University of Benghazi

Faculty of Science

Department of Statistics

The Application of The Block PCA on The Student’s GPA’s Database of Science Faculty of Benghazi University

By :

Osama .H. Othman

Supervised by :

DR.Rami Salah M. Gebril

A THESIS SUBMITTED FOR PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE MASTER OF SCIENCE IN STATISTICS

BENGHAZI –LIBYA

2011-2012

بسم الله الرحمن الرحيم

وَقُل رَّبِّ زِدْنِي عِلْمًا

سورة طه

(114)

Contents

Chapter 4:Application of Block Principle Component Analysis on Students' GPA'S Data

Dedication

I dedicate this thesis ...

To my precious prophet Mohammed (peace be upon him); I present this modest work to my great messenger and intercessor.

To the soul of my mother (May God bless her).

To my dear father.

To my beloved brother and sisters.

Acknowledgment

First of all, I would love to thank Allah for his help and support in finishing my postgraduate study and this thesis. And I would like to express my gratefulness to the Almighty for his generosity and for everything he has given to me.

I would like to thank my parents for taking care of me throughout the years of my life; from the day I was born to this moment. I appreciate their help and continuous encouragement to achieve my goals in my life.

I cannot ever forget the role of my older brother Ahmed. I will always be grateful for his help, his support in all areas; financially, emotionally, and scientifically. He has been with me in every step of my life.

My sincere appreciation goes to my supervisor DR.RAMI SALAH M.GEBRIL for his suggestions and his extremely significant help.

My appreciation also goes to the teaching staff members of the Statistics Department who taught me and my colleagues during our

postgraduate studies especially my colleague Magde.

There are some people like the programmer Dr.Abd Albaset who gave us the database and gave us a very understandable description of the nature of the database, also the programmer Ahmed Qasem who gave the study very significant help in data preparation phase, we thank them so much for their effort.

ABSTRACT

In many studies, it was found that the application of most inferential statistical techniques using huge data bases is difficult, inconvenient and in some situations unreliable. One of the modern solutions adopted to deal with large number of variables is the Block Principal Component Analysis (Block PCA). This modified technique is used to reduce data by selecting those variable containing most of the information. These selected variables can be then employed in further investigations and analyses.

In this thesis, the Block PCA is used to reduce the size of a huge database consisting of Grade Point Average (GPA) of the students in all courses of science faculty in Benghazi University.

In other words, constructing a substitutive database of the GPA's of the students with less variables containing most variation in the original database.

Block PCA is a multistage adapted technique of the original PCA. It involves the application of Cluster Analysis (CA) and variable selection throughout sub PC's. The application of Block PCA in this thesis is a modified version of the original work of Liu et al (2002) and Gebril (2005). The main objective was to apply PCA on variables in smaller groups instead of the whole large pack of variables which was proved to be unreliable.

In this study the number of variables which have the majority of the variation among all the included 251 variables are just 12. The 12 final selected variables(courses) has been selected by using the data reduction

and variable selection Block PCA, and hence we had the GPA'S data base of Science Faculty with smaller size and the most amount of the variation.

List of Tables

List of Figures

Chapter 1

Introduction to Multivariate Analysis

1.1 Introduction:

Multivariate analysis is considered to be one of the most important topics and branches of statistics. The importance of this set of techniques and methods of multivariate analysis comes from the variety of the phenomena that are present or appear in our lives in a way of multivariate nature.

In other words, those phenomena in the universe have many random varieties or variables that contribute to them or affect them in different degrees or levels. It is very rare to find a phenomenon or to conduct a study that involves one, two or little more variables; hence the need for multivariate analysis techniques arises and increases with the development in all the disciplines. Additionally, according to the nature of variables in the earlier decades of the 20th century, it was too difficult to conduct large and intensive researches and studies, because of the lack of computers and /or computers ability in the sense that they did not have sizeable capacity, speed and performance.

Nowadays most of the previous problems have been solved; in recent years, computers have seen a massive revolution in many aspects, and a large amount of software and packages have been invented which serve the needs and requirements of the multivariate analysis.

1.2 General Objectives of Multivariate Analysis:

The objectives of scientific investigations by using multivariate analysis methods are many, some of which will be mentioned here:

1.2.1 *Data reduction:*

The phenomena being studied is represented as simply as possible without sacrificing valuable information. It is hoped that will make the interpretation easier. In other words, the aim of a statistician undertaking multivariate analysis is to reduce the number of variables by employing suitable linear transformations to choose a very limited number of the resulting linear combinations in some optimum manner, disregarding the remaining linear combinations in the hope that they do not contain much significant information. The statistician thus reduces the dimensionality of the problem.

1.2.2 *Sorting and grouping:*

Groups of similar cases or variables are made according to measured attributes or characteristics, which means classifying cases or variables so that they belong to well-defined groups and are homogenous.

1.2.3 *Investigation of the dependence among variables:*

The researcher is always interested in the nature of the relationship between variables under consideration. Whether all of them depend on each other, and if so, to what extent? Or are all or some of them mutually independent?

1.2.4 *Prediction:*

Relationships between variables must be determined for the purpose of predicting the values of one or more variables on the basis of observations on the other variables.

1.2.5 *Inference about the population parameters (test of hypothesis and estimation):*

Statistical inference in its two classes is very useful in the sense that it can be used in the context of multivariate analysis to draw and conclude assumptions regarding the parameters of multivariate populations through the sample information.

1.3 Overview of Multivariate Techniques

There are several different techniques of multivariate analysis so it is difficult to get a complete acceptable classification. However, it could be classified into two classes according to the nature of the study.

If interest centers on the association between two sets of variables, where one set is the realization of a dependent or criterion measure then the appropriate class of techniques would be those designed as dependence methods. If interest centers on the mutual association across all variables with no distinction made among variables types, interdependence methods are used. Note that dependence methods seek to explain or predict one or more criterion measures based upon the set of predictor variables. Interdependence methods, on the other hand, are less predictive in nature and attempt to provide insights into the underlying structure of the data by simplifying the complexities, primarily through data reduction.

1.3.1 Dependence methods:

Depending on the nature and the number of variables the researcher wishes to study, there are several multivariate techniques that analyze dependence structure.

Dependence methods can be further classified according to:

1- The number of independent variables-one or more than one.

2- The number of dependent variables-one or more than one.

3- The type of measurement scale used for dependent variables (i.e., metric or nonmetric).

4- The type of measurement scale used for the independent variables (i.e., metric or nonmetric).

The following gives are the most common techniques.

1- Multivariate Regression.

2- Discriminant Analysis.

3- Logit Analysis.

4- Multivariate analysis of variance (MANOVA).

5- Canonical correlation Analysis.

6- Path Analysis:

1.3.2 Interdependence methods:

1- Principle Component Analysis.

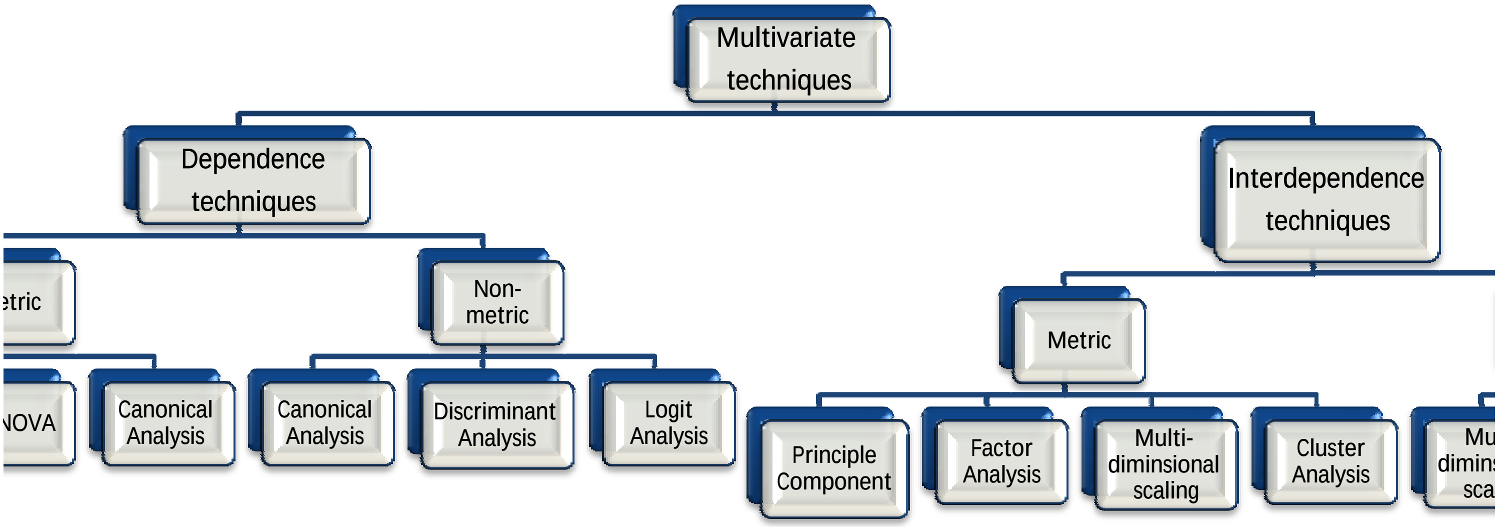
2- Factor Analysis.

3- Cluster Analysis.

4- Multidimensional Scaling Analysis.

5- Loglinear Modeling Analysis.

Figure (1.1) gives a visual presentation of the classification of Multivariate Analysis methods.



sification of multivariate techniques.

thew Golden.(1984). Multivariate Analysis (Methods and Applications).

1.4 Data Reduction:

Data or dimensionality reduction is a very common, desired and critical topic in statistics, because in the model building phase in any statistical research or study the researchers are always annoyed by the number of variables to be included or excluded in a model. As has been seen recently, many statistical studies have huge numbers of variables as a consequence of the evolution of all disciplines of sciences.

Some statisticians regard the huge number of variables as a merit or an advantage, because they want to express the factors that affect the underlying study implicitly. On the other hand, the huge number of variables would make a model very complex and tedious in the sense of getting inferences or results during the analysis phase. From that point of view, data reduction techniques and methods are now used in many aspects and ways and have many applications in all different domains of mathematics, statistics, data mining, computer sciences, artificial intelligence, economics, medicine and many other databases.

More typically, the data reduction process is applied to readings or measurements involving random errors. These are the intermediate errors inherent in the process of assigning values to observational quantities. In such cases, before data may be coded and summarized, the most probable value of a quantity must be determined. Provided the errors are normally distributed, the most probable (or central) value of a set of measurements is given by the arithmetic mean or, in the more general case, by the weighted mean. Data reduction may also involve operations of smoothing and interpolation, because the results of observations and measurements are always given as a discrete set of numbers, while the phenomenon being studied may be continuous in nature.(Micheline Kamber, Jiawei Han, 2006)

Data archiving and data compression can also reduce the amount of data needed to be stored on primary storage systems. Data archiving works by filing infrequently accessed data to secondary data storage systems. Data compression reduces the size of a file by removing redundant information from files so that less disk space is required.

In summary, data reduction techniques can be performed to get a reduced representation of the data set that is much smaller in size, and still almost maintains the integrity of the original data set. Moreover, mining on the reduced data set should be more efficient and produce the same or approximately the same results.

Strategies that are used for data reduction include the following

1- Data cube aggregation.

2- Dimensionality reduction.

3- Data compression.

4- Numerosity reduction.

5- Discretization and concept hierarchy generation.

The second strategy has been adopted to be practiced in this study; the objective of this study is dimensionality reduction and variable selection.

1.5 Literature Review:

This section has been set to present the historical facts about each of principle components analysis, cluster analysis and block principle components analysis. Due to the fact that the technique in this paper consists of a combination of two techniques, a brief historical notation for those techniques has been deemed necessary.

1.5.1 Principle Component Analysis (PCA) :

Principle component analysis was first introduced by Karl Pearson in the early (1900's). Formal treatment of the method is due to Hotelling in (1933). Hotelling's paper was in two parts.(see Jolliffe, 2010)

The first most important part, together with the Pearson's paper, is among the collection of papers edited by Bryant and Atchley (1975).The two papers adopted different approaches, with the standard algebraic derivation given above being close to that introduced by Hotelling (1933). Pearson (1901), on the other hand, was concerned with finding lines and plans that best fit a set of points in p-dimensional space, and the geometric optimization problems he considered also lead to PC's.

In the 32 years between Pearson's and Hotelling's papers, very little relevant material seems to have been published, although Rao (1964) indicates that Frisch (1929) adopted a similar approach to that of Pearson. Also, a footnote in Hotelling (1933) suggests that Thurstone (1931) was working along similar lines to Hotelling, but the cited paper, which is also in Bryant and Atchely (1975), is concerned with factor analysis.

A further paper by Hotelling (1936) gave an accelerated version of the power method for finding PCs; in the same year , Girshick (1936) provided some alternative derivations of PCs, and introduced the idea that sample PCs were maximum likelihood estimates of underlying population PCs. Girshick (1939) investigated the asymptotic sampling distributions of the coefficients and variances of PCs, but there appears to have been only a small amount of work on the development of different applications of PCA during the 25 years immediately following publication of Hotelling's paper.

Four papers will be mentioned; these appeared towards the beginning of the expansion of interest in PCA and have become important references within the subject.

The first of these, by Anderson (1963), is the most theoretical of the four. It discussed the asymptotic sampling distributions of the coefficients and

variances of the sample PCs, building on the earlier work by Girshick (1939), and has been frequently cited in subsequent theoretical developments. Rao's (1964) paper is remarkable for the large number of new ideas concerning uses, interpretations and extensions of PCA. Gower (1966) discussed links between PCA and various other statistical techniques, and also provided a number of important geometric insights. Finally, Jeffers (1967) gave an impetus to the really practical side of the subject by discussing two case studies in which the uses of PCA go beyond that of a simple dimension-reducing tool. Despite the apparent simplicity of the technique, much research is still being done in the general area of PCA, and it is very widely used. This is clearly illustrated by the fact that the web of science identifies over 2000 articles published in the two years 1999-

2000 that include the phrases 'principle component analysis' or 'principle components analysis' in their titles, abstract or keywords.

1.5.2 Cluster Analysis:

Clustering methods have been recognized throughout this century, but most of the literature on cluster analysis has been written during the past two decades. The major stimulus for the development of clustering methods was a book entitled *Principle of Numerical Taxonomy*, published in 1963 by two biologists, Robert Sokal and Peter Sneath .the listing of the contributions of Cluster Analysis is introduced in Roger K. Blashfield, Mark S. Aldenderfer (1984) .

Sokal and Sneath argued that an efficient procedure for the generation of biological classifications would be to gather all possible data on a set of organisms of interest, estimate the degree of similarity among these organisms, and use a clustering method to place relatively similar organisms into the same groups. Once groups of similar organisms were found, the membership of each group could be analyzed to determine if they represented different biological species. In effect, Sokal and Sneath assumed that "pattern represented process"; that is, the pattern of observed differences and similarities among organisms could be used as a basis for understanding the evolutionary process.

The literature on cluster analysis exploded after the publication of Sokal and Sneath's book. The number of published applications of cluster analysis in all scientific fields has doubled approximately once every three years from 1963 to 1975 (Blashfield and Aldendefer 1978). This rate of growth is much faster than that of even the most rapidly growing disciplines, such as biochemistry.

There are two reasons for the rapid growth of the literature on cluster analysis: (1) the development of high-speed computers and (2) the fundamental importance of classification as a scientific procedure. Before computers, clustering methods were cumbersome and computationally difficult when applied to the large data sets in which most classifiers were interested.

The social sciences have long maintained an interest in cluster analysis. Among the earliest of these studies were those by anthropologists who defined homogeneous culture areas by using methods of matrix manipulation (Czekanowski, 1991; Driver, 1965; Johnson, 1972). In psychology, cluster analysis was viewed as a "poor man's factor analysis" by one of its major proponents (Tryon, 1939). Other disciplines, most notably political science, were also involved in the early development of clustering in the social sciences. Although many of the theories and applications that served as the basis for clustering in the past have been repudiated by later generations of scholars, all social sciences now have strong modern traditions in the use of clustering methods.

1.5.3 Block Principle Component Analysis:

This technique has been proposed by (Liu et al, 2002) to solve the problem of extracting information from a database with large number of variables and a relatively small number of subjects without losing the ability of interpretation of the resulted principle components and that will be more meaningful than those which are resulted by ordinary principle components analysis.

Before the original paper of (Liu et al, 2002) with application to gene microarray data classification, there were some suggestions and papers introducing many ways to block PC'S. Most of those papers have been applied in the field of Image analysis. One of these previous works was by Qin et al (2001), about using multiblock PCA for decentralized process monitoring. Later, Qiu et al (2003) proposed blocking PCA in the field of image analysis for the image change detection by using Multi-Block PCA. Another application of the same context was introduced by Nishino et al (2005). This application was used for computer vision and image processing by using Clustered Block wise PCA to represent visual data. Another paper was introduced by Wang et al (2005) as an application of face recognition under the title The Equivalence of Two-Dimensional PCA to Line-Based PCA.

These previous papers are the most famous ones among papers published in this regard ; the technique has not had much application and most of these applications are in the field of image analysis and biostatistics. Our study will enrich the technique by adding another type of application using a new type of data which has not been used before.

1.6 Objectives of the Study:

In this study we have applied the Block Principle Component Analysis technique on a different type of data which may has not been applied before; all previous applications and studies that used this technique were in Biostatistics and in Image Analysis. The situation here is different and our application is in the context of using the statistics in problem solving in administration and management. Statistics have a very important role in administration. Therefore, we decided to choose the GPA'S from the Student Database at the Science Faculty of Benghazi University to check the problems faced by the administration of this faculty and to investigate whether the courses (variables) in this database have a considerable variation that is worth studying and looking for the causes of them. Moreover, this study is very meaningful, useful and helpful to our faculty in order to improve the way of teaching and support the faculty with modern education facilities after revealing the problems which can be found in terms of the variation.

The objectives in general can be summarized into three following points :

1- Dealing with the problem of large number of variables in PCA.

2- Monitoring the variation within clusters of variables.

3- Applying a methodology for variable selection in large databases.

4- discovering the most variant variables(courses) in order to know the reasons of the variations in future studies.

Next chapter is specified to the data preparation process and the topics which are related to that issue, we are going to reveal the stages of preparation our crude database from its nature as a raw data until the final data matrix that used in the main analysis.

