

**Spectral Clustering of
Economic Data**

by

Farag Hamed Ali Kuwil

Master degree, Electric and Computer Engineer, 2017

Submitted to the Graduate Faculty of
Science in partial fulfillment
of the requirements for the degree of
Master of Electric and Computer Engineer

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) _____

Prof. Dr. Osman Nuri UÇAN (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 Aydın

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.

Farag Hamed Ali Kuwil

[Signature]

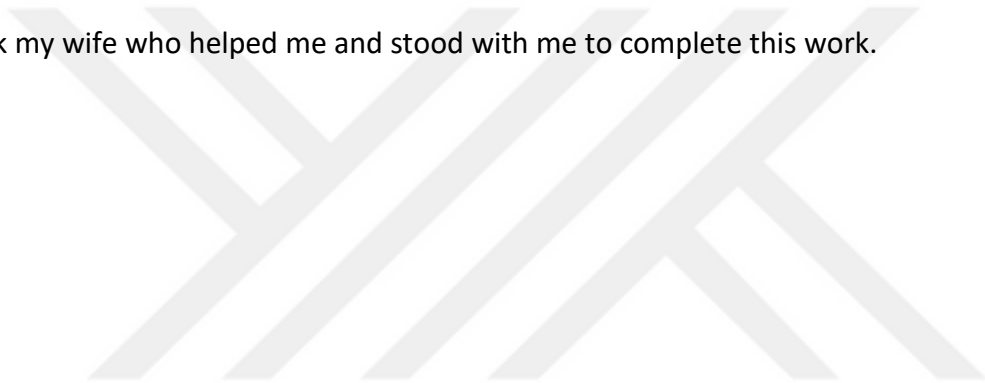
DEDICATION

My parents

First, I hope that Allah mercy them.

I could never have done this without faith, support and constant encouragement, thank you for teaching me to believe in Allah, in myself, in getting knowledge.

Thank my wife who helped me and stood with me to complete this work.



ACKNOWLEDGEMENTS

First and foremost, I thank my academic Professor Doç. Dr. Oğuz Bayat and Assist. Prof. Dr.Hüseyin Afşer, which help me to complete this thesis, Also for the jury

I also would like to thank all the professors' staff of university who taught me and gave me knowledge at the university.

I would like to thank all the people who contributed in some way to the work described in this thesis. And for everyone from friends or colleagues who helped me complete this work and who give a hand, even if give advice and make suggestions.

ABSTRACT

Spectral Clustering Algorithm

[Author's Kuwil, Farag],

M.S/Electronic and computer engineering/Altinbas University

Supervisor: Doç. Dr. Oğuz BAYAT

Date: 4/2017

In this thesis were implemented two of the most important algorithms in the data mining spectral clustering and k_mean ,try to find the variances between them where the study was conducted on the statistics and studies from data warehouse EEA (European Economic Association) in thirty one countries in the European Union in the seven years from 2008 to 2013, which was The first study classifies the People at risk of poverty or social exclusion, while the second one is Population tertiary by education attainment level And the last is Mean and median income by work intensity of the household.

In the first phase has been calibrated and format of the data in order to apply the algorithm, it has also been making some adjustments and guess some of the lost values of statistics according to the laws of statistics, The second stage, so we execute the algorithm program which wrote in matlab language version no14, then the results represented in 3D

Finally all the results analyzed, therefore the benefits and advantages, Find the relationship between three factors and which one effect to other then clustering the thirty one countries with year of studies in three different clusters.

In rich countries where people have well-being life and high living rate, The percentage for risk of poverty decreases and we noted rising the percentage of high education attainment, because people in those countries have Comfort and stability life ,While in the least well-being countries, we note rise the number of people at risk of poverty and decrease percentage educational attainment, so these people be more deal with reality and Approaching to the

obsessions and fears of the future, so they try to be Self-reliance by As a result, turning the labor market for attend to find a job opportunity at early age.



TABLE OF CONTENTS

1.INTRODUCTION	1
1.1CLUSTERING	1
1.2 DATA MINING	3
1.2.1Kinds of data can be mining.....	4
1.2.1.1 Database Data	5
1.2.1.2Data Warehouse	5
1.2.1.3Transactional Data	6
1.2.1.4 Other kind of Data	6
1.2.2 What Kind of patterns can be mined ?.....	6
1.2.2.1 Class/Concept Description.....	7
1.2.2.2 Mining Frequent Patterns, Associations, and Correlations	7
1.2.2.3 Classification and Regression for Predictive Analysis	8
1.2.2.4 Outlier Analysis	8
1.2.2.5 Cluster Analysis.....	8
1.2.3 Methods for Data Mining.....	8
1.2.3.1 Statistics	9
1.2.3.2Machinelearning	10
1.2.3.3 Database Systems and Data Warehouses	11

1.2.3.4 Information Retrieval.....	12
1.3 DATA.....	13
1.3.1 What is data?.....	13
1.3.2 Data point.....	13
1.3.3Data set.....	14
2. CLUSTERING ALGORITHM TECHNIQUES	16
2.1CLUSTER TECHNIQUE.....	17
2.1.1 K_mean cluster technique	17
2.1.2 Spectral clustering technique.....	18
1.2.3 The different between spectral clustering and K-mean technique.....	18
2.2 ALGORITHM.....	19
2.2.1 K_mean algorithm.....	19
2.2.2 Spectral clustering algorithm	20
3. METHODOLOGY	27
3.1 DATASET COLLECTION.....	28
3.1.1 Data collection	29
3.1.2 Data cleaning	29
3.1.3 Data transformation	30
3.1.4 Data integration	32

3.1.5 Data mining	35
3.2 IMPLEMENTATION	35
3.2.1 Stages of implementation	36
3.2.1.1 Load Dataset	36
3.2.1.2 Run Algorithm	37
3.2.1.3 Result Analysis	38
3.3 RESULT AND DISCUSSION.....	40
3.3.1 Result of spectral clustering algorithm	41
3.3.2 Discussion of spectral clustering algorithm	45
3.3.1 Result of k_mean algorithm	46
3.3.2 Discussionof K_mean algorithm	50
4.CONCLOSION	52
4.1 FUTURE WORK.....	54
REFERENCES.....	55
SOURCE CODE	57
DATASET CONTENTS	60

LIST OF TABLES

Table 1. Example of Dataset Attribute in Data Mining	14
Table 2. Dataset Attribute in S.C.A	20
Table3. Read Dataset From Excel File	36
Table 4. Outputs of Implementation S.C.A.....	39
Table 5. Combine Outputs of Implementation with Source Dataset.....	41

LIST OF FIGURES

Figure 1. Affinity Matrix Form in S.C.A	22
Figure 2. Degree Matrix Form in S.C.A	23
Figure3.Normalized Laplacian Matrix Form in S.C.A.....	23
Figure 4. Eigenvalues,Eigenvectors in S.C.A Form in S.C.A.....	24
Figure 5. Neigenvec Form in S.C.A	24
Figure6. Rearrange Neigenvec Form in S.C.A	25
Figure 7. Normalized Matrix U Form in S.C.A.....	25
Figure8.Flowchart of implementation	28
Figure 9.Cleaning Data in dataset 1 (Risk of Poverty).....	29
Figure10. Estimate Some Data in Dataset 1 (Risk of Poverty).....	30
Figure 11. Data Transformation in Dataset 1 (Risk of Poverty).....	31
Figure 12.Data Transformation in Dataset 2 (Education).....	32
Figure 13. Data Transformation in Dataset 3 (Income)	32
Figure 14.Integrate 3Dataset in one Dataset.....	33
Figure 15.Data Patterns.....	33
Figure 16.Dataset (final form of dataset).....	34
Figure 17. Implementation Flowchart	35
Figure 18. Three Avg values S.C.A with K=3	42
Figure 19. Four Avg values S.C.A with K=4	43

Figure 20. Five Avg values for S.C.A with K=5	43
Figure 21. Six Avg values for S.C.A with K=6	44
Figure 22. Sample of the S.C.A result	45
Figure 23. Three Avg values k_mean with K=3	47
Figure 24. Four Avg values k_mean with K=4	48
Figure 25. Five Avg values for k_mean with K=5	48
Figure 26. Six Avg values for k_mean with K=6	49
Figure 27. Standard deviation for all of S.C.A and k_mean	51

LIST OF CHART

Chart 1. Representing Dataset points in 3 D	40
Chart 2. Graph result for S.C.A with k=3	42
Chart 3. Representing S.C.A result with 3 clusters	42
Chart 4. Graph result for S.C.A with k=4	43
Chart 5. Representing S.C.A result with 4 clusters	43
Chart 6. Graph result for S.C.A with k=5	43
Chart 7. Representing S.C.A result with 5 clusters	43
Chart 8. Graph result for S.C.A with k=6	44
Chart 9. Representing S.C.A result with 6 clusters	44
Chart 10. Graph result for k_mean with k=3	47
Chart 11. Representing k_mean result with 3 clusters	47
Chart 12. Graph result for k_mean with k=4	48
Chart 13. Representing k_mean result with 4 clusters	48
Chart 14. Graph result for k_mean with k=5	49
Chart 15. Representing k_mean result with 5 clusters	49
Chart 16. Graph result for k_mean with k=6	49
Chart 17. Representing k_mean result with 6 clusters	50
Chart 18. Representing S.C.A and k_mean result with 6 clusters	50

LIST OF PICTURE

<u>Picture 1.</u> Idea of Clustering Technique.....	16
<u>Picture 2.</u> The Difference Between Spectral Clustering and k_mean	18



LIST OF ABBREVIATIONS

S.C.A Spectral clustering algorithm

EEA European Economic Association



1. INTRODUCTION

1.1 CLUSTERING

Clustering is an essential uncontrolled (unsupervised learning) algorithm. Nowadays Clustering technology exercised in many fields, like mining of data, pattern understanding, machine Learning, signal and image processing. Data spited into different clusters founded on some Standards by one of the clustering algorithms; points in the same cluster are comparable to each other and various in another case. Clustering ways can be almost concise as 2 types: hierarchical algorithm and division algorithm.. The first one is achieved by creating a structure of a tree. Dividing algorithms cluster data into K less than or equal N section (N: number of records or objects in the dataset). Classic algorithms, like k-mean, FCM, EM algorithm, perform perfect when operate context data, but it lead to local improvement when run non-context data.

Graph theory can be used to apply Spectral clustering and the SC algorithms process data as graph's vertex. In result clusters, there is high relationship within cluster and little border among clusters. SC problems actually are graph cut problems, there is many classical graph cut technique , like min cut, attribution cut, normalized cut, minimum/maximum cut etc. In this type of problem ,we can evidence that traditional information is included in eigenvectors of similarity matrix of the data(Chung, 1997; Fiedler, 1973). SC doesn't make any hypothesis on the framework of data. So SC algorithms can locate global optimal results when operate non-convex data (Ding,etal,2012).

A few years ago, S.C.A attracted a lot and many specialists attention, because of the clear and sound theoretical basis and good results are easy to analyze and utilize.(Nascimento& De Carvalho, 2011).Also many years ago, a lot of S.C.A suggested by specialists. Normalized cut Eigenvalues Eigenvector was suggested by (Shi& Malik, 2000). This standard reasons both inside and outside connections, which result in a more confident and stronger clustering result. (Ding,et al Simon, 2001) proposed minimum/maximum cut.

(Ng, et al, 2002) suggested the traditional NJW algorithm. Now S.C.A have been utilized, in many

fields and domains, like as computer vision (Malik et al, 2001; Zhang et al, 2008), integrated circuit design (Alpert& Kahng,1995), load balancing(Driessche& Roose, 1995;Hendrickson & Leland, 1995), biological information (Kluger et al, 2003; Paccanaro et al, 2006), text classification (Xie et al,2009) It can be said that deep research should be done in the in S.C.A .

Spectral clustering algorithm is one of the successful and effective algorithms. There can be no hypotheses and propositions on data, so the algorithms implement perfect on non-convex data and extract the outside optimal outputs. It can provide clustering problems in polynomial time and has a sound theoretical foundation. Even though it has a lot of usefulness, its research is still in the beginner phase. There are many challenges require being studied and researched (Nascimento De Carvalho, 2011).

Parameters choices issue: Firstly, when preparing resemblance array, scale factor or variable σ is chosen by manual. Clustering outputs is influenced greatly by the chosen of σ . It is a significant research trend. Secondly, how can we measure eigenvector and select the eigenvector are accelerated challenge to be simplified. Thirdly, the selection of the number of groups impacts the clustering output immediately. Fourthly, the chosen Laplacian array, 3 Laplacian matrices are aforesaid but these circumstances for ever matrix are completely not obvious..

Semi-supervised S.C.A: Limited prior awareness is not difficult to get and it's a type of supervise acquaintance. Now, a lot of scientists and researchers using prior knowledge in S.C.A to increase clustering effectiveness.

Fuzzy S.C.A: In real attitudes, most objects in many cases do not include the obvious attribute or feature. It is important to execute method of fuzzy in S.C.A to deal with S.C.A. Applied to big data. It is known that S.C.A also takes a long time to implement in addition to a large space of memory. However, large data were the topic of real research. Therefore, we find that applying the algorithm to large data is a valuable and important research. Kernel S.C.A and S.C.A ensemble. If the data is too large, combination and mixed, S.C.A will not be good choice.

This challenge can be solved by mixing kernel method and S.C.A stuff can solve with scale factor selection problem and the inherent accidental and unorganized of S.C.A.

1.2 DATA MINING

(Han, et al, 2011)

Today, people have been linked to technology; it becomes a basic necessity of life, which led to the accumulation of data in large quantities in warehouses, therefore how data mining can provide methods and techniques to extract the knowledge from data either in knowing the past or studying the present as well as predicting the future. Not surprisingly, dealing with data and do some operations, as a very knowledge base subject, is outlined in many alternative ways in which. Even the expression 'data processing' does not extremely gift all the main parts within the image. To learn how mining that gets in the rocks or sand to get gold, we are saying mining for gold rather than sand or rock mining. Analogously, data processing ought to are additional suitably named "knowledge mining from knowledge," that is sadly somewhat long. However, the shorter term, data mining might not mirror the stress on mining from massive amounts of information. Nonetheless, mining could be a vivid term description the method which locates a tiny low set of invaluable nuggets from a kindle deal of staple. Thus, such a name carrying each "data" and "mining" became a preferred selection. Additionally, several different terms have an analogous assuming to data processing, as an example, Knowledge extraction of data, knowledge extraction, data / pattern analysis, archeology data, and data dredging.

a lot of individuals handle data mining is synonymous with another term used popularly, the discovery of data knowledge, or KDD, while others see data mining as just an essential step in the process of knowledge discovery. As an iterative series of next steps:

- Data cleaning (to delete the noise and incompatible data).
- Data integration (we can merge multiple data sources).
- Data selection (data relevant to the analysis task are recovered).
- Data transformation (data are converted and consolidated into forms suitable).

Data mining (main steps where clever techniques are executed to locate data patterns).

