İSTANBUL TECHNICAL UNIVERSITY ★ INFORMATICS INSTITUTE

PARALLEL CLUSTERING ALGORITHMS
WITH APPLICATION TO CLIMATOLOGY

M.Sc. Thesis by

Halil BİŞGİN

Department : Informatics Department

Programme : Computational Science and Engineering

JANUARY 2008

İSTANBUL TECHNICAL UNIVERSITY ★ INFORMATICS INSTITUTE

PARALLEL CLUSTERING ALGORITHMS
WITH APPLICATION TO CLIMATOLOGY

M.Sc. Thesis by

Halil BİŞGİN

(702041007)

Date of Submission : 27 December 2007
Date of Examin : 28 January 2008

Supervisor	:	Prof. H. NÜZHET DALFES
Members of the Examining Committee		Assoc. Prof. Zehra ÇATALTEPE (İ.T.Ü.)
		Assoc. Prof. ÖMER LÜTFİ ŞEN (İ.T.Ü.)

JANUARY 2008

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ BİLİŞİM ENSTİTÜSÜ

İKLİM BİLİMİNDE UYGULAMASIYLA
PARALEL KÜMELEME ALGORİTMALARI

YÜKSEK LİSANS TEZİ

Halil BİŞGİN

(702041007)

Tezin Enstitüye Verildiği Tarih	:	**27 Aralık 2007**
Tezin Savunulduğu Tarih	:	**28 Ocak 2008**

Tez Danışmanı	:	**Prof. Dr. H. NÜZHET DALFES**
Diğer Jüri Üyeleri		**Doç. Dr. Zehra ÇATALTEPE (İ.T.Ü.)**
		Doç. Dr. ÖMER LÜTFİ ŞEN (İ.T.Ü.)

OCAK 2008

ACKNOWLEDGEMENT

First, I'd like to thank my supervisor Prof. H. Nüzhet Dalfes who guided me throughout this study. His contributions to not only my study, but also to my academic vision made me excited and hopeful for further studies. Besides him, I thank to dear faculty members in Computational Science and Engineering Program, Asst. Prof. Lale Ergene, Asst. Prof. F. Aylin Konuklar, Prof. Metin Demiralp, Prof. M. Serdar Çelebi as well as members of examining commitee Assoc. Prof. Zehra Çataltepe and Assoc. Prof. Ömer Lütfi Şen whose support and advices helped me to reach at this stage.

Secondly, I'd like to thank my doctors Hanife Can, Firdevs Atabey, Didem Görgün who took care for me during my treatment process in Yedikule Göğüs Hastalıkları Hastanesi and Associated Prof. Muharrem Çelik, Didem Güngör from 3. Cerrahi Servisi in Dr Siyami Ersek Göğüs Kalp ve Damar Cerrahisi Eğitim ve Araştırma Hastanesi.

I also hereby would like to thank HPCC group members, Ufuk Turunçoğlu, Özden Akıncı, Hakan Gürel and Kenan Gürsoy for their precious advices. Moreover, special thanks to my colleagues Şenol Pişkin, Muzaffer Ayvaz and Semra Bayat for their motivating encouragements.

Last but not least, thanks to my family without whom I could not go further.

JANUARY 2008 Halil BİŞGİN

TABLE OF CONTENTS

ABBREVIATIONS **v**
LIST OF TABLES **vi**
LIST OF FIGURES **vii**
LIST OF SYMBOLS **ix**
SUMMARY **x**
ÖZET **xi**

1. **INTRODUCTION** **1**
    - **1.1. Defining Climate Zones** 1
    - **1.2. Cluster Analysis** 1
        - 1.2.1. Clustering Techniques 3
        - 1.2.2. Previous Work 4

2. **METHODS** **12**
    - **2.1. Instruments for Cluster Analysis** 12
        - 2.1.1. Similarity 12
        - 2.1.2. Criterion Functions For Clustering 13
        - 2.1.3. Obtaining the minimum for The Sum-of-Squared-Error criterion 14
    - **2.2. Clustering Algorithms** 16
        - 2.2.1. K-means Algorithm 16
        - 2.2.2. Hierarchical Clustering Algorithms 17
    - **2.3. Parallel Clustering Algorithms** 19
    - **2.4. Principal Component Analysis** 19
    - **2.5. Visualization Techniques** 23

3. **METHODOLOGY** **24**
    - **3.1. Data** 24
        - 3.1.1. Quality Control 24
        - 3.1.2. Interpolation of Climate Variables 25
        - 3.1.3. Data Preprocessing 26
    - **3.2. K-means on Climatology** 27
        - 3.2.1. K-means procedure 28
        - 3.2.2. Number of Clusters 29
    - **3.3. Visualization** 31

4. **PARALLEL K-MEANS ALGORITHM** **32**
    - **4.1. Parallel k-means Procedure** 32
    - **4.2. Results and Conclusion** 33

5. **K-MEANS IMPLEMENTATION ON CLIMATOLOGY** **38**

5.1. Experiments with high dimensional data set 38
5.2. Utilizing PCA 52
5.3. Visualizing the distances from members to centroids 72
5.4. Results and Conclusion 74
5.4.1. With Cv and Elevation 74
5.4.2. With Cv Without Elevation 75
5.4.3. Without Cv and Elevation 75
5.4.4. Without Cv with Elevation 76

6. CONCLUSION **78**

REFERENCES **81**

CIRRICULUM VITAE **85**

ABBREVIATIONS

CV	:	Coefficients of Variation
MST	:	Minimum Spanning Tree
AVHRR	:	Advanced Very High Resolution Radiometer
hPa	:	hectopascals
NCEP	:	National Centers for Environmental Prediction
NCAR	:	The National Center for Atmospheric Research
SLP	:	Sea Level Pressure
GHz	:	Giga Hertz
HDF	:	Hierarchical Data Format
netCDF	:	network Common Data Form
GRIB	:	GRIdded Binary
CPU	:	Central Processing Unit
GB	:	GigaByte
MPI	:	Message Passing Interface
I/O	:	Input Output
HPCC	:	High Performance Computing Center

LIST OF TABLES

Page No

Table 4.1 Speed Up and Efficiency 34
Table 5.1 Validity Results for $30^o - 50^o$ N $3^o - 60^o$ E with CV and Elevation 39
Table 5.2 Validity Results for $34^o - 43^o$ N $23^o - 47^o$ E with CV and Elevation 40
Table 5.3 Validity Results for $30^o - 50^o$ N $3^o - 60^o$ E with CV without Elevation 43
Table 5.4 Validity Results for $34^o - 43^o$ N $23^o - 47^o$ E with CV without Elevation 43
Table 5.5 Validity Results for $30^o - 50^o$ N $3^o - 60^o$ E without CV and Elevation 46
Table 5.6 Validity Results for $34^o - 43^o$ N $23^o - 47^o$ E without CV and Elevation 46
Table 5.7 Validity Results for $30^o - 50^o$ N $3^o - 60^o$ E without CV with Elevation 49
Table 5.8 Validity Results for $34^o - 43^o$ N $23^o - 47^o$ E without CV with Elevation 49
Table 5.9 Eigenvalues for all data set 52
Table 5.10 Validity Results for 9 PC from $30^o - 50^o$ N $3^o - 60^o$ E with CV and Elevation 54
Table 5.11 Validity Results for 11 PC from $34^o - 43^o$ N $23^o - 47^o$ E with CV and Elevation 54
Table 5.12 Eigenvalues for all data set except elevation 57
Table 5.13 Validity Results for 9 PC from $30^o - 50^o$ N $3^o - 60^o$ E with CV without Elevation 59
Table 5.14 Validity Results for 11 PC from $34^o - 43^o$ N $23^o - 47^o$ E with CV without Elevation 59
Table 5.15 Eigenvalues for all data set without cv and elevation 62
Table 5.16 Validity Results for 9 PC $30^o - 50^o$ N $3^o - 60^o$ E without CV and Elevation 64
Table 5.17 Validity Results for 10 PC $34^o - 43^o$ N $23^o - 47^o$ E without CV and Elevation 64
Table 5.18 Eigenvalues for all data set without cv with elevation 67
Table 5.19 Validity Results for 9 PC from $30^o - 50^o$ N $3^o - 60^o$ E without CV with Elevation 69
Table 5.20 Validity Results for 10 PC from $34^o - 43^o$ N $23^o - 47^o$ E without CV with Elevation 69

LIST OF FIGURES

Page No

Figure 1.1 : Illustration of clustering [4] 2
Figure 1.2 : Holdridge Life Zone Model [8] 6
Figure 1.3 : Bayesian mixture model classification for southern Queensland with:(a) no prior, (b) moderate weighting on priors, and (c) strong weighting on priors. The priors were calculated from the existing sub-bioregions [10]. 7
Figure 1.4 : 50 distinct ecoregions for US by using 9 environmental conditions 11
Figure 2.1 : Agglomerative Clustering [33] 18
Figure 2.2 : Visualization of 3D Data Space [39] 21
Figure 2.3 : Standardized Data [39] 22
Figure 2.4 : Representation of the data set with 2 principal components [39] 22
Figure 3.1 : k-means algorithm [45] 28
Figure 4.1 : Speed Up for the 1^{st} data set with 6 clusters 35
Figure 4.2 : Speed Up for the 2^{nd} data set with 6 clusters 35
Figure 4.3 : Efficiency for the 1^{st} data set with 6 clusters 36
Figure 4.4 : Efficiency for the 2^{nd} data set with 6 clusters 36
Figure 5.1 : validity for every number of cluster 41
Figure 5.2 : validity for every number of cluster 41
Figure 5.3 : $30^o - 50^o$ N $3^o - 60^o$ E with cv and elevation 42
Figure 5.4 : $34^o - 43^o$ N $23^o - 47^o$ E with cv and elevation 42
Figure 5.5 : validity for every number of cluster 44
Figure 5.6 : validity for every number of cluster 44
Figure 5.7 : $30^o - 50^o$ N $3^o - 60^o$ E with cv without elevation 45
Figure 5.8 : $34^o - 43^o$ N $23^o - 47^o$ E with cv without elevation 45
Figure 5.9 : validity for every number of cluster 47
Figure 5.10 validity for every number of cluster 47
Figure 5.11 $30^o - 50^o$ N $3^o - 60^o$ E without cv and elevation 48
Figure 5.12 $34^o - 43^o$ N $23^o - 47^o$ E without cv and elevation 48
Figure 5.13 validity for every number of cluster 50
Figure 5.14 validity for every number of cluster 50
Figure 5.15 $30^o - 50^o$ N $3^o - 60^o$ E without cv with elevation 51
Figure 5.16 $34^o - 43^o$ N $23^o - 47^o$ E without cv with elevation 51
Figure 5.17 Eigenvalues from greatest to smallest 53
Figure 5.18 Eigenvalues from greatest to smallest 53
Figure 5.19 validity for every number of cluster 55
Figure 5.20 validity for every number of cluster 55
Figure 5.21 9 PCs from $30^o - 50^o$ N $3^o - 60^o$ E with cv and elevation . . 56
Figure 5.22 11 PCs from $34^o - 43^o$ N $23^o - 47^o$ E with cv and elevation . 56
Figure 5.23 Eigenvalues from greatest to smallest 58

Figure 5.24 Eigenvalues from greatest to smallest 58
Figure 5.25 validity for every number of cluster 60
Figure 5.26 validity for every number of cluster 60
Figure 5.27 9 PCs from $30^o - 50^o$ N $3^o - 60^o$ E with cv without elevation 61
Figure 5.28 11 PCs from $34^o - 43^o$ N $23^o - 47^o$ E with cv without elevation 61
Figure 5.29 Eigenvalues from greatest to smallest 63
Figure 5.30 Eigenvalues from greatest to smallest 63
Figure 5.31 validity for every number of cluster 65
Figure 5.32 validity for every number of cluster 65
Figure 5.33 9 PCs from $30^o - 50^o$ N $3^o - 60^o$ E without cv and elevation . 66
Figure 5.34 10 PCs from $34^o - 43^o$ N $23^o - 47^o$ E without cv and elevation 66
Figure 5.35 Eigenvalues from greatest to smallest 68
Figure 5.36 Eigenvalues from greatest to smallest 68
Figure 5.37 validity for every number of cluster 70
Figure 5.38 validity for every number of cluster 70
Figure 5.39 9 PCs from $30^o - 50^o$ N $3^o - 60^o$ E without cv with elevation . 71
Figure 5.40 10 PCs from $34^o - 43^o$ N $23^o - 47^o$ E without cv with elevation 71
Figure 5.41 $30^o - 50^o$ N $3^o - 60^o$ E without cv and elevation 73
Figure 5.42 $34^o - 43^o$ N $23^o - 47^o$ Ewithout cv and elevation 73

LIST OF SYMBOLS

$\mathbf{d}(.,.)$	:	Distance
Σ	:	Covariance Matrix
Σ^{-1}	:	Inverse Covariance Matrix
$\boldsymbol{\mu}_\mathbf{i}$	:	Mean of the i^{th} cluster
$\mathbf{J}_\mathbf{i}$	:	Squared error of the i^{th} cluster

PARALLEL CLUSTERING ALGORITHMS WITH APPLICATION TO CLIMATOLOGY

SUMMARY

How to determine the ecoregions or climate zones has been a controversial issue. Discussion appears from the debate if the selected method is objective or not. In order to prevent from subjective approaches, one has to utilize some formulations which are independent from such interferences. Cluster analysis, which is one of the famous pattern recognition tools and has hierarchical and non-hierarchical methods, contributes to the objectivity in this sense. Instead of relying on any expertise or personal interpretations, clustering methods provide a mathematical approach with the multivariate data set.

The aim of this work is to implement cluster analysis tools to climatology data in order to obtain climate zones with some other statistical techniques that will make the study more precise. In order to clarify, first we determine how many clusters or regions do we need for valid regionalization by posing a validation criterion on the algorithm.

While acquiring such a number of clusters, we have done experiments with both the high dimensional set where there are from 96 to 109 number of variables and the reduced dimensional data space which was obtained via Principal Component Analysis (PCA). Under the criterion we posed, in the region $30^o - 50^o$ N $3^o - 60^o$ E varying number of clusters obtained as the different variable combinations are used. Nevertheless, in $34^o - 43^o$ N by $23^o - 47^o$ E where Turkey covers almost all the frame, we consistently acquired 4 climate zones. During the cluster analysis (CA), besides the serial k-means algorithm we have also utilized parallel version. According to the time measurements, it is seen that whereas serial code performs better with the reduced dimensions, parallel version is good at dealing with high dimensional sets.

Consequently, the k-means algorithm suggests another point of view for the climate zones of both regions where it is possible to observe some climatic blocks that are generally stable. More precisely, 4 climate zones appear in all cases concerning the second frame which represents some differences from the preceding climate zone definitions which are based on conventional and hierarchical ideas.

İKLİM BİLİMİNDE UYGULAMASIYLA PARALEL KÜMELEME ALGORİTMALARI

ÖZET

Ekolojik sınırların nasıl belirleneceği, iklim sınıflandırmalarının nasıl yapılacağı uzun zamandır süregelen bir takım tartışmalara konu olmuştur. Tartışmanın çıkış noktası başvurulan yöntemin ne derece tarafsız olduğuna dair görüş ayrılıklarıdır. İşte bir takım yanlı olabilecek yaklaşımlardansa, böylesi müdahalelerin önlenebildiği formulasyonlar kullanılması gerekmektedir. Veri madenciliğinin önde gelen yaklaşımlarından olan, hiyerarşik ve hiyerarşik olmayan teknikleri de içeren kümeleme yöntemi bu açıdan bakıldığında bize objektif bir çözüm sunmaktadır. Yanlı kararlara neden olabilecek kişisel beceri veya yorumlara dayanmak yerine, kümeleme analizi metodunu kullanmak, elimizdeki çok değişkenli bir veri kümesi için matematiksel bir yaklaşım olacaktır.

Bu çalışmada, daha doğru ve kolay iklim bölgeleri edinmek için bazı istatistiksel enstrümanlarla beraber kümeleme yöntemi iklim verileri üzerinde uygulanmıştır. İlk olarak geçerli bir ayırma işlemi için algoritma üzerinde bir geçerlilik kriteri göz önüne alınmıştır. Değişken sayısının her bir deneyde 96 ile 109 arasında değiştiği hali ve Temel Bileşen Analizi (TBA) yoluyla indirgenmiş boyutlar için geçerlilik kriterinin onayladığı sayılarda iklim bölgeleri saptanmıştır. Değişken sayılarındaki bu değişim, ele aldığımız $30^o - 50^o$ K $3^o - 60^o$ D bölgesinde farklı sayılarda iklim bölgeleri önerirken, Türkiye'nin tamamına yakınını kapladığı $34^o - 43^o$ K $23^o - 47^o$ D bölgesinde devamlı olarak 4 iklim bölgesi saptamaktadır. Bu süreç ele alınırken, seri bir algoritmanın yanında paralelleştirilmiş k-ortalama uygulaması kullanılarak performansı gözlenmiştir. Uygulama neticesinde seri kodun TBA ile elde edilmiş veri kümesiyle çalışması daha kolayken, paralel prosedürün yüksek boyutlu küme ile daha iyi sonuçlar verdiğini görülmüştür.

Sonuç olarak k-ortalama algoritması $30^o - 50^o$ K $3^o - 60^o$ D ve $34^o - 43^o$ K $23^o - 47^o$ D bölgelerinin iklim sınıflandırmalarına yeni bir anlayış getirmiş, daha önce yapılmış olan bölgelendirmelerden farklı olarak Türkiye coğrafyasını 4 sınıfa ayırmıştır. Her iki çerçeveye ait deneylerde Türkiye üzerindeki sınıflar genelde aynı seviyede kendini göstermiştir.

1. INTRODUCTION

The question

How can one define climate zones with multivariate data clustering?

is the main problem that we have investigated in this study.

1.1 Defining Climate Zones

Defining the climate zones has been one of the primary research areas of many scholars. Koeppen and Thornthwaite are among the famous ones who came up with their own methods [1]. The enthusiasm which has attracted so many researchers is the aim of finding the more precise and reliable model after those frontiers. In this sense, many methods have been tried to be developed and implemented to obtain models which are close to debate. However, still some discussions are still going on.

As we have noted above, these determination ways are open to discussion since those classification rules have the possibility that they are subjectively formed. In those definitions, experts have tried to integrate and weigh all of the variables and put borders relying on this view. Often they confront the weakness of their reasonings [2].

1.2 Cluster Analysis

In pattern recognition problems, procedures are named whether if they have labeled data or not. Procedures that use labeled data are said to be *supervised.* On the other hand, if a procedure deals with unlabeled data, then it is called *unsupervised* [3]. Among the unsupervised learning methods, it can be claimed that clustering is the most important one. The goal here is to discover the relationship between the unlabeled data points. More precisely, it is organizing the data set in a way that similar ones form a cluster whereas dissimilar objects fall into separate groups [4]. Here is a simple graphical example:

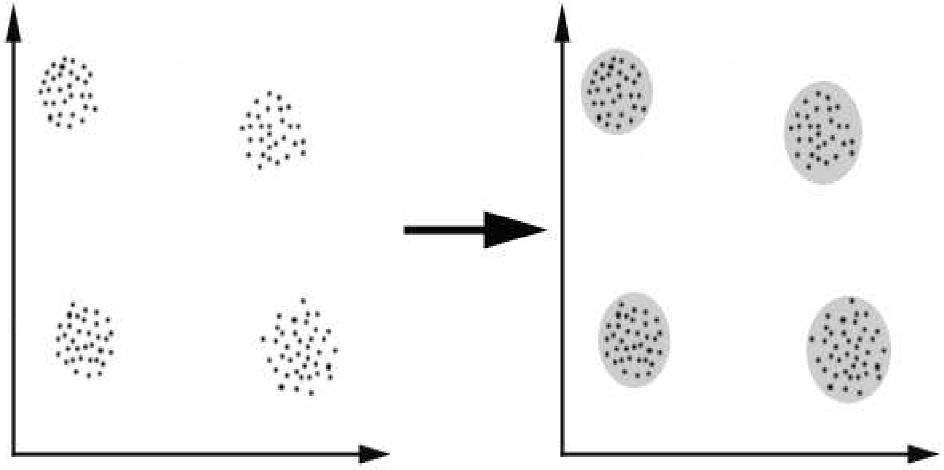


Figure 1.1: Illustration of clustering [4]

In the illustration above, we can easily realize that the data set can be partitioned into four groups. Each group consists of similar objects where similarity here is based on the distance between the data points.

Besides the example above which provides an intuition for clustering, there are some other techniques which have different point of views. These clustering methods can fit into several taxonomies. Such taxonomies are usually built by considering: (i) the input data representation,e.g numerical, categorical, special data structures, etc., (ii) the output representation, e.g a partition or hierarchy of partitions (iii) probability model used (if any), (iv) core search process, (v) clustering direction, e.g agglomerative or divisive. While many others are available, objective function of a clustering algorithm plays an important role. Actually, it can be said that objective function determines the output of the clustering procedure for a given set [5].

1.2.1 Clustering Techniques

In the previous section, we pointed out that there are several ways to cluster a data set. Moreover, all methods are capable of performing their jobs for a multivariate data set. While handling such a problem, every method uses its algorithm. Finally, we can list these algorithms which are also the names of the methods. Here is possible taxonomy suggested by Jain for Clustering Algorithms [5]:

- Heuristic-based
    - Pattern Matrix
        - Prototype-based (k-means, k-medoid)
    - Proximity Matrix
        - Linkage Methods (single link, complete link)
        - Graph Theoretic (MST, spectral clustering)
- Model Based
    - Spatial Clustering
    - Mixture Model (Gaussian Mixture)
- Density Based
    - Mode Seeking (mean shift)
    - Kernel Based (DENCLUE)
    - Grid Based (Wave-Cluster, STING)

All the listed algorithms are different approaches for an *unsupervised* procedure. Moreover, every algorithm has a cost or criterion function which they have to optimize. On the other hand, it does not mean that if they have the same input data, they have the same labeling at the end. As a matter of fact, objective functions causes big differences among the outputs of the algorithms. Jain also has a study which classifies different clustering algorithms in a way that takes into account the distances between their objective functions. $D(.,.)$ between the objective functions F_1 and F_2 on a data set X is estimated by the distance $d(.,.)$

between the data partitions $P_1(X)$ and $P_2(X)$ respectively. In other words, we have

$$D_x(F_1, F_2) = d(P_1(X), P_2(X))$$

where

$$P_1(X) = \arg\max_P F_1(P(X))$$

After defining such a procedure, Jain utilizes the classical Rand's index of partition similarity and Variation of Information (VI) distance which are both invariant w.r.t permutations of the cluster indices [5].

Finally, we observe that k-means algorithm shares the same cluster with Wards method which is one of the hierarchical methods. On the other hand, other hierarchical methods Single Linkage(SL), Average Linkage(CL), Complete Linkage (CL)are in the same group [5].

1.2.2 Previous Work

While studies were aiming contributions to zone definitions which would provide efficient policies on different areas, subjectivity of those qualitative methods was still disputable. Question arising in this dilemma was if one could use any quantitative method which would finalize such debates or drawing borders between zones should be relied on human expertise only [6]. A desire for preferring a quantitative method, of course, is not fully objective because data processing stage and interpreting the results also needs some human intervention. However, the goal here is to utilize a model which minimizes the bias from subjective point of views and develop a more transferable and consistent model.

Although those discussions may seem more recent, in 1947 a quantitative model was already on the scene. In his *The Holdridge Life Zone* model, Holdridge first comes up with a definition biotemperature. His assumption was that, from the perspective of plant physiology, there is no difference between 0 ^oC and its below because for these temperatures plants are dormant. Therefore, the life zones are defined first taking into account variable-degrees mean and annual temperature. In the second level, he combines other environmental variables. Boundaries was then determined with respect to logarithmic increases in mean annual temperature, logarithmic increases in total annual precipitation and the ratio mean annual potential evapotranspiration to mean total annual precipitation. After determining regions, he prefers hexagons in a triangular plot which can be seen on the next page [7].