Chapter 2

Data Preparation

2.1 Introduction

Raw data in most cases is an unordered, unorganized mass, not easily understandable and noisy. That mess is due to many reasons such as, missing values, outliers, extreme values and inconsistent and incomplete values. The noisy data contains errors and/or outliers values that deviate from the expected ones and inconsistent data are represented by discrepancies in the data. The data to be analyzed could be incomplete, and that means the lacking of some attribute values or certain attributes of interest. Those types of values are so common in the real world of huge databases and warehouses of data. There are many reasons for their occurrence.

As for incomplete data, the reason could be that some attributes are not always available about the phenomenon under consideration; the problem of missing value can occur due to the entry errors or due to errors of data collection and measurements. Noisy data which has incorrect attribute values may be found in the database because the data collection instrument used may be faulty or there might have been human or computer errors occurring at data entry. The necessity for data preparation or data preprocessing is that the raw data which is unordered cannot reveal any information, because the messy data hides the valuable information and the process of the extraction of that valuable information cannot be done without using data preparation tools. As sets of data become larger, getting a notation of their central tendency, variability or other general characteristic becomes more difficult as well as the interpretation of their results.

To overcome the difficulties of those previous problems, the tools of data preparation must be used. Data cleaning is one of the data preparation tools which works by filling in missing values, smoothing noisy data, identifying or removing outliers. In the sense that unorganized data can cause confusion and obstacles to the data analysis procedures or data mining procedures, data cleaning procedures are very important.

2.2 Data cleaning:

Data cleaning procedures are of big benefit and they have utilized to a large extent recently as a consequence of the improvement of the real world databases and warehouses. Several of them are used to fix problems like filling in missing values, smoothing out noise and identifying outliers, correcting inconsistencies and redundancies in the data.

In this chapter, only the tools that have been used will be introduced and explained in brief details.

2.2.1 Missing values:

Missing values is a very common and popular issue in databases nowadays. They occur due to many reasons. The missing values could be caused by the investigators themselves if the data collection phase was not done appropriately or if some mistakes were made in the data entry. Recently, the collection of various data types has spread considerably as a consequence of the improvement in all disciplines of sciences; as the data sets become larger, the number of missing values becomes higher. The problem of missing values if not fixed well can have a significant effect on the conclusions drawn from such data. Missing values can effectively reduce the representativeness of the sample and consequently can distort any inferences about the population because ignorance of this problem would surely lead to biased models and then the estimates wouldn't be correct. More than that, the missing values have bad effects in the data mining algorithms, where they affect the knowledge discovery operation in multi ways depending on the type of algorithm being used. The data mining algorithms that may be affected by them are K-nearest neighbor algorithm, decision trees, neural networks, and pattern recognition and association rules. The missing values are not of just one type, they have many types which are:

1- Data missing at random (MAR).
2- Data missing completely at random (MCAR).
3- Non-ignorable missing data.
4- Outliers treated as missing data.

There are some methods that can be used to solve the loss of the missing values.

1- Ignore the case.

2- Fill in the missing values manually.

3- Use a global constant to fill in the missing value.

4- Use the attribute mean to fill in the missing value.

5- Use the attribute mean for all samples belonging to the same class as the given tuple.

6- Use the most probable value to fill in the missing value.

Methods from 3 to 6 are biased to the data, the last method is the most popular one and is more efficient than the others, but it depends on some concepts like Bayesian formalism and Decision trees.

All of the previous methods are legal and can be used to fill in the data, they vary from one to another in their efficiency and accuracy.

In this study the fifth method was adopted since it is accurate and very popular. The topic of noisy data was not involved in this study as with the other topics of data cleaning because of the nature of our data, which does not involve and contain any of those problems. However, the detection of outliers was simply done by Box plot and will be introduced in the fifth chapter.

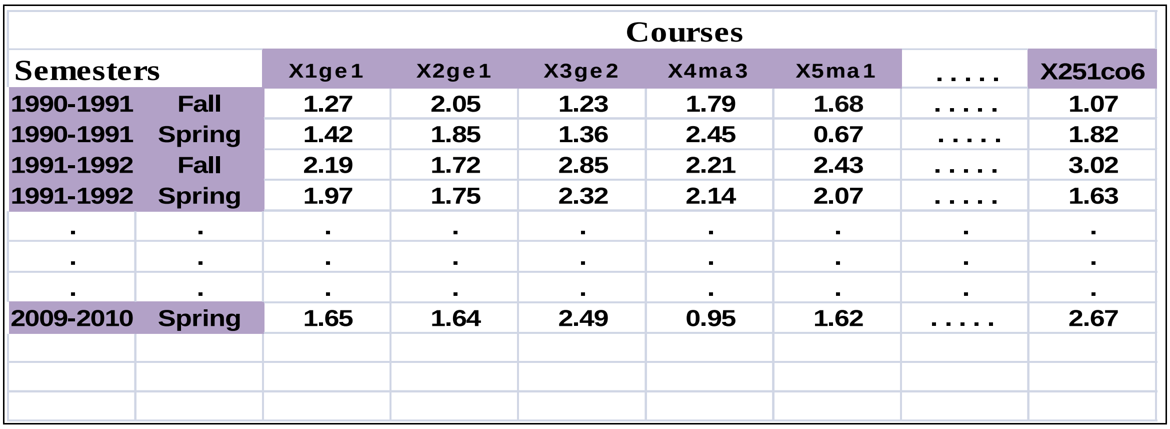
2.3 Handling the database

Here in this section the definition of the underlying database and its nature and the process of filtering it will be introduced and explained.

2.3.1 Description of the Students' GPA's Database:

The database being analyzed in this study is the student GPA database of Science Faculty in Benghazi University. As with any database of the same type, this one contains the information that is related to the departments and their subjects and their grades, also including the fields of heads of departments, specializations, students' markets, students' history, students' nationalities, subjects, years and many others. Three of those fields were utilized in this study, which are students' marks, years and subjects. The software that has been used to construct this database is the Structured Query Language (SQL) server 2005, which is a special program to manage and prepare the databases. That program was adopted by the Registration Administration in the Science Faculty to manage the GPA database.

Table (2.1) : ***Description of the used data matrix.***



		Courses						
Semesters		X1ge1	X2ge1	X3ge2	X4ma3	X5ma1	.....	X251co6
1990-1991	Fall	1.27	2.05	1.23	1.79	1.68	.....	1.07
1990-1991	Spring	1.42	1.85	1.36	2.45	0.67	.....	1.82
1991-1992	Fall	2.19	1.72	2.85	2.21	2.43	.....	3.02
1991-1992	Spring	1.97	1.75	2.32	2.14	2.07	.....	1.63
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
2009-2010	Spring	1.65	1.64	2.49	0.95	1.62	.....	2.67

Table (2.1) shows the way by which the data matrix was constructed. The data matrix's columns contains the courses (variables) and rows contains semesters (cases). the duration of the study was since the fall semester of the academic year (1990/1991) to spring semester of the academic year (2009/2010) such that the number of semesters is 41 semesters. As for the number of variables it was just chosen 251 courses out of 616 courses in the way that they are the most studied courses in the whole duration of the study. The courses which were neglected were not studied that much during the period of the study.

Some variations and considerations of our database were considered to be explained, for instance the number of courses which vary from one department to another, that difference does not matter because it was intended to have as much variables as possible regardless the membership of them. Another issue has appeared that shows the difference in the numbers of students from one course to another in each semester, that is. the sample size of the students for each course in every semester. Since our database was taken from the academic year of the Fall semester (1990/1991) the number of the students started being large and increase in stable way, the first semesters which were neglected the numbers of students in which for each semester were small.

Some other consideration were accounted which consider if the main technique Block Principle Component Analysis is applied in every department individually, this idea holds an ideal application if it is in the context of reaching the causes of the variation for the courses in each department, but the main aim in this study is about data reduction and variable selection, in other words, we look for a huge database consists of as many variable as possible.

2.3.2 Variable coding :

The process of variable coding was easy and interesting at the same time, because we wanted a way by which we can manage and distinguish between different variable**,** every variable has three indices as follows :

1- The First index refers to the serial number of the variable and the count is from 1 to 251.

2- The second index in variables which consists of two letters that refer to the Department’s membership, and the contraction was made as follows :

ge = General department.

ma= Mathematics department.

st= Statistics department.

ph= Physics department.

ch= Chemistry department.

go= Geology department.

bo = Botany department.

zo = Zoology department.

co = Computer department.

3- The third index refers to the semester's course, that is, the index gives an idea about the semester in which that course must be taken or studied.

2.3.3 Data preparation process:

When the GPA database was given to us kindly by the Registration Administration in the Science Faculty, it was in the form of SQL fields which needed to be changed into other forms that suited our needs. The GPA database was transferred from the form of SQL to the Excel spreadsheet form, which is presented as a matrix with columns representing the courses by averaging values taken throughout semesters that represent the cases.

The transfer operation to the underlying database was done by some procedures including query language which was involved in the SQL program. That procedure is used to make links between different fields and utilize them to make new fields or columns in a new form. That operation was done successfully and the desired data matrix form was obtained. Our objective was utilizing those links to calculate the averages of the scores or marks for each course (subject) for each semester. In other words, by taking the average value for the scores for each subject and putting it in a row in the data matrix, each row in the new data matrix represented a specified semester. It was decided that the study would start with semesters from the Fall semester of the academic year 1990/1991 until the Spring semester of the academic year 2009/2010. So the number of cases or rows was 41 which was enough to proceed with our technique.

Therefore , the data preparation started by using the Excel tools and ready functions such as ,"count" to know the number of observations in each row or column ,"min" and "max" to have an idea about the minimum and maximum values in each field as well as some measures of central tendency and dispersion such as mean, median and standard deviation. After having had access to all of those functions' results, some information was discovered and some other characteristics were clear. One piece of information is the missing values; that problem occurred in our database because of the lack of some fields or information due to some programmatic errors resulting from the programming of the original database or occurring during the data transfer operation. Those missing values filled up with the overall mean value. We have calculated the mean to the whole database vertically, i.e. to all variables (courses) and that was used to fill the missing values. The variables whose number of actual values was less than twenty were removed because they cannot be representative. In other words. The majority of them is not available and if they will be filled up by the overall average they would represent the values of the overall mean rather than their missing actual values.

In general, we did not face any other problems that need to be solved by data preprocessing procedures besides the problem of missing, which was solved easily and successfully. The second step was about the naming of the variables (courses); the name of each variable consists of three indices. The first index is about the serial number of the variable. The second index is about the membership of the course through which the variable can be distinguished by which department it belongs to. The third and final index is about deciding the regular time or semester for the course to be studied. The mathematical symbol that was taken is **X** to represent the variables as a recognized one. We have made an organized array of the whole database by naming the variables in such a way in order to distinguish between the courses through many factors.

After that, the database was ready for the statistical analysis or the main analysis. The database was transferred to some statistical software which are S-plus and Statistica. Those two items of software were chosen due to their proficiency and efficiency with statistical analysis and their skills in making such powerful graphs. Before the main analysis was performed, we made some graphs of the prepared database. Those graphs are bar charts that represent the averages of the departments throughout the semesters which show the general behavior of each department for the whole duration of study and give a general idea or notation about each department's performance by the students in figure (2.1).

Another graph was made to get more insight and more depth of understanding about the database, figure (2.2) which is bar chart for the averages for every department by group of years. In other words, the years of study, which are twenty one years, were divided into four parts and each part consists of five years except the last part which has 6 years. The notion of this extra graph is to clarify the behavior of the departments more precisely per a group of years and to know in which years the average of the scores were better or worse than other years.

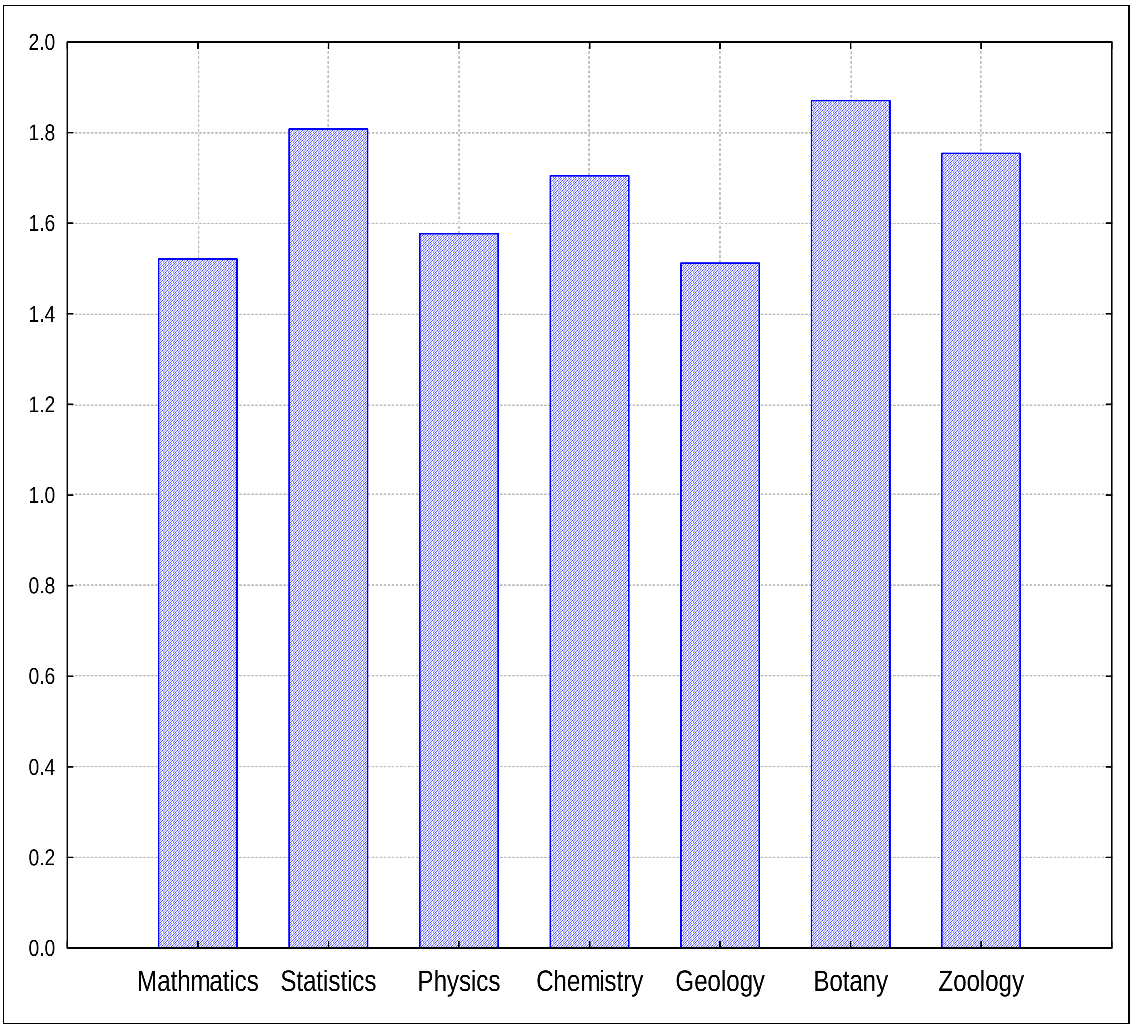


Figure (2.1): Bar chart of the departments' courses through the years of study

In figure (2.1) we can notice that the best performance in average of the students was in Botany department and Statistics department, whereas the worst performance was in Geology and Mathematics departments. on the other hand there is a general low average performance of the students in each department during the period of the study because all of the departments' averages are less than 2.00 which is the half of the general grade average .

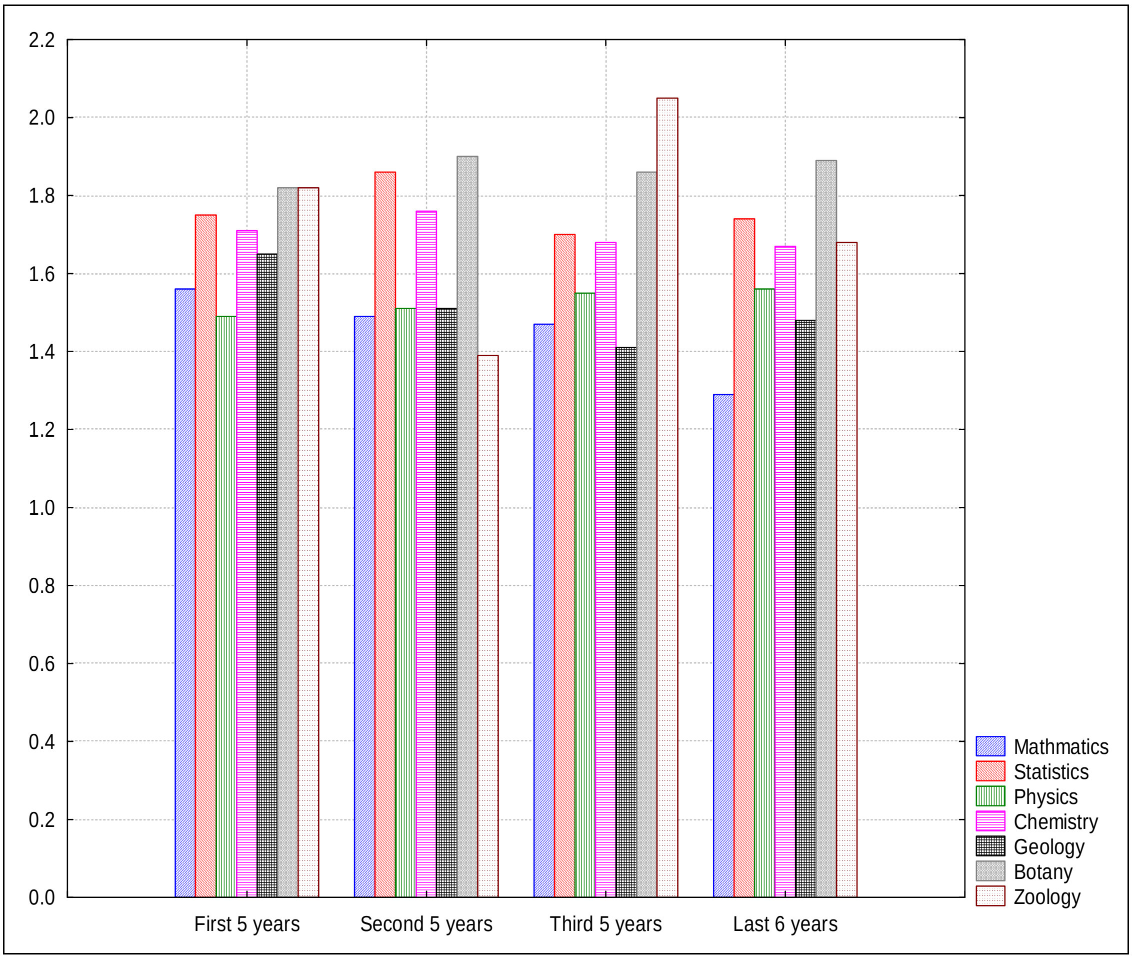


Figure (2.2): ***Bar chart of averages of the departments' courses by groups of years.***

In figure (2.2) we can observe some interesting patterns of the departments' behavior throughout the duration of study which was divided into groups of years, some departments like Mathematics was doing well in the first three groups of years but then the average of grades decreased, Geology department had kind of good performance in the beginning and then went down in the third 5 years and rise up a little bit again in the last 6 years. Physics department has almost a steady behavior throughout the whole duration of study, Zoology department has the highest average among all of the departments and it was in the third 5 years. Many other general characteristics can be detected in the figure (2.2).

