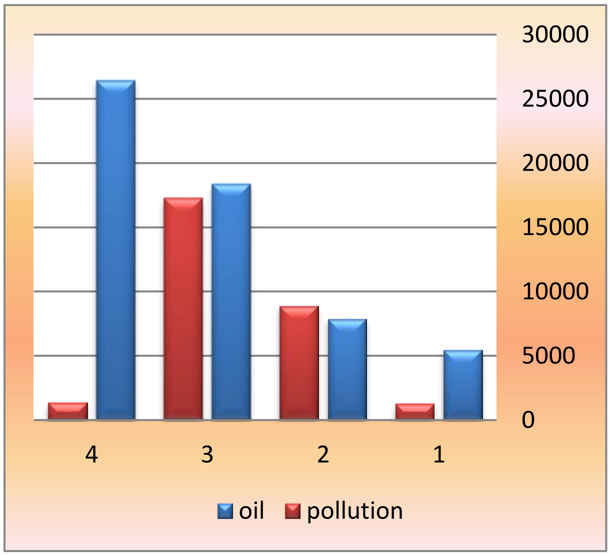


Figure 3.16. Graph result for S.C.A with k=4

Result S. C.A with K= 5

We note that the representation is not integrated. The reason is the fourth CLUSTER, where the percentage of pollution has increased. The increase in pollution may be caused by other factors, as in Figure 3.17.

No. Clusters	AVG.Pollution	AVG Oil	Avg Str
CLUSTER **1**	1550	14254	3046.294
CLUSTER **2**	846.55	1761.75	1384.536
CLUSTER**3**	15036.6	12255.51	8572.648
CLUSTER **4**	4574.183	5884.38	1341.554
CLUSTER**5**	3525	25902.21	1822.175

Figure 3.17 two Avg Values (pollution And oil) for S.C.A (K=5)

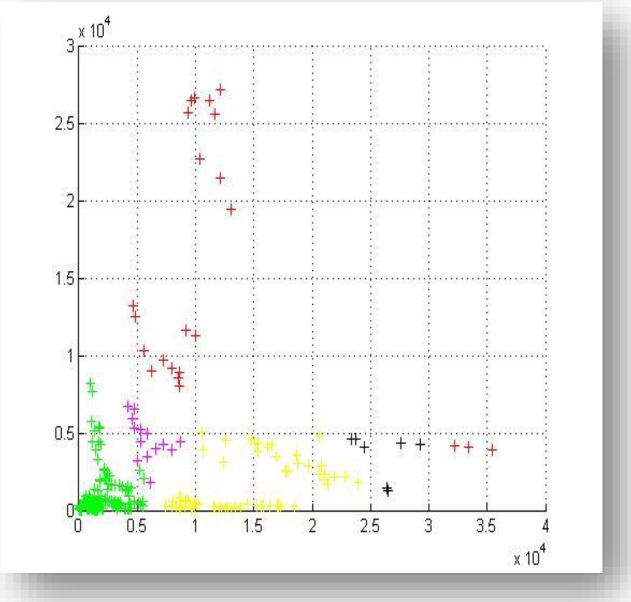


Figure 3.18. Representing S.C.A result with 5 Clusters

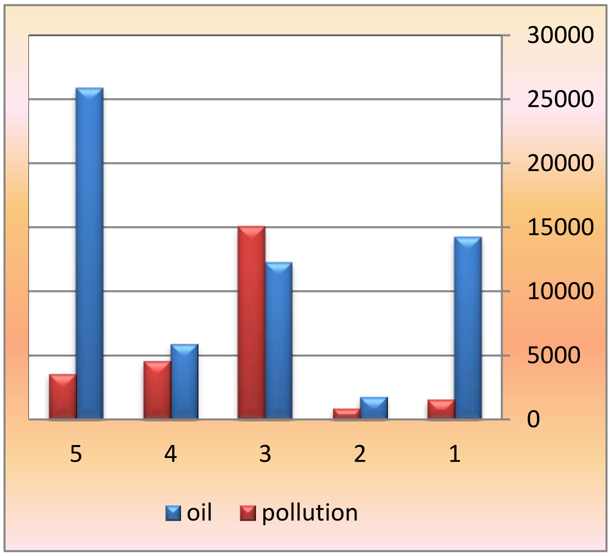
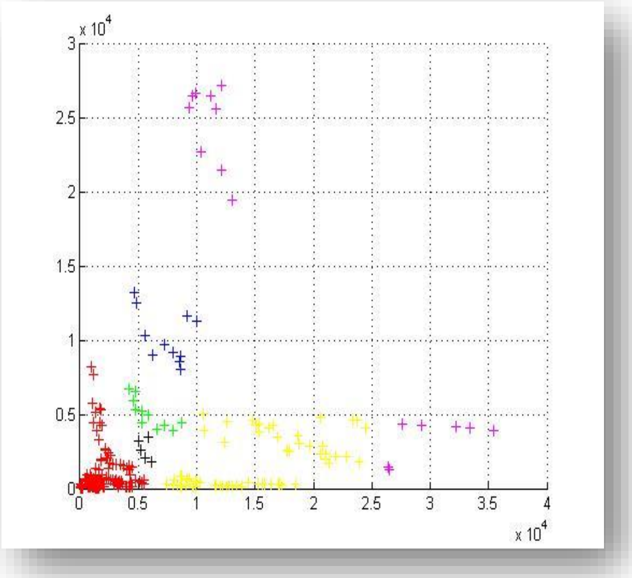


Figure 3.19 Graph result for S.C.A with k=5

Result S. C.A with k= 6

No. clusters	AVG.Pollution	AVG. oil	Avg. Str
CLUSTER **1**	1684	14697	3255.952
CLUSTER **2**	5053	5938.9	1247.022
CLUSTER**3**	828.5	1717.17	1358.242
CLUSTER**4**	15309.58	19383.9	10593.39
CLUSTER**5**	2611.642	5559.76	591.914
CLUSTER**6**	10204.45	7407.964	1772.304

figure 3.20. two Avg Values (pollution and oil) for S.C.A (K=6)