Pattern analysis (to locate the really fascinating patterns appearing data supported power measures—see Section one.4.6) data presentation (Where visualization and data illustration

techniques square measure accustomed gift strip-mined data to users) Steps one through 4 square calculate completely various varieties of input preprocessing, wherever data square measure fit for data-mining.

The Step mining information can move with the user or mental object. Fantastic measurement box patterns give the user new data must stay within the mental object.

The previous read explains data-mining mutually step within the data discovery operation, albeit a vital one as a result of it discover patterns which were hidden for analysis. Such as business, media, and within the analysis surroundings, the term data mining is usually accustomed asks the complete data discovery process (perhaps as a result of the term is less than data discovery from data). However, We tend to foster a broad reading of data-mining.

Functionality: Data-mining extraction is a process of extracting wonderful patterns and finding out huge amounts of information. The information root will embrace databases, data Warehouses, the Web, alternative data repositories, or information that square measure streamed into the System dynamically.

1.2.1 Kinds of data can be mining

As a global technology, data processing may be exercised to every quite knowledge as long because of the data square measure purposeful for a destination implementation. The foremost basic types of knowledge for mining-applications square measure information knowledge (Section one 3.1), knowledge warehouse knowledge (Section tw0.3.2), and transactional knowledge (Section three 3.3). The ideas and mechanism given during this thesis specialize in such knowledge. Data processing also can be executed to different types of knowledge (for instance data trends, instructed/series knowledge, graph or networked knowledge, locative knowledge, characters data, multimedia knowledge).

Techniques for mining of those sorts of knowledge square measure concisely introduced in Chapter thirteen. N depth Treatment is taken into account a sophisticated object. Data processing will definitely hold on to adopt new knowledge sorts as they protrude

1.2.1.1 Database Data

An information-system, additionally known as a system of management (DBMS), contain of a group of interconnected information, called info, and a group of code programs to access and organize the information. The source code give technique for outlining data structures and information warehouse; for determine and arrange synchronal, participated, or allocation of access information; and for making confirmed coherence and safety of the data stored regardless of system shatter and tries at not allowed access.

An electronic information service may be an assortment of tables, every of that is allotted a novel. Name. Every table consists of a group of attributes (columns or fields) and frequently stores A large set of tuples (records or rows). Every tuple associate degree exceedingly in a veryrelative table represents an object known by a novel key and delineated by a group of attribute values. A semantic data model, like associate degree entity-relationship (ER) information model, is commonly created for relational databases. The associate degree ER information model represents the info as a group of entities and their relationships.

1.2.1.2 Data Warehouse

A data warehouse essentially merges output data from many sources into one overall database. For instance, in the commercial world, a data warehouse might contain customer information from a company's point-of-sale systems (the cash registers), its website, its mailing lists and its comment cards. Instead of that, it could contain all the facts about employees, including time cards, demographic data, salary information, etc.

By merging all of this data and facts in one area, a company can analyze its customers in a more holistic approach, and to make sure that it has treaded all the information available. Data warehousing also led to easy to access data mining, which is the task of looking for patterns in the data that could facilitation to higher sales and profits.

1.2.1.3 Transactional Data

generally, every object or record in every group actionable information captivity a dealing, like a customer's buy, a booking of flight, or a mouse's clicks by users on an internet pages. A group action generally includes a novel group action identity range and an inventory of the things.

Forming up the group action, like the things purchased within the group action. A Database which contains transactions could have extra information as dataset format, that contain different info associated with the transactions, like item description, info regarding the employee or the branch, and so on.

1.2.1.4 Other Kinds of Data

Besides knowledge base electronic database on-line database computer database electronic information service} data, information warehouse information, and dealings information, there are many different type of knowledge that has multilateral forms and framework and to some extent perfectly different semantic significances.. Such forms of information may be seen in several applications: time related or sequence knowledge (for instance, historical records, securities market information, and time-series and biological sequence data), information streams (for instance, video police work and sensing element information, that square measure ceaselessly transmitted), spatial information (for instance, maps), engineering style information (for instance, the buildings structure, system-parts, or integrated-circuits), machine-readable text and multimedia system information (including text, picture, video files, and audio and signal data), diagram and network information (for instance, social and knowledge networks), and therefore the internet (a vast, cosmopolitan data repository created offered by the Internet).

1.2.2 What Kind of Patterns Can Be Mined?

Data mining operations are used to determine to be the types of patterns to be established in data mining tasks. Generally, such tasks can be distributed into 2 class: descriptive and predictive. Descriptive mining tasks describe attribute of the data in a destination data set. Predictive mining

tasks implement induction on the existing data in order to produce predictions awareness or knowledge.

Data mining operations and the types of patterns they can detect are described below, also patterns represent knowledge.

1.2.2.1 Class and Concept Description

Data entries are often related to categories or ideas. It is often helpful to explain individual categories and in brief, compendious, and however accurate terms. Such characterization of a category or a thought area unit known as concept /class characterization. They are often derived mistreatment (1) knowledge descriptions, by abstracting the information of the category below study (always known as the destination class) normally terms, or (2) knowledge recognition and determined , by rapprochement of the target category with one or a collection of comparative categories (often known as the different classes), or (3) each knowledge characterization and discrimination.

Data description may be an account of the overall description or options of a destination or target category of information. The information cherishes the user-particular category area unit generally collected by a question.

1.2.2.2 Mining Frequent Patterns, Associations, and Correlations

Frequent patterns, because the name suggests, square measure patterns that occur oftentimes in knowledge. There square measure several types of frequent patterns, as well as frequent item sets, frequent subsequences (also called ordered patterns), and frequent substructures. A frequent item set generally refers to a collection of things that usually seem along in an exceedingly transactional knowledge set. Mining repeated patterns result in the invention of fascinating associations and relation inside knowledge

1.2.2.3 Classification and Regression for Predictive Analysis

Classification is that the process of discovering a sample or model that describe and distinguishes information categories or concept. The sample springs supported the understandings of a group

of coaching information (i.e., information objects that the category labels are unit known). The sample or model is employed to foretell the category label of objects that the category label is strange.

1.2.2.4 Outlier Analysis

A dataset would contain objects that don't appropriate the general behavior or model of the information. That means these objects are outliers. Several data processing ways ignore outliers as noise or exclusion. However, in some cases (for instance fraud detection) the rare events are additional attention-grabbing than the additional often occurring a few times. The study and analysis of extremely information are seen as outlier analysis or anomaly mining.

1.2.2.5 Cluster Analysis

Reverse regression and classification, that analyzes class-labeled (practicing) information groups, Clustering analyzes information objects while not consulting category labels. In several cases, a category labeled information might merely not exist at the start. The clump will be wanted to generate category labels with all bunches of knowledge. The dataset-objects square measure clustered or sorted supported the standard of increasing the intra class rapprochement and reducing the interclass rapprochement. That is groups of dataset-objects square measure fashioned so objects among a cluster has a big likeness as compared to at least one to other, however square measure rather various to things in different groups. Every group thus fashioned will be sighted as a category of record or objects, from that principles, will be created. Clump also can facilitate taxonomy formation that is the arrangement of notice into a hierarchy of proportion that clusters similar procedure along.

1.2.3 Method for data mining

As an extremely implementation-driven domain, data processing has incorporated several techniques from alternative domains like machine learning, statistics. Pattern understanding, information and systems of information warehouse, data retrieval, image, algorithms, superior computing, and plenty of application domains.

The knowledge domain nature of information mining analysis and evolution contribution considerably to the prosperity of information mining and its intensive implementations. During (1.2.3 section), we have a tendency to offer samples of many corrections that powerfully impact the event of information mining strategies.

1.2.3.1 Statistics

It studies the gathering, analysis, translation or clarification, and presentation of information. Data processing has associate in nursing inherent reference to statistics. An applied math model could be a group of mathematical procedures that characterize the behavior of the dataset-objects in a very distention category whence of random variables and their connected chance distributions. Applied math models are wide wont to model information and information categories. For instance, in data treatment tasks like data description and classification, applied math models of target categories are often designed. In different words, such applied math models are often the end result of an information mining task. Instead, data processing tasks are often designed on high of applied math models. For instance, we are able to use statistics to the pattern noise and losing information values. Then, once mining model in a very massive information set, {the information the info the information} mining method will use the model to assist establish and handle noisy or missing values within the data. Statistics analysis develops tools for prediction and statement victimization information and applied math models.

Applied math strategies are often wont to summarize or describe a set of information. Statistics is beneficial for mining numerous models from information in addition as for discover the implicit mechanisms producing and touching the models. Inferential statistics (or prognosticative statistics) patterns information in a very means that represent for randomness and doubt within the notes and is employed to draw the conclusion concerning the method or population underneath investigation.

Statistical strategies may also be wont to verify data processing results. For instance, after Prediction and classification model are well-mined; the model ought to be verified by applied math hypothesis testing. An applied math hypothesis take a look at (sometimes referred to as confirmative information analysis) makes applied math selections victimization experimental

information. A result's referred to as statistically important if it's unlikely to possess occurred inadvertently. If the prediction and classification patterns stay true, then the descriptive statistics of the model will raise the truth of the patterns or model.

Applying applied math strategies in data processing is way from trivial. Often, a significant challenge is a way to proportion a method over an oversized information set. Several applied math strategies have high quality in computation. Once such strategies are executed on massive information sets that are divided into double logical or material sites, algorithms ought to be rigorously prepared and tuned to scale back the machine value. This problem will be more difficult for online implementation and application like on-line question proposals in search or seek engines, wherever data processing is needed to incessantly handle quick, and time period information streams.

1.2.3.2 Machine Learning

Machine learning explores, while, computers will be able to take decision (or progress in their effectiveness) supported information. The most analysis space is for laptop programs to mechanically learn to acknowledge complicated patterns and create intelligent choices supported information. For instance, a model machine learning downside is to prepare a laptop in order that it will mechanically acknowledge written communicating icon or symbol on Email when learn from a collection of examples.

Supervised learning is largely a word for classification. The oversight within the learning comes from the labeled examples within the coaching information set. For instance, in the postcode recognition downside, a collection of written postcode pictures and their corresponding machine-readable translations are used because the coaching examples, which supervise the educational of the classification model.

Unsupervised learning is actually a word for the cluster. The educational methods unsupervised since the input examples don't seem to be a category labeled. Usually, we might use clustering to find categories at intervals the information. For instance, associate unsupervised learning technique will suppose as input, a collection of pictures of written digits. We assume that it finds 10 groups of information. These groups or clusters could agree to the ten unique digits of zero to9,

severally. However, since the coaching information doesn't seem to be labeled, the learned pattern cannot give U.S.A. the linguistics which means of the clusters discovered.

Semi supervised learning may be a category of machine learning method that create use of each labeled and untagged examples once learning a model. In one approach, labeled examples are wanted to learn category models and untagged instance are wanted to refine the limits among categories.

Active learning may be a machine learning approach that lets users to move a vigorous role In the learning method. A vigorous learning approach will raise a user (for instance, a area expert) to label associate instance, that can be from a collection of untagged instance or synthesized by the educational system.

We can note there are many affinities between machine learning and data processing. For classification according to the category or affinity, machine-learning analysis typically cares on the significance of the pattern. Additionally, to significance, data processing analysis space a robust confirmation on the potency or measurability of data-mining ways on massive information sets, yet as on approaches that to treat complicated forms of information and discover new, various ways.

1.2.3.3 Database Systems and Data Warehouses

Database systems analysis focuses on the establishment, servicing, and use of databases for institutions and end-users. Significantly, information program researchers have found extremely familiar rules in information models, question languages, question process and improvement ways, information warehouse, and categorization and accessing method.

Database systems square measure typically documented for his or her high measurability in process terribly giant, comparatively structured information sets.

Many data processing tasks ought to handle giant information sets or maybe time period, quick streaming information. Therefore, data processing will keep the treat with scalable information technologies to realize big potency and measurability on giant information sets. However, data-

processing functions are often accustomed extend the potential of existing information systems to fulfill advanced users refined information analysis necessities.

Modern information systems have engineered methodical information analysis abilities on information Data victimization information deposit and data processing Accessories. An information warehouse integrates data constructing from various sources and varied time frames. It integrated and merged information in multi dimensional house to make part achieved information cubes. The information cube sample does not solely facilitate (OLAP) in multi dimensional databases however conjointly promotes multidimensional data mining.

1.2.3.4 Information Retrieval

Information retrieval (IR) is the study of searching for article or text or data in documents. It will be text or multimedia program and will establish on the online. The variances between ancient data retrieval and info systems square measure twofold:

Information retrieval suppose that (1) the information beneath seek square measure unorganized; and (2) the questions square measure fashioned chiefly by keywords, that do not have complicated structures (opposite SQL queries in information systems).

The model approaches in data retrieval accept potential models moreover, a subject in a very collection of article documents will be sculpture squat as a likelihood allocation of the vocabulary, that is termed a subject model. A text document, which may include one or more topics, will be considered a combination of numerous theme models.

1.3 DATA

(Delen, 2015)

The word "data" may be a general purpose word denoting a group of measurements. "Data points" are individual instances of knowledge. A "data set" may be a well-structured set of knowledge points. Information points will be of many "data varieties," like numbers, or text, or date-times. After we collect information on similar objects in similar formats, we have a tendency to bundle the information points into a "variable." We have a tendency to may provide a variable a reputation like 'age,' that may represent the list of ages of everybody in a very space. The information points related to a variable are unit known as the "values" of the variable. These ideas are a unit foundational to understanding information science.

1.3.1 What is Data

The defines information because of the descriptor of datum; as items of information; and as a group of object-units that are unit distinct from each other.

For our functions, the key elements of those definitions are unit that information are unit observations that are unit measured and communicated in such how on be intelligible to each the recorder and therefore the reader. So, you as someone don't seem to be information, however, recorded observations regarding your information. As an example, your name once was written down is information; or the digital recording you speaking your name is data or a digital photograph of your face or video of you recreation are unit information.

1.3.2 Data point

A data purpose may be a distinct unit of knowledge. During a general sense, any single reality may be information. During an applied mathematics or analytical context, an information purpose is typically derived from a measuring or analysis and might be drawn numerically and/or diagrammatically. The term information is roughly as a data point, the signifier of data.

1.3.3 Data set

In most cases, data mining (sometimes referred to as knowledge or information discovery) is the mode of analyzing knowledge from completely many views and briefing it into helpful data - data that may be accustomed grow revenue, cuts costs, or both. Data processing computer code is one among a variety of analytical equipment or analyzing knowledge. It permits users to investigate knowledge from many alternate angles and dimensions, categorize it, and brief the relationships were known. Technically, data mining is that the process of finding correlations or patterns among dozens of fields in giant relative databases, within the table (1) below illustrates the dataset format.

No	Name	Department	semester
10001	ALI	accountancy	20151
10002	Emra	computer	20162
10003	Rjeb	Chemistry	10172

Table 1. Example of Dataset Attribute in Data Mining

Gathering of objects and their attributes or a gathering of attributes describe an object.