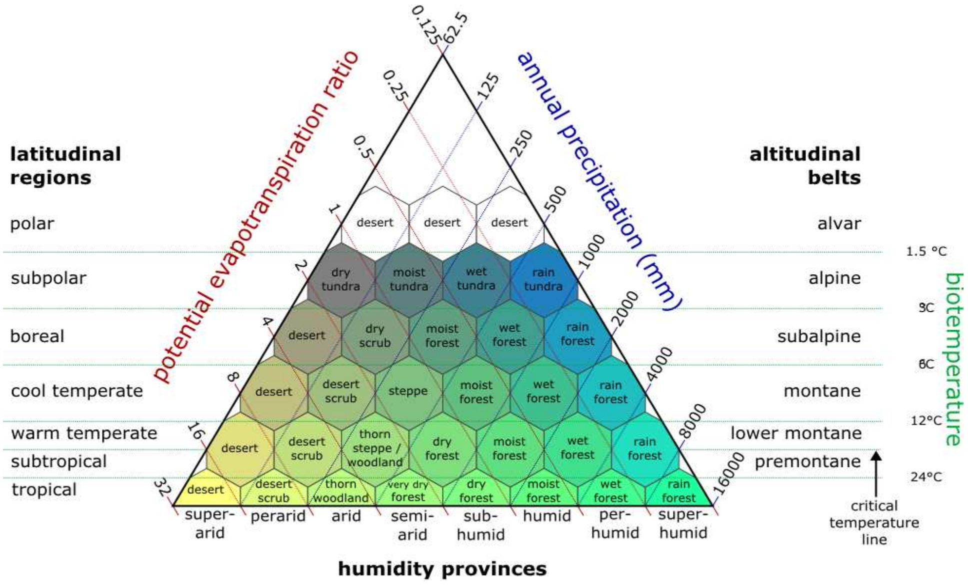


Figure 1.2: Holdridge Life Zone Model [8]

Both in the past and recent studies, many researchers have adopted quantitative ways while dealing their multivariate data sets. Several local or global investigations can be listed which considers a quantitative idea as a main approach. While determining borders between groups, neural networks, Bayesian approaches, clustering analysis and principal components analysis for a preprocessing stage have been the pioneering quantitative patterns. A local study regarding Puerto Rico was done via neural networks along with principal component analysis [9]. Climate data from 18 stations were reduced to five principal components of seasonal averages of climate data. Then they succeeded in splitting Puerto Rico into four climate zones.

Dealing with an elicitation problem, one can also handle it with the help of Bayesian statistical modeling. On the environmental variables from the target region Normal density estimate was done in Pullar's study [10]. Pullar and others, balance the prior information with the data classification of environmental data sets using Bayesian statistical modeling approach. That method is called model-based clustering which fits a Normal mixture model to the clusters associated with regions. Results of the research were applied to eastern bioregions within the state of Queensland in Australia. The delineation they obtained is the following:

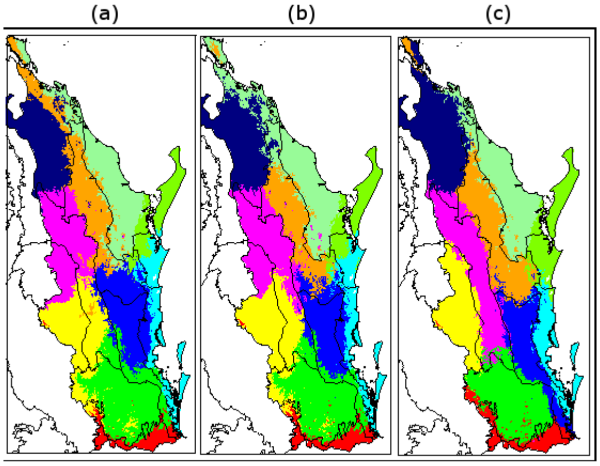


Figure 1.3: Bayesian mixture model classification for southern Queensland with:(a) no prior, (b) moderate weighting on priors, and (c) strong weighting on priors. The priors were calculated from the existing sub-bioregions [10].

For identification of representative areas of national ecosystem, climatic and edaphic factors were exploited since they were also helpful in modeling species distribution [11]. Bernet and others chose using small-scale digital data to quantify spatial relation among the environmental attributes. They grouped the data via cluster analysis and multidimensional scaling. The method was applied on previously defined ecological zone, the Western Belt of the Central United States with the inputs soil associations, AVHRR remote imagery and a combined data partition of landform, forest and soils data [12].

Host and others also among the ones who were complaining about the map units of the classification systems which had resulted from subjectively drawn boundaries [13]. Moreover, they suggested that those delineations were derived by consensus and with unclear selection and weighting inputs. Then they utilized geographic information systems (GIS) with multivariate statistical analysis combined another statistical tool, principal component analysis (PCA) for the input data from northwestern Wisconsin.

In classification of geological objects, multivariate clustering was also used. Harff and Davis, obtained homogenous regions of geology using some environmental characteristics [14]. A local application took place for western Kansas. Variables are the geologic properties and the measure for similarity was Mahalanobis distances. To have regions of uniform crop yields, Lark refers to clustering a multivariate data set to have classes with a spatially coherent distribution [15]. Another local study is about western Kenya where soil fertility management is the main theme [16]. In addition to environmental variables, human variables were also processed for this research. In the first part of the work, variable selection was done whereas in the second part multivariate statistical techniques were implemented to construct the stratification.

Jensen and his colleges came up with the map that had 84 subregions within Colorado River Basin [17]. What they preferred was hierarchical clustering algorithm which is Ward's method in particular. In this work, 19 indirect biophysical variables identified were benefited to produce an ecological clustering of 7462 subwatersheds in the region. Implementing this agglomerative technique 84 hydrologic subregions were obtained.

To be able to reduce the judgement biases and uncertainty of manual analyses, Zhou and others were after an objective mapping [18]. The aim was to be able to delineate ecoregions at multiple levels. By using not only the satellite data, but also climate and soil information they concluded the hierarchical processing of 2024 polygons with 66 and 23 regions in Nebraska.

An environmental domain classification was also done in New Zeland [19]. To be more specific, environmental domains were created at 1 km resolution. Procedure here was done in two stages of multivariate clustering with a data set which included 10 climatic and landform variables affecting plant physiological process. In the first stage, Leathwick and others used non-hierarchical clustering to get 350 zones. In the second part, they applied agglomerative clustering which was a hierarchical pattern. Finally, they produced a tree showing 20 domains. In both clustering technique, they take Gower metric to calculate the distances.

Esteban and others investigated the atmospheric circulation patterns related to heavy snowfall days in Andorra, Pyrenees [20]. Synoptic-scale atmospheric events were thought the cause of some avalanches. Referring to the intensity of at least 30 cm of snow in a 24 h period, they first applied PCA, then the clustering technique *k-means*. Rejection of the iterations were also proposed. In this study, the synoptic-classification of every heavy snowfall day, and composite maps were constructed for sea level pressure, 500 hPa geopotential height and 1000-500 m thickness. Study showed that there were seven circulation patterns, which mostly had Atlantic component of wind. In a snowfall forecasting case those results were expected to assist.

Another study regarding atmospheric circulation was also done by Esteban and his colleagues like the work above. Again, they applied both PCA and CA. In contrast to the previous research, they aimed to characterize the daily surface synoptic circulation patterns over the region 30^o N-60^o N by 30^o W-15^o E for the time interval 1960-2001 [21]. NCEP/NCAR Reanalysis Project Data was the input. K-means algorithm was the clustering instrument imposing a criterion for the number of clusters. Twenty SLP circulation patterns were acquired. Furthermore, the composite maps for SLP and 500 hpa geopotential height, the monthly distribution and long term variability for every circulation patterns was obtained.

In addition to heavy snowfall research, daily rainfall patterns in a Mediterranean area had become a source of curiosity for Penãrrocha and others [22]. They were pursuing the classification of daily rainfall patterns in the Valencia region which gets a high level of rainfall with a precipitation level 800 mm. In this work, one of the two perspectives were emphasizing on torrential rain events between 1971-95 acquired geographical distributions of the daily precipitation maxima whereas in the second one PCA and CA were visited. From both perspectives, they got consistent results reflecting the main characteristics of the daily precipitation patterns.

Climate clustering via k-means method was employed on temperature data besides precipitation for Canada [23]. Spatially coherent patterns of variations between two decades 1976-85 and 1986-95 was a conclusion of previously done research. In this, works the aimed to expand previous results with finer time intervals which are also varying. Moreover, to get more accurate results they had a greater number of stations. During the grouping process, k-means algorithm was implemented here, too. Authors, then had variations in seasonality of temperature and precipitation.

Hargrove, whom we were initially inspired from has also several studies which are mainly about defining ecoregions. With Luxmoore, he has a research on a spatial clustering technique for the identification of customizable ecoregions where they used statistical analysis package (SAS) linked with GIS with 50-year mean monthly precipitation, total plant-available water content of soil, total organic matter in soil, total Kjeldahl soil nitrogen and elevation as input variables [24]. Hargrove's another study with Hoffman on defining the ecoregions is involving another contribution. At this time, they parallelized an existing k-means algorithm for a huge data set. They developed a code on a highly heterogeneous Beowulf-class parallel machine constructed from surplus 486- and Pentium-based PCs. Not only reducing the time consumption via parallel implementation, but also quantifying representativeness and edge characteristics were other components that made this study different [25]. Figure 1.4 is an example from their ecoregion classification approach. In addition to this study, Hargrove was involved another research with Mahinthakumar, Hoffman and Karonis where GLOBUS which is a metacomputing software toolkit was used to achieve the parallel version [26].

Within the region where Turkey is focused on, there are some previous definitions of zones also. First is the one which was accepted by Turkish scholars and was defined via conventional approaches and still in use [27]. Another study done by Unal and others has suggested 8 climatic zones which were obtained relying on max, min, mean temperatures and precipitation variables processed via Ward's method [27]. Whereas in the previous studies, the Aegean and Marmara regions were considered as separate zones, in this cluster analysis they have found that those two districts are in the same climatic division. As they used 4 variables, they also made experiments with mean temperature and precipitation separately.

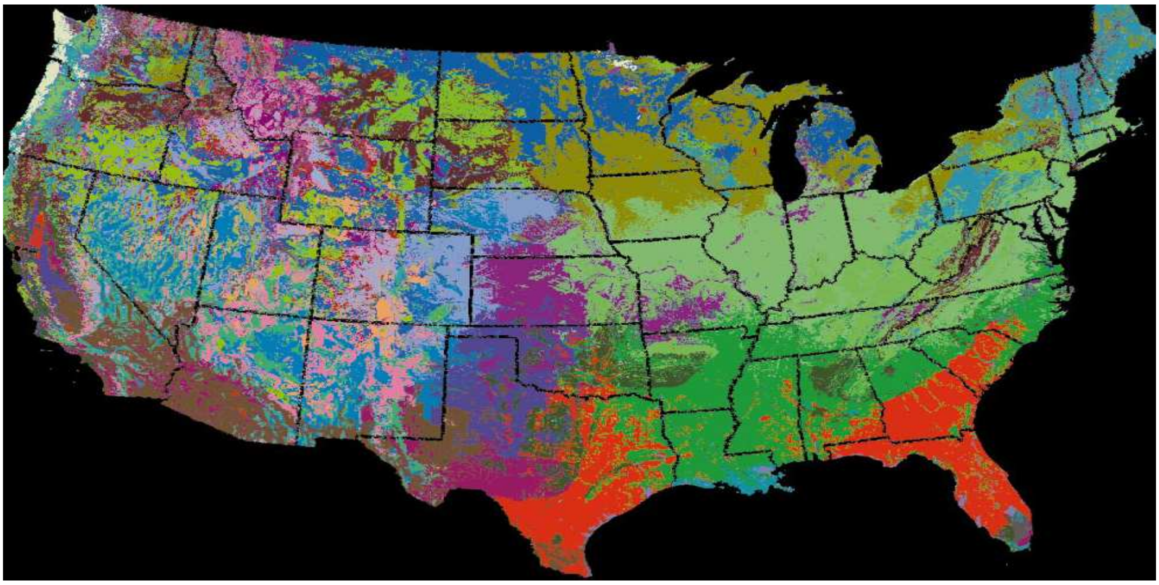


Figure 1.4: 50 distinct ecoregions for US by using 9 environmental conditions

2. METHODS

In this chapter, we are going to introduce some basic statistical instruments that we used during our study. First, the cluster analysis will be discussed by explaining the details of two main clustering algorithms, k-means algorithm and hierarchical algorithm. Secondly, since those methods are for dealing with multivariate data sets, principal component analysis is another mathematical modeling to achieve a reduction in the dimension.

2.1 Instruments for Cluster Analysis

2.1.1 Similarity

Since our job here is to decide on the membership of a sample from the data set, we first have to be able to measure the similarity. The most apparent answer for this question is to select distance metric d where d can be defined in many ways. Here are some of the best known distance measures:

- Minkowski Metric
- Euclidean Metric
- Manhattan Metric
- Mahalanobis Distance

Minkowski metric is a more general form where some others can be extracted from.

$$d(x,x') = (\sum_{k=1}^{d} |x_k - x'_k|^q)^{1/q}$$

where $d(x,x')$ is the distance between x and x'.

The Euclidean metric is a particular case of Minkowski metric. In this case we have the distance as following:

$$d(x,x') = (\sum_{k=1}^{d} |x_k - x'_k|^2)^{1/2}$$

While doing the calculations, considering the similarity in terms of the squared Euclidean does not matter.

Manhattan metric which can also be intuitively seen from Minkowski metric is

$$d(x,x') = \sum_{k=1}^{d} |x_k - x'_k|$$

It is also known as *taxicab distance.*

Mahalanobis distance is

$$d(x,x') = (x - x')^t \Sigma^{-1} (x - x')$$

where Σ^{-1} is the inverse of the covariance matrix [28].

2.1.2 Criterion Functions For Clustering

As we pointed out before every clustering algorithm has criterion or cost function that is to be optimized. We also know that every clustering may have different outcomes depending on the objective function. The Sum-of-Squared-Error criterion, related Minimum Variance Criteria, Scatter Criteria, The Trace Criterion, The Determinant Criterion and Invariant Criterion are the criteria that serve for different clustering techniques [3]. In this part we are going to depict The Sum-of-Squared-Error criterion which plays an important role for grasping the idea behind k-means clustering.

Assume that we have set $X = x_1, ..., x_n$ which is *unsupervised* and want to split this set into k partitions. Let n_i be the number and μ_i be the means of the samples belonging to G_i where

$$\mu_i = \frac{1}{n_i} \sum_{x \in G_i} x \qquad (2.1)$$

So we have the squared error summation in the following form:

$$J_e = \sum_{i=1}^{k} \sum_{x \in G_i} \|x - \mu_i\|^2 \tag{2.2}$$

As one can realize, first the differences are taken within every partition. Secondly, squared differences or distances are summed up over the whole set and this is called the *Sum-of-Squared-Error* The goal in the following section will be to *minimize* this error.

2.1.3 Obtaining the minimum for The Sum-of-Squared-Error criterion

After deciding on the appropriate criterion function a discrete optimization takes place to find the minimum. One of the iteration steps, we obtain an allocation which provides sum of squared errors at a desired level. Although the possibilities of being trapped in a local minima, accepting a solution which may not be the best and finding different solutions depending the initial points exist, computational aspect which seems to be handled easily makes iterative optimization appealing.

The notations from preceding part are put into this form:

$$J_e = \sum_{i=1}^{k} J_i \tag{2.3}$$

where we have J_i's like before

$$J_i = \sum_{x \in G_i} \|x - \mu_i\|^2 \tag{2.4}$$

and the μ_i, mean of each partition G_i, is

$$\mu_i = \frac{1}{n_i} \sum_{x \in G_i} x \tag{2.5}$$

Suppose that $\hat{x}$ is in G_i at some time of the iteration and may be moved to G_j. With a new member, mean of G_j becomes

$$\mu_j = \mu_j + \frac{\hat{x} - \mu_j}{n_j + 1} \tag{2.6}$$

which leads a change in the sum of squared errors.

$$
\begin{aligned}
J_j^* &= \sum_{x\in G_j} \|x-\mu_j^*\|^2 + \|\hat{x}-\mu_j^*\|^2 && (2.7)\\
&= (\sum_{x\in G_j} \|x-\mu_j - \frac{\hat{x}-\mu_j}{n_j+1}\|^2) + \|\frac{n_j}{n_j+1}(\hat{x}-\mu_j)\|^2 && (2.8)\\
&= J_j + \frac{n_j}{n_j+1}\|\hat{x}-\mu_j\|^2 && (2.9)
\end{aligned}
$$

Since the sample $\hat{x}$ is excluded from G_i, both μ_i and J_i is affected. In a similar fashion above, one can update μ_i. Let μ_i^* be the new mean after move. Then,

$$\mu_i = \mu_i - \frac{\hat{x}-\mu_i}{n_i+1} \tag{2.10}$$

and

$$J_i^* = J_i - \frac{n_i}{n_i-1}\|\hat{x}-\mu_i\|^2 \tag{2.11}$$

Finally, we are about to decide whether to move the sample from one cluster to another is worthwhile. To accept such a movement from G_i to G_j, an increase in the sum of squared errors must not occur. Namely if the following comparison holds, one can have the advantage of moving sample $\hat{x}$.

$$\frac{n_i}{n_i-1}\|\hat{x}-\mu_i\|^2 > \frac{n_j}{n_j+1}\|\hat{x}-\mu_j\|^2 \tag{2.12}$$

From the equation above we can construct a motivation for k-means algorithm by concentrating on the factors regarding the means. It can be claimed that such a condition is mostly have a tendency to occur when $\hat{x}$ is closer to μ_j than μ_i. Here is the sketch of the procedure [3]

```
begin initialize $n, k, \mu_1, \mu_2, ..., \mu_k$
        do randomly select a sample $\hat{x}$
          $i \leftarrow \arg\min_{i'} \|\mu_{i'} - \hat{x}\|$ (classify $\hat{x}$)
            if $n_i \neq 1$ then
                            $\rho_j = \frac{n_j}{n_j+1}\|\hat{x}-\mu_j\|^2 \qquad j \neq i$
```

```
                    $\rho_j = \frac{n_j}{n_j-1}\|\hat{x}-\mu_i\|^2 \qquad j=i$
                if $\rho_\ell \leq \rho_j \ \forall j$ then move $\hat{x}$ to $G_\ell$ and calculate $J_e$, $\mu_i$, $\mu_\ell$
            until no change in $J_e$ in $n$ attempts
        return $\mu_1, \mu_2, ..., \mu_k$
end
```

2.2 Clustering Algorithms

2.2.1 K-means Algorithm

In the previous section, we acquired a stepwise algorithm where update is done after each sample is reclassified. On the other hand with the same idea an update is also possible after n samples are reclassified. This method is called $k-means$ which has advantages with respect to the iterative optimization procedure which sustains some disadvantages. It is more preferable due to the stuck risk of iterative optimization approach. Moreover, problems depending on the order in which the candidates selected are also the weak sides of iterative optimization when compared to $k-means$.

Relying on the squared error approach k-means is the simplest fashion for clustering [28]. In addition to its simplicity, as the sample size increases, conditions are found that ensure the almost sure convergence [29].

The procedure starts with a random initial assignments for clusters centers and the classifications continues with respect to the similarity between the present centroid and the pattern. After the comparison of the vicinity of the pattern to all centroids, it is moved to the group which has the nearest one and a new *mean* is computed. This process is allow to go on until the convergence criterion is satisfied [30].

```
begin initialize $n, k, \mu_1, \mu_2, ..., \mu_k$
        do classify $n$ samples according to nearest $\mu_i$
            update $\mu_i$
        until the convergence criterion is met
    return $\mu_1, \mu_2, ..., \mu_k$
end
```

In addition to the superiorities of k-means on squared error procedure, it has also a attractive side in terms of complexity when compared some other clustering techniques. Having a more bearable computational burden $O(n)$, it is seen an easily implemented algorithm whereas, for instance, hierarchical methods require $O(n^2)$ where n is the number of items to be clustered.

A major problem with k-means is the outcome also depends on the initialization procedure. In addition, distribution of the data sometimes leads different outcomes with different initial assignments [28]. Another, point which is open to discussion is the validity of the cluster numbers. Namely, we should also decide how well the cluster arrangement is appropriate to finalize the procedure. In the next chapter, we will also focus on this problem.

2.2.2 Hierarchical Clustering Algorithms

Hierarchical clustering is often portrayed as the better quality clustering approach, but is limited because of its quadratic time complexity [31]. Its quality may be observed if an appropriate distance metric can be defined to obtain the similarity, in this case a *distance matrix*. From this point of view, hierarchical clustering is the ideal method of clustering, but has not been preferred much due to an $O(n^2)$ complexity [32].

There are two major types of hierarchical techniques: divisive and agglomerative. Agglomerative hierarchical techniques are the more commonly used. The idea behind this set of techniques is to start with each cluster comprising of exactly one object and then progressively agglomerating (combining) the two nearest clusters until there is just one cluster left consisting of all the objects. Nearness of clusters is based on a measure of distance between clusters. All agglomerative methods require as input a distance measure between all the objects that are to be clustered. This measure of distance between objects is mapped into a metric for the distance between clusters (sets of objects) metrics for the distance between

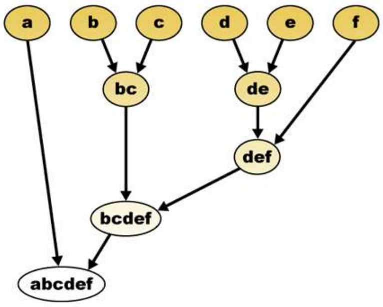


Figure 2.1: Agglomerative Clustering [33]

two clusters. The only difference between the various agglomerative techniques is the way in which this inter-cluster distance metric is defined. [32]

In order to measure inter-clusters distances, there are three graphical tools. These graphical methods are single linkage, complete linkage and average linkage methods.

a. Single Link: The distance between any two clusters is the minimum distance between two points such that one of the points is in each of the clusters.

b. Complete Link: The distance between any two clusters is the maximum distance between two points such that one of the points is in each of the clusters.

c. Average Link: The distance between any two clusters is the average distance between two points such that each pair has a point in both clusters [34].

For the Single Linkage case we have the procedure below:

1. Determine and store the distance between each pair of clusters. (Initially, each point is considered a cluster by itself) Also, for each cluster determine its nearest neighbor.
2. Determine the pair of clusters with the smallest distance between them and agglomerate them.
3. Update the pairwise distances and the new nearest neighbors.
4. If more than one cluster still exits goto Step2.

2.3 Parallel Clustering Algorithms

Although hierarchical clustering algorithms have a higher computational complexity w.r.t k-means algorithm, there have been many studies to reduce complexities of both techniques via parallelization. Besides the technical report prepared by Olson in which it is stated that complexity of the hierarchical clustering procedure is reduced to $O(n \log n)$ on an appropriate architecture and network, Zhihua and Lin have acquired the same complexity in a recent study with $n/\log n$ number of processors [35]. Nearby the efforts for handling hierarchical clustering, k-means clustering has been also parallelized despite its low complexity.

In a physical problem concerning N-body simulation, k-means algorithm has been utilized in a parallel fashion where weights of the observations were also included [36]. Furthermore, studies have also done to enhance the performance of k-means method. Jinlan and others have aimed to reduce the number of iterations via refining the initial centroids with a parallel version also [37].

Likely, we will also prefer using parallel k-means algorithm where we become capable of handling a large data set more easily. For a high dimensional variable collection, data partitioning will lower the size of work per processor which provides us save time.

2.4 Principal Component Analysis

Multivariate data analysis may seem to be a tedious job while dealing with a high dimensional data space. Visualizing, processing such a huge data set becomes a cumbersome and time consuming process many times. Therefore, *Principal Component Analysis* is preferred as an easy and reliable way to be able to overcome such obstacles. Moreover, it provides a reduction which is a loss of information at a minimized level [38]. It has a wide range spectrum that can be utilized. From neuroscience to computer graphics; from environmental sciences to bioinformatics *PCA* is an appealing statistical tool due to its simple nature and the ability to extract relevant information from confusing data sets.

The solution that we will sketch was derived Hotelling in 1933 [40]. The idea behind *PCA* is *projection.* We project our raw data to another space. Namely, let X be our data set in this case, then the projection of X the direction of w is: $Z = w^T X$. Since we are looking for the most representative bases, we have to find w such that $Var(z)$ is maximized. After some algebra as follows, we obtain an equation where the selections of w is seen.

$$\begin{aligned} Var(Z) = Var(w^T X) &= E[(w^T X - w^T \mu)^2] && (2.13) \\ &= E[(w^T X - w^T \mu)(w^T X - w^T \mu)] && (2.14) \\ &= E[w^T (X - \mu)(X - \mu)^T w] && (2.15) \\ &= w^T E[(X - \mu)(X - \mu)^T] w = w^T \Sigma w && (2.16) \end{aligned}$$

where we conclude that $Var(X) = E[(X - \mu)(X - \mu)^T] = \Sigma$

To get a unique solution, assume that $\|w\|$=1 and solve the equation below:

$$\max_{w_1} w_1^T \Sigma w_1 - \alpha(w_1^T w_1 - 1) \tag{2.17}$$

If we take the derivative of the equation (2.17) and equate it to zero, we have

$$2\Sigma - 2\alpha w_1 = 0 \tag{2.18}$$

$$\Sigma w_1 = \alpha w_1 \tag{2.19}$$

One can realize from (2.19) that we have a *eigenvalue* equation. Namely, w_1 is *eigenvector* of Σ

Finally, to make $Var(z)$ maximized $Var(w^T X)$ must be maximized by choosing the largest *eigenvalue* since

$$Var(z) = Var(w^T X)$$

When we want to get the second principal component, it means that we are now after w_2 which makes $Var(z_2)$ maximized under the constraints $\|w_2\| = 1$ and, of course, orthogonal to w_1. So,

$$\max_{w_2} w_2^T \Sigma w_2 - \alpha(w_2^T w_2 - 1) - \beta(w_2^T w_1 - 0) \tag{2.20}$$

Again taking the derivative w.r.t w_2^T in this case and multiplying by w_1, what we have now is the *eigenvalue* equation for w_2.

$$\Sigma w_2 = \alpha w_2 \tag{2.21}$$

Iteration above is done as until the desired number of orthonormal bases are obtained. All the work done above can be summarized in $Z = W^T(X - \mu)$ where the columns of W are the eigenvectors of Σ and μ is the mean of sample X [38]. Figures below also illustrates the process.