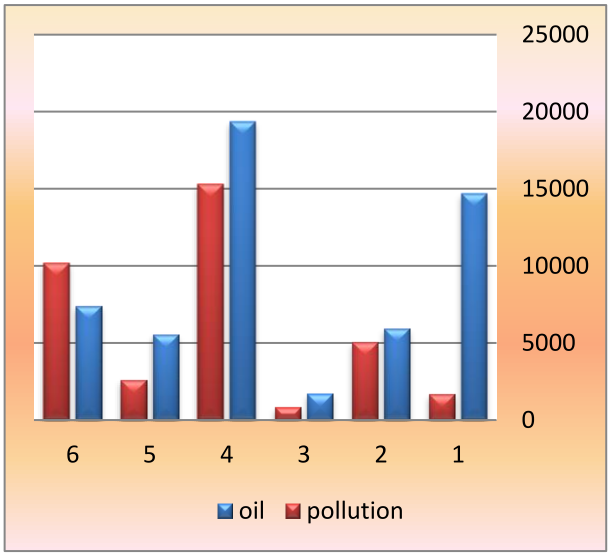


figure 3.21. Representing S.C.A result with 6 Clusters

figure 3.22. Graph result for S.C.A with k=6

3.5.2 Results of K-mean Algorithm (Data set of Oil and Pollution)

K-means will be executed more than once to obtain the best result.

- **Result k-mean with *K=3***

Through implementation, we obtained good classification only. However, there was no logical correlation relation between them, as in Figure **3.23.**

No. Clusters	AVG.Pollution	AVG Oil	Avg.Str
CLUSTER **1**	6837.37	19040.65	7488.683
CLUSTER **2**	828.5	1717.17	1358.242
CLUSTER **3**	3452	8768.747	3324.299

Figure 3.23. Avg. Values (Pollution and Oil) for k-mean (K=3)

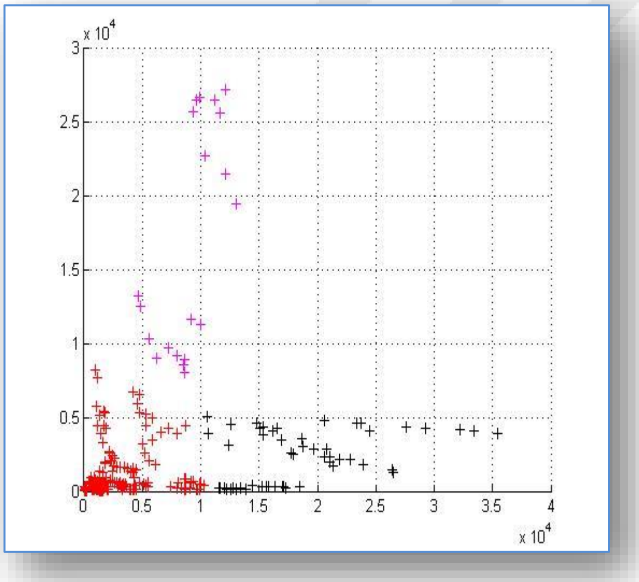


figure 3.24. Representing k-mean with 3 Clusters

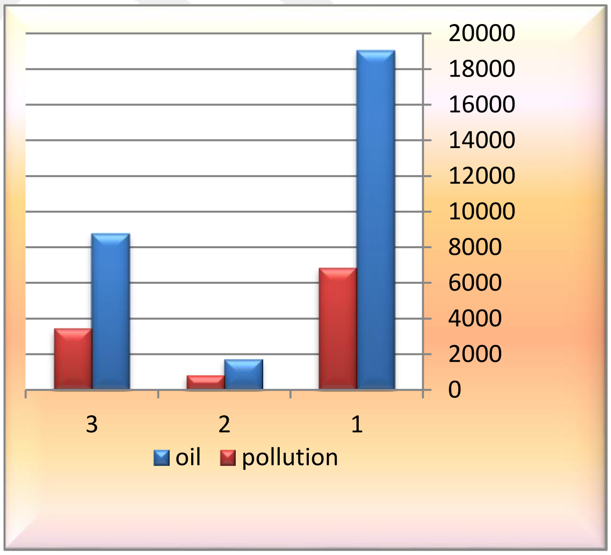


figure 3.25. Representing k-mean result with k= 3

Result k-mean with *K=4*

Through implementation, we obtained good classification and correlation relation, as in Figure **3.26.**

No. clusters	AVG.Pollution	AVG Oil	Avg.str
CLUSTER **1**	2627.288	20899.46	3407.209
CLUSTER **2**	24609.23	11037	2021.371
CLUSTER**3**	828.5	1717.17	1358.242
CLUSTER**4**	3522.8	8761	3305.765

figure 3.26 Avg. Values (Pollution and Oil) for k-mean (K=4)