Now the database is ready for the main analysis according to our study. The steps and the procedure followed as well as the results will be presented in the fifth chapter.

In the third chapter we are going to introduce the Cluster Analysis technique as an intermediate step in our study. Some definitions and the classification of the clustering methods with brief details are given, as well as a description of some used graphs of this technique.

Chapter 3

Block Principle Components Analysis (BPCA)

3.1 Introduction

In this chapter the two techniques which are merged to constitute what is called Block Principle component analysis are introduced briefly.

3.2 Principle Component Analysis

is one of the most important techniques of multivariate analysis. The significance of this technique lies in its objectives which are:

(1) Data reduction.

(2) Interpretation.

Nowadays, we are very concerned with the first objective, data reduction, for its benefit in statistical research. The data or dimensionality reduction is a very common, desired and critical topic in statistics.

This is due to the fact researchers at the model building phase in any statistical research or study are annoyed by the number of variables to be included or excluded in a model (the curse of dimensionality). As has been seen recently, many statistical studies have huge numbers of variables as a consequence of the evolution of all disciplines of sciences. Some statisticians regard the enormity of the number of variables as a merit or an advantage, as it is desirable to express the factors that affect the underlying study implicitly. On the other hand, the great number of variables would make a model very complex and tedious with regards to getting inferences or interpreting results during the analysis phase.

For instance, if there are two models, one of which is very accurate, but at the same time is complex (i.e. it has many variables) and the other one is not as accurate as the first one but it is a simple model, which model would be preferable or should be chosen?

The increase in accuracy usually raises substantially the costs of data acquisition, computer time and personal time. If a small loss in accuracy is not too critical, and if it lowers costs substantially, the simpler model may be preferable to the more accurate but more complex. This is because an elaborate and complex model may lead to more accurate forecasts or inferences but may be more costly and difficult to implement and also to interpret. Unfortunately, that differentiation may not be available in many cases; that is the researchers find themselves expressing data in one complex model and cannot choose between many models. From this point of view, principle component analysis can be used as a reduction technique in multivariate analysis to reduce and eliminate the unnecessary information (variables) without much lack or loss in the valuable information. Principle component analysis reproduces the system of the data, That is, instead of p variables measured on n observations, it abbreviates the p variables by k linear combinations of the original variables, where $k \leq p$. Therefore, the new variables or the linear combinations are completely uncorrelated. It should be stressed that the lack of association between the components (linear combinations) is of great benefit; i.e. one of the biggest problems that face researchers in the model building phase is the issue of multicollinearity and principle component analysis can be used to get rid of this problem completely. The usage of PCA in this case is not as a reduction technique but it is used as a tool to solve that problem by making a rotation to the original axes (variables) without any loss or elimination of the original information. However, it can be used as a reduction tool at the same time.(Bovas Abraham, Johannes Ledolter, 1983).

So it can be noted that the PCA is not a means to an end in itself but it can be used as an intermediate step in many studies; for example with cluster analysis as is the case in this study, discriminant analysis and regression analysis. Additionally, it can be used to make a rotation to the original axes to make the structure of the data easier to understand by identifying new orthogonal axes because one of the PCA objectives is interpretation. PCA can be applied without the need to the assumption of normality(Johnson, Richard A. & Dean W. Wichern , 2007), but if the study is in the context of statistical inference, i.e. if one wants to conduct any procedure of statistical inference such as estimating some parameter or testing a hypothesis regarding some unknown parameter then the assumption of normality is compulsory. In other cases such as our study where the goal is just data reduction and variable selection, it should not be said that the normality is useless, but it is not necessary. If normality and the system of data (i.e. the matrix of the data **X**) belong to the multivariate normal distribution, after transforming this set of independent univariate normal distributions, each of the components will belong to the normal distribution. This is because any linear combination of a set of variables normally distributed will be normally distributed. And those components will be independent, since under normality, zero correlation implies independence. (Timm, Neil H. 2002).

3.3 Definition and Derivation of Principle Components

In this chapter a brief mathematical presentation will be illustrated regarding the definition and derivation of PC'S. Suppose that **x** is a matrix of p variables, measured on n cases. The variances and covariances or correlations of those p variables are of interest. Suppose too that p is a very large number , in the sense that the variance –covariance structure of **x** is not clear enough due to the multidimensionality. PCA can sort out this dilemma by arranging or transforming the original data set to another appearance or structure that makes it easier to understand and interpret, also to reveal and get good knowledge about the unknown structure of the data system. The transformation process of the original data through these few linear combinations can be done mathematically as follows: Suppose that the variance-covariance matrix of the random vector

x$'$ = $[x_1, x, \dots x_p]$ is $\sum$ and the Eigen values of that matrix are :

$\lambda_1 \geq \lambda_2 \geq, \dots, \geq \lambda_p \geq 0.$

Suppose also that we have the linear combinations

$Y_1 = \alpha_1'\mathbf{x} = \propto_{11} x_1 + \propto_{21} x_2 + \cdots + \propto_{p1} x_p$

$Y_2 = \alpha_2'\mathbf{x} = \propto_{12} x_1 + \propto_{22} x_2 + \cdots + \propto_{p2} x_p$ (3.1)

.

.

.

$Y_p = \alpha_p'\mathbf{x} = \propto_{1p} x_1 + \propto_{2p} x_2 + \cdots + \propto_{pp} x_p$

(Where α_i refers to the Eigen vectors of the variance-covariance matrix)

These linear combinations are uncorrelated so that

$\mathrm{Var}(Y_i) = \alpha_i'\sum\alpha_i \qquad i = 1,2,\ldots.,p$ (3.2)

$Cov(Y_i, Y_k) = \alpha_i'\sum\alpha_k = 0 \qquad i, k = 1,2,\ldots.,p$ (3.3)

The PC'S are simply those linear combinations which are uncorrelated to each other and their variances are as large as possible.

The first PC is the linear combination that has the largest variance. It maximizes $Var(Y_1) = \propto_1' \sum \propto_1$. That variance can be maximized by multiplying the vector $\propto_1$ in a suitable constant magnitude. To solve that problem, it would be convenient to focus on the vectors' coefficients each of which has the unit length. So the next definitions must be introduced as follows:

The first PC is the linear combination $\propto_1' \mathbf{x}$ which maximizes the variance $Var(\propto_1' \mathbf{x})$, so that $\propto_1'\propto_1 = 1$.

The second PC is the linear combination $\propto_2' \mathbf{x}$ which maximizes the variance $Var(\propto_2' \mathbf{x})$, so that $\propto_2'\propto_2 = 1$ and $Cov(\propto_1' \mathbf{x}, \propto_2' \mathbf{x}) = 0$.

Then the step i is:

The ith PC is the linear combination $\propto_i' \mathbf{x}$ which maximizes the variance $Var(\propto_i' \mathbf{x})$, so that $\propto_i' \propto_i = 1$ and $Cov(\propto_i' \mathbf{x}, \propto_k' \mathbf{x}) = 0$, $k < i$.

In short, the coefficients $\propto_1', \propto_2', \dots., \propto_p'$ have been chosen in order to maximize the variances of the PC'S to the total variation further. Those coefficients are subject to the constraint that $\sum_{i=1}^{p} \propto_{ij}^2 = 1$.

3.4 Geometrical View of Principle Component Analysis

The main objective of PCA is to reduce the dimensionality of a dataset and to reproduce the graphical representation of the original dimensions to other new ones by transforming them via a set of linear combinations. In order to understand the relationship between two variables or more, PCA can be conducted by drawing these variables in two dimensions (plane) or in three dimensions (hyper plane). Additionally, there are many methods to represent high dimensional data in two or three dimensions, but more than that is impossible as it is known.

Trying to solve the problem of graphical representation limitation has led to many solutions and techniques; such as principle coordinate analysis, correspondence analysis, and Biplot. In our study, the Biplot will be used as the best suggested technique to demonstrate the nature of the relationship between each of variables, cases, and PC'S at the same time. That mission can be done by displaying n observations and p variables simultaneously in just two dimensions. The plot is identical to a plot with respect to the first two PC'S, but it gives a simultaneous informative representation of the relationship between those two PC'S and the relative positions of the variables and observations.

3.4.1 Plotting Data with Respect to the First Two or Three Principle Components

As usual, the first two or three PC'S are the most important and informative PC'S because they always have the maximum variation of the dataset and plotting them gives a general and good impression for what the data looks like and the patterns of the data can be detected. By plotting the first two or three PC's, some information can be extracted and concluded and the relationships between variables and those components can be demonstrated too. Furthermore, the interpreting of the first three PC'S is less easy than interpreting the first two PC'S. PCA, in small number of dimensions, can easily find any good representation of the dataset if it is really existed, in the sense that the first few dimensions have the best fitting or the maximum amount of variation.

There is an argument about whether the first two or three PC'S are appropriate to represent the majority of the variation in the data set, or if there must be more than three to obtain the most of the variation. In most cases and studies, it has been noted and proved that the two or three first PC'S are sufficient to represent the most of the total variation in the data. The revealing of some range of structures of data sets may be done by plotting PC'S. It is limited by the fact that the PC'S are uncorrelated. Therefore, some structures of some data sets that have outliers or nonlinear relationships between PC'S can be visible, but linear relationships between PC'S are impossible(Jolliffe, I.T, 2010).

3.4.2 Biplots

The difference between the Biplot and the later type of plots is that the Biplots similarly construct plots of *n* observations, but simultaneously they give plots of relative positions of the *p* variables in two dimensions. Furthermore, it provides significant information about the relationship between variables and observations with the related PC'S, which cannot be available in each individual plot. Since the original publication of Biplots, many substantial developments have been made by many authors. It should be noted that the monograph by Gower and Hand (1996) made a considerable extension to the definition of Biplots.(Jolliffe, I.T, 2010)

Biplots have many types and applications; in this study we will focus on the classical type, which is also called ' Principle Component Biplots', which were essentially published by Gabriel (1971, and several subsequent papers). The principle notion had been given by Jolicoeur and Mosimann (1960) as an example of similar diagrams to Biplots. The principle idea of Biplots is very easy, and like all simple solutions to tedious problems, it is both powerful and very useful. Biplots reveal the structure in the data in a methodical way; e.g. correlations between variables or similarities between observations. Biplots can deal with many types of matrices in order to reveal the inherent structures in the data. Geometrically, Biplots is a generalization of the well-known graph *scatter plot* of observations on two variables. Biplots can be defined as a multidimensional scatter plot that expresses the data matrix in terms of its singular value decomposition (SVD).(Gower, J, Sugnet Lubbe & Niel Le Roux, 2010).

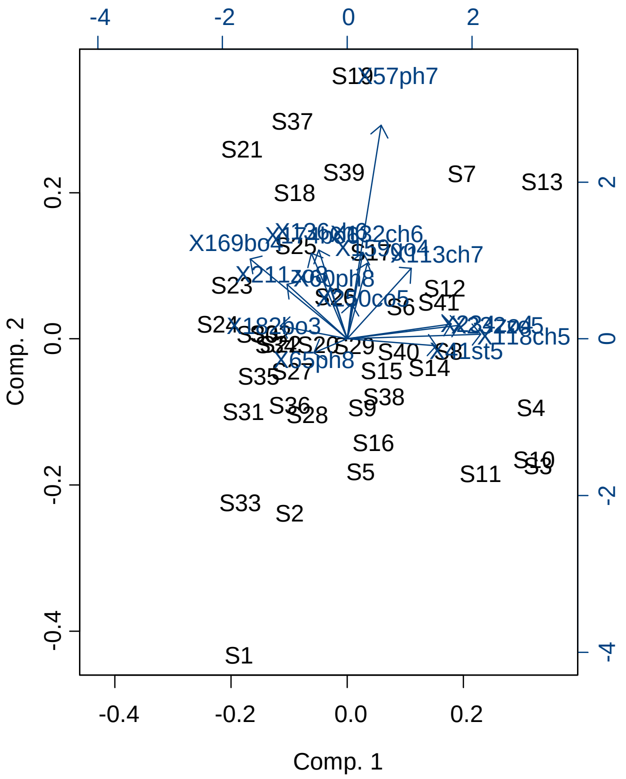


Figure (3.1): ***Biplot of the selected variables from the first block***

figure (3.1) shows an illustrating example of the Biplot representing the variables (x) and the cases (s) giving us a very informative relationships between them as well as the first and second PC'S.

3.5 Properties of Principle Components

Some mathematical and applied properties of PCA are considered in this section. Suppose that the variance-covariance matrix of the random vector $\mathbf{x}' = [x_1, x, \dots x_p]$, assuming that ∑ is a positive definite matrix with distinct roots (eigenvalues) and symmetric matrix, such that

$$\sum = P\Lambda P' \qquad (3.4)$$

Where Λ is a diagonal matrix whose diagonal elements are $\lambda_1, \lambda_2, \dots ., \lambda_p$

And P is an orthogonal matrix of order p whose j th column is the eigenvector corresponding to λ_j. The elements of P are the covariance loadings α_{ij}, which give the contribution of the *j* th variable with the *i*th PC. Assume that the vector of principle components $Y' = (Y_1, Y_2, \dots, Y_p)$ can be written as

$$Y = P'\mathbf{x} \qquad (3.5)$$

The variance-covariance matrix of *Y* is given by

$$Var(Y) = P'\sum P \qquad (3.6)$$

By substituting for ∑ we find

$$Var(Y) = P'(P\Lambda P')P = \Lambda \qquad (3.7)$$

Since $P'P = I$,which shows that the PC'S are uncorrelated and the variances of them are given by λ_j.Another property of PC'S can be demonstrated in terms of ∑

$$tr(\Sigma) = \sum_{j=1}^{p} \sigma_{jj}{}^2 \qquad (3.8)$$

σ_{jj}^2 is known as the variance of the *j*th variable where

$$tr(P\Lambda P') = tr(P'P\Lambda) \qquad (3.9)$$

$$tr(\Lambda) = \sum_{j=1}^{p} \lambda_j \qquad (3.10)$$

That mathematical property means that the total variance of the original variables is equal to the sum of the variances for the all PC'S.

3.6 Principle Components Using Correlation or Covariance Matrices

Principle component analysis can be conducted by either variance-covariance matrix or correlation matrix in order to evaluate eigenvalues and eigenvectors to constitute the PC'S. Many arguments have been considered about the use of covariance or correlation matrices, but the major one about the use of correlation instead of covariance matrices to construct PC'S is that the results of analysis for different sets of random variables are more directly comparable than for analysis based on covariance matrices. The biggest drawback of PCA that is based on covariance matrices is that the PC'S are very sensitive to the units of measurements for the variables. When the variances of the individual variables are widely different that would cause some kind of problem in PCA, which means that the PC'S can be seen as the original variables that are rearranged in decreasing order of the size of their variances. Moreover, the first PC'S account for little of the off-diagonal elements of $\sum$. In most cases such a transformation to PC'S is of little benefit.

It is inappropriate to use the covariance matrix in the case of different types of measurements, except if there is a strong reason in that the units of measurements that are chosen are the only ones that make sense. Even though, in that case, the PC'S provided by means of covariance matrix will not be informative if the variables have widely different variances. Furthermore, the PC'S scores would be difficult to interpret with covariance matrices that have non-commensurable variables. This problem can be solved easily with correlation matrices by standardizing the variables. Another disadvantage of the use of covariance matrices is that it is more difficult than correlation matrices with regards to comparing the results from different analysis.(Jolliffe, I.T, 2010).

Generally speaking, covariance matrices do have just one advantage over correlation matrices; that advantage can be seen particularly in a special case. In the context of statistical inference regarding population PC'S based on sample PC'S, it is easier for covariance matrices than for correlation matrices. Which means it is more difficult to base PCA on correlation matrices rather than covariance matrices in statistical inference. In the main, PCA is considered as a descriptive tool rather than an inferential tool. Thus, that potential advantage of PCA based on covariance matrices is irrelevant. Another advantage of covariance matrices is when all elements of X have the same units; then covariance matrices can be used instead of correlation matrices. Doubtless, the analyst would get different results from using either covariance or correlation matrices. And that is because of their different nature which is expressed by the eigenvalues and eigenvectors.

The drawbacks of using correlation matrix are few; one of them is that correlation matrix PC'S give coefficients (eigenvectors) for standardized variables and thus they are more difficult to interpret. Moreover, , it is more difficult to base statistical inferences regarding population PC'S on correlation matrix.

3.7 The Extraction and Interpretation of Principle Components

The extraction or retaining of the optimal number of PC'S is of great interest in many applications of PCA. There are many criteria which can decide approximately how many PC'S to retain. The PC'S that the analyst would like to retain must have as much as of total variation in the data under consideration, in other words. The PC'S that account for most of the variation in the data are those first few selected PC'S .The decision of choosing the PC'S is a very crucial one to decide how many small PC'S to retain without serious loss in information. A varity of rules and ad hoc methods have been proposed to determine an appropriate value of *m* (such as $m \le p$) .