1.3.3.1 Dataset have two parts

Attributes a property or characteristic of an object, we can also say an Attribute is a variable, field, characteristic, or feature.

Object it could be record, point, case or sample

Attribute values

They are amount or character assigned to an attribute.

The difference between attributes and attribute values, the same attribute can be a chart with different attribute values, different attributes can be mapped to the same set of value.

Types of Attributes

the most important types of attributes:-

- Nominal Examples: ID num, hair color, post codes.
- Ordinal Examples: rankings (e.g., taste of cheese on a scale from 1-10), grades, height in {long, medium, low}.
- Interval Examples: calendar dates, temperatures in centigrade or Fahrenheit.
- Ratio Examples: temperature in Rome, length, time, counts n.

1.3.3.2 Discrete and Continuous Attributes

Discrete Attribute: it has only a finite or countable infinite set of values and often represented as integer variables.

Note: binary attributes are a special case of discrete attributes.

Continuous Attribute: Practically, real values (floating-point) can only be measured and represented using a finite number of digits.

2. CLUSTERING (ALGORITHMS&TECHNIQUES)

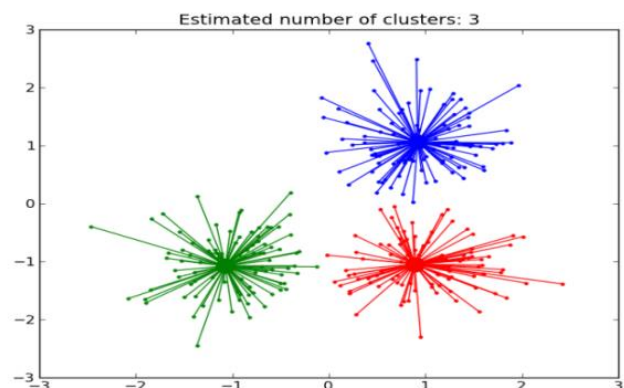
It is one of the most important algorithms in data mining, so it process to find and search for specific, meaningful and useful information within a large size of data, and this is done by the process of linking analysis of these data and methods of artificial intelligence to become better efficient in the process of research.

Clustering is the process of developing data from similar gatherings, which is a branch of data mining. Aggregation algorithm divides data sets into several clusters, as the similarities between the points within a certain grouping larger than the similarity between two points within the different two communities. The idea of compiling data simple 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 the vast amount of data to summarize a few of the groups or categories, in order to facilitate the process of analysis.

Does not use only to organize and classify the data, but it using in data compression and build a model arrangement of data aggregation algorithms on a large scale. If we can find clusters of data, it is possible to build a model of the problem on the basis of those gatherings.

Clustering, in data mining, is helpful for detect collections and identifying interesting distributions in the implicit data. Traditional clustering algorithms either support clusters with Spherical shapes and similar sizes, or are very weak in the existence of extreme values . and that is clear during the picture (1).

Data points as nodes of coherence
Graph and clusters are found by
separating this graph, depend on
Its spectral decomposition, into
sub graphs.



Picture1. Idea of Clustering Technique

2.1 CLUSTERING TECHNIQUES

There is much type of clusters, but we focused on two type of them which are used in this thesis K-mean cluster and spectral cluster.

2.1.1 K-means cluster technique

Divide the objects into k clusters such that some metric relative to the cancroids of the clusters is minimized.

Advantages

- Fast, robust and easier to understand.
- Relatively efficient.
- Gives best result when data set are distinct or well separated from each other.
- It is require calculating the distance between the data points and cluster center and do not need to calculate it between data points itself.

Disadvantages

- The knowledge algorithm needs a priority apportionment of the number of cluster.
- Euclidean distance gauge does not equal weight implied factors.
- The discovery algorithm supplies the local optimum of the squared error procedure.
- Randomly selecting the cluster center cannot advance us the useful result..
- when mean is defined we can deal with it i.e. unsuccessful for categorical data.
- It cannot be handle with noisy and outliers data.
- It cannot be handling for non-linear data set.

2.1.2 Spectral clustering technique

It studies the relation between the data points itself then divide it into some groups where the data in the same group must be connected and coherent.

Pros and cons of spectral clustering:-

Advantages:

- Does not make strong assumptions on the statistics of the clusters .
- Easy to implement.
- Good clustering results.
- Reasonably fast for sparse data sets of several thousand elements.

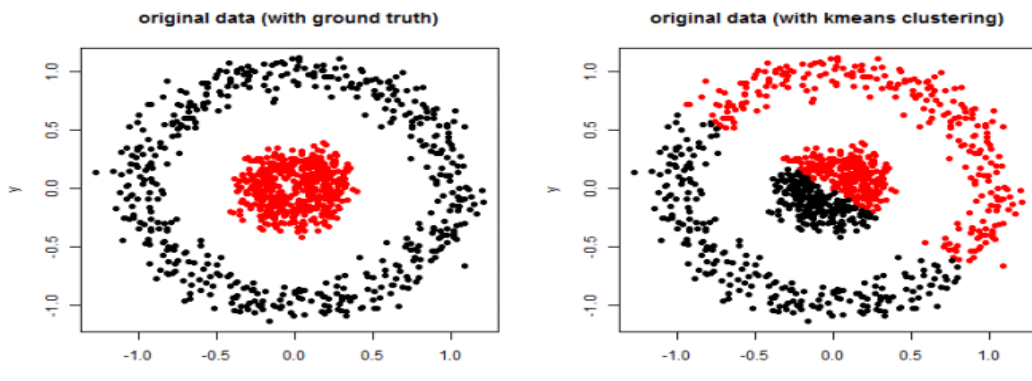
Disadvantages:

- May be sensitive to choice of parameters.
- Computationally expensive for large datasets.
- Dataset that can be able to execute is not public and common in real life.

2.1.3 The different between spectral clustering and K-mean technique

It is very clear in the picture (2) what is the variance between them, in k-mean graph, it is clear that the points are divided into two groups, so that all cluster points are close to each other regardless of the coherence and interconnectivity of points in every cluster.

While noting in S.C.A that separate the data into two clusters according to nearing all points to their cluster center, all the points that are connected and coherence in the same cluster, in addition all data points can reach to others in the same group easily and without having to jump from one point to another



Picture 2. The Different Between Spectral Clustering and k_mean

2.2 ALGORITHMS

There is a close relationship between S.C.A and K-mean, where the last one already included in S.C.A. So when you want to execute it, it is necessary to apply K-mean . After performing many operations firstly .We will first begin to illustrated the processing steps of K-mean algorithm

2.2.1 K_mean clustering algorithm

Suppose that:

$p_i = \{p_1, p_2, \dots, p_n\}$ be the set of data points and $c_j = \{c_1, c_1, \dots, c_k\}$ be the set of centers

K number of clusters.

c_j Cluster center, so c_1 cluster center in cluster 1, c_2 cluster center in cluster 2 and so on.

n no of all data points, n_1 no of data points in cluster 1, n_2 no of data points in cluster 2.

Now the following steps can explain how algorithm works

Randomly select center for every cluster $c_j = (x_{c_j}, y_{c_j}, z_{c_j})$.

$c_1 = (x_{c_1}, y_{c_1}, z_{c_1})$ it is a center for cluster 1

$c_2 = (x_{c_2}, y_{c_2}, z_{c_2})$ it is a center for cluster 2,..... And so on

Calculate the distance between each data point and cluster centers.

First distance between all data points and c_1

$\text{dis}(p_1, c_1) \& \text{dis}(p_2, c_1) \& \text{dis}(p_3, c_1) \& \dots \& \text{dis}(p_n, c_1)$

Second distance between all data points and c_2

$\text{dis}(p_1, c_2) \& \text{dis}(p_2, c_2) \& \text{dis}(p_3, c_2) \& \dots \& \text{dis}(p_n, c_2)$

Then continuous for all clusters center until

Second distance between all data points and c_k

$dis(p_1, c_k) \& dis(p_2, c_k) \& dis(p_3, c_k) \& \dots \& dis(p_n, c_k)$

Note that to find the distance between two point in the space 3D

For instance the distance between any point $p_i(x_i, y_i, z_i)$ and the cluster center number c_k

(x_{cj}, y_{cj}, z_{cj}) is:

$$dis_{p_i, c_j} = \sqrt{(x_i - x_{cj})^2 + (y_i - y_{cj})^2 + (z_i - z_{cj})^2}$$

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.

4) Recalculate the new cluster center using:

$$c_j = ((1/n_i) \sum_{i=1}^{n_i} x_i, (1/n_i) \sum_{i=1}^{n_i} y_i, (1/n_i) \sum_{i=1}^{n_i} z_i) \implies (x_{cj}, y_{cj}, z_{cj})$$

Where, 'n_i' represents the number of data points in j^{th} cluster.

5) Recalculate the distance between each data point and new obtained cluster centers (c_j).

6) If no data point was reassigned then stop, otherwise repeat from step 3, whose is called iteration, it is meaning how many times the steps from 3 to 5 will be repeated.

2.2.2 Spectral clustering algorithm

➤ Given data from, read it from external file to the matrix

$$A = \begin{matrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & x_3 & x_3 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ x_n & y_n & z_n \end{matrix}$$

S	Variable1(x)	Variable2(y)	Variable3(z)
1	----	----	----
2	----	----	----
↓	↓	↓	↓
n	----	----	----

Table 2. Dataset Attribute in S.C.A

Then we need to do pre processing steps as following

➤ Find affinity matrix A [n,n] from matrix A[n,3]

Affinity = $\frac{1}{e^{dis/2\sigma^2}}$ where dis is the distance between every

Element in and others in the same matrix A by the low

$$dis_{p_i,p_j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$$

Where $p_i = (x_i, y_i, z_i)$ & $p_j = (x_j, y_j, z_j)$

K is a parameter express a number of a appropriate vectors will be choose from eigenvector in general equal to number of clusters +1 to provide neigvec.

Here the scaling parameter σ^2 controls how rapidly this affinity falls off with the distance between p_i, p_j .

Two Local Scaling the scaling parameter is some test of when two cases are considered comparable (similar). This gives an intuitive way of choosing potential values for σ . The chosen of σ is generally done manually. Suppose chosen σ automatically by running their clustering algorithm repeatedly for a number of values of σ and choice the one which provides best clusters.

This rises significantly the calculation time. In addition, the extent of values to be tested still has to be set manually. However, when the input data includes clusters with different local statistics there may not be an odd value of σ that works well for all the data. The high impact σ has on clustering. When the data have multiple scales, even using the optimal σ fails to give perfect clustering, the impact of local scaling. The Convergence across clusters are now significantly lower than the affinities within any unique cluster.

To determine the value of σ , we calculate it for finding the Affinity of every two points i and j

First we should analysis σ^2 to σ_i, σ_j

$$\sigma_i = \text{dis}(s_i - s_k)$$

Where s_k is the K'th neighbor of points s_i . The selection of K is independent of scale and is a function of the data dimension of the embedding space. Nevertheless, in most experiments (both on synthetic data and on images) a single value of $K = 7$ was used, which give good results even for high-dimensional data.

Also for σ_j

$$\sigma_j = \text{dis}(s_j - s_k)$$

The Matrix A will be on the form

1	A(p1,p2)	A(p1,p3)	→	A(p1,pn)
A(p2,p1)	1	A(p2,p3)	→	A(p2,pn)
A(p3,p1)	A(p3,p2)	1	→	A(p3,pn)
↓	↓	↓		↓
A(pn,p1)	A(pn,p2)	A(pn,p3)	→	1

Figure 1. Affinity Matrix Form in S.C.A

➤ **Compute** the degree matrix

The D matrix degree is the same degree of the A matrix with the following condition in mind

Calculate summation of each row of the A matrix in a position equal to the row and column of the D matrix

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

$$D[i,j] = \text{Zero}$$

Where n is matrix degree

All other elements in D matrix will be equal to zero. The figure2. Shows final form of D matrix

Sum 1 st row	Zero	Zero	----->	Zero
Zero	Sum 2 st row	Zero	----->	Zero
Zero	Zero	Sum 3 st row	----->	Zero
↓	↓	↓		
Zero	Zero	Zero		Sum n st row

Figure 2. Degree Matrix Form in S.C.A

- **Calculate** the normalized Laplacian affinity matrix by this lows and it is using to find every element in the N.Lap matrix

$$N. L_{(i,j)} = \frac{Affinity[i,j]}{\sqrt{D[i,i]} * \sqrt{D[j,j]}}$$

N. lap _(1,1)	N. lap _(1,2)	N. lap _(1,3)	----->	N. lap _(1,n)
N. lap _(2,1)	N. lap _(2,2)	N. lap _(2,3)	----->	N. lap _(2,n)
N. lap _(3,1)	N. lap _(3,2)	N. lap _(3,3)	----->	N. lap _(3,n)
↓	↓	↓		↓
N. lap _(n,1)	N. lap _(n,2)	N. lap _(n,3)	----->	N. lap _(n,n)

Figure 3. Normalized Laplacian Matrix Form in S.C.A

- **Decomposition**

Linear algebra technique has been used to find both Eigen vectors V and Eigen Values λ

Such that $(N \cdot \text{lap}^* V = \lambda \cdot v)$

Where v is the eigenvector of $N \cdot \text{lap}$ corresponding to λ

We got 186 Eigenvectors and Eigenvalues so for λ there is V

$(\lambda_1 \dots V_1, \lambda_2 \dots V_2, \dots, \lambda_n \dots V_n)$

Eigenvalues Eigenvectors

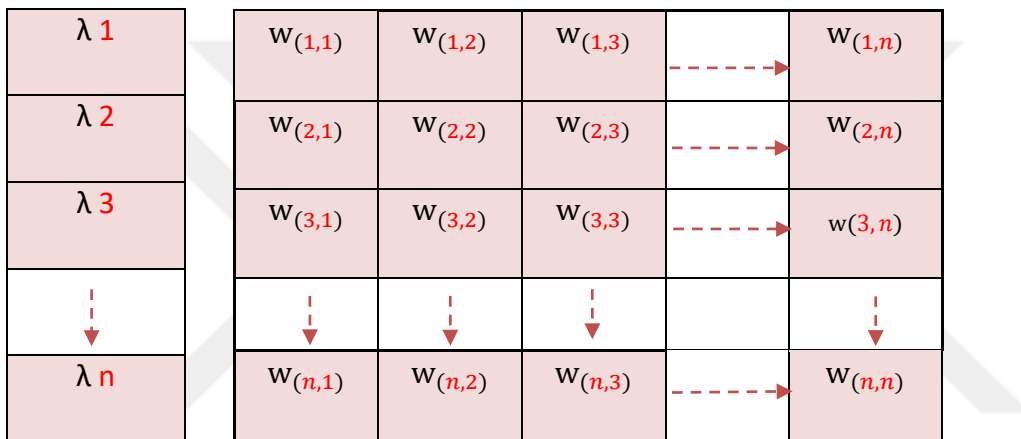


Figure 4. Eigenvalues, Eigenvectors in S.C.A Form in S.C.A

From the Eigenvectors we choose $(k+1)$ such that k no of clusters.