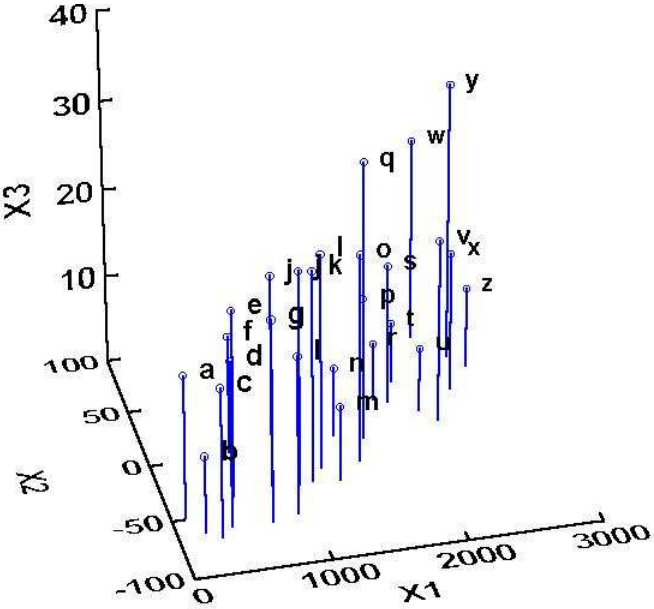


Figure 2.2: Visualization of 3D Data Space [39]

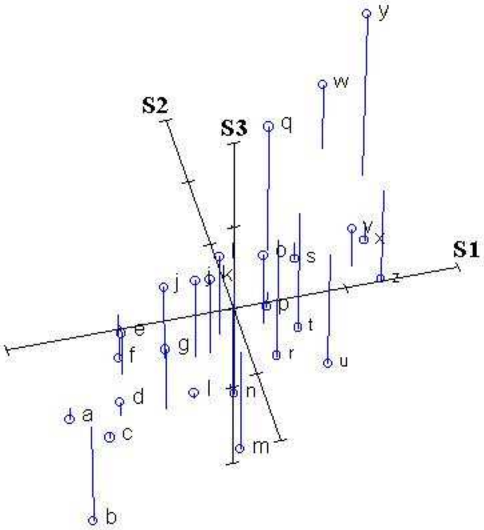


Figure 2.3: Standardized Data [39]

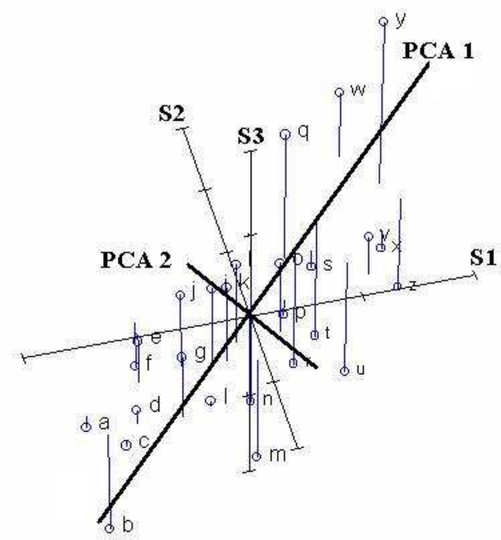


Figure 2.4: Representation of the data set with 2 principal components [39]

2.5 Visualization Techniques

In visualization process, we preferred **NCL** toolbox where it stands for The NCAR Command Language. NCL is a product of the Computational and Information Systems Laboratory at the National Center for Atmospheric Research (NCAR), is a free interpreted language designed specifically for scientific data processing and visualization.

NCL has a wide variety input output options. It can read an write netCDF-3, netCDF-4, HDF4, binary, and ASCII data, and read HDF-EOS2, GRIB1 and GRIB2.

It can be run not only in Linux environment, but also can be run on Solaris, AIX, IRIX, MacOSX, Dec Alpha, and Windows with Cygwin/X platform.

NCL can be run in interactive mode, where each line is interpreted as it is entered at your workstation, or it can be run in batch mode as an interpreter of complete scripts. One can also use command line options to set options or variables on the NCL command line.

In the areas of file input and output, data analysis, visualization, the power and utility of the language are noticeable.

NCL has many common features with other programming languages, in terms of types, variables, operators, expressions, conditional statements, loops, and functions and procedures. In addition to common programming features, NCL also has features that are not found in other programming languages, including features that handle the manipulation of metadata, the configuration of the visualizations, the import of data from a variety of data formats, and an algebra that supports array operations [41].

3. METHODOLOGY

3.1 Data

In our experiments, we benefit from the Climatic Research Unit data set which has $10^{'}$ latitude/longitude resolution of monthly mean surface climate over global land areas except Antarctica [42]. Within climatology, there are eight climate components which are *precipitation*, *wet-day frequency*, *temperature*, *diurnal temperature range*, *relative humidity*, *sunshine duration*, *ground frost frequency* and *wind speed*. In addition to those variables, coefficients of variation of precipitation is also calculated. From the same study, it is noticed that data are interpolated from a data set of station means for the period 1961-1990.

A similar study were done in a lower resolution. In other words, this work represents an improvement on an earlier gridded climatology at $30^{'}$ lat/lon resolution through an increased spatial resolution. Contribution of additional station data and inclusion of precipitation variability which was calculated through a probability distribution of monthly precipitation [42].

3.1.1 Quality Control

The CRU work has also used World Meteorological Organization and National Meteorology Agency data collections. The data obtained from those sources were subjected to a comprehensive series quality control checks by National Climatic Data Center (NCDC) and NMA respectively. However, all data were also checked through a 2-stage check [42].

1. Standard series of automated tests were done on individual station normals

    (a) internal consistency checks, e.g ensuring that the monthly mean follow a consistent seasonal cycle and the predefined absolute limits are not exceeded

(b) between-variable consistency tests, e.g ensuring that monthly minimum, mean and maximum temperatures are consistent and that months with zero precipitation have zero wet-days.

2. During the interpolation of station data interpolation diagnostics were enabled in order to identify station-months that had large residuals, and were potentially in error. As a general rule, data that failed these QC tests were removed from the interpolation. In some cases, however, the data could be compared and replaced with normals calculated from the CRU monthly station time-series which was also described in this work.

3.1.2 Interpolation of Climate Variables

The station climate statistics were interpolated using thin-plate smoothing splines (ANUSPLIN) developed by Mike Hutchinson at the Australian National University. It is a technique which was originally described by Wahba in 1979 and its robustness may be observed in areas with sparse or irregularly spaced points [42].

Trivariate thin-plate spline surfaces were fitted as functions of latitude, longitude and elevation to the station data over several regional domains. Taking into account the elevation, topographic controls were also enabled over climate. Moreover, for some variables, for instance precipitation, there is a huge number of station. Due to the memory constraints of the computers, authors preferred to do the interpolation over sub-areas instead of the large continental domains. Since sub-areas were determined considering the overlapping areas, discontinuities were avoided after merging process [42].

The spline-fitting program made them have the stations where the largest residuals from the fitted surfaces were observed. Via this labeling, they were able to identify and check potentially erroneous stations. Nonetheless, in some cases those were accepted to be correct and their positions as outliers were seen as some results that come from local climatological variations which could not be resolved with available network. On the other hand, a number of stations that apparently have some mistakes including inaccurate locational or elevation information, typographic errors in which data for a single month did not fit in with the seasonal pattern. Stations in this pattern were corrected or, excluded from the interpolation set if not possible to overcome it.

Like many numerical approaches, interpolation done here also has some errors in it. For geostatistical interpolation to be considered well, the data should have some spatial predictability. Variability in the data that is not predictable (in this case, variability that is not a function of latitude, longitude and elevation) is accepted as noise. If the data set has much noise, its predictive error also will be greater as one moves from control stations.

In the derivation of CRU data set, Lume and others have used *generalized cross validation* technique (*GCV*) to obtain an estimate of predictability. GCV procedure is simply removing each data point in turn and summing, with appropriate weighting, the square of the differences between the omitted point and that predicted by a surface fitted using all the other points [42].

3.1.3 Data Preprocessing

All the variables listed above are in ASCII file format with $10^{'}$ latitude/longitude resolution in a gridded pattern as follows:

1. All grid files except elevation (elv) and precipitation includes latitude, longitude, 12 monthly values (Jan to December) where latitudes and longitude are in degrees.
   format='(2f9.3,12f7.1)'
   Example (first line of temperature file):
   -59.083 -26.583 0.2 0.3 0.2 -1.9 -6.0 -9.8 -13.6 -9.2 -8.1 -5.3 -2.3 -1.1

2. The file for precipitation includes latitude, longitude, 12 monthly means of precipitation, 12 monthly CVs of precipitation
   format='(2f9.3,24f7.1)'
   Example (first line of precipitation file):
   -59.083 -26.583 105.4 121.3 141.3 146.7 159.6 162.4 141.5 151.1 141.6 124.9 110.0 93.9 35.2 38.7 38.4 27.5 49.5 40.8 50.8 33.5 42.2 56.6 35.5 43.4

3. The file for elevation has latitude, longitude, elevation
   format='(3f9.3)'
   Example (first line of elevation file):
   -59.083 -26.583 0.193

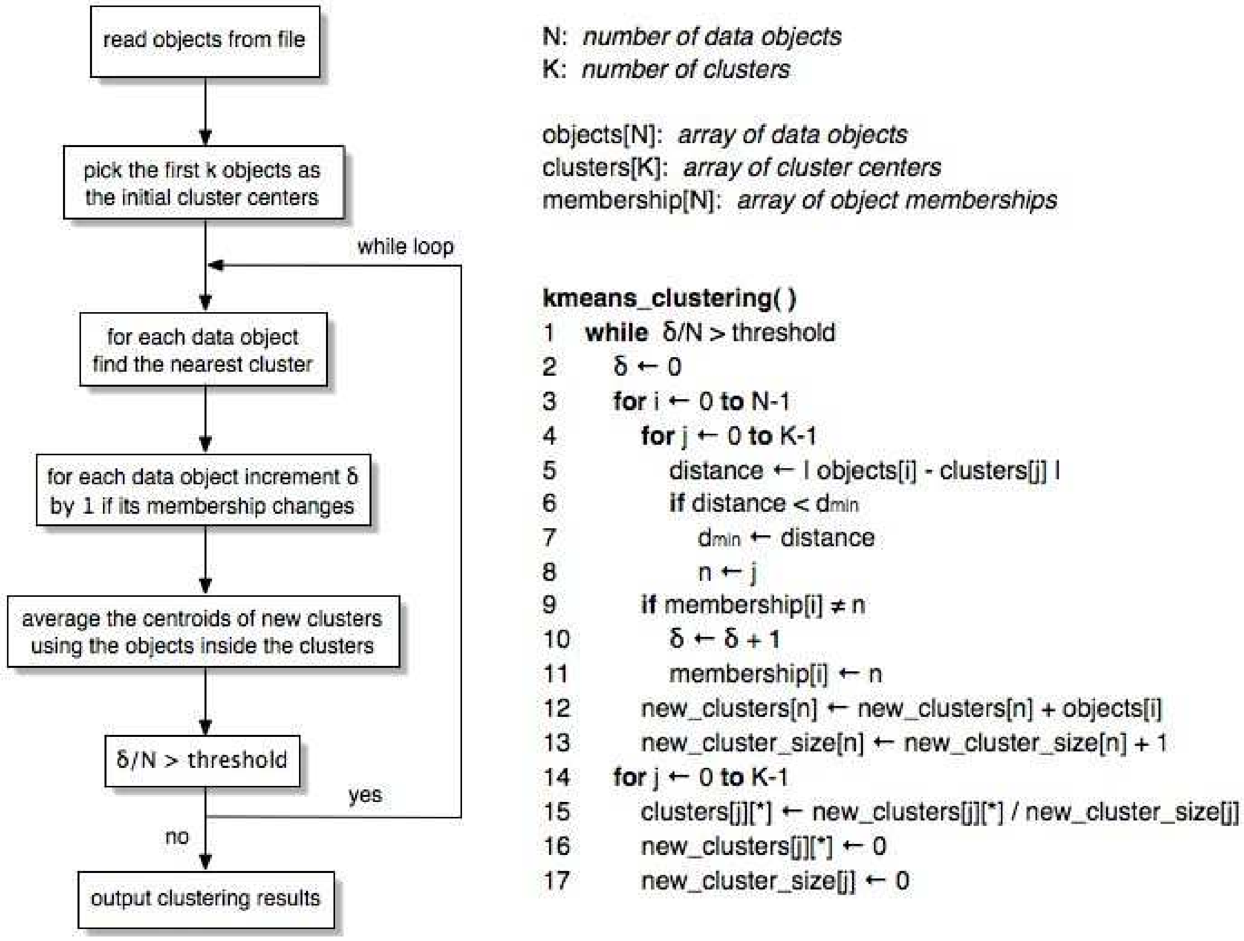
For our study we select reference points that we exclude from the raw data. Namely, since we are aiming to determine climatic characteristics of both 30^o-50^o N by 3^o-60^o E and 34^o 43^o N by 23^o 47^o E regions whose common aspect is that they include Turkey.

Exclusion has been done by using `Matlab` for both regions from the data set of every variable which means that we have elevation, 24 columns for precipitation and remaining variables 7x12 columns in each row. In addition to this 97 columns we have also 2 columns for latitude and longitude values at the beginning of our set.

Since the data set is huge with 97 columns and 31208 row for the first region defined above, we have also worked with a reduced dimension. Acquiring a reduced dimension was possible finding the principal components of our data set. The principal component analysis procedure was also handled with `Matlab` via its `statistical toolbox`. In both dimensions, initial and reduced, we have not omitted the standardization. In other words, we first utilized `autosc` function of `Matlab` over our initial data set and then `princomp` for PCA .

3.2 K-means on Climatology

In the previous chapters, we had pointed out that there are several applications of quantitative techniques for defining zones. In particular, k-means has also a wide range application examples in semioriental sciences. Hargrove and others are among the ones who utilized k-means algorithm and its parallel version in defining ecoregions. Since they have also some additional data besides the climatic data set, they were able to determine those zones [25]. On the other hand, we also employed k-means algorithm with only climate data. *Slope, soil bulk density, mineral soil depth, bedrock depth* variables does not exist within our variables


read objects from file
pick the first k objects as the initial cluster centers
while loop
for each data object find the nearest cluster
for each data object increment δ by 1 if its membership changes
average the centroids of new clusters using the objects inside the clusters
δ/N > threshold
yes
no
output clustering results

```
N: number of data objects
K: number of clusters

objects[N]: array of data objects
clusters[K]: array of cluster centers
membership[N]: array of object memberships

kmeans_clustering( )
while δ/N > threshold
    δ ← 0
    for i ← 0 to N-1
        for j ← 0 to K-1
            distance ← | objects[i] - clusters[j] |
            if distance < dmin
                dmin ← distance
                n ← j
        if membership[i] ≠ n
            δ ← δ + 1
            membership[i] ← n
        new_clusters[n] ← new_clusters[n] + objects[i]
        new_cluster_size[n] ← new_cluster_size[n] + 1
    for j ← 0 to K-1
        clusters[j][*] ← new_clusters[j][*] / new_cluster_size[j]
        new_clusters[j][*] ← 0
        new_cluster_size[j] ← 0
```

Figure 3.1: k-means algorithm [45]

whereas we have *relative humidity*, *sunshine duration*, *ground frost frequency* and *wind speed* in contrast to their study.

3.2.1 K-means procedure

The algorithm we used was written Wei-keng Liao who is a Research Assistant Professor at Electrical Engineering and Computer Science Department at Northwestern University. The procedure in the algorithm relies on the notion that was developed by J. MacQueen, in 1967.

In the general description of k-means algorithm it is stated that initially random points are selected as many as the desired number of clusters. However, if we are to determine *k* clusters, without violating randomness, our algorithm takes the first *k* data points as initial cluster centers or *centroids*.In our case, clusters can be considered as climatic zones or regions.

After the initialization where the centroids are from the first k points, each data row is examined. In fact, distance from every point to the centroids are computed. A sample is then moved to the nearest centroid's cluster. If the sample examined moves from one group to another, the control variable δ is incremented by 1. Until the stability, there will be shifts;therefore, after all the members are checked in every step, a new mean is calculated. What we have obtain is the new centroid that will be used in the calculation of the distance.

Procedure above stops when the stopping criterion is satisfied. As a stability measure, a threshold, which is .0001 in our case, is given by the user in the code. When the ratio δ/N is remains under the threshold, process ends with the final configuration of memberships.

3.2.2 Number of Clusters

Deciding on the number of clusters is another area for some researchers. Among the separation measures, validation techniques developed by Davies-Bouldin and Dunn are the best known ones. Nonetheless, there are studies which are capable of illustrating deficiencies of those indices and pursuing new ones [43]. In our experiments, besides the *threshold* which provides a stabilized clustering configuration, we also pose a criterion which was developed by Siddheswar and Rose [44].

Their method suggests another way of overcoming the obstacles for determining the number of clusters by incorporating a technique based on the *intra-cluster* and *inter-cluster* distance measures. In an image-segmentation experiment with k-means clustering algorithm, they used their method and compared the results with Davies-Bouldin and Dunn indices. In conclusion, their measure worked more consistently for the natural images than both indices. Namely, the number of clusters produced by this measure produced good segmentation results for each of the natural images as opposed to the previous measures.

In this mentioned measure compactness is being questioned. Because k-means intends to minimize the sum of squared distances from all points to their centroids, normally compact clusters should be obtained. Therefore, distances from the points to centroids are used to determine if the present configuration is compact. Naturally, distance between a point and its cluster center is computed first. Secondly, average of these distances calculated in the following fashion:

$$intra = \frac{1}{N}\sum_{i=1}^{k}\sum_{x \in C_i} \|x - z_i\|^2 \tag{3.1}$$

where N is the number of samples, k is the number of clusters, and z_i is the cluster center of cluster C_i. At first glance, one can assert that this quantity should be minimized.

Another component of this method is *inter-cluster* distance which is to be maximized as opposed to *intra-cluster* quantity. It has to be maximized because they aim to obtain clusters centers which are as far as possible away from each other. Relying on this statement, they calculate mutual distances of centroids. In order to work with the smallest one, minimum of those distances are also calculated.

$$inter = \min(\|z_i - z_j\|^2), i = 1, 2, ..., k-1 \quad j = i+1, ..., k \tag{3.2}$$

It is because, if we take the smallest one and maximize it other distances will automatically be bigger than this value.

Both complements are combined in a way that relation between them roughly explains how well the clustering is finalized. While minimizing the distances within the clusters is enabling to have compact groups, maximizing the inter-cluster gives us well separated clusters. A ratio of those two results gives us the validity index. In other words,

$$validity = \frac{intra}{inter} = \frac{\frac{1}{N}\sum_{i=1}^{k}\sum_{x \in C_i} \|x - z_i\|^2}{\min(\|z_i - z_j\|^2)} \tag{3.3}$$

Obviously, *validity* defined above is going to be minimized where we will have minimized *intra* and a maximized *inter* value. Likewise, in our experiments we preferred the same validity measure while deciding about the number of climate zones which would have an optimal configuration.

3.3 Visualization

At the end of the clustering process, we get labeled samples such that each has number varying from 1 to the desired number of clusters. At this level, it is not possible to visualize the labeled data since our matrix has missing values due to the nature of data set which only cover the land area. In order to be able to cover all grid for the mentioned latitudes and longitudes, we fill zeros for sea areas in the matrix. Finally, we have 120x342 matrix for the 30^o-50^o N by 3^o-60^o E and 54x144 matrix for 34^o-43^o N by 23^o-47^o E regions in which coordinates of sea areas have zeros.

After forming matrices consisting of numbers from zero to the number we desired, we construct contours over the maps confined by the coordinates above. For drawing those contours, we used the toolbox **NCL**. NCL first reads the input from an ASCII file which contains the matrices that we formed beforehand. Secondly, it reads the specified files containing latitudes and longitudes again from ASCII files. When I/O stage finishes, it generates a map regarding our assignments for the resources of associated variable.

4. PARALLEL K-MEANS ALGORITHM

In most cases where one has to handle a large data set or intensive computation takes place, parallel computing appears to be the first choice of the user. Due to its characteristics of dividing the job into small pieces, time consuming jobs become easy to achieve in a shorter time period. Besides reducing the size of the data set via data partitioning, computational complexity of a particular problem can also be reduced considerably [46].

In our case, we do not have a large number of points, but we have a high dimensional data space. Dealing with a 40140x109 matrix may not be always easy depending on the computation that we try to achieve. With the algorithm having an $O(n)$ complexity, parallelization process saves time.

The platform that we have run our code is Redhat 3.0 AS + SFS operating system in **ITU HPCC** lab where we utilized ***hp*** clusters. It is a distributed environment having nodes with Intel(R)Xeon(TM) 3.4 GHZ processors and 2x2 GB RAM. The communication infra-structure is myranet.

4.1 Parallel k-means Procedure

In contrast to the serial procedure, I/O is also done in parallel. We used Intel MPI library for the communication between the nodes. Every processor participates the reading process by reading the assigned partition which was previously arranged regarding the information in binary input files. After getting the data, first members of the whole data set, which are naturally in processor 0, are broadcasted to all processors as the initial centroids by using `MPI_Bcast`. In every processor, standard k-means procedure takes place and every sample is associated with a group. Meanwhile, whenever a shift occurs from one cluster to another our control variable δ is incremented by 1. Once the assignments are completed, members of each group is summed separately. During the summation process, members of the clusters are also counted [45].

It is dependent on the δ if we need one more iteration. Namely, via `MPI_Allreduce` function each processor has the sum of all δs. Likewise, every processor becomes able to take the average of summed coordinates and summed counts which helps us to calculate the new centroids. Since total number of the objects is also known, call it N in this case, $\frac{\sum_{k=1}^{nproc} \delta_k}{N}$ is compared with the *threshold*. If the comparison leads to another iteration, updated cluster centers are put into process again. Iterations go on until the threshold is reached.

4.2 Results and Conclusion

As we mentioned before, we have two data sets used in computations. In addition to low complexity of k-means algorithm, sizes of those inputs give idea about the scalability of the algorithm. When we observe the outputs with a variety of input types where number of clusters and the dimension of the sets are separately examined, it becomes apparent under which conditions our code performs better.

More precisely, two point of views make us be able to compare the facts behind the performance. One of those is computational burden which is related to the number of clusters. As the number of clusters is increased, more distances are calculated. In other words, all the samples are examined as many as the specified number. Whenever the number is low, less computation is required as opposed to greater ones. Therefore, for a particular data set, computation time may increase for large number of clusters. The second approach is to consider about the dimension. For a given number of points, if the dimension increases, time consumption also increases.

Regarding the criterion above, we pay attention how well the parallel implementation performs. To be able to gain knowledge depending on the size of the data sets, we did experiments for two cases. To see what happens if we work a larger data set for the same number of clusters, we run the code with both 31208x109 and 5465x109 matrices. Since the number of centroids that are to be examined for each member of both data sets, the only independent variable that is to be observed is the size. Outputs from two experiments show that algorithm is scalable when the large data set is taken as the input. In addition to scalability, efficiency does not drop below % 76 whereas 5465x109 matrix has an efficiency within the range from % 98 to % 39 which is drastically bad as it can be realized from the tables and the figures below.