The reduction can be done by using *m* PC'S instead of *p*, but the question is how can the reduction be done after choosing the optimal number of PC'S? That can be done by looking at the contribution of each variable in each PC and then deciding if it should be excluded or included. That contribution of variables in each PC is simply the loadings of the variables which are the correlations between each variable and the corresponding PC. Both procedures of choosing the PC'S and variable selection need experience, skill and common sense on the part of the analyst and ultimately the type of data will impose itself in each method of choosing the optimal number of PC'S and that is why opinions differ from one author to another. The best way is to" let the data speak about itself " and then use skill and common sense, because every data set has its own features and nature and that will contradict with some recommendations about retaining the components. In this section some of the most common and used rules and ad hoc methods will be introduced shortly to clarify the similarities and dissimilarities between them and the situations for using them as well as to present some methods that can be used for some types of data.(Jolliffe, I.T, 2010).

3.7.1 Cumulative Percentage of Total Variation

This is the most famous and used ad hoc method to decide how many PC'S to retain because of its simplicity and efficiency. The way this method works is to select the cumulative percentage of total variation that one desires to use. Many opinions have appeared for selecting the suitable percentage of total variation; Mardia, Kent, and Bibby (1979) have indicated that 90% of total variation would be useful, I.T.Jolliffe(1986) pointed out that the selected percentage of total variation should be between 70% and 90%. There are many opinions and suggestions about that issue, but the most important opinion without forgetting those standard opinions is that of the analyst himself. Because the underlying data will suggest some cut-off point for the ideal percentage of total variation and here is the role of the analyst who can decide that depending on the nature of the underlying data. This method will be adopted in our research for the reason that we are looking for "the most important variation" and reducing a small amount of variation that is not useful.

3.7.2 Size of Variances of Principle Components

It was proposed by Kaiser (1960). This method simply works by excluding those principal components with eigenvalues below the average after calculating the average for the variances of PC'S. For principal components calculated from a correlation matrix, this criterion excludes components with eigenvalues less than 1. This method can be applied also by using geometric mean instead of simple average or arithmetic mean of the eigenvalues with the same considerations as before(Timm, Neil H, 2002).

3.7.3 The Scree Graph (Cattell's Criteria)

This method is a graphical one, and it involves more subjectivity than the two previous methods. It was named for Cattell (1966) and it simply involves looking at a plot of two dimensions of the eigenvalues λ_k against k which is the number of components. In this method the way of deciding the number of components to be retained is by looking for what is called "the Elbow". The Elbow can be seen as a sharp angle which involves some value of k. This value then can be taken as a number of components m. The name of the Scree plot was derived from the similarity of its shape to that of an accumulation of loose rubble at the foot of a mountain slope.(I.T.Jollife). Another method which has been suggested and which the Scree plot is compared to is known as the log-eigenvalue or (LEV) diagram. This method was developed in atmospheric sciences, and can be applied by plotting the logarithm of eigenvalues log(λ_k) rather than λ_k against number of components k .

Mardia, *et al.* point out that using Cattell's criterion typically results in too many included components, while Kaiser's criterion typically includes too few. The 90% criterion is often a useful compromise. As was indicated before, those methods are the most common and most used in applications of PCA. And they have the advantage of simplicity in performing and interpretation.

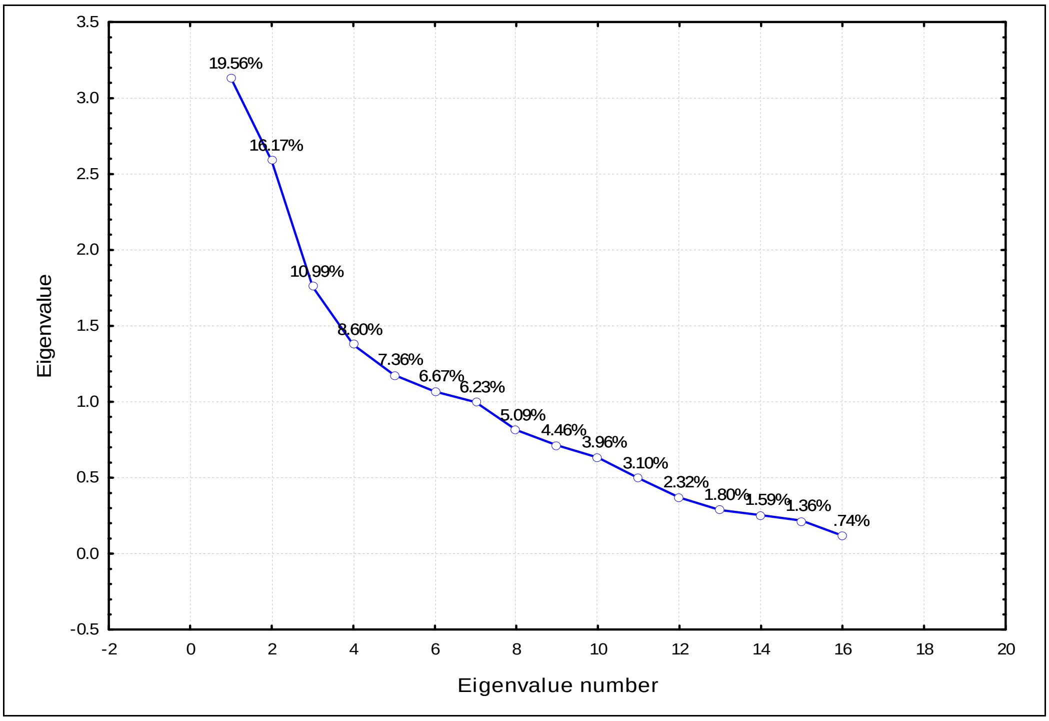


Figure (3.2): ***scree plot of the first block .***

In figure (3.2) the scree plot is illustrated by showing in the horizontal axis the PC'S whereas in the vertical axis the variances of the PC'S, according to Cattell's criteria it cannot be allowed to retain number of components less than two, that is, before the elbow. The subjectivity here can be used in retaining number of components greater than two or when the stability of the steep can be detected. The percentages that are shown at different levels are the percentages of variation that explained by each PC.

3.8 Cluster Analysis :

Cluster analysis is the art of finding groups in data. This technique aims at classifying data points into different groups or clusters, so that members of the same cluster are as similar as possible, while members of different clusters are as dissimilar as possible. More specifically, Cluster analysis is a multivariate statistical procedure that starts with a data set containing information about a sample of entities and attempts to recognize these entities into relatively homogeneous groups. The classification of similar objects into groups is an important and basic human activity for many ages. Cluster analysis can be found in many aspects, and can be applied in many fields of life and disciplines of sciences like mathematics, pattern recognition, computer sciences, biology, astronomy, artificial intelligence and other fields.(Kaufman, Leonard. & Peter J. Rousseeuw, 2005).

The term cluster analysis does not identify a particular statistical method or model, as do discriminant analysis, factor analysis, and regression analysis. You often don’t have to make any assumptions about the underlying distribution of the data. On the whole, Cluster analysis has two types: parametric and non-parametric. Many methods and algorithms have been proposed throughout the recent years of the last century. The choice between them depends on the goal of the study and the nature of the data, beside their accuracy and effectiveness.

Generally speaking, Cluster analysis methods can be classified into two categories (1) partitioning methods (k-clustering) (2) hierarchical methods. The first type assigns each data point to one and exactly one group or cluster, while the later one has some flexibility in the assigning operation. The second type is the one that is relevant to our study and it will be discussed in detail later.

3.9 Proximity Measures

Proximity measures are used to represent the nearness of two objects. If a proximity measure represents similarity, the value of the measure increases as two objects become more similar. Alternatively, if the proximity measure represents dissimilarity (distance) the value of the measure decreases in value as two objects become more alike. The operation of making a proximity matrix is fundamental to the use of Cluster analysis, and the selection of similarity or dissimilarity measure present an interesting problem in clustering. The choice of the proximity measure is based upon the quality of the underlying data. With data having metric properties, a dissimilarity measure can be used, whereas with data having qualitative properties, a similarity measure is appropriate.

3.9.1 Dissimilarity Measures

Assume that a data set have been collected on ***n*** objects or individuals. Each object will be represented by a vector of observations $\mathbf{X}'$

=(x_1,x_2,…..,x_p) on the p variables. For two objects in that p dimensional space, a dissimilarity measure satisfies the following conditions:

1- $d_{rs} \geq 0$ for all objects x_r and x_s .

2- $d_{rs} = 0$ if and only if x_r=x_s .

3- d_{rs}= d_{sr} .

4- $d_{rs} \leq d_{rh}$+d_{hs} .(where d is the distance coefficient, r,s and h are objects "numbers")

Condition (1) implies that the measure is never negative. Condition (2) requires the measure to be zero whenever object **r** equals object **s**. Condition (3) implies that the measure is symmetric. Condition (4) implies the triangle inequality which says essentially that going directly from r to s is shorter than making a detour over object h.

The general case of distance measures is the Minkowski distance which is defined by

$$d_{rs} = \left(\sum_{h=1}^{p} |x_{rh} - x_{sh}|^{r}\right)^{\frac{1}{r}} \qquad (3.11)$$

If we set r=2, then we have the familiar Euclidean distance between objects r and s :

$$d_{rs} = \left(\sum_{h=1}^{p} (x_{rh} - x_{sh})^{2}\right)^{\frac{1}{2}} \qquad (3.12)$$

The Euclidean distance is the most common dissimilarity measure and is suitable for continuous (interval, ratio scale) variables. The Euclidean distance matrix is the most effective for variables that are commensurate.

When variables are not commensurate, one may weigh the squared differences.

If r=1, then we have

$$d_{rs} = \sum_{h=1}^{p} |x_{rh} - x_{sh}| \qquad (3.13)$$

This is referred to as the absolute or City-Block distance (Manhattan distance).

3.9.2 Similarity measures

Similarity measure which are perhaps better known as association coefficients (Matching –type measures) are appropriate when the data is nominally scaled. These types of similarity measures generally take on values in the range 0 to 1, and are based on the reasoning that two individuals should be viewed as being similar to the extent that they share common attributes. Where 0 means that ***r*** and ***s*** are not similar at all and 1 reflects maximal similarity. Values between 0 and 1 indicate various degrees of resemblance.

For two objects in the *p* dimensional space, a similarity measure satisfies the following condition.

1- $0 \leq d_{rs} \leq 1$ for all objects x_r and x_s .

2- $d_{rs} = 1$ if and only if $x_r = x_s$.

3- $d_{rs} = d_{sr}$

Condition (3) implies the measure is symmetric while conditions (1) and (2) ensure that it is always positive and identically one only if objects ***r*** and ***s*** are identical. There are several measures of association which reflect matching type argument. In this chapter, we are not going to write about their mathematical formulas because they are not used.

Many authors have argued about the most common and used association coefficients in application, many books suggested many association coefficients to be the most common, but once again it depends on the experience of the researcher and the nature of data. The proximity measures that have been discussed are the most common and used measures in statistics. From our point of view, they cannot be covered all in this study because there are as many different measures as the number of measures itself. For more details, the reader should look at the references which are related to cluster analysis.

3.10 Types of Clustering Methods

As was indicated earlier, a rather large number of clustering algorithms have been proposed. One of them which is used in this study called hierarchical techniques will be discussed here. They work by clustering the clusters themselves at various levels. Some other clustering techniques have been avoided since they were not used in this study.

3.10.1 Hierarchical Methods

Hierarchical techniques perform successive fusions or divisions of the data so that all the data points are either merged together in one big group or cluster or fused to many single groups or clusters. Hierarchical clustering is one of the most straightforward methods, i.e. .it doesn't require any prior knowledge about the type of the distribution of the underlying data, which frees this method from the assumption of normality in the context of statistical inference. One of the primary features distinguishing hierarchical techniques from other clustering algorithms is that allocation of an object to a cluster is irrevocable; that is, once an object joins a cluster it is never removed and fused with other objects belonging to some other cluster.(Kaufman, Leonard. & Peter J. Rousseeuw, 2005)

In the main, there are broadly two kinds of hierarchical clustering, agglomerative and divisive. The concept of the first type is simply each datum represents a cluster by itself, successively. Those data points start merging on stages, the merging process proceeds until all the data points are merged in just one cluster. Conversely, the divisive algorithm starts with one big cluster containing all the data points, and then that cluster will be disjointed and makes smaller clusters, the fusion process proceeds till every datum represents a single cluster.

As was indicated previously, the first step in Cluster analysis is to work out the proximity matrix. The second step is to choose the appropriate clustering algorithm including the metric used to measure the distances among clusters. There is a little bit confusion between the first type of proximity measures and the latter type. The latter is considered to measure the distance among the clusters (using one of the proximity measures such as Euclidean) after the distances between the original data have been measured.

3.11 Linkage Methods

There is no specific or known criteria which clustering distance must use; many studies suggest that the data will choose the method by itself, in other words. After running all of the clustering distances, there will be just one best method which fits the underlying data. The most used hierarchical clustering distance could be explained briefly as follows:

Firstly, the linkage methods will be introduced briefly with their principle notion; these methods have been very popular throughout the last decades. The principle idea of these methods lies on the notion of distance whether it is the nearest or the furthest or by taking the average of distances, the considerations are different in each, if the calculation will be made via the algorithms by looking at smallest distances between data points or largest distances or by taking their average as a compromise between them.

1- Single linkage clustering (minimum or nearest neighbor method)

This method is the easiest and earliest one which was introduced by Florek et al.(1951) and Sneath (1957). (Kaufman, Leonard. & Peter J. Rousseeuw, 2005)The implementation of this method is to combine objects in clusters using the minimum dissimilarity between clusters; it starts out by first finding those objects having the shortest distance, the process continues until all objects belong to a single cluster. The obvious demerit of this method is that the objects tend to join the clustered objects rather than initiate new clusters; this problem is known as the chaining problem which produces long chains of heterogeneous clusters. The mathematical formula of this method can be defined as follows:

$$d(R,Q) = min_{\substack{i\in R \\ j\in Q}}(i,j) \qquad (3.14)$$

where *R* and *Q* are any two clusters.

2-Complete linkage clustering (maximum or furthest neighbor method)

This method was described by Mcquitty (1960), Sokal and Sneath(1963) and Macnaughton-Smith (1965).(Kaufman, Leonard. & Peter J. Rousseeuw, 2005)This method is definitely the opposite of the Single linkage method. Instead of starting to look for the minimum distances between objects, it works by searching for the largest distance. Complete linkage is slightly affected by noise and outliers and appears to form spherical shaped clusters. When the researcher is looking for compact and tight clusters, in which none of any two data points are far apart, this method would be the best one.

The mathematical formula can be defined as follows:

$$d(R,Q) = max_{\substack{i\in R \\ j\in Q}}(i,j) \qquad (3.15)$$

3- Average linkage clustering

Average linkage clustering works by taking the average distances between all data points so that the average distance between all cases in the new cluster is as small as possible. This method can be considered as the compromise between the Single linkage and Complete linkage clustering.

$$d(R,Q) = \frac{1}{|R||Q|}\sum_{i\in R, j\in Q} d(i,j) \qquad (3.16)$$

4- Centroid linkage clustering

It was introduced by (Sokal and Michener, 1958; Lance and Williams, 1966; Gower, 1967).(Kaufman, Leonard. & Peter J. Rousseeuw, 2005)This method calculates the distance between two clusters by calculating the distance between the centroids of the clusters, the centroids can be the mean or the median of the cluster. It is considered to measure interval scaled data. The demerit of this method is that the distance at which clustered are merged can decrease from one step to the next. This is an undesirable property because clusters combined at later stages are dissimilar to those merged at earlier stages. And its mathematical formula can be defined as follows:

$$\bar{x}_f(R) = \frac{1}{|R|}\sum_{i\in R} x_{if} \qquad (3.17)$$

$$\bar{x}_f(Q) = \frac{1}{|Q|}\sum_{j\in Q} x_{jf} \qquad (3.18)$$

$$d(R,Q) = \|\bar{x}(R) - \bar{x}(Q)\| \qquad (3.19)$$

x_{if} refers to the f th measurement of the i th object x_i, where f ranges from 1 to p.

5- Ward's method

The basic notion of Ward's method is similar to the objective function of K-means clustering, and from the mathematical point of view is similar to Average linkage clustering. The way of calculating the distances is based on the lack of information resulting from the clustering cases into clusters as measured by the total sum of squared deviations of every observation from the mean of the cluster to which it belongs.

$$d^2(R,Q) = \frac{2|R||Q|}{|R| + |Q|} \|\bar{x}(R) - \bar{x}(Q)\|^2 \qquad (3.20)$$

Where $\|\,.\,\|$ refers to the norm.

Hierarchical clustering techniques have the advantage of simplicity in both application and interpretation. That simplicity can be represented by not needing the prior number of clustering. This is a definite advantage over non-hierarchical methods. As was indicated previously hierarchical methods also have another advantage; i.e., they do not require the type of the distribution of the underlying data.

On the other hand, hierarchical techniques have the disadvantage that once a data point located to some cluster, it is impossible to relocate it again to another cluster. Due to that disadvantage, hierarchical methods can be used as an exploratory analysis technique to figure out the features of the data and the possibilities of the cluster solution, in that sense hierarchical, and non-hierarchical methods could be seen as complementary techniques to each other rather than as competing techniques. In other words, hierarchical clustering methods are often called exploratory while non-hierarchical methods are often called confirmatory. In the following section, the algorithms that are used in Hierarchical methods will be referred to, explaining their way of working and their properties and features.(Subhash Sharma, 1996).

3.12 Agglomerative Nesting Clustering (AGNES)

The AGNES algorithm make a hierarchy of clusters, firstly each observation represents a cluster by itself. In the next step the algorithm starts the merging process for the whole single clusters by uniting the two closest observations . In the succeeding steps, a new observation joins any formed cluster, or it forms a new cluster with another observation. At each stage the previous steps are based on (1) Single linkage, (2) Complete linkage, (3) Average linkage, (4) Centroid method, or (5) Ward's method. That operation is repeated until eventually all data points are combined to only one cluster.

3.13 Divisive Analysis (DIANA)

DIANA algorithm starts out with the total pool of the observations in order to split them up into smaller groups or clusters; likewise the AGNES algorithm constructs a hierarchy of clusters, but it starts from the top to down, in other words. Starting with one cluster containing all observations, the fusion process begins at that stage until each cluster contains only one observation. The process begins by splintering out the observation having the largest distance from the other observations. Thus two clusters are formed. The distances are computed at each step; the average distance of each observation in the main to the splinter group, and the average distance of each observation in the main group to the other observations in the main group. The cluster splitting process is repeated until each observation represents a cluster.