Then chose the biggest Eigenvectors according to k Which is in front the same number of biggest Eigenvalues, suppose $k=3$ then we will choose 4 eigenvectors, then the result was on form neigvec matrix.

w_{n-3}	w_{n-2}	w_{n-1}	w_n
$w_{(1,n-3)}$	$w_{(1,n-2)}$	$w_{(1,n-1)}$	$w_{(1,n)}$
$w_{(2,n-3)}$	$w_{(2,n-2)}$	$w_{(2,n-1)}$	$w_{(2,n)}$
⋮	⋮	⋮	⋮
$w_{(n,n-3)}$	$w_{(n,n-2)}$	$w_{(n,n-1)}$	$w_{(n,n)}$

Figure 5. Eigenvec Form in S.C.A

➤ Construct the normalized matrix [U] from the obtained eigvec matrix by using the equation

$$W_{(i,j)} = \frac{w_{(i,j)}}{\sqrt{w_{(i,1)}^2 + w_{(i,2)}^2 + w_{(i,3)}^2 + w_{(i,4)}^2}} \quad \text{For every element in the U matrix}$$

$w_{(1,1)}$	$w_{(1,2)}$	$w_{(1,3)}$	$w_{(1,4)}$
$w_{(2,1)}$	$w_{(2,2)}$	$w_{(2,3)}$	$w_{(2,4)}$
⋮	⋮	⋮	⋮
↓	↓	↓	↓
$w_{(n,1)}$	$w_{(n,2)}$	$w_{(n,3)}$	$w_{(n,4)}$



Figure 6. Rearrange Eigenvec Form in S.C.A

$\frac{w_{(1,1)}}{\sqrt{w_{(1,1)}^2 + w_{(1,2)}^2 + w_{(1,3)}^2 + w_{(1,4)}^2}}$	$\frac{w_{(1,2)}}{\sim}$	$\frac{w_{(1,3)}}{\sim}$	$\frac{w_{(1,4)}}{\sim}$
$\frac{w_{(2,1)}}{\sqrt{w_{(2,1)}^2 + w_{(2,2)}^2 + w_{(2,3)}^2 + w_{(2,4)}^2}}$	$\frac{w_{(2,2)}}{\sim}$	$\frac{w_{(2,3)}}{\sim}$	$\frac{w_{(2,4)}}{\sim}$
⋮	⋮	⋮	⋮
↓	↓	↓	↓
$\frac{w_{(n,1)}}{\sqrt{w_{(n,1)}^2 + w_{(n,2)}^2 + w_{(n,3)}^2 + w_{(n,4)}^2}}$	$\frac{w_{(n,2)}}{\sim}$	$\frac{w_{(n,3)}}{\sim}$	$\frac{w_{(n,4)}}{\sim}$

Figure 7. Normalized Matrix U Form in S.C.A

➤ then apply K_mean algorithm on U matrix to normalized it to find the K clusters of the given n point in 3D ,therefore It will return the cluster address ID of each point which represent the number of cluster each point belongs to.



3. METHODOLOGY

It is not easy to apply the algorithm on real data, because that application may refer to the specificity and advantage of this algorithm. We can split the data into a number of clusters based on the convergence of these points (even if it is one or more dimension). In addition, the cohesion and the continuity that is uncommon in the practical life, it can be applied in several studies, researches and statistics and engineering applications, and that it depends on the nature of the data type.

In this study, we have done with a lot of collection, testing, calibration of dozens of statistics gotten from warehouse EEA database (European Economic Association), and those data can be applied by one important algorithm in data mining system or it can be applied by k_mean algorithm to illustrate the advantages. We can assume in the first stage that the result can be analyzed and find the clustering benefits, according to the following parameters: The first study classifies the people at risk of poverty or social exclusion, while the second study is focusing on population tertiary education attainment level and median income by work intensity of the household. All studies were conducted between 2007 and 2014 in thirty-one European countries.

What Is Tertiary Education?

Tertiary education refers to any type of education that is pursued beyond the high school level. This includes diplomas, undergraduate and graduate certificates, associates, bachelor's, masters and doctoral degrees. I used matlab language to execute the S.C.A, so the program reads the dataset from excel file which is provided in the dataset collection stage. The stage is to analyze the results by using a few statistic techniques, such as average, variance and graph representation to get the acceptable, persuading and satisfactory benefits.

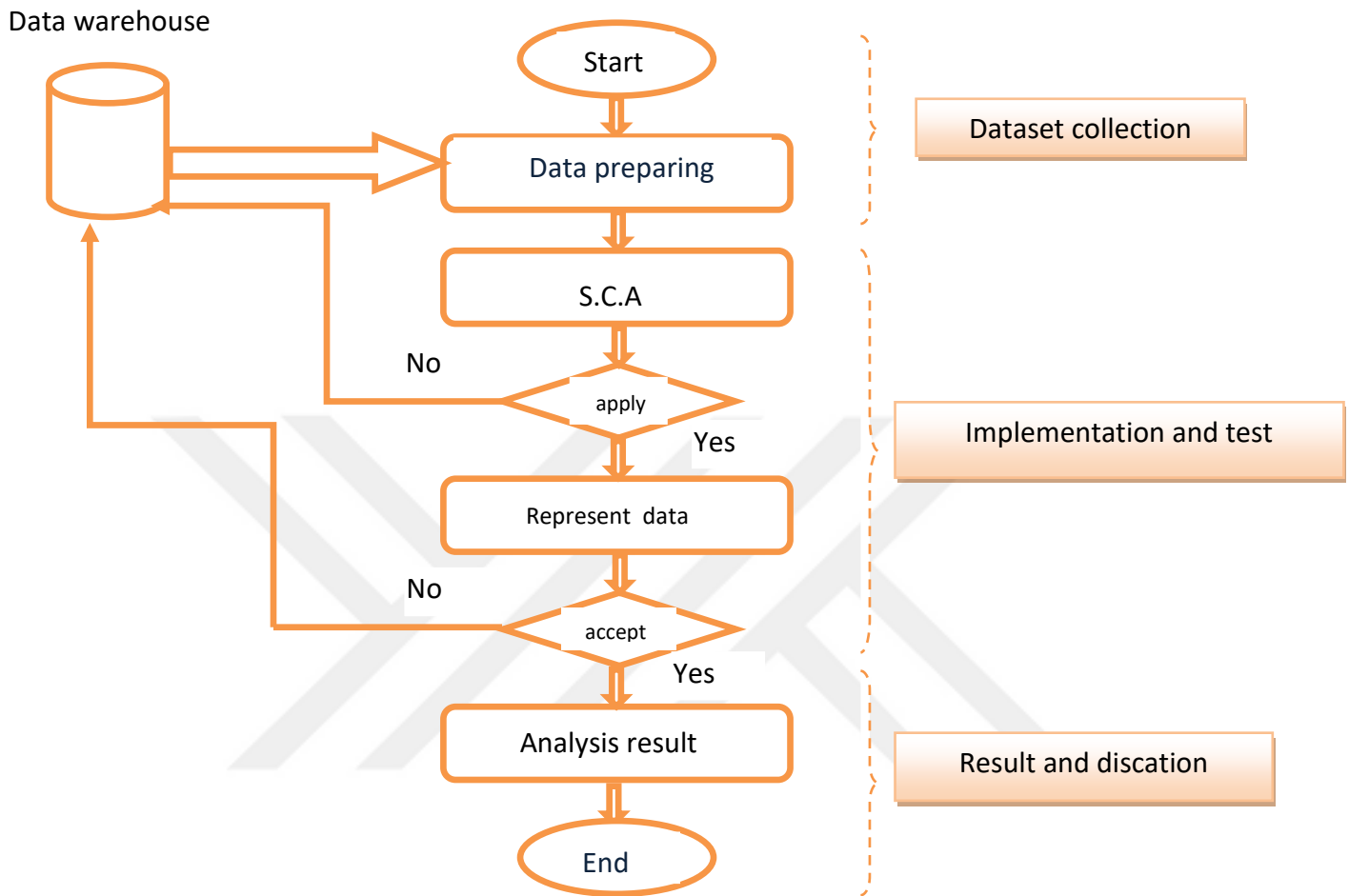


Figure 8. Flowchart of Implementation

3.1 DATASET COLLECTION

Some technologies have been used to coordinate and standardize data from warehouse EEA database- knowledge discovery from data (KDD). So, it is necessary to implement any algorithm in data mining for executing some KDD's processing to provide and prepare a dataset in appropriate ways. In the following steps we are explaining the most important operations.

3.1.1 Data selection

In this study, selecting appropriate way for dataset is necessary and relevant to the analysis task. We are focusing repeatedly in warehouse data until we got an adequate database in order to achieve the objectives and satisfactory results for this study.

3.1.2 Data cleaning

It is for removing noise and inconsistent data in dataset process.

First dataset 1

Total percentage of people with a risk of poverty (same countries and same period).

geo\time	2003	2008	2014	2015	2016
Euro area (19 countries)	:	21.7	23.5	23.1	:
Euro area (18 countries)	:	21.6	23.5	23	:
Belgium	23.4	20.8	21.2	21.1	:
Bulgaria	:	44.8	40.1	41.3	:
Norway	12.9	15	13.5	15	:
Switzerland	:	18.1	16.4	:	:
Serbia	:	:	43.1	41.3	:

Figure 9. Cleaning Data in Dataset 1 (Risk of Poverty)

There are lots of inconsistent data between the three statistics, for example, we found that some values which belong to the European Union are combined years and number of countries. In other words, they are focusing on individual country and not on the general statistics of the European Union. There are also many countries that have no values in some years and sometimes for most of the years, so we ignored them as shown in figure

9. The attribute (field) that don't contain data has been estimated by using average and variance as we noted that in yellow color in figure 9.

s	stat/time	2008	2009	→	2013	2014
1	Belgium	20.80	20.20		20.80	21.20
2	Bulgaria	44.80	46.20		48.00	40.10

30	Norway	15.00	15.20		14.10	13.50
31	Switzerland	18.10	17.90		16.30	16.40

Figure 10. Estimate Some Data in Dataset 1 (Risk of Poverty)

3.1.3 Data transformation

Where data are transformed and consolidated into appropriate forms for mining by performing summary or aggregation operations. All columns were converted by starting from the 4th column while preserving the sequence of the 1st column. The second column is referring to the state, and the 3rd column contains the 1st year (2007). The entire fourth column was moved down, the third column is taking a copy from the second column (countries) and pasted below the same column to match each country with the corresponding value. This process was done in order to avoid overlap the dataset. That operation was clear in when the dataset is changing from figure 10 to figure 11.

ser	risk of poverty	year	country
1	20.80	2008	Belgium
2	44.80	2008	Bulgaria
↓			
31	18.10	2008	Switzerland
32	20.20	2009	Belgium
33	46.20	2009	Bulgaria
↓			
62	17.90	2009	Switzerland
↓			
186	16.30	2013	Switzerland

Figure 11. Data Transformation in Dataset 1 (Risk of Poverty)

Thus, there are three files; each file is containing the $(6 * 31) = 186$ Rows (Object or record) which 31 represent countries and 6 represent the years. The figure10 is showing the calibrations; percentage of people in educational attainment at different levels in thirty one European countries in the period from 2007 to 2015. The same methods are used for data cleansing and data conversion with dataset 2 (percentage of people in educational attainment) and dataset 3 (average income and average work intensity in the household). Therefore, the results are obtained in Figure 11. represents the dataset 2 and Figure 12. Represents the data set 3.

ser	education	year	country
1	2008	Belgium
2	2008	Bulgaria
↓			
31	2008	Switzerland
32	2009	Belgium
33	2009	Bulgaria
↓			
62	2009	Switzerland
↓			
156	2013	Belgium
157	2013	Bulgaria
↓			
186	2013	Switzerland

ser	income	year	country
1	2008	Belgium
2	2008	Bulgaria
↓			
31	2008	Switzerland
32	2009	Belgium
33	2009	Bulgaria
↓			
62	2009	Switzerland
↓			
156	2013	Belgium
157	2013	Bulgaria
↓			
186	2013	Switzerland

Figure 12: Data Transformation in Dataset 2 (Education)

Figure 13: Data Transformation in Dataset 3 (Income)

3.1.4 Data integration

Multiple data sources may be combined, all the three datasets from warehouse database (EEA), and they are from the main branch (Living condition and welfare). But, there are different sub-branch. The figure not shows how that tree dataset are combined into one dataset.

ser	risk of poverty	education	income	year	country
1	20.80	28.40	20,230	2008	Belgium
2	44.80	18.90	2,554	2008	Bulgaria
186	16.30	33.20	43,051	2013	Switzerland

Figure 14. Integrate Three Dataset in One Dataset

3.1.5 Data mining

It is an essential process where intelligent methods are applied to extract data patterns. The order should be in the figure, so the year and the state are not included, but it includes (risk of poverty, education, and income). That can keep the order of the objects or points; therefore the last form of the dataset will be 186 rows and 3 columns as shown in figure 15.

ser	risk of poverty	education	income	cluster
1	20.80	28.40	20,230	?
2	44.80	18.90	2,554	?
186	16.30	33.20	43,051	?

Figure 15. Data Patterns

For the appropriate data, the attribute value of income is divided by 1000, then the Final dataset which contains 3 variables (attribute) and 186 points (objects), as it is illustrating in figure 14.

3.2 IMPLEMENTATION

After preparing the data and put it in an appropriate pattern by technical ways. The steps for data collection are following as explained in the previous section 3.1.