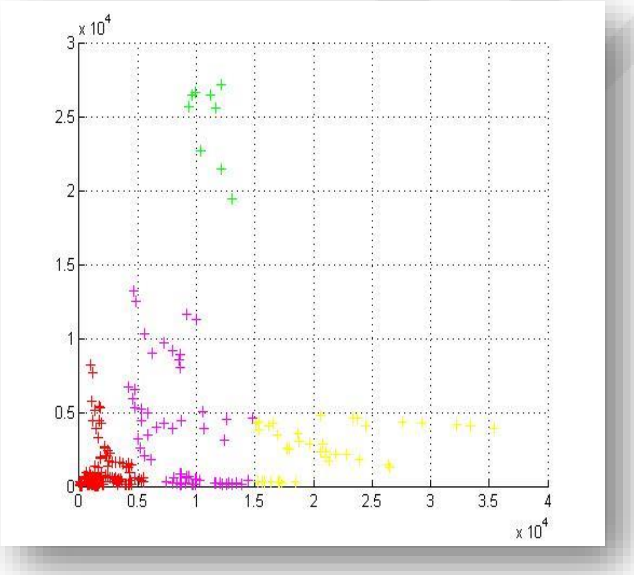


figure 3.24. Representing k-mean with 4 Clusters

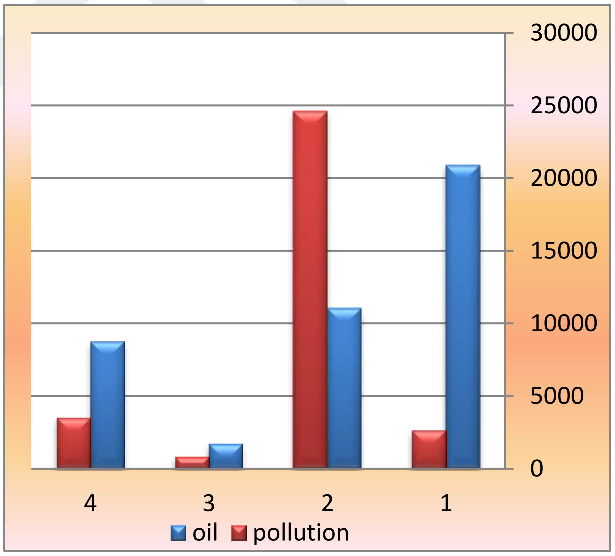


figure 3.25. Representing k-mean result with k= 4

Result k-mean with k=5

Through implementation, we obtained a good classification. However, there was no logical correlation relation between them, as in Figure **3.26**.

No. clusters	AVG.Pollution	AVG oil	Avg.str
CLUSTER **1**	678.3257	1802.369	1192.084
CLUSTER **2**	24609.23	22037	2021.371
CLUSTER**3**	3107.55	24244.42	2962.641
CLUSTER**4**	1291.49	12655	2484.558
CLUSTER**5**	7131.5	5669.5	2715.822

Table 3.26 (Pollution &Oil) for k-mean (K=5)

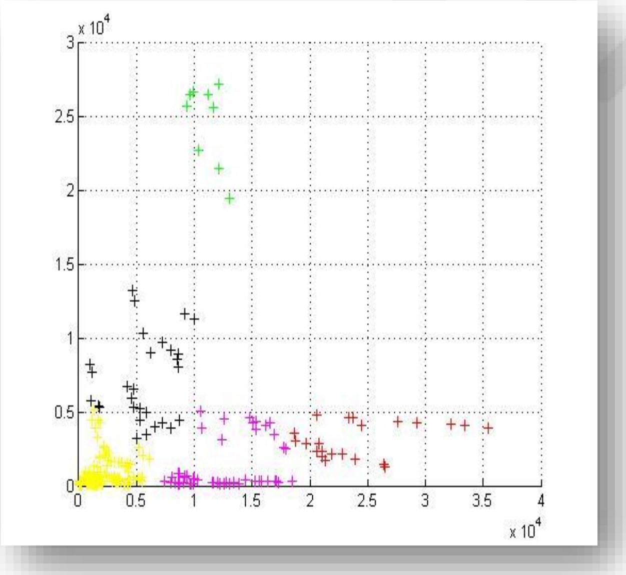


figure 3.27. Representing k-mean with 5 Clusters

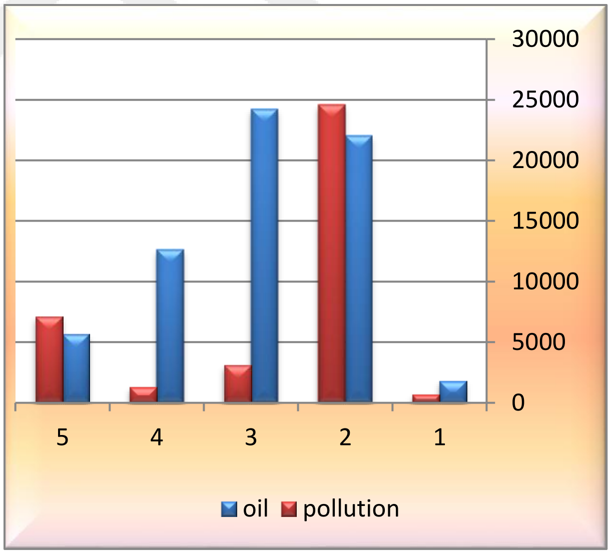


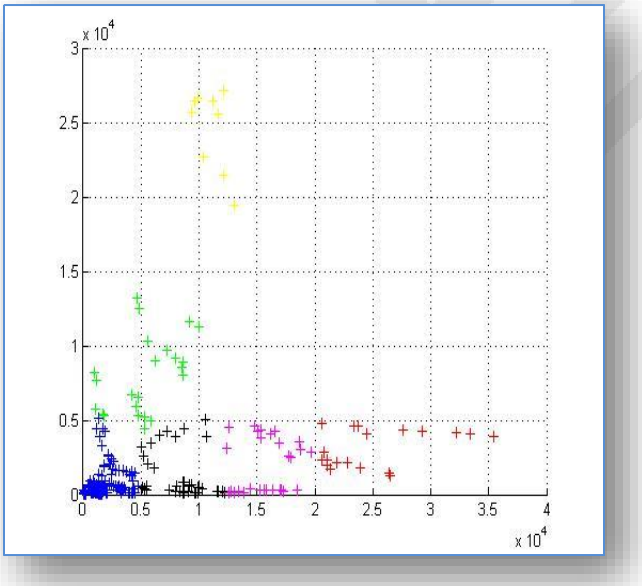
figure 3.28. Representing k-mean result with k= 5

Result k-mean at k=6

Through implementation, we obtained good classification only, but there was no logical correlation relation between them, as in Figure **8**.