Additional comments:

Generally speaking, in both the agglomerative and divisive processes, a tree diagram, or Dendogram, is created as a map of the process. The agglomerative procedures broadly fall into three categories: Linkage, Centroid, and Error variance methods. Among them, only linkage could be used to cluster both items (cases) or variables. The other two methods can be used to cluster only items, but that classification between those methods is not based on theorems, in other words it is based on empirical studies (simulation studies) that suggested those rules for clustering. Virtually , the type or nature of data will choose which method fits it.

Dubes and Jain (1976) have indicated that there is no single best clustering procedure, and some procedures will work better for a particular application than others. Once again the empirical studies that Cunningham and Ogilvie (1972) also Sneath (1966), Milligan and Issac (1980) have conducted, suggested that the average linkage clustering method was the preferred method over the others.

On the whole, all Hierarchical methods are affected by the chaining problem. The observations tend to be located to the formed cluster instead of constituting a new one. This problem usually rises with the single linkage method, but on the other hand it is robust to outliers. Sibson (1973) and others recommended single linkage for large data sets efficiently. Compared to the complete linkage method, the single linkage method is more affected by outliers and noise. The complete linkage method typically constructs very compact and tight clusters of similar cases. Moreover, Ward's method appears to identify compact clusters that are approximately of equal size and shape.

3.14 Geometrical View of Cluster Analysis

In this section we will introduce the graphs that are used by Hierarchical clustering, which are : the Dendogram plot and the Banner plot.

3.14.1 Dendogram plot

The word Dendogram is originally the composition of two Greek words which are: *Dendron* means "tree" and *Gramma* means "drawing". It is a tree diagram that is often used to represent the arrangement of the clusters formed by Hierarchical clustering .The Dendogram is directly represented as a nested list where each component is equivalent to a branch or twig of the tree. Each node of the tree carries some information needed for efficient plotting or cutting as attributes. Each leaf in the Dendogram corresponds to one data point.

The majority of people who use Hierarchical clustering always display their results by means of Dendogram. Many types of Dendograms exist, but the most used type will be illustrated, which can be read vertically from top to bottom or vice versa and horizontally from left to right or vice versa. The way of reading the Dendogram depends on the chosen algorithm either agglomerative or divisive. One of the advantages of the Dendogram is its aesthetical appearance that makes it attractive and easily interpreted.

Despite the simplicity of the Dendogram, understanding how to construct and interpret it is one of the most important aspects of Cluster analysis. In Figure (3.3) which is set to be as a illustrating graph, it can be seen the general shape and features of the Dendogram.

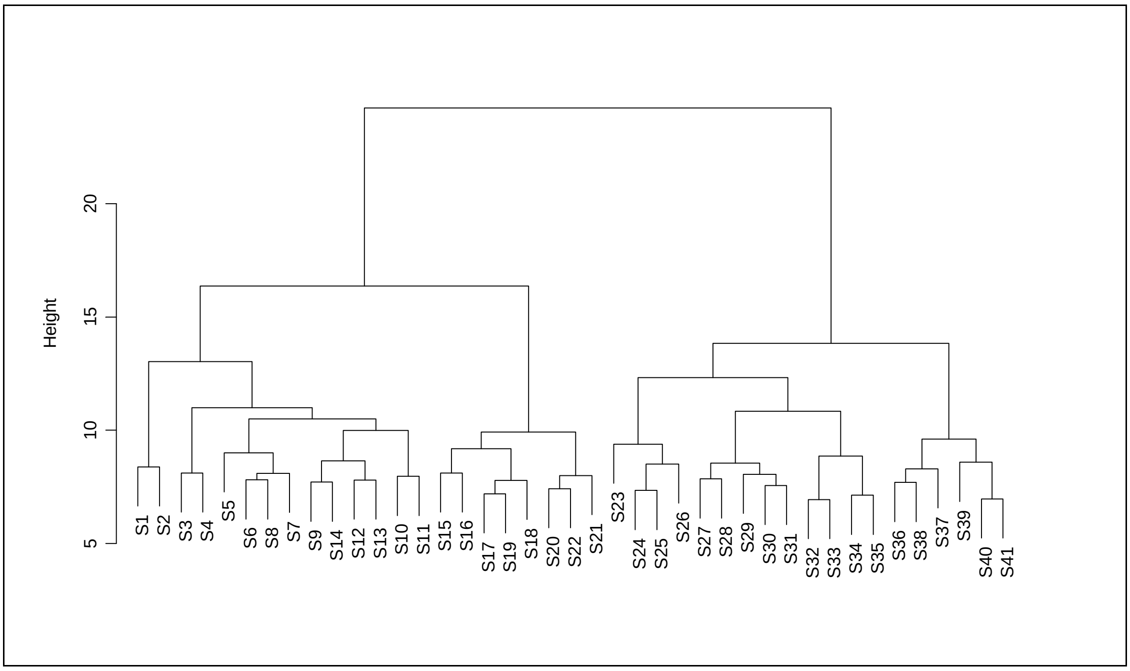


Figure (3.3): ***Dendogram of the semesters for all variables.***

3.14.2 Banner plot

Ward (1963) invented a hierarchy graph that shows the schemes of clustering and demonstrates the quality of the clustering structure. It is simply a horizontal barplot visualizing (agglomerative or divisive) hierarchical clustering. Banner is simple to construct and to understand too, it doesn't require any complicated or sophisticated plotting machinery. This graph can be read either from right to left or from left to right. Again it depends on the type of algorithm that is being used agglomerative or divisive. Johnson (1967) improved the Banner plot. The improved one can be read from top to bottom. The Banner plot is very related to the icicle plot (Kruskal and Landwehr) and the resemblance between them is great. The characters of resemblance are simply represented in some graphical features. On the other hand, there are some differences.(Kaufman, Leonard. & Peter J. Rousseeuw,2005).

Figure (3.4) shows an example of the data set of the selected variables from the first block constructing by the Banner plot and it can be seen that the quality of the clustering structure is very bad via the agglomerative coefficient (AC = 0.3).

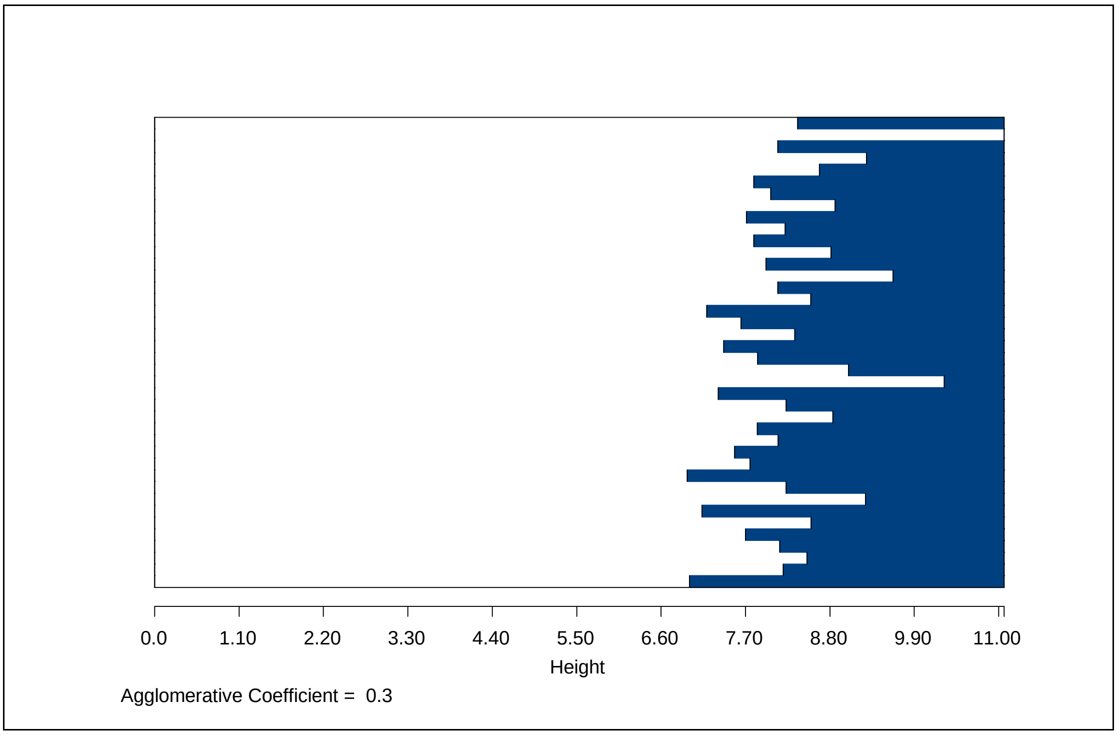


Figure (3.4): ***Banner plot of the selected variables from the first block.***

In next chapter our intention is to introduce the PCA technique and also the Block PCA technique in the same context, explanations and descriptions of those techniques will be given in short details by focusing on the most important items and issues.

3.15 Block Principle Component Analysis

The ordinary PCA procedure cannot be relied on when the number of variables is numerous. Because the results would be of little benefit and their interpretation would be also meaningless and unreliable. In other words, in PCA, our desire is to either reduce the dimensionality for cases or variables (it should be indicated that the first case is not usual or common in the applications of PCA) without losing much of the potential information in the data, that information can be expressed in terms of variation and as it is known the objective of statistics is to study the variation, so the ordinary PCA will just choose the first few PC'S which reserve or have the most variation in the data; but the question is this: what about the rest of the PC'S? It is unwise to simply get rid of or just delete them without any consideration because they may have some important information (variation) and that variation must be considered. The ordinary PCA cannot do that mission(Liu, Aiyi, Ying Zhang, Edmund Gehan & Robert Clarke, 2002).

An alternative way to keep most of the information in the data without any serious loss is by performing what is called" Block Principle Component Analysis". This is a new technique which can be performed in the context of data reduction and variable selection to extract information form date sets whose number of variables are large and with relatively small number of cases. This technique has many advantages such as its computational simplicity in application as it is derived from a mathematical theorem. It does not need any simulation or any empirical study for any type of data. Which guaranties its validity and adequacy for any type of data, as well as results which can be relied on with much more confidence than the ordinary PCA.

In short PCA is an exploratory multivariate technique for data reduction that transforms a set of p correlated variables into set of m uncorrelated variables (linear combinations). However, PCA with an extremely huge number of variables (for example 500 or more) cannot be efficient and the computations can be really intensive, and that due to the finding the eigenvalues and eigenvectors of the original data matrix which would be time consuming and computationally tedious, beside such a huge number of the PC'S is meaningless to the analyst since the high dimensionality would make it hard to extract any useful information to interpret the PC'S. In dealing with such high dimensional data, it is better to perform PCA in stratified way instead of the direct PCA. Firstly, the original variables are grouped into several groups (blocks) in the sense that each block contains variables that are homogenous, so that variables in the same block are more correlated than variables in different block. After that, PCA can be performed in each block and then extracts the most important information (variation) from each block which can be summarized in the first few PC'S in each block. Next the final step can be done easily by merging the whole PC'S that were chosen from each block to form a new data set in order to perform the last PCA to get the most variation concentrated PC'S. In the previous step, the variables with high coefficients would be retained and selected, because those high coefficients reflect their importance to the corresponding retained PC'S. The operation of classifying the variables into homogenous groups can be done by Hierarchical cluster analysis which was explained in the previous chapter in detail.

The application of our suggested technique on GPA'S student database with brief presentation of the obtained results are introduced in the next chapter. The fifth chapter does not have all the results of the study, it is just has some of them and the rest is presented in the appendices.

Chapter 4

Application of Block Principle Component Analysis on Students' GPA's Data

4.1 Introduction

In this chapter we will introduce the results that we obtained throughout the whole study including the final result and conclusion. Starting from the first phase which explains the data clustering, we will illustrate the chosen algorithm (method) and the differentiation between several criteria's to get the best solution. Our study does not involve any simulation study. We followed a mathematical theorem with state and prove, hence there was no need for simulation to check out the validity, adequacy and the efficiency of the method being used. In general, the results in this chapter will be represented briefly without too many details, because of the similarity between the obtained results and to avoid the repetition of the results also the enormity of the pages. Some of the results will be introduced in the appendices of the thesis.

4.2 Clustering Data Variables (Courses)

After the database was prepared, it became ready for the first step of the main analysis which is the clustering. The clustering methods and algorithms were introduced in the third chapter generally and briefly. This is due to the fact that the cluster analysis is just used here as a supportive or helpful tool and not as a basic method. The method that we adopted is the Hierarchical method as the original technique involved. The algorithm type that was selected for the mission of clustering is the agglomerative and that selection was after trying both the agglomerative and the divisive algorithms. The agglomerative gave us better a clustering structure according to both the Banner plot and the agglomerative coefficient (AC) which express the clustering quality. The second selection was about the differentiation of the criterias of between-cluster dissimilarity; five measures were considered and we tried all of them to see which one would give us the best clustering structure. The Ward method was the best one. The way of selection between those measures was according to the agglomerative coefficients(AC) and the Banner plot. The Ward method had the best result among the other methods therefore it was selected for the clustering mission.

Table (4.1): ***The comparison between linkage methods*.**

Linkage Methods	Single	complete	average	ward	weighted
AC	0.5156731	0.8218335	0.702347	0.9213602	0.7550906

It can be seen from the table (4.1) that the Ward method has the highest (AC). Consequently, it was chosen to be the essential method for the main Cluster Analysis among all the other examined methods.

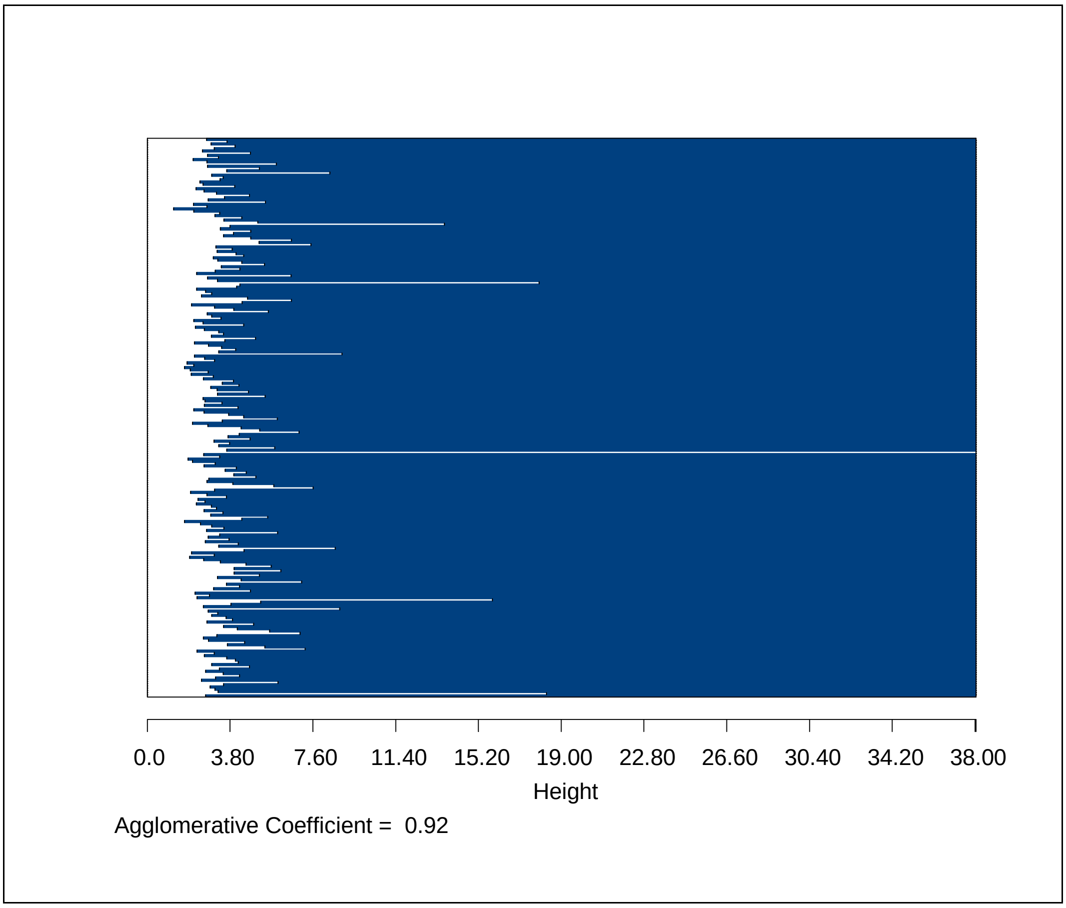


Figure (4.1): ***Banner plot for the Ward method*.**

The banner plot in figure (4.1) can be a complementary tool to the (AC) or a visual tool to display the clustering structure that is done by the Ward method.

We adopted the Ward method for the mission of clustering. After that we faced the issue of the optimal number of clusters, but we were subjective; in other words, we examined many numbers of clusters and then the number 10 was chosen. Because the other numbers of clusters gave us inconsistent distribution of the number of the variable in each cluster. That is, some clusters have many variables whereas the others have a small number of variables. The number 10 of clusters gives the most convenient distribution of the number of variables.

Table(4.2): ***Summary of the blocks***

Block	1	2	3	4	5	6	7	8	9	10
Variables	X3ge2	X4ma3	X34st6	X5ma8	X15ma4	X6ma3	X1ge1	X9ma3	X7ma5	X101ch8
	X14ma5	X8ma4	X38st6	X35st8	X18ma4	X13ma4	X2ge1	X11ma6	X44st7	X241bo8
	.	.	.	.	.	.	.	.	.	
	.	.	.	.	.	.	.	.	.	
	X250co5	X186bo4	X237zo5	X251co6	X236zo6	X248co3	X246st7	X249co3	X233zo8	
Number of variables	43	26	38	23	44	32	16	23	4	2

The known methods and statistics to decide the optimal number of clusters were not used. This is because they have the same problem as our data, which is the inconsistency of the number of variables in each cluster, and hence we were subjective in deciding the suitable number of clusters. In table (4.2) there is a summary showing the number of variables in each block; for instance the variables included in each block and the number of variables in each one.

4.3 Steps of Block Principle Component Analysis

After the mission of clustering was done and we had a homogenous clusters (blocks), the second step was to perform the PCA in each block. As it is known the PCA does not require any prior assumptions about the data, which released us from any constraint about our database.