Table 4.1: Speed Up and Efficiency

	1^{st} Region		2^{nd} Region	
# of procs	speed up	efficiency	speed up	efficiency
2	1.87	0.93	1.96	0.98
3	2.77	0.92	2.85	0.95
4	3.68	0.92	3.76	0.94
5	4.52	0.90	4.48	0.90
6	5.36	0.89	4.95	0.82
7	6.06	0.87	5.22	0.75
8	7.10	0.89	5.53	0.69
9	7.67	0.85	6.27	0.70
10	8.34	0.83	6.27	0.63
11	9.14	0.83	6.27	0.57
12	9.94	0.83	6.27	0.52
13	10.19	0.78	6.27	0.48
14	11.00	0.79	6.27	0.45
15	11.40	0.76	6.27	0.42
16	12.19	0.76	6.27	0.39

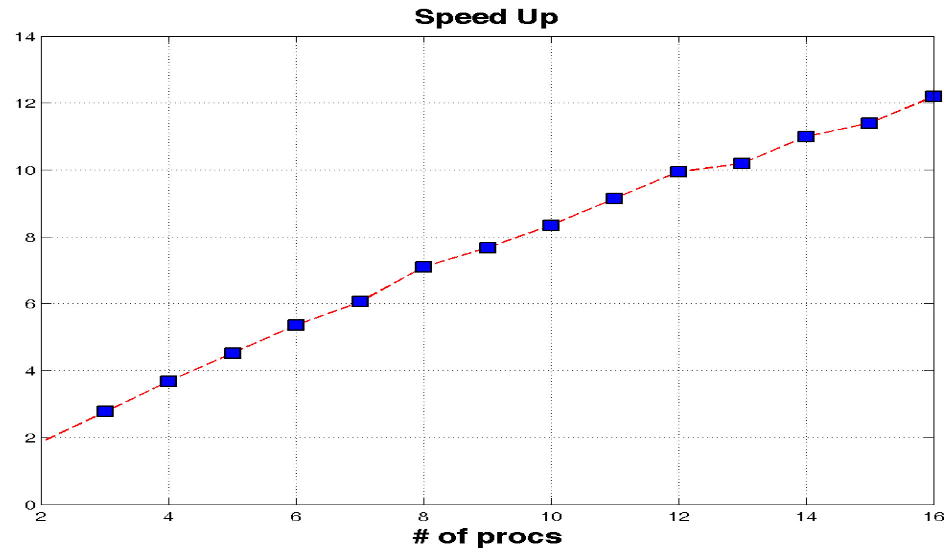


Figure 4.1: Speed Up for the 1^{st} data set with 6 clusters

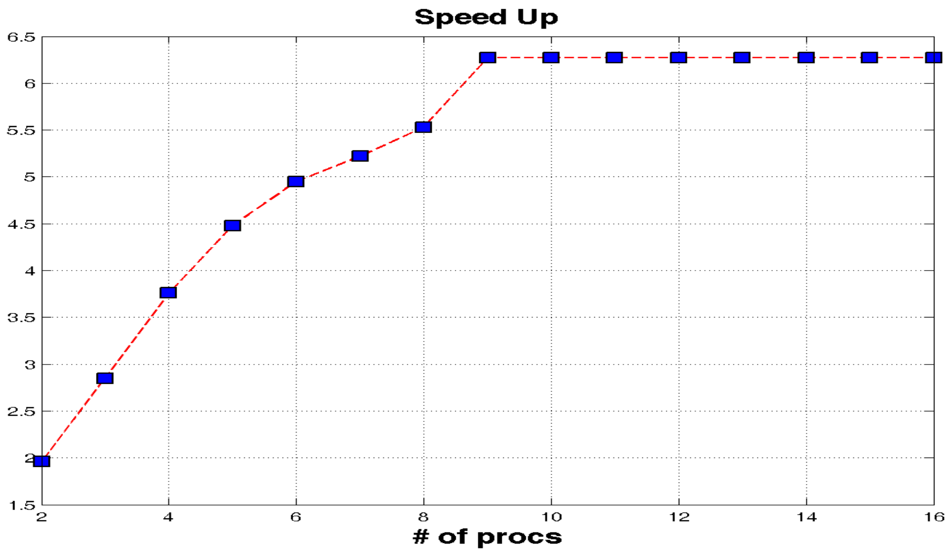


Figure 4.2: Speed Up for the 2^{nd} data set with 6 clusters

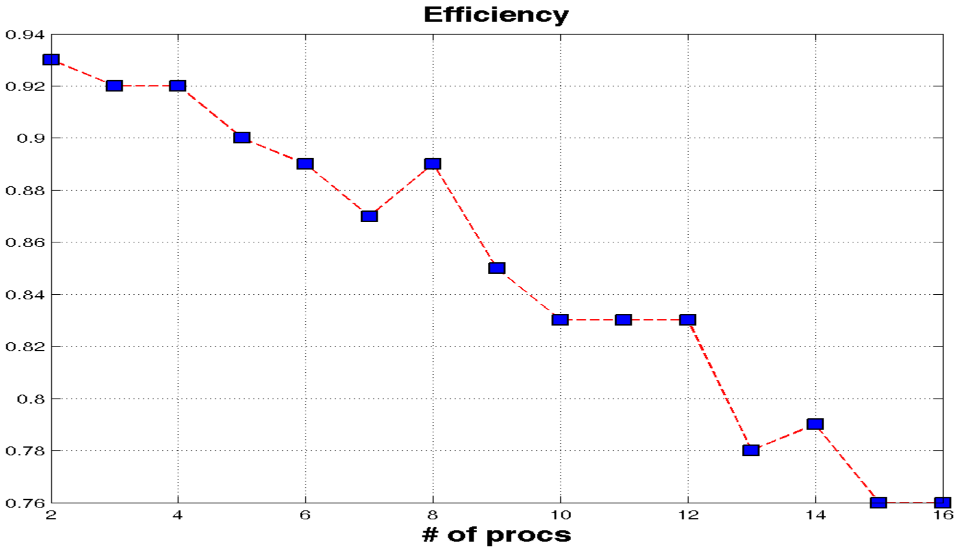


Figure 4.3: Efficiency for the 1^{st} data set with 6 clusters

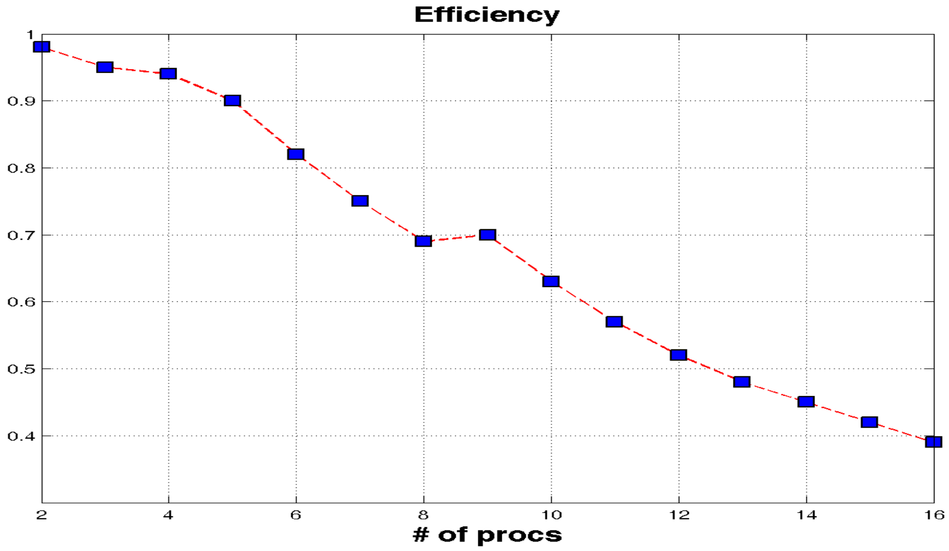


Figure 4.4: Efficiency for the 2^{nd} data set with 6 clusters

It can be stated that although we have an embarrassingly parallel algorithm with an $O(n/p)$ complexity, the time for communication dominates over the time for computation. In other words, after a point the messaging load becomes so heavy that parallel processing does not speed up the computation. As a result, we can claim that as the dimension of the data is set increased, we can have more scalable and efficient outcomes in further applications.

5. K-MEANS IMPLEMENTATION ON CLIMATOLOGY

Until this section we have given some preliminary work that we are going to utilize in this section. Our work is maintained with both data sets belonging to the regions 30^o-50^o N by 3^o-60^o E and 34^o-43^o N by 23^o-47^o E. The purpose behind these selections comes from the fact that those frames both include Turkey in narrower and broader views. Generally speaking, experiments takes into account first the whole data set which has 109 variables where combinations of cv and elevation are observed. Secondly, the reduced version of them for those different data sets via PCA are used to obtain zones. Not only the memberships, but also the intermediate steps are also colored by using `NCL` toolbox in the further applications which will be explained later.

5.1 Experiments with high dimensional data set

Through a standardization process, we have a scaled data set where we prevent a column from dominating others while doing the calculations. We do our experiments with the data set gathered from the region 30^o-50^o N by 3^o-60^o E and 34^o-43^o N by 23^o-47^o E. Since our data are confined only with land areas, we have 31208x109 data matrix instead of 41040x109 which corresponds to all the points within this frame. Likely, for the second area we have 5465x109 where actual grid consists of a 7776x109 matrix.

The algorithm originally does not deal with how to obtain intra and inter distances. We modified the code such a way that it runs a reasonable number of times. More precisely, we pose an upper bound for the number of clusters and we make the code run until the upper limit. We determined this upper limit as 20. However, when we confront with very similar validity indices we take one which indicates the greatest number of clusters.

For the region 34^o-43^o N by 23^o-47^o E, to be able to compare the current 7 geographic zones for Turkey we constrain it with 10 in the second frame.

Instead of a straightforward clustering, we have preferred combining the environmental variables. Namely, besides taking the principal components of the data set, we have also tried to observe affect of coefficient of variations for precipitation and the topographic component, elevation, on clustering.

Initially, all the variables including both cv and elevation were are tested for two regions. For the matrix 41040x109 we obtain 6 clusters whereas our algorithm suggests 4 as the optimum number of clusters for 5465x109 data set. Ratios for those sets are 0.18539 and 0.53898 which can be compared with the other validity results from the tables below:

Table 5.1: Validity Results for $30^o - 50^o$ N $3^o - 60^o$ E with CV and Elevation

# of clusters	Intra Distance	Inter Distance	Validity
3	55.304	123.655	0.44724
4	45.735	99.112	0.46145
5	41.467	76.839	0.53966
6	**37.621**	**202.930**	**0.18539**
7	35.267	99.254	0.35532
8	34.328	55.137	0.62261
9	32.779	71.268	0.45994
10	28.029	55.528	0.50477
11	26.608	56.659	0.46962
12	26.089	37.250	0.70038
13	24.701	43.530	0.56744
14	24.470	70.168	0.34874
15	24.046	67.537	0.35604
16	22.268	25.133	0.88602
17	21.406	35.247	0.60733
18	21.567	24.299	0.88758
19	19.633	31.997	0.61357
20	19.923	24.746	0.80508

Table 5.2: Validity Results for $34^o - 43^o$ N $23^o - 47^o$ E with CV and Elevation

# of clusters	Intra Distance	Inter Distance	Validity
3	63.675	72.557	0.87758
4	**54.116**	**100.405**	**0.53898**
5	48.430	73.841	0.65587
6	45.951	36.083	1.27350
7	41.875	37.285	1.12310
8	40.435	41.576	0.97254
9	36.179	44.525	0.81254
10	34.413	43.903	0.78385

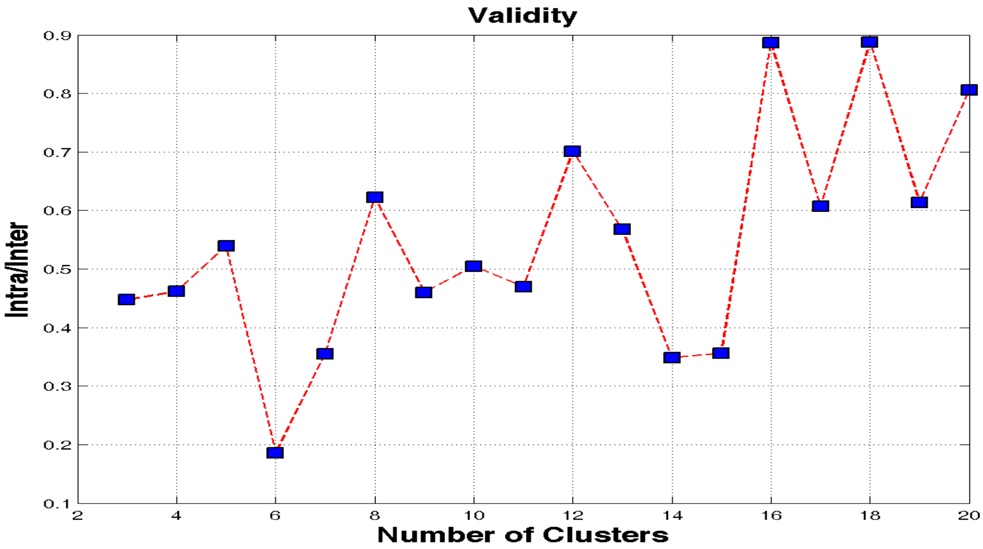


Figure 5.1: validity for every number of cluster

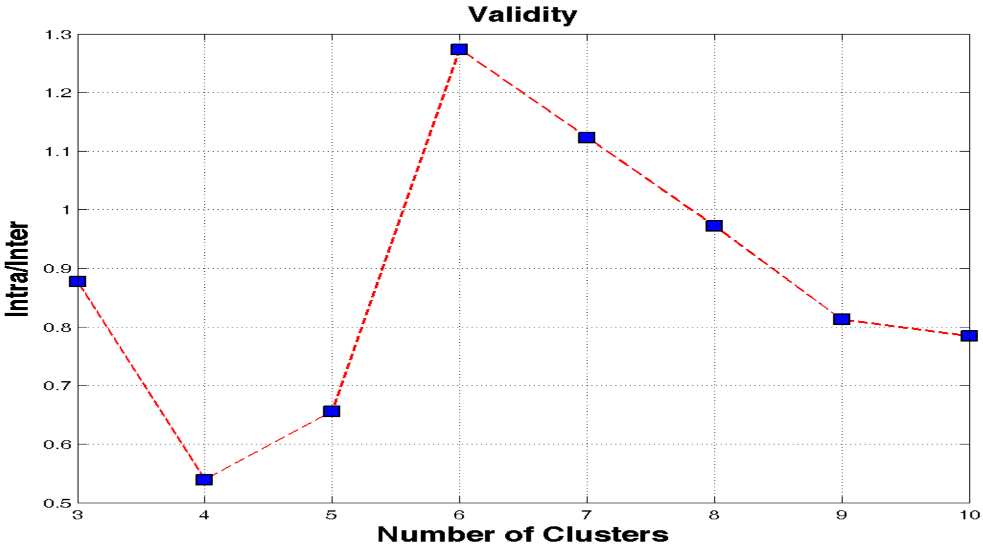


Figure 5.2: validity for every number of cluster

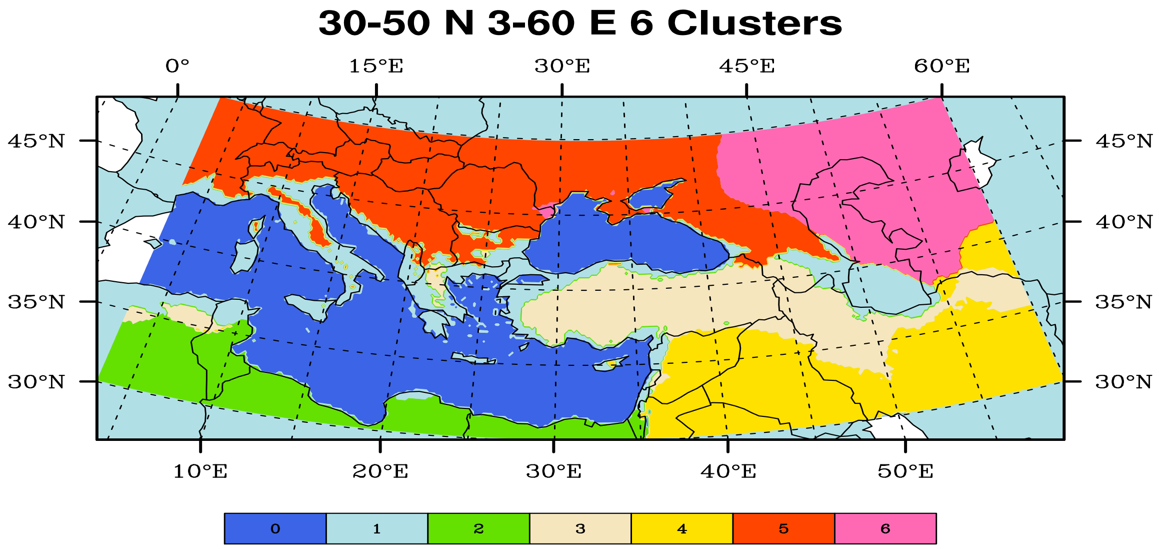


Figure 5.3: $30^o - 50^o$ N $3^o - 60^o$ E with cv and elevation

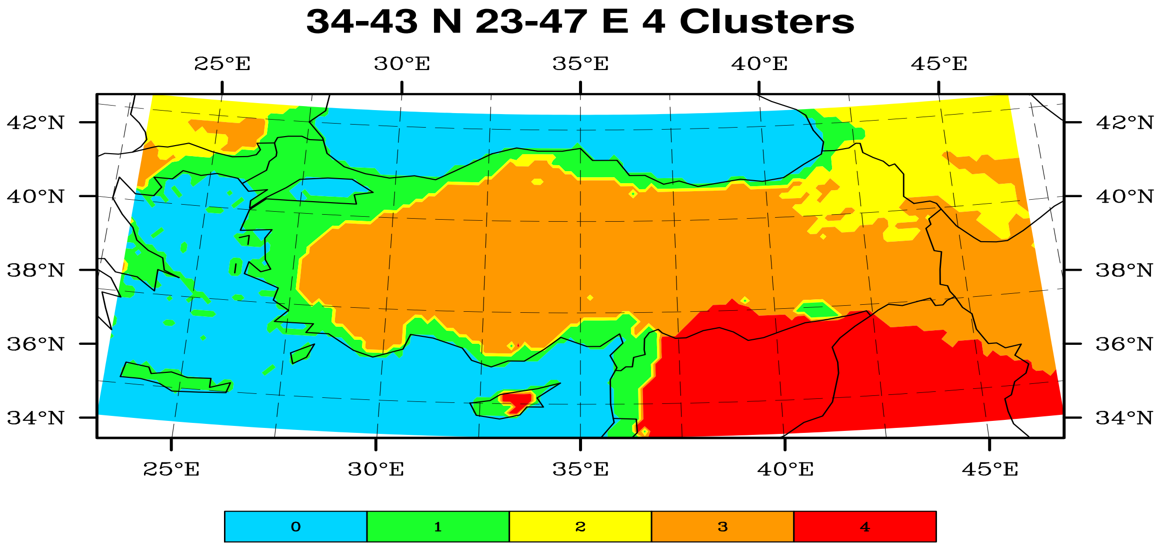


Figure 5.4: $34^o - 43^o$ N $23^o - 47^o$ E with cv and elevation

If elevation is excluded, then we have a new valid number of clusters. In this case, we have 108 columns for each data set and 10 zones with 0.11552 whereas 4 again with 0.54367 respectively. It can be easily seen from the graphs that at 10 and 4 we obtain the minimum values. Thus, it gives us the maps accordingly.

Table 5.3: Validity Results for $30^o - 50^o$ N $3^o - 60^o$ E with CV without Elevation

# of clusters	Intra Distance	Inter Distance	Validity
3	54.316	123.706	0.43908
4	44.987	98.984	0.45449
5	40.689	75.713	0.53742
6	37.103	70.108	0.52923
7	34.650	98.844	0.35055
8	33.713	55.587	0.60648
9	29.677	73.249	0.40515
10	**28.218**	**244.263**	**0.11552**
11	26.173	53.949	0.48515
12	24.820	34.051	0.72889
13	24.314	58.960	0.41239
14	24.088	41.447	0.58116
15	22.615	44.851	0.50422
16	21.773	25.124	0.86662
17	21.154	32.249	0.65596
18	20.177	26.218	0.76957
19	20.001	37.248	0.53697
20	19.682	33.431	0.58873

Table 5.4: Validity Results for $34^o - 43^o$ N $23^o - 47^o$ E with CV without Elevation

# of clusters	Intra Distance	Inter Distance	Validity
3	63.128	70.440	0.89620
4	**53.614**	**98.616**	**0.54367**
5	48.091	71.254	0.67493
6	45.631	35.649	1.28000
7	41.568	37.306	1.11426
8	39.625	51.109	0.77531
9	35.896	44.102	0.81393
10	34.135	43.684	0.78141

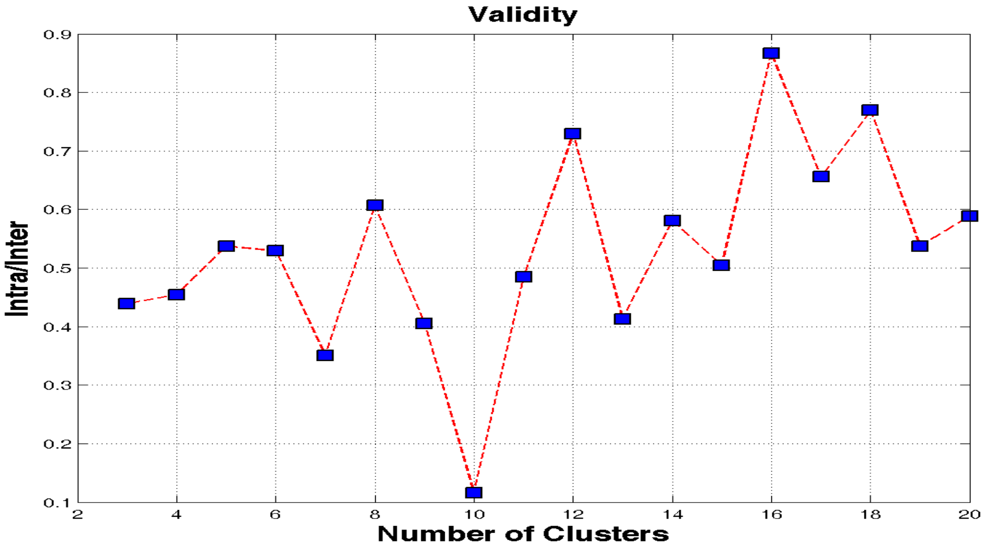


Figure 5.5: validity for every number of cluster

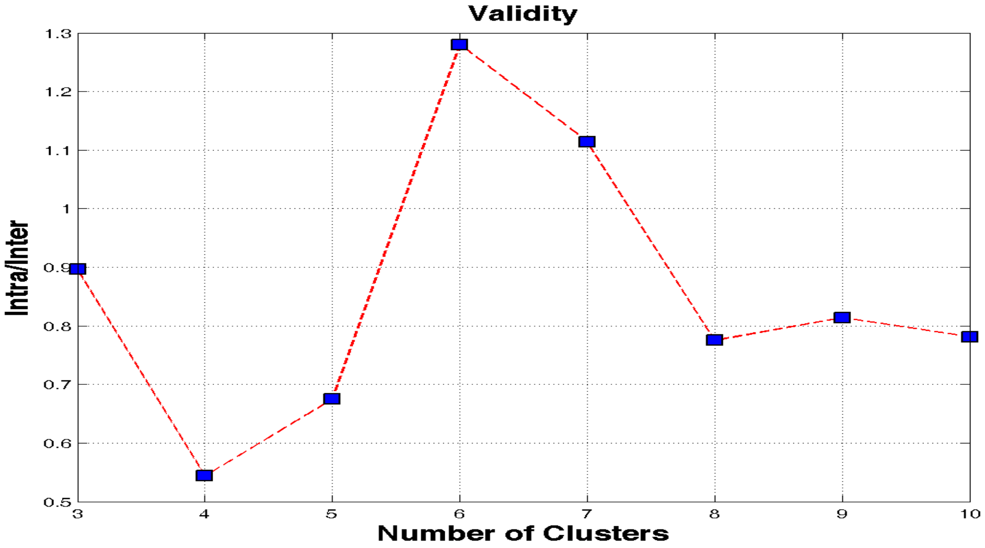


Figure 5.6: validity for every number of cluster

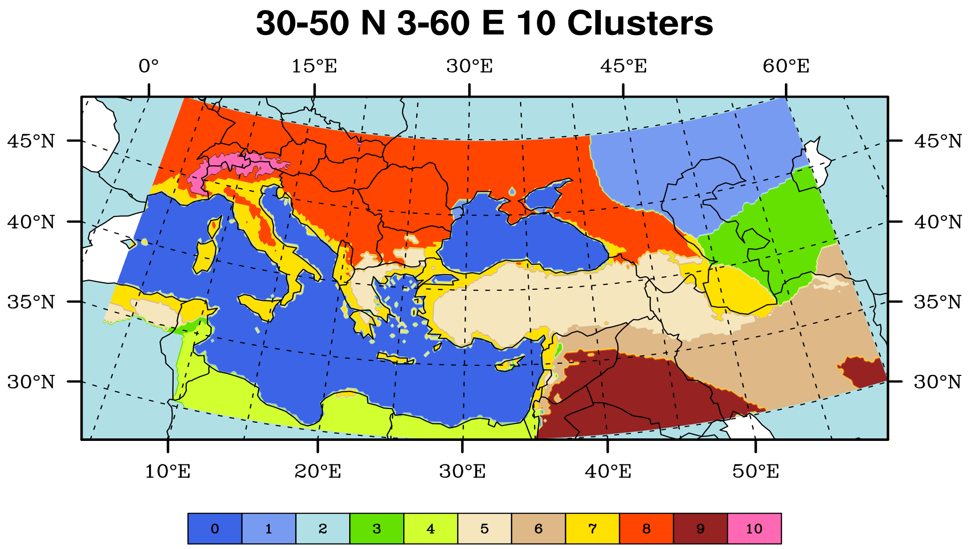


Figure 5.7: $30^o - 50^o$ N $3^o - 60^o$ E with cv without elevation .