No. clusters	AVG.Pollution	AVG oil	Avgstr
CLUSTER **1**	342.7118	970.72	505.1815
CLUSTER **2**	1350.39	12965.68	2431.155
CLUSTER**3**	24609	11037	2021.371
CLUSTER**4**	7965	6231	2599.006
CLUSTER**5**	1635	3690	1599.214
CLUSTER**6**	3107	24244	2962.641

Table 3.29 Avg. Values (pollution and Oil) for k-mean (K=6)



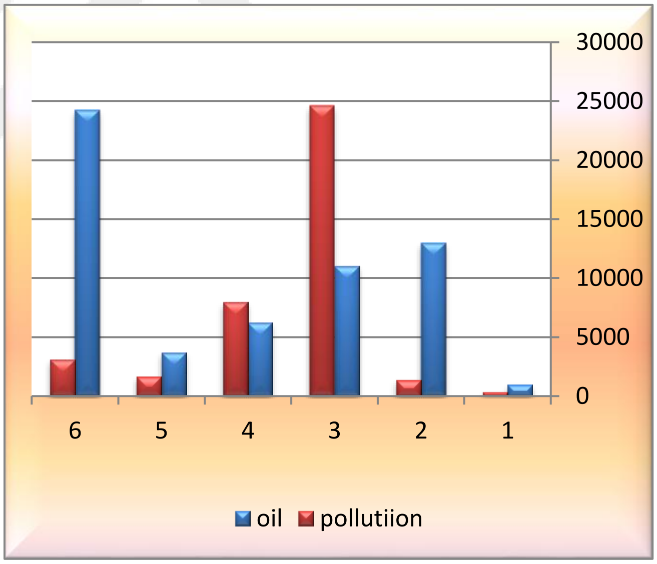


figure 3.27. Representing k-mean with 6 Clusters

figure 3.28. Representing k-mean result with k= 6

3.6 Evaluation of Results.

After the algorithm is applied to several attempts, we acquired the best results when **k = 3.**

The classification is good and there is also a correlation. It also follows the average standard deviation Table 3.19. We find the lowest value for cluster 3 and cluster 5. It is known that the lower the value of the standard deviation, the better the grouping of data and the better the classification.

No cluster. K-means	AVG Str
CLUSTER **3**	4057.075
CLUSTER **4**	2523.147
CLUSTER 5	2275.295
CLUSTER**6**	2019.761

figure 3.29 . number of k and stander deviation

No cluster.	AVG Str
CLUSTER **3**	5436.77
CLUSTER **4**	3947.505
CLUSTER 5	3233.441
CLUSTER**6**	3136.47

Figure 3.30 . number of k and stander eviation

When k-mean, k=3

We note that through our implementation of data (pollution and petroleum), the best results were obtained with k-means in many ways:

1 - in terms of the relationship between quantities of emissions and quantities of petroleum consumption when they were k = 3.

petroleum consumption, when they were k = 3, the relationship is the greater, The consumption of petroleum as pollution increases, as we have noticed in the following figure:

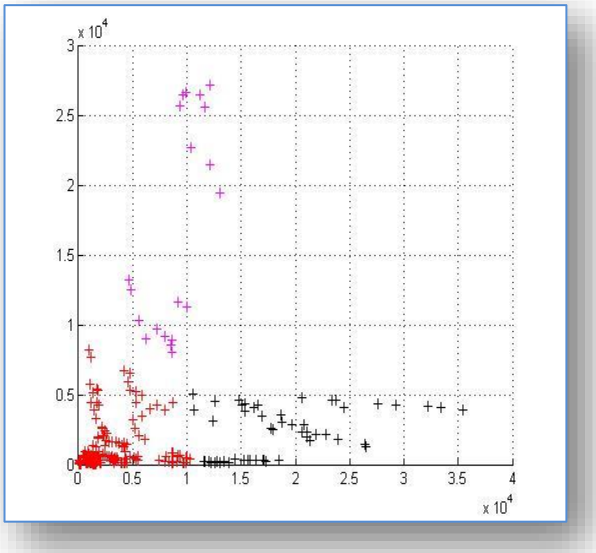


Figure 3.31 Sub Optimal clusters with k=3

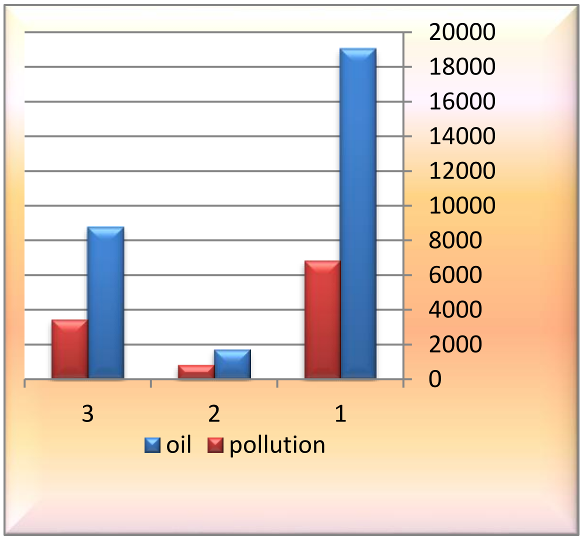


Figure 3.32 BEST Correlation relation. k=3

1. Shapes: One of the defects of k-means that this algorithm favors circular shapes. We note that the clustering has not been organized.
2. Size: We note a clear difference in the size of the cluster.
3. The average standard deviation for the value 4057 is greater than the remainder of the cluster. This indicates that the data representation is not strong.
4. Density: The density of cluster is different from one cluster to another; some clusters have higher densities and other shave lower densities.

   In brief, when it was applied to the data (pollution, petroleum) on the k-means algorithm, the value of k = 3 was very logical and the cognitive pattern was good.

   The main problem was the classification. The classification was not the best of solutions or attempts because of the lack of this algorithm, as mentioned earlier.

When the k-means **k=6 ,** in terms of classification and data representation, the best is k = 6 due to the average standard deviation of 2019.76 being Smaller than the remainder of the cluster. Moreover, it decreases the disadvantages of k-means of density and size, as in Figure 3.20 and higher values for the k-means. This makes the results better on this data.