The main technique in this study is PCA which was performed in each block in order to extract the most important variables from each block, which basically means to obtain the variables that contain the most variation in the blocks.

As we had explained previously, PCA has many methods to decide how many PC'S to retain. In this study we chose the method of cumulative percentage of total variation, and that choice was due to our goal to get the variables that have the most variation among the others and equivalently reduce those variables having a little variation and those have the behavior of stability in the context of data reduction which is the other goal of our study.

Table(4.3): ***Applying PCA on the blocks***

Block	1	2	3	4	5	6	7	8	9	10
variables	X41st5	X4ma3	X34st6	X5ma8	X18ma4	X6ma3	X2ge1	X28ma7	X7ma5	X101ch8
	X57ph7	X10ma5	X47st4	X35st8	X19ma5	X13ma4	X21ma2	X40st8	X44st7	X241bo8
	.	.	.	.	.	.	.	.		
	.	.	.	.	.	.	.	.		
	X250co5	X186bo4	X237zo5	X251co6	X200bo3	X244co3	X223zo1	X229zo3		
Number of variables	16	16	15	11	22	14	11	8	2	2

Table (4.3) shows the numbers of the variables in each block after applying PCA.

Table (4.4): ***PCA performed on the first block*.**

```
                              *** Principal Components Analysis ***

Importance of components:
                         Comp. 1   Comp. 2   Comp. 3    Comp. 4    Comp. 5    Comp. 6    Comp. 7    Comp. 8
    Standard deviation 1.0023673 0.9133875 0.7773251 0.64550041 0.62744385 0.59339868 0.56593396 0.54805256
Proportion of Variance 0.1483343 0.1231680 0.0892058 0.06151497 0.05812159 0.05198534 0.04728455 0.04434372
 Cumulative Proportion 0.1483343 0.2715023 0.3607081 0.42222303 0.48034462 0.53232996 0.57961451 0.62395823

                          Comp. 9  Comp. 10   Comp. 11   Comp. 12   Comp. 13  Comp. 14   Comp. 15   Comp. 16
    Standard deviation 0.52899791 0.5085695 0.46707166 0.44269266 0.42227447 0.3964279 0.38285643 0.35811087
Proportion of Variance 0.04131385 0.0381846 0.03220733 0.02893293 0.02632555 0.0232015 0.02164012 0.01893314
 Cumulative Proportion 0.66527208 0.7034567 0.73566402 0.76459695 0.79092249 0.8141240 0.83576411 0.85469726

                         Comp. 17   Comp. 18  Comp. 19  Comp. 20   Comp. 21   Comp. 22    Comp. 23  Comp.24
    Standard deviation 0.34433141 0.32447580 0.3094514 0.2961780 0.28659538 0.27288122 0.2570016850.233834094
Proportion of Variance 0.01750415 0.01554363 0.0141375 0.0129507 0.01212624 0.01099348 0.0097512360.008072414
 Cumulative Proportion 0.87220141 0.88774504 0.9018825 0.9148332 0.92695947 0.93795295 0.9477041850.955776598

                          Comp. 25    Comp. 26    Comp. 27    Comp. 28    Comp. 29    Comp. 30    Comp. 31
    Standard deviation 0.228886922 0.215039979 0.197562848 0.186339002 0.163530110 0.149556893 0.144353351
Proportion of Variance 0.007734455 0.006826941 0.005762332 0.005126197 0.003948055 0.003302179 0.003076391
 Cumulative Proportion 0.963511053 0.970337994 0.976100326 0.981226523 0.985174578 0.988476757 0.991553148

                          Comp. 32    Comp. 33    Comp. 34    Comp. 35     Comp. 36     Comp. 37     Comp. 38
    Standard deviation 0.133331673 0.120706543 0.108288233 0.084850186 0.0487457879 0.0429590042 0.0342074182
Proportion of Variance 0.002624547 0.002151044 0.001731212 0.001062902 0.0003508019 0.0002724559 0.0001727541
 Cumulative Proportion 0.994177695 0.996328740 0.998059952 0.999122854 0.9994736558 0.9997461117 0.9999188658

                            Comp. 39      Comp. 40      Comp. 41     Comp. 42     Comp. 43
    Standard deviation 0.02223381717 7.43094000000 6.60413600000 4.83993400000        0
Proportion of Variance 0.00007298201 8.15220800000 6.439021e-018 3.45833200000        0
 Cumulative Proportion 0.99999184779 1.00000000000 1.00000000000 1.00000000000        1
```

The results in table (4.4) show the applying of PCA on the first block, and those results basically give the importance of PC'S from many points; which are the standard deviation of each PC, the proportion of variance of each PC and the most important which is the cumulative proportion of total variation that is explained by those PC'S and we used it to decide how many PC'S to retain. As was indicated before, the number of PC'S is equal to the number of the original variables. As can be seen in the previous results that number of the variables in the first block is 43.

We have taken the 70% cut off point in the use of the cumulative percentage of total variation method to choose the most effective and variant PC'S. It is obvious from the results above that the tenth component is where we can cut off and choose just ten components. This is due to the fact that the cumulative percentage at the tenth component is 70%. There is a graph that can help or give an idea about the number of components that should be chosen which is scree plot. We have explained it before in the fourth chapter and will here just mention its use and result.

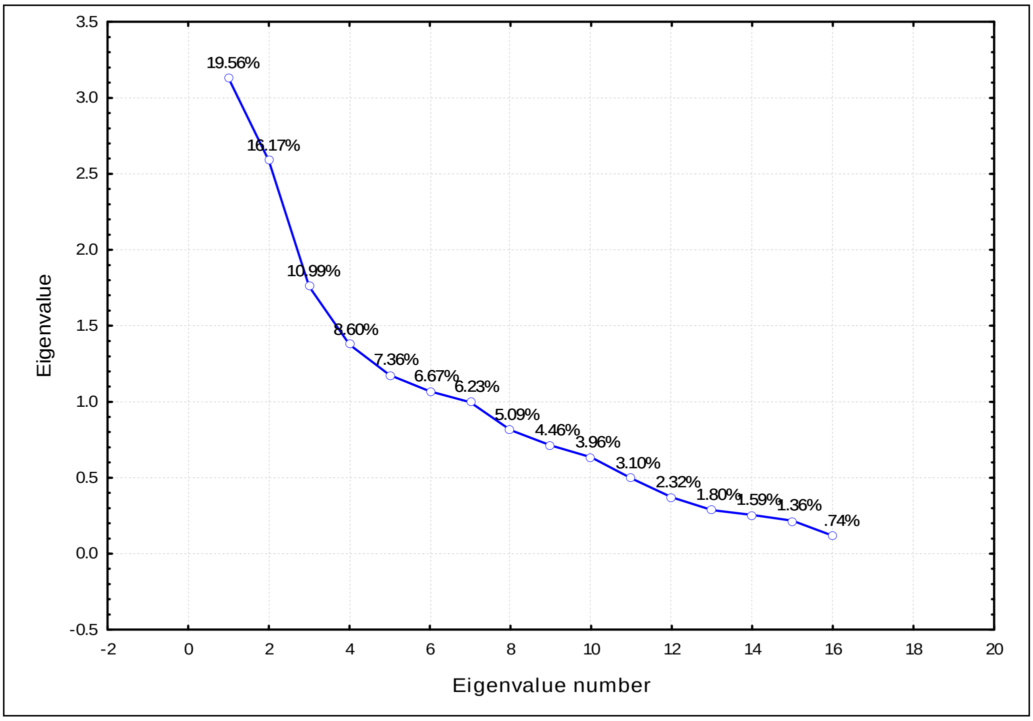


Figure (4.2): ***Scree plot of the first block*.**

It can be noted in figure (4.2) that the "elbow" or the angle is at the fourth component which is to say we cannot retain any number of PC'S before the fourth component and we can subjectively choose any number of PC'S after the fourth component. Once again the later method does not depend on the notion of variation like the method cumulative percentage of total variation. Basically we do not depend completely on screw plot; it is just another opinion we would like to take and know about, the original consideration in this study is about the cumulative percentage of total variation method. Where the percentages that displayed at different levels are the percentages of variation which are explained by each PC individually.

After the first few and most important PC'S have been chosen, it is time to move to the issue of variable selection. In this issue, we relied on the loadings of the PC'S which reflect the importance of each variable in the corresponding PC. That importance can simply be represented by the correlations between each variable and component. It is known that the value of correlation coefficient vary from -1 to +1; that is, those loadings can tell us by means of correlation coefficients the direction and the strength between the variables and the corresponding PC'S.

Table (4.5): ***The loadings of variables and the corresponding PC'S in the first block.***

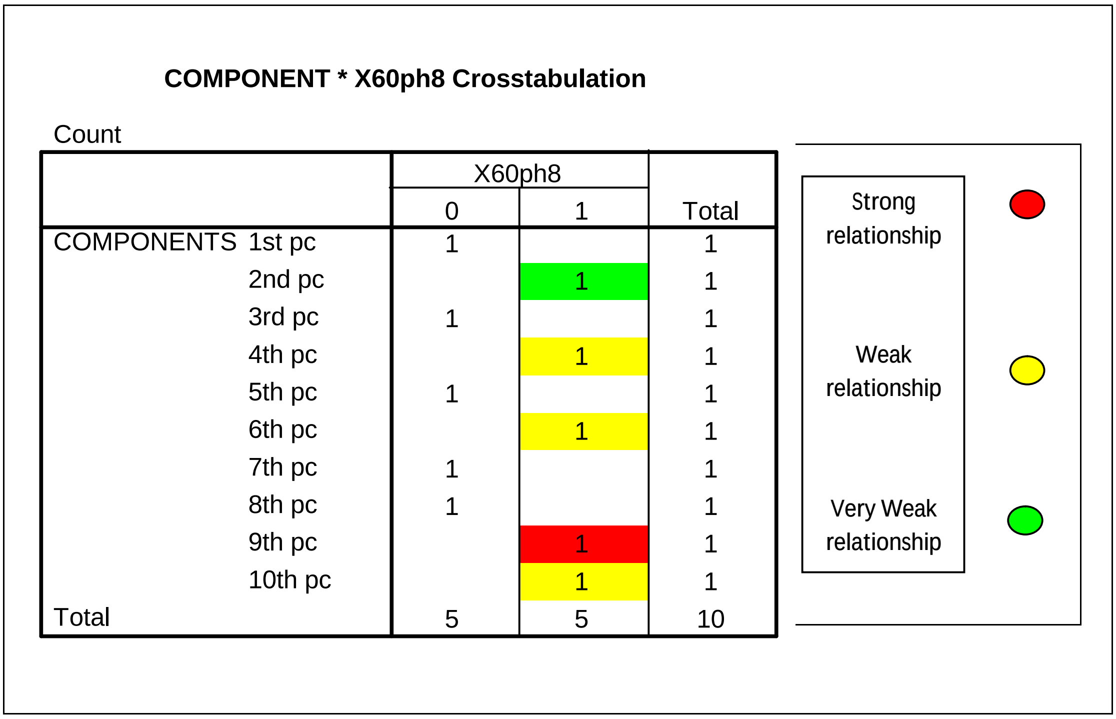
```
Loadings:

        Comp. 1 Comp. 2 Comp. 3 Comp. 4 Comp. 5 Comp. 6 Comp. 7 Comp. 8 Comp. 9 Comp. 10 Comp. 11 Comp.
12
  X3ge2          0.162                  -0.312                                  -0.140    0.172    0.116
 X14ma5          0.113           0.183                          -0.337  -0.109           -0.295
 X41st5  0.299  -0.106           0.219                           0.460                             0.225
 X57ph7  0.154   0.312   0.625  -0.314  -0.300   0.117   0.177  -0.102                            -0.145
 X60ph8          0.201           0.302           0.204                  -0.404   0.204    0.142
 X65ph8          0.212  -0.494  -0.241  -0.178   0.324   0.235   0.124                   -0.147   -0.108
 X98ch6          0.180          -0.117                                           0.158
X102ch4                                  0.174           0.146           0.134           -0.184
X113ch7  0.213   0.149   0.108           0.338   0.143                  -0.160           -0.214    0.147
X115ch5  0.102                   0.107           0.140                          -0.165
X118ch5  0.412          -0.116           0.204  -0.396          -0.114           0.176   -0.170   -0.395
X119ch5  0.220   0.133          -0.175   0.214                                           -0.229    0.198
X131ch6  0.143   0.220  -0.180  -0.109                           0.149           0.119
X132ch6          0.281          -0.104   0.132  -0.103           0.233  -0.216  -0.140             0.263
X136ch6          0.373  -0.161  -0.118  -0.158          -0.252           0.193  -0.131
X137ch6          0.214          -0.117                                           0.117
X144go6                                         -0.245   0.148                           -0.145
X145go6
X151go6                 -0.124                          -0.159  -0.228  -0.172           -0.197    0.199
X154go5                         -0.187   0.106                   0.239  -0.136  -0.117    0.116   -0.138
X159go4                  0.188  -0.162   0.197   0.446  -0.305           0.270  -0.123            -0.152
X169bo4 -0.279   0.207   0.133   0.185                  -0.399   0.182  -0.218                    -0.355
X174bo6          0.225   0.177   0.316          -0.145   0.179   0.107   0.142  -0.146    0.168   -0.108
X177bo7 -0.160   0.223           0.144           0.120          -0.197                             0.159
X182bo3 -0.181           0.107   0.258                  -0.137   0.138   0.235  -0.291   -0.333
X185bo6                                                                         -0.134
X187bo4 -0.142                                  -0.103  -0.194  -0.242           0.141    0.179    0.289
X195bo3                         -0.120  -0.199  -0.255   0.101                  -0.205   -0.379
X198bo7 -0.109                           0.143                  -0.171  -0.179            0.147   -0.359
X210zo7                         -0.114   0.172  -0.107  -0.289          -0.131   0.208
X211zo8 -0.163   0.220                   0.255  -0.204                   0.102  -0.229    0.172
X212zo7                                 -0.105  -0.134  -0.232                           -0.133
X214zo7  0.120          -0.105          -0.150   0.130
X217zo8                          0.106                   0.136          -0.241  -0.215
X218zo7                         -0.232   0.116  -0.186  -0.103           0.297            0.162
X224zo7                                                          0.161  -0.206
X225zo7  0.183                   0.157   0.145   0.127                   0.102           -0.138
X228zo5         -0.149  -0.110  -0.140           0.160  -0.193   0.210  -0.124                    -0.238
X232zo5  0.347          -0.135                                  -0.348  -0.111  -0.562    0.246   -0.106
X234zo4  0.343                          -0.425          -0.326                   0.141    0.110
X238zo5                                                 -0.186          -0.104           -0.161
X240zo4          0.235           0.165                   0.166   0.113                   -0.205   -0.142
X250co5          0.248  -0.225   0.253           0.107                   0.332   0.182            -0.142
.
.
.
        Comp. 43

 X14ma5  0.100
 X41st5
 X57ph7
 X60ph8 -0.156
 X65ph8  0.113
 X98ch6 -0.245
X102ch4
X113ch7
.
.
.
.
X232zo5
X234zo4
X238zo5
X240zo4
X250co5 -0.233
```

The gaps in the table (4.6) mean that the relationship between the variables and the components is very weak. Consequently, it was neglected. We organized a crosstabulation between each variable and the selected PC's, to clarify the contributions of the variables in the PC'S. In the table (4.5) we can see the role of the variable $\mathbf{X_{60ph8}}$ and its contribution for the PC'S which was selected from the first block. The first column's heading "1" means the existence or the presence of the variable in the corresponding PC; by contrast the second column's heading "0" means the absence of the variable in the corresponding PC. The colors show different degrees of the strength of the relationship between the variable and the PC'S. Here it can

Table (4.6): ***Summary of one of the variables and the chosen PC'S from the first block.***



COMPONENT * X60ph8 Crosstabulation

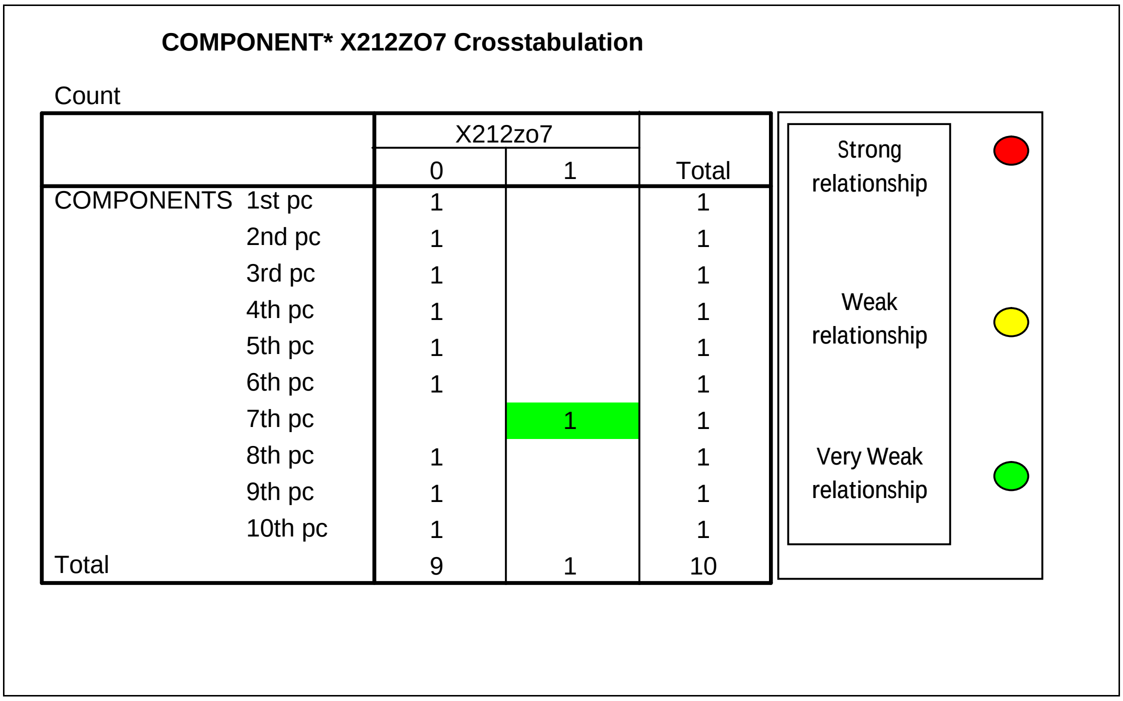
Count

		X60ph8		Total
		0	1	
COMPONENTS	1st pc	1		1
	2nd pc		🟩 1	1
	3rd pc	1		1
	4th pc		🟨 1	1
	5th pc	1		1
	6th pc		🟨 1	1
	7th pc	1		1
	8th pc	1		1
	9th pc		🟥 1	1
	10th pc		🟨 1	1
Total		5	5	10

Strong relationship	🔴
Weak relationship	🟡
Very Weak relationship	🟢

be seen from the table above that the variable has a considerable contribution in five PC'S which deserve to be chosen from the first block.