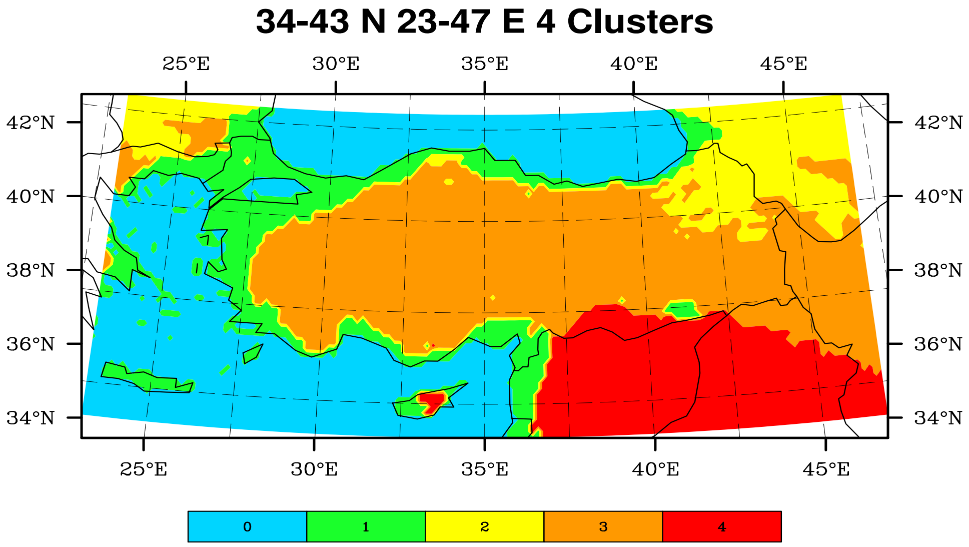


Figure 5.8: $34^o - 43^o$ N $23^o - 47^o$ E with cv without elevation .

Now we consider another case in which both cv and elevation are excluded from the data sets. However, what we confront here is that there are competing numbers of clusters for the broader region. After the computation, which can also be noticed from the graphs, 4 and 14 seem to compete with the values 0.43134 and 0.44215 respectively. Although 19 seems the strongest candidate, we reject it since we do not prefer such a high number of zones. On the other hand, for the second region we have 4 zones with validity 0.48718.

Table 5.5: Validity Results for $30^o - 50^o$ N $3^o - 60^o$ E without CV and Elevation

# of clusters	Intra Distance	Inter Distance	Validity
3	49.680	98.466	0.50455
4	**40.555**	**94.021**	**0.43134**
5	36.581	68.616	0.53313
6	33.793	66.850	0.50551
7	32.336	58.465	0.55308
8	30.924	67.349	0.45916
9	27.095	40.074	0.67612
10	25.804	43.730	0.59006
11	23.800	33.680	0.70665
12	22.533	34.133	0.66016
13	21.657	32.825	0.65978
14	**20.505**	**46.375**	**0.44215**
15	20.026	39.915	0.50171
16	19.248	35.755	0.53832
17	18.330	33.349	0.54965
18	17.582	35.462	0.49578
19	17.404	41.417	0.42022
20	17.287	16.497	1.04789

Table 5.6: Validity Results for $34^o - 43^o$ N $23^o - 47^o$ E without CV and Elevation

# of clusters	Intra Distance	Inter Distance	Validity
3	57.706	68.896	0.83758
4	**48.113**	**98.759**	**0.48718**
5	43.173	70.802	0.60978
6	39.219	51.145	0.76683
7	36.287	52.183	0.69539
8	34.214	42.046	0.81372
9	31.931	41.649	0.76668
10	30.383	40.941	0.74213

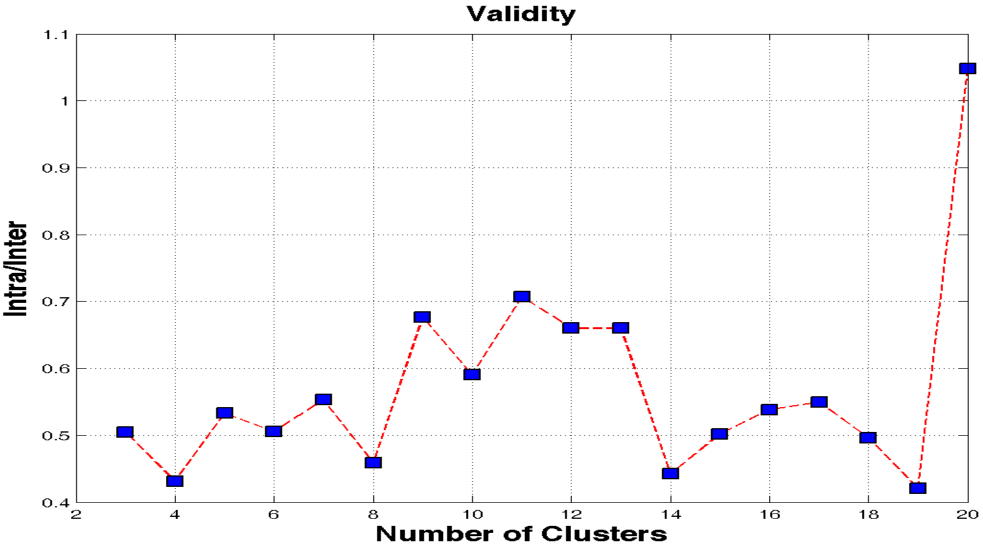


Figure 5.9: validity for every number of cluster

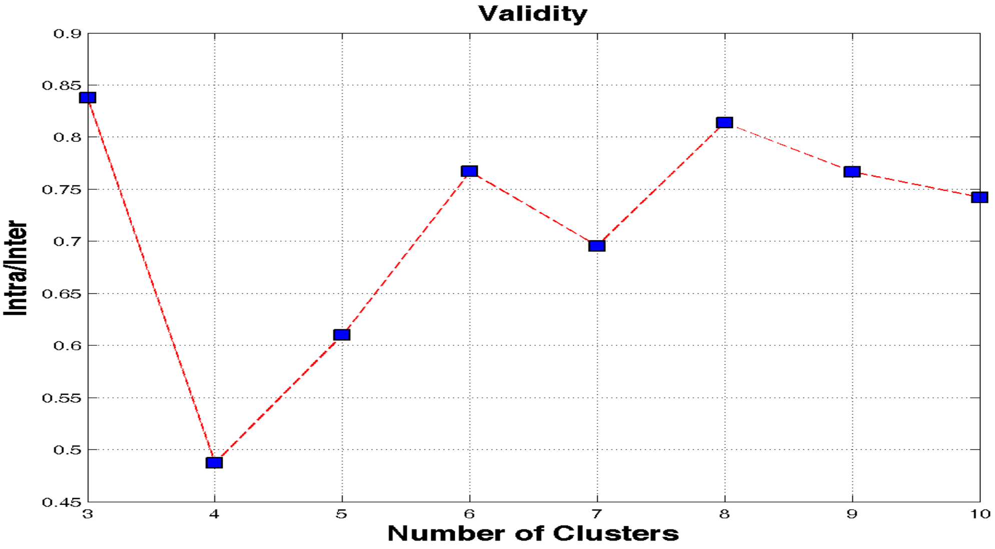


Figure 5.10: validity for every number of cluster

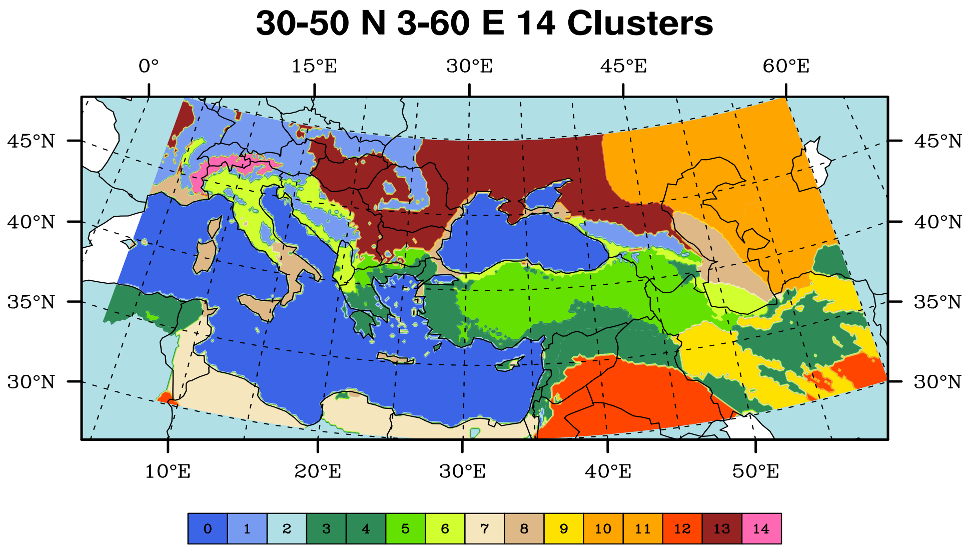


Figure 5.11: $30^o - 50^o$ N $3^o - 60^o$ E without cv and elevation .

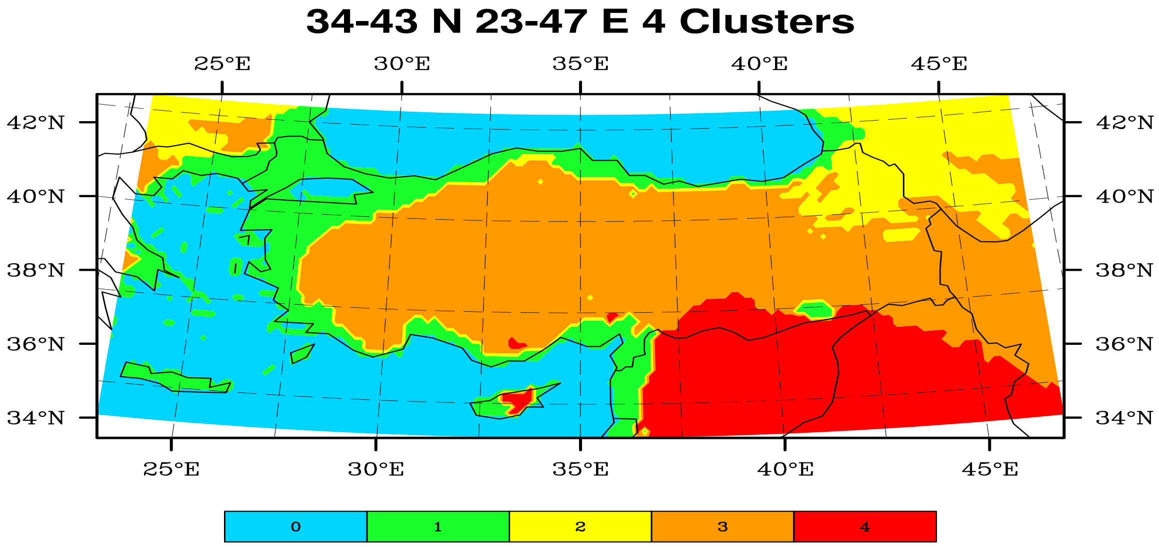


Figure 5.12: $34^o - 43^o$ N $23^o - 47^o$ E without cv and elevation .

Another combination that we have dealt with is the case where we take into account the elevation, but not cv. As a result, 97 columns are used to define a valid separation. 15 zones with 0.50876 and 4 zones with 0.48743 values came up with same computations.

Table 5.7: Validity Results for $30^o - 50^o$ N $3^o - 60^o$ E without CV with Elevation

# of clusters	Intra Distance	Inter Distance	Validity
3	50.670	98.431	0.51478
4	41.292	93.875	0.43987
5	37.150	96.810	0.38374
6	34.343	45.134	0.76091
7	32.875	48.786	0.67387
8	31.361	70.635	0.44399
9	30.346	44.984	0.67459
10	29.112	45.943	0.63365
11	24.226	33.710	0.71866
12	23.231	52.846	0.43959
13	22.571	39.746	0.56789
14	20.702	34.624	0.59792
15	**19.941**	**120.530**	**0.16544**
16	19.794	38.906	0.50876
17	18.657	19.784	0.94304
18	18.636	38.727	0.48123
19	17.571	39.002	0.45052
20	16.741	91.716	0.18253

Table 5.8: Validity Results for $34^o - 43^o$ N $23^o - 47^o$ E without CV with Elevation

# of clusters	Intra Distance	Inter Distance	Validity
3	58.223	70.622	0.82444
4	**48.590**	**99.687**	**0.48743**
5	43.518	72.930	0.59671
6	39.544	52.136	0.75847
7	36.618	53.156	0.68887
8	34.479	42.503	0.81121
9	32.204	42.199	0.76315
10	30.625	41.560	0.73688

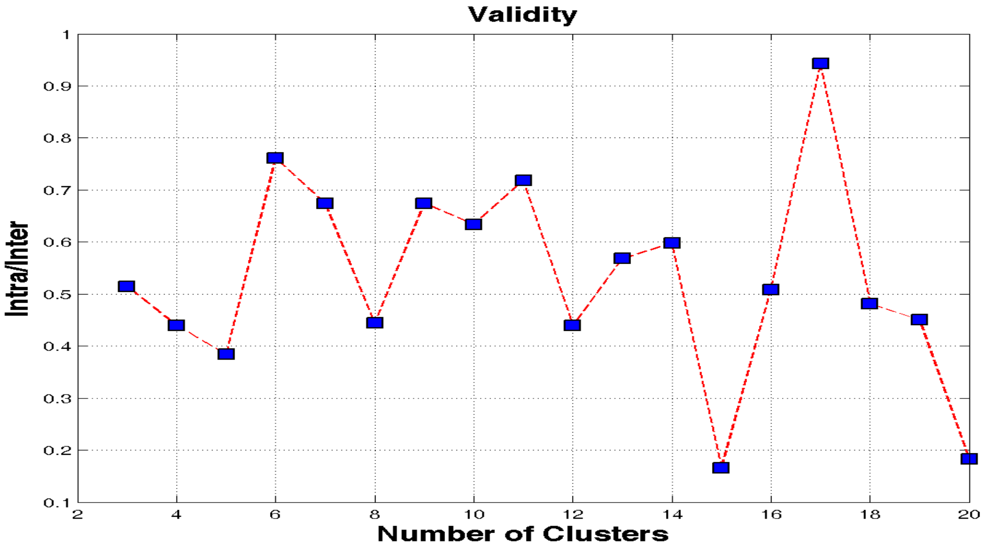


Figure 5.13: validity for every number of cluster

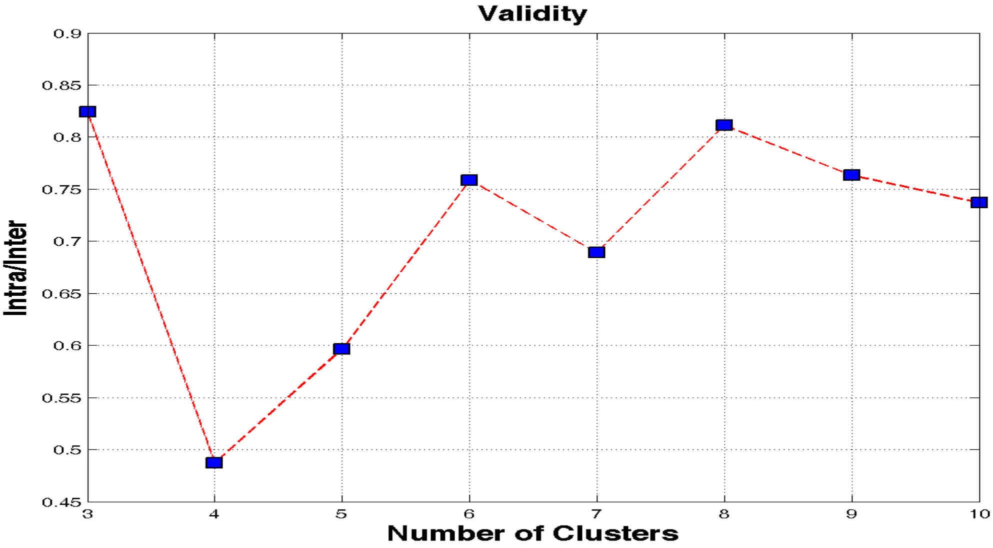


Figure 5.14: validity for every number of cluster

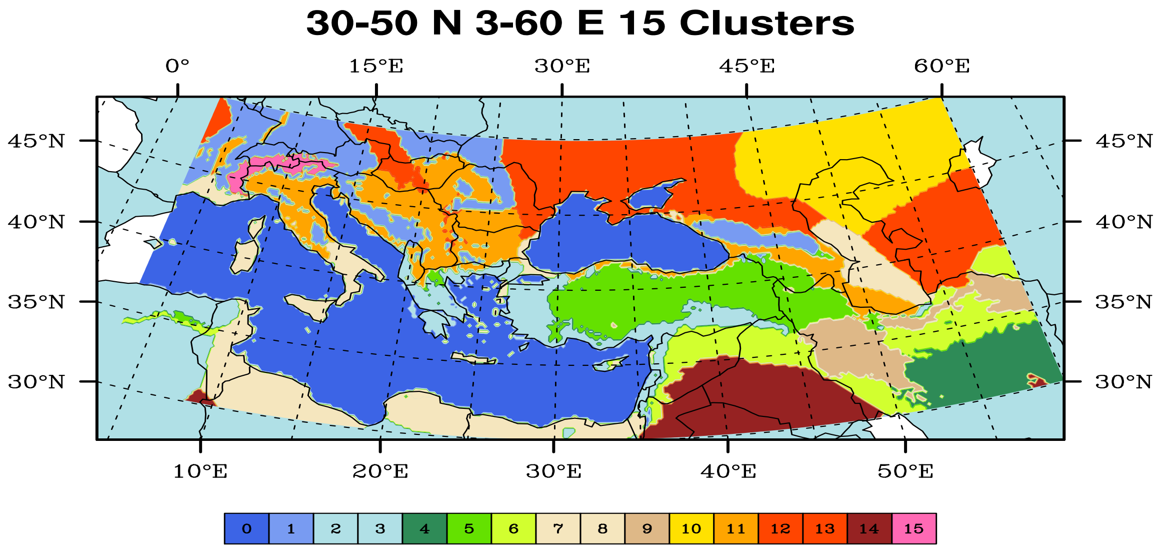


Figure 5.15: $30^o - 50^o$ N $3^o - 60^o$ E without cv with elevation

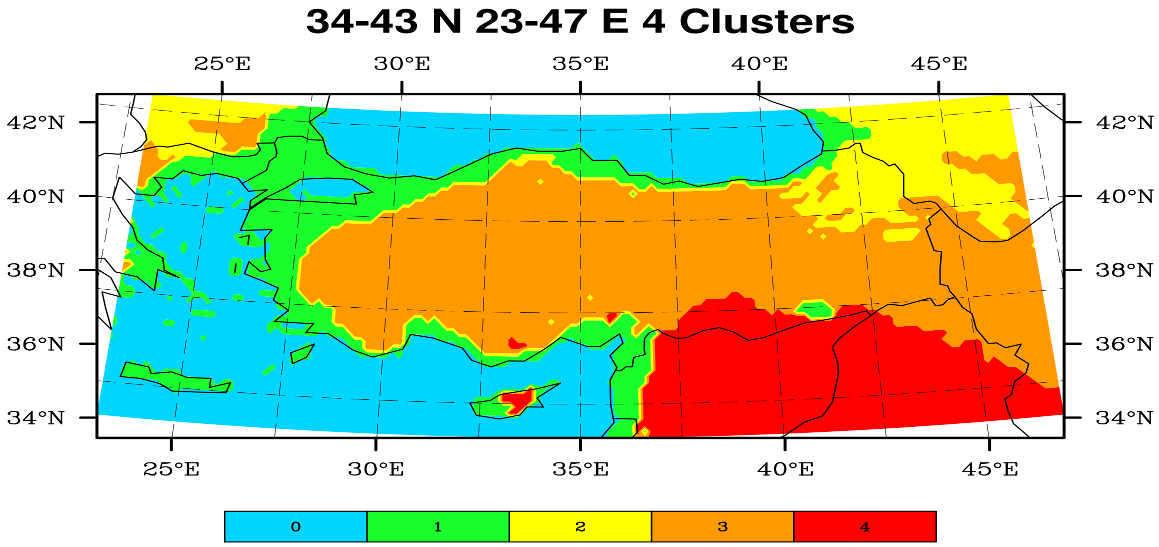


Figure 5.16: $34^o - 43^o$ N $23^o - 47^o$ E without cv with elevation .

5.2 Utilizing PCA

In the preprocessing stage of data, we use `princomp` function of Matlab. Observing the output of this function we notice that first eigenvalue is extremely dominant. Following ones are also dominant on the remaining, but in PCA 1 is considered as the cut-off eigenvalue and dominance of below 1 is omitted in this case.

Like in the previous part, we will also think about different variations depending on cv and elevation. As we modify the data set, principal components will also vary accordingly.

When we attempt to do principal component analysis with the whole data set, we acquire 109 eigenvalues. PCA puts those values in order. We present the greater ones in the table and bar plot for all.