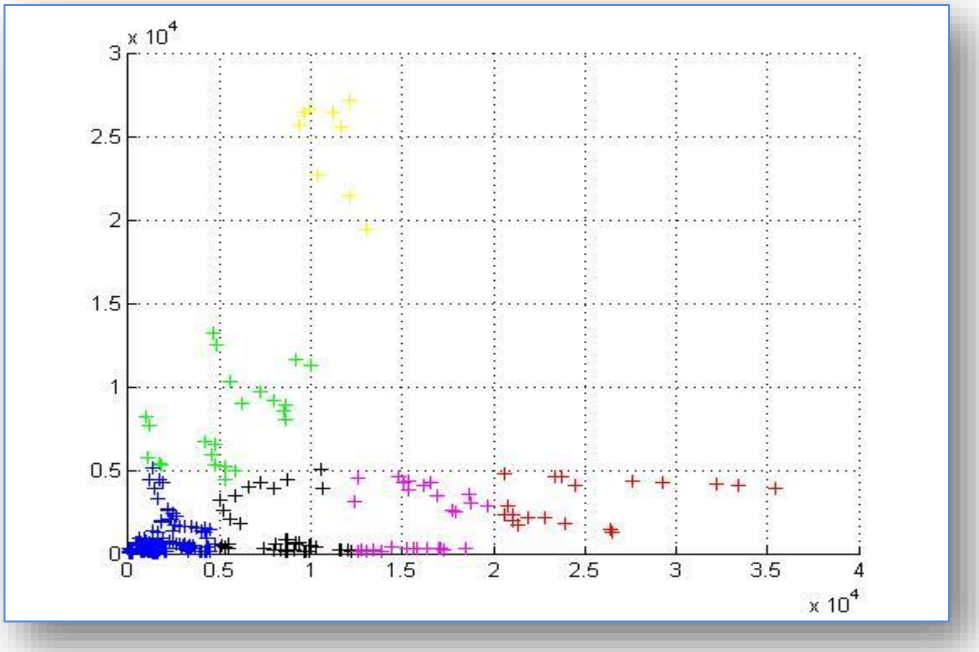


Figure 3.33 Optimal Clusters

4- CONCLUSION

The purpose of this study is to clarify the pros and cons of spectral clustering and to compare results with the real world and apply the truth data and process them before use. We attempted to find a good classification and intelligent style in one of the algorithms and compare the performance of the algorithms. Through our analysis, we see useful results for using the spectral clustering algorithm where determining and arranging sets of any particular statistics or studies in many fields, such as the economy, education, people activities, social studies, criminal fields, finance, industry, health, pollution, environment and oil energy. We also saw their effects on each other. By doing so, we observe through the application of an algorithm, or technique such as regression and correlation that it is difficult to determine the independent or dependent variable, or determine which one is linear or nonlinear. The thesis summarizes the application of spectral and quantification of four variables based on real data from an official source in the 31 countries in the European Union in the seven years from 2006 to 2014. The first study ranked between tons of pollution and oil production or oil energy, according to quantities of harmonization of pollution and oil. The second study was between the quantities of pollution and the level of consumption of clean energy and oil, which was classified as quantities of pollution and the amount of energy consumed.

This algorithm makes it easier for us to classify countries in terms of quantities of pollution over the years, the magnitude of the oil consumption, and the extent to which these countries develop in terms of oil or oil production.

The result was that there was a direct effect of all these variables on each other by dividing these data into four different groups in their characteristics and specifications.

Analysis and study of the application of the S.C. algorithm showed a strong correlation between the four variables, as follows:

Classification of countries in terms of quantities of pollution and production of clean energy or oil energy to seven groups according to the average of both pollution and oil energy; and

The classification of countries in terms of quantities of pollution and consumption of petroleum also into seven groups according to the average of both pollution and consumption of petroleum.

The first study is the pollution and oil energy production of the EU countries for 2006-2014.

5. Future Recommendations

While executing the algorithm(spectral and k-means) on a large amount of data, we obtained different data and when comparing around, we found that the spectral clustering was better thanks to the k-means ,because k-means depends on the center (centroid).Therefore, in any future study, the best comparison can be made between the spectra clustering with the Bisecting k-means algorithm for clustering, such as Bisecting k-means, which results in less damaging defects.

We also have difficulty determining the sigma value and k clusters, which is heavily influenced by the data (depending on the resolution).

Note. Bisecting k-means a combination of k-means and hierarchical clustering.

It starts with all objects in a single cluster.

REFERENCES

Abubaker, M., & Ashour, W. (2013). Efficient data clustering algorithms: Improvements over Kmeans. *International Journal of Intelligent Systems and Applications, 5*(3), 37.

Ackerman, M. (2012). Towards Theoretical Foundations of Clustering.

Alfraih, A. S. (2015). *Feature extraction and clustering techniques for digital image forensics.* University of Surrey (United Kingdom).

Cameron, E. T. (2012). Optimal clustering techniques for metagenomic sequencing data.

Farmani, M. R. (2016). *Clustering analysis using Swarm Intelligence.* Universita'degli Studi di Cagliari.

Gordon, S. (2003). *Unsupervised image clustering using probabilistic continuous models and information theoretic principles*: Tel-Aviv University.

Heller, K. A. (2008). *Efficient Bayesian methods for clustering*: University of London, University College London (United Kingdom).

Kumar, D. (2009). *Study On Clustering Techniques And Application To Microarray Gene Expression Bioinformatics Data.*

Lämsä, V. (2008). Spectral/K-means clustering.

Nemala, V. (2009). Efficient clustering techniques for managing large datasets.

Panda, M., Hassanien, A. E., & Abraham, A. (2017). Hybrid Data mining approach for image segmentation based Classification *Biometrics: Concepts, Methodologies, Tools, and Applications* (pp. 1543-1561): IGI Global.

Singh, S., & Polous, K. Master Thesis Spatial Temporal Analysis of Social Media Data.

Vejmelka, M. (2009). *Spectral graph clustering.* Paper presented at the Seminar z Umele Inteligence, Prague.

Dhillon, I. S., Guan, Y., & Kulis, B. (2004). *Kernel k-means: spectral clustering and normalized cuts.* Paper presented at the Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.