Table (4.7): ***Summary of one of the variables and the chosen PC'S from the first block.***



COMPONENT* X212ZO7 Crosstabulation

Count

		X212zo7		Total
		0	1	
COMPONENTS	1st pc	1		1
	2nd pc	1		1
	3rd pc	1		1
	4th pc	1		1
	5th pc	1		1
	6th pc	1		1
	7th pc		1	1
	8th pc	1		1
	9th pc	1		1
	10th pc	1		1
Total		9	1	10

Strong relationship	🔴
Weak relationship	🟡
Very Weak relationship	🟢

Table (4.7) shows another example that can be noted easily the weak contribution of the variable $\mathbf{X}_{\mathbf{212Z07}}$ for the corresponding PC'S selected from the first block. That contribution was only with the 7th component and the green color explains the degree of that relationship, i.e. the importance of the variable in this case is slight and due to that fact it would not be wise to retain it or choose it for the next or final step.

We have used the crosstabulation as a helpful summary tool that gave us an arrangement between the selected PC'S and the variables in each block in a concise way. It can comprehensibly clarify and simplify the role of the variables in the retained PC'S from the blocks as well as the type of the relationships between them. The later graph gives a visual representation of the relationships between the variables and the first few selected PC'S from the first block.

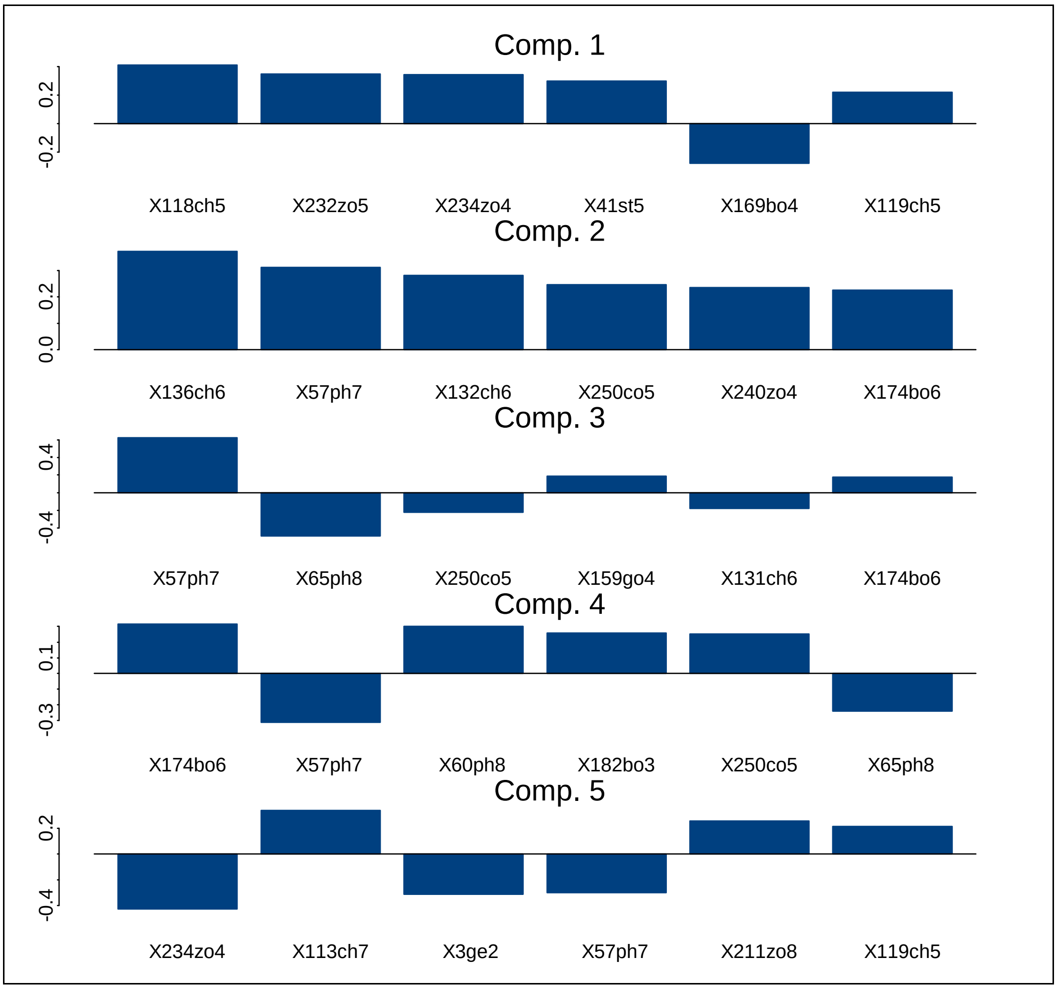


Figure (4.3): ***Visual representation of the variables' loadings with the corresponding PC'S.***

From graph (4.3) we can notice the variety of the bars that represent the loadings of the variables; variables whose direction is upwards mean a positive relationship or positive correlation coefficient and vice versa for the bars directed downwards. The previous graph is limited because it can just show us five components whereas in many cases the retained components are more than five, but the design of that graph includes the first few five PC'S which are usually the most retained PC'S in most applications. The operation of selecting the variables from the blocks was absolutely subjective to us as it only considered the variables that have loadings values starting from 0.200 and ignored any other variables with loadings less than 0.200. This was because we were looking for the best and the strongest relationships within the PC'S in each block. The table (4.8) gives a summary of the selected variables from the first block which are sixteen variables out of 43 variables.

Table (4.8): ***Summary of the Selected Variables from the first block*.**

Variables	Min	Max	Mean	Median	Standard Deviation	C.V
1.Exp.Design 2.	0.82	2.51	1.68	1.68	0.48	28.90
2.Selected topics "Physics".	0.53	3.62	1.84	1.68	0.68	37.06
3.Phys Laboratory 7 .	0.84	2.84	2.00	1.89	0.44	21.75
4.Solid State Phys 2.	0.33	3.00	1.79	1.70	0.58	32.58
5.Phys Chemistry 8.	0.49	2.97	1.73	1.68	0.44	25.15
6.Inorganic Chemistry 8.	0.27	3.03	1.83	1.68	0.59	31.92
7.Bio Chemistry 2.	0.50	2.50	1.71	1.68	0.41	23.82
8.Analytical Chemistry 5.	0.45	2.90	1.82	1.68	0.47	25.78
9.Exploration Geophysics.	0.58	2.73	1.71	1.68	0.47	27.82
10.Soil Microbiology.	0.50	3.32	1.82	1.87	0.52	28.71
11.Physiology of Microbes.	1.11	3.04	2.01	1.93	0.43	21.14
12.Morphological Science.	0.78	3.14	1.84	1.83	0.42	22.92
13.Sci and Comparative Physiology.	0.87	2.80	2.04	2.01	0.41	19.99
14.Animal Eco.	0.59	3.00	1.73	1.68	0.55	31.88
15.Higher Invertebrates.	0.78	2.85	1.72	1.71	0.51	29.88
16.Meth of Numerical Analysis 2.	0.78	2.65	1.64	1.62	0.46	27.82

Some central tendency and dispersion measurements were calculated to get an overview of the characteristics and features of the selected variables.

Table (4.9): ***Summary of the Selected Variables from the final PCA.***

Variables	Min	Max	Mean	Median	Standard Deviation	C.V
1.Selected Topics "Phys".	0.53	3.62	1.84	1.68	0.68	37.06
2.Exploration Geophysics.	0.58	2.73	1.71	1.68	0.47	27.82
3.Geological field work.	0.06	3.08	1.04	0.83	0.71	67.87
4.Bacteriology.	0.18	2.84	1.12	1.09	0.58	51.58
5.Nuclear Physics 1.	0.53	4.00	1.76	1.68	0.70	39.82
6.Molecular Biology.	0.85	3.47	2.02	1.68	0.65	32.17
7.Graduation Project "Math".	1.47	3.78	1.98	1.68	0.60	30.20
8.Mechanics 1.	0.59	2.67	1.53	1.48	0.49	32.21
9.Quantum Mechanics 1	0.33	2.87	1.39	1.44	0.55	39.66
10.Mechanics 2.	0.51	2.57	1.57	1.61	0.52	33.16
11.Inorganic Chemistry 1.	0.25	3.16	1.41	1.48	0.80	57.17
12.Independent study "Math".	0.98	4.00	2.45	2.61	0.71	29.03

The same statistics summary was done for the last selected 12 variables in table (4.9).

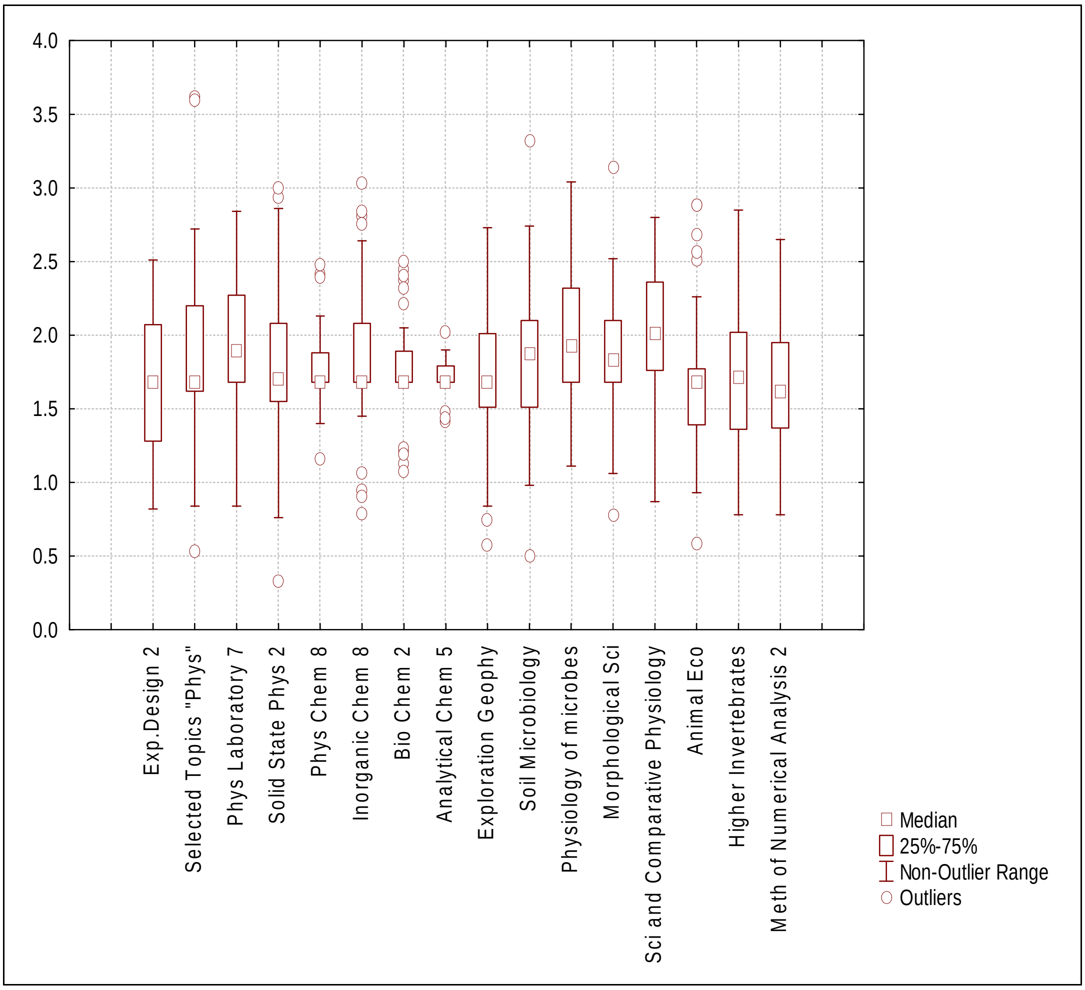


Figure (4.4): ***Box plot of the selected variables from first block*****.**

The box plot graph is of great importance in revealing the behavior of the variation between variables as well as within variables. This graph can be used and utilized to get some measures and characteristics about the data being studied, such as: minimum value, maximum value, quartiles, percentiles and median and can be employed to detect the outliers and extreme values. From graph (4.4), we can observe the distribution of the variation between variables via the medians of different variables; in other words, there is a kind of grouping for the variables that have the same amount of variation in each block or that have the same behavior or distribution of variation.

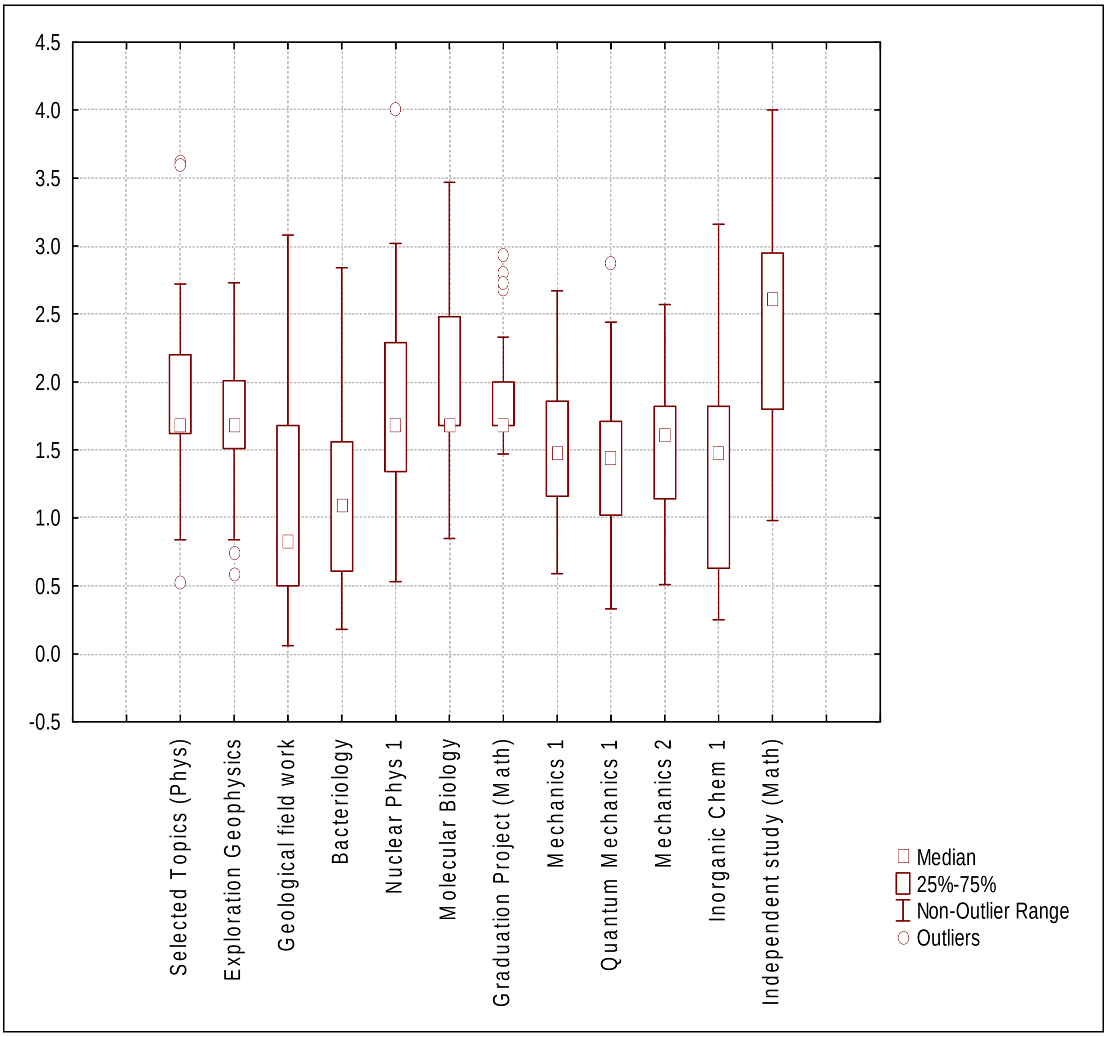


Figure (4.5): ***Box plot of the selected variables from the final analysis*.**

Figure (4.5) was drawn for the last selected variables in the final step of the analysis; there is an obvious pattern of the variation between the variables which can be explored easily by the positions of the medians of each variable. Looking at the graph, we can see that the first two variables have the same pattern of variation due to their medians positions, the same is for the second two variables which are "Geological field work" and "Bacteriology". Similarly, the same can be said of "Nuclear Phys 1","Molecular Biology", "Graduation project (Math)". The variables "Mechanics 1", "quantum Mechanics 1", "Mechanics 2" and "Inorganic Chem 1" have the same amount of variation and they distributed in one homogenous group according to their variation. Finally, the last variable "Independent Study (Math)" which is located alone because of its different pattern that is not similar to any other variable.