Table 5.9: Eigenvalues for all data set

#	1^{st} Region	2^{nd} Region
1	58.08	48.54
2	15.64	18.82
3	11.49	10.64
4	6.65	7.31
5	4.46	4.89
6	2.49	4.25
7	1.81	2.63
8	1.36	2.33
9	**1.06**	1.37
10	0.93	1.21
11	0.69	**1.14**
12	0.56	0.99

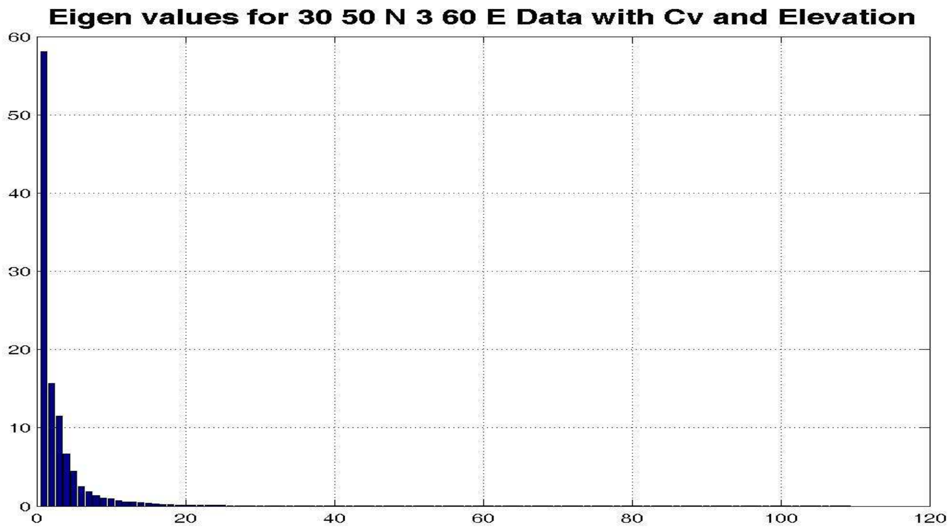


Figure 5.17: Eigenvalues from greatest to smallest

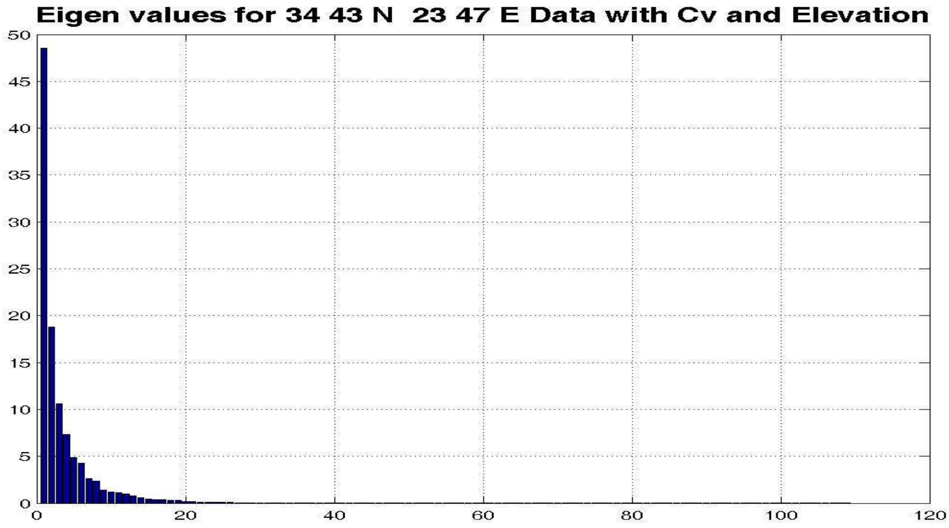


Figure 5.18: Eigenvalues from greatest to smallest

Table 5.10: Validity Results for 9 PC from $30^o - 50^o$ N $3^o - 60^o$ E with CV and Elevation

# of clusters	Intra Distance	Inter Distance	Validity
3	49.401	123.444	0.40019
4	39.820	98.585	0.40392
5	35.588	230.643	0.15430
6	31.807	123.524	0.25750
7	31.003	72.830	0.42569
8	28.696	127.985	0.22421
9	**24.627**	**260.328**	**0.09460**
10	22.571	37.261	0.60577
11	21.194	102.004	0.20777
12	20.808	52.991	0.39267
13	19.369	32.518	0.59562
14	19.130	32.553	0.58765
15	18.379	58.071	0.31650
16	18.151	43.171	0.42045
17	16.993	23.086	0.73607
18	15.922	35.304	0.45098
19	15.289	28.708	0.53259
20	15.050	27.354	0.55020

Table 5.11: Validity Results for 11 PC from $34^o - 43^o$ N $23^o - 47^o$ E with CV and Elevation

# of clusters	Intra Distance	Inter Distance	Validity
3	57.834	72.420	0.79859
4	**48.319**	**99.623**	**0.48502**
5	42.748	73.599	0.58082
6	40.366	34.203	1.18018
7	36.305	35.374	1.02632
8	33.890	33.827	1.00186
9	31.980	49.673	0.64382
10	30.298	46.586	0.65037

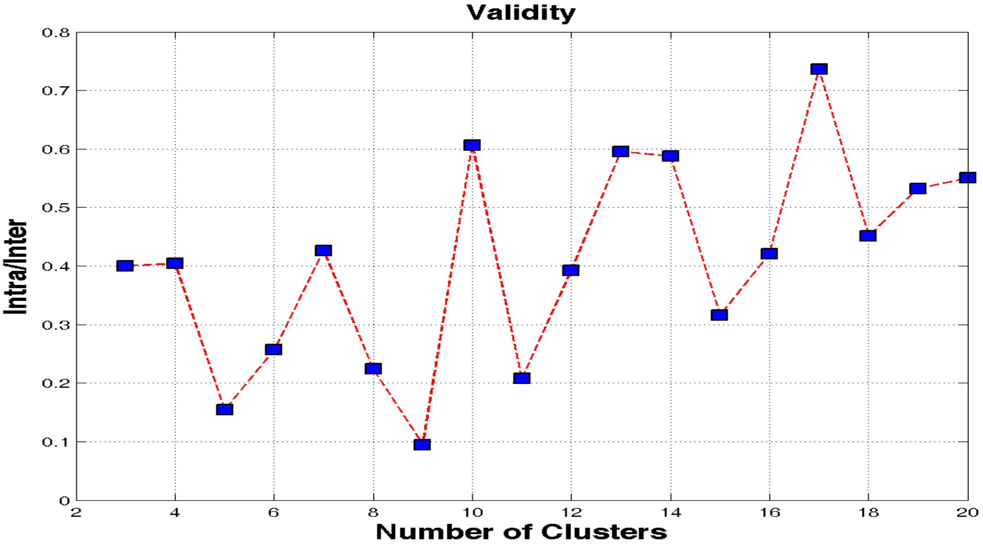


Figure 5.19: validity for every number of cluster

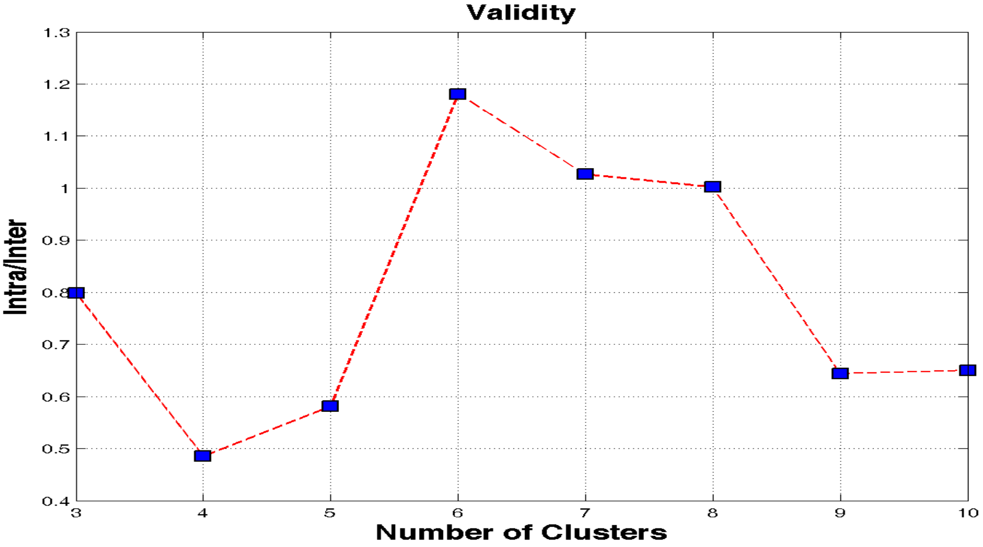


Figure 5.20: validity for every number of cluster

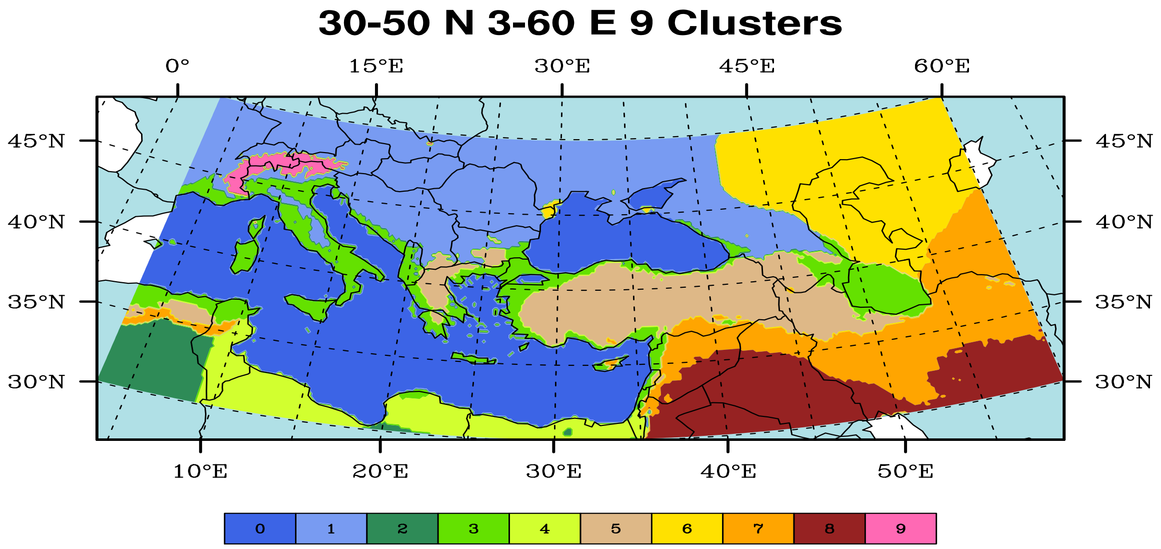


Figure 5.21: 9 PCs from $30^o - 50^o$ N $3^o - 60^o$ E with cv and elevation

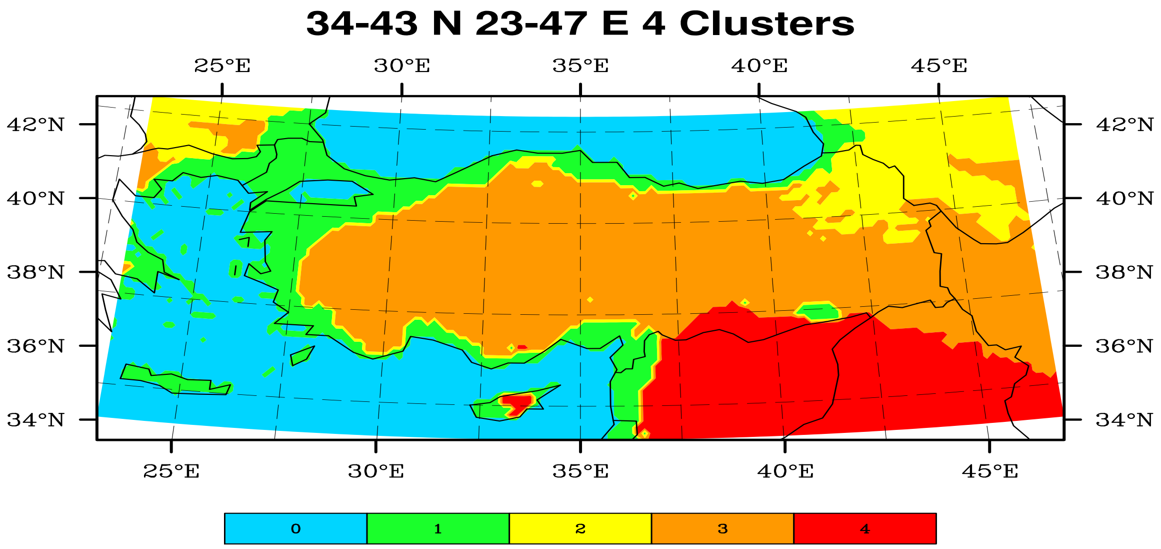


Figure 5.22: 11 PCs from $34^o - 43^o$ N $23^o - 47^o$ E with cv and elevation

If we omit elevation again and do the PCA, 9 components for the first region and 11 components for the second region are able to represent the nature of the data sets that they belong to. Nonetheless, deciding for the valid number of clusters is a problem again since the values do not differ much. To clarify, it can be seen that 5, 8, 13 competes with the values 0.10609, 0.11040 and 0.10631 respectively. On the other hand, for the second region validity measure determines the number of clusters as 4 again.

Table 5.12: Eigenvalues for all data set except elevation

#	1^{st} Region	2^{nd} Region
1	58.08	48.25
2	15.43	18.40
3	11.15	10.64
4	6.43	7.10
5	4.45	4.89
6	2.47	4.21
7	1.81	2.62
8	1.23	2.33
9	**1.05**	1.36
10	0.93	1.21
11	0.69	**1.13**
12	0.55	0.99

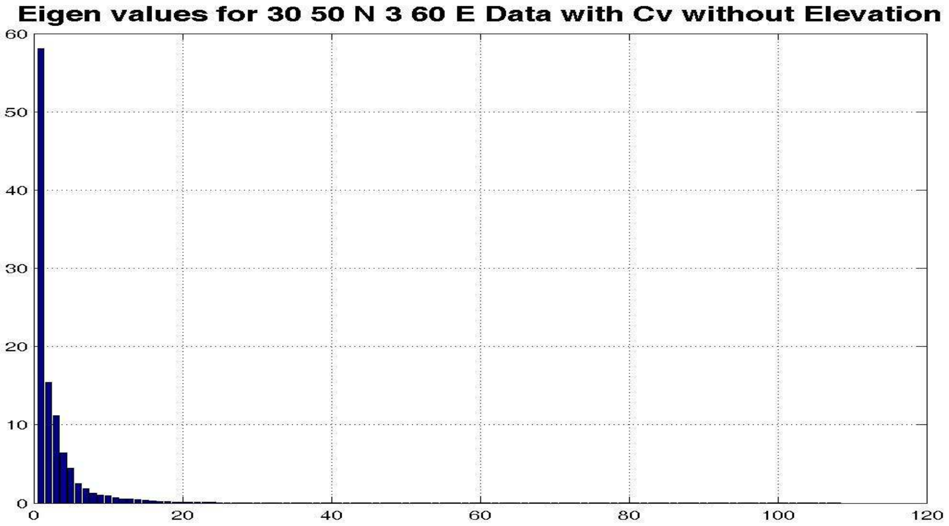


Figure 5.23: Eigenvalues from greatest to smallest

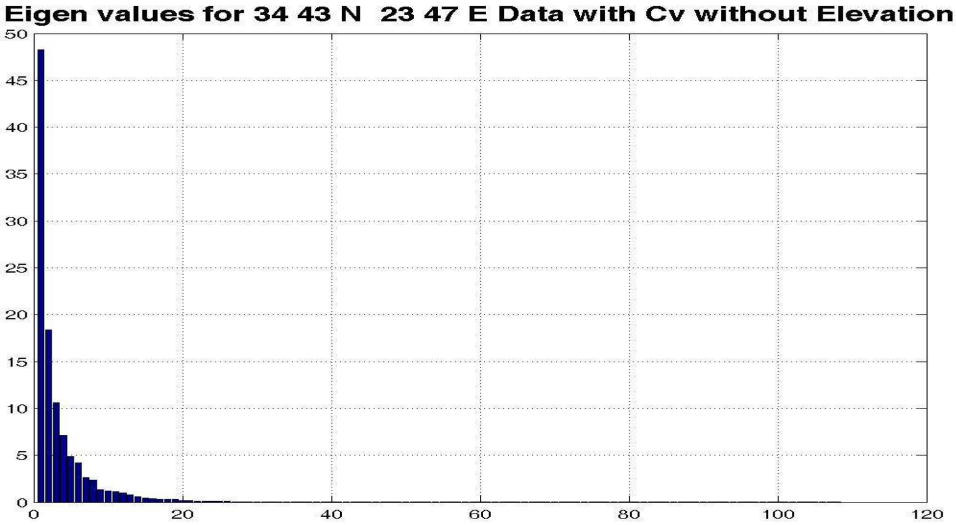


Figure 5.24: Eigenvalues from greatest to smallest

Table 5.13: Validity Results for 9 PC from $30^o - 50^o$ N $3^o - 60^o$ E with CV without Elevation

# of clusters	Intra Distance	Inter Distance	Validity
3	48.474	123.497	0.39251
4	39.136	98.222	0.39844
5	**34.878**	**328.767**	**0.10609**
6	31.309	121.894	0.25685
7	30.500	72.679	0.41966
8	**28.153**	**255.015**	**0.11040**
9	26.415	114.195	0.23131
10	22.811	70.303	0.32446
11	20.809	99.578	0.20897
12	19.523	60.467	0.32286
13	**19.058**	**179.265**	**0.10631**
14	18.267	33.674	0.54247
15	17.580	31.800	0.55283
16	17.441	23.360	0.74664
17	16.619	23.223	0.71560
18	16.302	24.650	0.66132
19	15.288	31.024	0.49277
20	14.127	54.253	0.26040

Table 5.14: Validity Results for 11 PC from $34^o - 43^o$ N $23^o - 47^o$ E with CV without Elevation

# of clusters	Intra Distance	Inter Distance	Validity
3	57.305	70.565	0.81208
4	**47.831**	**98.418**	**0.48600**
5	42.421	71.075	0.59685
6	40.059	34.374	1.16541
7	36.011	35.517	1.01391
8	33.578	33.654	0.99776
9	31.742	48.864	0.64960
10	30.050	46.554	0.64550

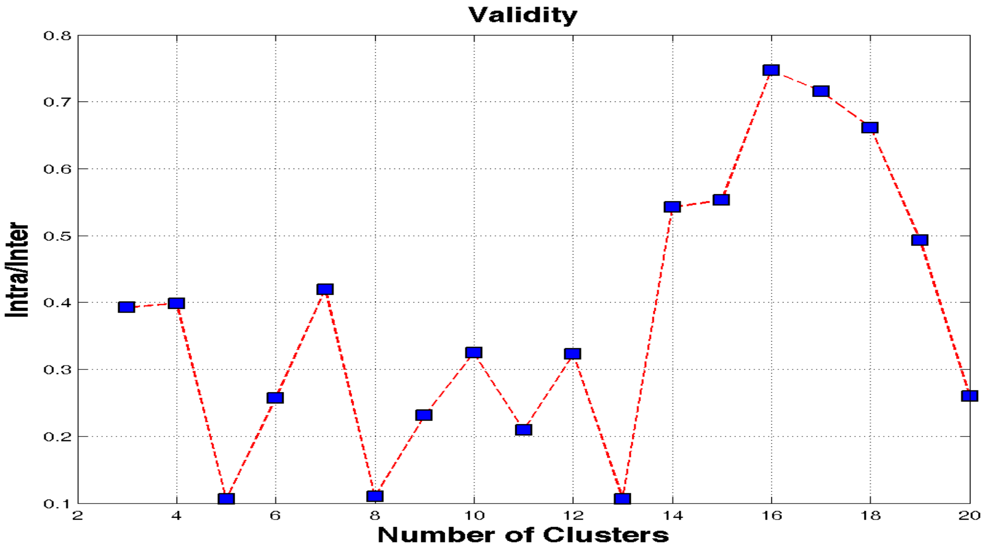


Figure 5.25: validity for every number of cluster

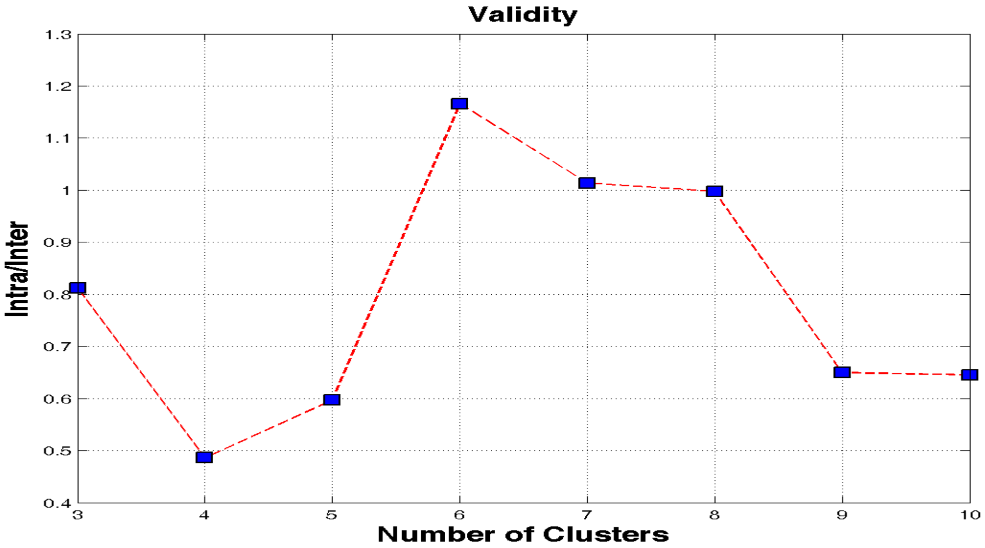


Figure 5.26: validity for every number of cluster

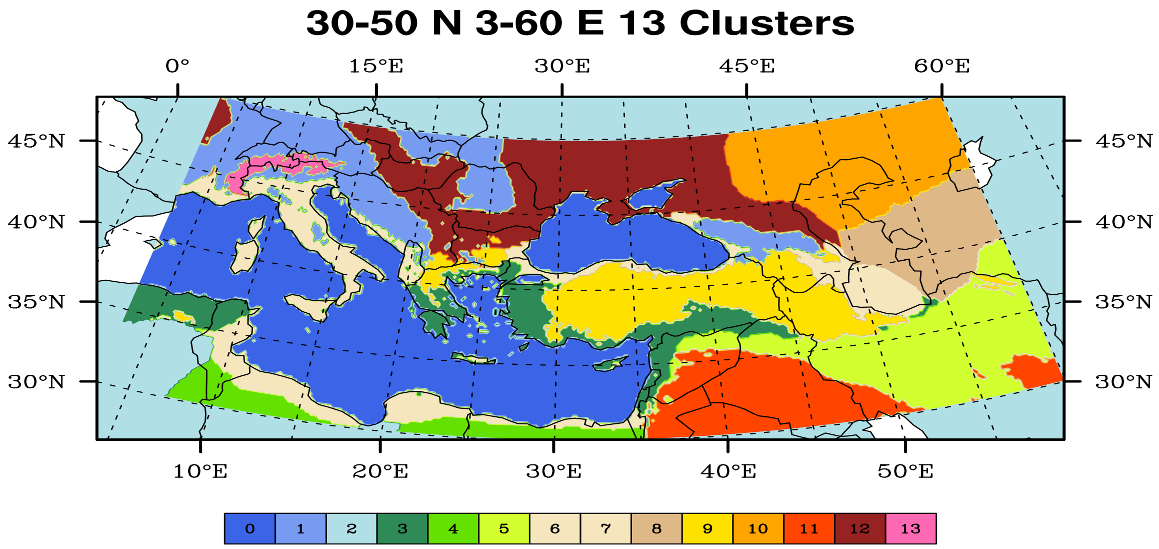


Figure 5.27: 9 PCs from $30^o - 50^o$ N $3^o - 60^o$ E with cv without elevation

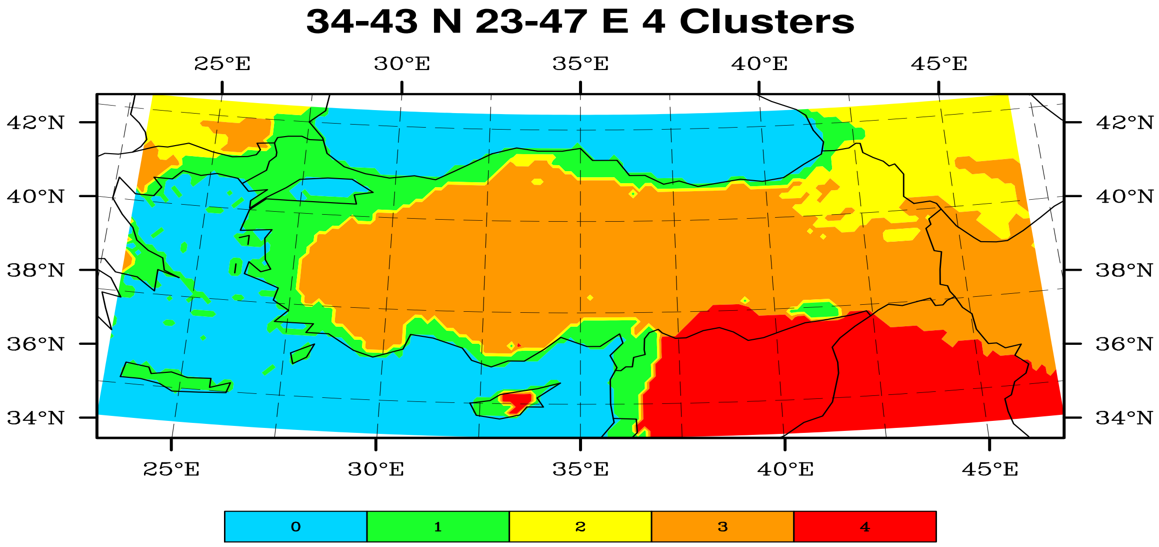


Figure 5.28: 11 PCs from $34^o - 43^o$ N $23^o - 47^o$ E with cv without elevation

Once we exclude both variables, 97 columns remains to be analyzed. After the analysis, first 9 and 10 components represents main information about the data sets. Therefore, validity ratios and the zones are determined accordingly. Namely, 0.12017 leads to 13 clusters whereas 0.44045 points 4.