"*A comparison of document clustering techniques", M. Steinbach, G. Karypis and V. Kumar. Workshop on Text Mining, KDD, 2000.*

APPENDIX A

SOURCE CODE FOR spectral cluster algorithm

We used Mat lab version 14 to implement the algorithm consists of the 4 programs

1. main program (Jordan_Weiss1)
2. read data from an external file (Gen_Data1)
3. Calculate Affinity matrix (Cal_Affinty1)

Firstly, main program (Jordan_Weiss1)

```
clear all;
close all;
dim=2
data = Gen_Data1;
affinty1 = Cal_Affinty1(data1, dim);
for i=1: size(affinty1,1)
D(i,i) = sum(affinty1(i,:));
end
for i=1: size(affinity,1)
for j=1: size(affinty1,2)
NL1(i,j) = affinty1(i,j) / (sqrt(D(i,i)) * sqrt(D(j,j)));
end
end
[eigVector, eigValues1] = eig1(NL1);
k =7
nEigVec = eigVector (:(size(eigVector,1) -(k-1)): size(eigVector,1));
for i=1: size(nEigVec,1)
n = sqrt (sum (nEigVec(i, :).^2));
U (i, :) = nEigVec(i,:) ./ n;
end
[IDX,C,sumb,d] = kmeans(U,3);
```

```
%data = display1(IDX,data,dim);
if dim==2
figure, plot3(data (:1), data (:2), data(:,3),'r+'), title('Original Data Points'); grid on;
figure, plot3(data (1,1), data (1,2), data(10,3),'w+'); grid on;
hold on 54
for i=1:size(IDX,1)
if IDX(i,1) == 1
plot3(data(i,1),data(i,2),data(i,3),'y+');
data(i,4)=1
elseif IDX(i,1) == 2
plot3(data(i,1),data(i,2),data(i,3),'g+');
data(i,4)=2
elseif IDX(i,1) == 3
plot3(data(i,1),data(i,2),data(i,3),'r+');
data(i,4)=3
elseif IDX(i,1) == 4
plot3(data(i,1),data(i,2),data(i,3),'b+');
data(i,4)=4
elseif IDX(i,1) == 5
plot3(data(i,1),data(i,2),data(i,3),'m+');
data(i,4)=5
elseif IDX(i,1) == 6
plot3(data(i,1),data(i,2),data(i,3),'k+');
data(i,4)=6
end
end
elseif dim==2
figure,plot(data(:,1), data(:,2),'b+'), title('Original Data Points'); grid on;shg
figure
```

```
hold on
for i=1:size(IDX,1)
if (IDX(i,1) == 1)
plot(data(i,1),data(i,2),'r+');
data(i,3)=1
elseif IDX(i,1) == 2
plot(data(i,1),data(i,2),'k+');
data(i,3)=2
elseif IDX(i,1) == 3
plot(data(i,1),data(i,2),'m+');
data(i,3)=3
elseif IDX(i,1) == 4
plot(data(i,1),data(i,2),'r+');
data(i,3)=4
elseif IDX(i,1) == 5
plot(data(i,1),data(i,2),'m+');
data(i,3)=5
else
plot(data(i,1),data(i,2),'y+');
data(i,3)=6 55
end
end
end
xlswrite('D:\CLUTUR AND CRIM\NEW.xlsx',data,'sheet6');
hold off;
grid on;
tt=1;
```

secondly read data from an external file (Gen_data)

```
function [data] = GenerateData()
```

```
data=xlsread('E:\kemerburgaz\thises\data_try\input_data.xlsx','sheet1','a:c');
end
```

thirdly calculate Affinity matrix (Cal_Affinity)

```
function [affinity] = Cal_Affinity(data,di)
sigma =0.572
if di==3
for i=1:size(data,1)
for j=1:size(data,1)
for K=1:size(data,1)
dist = sqrt((data(i,1) - data(j,1) )^2 + (data(i,2) - data(j,2))^2+ (data(i,3)- data(j,3) )^2 );
affinity(i,j) = exp(-dist/(2*sigma^2));
else
for i=1:size(data,1)
for j=1:size(data,1)
dist = sqrt((data(i,1) - data(j,1) )^2 + (data(i,2) - data(j,2))^2);
affinity(i,j) = exp(-dist/(2*sigma^2));
end
```

APPENDIX B

The table below consists of data representing the first study conducted using algorithms and their results.

country	year	POULATION	Oil	S.C.A
Belgium	2006	133,599	1,364.6	1
Bulgaria	2006	765,131	1,140.5	1
Czech Republic	2006	203,163	2,213.1	1
Denmark	2006	29,958	2,889.1	1
Germany	2006	476,336	20,580.7	3
Estonia	2006	69,900	530.8	1
Ireland	2006	62,955	423.1	1
Greece	2006	533,230	1,781.6	1
Spain	2006	1,161,121	9,163.9	4
France	2006	438,065	15,314.0	3
Croatia	2006	54,286	1,732.5	1
Italy	2006	384,948	15,326.5	3
Cyprus	2006	31,471	55.8	1
Latvia	2006	8,609	1,430.1	1
Lithuania	2006	30,517	930.9	1
Luxembourg	2006	2,819	75.8	1
Hungary	2006	39,305	1,231.8	1
Malta	2006	11,495	0.6	1
Netherlands	2006	64,140	2,414.1	1
Austria	2006	27,120	7,373.3	7
Poland	2006	1,321,241	4,694.6	4
Portugal	2006	151,889	4,212.1	1
Romania	2006	654,533	4,780.9	6
Slovenia	2006	16,629	767.7	1
Slovakia	2006	87,785	834.8	1