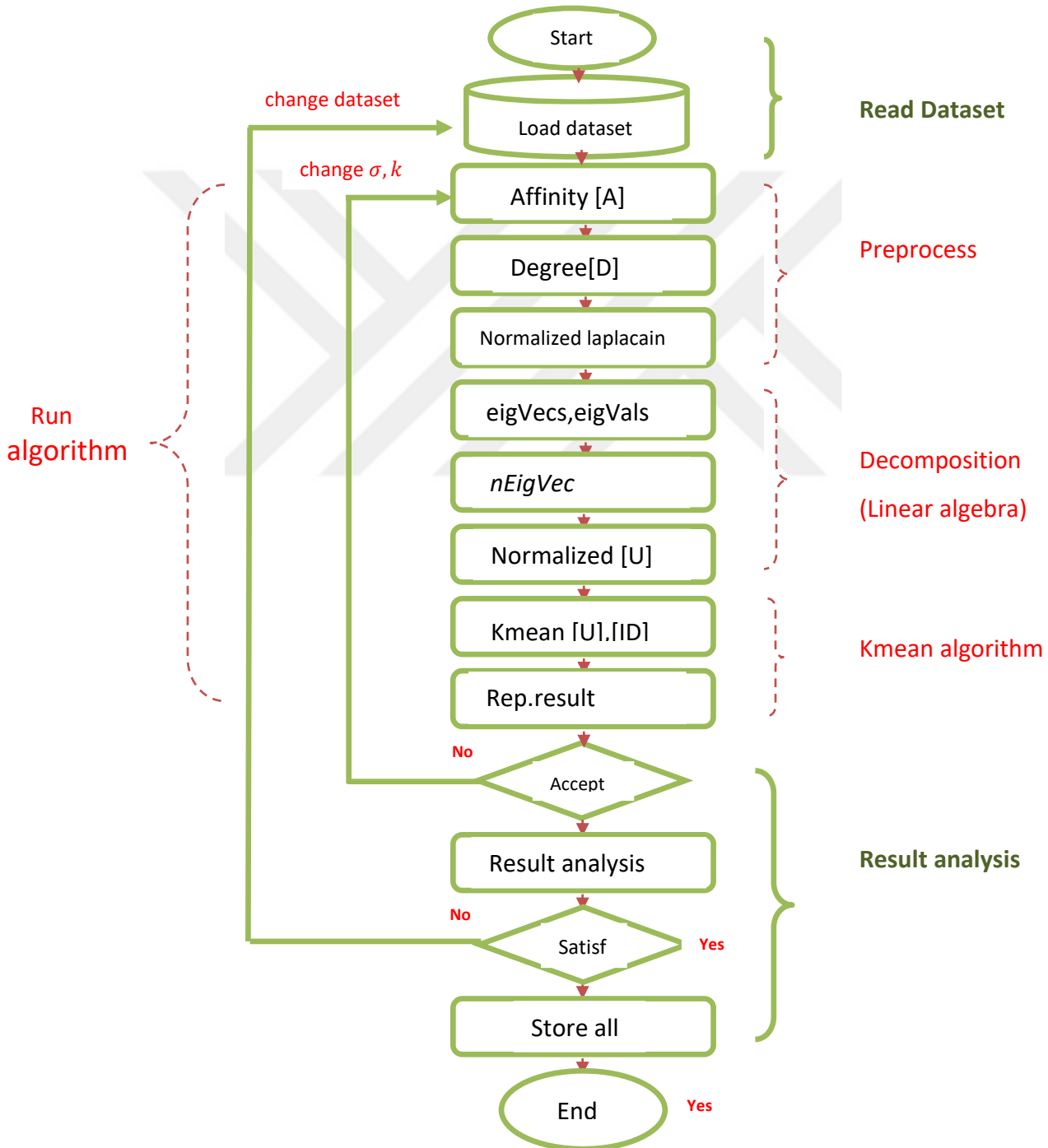


Figure 17. Implementation Flowchart

It is getting times by trying to obtain an acceptable data for having useful and satisfying results. And this will led to the compilation of dataset collection within the implementation of the algorithm as shown in Figure17. , This process is more complicated when we are using an attribute which is consisted of 3 factors, and the stages of implementation are as follows:-

3.2.1 Stages of Implementation

The language used for implement is matlab language version 14 in the implementation of the algorithm because of its features and advantages which distinguish it than others for the execution of complex engineering programs.

In order to facilitate the idea of implementing the program, the program was divided into 3 stages where each stage depending on the other stage and can be returned from one stage to another. Whenever necessary or the quality of results was not acceptable for the implementation. The following is an explanation of these stages

3.2.1.1 Load Dataset

Read the data from excel file to the matrix

$$A = \begin{matrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & x_3 & x_3 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ x_{186} & y_{186} & z_{186} \end{matrix}$$

S	Risk	Education	income
1	20.80	28.40	20.23
2	44.80	18.90	2.55
↓	↓	↓	↓
186	16.30	33.20	43.05

Table 3. Read Dataset From Excel File

The data points are stored in the array [186, 3], so every point is represented in 3 dimension x,y,z the first step in the program after clear all memory from any previous variables read excel file in the array inside the code.

3.2.1.2 Run Algorithm

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

- Preprocessing.
- Decomposing.
- Apply k_mean algorithm.

Firstly preprocessing consists of some operations as follows:-

1) Find affinity matrix A [186,186] from matrix A [186,3]

Affinity = $\frac{1}{e^{dis/2\sigma^2}}$ where dis is the distance between every element in dataset and other elements in the same matrix A by the low.

$$dis_{p_1,p_2} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$$

Where $p_1 = (x_1, y_1, z_1)$ & $p_2 = (x_2, y_2, z_2)$

As we illustrated in chapter two (2.2.2), how we can choose both parameters (k and σ), so in our study they are as following (k=4) , ($\sigma = 0.572$), where parameter (K) is expressing a number of appropriate vectors and that will be chosen from eigenvector and σ^2 controls how rapidly this affinity falls off with the distance between every two points p_i, p_j in dataset points.

2) Compute the degree matrix

The D matrix degree is the same degree of the A matrix with the following condition in mind.

Calculate summation of each row of the A matrix in a position equal to the row and column of the

D matrix

$$D[i, j] = \sum_{j=1}^n D[i, j] \quad \left\{ \begin{array}{l} \text{If } i = j \\ \text{if } i \neq j \text{ Where } n=186, \text{ it is matrix degree} \end{array} \right.$$

All other elements in D matrix will be equal to zero

3) Calculation is to normalize Laplacian affinity matrix by this lows and it is using to find every element in the N.Lap matrix.

$$N. L_{(i,j)} = \frac{Affinit[i,j]}{\sqrt{D[i,i]} * \sqrt{D[j,j]}}$$

Secondly decomposing

The technique for Linear algebra has been used to find both Eigenvectors V and Eigen Values λ Such that (n.lap* V= λ *v). Where v is the eigenvector of N.lap corresponding to λ . We got 186 Eigenvectors and Eigen values, so for every λ there is V ($\lambda_1 ..V_1, \lambda_2 ...V_2,$, $\lambda_{186} .. V_{186}$). From the Eigenvectors we choose (k+1) =4 such k of no clusters. And we choose the biggest Eigenvectors according to k which corresponds to the same number of biggest Eigenvalues, because k=3 equals to 4 eigenvectors.

Constructing the normalized matrix [U] from the obtained neigvec matrix [186,4] to find every element of this matrix using the law :-

$$U_{(i,j)} = \frac{v_{(i,j)}}{\sqrt{v_{(i,1)}^2 + v_{(i,2)}^2 + v_{(i,3)}^2 + v_{(i,4)}^2}}$$

Thirdly apply k_mean algorithm

As we explained in chapter 2 (2.2.1) K_mean clustering algorithm is executing on [U] matrix, then the data will be divided in three clusters according to S.C.A.

3.2.1.3 Results Analysis

The data is represented in three dimensions; every data of cluster is colored to differentiate one from others in order to determine to extent if these results are acceptable and allowing to get to subsection 3.2.2 but, the results cannot be acceptable somehow within the paradigms of changing parameters for k and σ values, or one of them. to obtain useful and meaningful results, the

characteristics of each group should distinguish between the rests of the clusters ending the execution and the other cases. That recommends otherwise the necessity to return to the first subsection 3.2.1.1 Load Dataset.

Outputs of spectral clustering algorithm implementation will be on table 4 where every point carries cluster number which belongs to.


ser	risk of poverty	education	income	cluster
1	20.80	28.40	20,230	1
2	44.80	18.90	2,554	2
				
186	16.30	33.20	43,051	1

Table 4 .Outputs of Implementation S.C.A

3.3 RESULT AND DISCUSSION

The chart explains distributing dataset points in three dimensions before applying any algorithm on them.

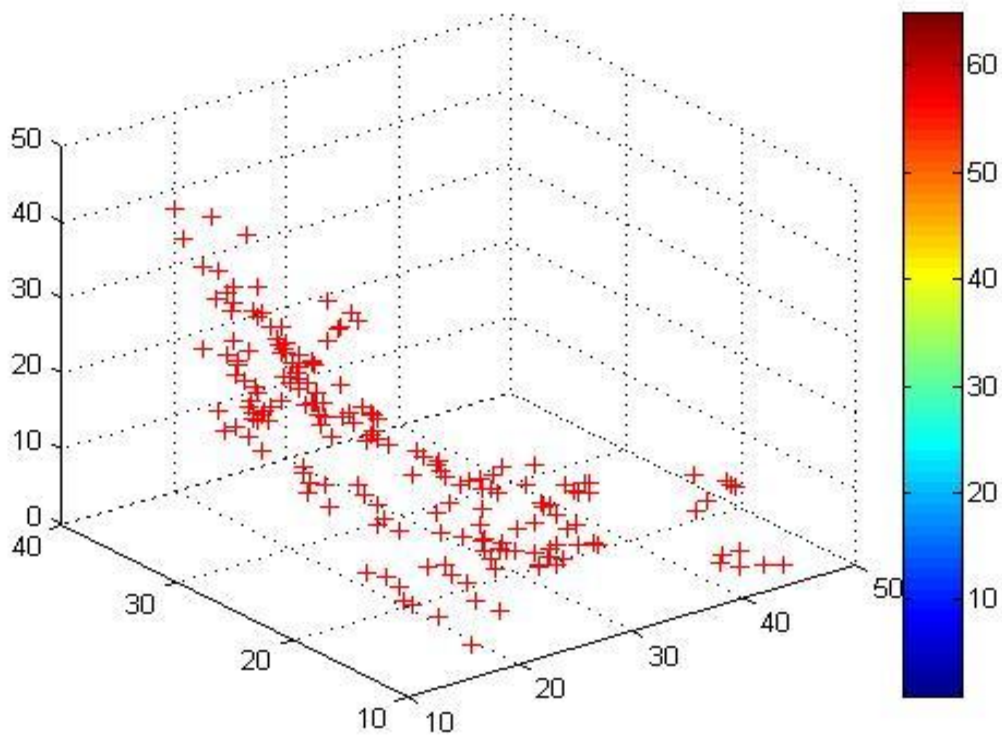


Chart 1. Represent Dataset Points in 3D

To make my project clearer and to recognize the advantages and the characteristics of the use of S.C.A , we implemented the K_mean algorithm on the same dataset.

To clarify the idea of the work for getting the basic differentiation in the mathematical and logical concept, or to have the difference in the general concept through the definition, where the last concept is focused on the relationship between the centered cluster of data, and the rest data points. In other words, it divides the data into a group of common groupings with some characteristics, regardless of the interconnection and modulation of those data within the clustering (grouping).

As for the S.C.A the dataset points are divided into clusters with similar characteristics by taking into account the strength of the relationship between them within the cluster.

3.3.1 Results of S.C.A

The results of the algorithm are influenced by the number of clusters. Therefore, we implemented the program many times with a change K each time, then we studied these results to get the best.

Implement with K=3

3D point	risk	education	income	year	country	cluster
1	20.8	28.4	20.23	2008	Belgium	1
2	44.8	18.9	2.554	2008	Bulgaria	2
3	15.3	12.4	6.539	2008	Czech	3
4	16.3	26.3	26.422	2008	Denmark	1
185	14.1	34.2	44.5	2013	Norway	1
186	16.3	33.2	43.051	2013	Switzerland	1

Table 5. Combine Outputs of Implementation with Source Dataset

Table 5. is combining the result of algorithm implementation and original dataset. In the common three groups we used some statistical measures like mean and variance as shown in the figure 18. to easily and quickly find the Characteristics of each cluster, Then we represent the dataset according to the result , the clusters will be illustrated on chart 2.

	cluster1	cluster2	cluster3
Avg of risk of poverty	19.49	45.1	26.24
Avg of Education	28.93	16.23	18.32
Avg of income	23.26	2.63	8.48

Figure 18. Three Avg Values (Poverty,Education,Income) for S.C.A (K=3)

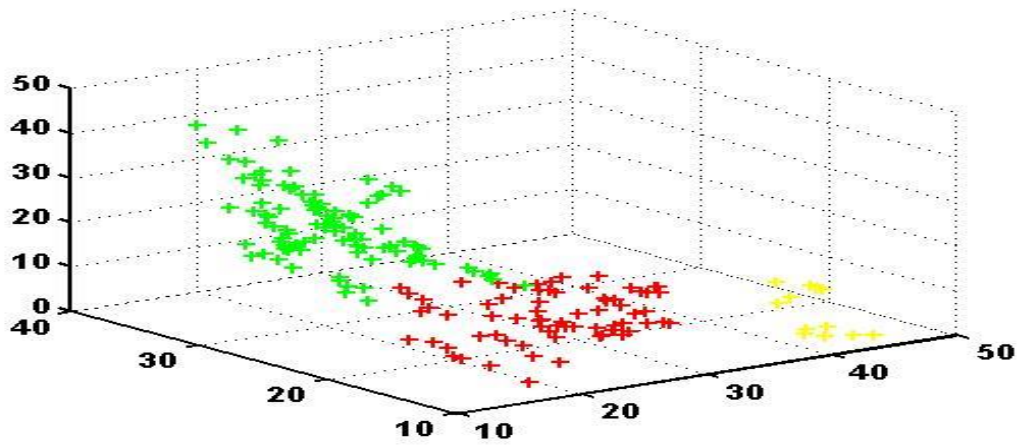


Chart 2. Graph result for S.C.A with k=3

To facilitate the analysis of the results, we can find mean value for risk of poverty, education, and income in each cluster; it is clear in figure 18. ,Chart 3. shows three values for every cluster.