Table 5.15: Eigenvalues for all data set without cv and elevation

#	1^{st} Region	2^{nd} Region
1	50.68	41.32
2	15.13	18.09
3	10.57	9.37
4	5.88	6.91
5	3.65	4.78
6	2.18	3.94
7	1.50	2.56
8	1.16	2.06
9	**1.03**	1.18
10	0.64	**1.00**
11	0.59	0.91
12	0.49	0.89

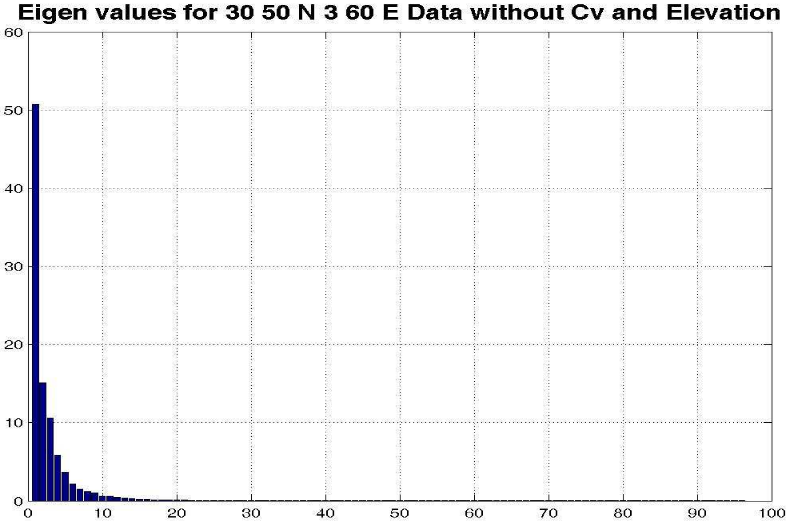


Figure 5.29: Eigenvalues from greatest to smallest

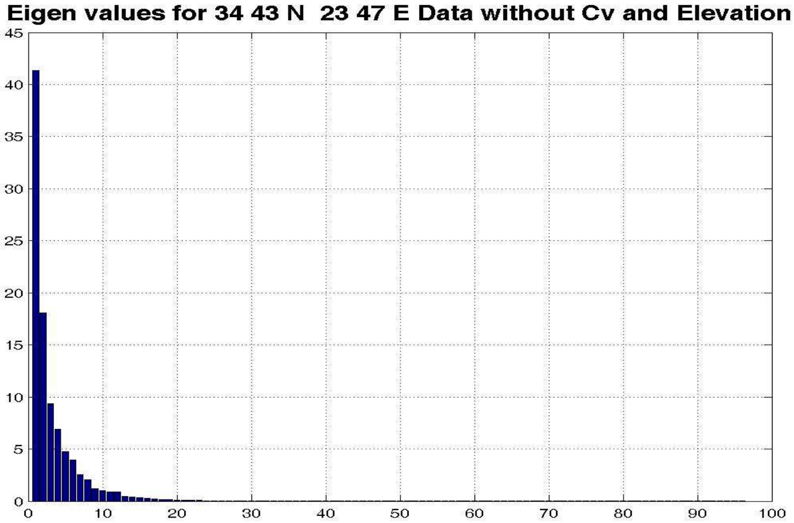


Figure 5.30: Eigenvalues from greatest to smallest

Table 5.16: Validity Results for 9 PC $30^o - 50^o$ N $3^o - 60^o$ E without CV and Elevation

# of clusters	Intra Distance	Inter Distance	Validity
3	45.498	98.535	0.46174
4	36.370	93.696	0.38817
5	32.443	67.486	0.48073
6	29.682	44.046	0.67389
7	27.911	56.919	0.49036
8	26.187	94.837	0.27612
9	25.815	43.021	0.60004
10	21.086	40.806	0.51674
11	20.304	42.683	0.47569
12	19.103	34.772	0.54938
13	**17.882**	**148.811**	**0.12017**
14	16.738	33.905	0.49368
15	16.398	41.661	0.39360
16	15.859	18.880	0.83997
17	14.968	18.668	0.80178
18	14.283	31.713	0.45039
19	15.015	15.986	0.93927
20	13.957	16.150	0.86421

Table 5.17: Validity Results for 10 PC $34^o - 43^o$ N $23^o - 47^o$ E without CV and Elevation

# of clusters	Intra Distance	Inter Distance	Validity
3	52.960	68.757	0.77026
4	**43.402**	**98.538**	**0.44045**
5	38.512	70.184	0.54873
6	34.623	51.126	0.67721
7	32.614	24.519	1.33016
8	29.802	41.400	0.71984
9	27.609	40.380	0.68374
10	26.266	39.701	0.66160

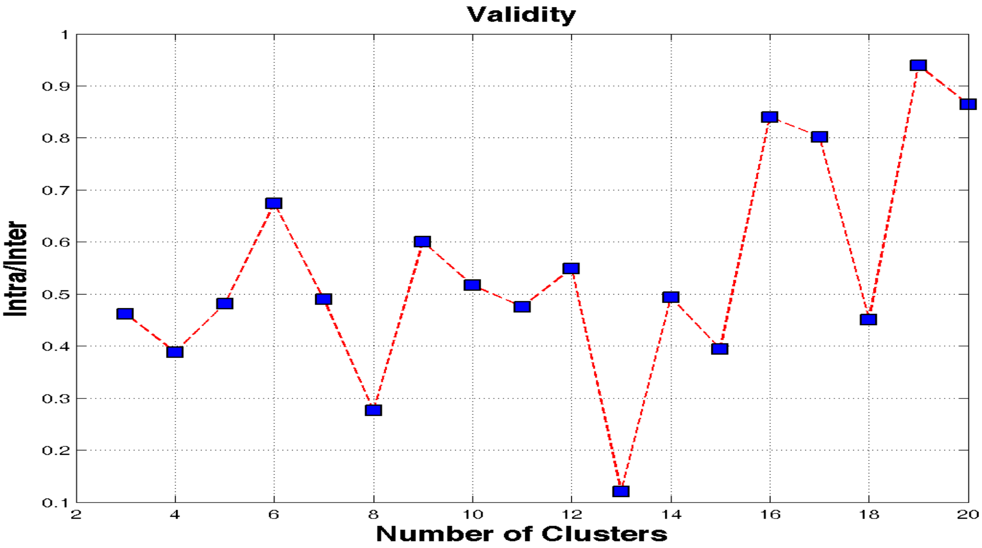


Figure 5.31: validity for every number of cluster

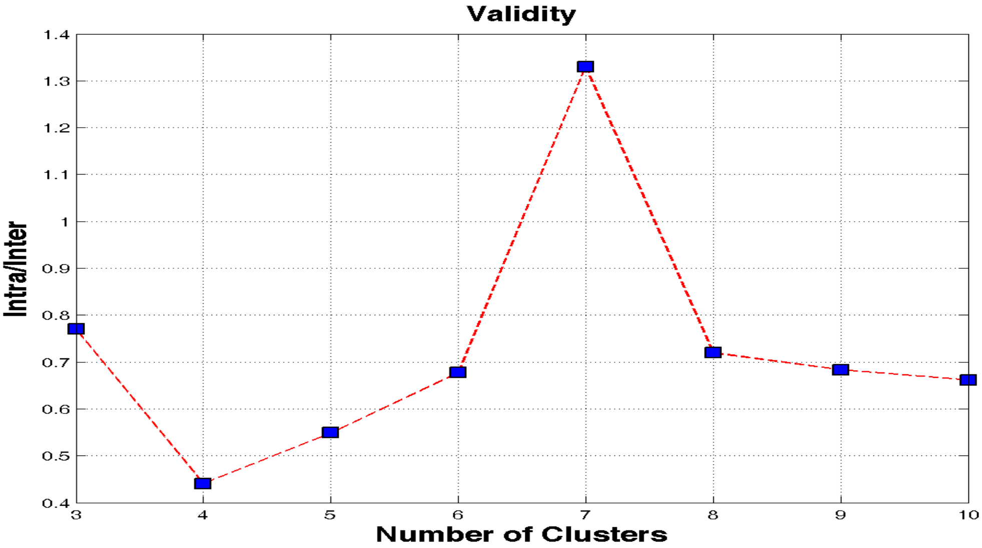


Figure 5.32: validity for every number of cluster

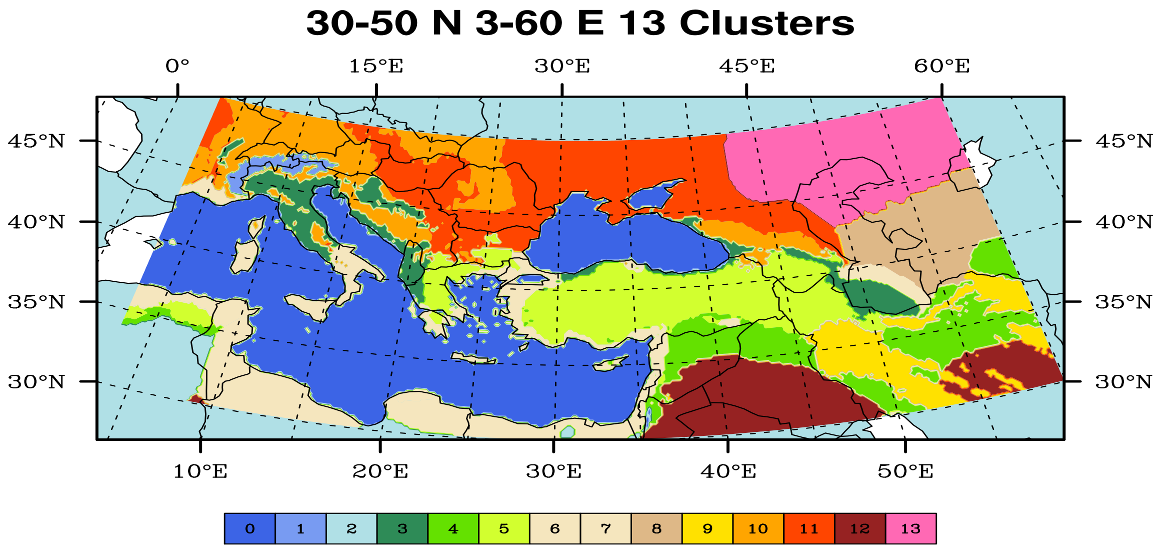


Figure 5.33: 9 PCs from $30^o - 50^o$ N $3^o - 60^o$ E without cv and elevation

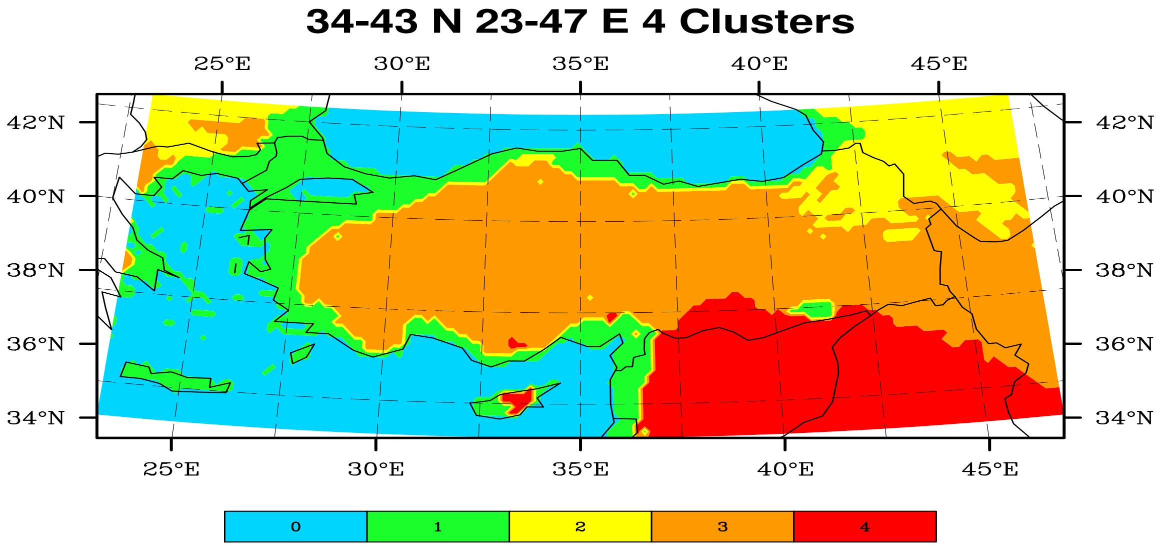


Figure 5.34: 10 PCs from $34^o - 43^o$ N $23^o - 47^o$ E without cv and elevation

Lastly, in the final combination elevation is included, but cv is removed from the set. When the analysis is done, representative dimensions are acquired as 9 and 11 by evaluating the eigenvalues corresponding to our data sets. Moreover, those principal components suggests 9 and 4 partitions where the optimum values are 0.09460 and 0.48502.

Table 5.18: Eigenvalues for all data set without cv with elevation

#	1^{st} Region	2^{nd} Region
1	50.68	41.62
2	15.30	18.5
3	10.99	9.37
4	6.06	7.11
5	3.66	4.78
6	2.20	3.98
7	1.50	2.58
8	1.29	2.06
9	**1.03**	1.19
10	0.65	**1**
11	0.60	0.91
12	0.49	0.89

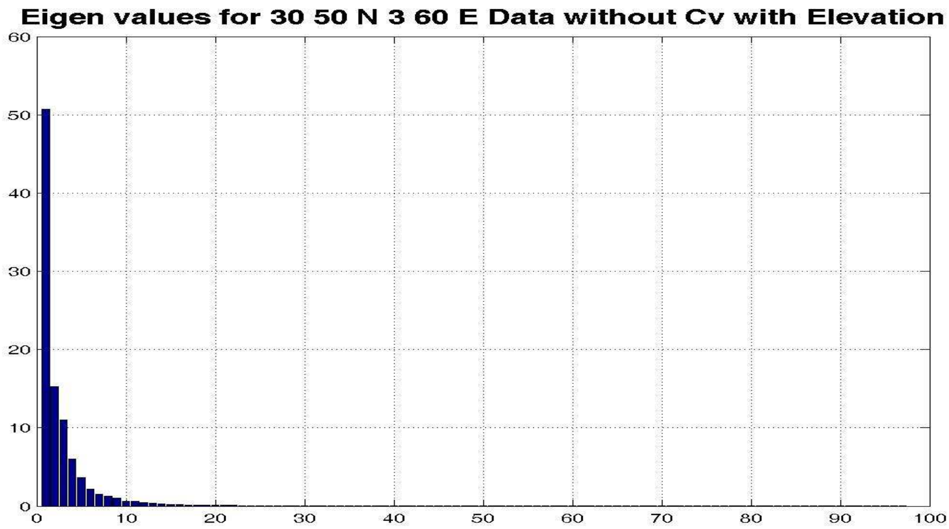


Figure 5.35: Eigenvalues from greatest to smallest

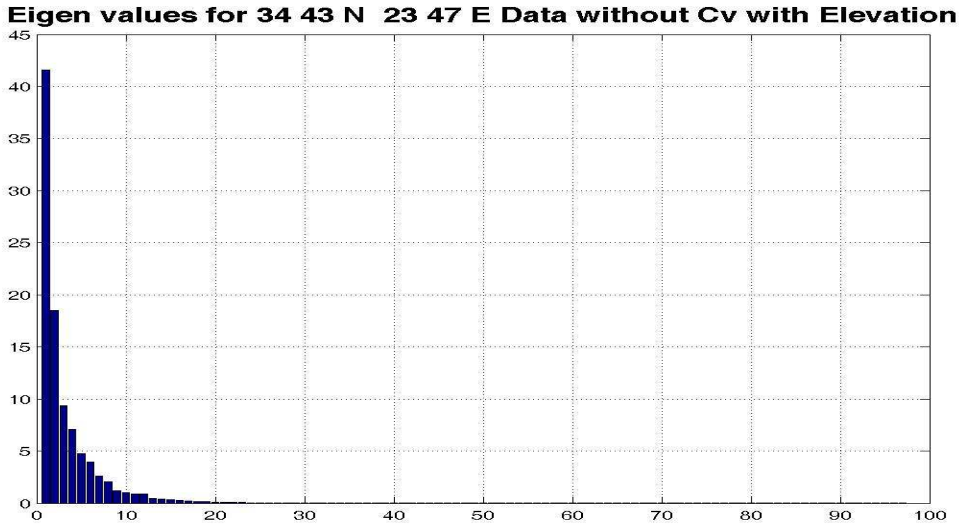


Figure 5.36: Eigenvalues from greatest to smallest

Table 5.19: Validity Results for 9 PC from $30^o - 50^o$ N $3^o - 60^o$ E without CV with Elevation

# of clusters	Intra Distance	Inter Distance	Validity
3	46.428	98.380	0.47192
4	37.044	93.603	0.39575
5	**32.949**	**96.431**	**0.34168**
6	30.171	44.591	0.67661
7	28.440	60.737	0.46825
8	24.367	47.846	0.50927
9	23.549	38.389	0.61343
10	21.524	40.683	0.52908
11	20.510	35.253	0.58179
12	19.321	32.242	0.59926
13	18.202	32.556	0.55910
14	16.969	33.523	0.50621
15	17.299	40.653	0.42553
16	16.080	19.028	0.84507
17	15.638	27.596	0.56668
18	14.393	31.776	0.45295
19	13.732	18.838	0.72893
20	14.768	16.945	0.87154

Table 5.20: Validity Results for 10 PC from $34^o - 43^o$ N $23^o - 47^o$ E without CV with Elevation

# of clusters	Intra Distance	Inter Distance	Validity
3	53.464	70.444	0.75895
4	**43.863**	**99.390**	**0.44133**
5	38.835	72.393	0.53645
6	34.927	52.582	0.66424
7	32.925	24.275	1.35636
8	30.056	42.374	0.70929
9	27.869	41.010	0.67955
10	26.502	40.674	0.65156

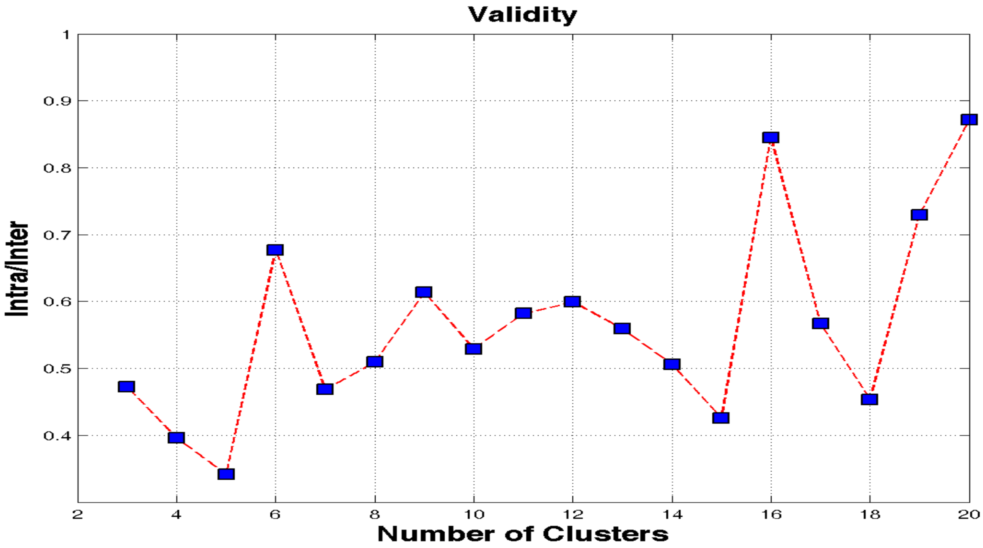


Figure 5.37: validity for every number of cluster

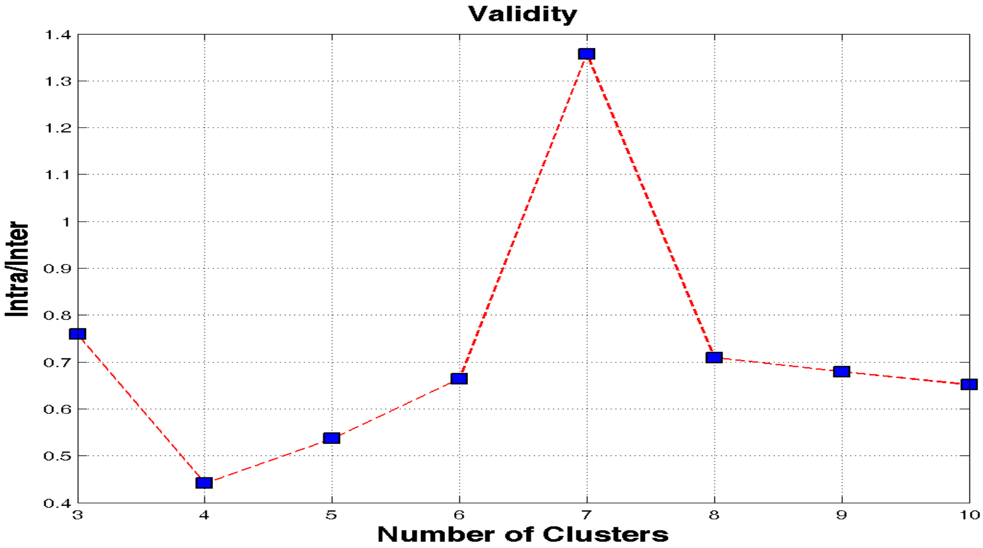


Figure 5.38: validity for every number of cluster

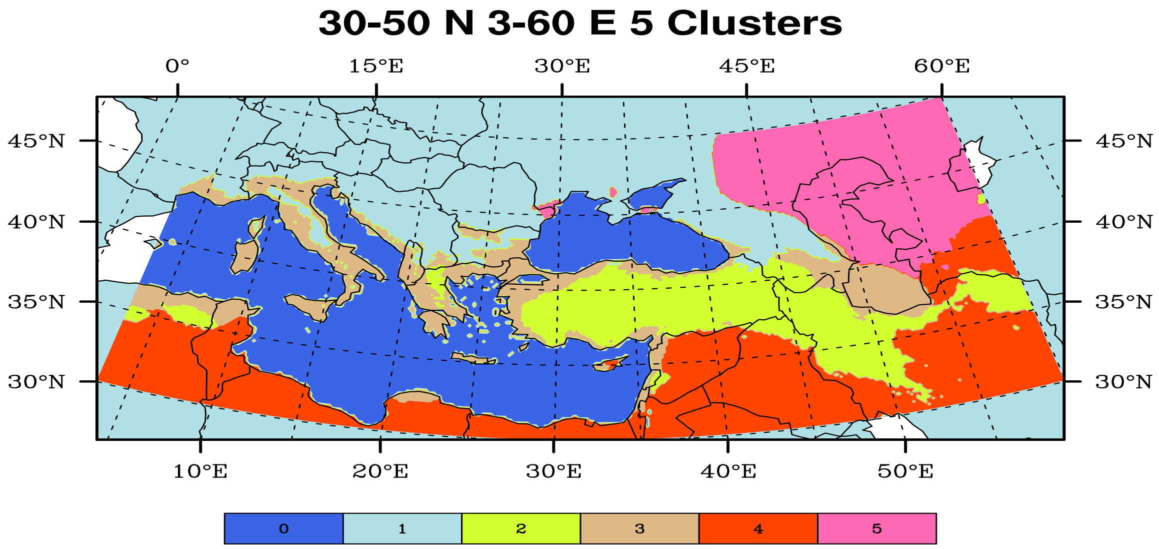


Figure 5.39: 9 PCs from $30^o - 50^o$ N $3^o - 60^o$ E without cv with elevation

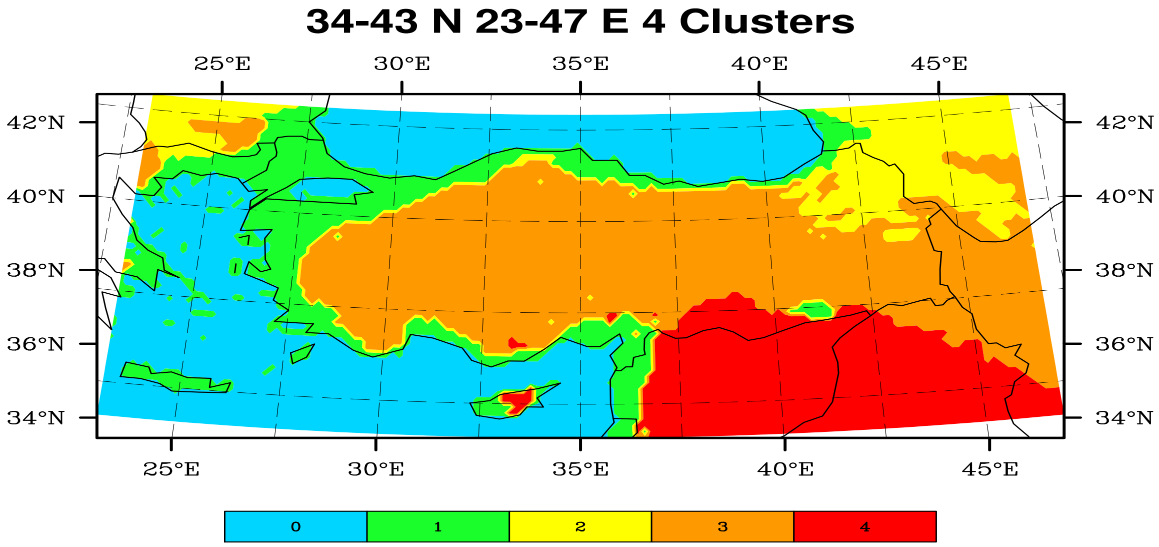


Figure 5.40: 10 PCs from $34^o - 43^o$ N $23^o - 47^o$ E without cv with elevation

5.3 Visualizing the distances from members to centroids

In the previous part, coloring was done according to the labels of each sample. In other words, random colors were assigned to each label that we obtained after cluster analysis. What we investigate in this section is how far the members away from the mean that belongs to their partition. To obtain the distances, we have modified the code where we obtain the final configuration. When the algorithm gives the final clusters that we indicated beforehand, we compute the Euclidean distances. Let N_j be the order of n dimensional cluster C_j. Distance between $x_i \in C_j$ and μ_j, which is the mean of cluster C_j, can be calculated as follows:

$$d(x_i,\mu_j) = (\sum_{k=1}^{k=n} \|x_{ik} - \mu_{jk}\|^2)^{1/2} \quad i \in \{1,2,...,N_j\} \tag{5.1}$$

In contrast to previous work, where we assigned a color for each cluster number, our aim here is to be able to present the intermediate values. In order to clarify, after associating a point with a cluster number which comes with a particular color for that cluster, we are going to add the position within the cluster to its group label. Position of every sample is decided relying on a normalization process. We first normalize the distances within their partitions, then add their normalized value to their class labels which were cluster numbers as we mentioned above. Normalization is done as follows:

Let $d_{ij} = d(x_i,\mu_j)$ and $d_{ij}^{'}$ is the normalized values.

$$d_{ij}^{'} = \frac{d_{ij} - min(d_{ij})}{max(d_{ij}) - min(d_{ij})} \tag{5.2}$$

From (5.2) one can easily realize that all $d_{ij}^{'}$ will vary from 0 to 1. Since we are regarding the numeric value of each sample while coloring and the toolbox provides contours for those values, we simply add the normalized distance to its class label, i.e., $j + d_{ij}^{'}$. Instead of contouring for discrete values, NCL will draw contours over the map for floating numbers. Consequently, we will obtain a stepwise coloring of our climate regions. Under the same criteria like in the previous work, we have maps that are able to represent how well a point exhibits its climatic characteristics with respect to others in the same cluster. As the color gets darker, we derive that it's far away from the centroid and among the ones that are in the vicinity of next cluster. However, we cannot claim that representation is strong enough because NCL cannot assign colors for every floating number which means that it assigns colors for intervals. Although there are many steps, it reduces the step size which causes a less representative contouring. Here are two maps based on this process from two regions where cv and elevation variables are excluded.