Finland	2006	84,592	8,690.9	7
Sweden	2006	35,891	14,388.2	7
United Kingdom	2006	669,948	4,202.5	6
Norway	2006	21,143	11,562.3	7
Turkey	2006	2,269,951	10,359.1	2
Belgium	2007	124,092	1,590.7	1
Bulgaria	2007	820,614	961.6	1
Czech Republic	2007	208,654	2,320.0	1
Denmark	2007	27,268	3,206.0	1
Germany	2007	459,942	23,728.1	3
Estonia	2007	88,034	601.6	1
Ireland	2007	56,936	477.9	1
Greece	2007	537,944	1,726.8	1
Spain	2007	1,124,701	10,007.5	4
France	2007	422,582	16,550.3	3
Croatia	2007	59,545	1,537.1	1
Italy	2007	342,661	16,946.0	3
Cyprus	2007	29,423	73.3	1
Latvia	2007	8,140	1,406.8	1
Lithuania	2007	26,761	964.2	1
Luxembourg	2007	2,389	127.8	1
Hungary	2007	34,929	1,366.6	1
Malta	2007	11,807	0.8	1
Netherlands	2007	60,592	2,587.1	1
Austria	2007	24,151	7,938.3	7
Poland	2007	1,254,293	4,823.8	4
Portugal	2007	144,906	4,480.5	1
Romania	2007	528,158	4,748.2	6
Slovenia	2007	14,848	726.7	1
Slovakia	2007	70,595	942.7	1
Finland	2007	82,930	8,657.3	7

Sweden	**2007**	**32,674**	**15,293.7**	**7**
United Kingdom	**2007**	**589,077**	**4,530.2**	**6**
Norway	**2007**	**20,097**	**12,806.1**	**7**
Turkey	**2007**	**2,647,732**	**9,603.9**	**2**
Belgium	**2008**	**96,376**	**1,910.2**	**1**
Bulgaria	**2008**	**571,696**	**1,062.2**	**1**
Czech Republic	**2008**	**168,589**	**2,518.4**	**1**
Denmark	**2008**	**21,187**	**3,247.9**	**1**
Germany	**2008**	**460,367**	**23,352.1**	**3**
Estonia	**2008**	**69,477**	**646.1**	**1**
Ireland	**2008**	**47,499**	**575.0**	**1**
Greece	**2008**	**445,156**	**1,709.5**	**1**
Spain	**2008**	**503,189**	**10,552.3**	**7**
France	**2008**	**355,773**	**18,619.8**	**3**
Croatia	**2008**	**53,101**	**1,631.2**	**1**
Italy	**2008**	**288,137**	**19,706.9**	**3**
Cyprus	**2008**	**22,433**	**94.7**	**1**
Latvia	**2008**	**6,674**	**1,377.6**	**1**
Lithuania	**2008**	**22,543**	**1,021.0**	**1**
Luxembourg	**2008**	**1,747**	**132.9**	**1**
Hungary	**2008**	**35,419**	**1,587.9**	**1**
Malta	**2008**	**10,778**	**0.9**	**1**
Netherlands	**2008**	**50,724**	**3,008.3**	**1**
Austria	**2008**	**21,810**	**8,360.5**	**7**
Poland	**2008**	**1,032,167**	**5,559.9**	**4**
Portugal	**2008**	**95,603**	**4,329.4**	**1**
Romania	**2008**	**524,774**	**5,343.3**	**6**
Slovenia	**2008**	**13,047**	**837.9**	**1**
Slovakia	**2008**	**69,447**	**938.1**	**1**
Finland	**2008**	**70,136**	**9,141.4**	**7**
Sweden	**2008**	**30,501**	**15,619.9**	**7**

United Kingdom	2008	491,595	5,851.7	6
Norway	2008	20,027	13,366.0	7
Turkey	2008	2,561,412	9,311.9	2
Belgium	2009	74,234	2,268.2	1
Bulgaria	2009	443,750	1,111.3	1
Czech Republic	2009	165,748	2,842.1	1
Denmark	2009	15,577	3,299.9	1
Germany	2009	411,369	24,481.3	3
Estonia	2009	54,876	717.5	1
Ireland	2009	34,393	663.0	1
Greece	2009	425,553	1,865.7	1
Spain	2009	452,139	12,569.4	7
France	2009	305,602	18,749.6	3
Croatia	2009	55,743	1,829.0	1
Italy	2009	235,689	21,026.2	3
Cyprus	2009	17,742	100.3	1
Latvia	2009	6,447	1,566.7	1
Lithuania	2009	21,245	1,052.3	1
Luxembourg	2009	1,781	124.7	1
Hungary	2009	29,918	1,836.8	1
Malta	2009	8,003	0.9	1
Netherlands	2009	37,393	3,276.5	1
Austria	2009	16,407	8,736.0	7
Poland	2009	899,224	6,244.7	4
Portugal	2009	61,071	4,785.8	1
Romania	2009	444,827	5,268.7	6
Slovenia	2009	10,820	1,079.2	1
Slovakia	2009	64,113	1,131.1	1
Finland	2009	59,379	8,051.5	7
Sweden	2009	29,618	15,819.1	7
United Kingdom	2009	400,666	6,566.7	6

Norway	2009	15,494	12,165.9	7
Turkey	2009	2,665,176	9,916.1	2
Belgium	2010	60,615	2,832.6	1
Bulgaria	2010	388,794	1,456.6	1
Czech Republic	2010	160,272	3,130.1	1
Denmark	2010	15,587	3,919.5	1
Germany	2010	432,228	27,570.7	5
Estonia	2010	83,273	846.5	1
Ireland	2010	28,257	662.7	1
Greece	2010	265,410	2,131.3	1
Spain	2010	421,131	15,047.5	3
France	2010	286,215	20,801.1	3
Croatia	2010	34,743	2,064.5	1
Italy	2010	216,952	21,864.4	3
Cyprus	2010	21,919	105.1	1
Latvia	2010	4,517	1,434.5	1
Lithuania	2010	21,120	1,064.6	1
Luxembourg	2010	1,756	128.5	1
Hungary	2010	31,353	1,953.9	1
Malta	2010	8,090	5.2	1
Netherlands	2010	33,800	3,188.9	1
Austria	2010	17,901	9,107.9	7
Poland	2010	969,538	7,269.1	4
Portugal	2010	53,276	5,459.4	1
Romania	2010	349,465	5,860.4	6
Slovenia	2010	10,138	1,121.0	1
Slovakia	2010	71,618	1,324.8	1
Finland	2010	66,957	9,352.1	7
Sweden	2010	32,131	16,996.8	7
United Kingdom	2010	422,975	7,277.3	6
Norway	2010	19,679	11,675.0	7