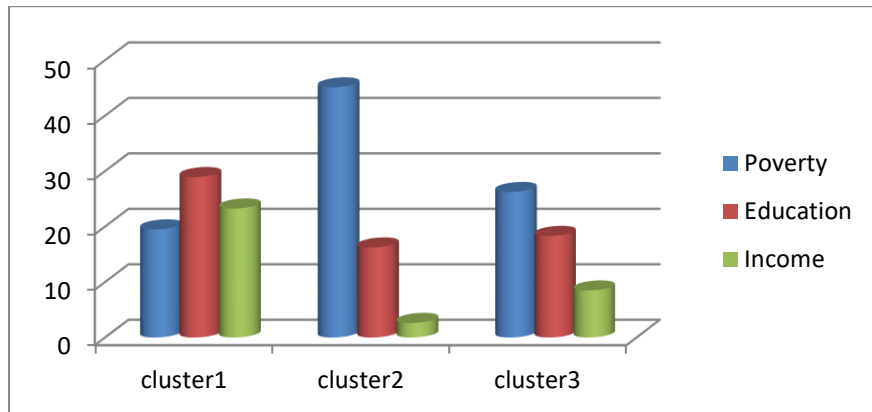


Chart 3. Representing S.C.A result with 3 Clusters

Implement with K=4

	cluster1	cluster2	cluster3	cluster4
Poverty	37.67	19.29	19.18	26.24
Education	21.05	29.65	16.35	18.32
Income	4.02	24.44	22.03	8.48

Figure 19. Four Avg Values (Poverty,Education,Income) for S.C.A (K=4)

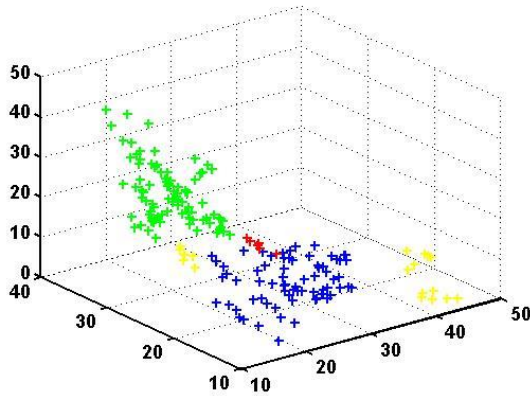


Chart 4. Graph result for S.C.A with k=4

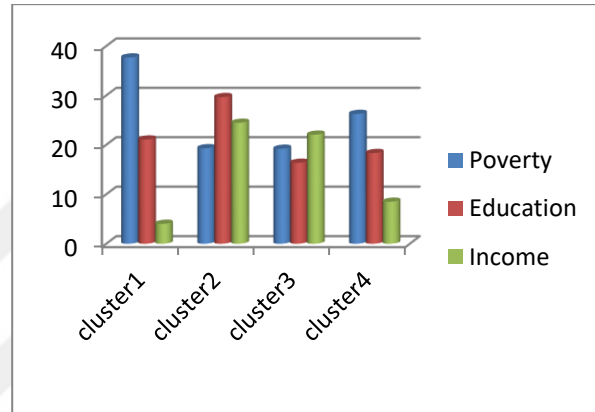


Chart 5. Representing S.C.A result with 4 Clusters

Implement with K=5

	cluster1	cluster2	cluster3	cluster4	cluster5
Poverty	19.29	19.18	26.24	22.81	45.1
Education	29.65	16.35	18.32	30.7	16.23
Income	24.44	22.03	8.48	6.8	2.63

Figure 20. Five Avg Values (Poverty,Education,Income) for S.C.A (K=5)

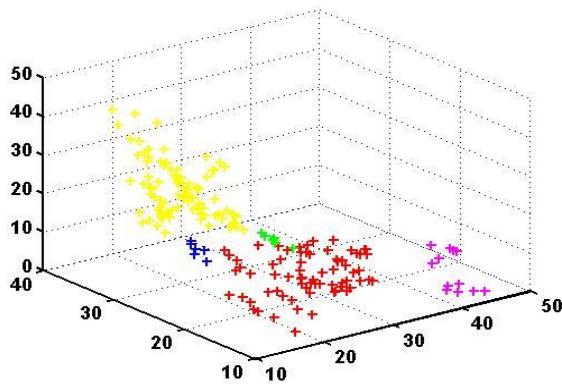


Chart 6. Graph result for S.C.A with k=5

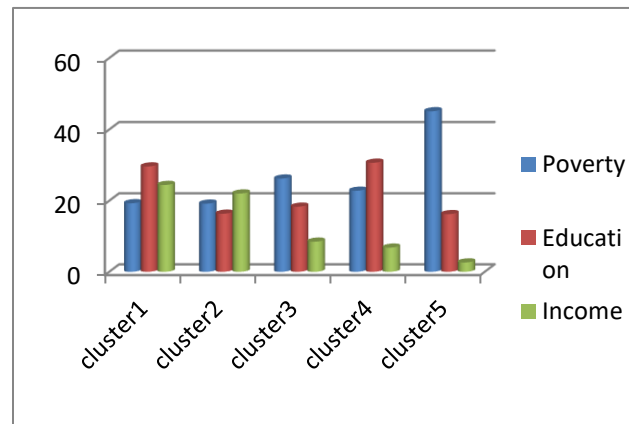


Chart 7. Representing S.C.A result with 5 Clusters

Implement with K=6

	cluster1	cluster2	cluster3	cluster4	cluster5	cluster6
Poverty	22.81	26.24	19.18	16.5	20.49	45.1
Education	30.7	18.32	16.35	30.07	29.47	16.23
Income	6.8	8.48	22.03	33.083	20.74	2.639

Figure 21. Six Avg Values (Poverty,Education,Income) for S.C.A (K=6)

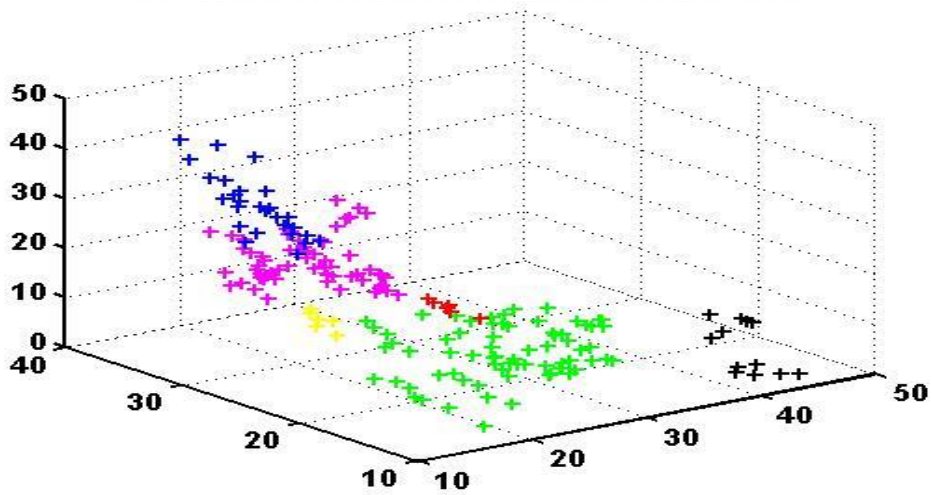


Chart 8. Graph result for S.C.A with k=6

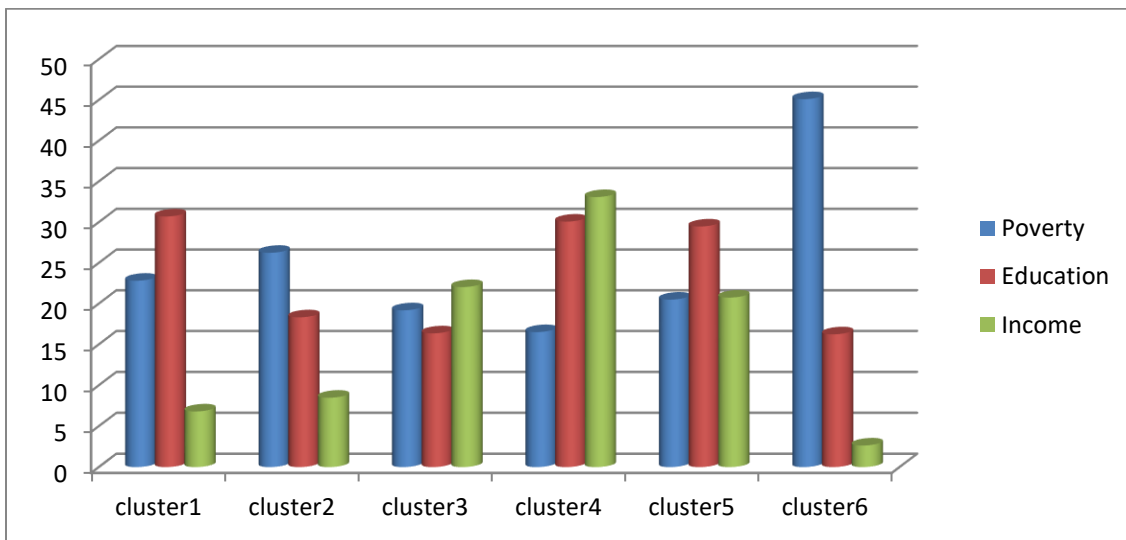


Chart 9. Representing S.C.A result with 6 Clusters

3.3.2 Discussion Spectral Clustering Algorithm

In view of the implementation results, a logical and realistic relationship is observed when k=3 and that is not verified when k=4,5,6 ,so the relation is as following:-

For S.C.A. we can express three factors and their relationships:

(Income & Risk of poverty)  negative relation.

(Income & Education)  positive relation.

In this analysis for using S.C.A , it classified European countries according to 3 different statistical variables and their impact on each other; the medium income by work intensity of the household (Euro), the risk of poverty or social exclusion, and percentage of Tertiary education attainment. In this case, by applying any algorithm, or technique such as: regression and correlation, it is difficult to determine which variable is independent or dependent, or to determine which one is linear or nonlinear. In rich countries where people have luxury, well-being life and high living rate, the percentage for risk of poverty decreases. In the same time.

Risk of poverty	Education	income	cluster	year	country
16.3	26.3	26.42	1	2008	Denmark
20.1	21.4	20.25	1	2008	Germany
21.8	28.3	6.151	1	2008	Estonia
23.7	30.3	25.82	1	2008	Ireland
19.49	28.93	23.26			Average
43	11.2	2.245	2	2009	Romania
49.2	19.7	3.329	2	2010	Bulgaria
41.5	11.9	2.068	2	2010	Romania
49.1	20.1	3.335	2	2011	Bulgaria
41.64	14.98	2.44			Average
24	17.2	13.29	3	2013	Malta
25.8	22.6	5.342	3	2013	Poland
27.5	17.6	8.757	3	2013	Portugal
20.4	24.4	12.5	3	2013	Slovenia
26.24	18.32	8.48			Average

Figure 22. Sample of the S.C.A Results

The percentage of educational attainment is higher for reason that those countries people have Comfort and life stability. But, in the least well-being and least luxury countries, the number of risk of poverty rises, while the percentage educational attainment decreases. As a result, the last

group is trying to rely on themselves and return to labor market for attempt to find a job opportunity.

More detail

The below Figure 19. shows sample of the result how those three variables are interchanged within the cluster. It shows that the increase of the average income value with the level of educational attainment get the number of risk poverty or social exclusion decreasing. Three cases can satisfy that approach; first, the Estonian case with increasing of 6.151 incomes and increasing as well the percentage of education and reducing of risk of poverty. In second cluster all the sample are compatible with each other: in clustering countries, the percentage of income is decreasing sharply with a significant decrease in educational attainment and raising the number of risk of poverty. The last one illustrated the same relation in pervious clusters with clear illustrating the advantages of this algorithm to find a complex and intertwined relationship between the variables, studies or statistic relations.

From the data we can spilt the clusters into three and the algorithm may be applied fully to cluster number two; it is clearly evident from the next graph where all the points are closed to each other within the cluster. In addition, the cluster is isolated from the rest of other groupings. But we separate the clusters from one another when is not possible to gather the points for reason that overlapping between the points can only be applied to a large extent.

3.3.3 Results of K_mean

As in case S.C.A we implemented the program many times with a change K each time, then we choose the best result, according to the achievement any relationship between variables.

Implement k-mean with K=3

The result for implementing the k_mean can be represented in 3 D (x,y,z) by chart 10.

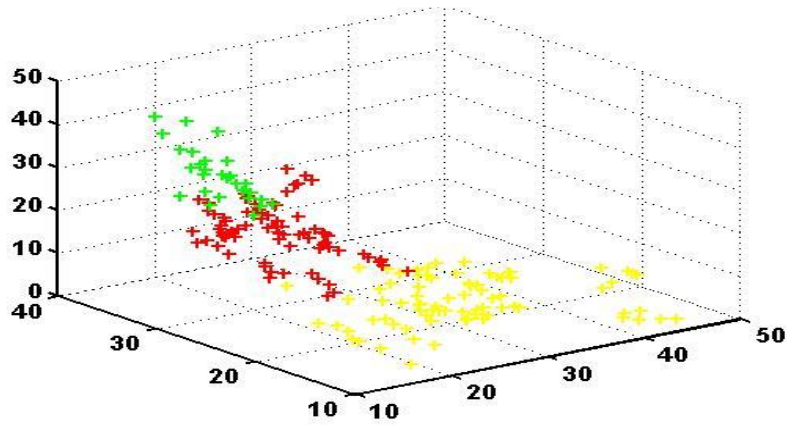


Chart 10. Graph result for k_mean with k=3

The procedures which taken with S.C.A repeating with K_mean implementation, is to find the average of three factors (risk,education,income) for all clusters as shown the figure 22.

	cluster1	cluster2	cluster3
Poverty	29.48	16.48	20.51
Education	17.92	30.19	28.03
Income	7.32	32.81	19.37

Figure 23. Three Avg Values (Poverty,Education,Income) for k_mean (k=3)

The chart 11. illustrates the coloration relation between the 3 variables.

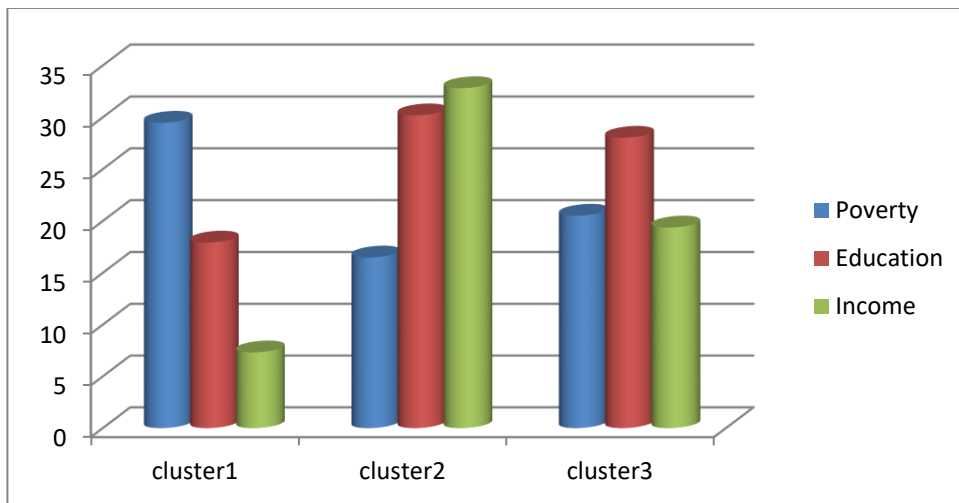


Chart 11. Representing k_mean result with k=3

Implement k-mean with K=4

	cluster1	cluster2	cluster3	cluster4
Poverty	35.86	21.665	20.749	16.48
Education	19.88	16.44	29.54	30.19
Income	4.81	11.81	19.64	32.81

Figure 24. Four Avg Values (Poverty,Education,Income) for k_mean (k=4)