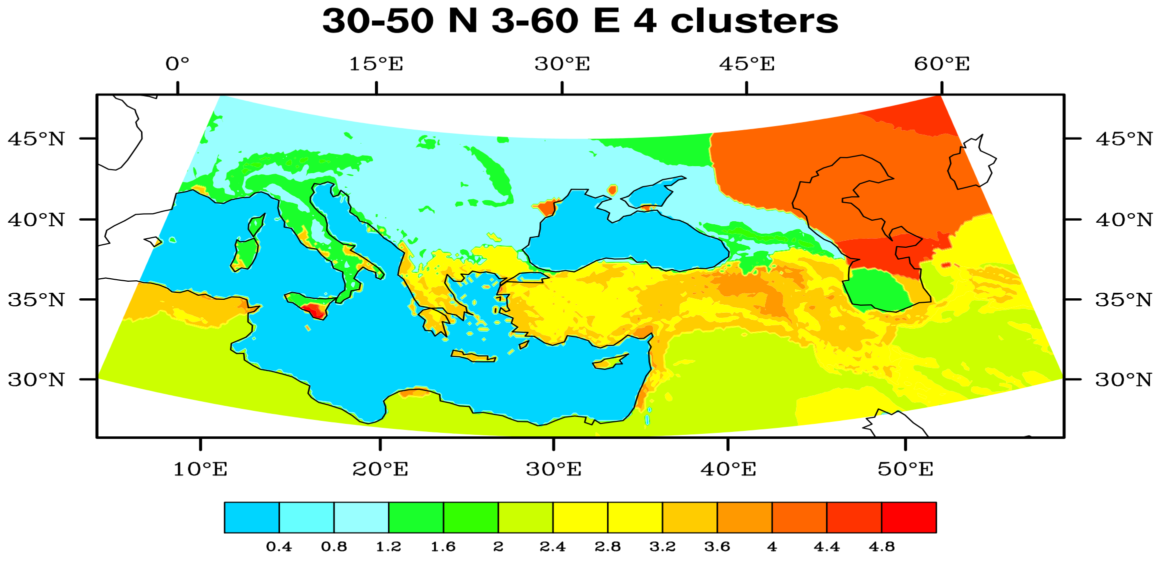


Figure 5.41: $30^o - 50^o$ N $3^o - 60^o$ E without cv and elevation .

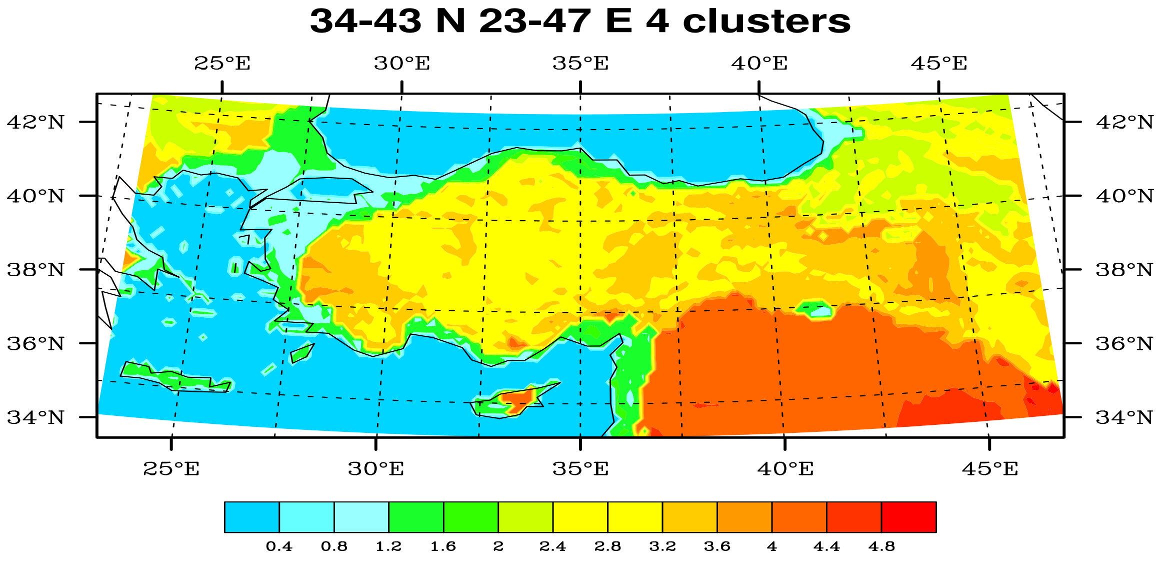


Figure 5.42: $34^o - 43^o$ N $23^o - 47^o$ Ewithout cv and elevation .

5.4 Results and Conclusion

As we mentioned earlier, different cases have been examined through the cluster analysis. Not only the clustering process, but also the visualization via maps provide us different climate zone definitions. It is worthwhile to bear in mind that there may be apparent differences between the representation of broader and narrower zones. For each region, we are going to interpret the results with different variable combinations as well as principal components.

5.4.1 With Cv and Elevation

When we work with the whole data sets in which coefficient of variations and elevation included, for the first regions we 6 climate zones where most noticeable thing is the places seaside which share the same characteristics. Lands neighboring Mediterranean, Black Sea and South Caspian Sea are the ones from this cluster. While a wide range area in Italy and Greece has this property, in Turkey only the parts nearby sea and Thrace are put in the same zone. Moreover, whole inner Anatolia is in another cluster which has also common property with the areas below South East Caspian Sea and a small part in North Africa. Remaining clusters are the blocks including West and South Europe as well as Ukraine with South Russia.

If we try to represent the data set with 9 principal components, our algorithm suggests 9 partitions over the same area. At the first glance, similarity with the previous partitioning can be realized because areas nearby sea are mostly preserved in this map also. Blocks are also conserved, but some new partitions occur over South Caspian Sea, inner Libya and Switzerland. In contrast to the previous one, the region surrounded by Alps appears as another climatic region.

In the second region that is to be split into zones we have both types of variables. PCA suggests 11 components to represent the whole set. For both initial and the reduced dimensions, computations resulted in 4 clusters which are nearly the same. Like in the broader region we see that seaside areas have common climatic characteristics. In addition, in both representations, between the inner Anatolia and the seaside a thin line draws attention that has similar characteristics with North East and North West of the area.

5.4.2 With Cv Without Elevation

Blocks that we acquired in the previous map are roughly preserved again in this case. In contrast to the 6 clustered version, Greece has a greater area that is similar to inner Anatolia. Alps and its surroundings present a separate nature. Furthermore, around the Caspian Sea it is possible to notice three climatic zones. The first one lies in the South; the second one covers the middle and expands towards East and the last one is in the North and continues in the same direction. Nine principal components with 13 zones do not affect the distribution of the zones around the Caspian Sea at all. Whereas all the seaside areas were in the same cluster, vicinity of Black Sea in North Anatolia, most of Italy and South Mediterranean Sea are now a separate group. Similarity between Greece and Turkey changes in a way that both have same characteristics in inner regions and in the neighboring areas of the Mediterranean Sea. Alps still preserve its climatic borders. In addition to this new formation, the Balkans exhibits a different characteristics with a split version of the previous partitioning.

Concerning the second region, we can claim that both high and reduced dimensions with 11 principal components provide the same map with 4 clusters. Only in a few places we notice differences which do not affect the generalizations at all.

5.4.3 Without Cv and Elevation

The neighboring lands of Caspian Sea from Turkmenistan, Kazakhistan and Iran forms the same blocks again with some insignificant differences. Inner parts of Iran, Northern Iraq and Northern Syria fall into the same cluster.

In this configuration, Southern Syria is located in a cluster with some members from Southern Iran and small area from Libya. Another evident separation occurs around the seasides. Namely, in the Mediterranean, Northern Egypt, Northern Libya and Italy excluding inner regions show the similar characteristics. However, in the same region Southern Anatolia, West Anatolia and Greece are members of another division. Alps are located like in the previous experiments. Commonality between Moldova, Romania, Ukraine and some parts of Russia that are next to Georgia is sustained.

Now, we let 9 principal components help us to do another cluster analysis. Divisions including the regions Caspian Sea and the areas below the Northern Iraq are represented nearly same as the previous one. Alps form a detached cluster. As opposed to all previous maps, Italy is divided into three main zones. South Italy with North Libya, North Egypt, Mediterranean and Aegean side of Anatolia are similar, whereas East Greece, East Bulgaria and Central Anatolia are the members of the same group. Moreover, a small area in the Central Italy has a commonality with Germany and France. Membership of the third part that is the area from Central Italy to the North of the country and seaside of the Balkans also overlaps.

When we focus on the narrower region, we do not have significant difference from partitioning done before. Some small changes begins to be appear. For instance, in the Southern part of Turkey, two small portions that are in the cluster of Syria are noticed within another zone. Likely, again with same number of clusters 10 principal components give almost same distribution.

5.4.4 Without Cv with Elevation

In the experiment where we excluded both cv and elevation we had 14 climatic zones defined. If let the elevation take place in the cluster analysis, we have 15 clusters as an outcome. The most apparent change occurs in Bulgaria and Romania which were together with Ukraine and South Russia until this experiment. Therefore, towards the North West a new cluster appears when compared the 14 clustered map. In addition, inner parts of Iran exhibits a compact picture with respect to the last delineation again.

After the PCA, we decide on 9 principal components which were responsible for the clustering stage. As a result, we are to delineate the region in 5 climatic zones which come up with 5 blocks. In order to clarify, Inner and East Anatolia, some parts of Northern Iraq and North West of Iran with some inner areas form a block along the line from the West to the East. Another block extends from Georgia to France along the upper side of Black Sea including Russia, Ukraine, Moldova and Romania. An apart member of this group is located on a stripe from North Italy to South. Libya, Syria, Iraq and West Turkmenistan remains to be another ostensible band below the region.

Lastly, due to the lack of cv with the small portions appearing in the South, we have again our 4 clustered map which consistently remains to be same except a few details. Same generalization can also be made for the distribution based on 10 principal components.

6. CONCLUSION

Without involving any discussion concerning objectivity issue for definition of climatic zones, in this study we have showed that k-means clustering technique is capable of dividing the regions that we worked on independent from any personal expertise or interpretation. Thus, we have examined different combinations of variables based on an objective mathematical modeling which has been employed with its parallel version also. In this study, other than handling whole data set which is high dimensional, we also utilized PCA to work with reduced but representative dimensions. Evaluations of the eigen spectrum of covariance matrices of our data sets have determined the number of principal components. In general, we attained a variety of dimensions from 9 to 11.

Number of clusters, data points and dimension have shown time consuming pattern of the algorithm. We obtained a scalable work with the high dimensional data set. Moreover, with the same data set, greater number of partitions which leads to an intensive computation made the code give more scalable outcomes. The reason behind those two facts is the low level communication overhead which is dominated when the size and the computational burden is increased.

In the combinations where coefficient of variation for precipitation is excluded, we try to observe how elevation effects the distributions. If the elevation is considered as an input parameter, we notice that the Balkans are evidently separated from the block above the Black Sea. Furthermore, Greece and seaside parts of Western Anatolia are in the same climatic zone. When we look at Northern Italy, we realize another dissemblance w.r.t the case where elevation is not taken into account. However, when it is tried to be delineated based on 9 principal components a similar partition occurs. On the other hand, if coefficient of variation is put into experiments, North Libya and North Egypt come into scene as a climatic block, but some parts of Libya present another climatic property in the absence of elevation. Actually, in the exclusion of elevation 4 more clusters appear. Besides the new formations in the South Mediterranean, around the Caspian Sea, Alps and in the North Western direction of Thracia other arrangements draw attention. In addition to all comparisons above, we note that a consistent climatic zone, which has small modifications in all cases, goes along the direction from West Anatolia to the region below South Caspian Sea.

Whereas there are different number of clusters for the first region, as we noted earlier, we have 4 climatic zones for the second region where Turkey is the main experimental area within the frame based on a data set in which elevation and coefficients of variation are added and removed. What pays attention is the climatic block which lies on the line from West to East like in the previous frame that we mentioned above. Remaining divisions are generally compact, but occasionally they may include members of other zones. It is also another fact that in both North East and North West areas of the region, there are considerable amount of members of the same climatic characteristics. When those results are compared with the ones that Karaca and others found out, it is noticed that their results are similar to the current configuration of the zones [27]. In particular, our model concludes that all the seasides are the same whereas in Karaca's work Black Sea and Mediterranean parts are seen different. On the other hand, a common outcome of our studies is that the Aegean and Marmara regions were considered as the same region. Moreover, characteristics of South Eastern Anatolia and Eastern Anatolia are found similar in both studies.

Finally, k-means clustering algorithm provides us a variety of zone definitions with respect to the criterion we have posed on it. Variety is not only dependent on the criterion we put, but also dependent on the nature of the data sets having different combinations of variables as well as the PCA. Regarding all the facts concerning both PCA and variable combinations, we can conclude that within the frame we have focused Turkey has 4 climatic zones.

REFERENCES

[1] **D. B. Carter and J. R. Mather**, 1966. Climatic Classification for Environmental Biology, *Publications in Climatology*, **19**(4)

[2] **W. W. Hargrove and F. M. Hoffman**, 2005. Potential of Multivariate Quantitative Methods for Delineation and Visualization of Ecoregions, *Environmental Management* **34**(1), S39-S60

[3] **R.O. Duda, P.E. Hart and D.G. Stork**, Pattern Classification, *Wiley-Interscience Publication*, 2000.

[4] **M. Matteo**, A Tutorial on Clustering Algorithms, URL(cited on 02/17/2008): *http://home.dei.polimi.it/matteucc/Clustering/tutorial_html*

[5] **A. K. Jain, A. Topchy, M. H. C. Law, and J. M. Buhmann**,2004. Landscape of Clustering Algorithms, *Proc. IAPR International Conference on Pattern Recognition*, Cambridge, UK.

[6] **G. McMahon, S. M. Gregonis, S. W. Waltman, J. M. Omernik, T. D. Thorson, J. A. Freeouf, A. H. Rorick, and J. E. Keys**,2001. Developing a spatial framework of common ecological regions for the conterminous United States, *Environmental Management*, **28**, 293-316.

[7] **L. R. Holdridge**,1947. Determination of world plant formations from simple climatic data, *Science*, **105**, 367-368.

[8] **P. Halasz**, Holdridge life zones, URL (cited on 02/17/2008): *http://en.wikipedia.org/wiki/Holdridge_life_zones*

[9] **B. A. Malmgren and A. Winter**, 1999. Climate zonation in Puerto Rico based on principal components analysis and an artificial neural net, *Journal of Climate*, **12**, 977-985

[10] **D. Pullar, S. Low Choy and W. Rochester**, 2004. Ecoregion Classification Using a Bayesian Approach and Model-based Clustering, *International Congress on Environmental Modelling and Software.*

[11] **L. Belbin**, 1993. Environmental representativeness: regional partitioning and reserve selection, *Biological Conservation*, **66**, 223-230.

[12] **J. A. Bernert, , J. M. Eilers, T. J. Sullivan, K. E. Freemark, and C. Ribic**, 1997. A quantitative method for delineating regions: an example for the Western Corn Belt Plains ecoregion of the USA, *Environmental Management*, **21**, 405- 420.

[13] **G. B. Host, , P. L. Polzer, D. J. Mladenoff, M. W. White, and T. R. Crow**, 1996. A quantitative approach to developing regional ecosystem classifications, *Ecological Applications*, **6**,608-618

[14] **J. Harff and J. C. Davis**, 1990. Regionalization in geology by multivariate classification, *Mathematical Geology*, **22**, 573-588

[15] **R. M. Lark**, 1998. Forming spatially coherent regions by classification of multi-variate data: An example from the analysis of maps of crop yield. *International Journal of Geographic Information Science*, **12**, 83-98.

[16] **S. E. Carter**, 1997. Spatial stratification of western Kenya as a basis for research on soil fertility management, *Agricultural Systems*, 55, 45-70

[17] **M. E. Jensen, I. A. Goodman, P. S. Bourgeron, N. L. Poff, and C. K. Brewer**, 2001. Effectiveness of biophysical criteria in the hierarchical classification of drainage basins. *Journal of the American Water Resources Association*, **37**, 1155-1167

[18] **Y. Zhou, S. Narumalani, W. J. Waltman, S. W. Waltman, and M. A. Palecki**, 2003. A GIS-based spatial pattern analysis model for ecoregion mapping and characterization, *International Journal of Geographic Information Science*, **17**, 445-462

[19] **J. R. Leathwick**, 2001. New Zealand's potential forest pattern as predicted from current species-environment relationships, *New Zealand Journal of Botany*, **39**, 447-464

[20] **P. Esteban, P. D. Jones , J. Mart´ın-Vide and M Mases**, 2005. Atmospheric circulation patterns related to heavy snowfall days in Andorra, Pyrenees, *International Journal Of Climatology*, **25**, 319-329

[21] **P. Esteban, P. D. Jones, J. Mart´ın-Vide and M Mases**, 2006. Daily Atmospheric Circulation Catalogue For Western Europe Using Multivariate Techniques, *International Journal Of Climatology*, **26**, 15011515

[22] **D. Peñarrocha, M. J. Estrela and M. Millán**, 2002. Classification of Daily Rainfall Patterns In A Mediterranean Area With Extreme Intensity Levels: The Valencia Region, *International Journal Of Climatology*, **22**, 677-695

[23] **P. H. Whitfield, K. Bodtkerb and A. J. Cannona**, 2002. Recent Variations In Seasonality Of Temperature And Precipitation In Canada, 197695, *International Journal Of Climatology*, **22**, 1617-1644

[24] **W. W. Hargrove and R. J. Luxmoore**, 1998. A clustering technique for the generation of customizable ecoregions, *Proceedings, ESRI Arc/INFO Users Conference*

[25] **W. W. Hargrove and F. M. Hoffman**, 1999. Using multivariate clustering to characterize ecoregion borders, *Computers in Science & Engineering*, **1**, 18-25

[26] **W. W. Hargrove, F. M. Hoffman, G. Mahinthakumar, N. T. Karonis**, 1999. Multivariate Geographic Clustering in A Metacomputing Environment Using Globus, *Proc ACM/IEEE SC Conference*, 5

[27] **Y. Unal, T. Kindap and M. Karaca**, 2003. Redefining The Climate Zones Of Turkey Using Cluster Analysis, International Journal Of Climatology, **23**, 1045-1055

[28] **A. K. Jain, M. N. Murty and P. J. Flynn**, 1999. Data Clustering: A Review, *ACM Computing Surveys*, 31(3)

[29] **D. Pollard**, 1981. Strong Consistency Of K-Means Clustering, *The Annals of Statistics*, **9**(1), 135-140

[30] **J. MacQueen**, 1967. Some Methods for Classification and Analysis of Multivariate Observations, *In the 5th Berkeley Symposium on Mathematical Statistics and Probability*

[31] **M. Steinbach, G. Karypis and V. Kumar**, 2000. A Comparison of Document Clustering Techniques, *KDD Workshop on Text Mining*, University of Minnesota

[32] **C. F. Olson**,1994. Parallel Algorithms for Hierarchical Clustering, *Technical Report UCB//CSD-94-786*, *University of California at Berkeley*

[33] **S. Sideris**,URL (cited on 02/17/2008):*http://en.wikipedia.org/wiki/Data_ clustering*

[34] **P. Tan, M. Steinbach and V. Kumar**, 2005. Introduction to Data Mining, *Addison Wesley*

[35] **Z. Du and F. Lin**, 2005. A novel parallelization approach for hierarchical clustering, *Parallel Computing* , **31**

[36] **Y. M. Marzouk, A. F. Ghoniem**, 2005. K-means clustering for optimal partitioning and dynamic load balancing of parallel hierarchical N-body simulations, *Journal of Computational Physics*, **207**, 493-528

[37] **T. Jinlan, Z. Lin , Z. Suqin and L. Lu**, 2005. Improvement and Parallelism of k-Means Clustering Algorithm, *Tsinghua Science And Technology*, **10**(3), 277-281

[38] **E. Alpaydın**, 2004. Introduction to Machine Learning (Adaptive Computation and Machine Learning), *The MIT Press*

[39] **M. Palmer**, URL (cited on 02/17/2008):*http://ordination.okstate.edu/PCA.htm*

[40] **W. W. Cooley and P. R. Lohnes**, 1971. Multivariate Data Analysis, *John Wiley & Sons, Inc.*

[41] **The NCAR Command Language (NCL)**, URL (cited on 02/17/2008): *http://www.ncl.ucar.edu*

[42] **M. New, D. Lister, M. Hulme, I. Makin**, 2002. A high-resolution data set of surface climate over, global land areas, *Climate Research*, **21**, 1-25

[43] **J. C. Bezdek, N. R. Pal**,1998. Some new indexes of cluster validity, *Systems, Man, and Cybernetics, Part B, IEEE Transactions*, **28**(3)

[44] **S. Ray and R. H. Turi**, 1999. Determination of Number of Clusters in K-Means Clustering and Application in Colour Image Segmentation, *The 4th International Conference on Advances in Pattern Recongnition*

[45] **W. Liao**, The Software Package of Parallel K-means, URL(cited on 02/17/2008):*http://www.ece.northwestern.edu/˜wkliao/Kmeans/index.html*

[46] **A. Grama, G. Karypis, V. Kumar and A. Gupta**, 2003. An Introduction to Parallel Computing: Design and Analysis of Algorithms, Second Edition, *Addison-Wesley*

CIRRICULUM VITAE

Halil Bişgin was born in 1980 in Tokat. He graduated from Sivas Lisesi in 1998 and received his B.Sc degree from Mathematics Department in Koc University in 2003. He has been a M.Sc. candidate and an employer as a research assistant in the Computational Science and Engineering Department of Informatics Institute in Istanbul Technical University since October, 2004.