Turkey	2010	2,561,012	11,627.0	2
Belgium	2011	53,015	3,119.8	1
Bulgaria	2011	516,175	1,362.6	1
Czech Republic	2011	160,410	3,440.0	1
Denmark	2011	14,375	4,001.8	1
Germany	2011	427,994	29,300.5	5
Estonia	2011	72,768	835.4	1
Ireland	2011	26,646	767.3	1
Greece	2011	262,162	2,139.9	1
Spain	2011	457,337	14,832.2	3
France	2011	249,664	17,892.2	3
Croatia	2011	28,835	1,704.1	1
Italy	2011	194,886	21,025.4	3
Cyprus	2011	20,919	121.3	1
Latvia	2011	4,303	1,417.3	1
Lithuania	2011	24,278	1,056.6	1
Luxembourg	2011	1,296	125.2	1
Hungary	2011	34,192	1,887.4	1
Malta	2011	7,921	8.4	1
Netherlands	2011	33,463	3,420.8	1
Austria	2011	16,797	8,623.7	7
Poland	2011	917,062	7,936.5	4
Portugal	2011	48,065	5,137.9	1
Romania	2011	320,090	5,068.1	6
Slovenia	2011	11,890	1,039.4	1
Slovakia	2011	68,724	1,292.8	1
Finland	2011	61,112	9,143.7	7
Sweden	2011	29,228	16,324.2	7
United Kingdom	2011	392,774	8,009.2	6
Norway	2011	18,794	11,998.3	7
Turkey	2011	2,640,511	11,222.4	2

Belgium	2012	47,309	3,366.1	1
Bulgaria	2012	329,980	1,627.2	1
Czech Republic	2012	154,692	3,687.9	1
Denmark	2012	12,634	4,183.1	1
Germany	2012	412,669	32,251.7	5
Estonia	2012	40,626	861.5	1
Ireland	2012	25,182	779.4	1
Greece	2012	244,900	2,447.8	1
Spain	2012	404,201	16,134.8	3
France	2012	235,324	20,617.2	3
Croatia	2012	24,827	1,747.6	1
Italy	2012	176,448	23,885.2	3
Cyprus	2012	16,222	129.1	1
Latvia	2012	4,286	1,651.4	1
Lithuania	2012	20,995	1,161.4	1
Luxembourg	2012	1,500	138.3	1
Hungary	2012	31,213	1,772.0	1
Malta	2012	7,738	11.7	1
Netherlands	2012	33,786	3,616.0	1
Austria	2012	16,123	9,722.0	7
Poland	2012	892,014	8,610.0	4
Portugal	2012	43,030	4,354.2	1
Romania	2012	257,679	5,195.2	6
Slovenia	2012	10,842	1,069.1	1
Slovakia	2012	57,480	1,358.7	1
Finland	2012	51,437	9,990.6	7
Sweden	2012	28,339	18,524.4	7
United Kingdom	2012	439,147	8,750.0	6
Norway	2012	17,300	13,826.3	7
Turkey	2012	2,715,911	12,151.0	2
Belgium	2013	44,702	3,504.4	1

Bulgaria	2013	195,832	1,813.6	1
Czech Republic	2013	137,909	4,049.9	1
Denmark	2013	13,012	4,315.7	1
Germany	2013	410,425	33,397.4	5
Estonia	2013	36,533	851.2	1
Ireland	2013	25,389	842.7	1
Greece	2013	226,513	2,616.3	1
Spain	2013	258,671	17,744.0	3
France	2013	217,193	22,775.5	3
Croatia	2013	16,507	2,081.9	1
Italy	2013	145,140	26,370.6	5
Cyprus	2013	13,756	134.3	1
Latvia	2013	3,850	1,611.2	1
Lithuania	2013	19,821	1,212.4	1
Luxembourg	2013	1,579	154.5	1
Hungary	2013	30,459	1,866.7	1
Malta	2013	5,028	12.3	1
Netherlands	2013	29,596	3,500.4	1
Austria	2013	15,877	9,952.8	7
Poland	2013	853,438	8,568.9	4
Portugal	2013	38,743	5,300.5	1
Romania	2013	202,760	5,550.9	6
Slovenia	2013	11,590	1,174.1	1
Slovakia	2013	53,474	1,409.3	1
Finland	2013	47,379	9,912.2	7
Sweden	2013	26,802	17,082.6	7
United Kingdom	2013	385,982	10,635.4	7
Norway	2013	16,664	12,596.0	7
Turkey	2013	1,939,104	13,061.9	2
Belgium	2014	42,253	3,397.8	1
Bulgaria	2014	188,937	1,788.8	1

Czech Republic	2014	126,953	4,176.3	1
Denmark	2014	11,419	4,466.8	1
Germany	2014	388,034	35,406.3	5
Estonia	2014	40,839	858.9	1
Ireland	2014	19,343	961.4	1
Greece	2014	138,109	2,445.0	1
Spain	2014	254,614	17,768.3	3
France	2014	169,376	21,354.3	3
Croatia	2014	15,559	2,007.6	1
Italy	2014	130,522	26,512.2	5
Cyprus	2014	16,795	132.5	1
Latvia	2014	3,795	1,613.1	1
Lithuania	2014	17,832	1,277.0	1
Luxembourg	2014	1,589	189.4	1
Hungary	2014	27,146	1,885.1	1
Malta	2014	4,671	17.7	1
Netherlands	2014	29,060	3,399.7	1
Austria	2014	16,019	9,603.4	7
Poland	2014	800,101	8,608.7	4
Portugal	2014	34,841	5,513.0	1
Romania	2014	175,827	6,124.2	6
Slovenia	2014	8,816	1,202.4	1
Slovakia	2014	45,273	1,420.0	1
Finland	2014	43,556	10,301.0	7
Sweden	2014	23,973	17,317.6	7
United Kingdom	2014	307,638	12,357.4	7
Norway	2014	16,633	13,058.8	7
Turkey	2014	2,147,499	12,080.5	2