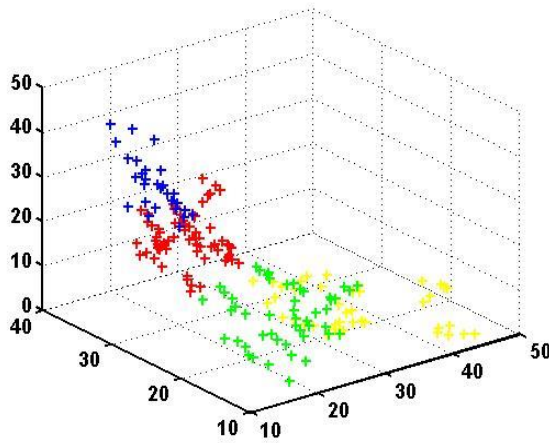


Chart 12. Graph result for k_mean with k=4

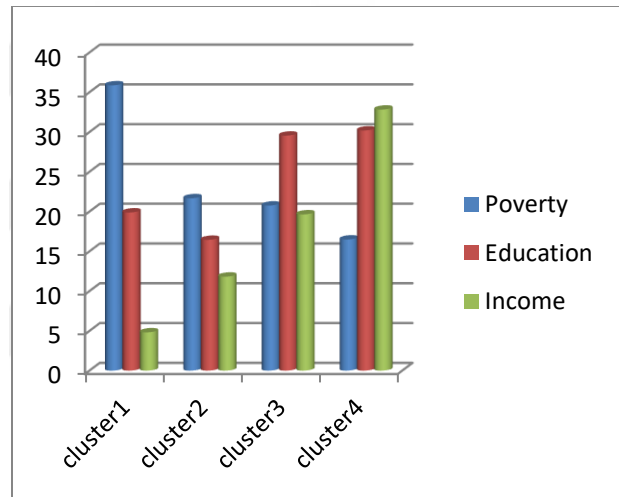


Chart 13. Representing k_mean result with k=4

Implement k-mean with K=5

	cluster1	cluster2	cluster3	cluster4	cluster5
Poverty	20.35	21	44.72	29.96	16.54
Education	29.43	15.73	16.8	22.54	30.22
Income	20.19	12.133	2.81	6.44	33.62

Figure 25. Five Avg Values (Poverty, Education, Income) for k_mean (k=5)

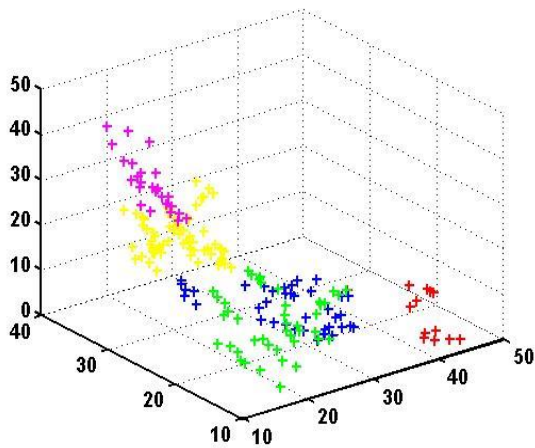


Chart 14. Graph result for k_mean with k=5

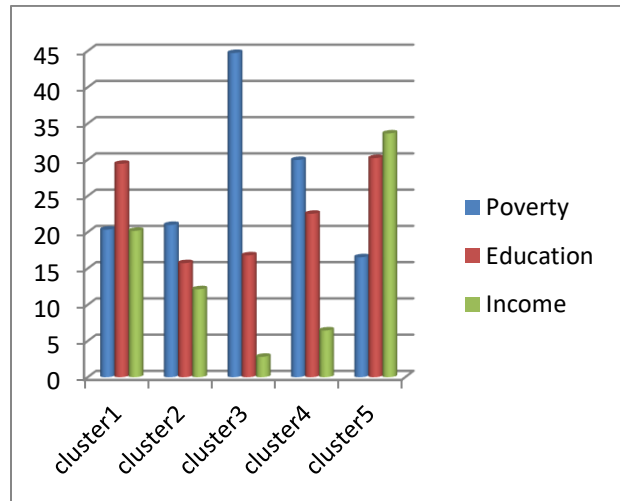


Chart 15. Representing k_mean result with k=5

Implement k-mean with K=6

	cluster1	cluster2	cluster3	cluster4	cluster5	cluster6
Poverty	20.19	22.45	44.72	17.07	24.72	31.22
Education	15.8	16.3	16.8	29.26	31.28	21.1
Income	9.21	19.61	2.81	26.69	16.18	6.37

Figure 26. Six Avg Values (Poverty, Education, Income) for k_mean for k=6

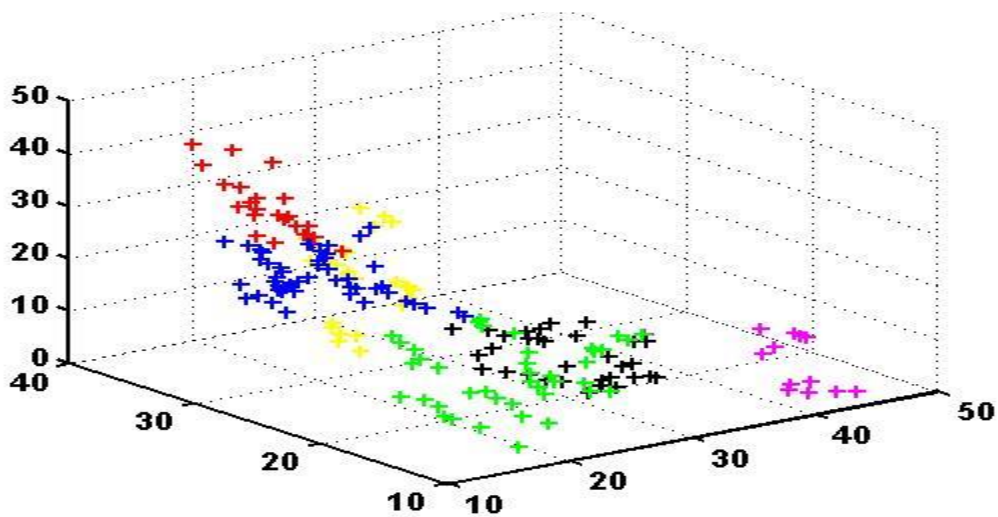


Chart 16. Graph result for k_mean with k=6

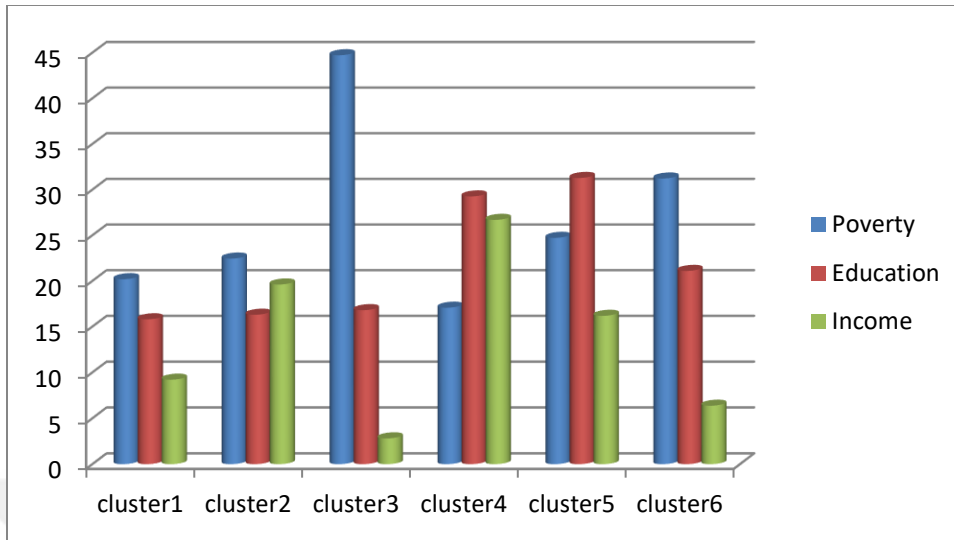


Chart 17. Representing k_mean result with k=6

3.3.4 Discussion of K_mean

For the k_mean results the same relationship was achieved using the S.C.A with k=3 only, where there is no relationship fixed between the clusters in all cases (k=4,5,6).

Through the above, it is cleared that the coloration relation was achieved with K = 3 in both algorithm and to select the most efficient algorithm with the data used we look to the chart 11.

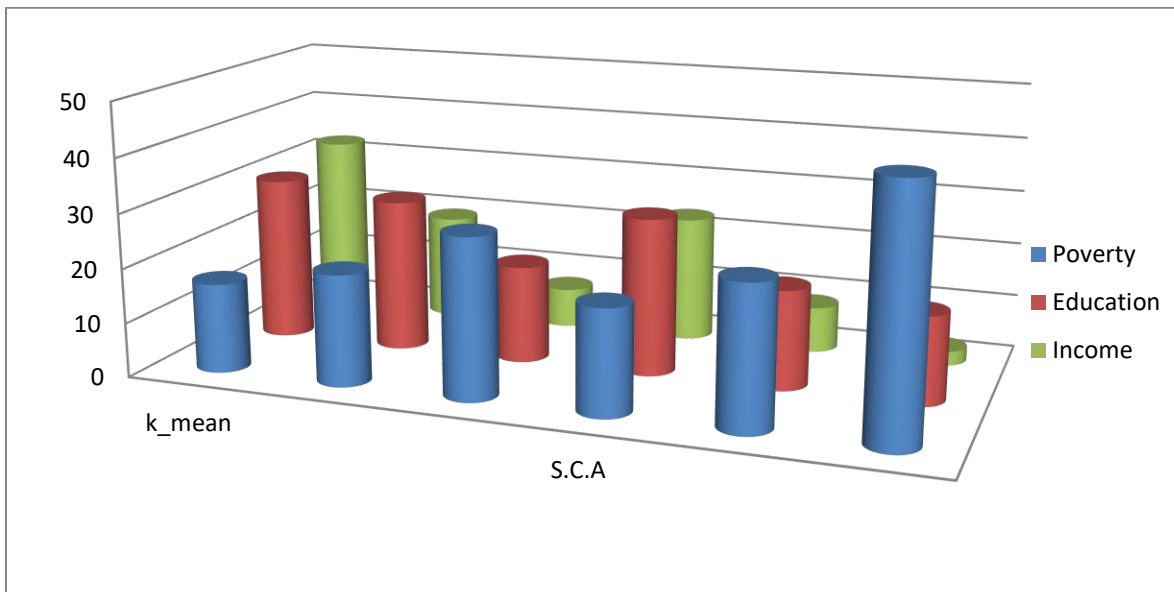


Chart 18. Representing both k_mean , S.C.A results with k=3

From the chart 18. the correlation relation of risk of poverty with both income and education level in the case of S.C.A is found to be stronger than k_mean algorithm, Although it also achieved the relationship.

It is the nature of the data that compels the user to select the appropriate algorithm. When it is possible to separate data in many clusters, it is appropriate to use S.C.A, as in yellow cluster in chart 2. while in results of k mean, we note in chart 4. that yellow cluster contains points which are not connected and coherent and that is compatible with k_mean concepts.

Also, both results can be compared using the standard deviation in all clusters to recognize how closely and connect each data is the cluster.

	cluster1	cluster2	cluster3	over all
k_mean	5.872	3.178	4.602	4.551
S.C.A	5.389	2.666	4.955	4.337

Figure 27. Standard deviation for all clusters of S.C.A ,k_mean

Figure 24. shows that the results of S.C.A have a standard deviation value(4.337) which less than k_mean(4.551), that is mean the points in S.C.A are more connected and consistent. , and this leads to the coloration relation being stronger.

4. CONCLUSION

The practitioners for spectral clustering are usually showing the adequate ways for selecting good parameters to tune the clustering process. This study was aiming to illustrate the advantage and characteristics of spectral clustering, and to make the k_mean algorithm clearly be applied. Therefore, there are three key points for doing some comparisons with k_mean dataset process: first, using a local scale rather than a global one, Second, estimating the scale from the data, Third, use of some statistical measures such as mean and standard deviation to differentiate between the algorithms results k means and S.C.A.

One of the priorities of his interest is to know the relationship between all the points in the datasets and to study the strength relationship for classification process and the compilation of each group in the special cluster. Through our analysis we have seen useful results for using spectral clustering algorithm where, It is identifying and ranking the clusters of any given statistics or studies in many fields such as: economic, education, people activity, social studies, criminal, finance ,industry ,healthy,Etc. and we have seen as well the impact on each other. By doing so, we have noted by applying any algorithm, or technique such as: regression and correlation, it is difficult to determine which variable is independent or dependent, or to determine which one is linear or nonlinear.

The thesis summarizes the application of two algorithms spectral clustering and k_mean on three variables based on real data from an official source in thirty-one countries in the European Union in the seven years from 2008 to 2013, which was the first study classifies the People at risk of poverty or social exclusion, while the second one is Population tertiary by education attainment level And the last is Mean and median income by work intensity of the household.

The result was that there is a direct effect of all these variables on each other by dividing these data into three different clusters into their characteristics and Specifications.

The results showed through the analysis and study for apply S.C algorithm that there is a strong correlation between the three variables as follows:

- Income and Risk of poverty: negative relation.
- Education α income: positive relation.

In rich countries where people have well-being life and high living rate, the percentage for risk of poverty declines and we noted rising the percentage of high education attainment, because people in those countries have comfort and stability life, while in the least well-Being countries, we note rise the number of people at risk of poverty and decrease percentage educational attainment, so these people be more deal with reality and Approaching to the obsessions and fears of the future, so they try to be Self-reliance by, As a Result, turning the labor market for Attempt to find a job opportunity at early age.

When we used k-mean algorithm the result was different and analogical, because the data is divided into three clusters, so that the data of each cluster is related to a specific relationship with the disregard of its coherence and continuity. So that such a group difficult to find common characteristics of all its data.

According to these results we can say S.C more comprehensive and general than k_mean, because in some cases it divide the data into some clusters but maybe one or more cluster does not include similarity and compatible characteristic data, however S.C.A have some challenges to apply like :

- May be sensitive to the choice of parameters sigma and k as we illustrated in chapter 3.
- Computationally expensive for large datasets.
- It is not easy to find real data can be applied by the algorithm .

4.1 FUTURE WORK

A lot of time had been dedicated finding a real dataset that can analyze adequately the algorithm. Therefore the future work is to develop an algorithm that can measure the percentage clearly and apply its variables on the different real dataset. Below is a brief advantage of this algorithm.

- Saving time and effort to researchers and studies.
- Obtaining the efficiency of applying the algorithm to the dataset.
- Facilitating the study and analysis of results for researchers and statisticians.
- Increasing reliability in results.

REFERENCES

- Alpert CJ, Kahng AB. (1995) Multi-way partitioning via geometric embeddings, orderings and dynamic programming. *IEEE Trans Comput-Aaid Des Integer Circuits Syst* 14(11):1342–1358.
- Chung, F. R. (1997). *Spectral graph theory* (Vol. 92). American Mathematical Soc.
- Delen, D. (2015) *Real-World Data Mining*. Pearson Education, Inc. New Jersey: USA.
- Ding, C. H., He, X., Zha, H., Gu, M., & Simon, H. D. (2001). A min-max cut algorithm for graph partitioning and data clustering. In *Data Mining, 2001. ICDM2001, Proceedings IEEE International Conference on* (pp. 107-114). IEEE.
- Ding SF, Jia HJ, Zhang LW et al. (2012) Research of semi-supervised spectral clustering algorithm based on pairwise constraints. *Neural Comput Appl*. doi:10.1007/s00521-012-1207-8.
- Driessche RV, Roose D. (1995) an improved spectral bisection algorithm and its application to dynamic load balancing. *Parallel Compute* 21(1):29–48.
- Fiedler, M. (1973). Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(2), 298-305.
- Han, J, Kamber, M, Pei, J.(2011) *Data Mining: Concepts and Techniques*. 3rd Morgan Kaufmann, Waltham: USA.
- Hendrickson B, Leland R. (1995) an improved spectral graph partitioning algorithm for mapping parallel computations. *SIAM J SciComput* 16(2):452–459.
- Kluger Y, Basri R, Chang JT et al. (2003) Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Res* 13(4):703–716.
- Malik J, Belongie S, Leung T et al. (2001) Contour and texture analysis for image segmentation. *Int J Comput Vis* 43(1):7–27.
- Nascimento, M. C., & De Carvalho, A. C. (2011). Spectral methods for graph clustering—A survey. *European Journal of Operational Research*, 211(2), 221-231.
- Ng AY, Jordan MI, Weiss Y. (2002) On spectral clustering: analysis and an algorithm. *Adv Neural Inf Process Syst* 14: 849–856.
- Paccanaro A, Chennubhotla C, Casbon JA. (2006) Spectral clustering of protein sequences. *Nucl Acids Res* 34(5):1571–1580.

Shi J, Malik J (2000) Normalized cuts and image segmentation. *IEEE Trans Patt Anal Mach Intell* 22(8):888–905

Xie YK, Zhou YQ, Huang XJ. (2009) A spectral clustering based conference resolution method. *J Chin Inf Process* 23(3):10–16

Zhang XR, Jiao LC, Liu F. (2008) Spectral clustering ensemble applied to SAR image segmentation. *IEEE Trans Geosci Rem Sens* 46(7):2126–2136 Zhang XR, Jiao LC, Liu F (2008) Spectral.



APPENDIX A

SOURCE CODE FOR S.C.A

The appendix includes a program written in Matlab version 14 to implement the algorithm consists of the 4 programs

1. main program (Jordan_Weiss)
2. read data from an external file (Gen_Data)
3. Calculate Affinity matrix (Cal_Affinity)

Firstly main program (Jordan_Weiss)

```
clear all;
close all;
dim=3
data = Gen_Data;
affinity = Cal_Affinity(data,dim);
for i=1:size(affinity,1)
    D(i,i) = sum(affinity(i,:));
end
for i=1:size(affinity,1)
for j=1:size(affinity,2)
    NL1(i,j) = affinity(i,j) / (sqrt(D(i,i)) * sqrt(D(j,j)));
end
end
[eigVectors,eigValues] = eig(NL1);
k =4
nEigVec = eigVectors(:,(size(eigVectors,1)-(k-1)):
size(eigVectors,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==3
figure,plot3(data(:,1), data(:,2),data(:,3),'r+'), title('Original
Data Points'); grid on;shg

figure,plot3(data(1,1), data(1,2),data(10,3),'w+'); grid on;shg

hold on
```

```

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

```

```
end
end
end
```

```
xlswrite('E:\kemberburgaz\thises\data_try\output_data.xlsx',data,'sheet1');
hold off;
title('Clustering Results using spectral clustering');
grid on;shg
tt=1;
```

secondly read data from an external file (Gen_data)

```
function [data] = GenerateData()
data=xlswrite('E:\kemberburgaz\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));
end
end
end
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
end
end
```

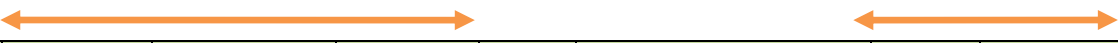
APPENDIX B

DATASET CONTENTS

The table below contains the data which have been applied by algorithms and the their results.

Input

output



ser	poverty	education	income	year	country	S.C.A	K_mean
1	20.8	28.4	20.23	2008	Belgium	3	1
2	44.8	18.9	2.554	2008	Bulgaria	1	3
3	15.3	12.4	6.539	2008	Czech Republic	2	3
4	16.3	26.3	26.422	2008	Denmark	3	2
5	20.1	21.4	20.248	2008	Germany	3	1
6	21.8	28.3	6.151	2008	Estonia	3	3
7	23.7	30.3	25.822	2008	Ireland	3	1
8	28.1	19.8	11.8	2008	Greece	2	3
9	23.8	27	14.842	2008	Spain	3	1
10	18.5	24.8	19.49	2008	France	3	1
11	19.5	13.6	0.018	2008	Croatia	2	3
12	25.5	12.7	16.743	2008	Italy	2	3
13	23.3	31	17.127	2008	Cyprus	3	1
14	34.2	20.7	5.428	2008	Latvia	2	3
15	28.3	25.3	4.644	2008	Lithuania	2	3
16	15.5	23.7	31.527	2008	Luxembourg	3	2
17	28.2	16.4	4.564	2008	Hungary	2	3
18	20.1	12.1	11.115	2008	Malta	2	3
19	14.9	27.8	20.831	2008	Netherlands	3	1
20	20.6	15	20.306	2008	Austria	3	1
21	30.5	16.5	4.331	2008	Poland	2	3
22	26	12.6	8.697	2008	Portugal	2	3
23	44.2	10.7	2.1	2008	Romania	1	3
24	18.5	19	11.455	2008	Slovenia	2	3
25	20.6	12.3	5.116	2008	Slovakia	2	3
26	17.4	30.2	21.626	2008	Finland	3	1
27	14.9	26.9	21.58	2008	Sweden	3	1
28	23.2	28.7	21.196	2008	United Kingdom	3	1
29	11.8	25.5	33.266	2008	Iceland	3	2
30	15	30.2	33.456	2008	Norway	3	2
31	18.1	28.5	27.868	2008	Switzerland	3	2
32	20.2	29.4	21.519	2009	Belgium	3	1

33	46.2	19.2	3.194	2009	Bulgaria	1	3
34	14	13.4	7.907	2009	Czech Republic	2	3
35	17.6	26.9	27.496	2009	Denmark	3	2
36	20	22.3	20.528	2009	Germany	3	1
37	23.4	30.2	6.909	2009	Estonia	3	1
38	25.7	31.4	25.68	2009	Ireland	3	1
39	27.6	19.9	12.333	2009	Greece	2	3
40	24.7	27.4	15.684	2009	Spain	3	1
41	18.5	25.9	20.235	2009	France	3	1
42	26	14.5	12.56	2009	Croatia	2	3
43	24.9	12.8	16.573	2009	Italy	2	3
44	23.5	30.5	17.393	2009	Cyprus	3	1
45	37.9	21.4	6.241	2009	Latvia	2	3
46	29.6	25.5	5.386	2009	Lithuania	2	3
47	17.8	30.2	31.978	2009	Luxembourg	3	2
48	29.6	16.9	4.879	2009	Hungary	2	3
49	20.3	12.8	11.729	2009	Malta	2	3
50	15.1	28.4	21.44	2009	Netherlands	3	1
51	19.1	16	21.497	2009	Austria	3	1
52	27.8	18.1	5.364	2009	Poland	2	3
53	24.9	13.1	8.665	2009	Portugal	2	3
54	43	11.2	2.245	2009	Romania	1	3
55	17.1	19.6	12.445	2009	Slovenia	2	1
56	19.6	13.4	6.089	2009	Slovakia	2	3
57	16.9	30.9	22.839	2009	Finland	3	1
58	15.9	27.6	22.437	2009	Sweden	3	1
59	22	30	18.385	2009	United Kingdom	3	1
60	11.6	26.5	22.58	2009	Iceland	3	1
61	15.2	30.7	35.369	2009	Norway	3	2
62	17.9	29.6	30.608	2009	Switzerland	3	2
63	20.8	30.7	21.792	2010	Belgium	3	1
64	49.2	19.7	3.329	2010	Bulgaria	1	3
65	14.4	14.5	7.586	2010	Czech Republic	2	3
66	18.3	27.5	28.38	2010	Denmark	3	2
67	19.7	22.7	20.771	2010	Germany	3	1
68	21.7	30	6.598	2010	Estonia	3	1
69	27.3	32.7	24.088	2010	Ireland	3	1
70	27.7	20.9	13	2010	Greece	2	3
71	26.1	28.4	15.811	2010	Spain	3	1
72	19.2	26.2	20.711	2010	France	3	1
73	31.1	15.7	6.642	2010	Croatia	2	3
74	25	13	16.958	2010	Italy	2	3

75	24.6	32.1	17.039	2010	Cyprus	3	1
76	38.2	22.6	5.224	2010	Latvia	2	3
77	34	26.9	4.286	2010	Lithuania	2	3
78	17.1	30.3	32.48	2010	Luxembourg	3	2
79	29.9	17.1	4.399	2010	Hungary	2	3
80	21.2	14.2	11.433	2010	Malta	2	3
81	15.1	27.7	21.735	2010	Netherlands	3	1
82	18.9	16.2	22.117	2010	Austria	3	1
83	27.8	19.4	4.632	2010	Poland	2	3
84	25.3	13.9	9.358	2010	Portugal	2	3
85	41.5	11.9	2.068	2010	Romania	1	3
86	18.3	20.2	12.318	2010	Slovenia	2	1
87	20.6	15.1	6.576	2010	Slovakia	2	3
88	16.9	31.6	23.28	2010	Finland	3	1
89	15	28.2	20.72	2010	Sweden	3	1
90	23.2	31.6	19.303	2010	United Kingdom	3	1
91	13.7	26.3	18.536	2010	Iceland	3	1
92	14.9	31.4	34.186	2010	Norway	3	2
93	17.2	30	31.834	2010	Switzerland	3	2
94	21	30.4	22.456	2011	Belgium	3	1
95	49.1	20.1	3.335	2011	Bulgaria	1	3
96	15.3	15.8	8.05	2011	Czech Republic	2	3
97	17.6	27.9	29.901	2011	Denmark	3	2
98	19.9	24.3	21.063	2011	Germany	3	1
99	23.1	31.3	6.508	2011	Estonia	3	1
100	29.4	33.3	23.57	2011	Ireland	3	1
101	31	22.2	12.144	2011	Greece	2	3
102	26.7	29.3	15.26	2011	Spain	3	1
103	19.3	26.7	20.537	2011	France	3	1
104	32.6	15.4	6.192	2011	Croatia	2	3
105	28.1	13.2	16.93	2011	Italy	2	3
106	24.6	33.7	17.747	2011	Cyprus	3	1
107	40.1	23.6	4.867	2011	Latvia	2	3
108	33.1	27.9	4.394	2011	Lithuania	2	3
109	16.8	31.7	32.587	2011	Luxembourg	3	2
110	31.5	18	4.766	2011	Hungary	2	3
111	22.1	15.1	11.922	2011	Malta	2	3
112	15.7	28	21.634	2011	Netherlands	3	1
113	19.2	16.3	22.522	2011	Austria	3	1
114	27.2	20.3	5.22	2011	Poland	2	3
115	24.4	15.5	8.912	2011	Portugal	2	3
116	40.9	12.9	2.118	2011	Romania	1	3

117	19.3	21.6	12.71	2011	Slovenia	2	1
118	20.6	16.4	6.768	2011	Slovakia	2	3
119	17.9	32.5	23.981	2011	Finland	3	1
120	16.1	29.1	23.815	2011	Sweden	3	1
121	22.7	33.2	19.27	2011	United Kingdom	3	1
122	13.7	27.4	19.537	2011	Iceland	3	1
123	14.5	32.1	38.145	2011	Norway	3	2
124	17.2	30	35.964	2011	Switzerland	3	2
125	21.6	31.3	22.881	2012	Belgium	3	1
126	49.3	20.7	3.25	2012	Bulgaria	1	3
127	15.4	17	8.421	2012	Czech Republic	2	3
128	17.5	28.6	30.015	2012	Denmark	3	2
129	19.6	24.9	21.677	2012	Germany	3	1
130	23.4	32.1	7.042	2012	Estonia	3	1
131	30.3	34.7	23.194	2012	Ireland	3	1
132	34.6	22.9	10.133	2012	Greece	2	3
133	27.2	30	14.982	2012	Spain	3	1
134	19.1	27.7	21.123	2012	France	3	1
135	32.6	15.8	6.065	2012	Croatia	2	3
136	29.9	13.9	16.686	2012	Italy	2	3
137	27.1	35	17.892	2012	Cyprus	3	1
138	36.2	25.2	5.133	2012	Latvia	2	3
139	32.5	28.6	5.048	2012	Lithuania	2	3
140	18.4	33.4	32.75	2012	Luxembourg	3	2
141	33.5	19	5.034	2012	Hungary	2	3
142	23.1	16.4	12.352	2012	Malta	2	3
143	15	28.6	22.045	2012	Netherlands	3	1
144	18.5	16.9	22.811	2012	Austria	3	1
145	26.7	21.5	5.283	2012	Poland	2	3
146	25.3	16.7	8.797	2012	Portugal	2	3
147	43.2	13.5	2.072	2012	Romania	1	3
148	19.6	23	12.746	2012	Slovenia	2	1
149	20.5	17	7.555	2012	Slovakia	2	3
150	17.2	32.8	24.925	2012	Finland	3	1
151	15.6	30.1	26.262	2012	Sweden	3	2
152	24.1	34.6	21.146	2012	United Kingdom	3	1
153	12.7	28.5	20.048	2012	Iceland	3	1
154	13.7	33	41.478	2012	Norway	3	2
155	17.5	31.2	41.686	2012	Switzerland	3	2
156	20.8	31.5	24.415	2013	Belgium	3	1
157	48	22.2	3.382	2013	Bulgaria	1	3
158	14.6	18.1	8.25	2013	Czech Republic	2	3

159	18.3	29.1	30.796	2013	Denmark	3	2
160	20.3	25.2	21.51	2013	Germany	3	1
161	23.5	32.3	7.593	2013	Estonia	3	1
162	29.9	36.3	23.649	2013	Ireland	3	1
163	35.7	24	8.97	2013	Greece	2	3
164	27.3	30.9	14.824	2013	Spain	3	1
165	18.1	28.9	21.301	2013	France	3	1
166	29.9	17	5.738	2013	Croatia	2	3
167	28.5	14.4	16.601	2013	Italy	2	3
168	27.8	35.4	16.852	2013	Cyprus	3	1
169	35.1	27	5.463	2013	Latvia	2	3
170	30.8	29.8	5.325	2013	Lithuania	2	3
171	19	35.2	33.172	2013	Luxembourg	3	2
172	34.8	19.5	4.674	2013	Hungary	2	3
173	24	17.2	13.29	2013	Malta	2	3
174	15.9	29.3	22.075	2013	Netherlands	3	1
175	18.8	17.7	22.924	2013	Austria	3	1
176	25.8	22.6	5.342	2013	Poland	2	3
177	27.5	17.6	8.757	2013	Portugal	2	3
178	41.9	13.8	2.02	2013	Romania	1	3
179	20.4	24.4	12.501	2013	Slovenia	2	1
180	19.8	17.7	7.151	2013	Slovakia	2	3
181	16	33.6	25.619	2013	Finland	3	2
182	16.4	31.4	28.062	2013	Sweden	3	2
183	24.8	35.6	20.749	2013	United Kingdom	3	1
184	13	29.3	21.807	2013	Iceland	3	1
185	14.1	34.2	44.5	2013	Norway	3	2
186	16.3	33.2	43.051	2013	Switzerland	3	2