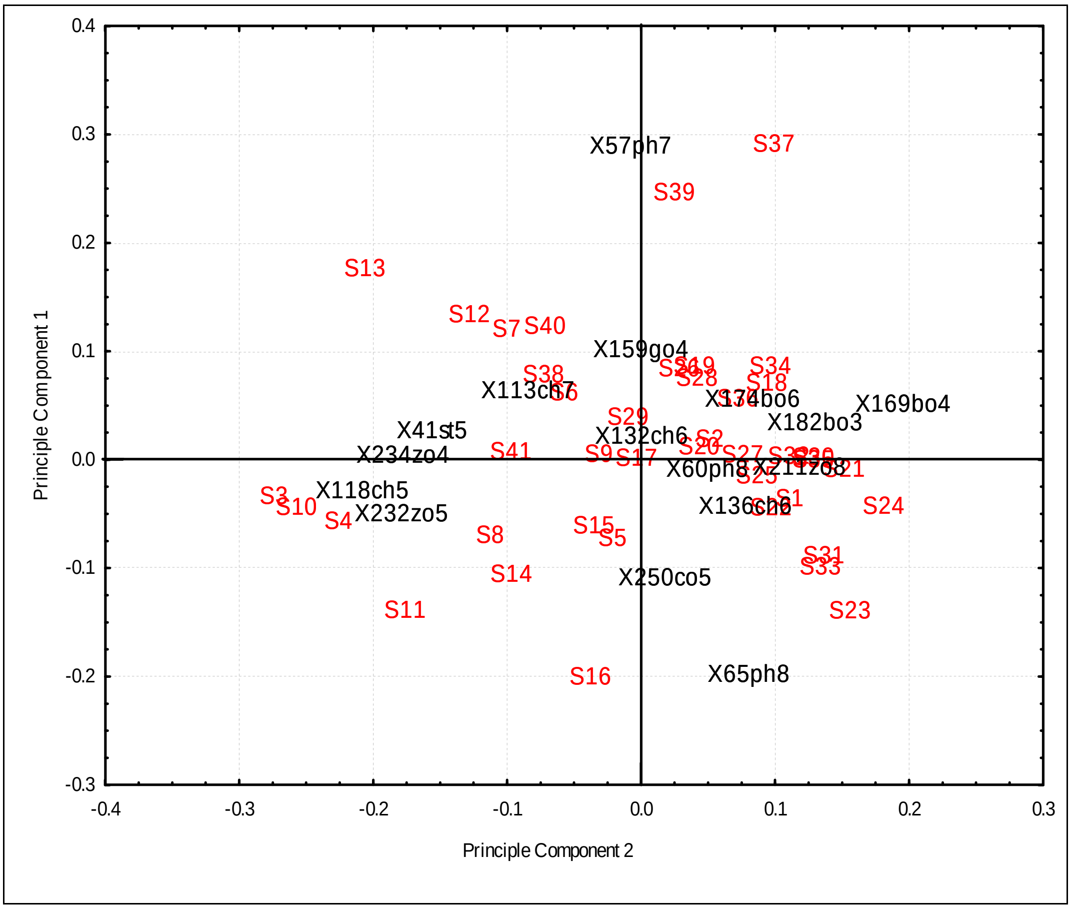


Figure (4.6): ***Biplot of the selected variables from the first block.***

The figure (4.6) is the Biplot that was drawn for the selected variables from the first block. From which we can conclude the relative positions of the variables and the cases and the corresponding drawn two PC'S simultaneously. That is, this graph can show the relative positions between the variables themselves and the variables with cases as well as the variables with the drawn PC'S, and it is the same with the cases.

From the previous Biplot, we can notice many facts about the relationships between the first and the second components and all of the selected variables from the first block as well as the corresponding cases (semesters).The Biplot can be made with any components of interest according to the desire of the researcher, but we have decided to plot the Biplots with the first two components, because they are the most important components. From the third quadrant it can be seen that the variables (courses) $\mathbf{X_{118ch5}}$ and $\mathbf{X_{232zo5}}$ are close together, reflecting their relatively small positive correlation, and the variables $\mathbf{X_{169bo4}}$ and $\mathbf{X_{232zo5}}$ are in opposite quadrants, confirming their fairly large negative correlation. Considering the positions of some cases (semesters) with the variables, looking at the cases $\mathbf{S_3}$ and $\mathbf{S_{10}}$ which are close to each other, gives us an idea about their similar behavior. However, by looking at these two cases that are so close to the variables $\mathbf{X_{118ch5}}$ and $\mathbf{X_{232zo5}}$ and at the same time are far from the position of the variable $\mathbf{X_{169bo4}}$, implies that they should have a higher and stronger relationship with $\mathbf{X_{118ch5}}$ and $\mathbf{X_{232zo5}}$, and a weaker relationship with $\mathbf{X_{169bo4}}$. In other words, the averages of the scores of these two courses are approximately the same in two of these semesters. Conversely, the variable $\mathbf{X_{169bo4}}$ has an average value that is different from the value of the variables $\mathbf{X_{118ch5}}$ and $\mathbf{X_{232zo5}}$ as well as the semesters $\mathbf{S_3}$ and

S_{10}, but it has approximately a case of similar behavior with the close cases and variables in the same quadrant.

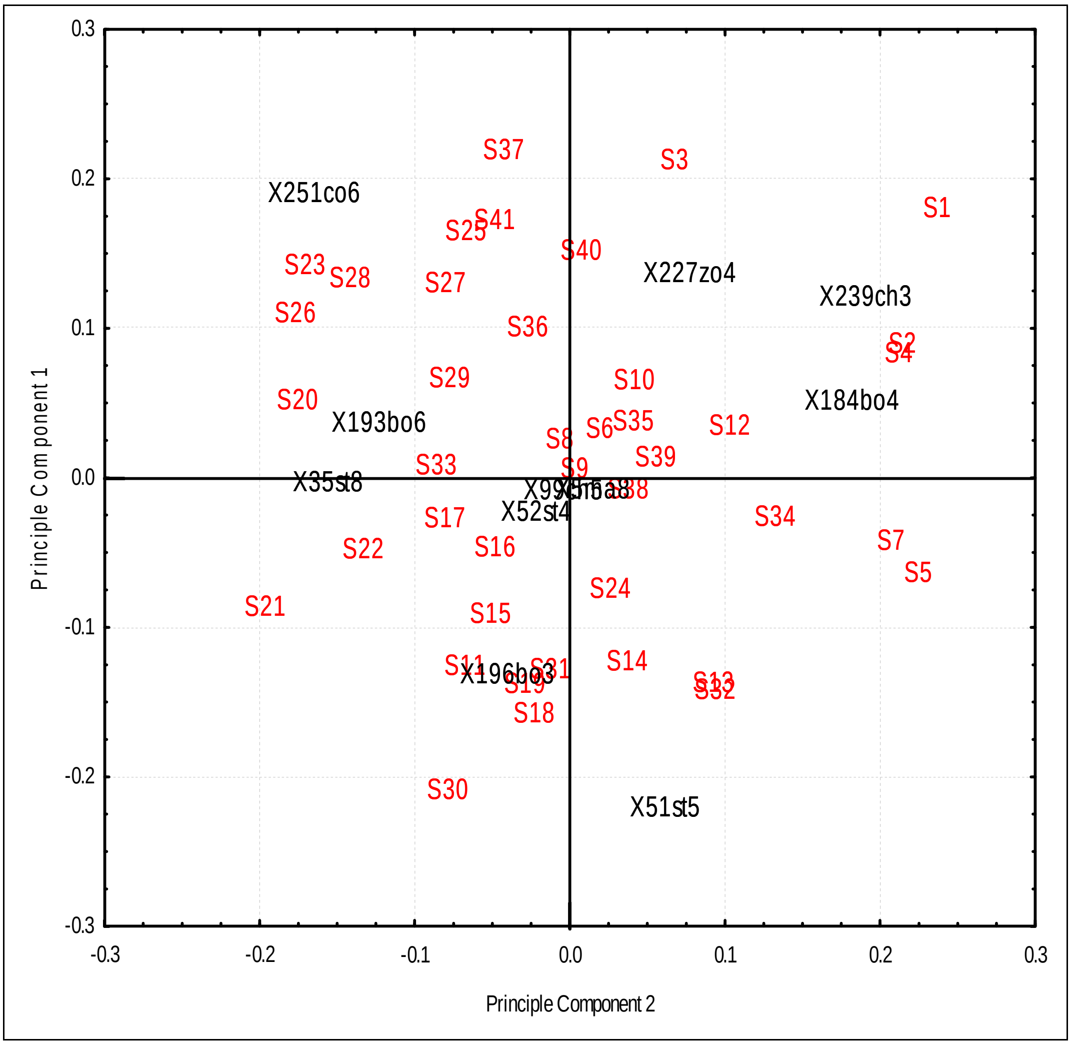


Figure (4.7): ***Biplot of the selected variables from the fourth block.***

Similar considerations can be drawn from figure (4.7), which is made for the fourth block. It is obvious the relationships between the cases (semesters) **S_{11},S_{31},S_{19}** and **S_{18}**in the third quadrant are the same, implying that these four semesters have the same score averages with the variable (course) **X_{169bo3}**. Oppositely, the latter semesters and variables have definitely different behavior or average scores with the cases S_2 and S_4 and with their related variables $\mathbf{X_{239ch3}}$ and $\mathbf{X_{184bo4}}$.

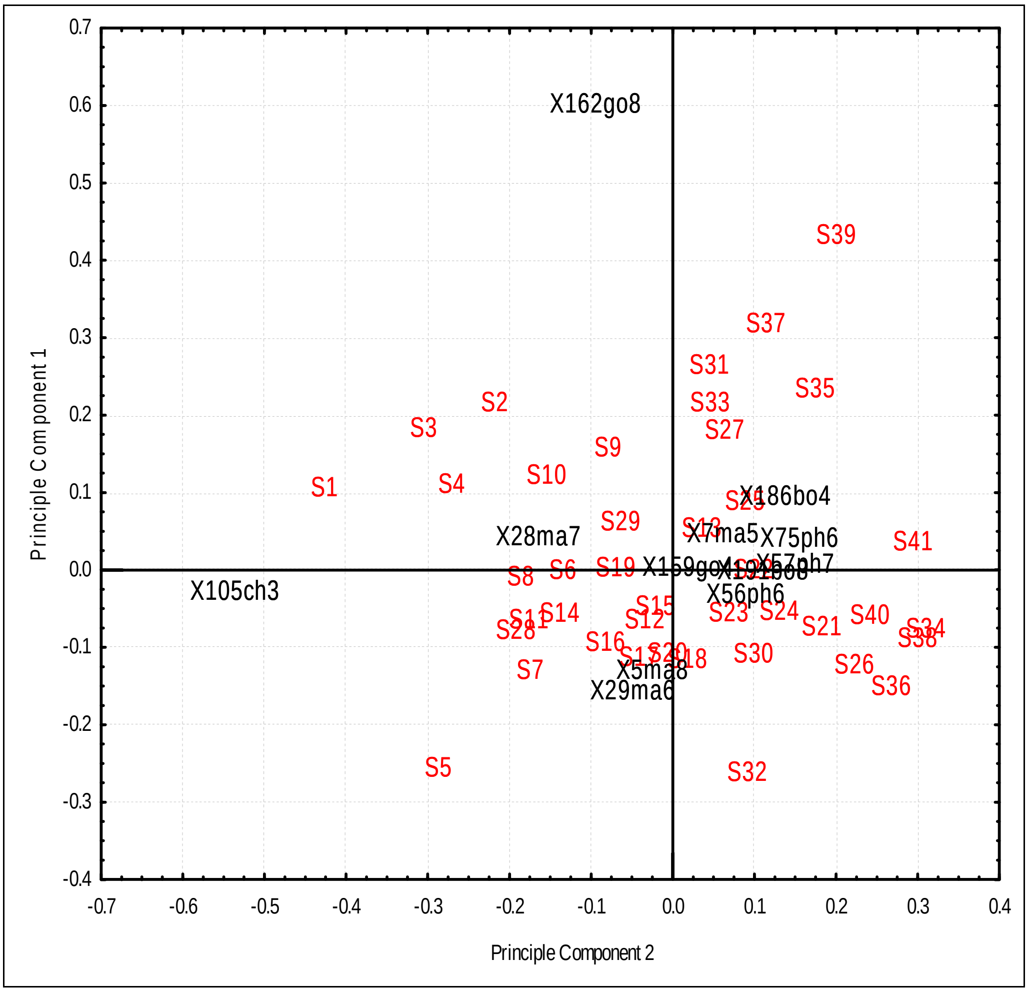


Figure (4.8): ***Biplot of the selected variables from the last PCA.***

This figure (4.8) was made for the best of the best of the variables; the variables that were selected in the final step of the analysis. In this graph, there are some strange behaviors for the variables and cases. As for the variables, it is apparent that the variables $\mathbf{X_{105ch3}}$ at the top of the graph in the second quadrant and $\mathbf{X_{162go8}}$ at left in the third quadrant are outliers, also the cases $\mathbf{S_5}$,$\mathbf{S_{32}}$ at the bottoms of the third and fourth quadrants and $\mathbf{S_{39}}$ at the top of the first quadrant. These outliers' variables and cases have some different and odd behavior compared to the others. For instance, the variable $\mathbf{X_{105ch3}}$ has a large negative correlation with all other variables because it stands alone in an opposite way with the other variables. This means that this course has marks in all semesters which are different from the other variables. As for the point (semester) $\mathbf{S_{39}}$, it has very low average scores with the variables in the second and the third quadrants except the only variable in the top left of the second quadrant which is $\mathbf{X_{162go8}}$. The case $\mathbf{S_5}$ has high average scores with the variables that are close to it which are $\mathbf{X_{105ch3}}$ and $\mathbf{X_{29ma6}}$, but it has low average scores with the other courses that are far from it. In general, the interpretation of these outliers' courses and semesters may be due to the performance of the students which was different from other courses and semesters.

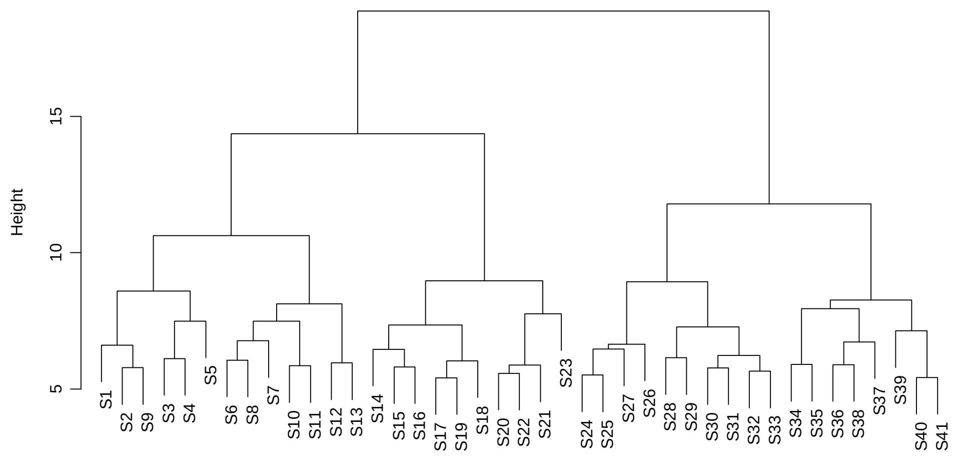


Figure (4.9): ***The Dendogram of the 117 first selected variables from the 10 blocks.***

The last type of graphs is what is called the Dendogram, which is simply a visual simplification of the classification process or the tree that gives all the stages of the classification procedure. It has been taken as a justification tool for the purpose of comparison between the classification of cases (semesters) for the first selected 117 variables in the first step of BPCA and later with the final selected 12 variables in the last step of BPCA, and to make sure that our classification was done correctly.

These two graphs which are (4.9) and (4.10) show that most of the cases belonged to the same clusters. The graphs are not completely the same due to some variations that are not known and considered.

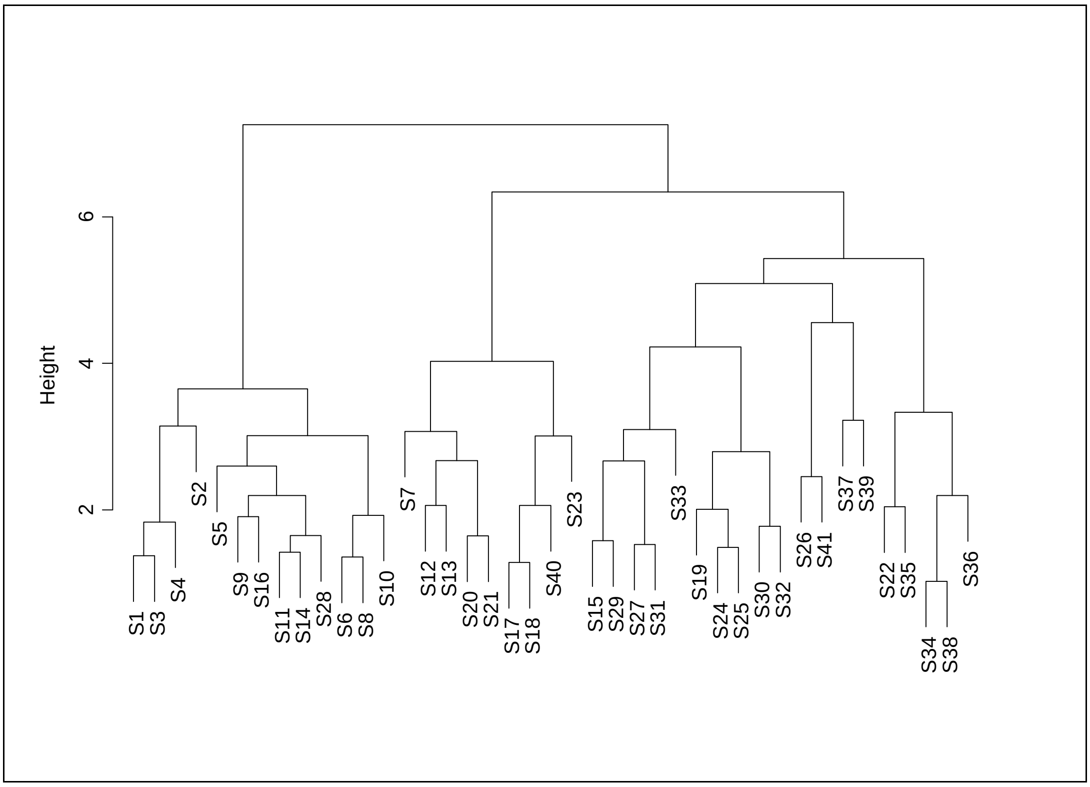


Figure (4.10): ***The Dendogram of the last 12 selected variables from the last analysis.***

4.4 Summary and Conclusion

Applying Block Principle Component Analysis technique consists of merging two techniques, PCA and cluster analysis. After the data preparation procedure was successfully done , the database was ready for the first step of the BPCA, which is clustering the variables on the basis of homogeneity. That is, the variation inside clusters is as small as possible whereas the variation between clusters is as much as possible. Cluster analysis was just used as an intermediate classification tool in order to perform PCA in each homogeneous cluster (Block) and this is the new and original notion of the whole technique, instead of applying PCA directly on a very large number of variables which may result in meaningless and unreliable outcomes and many variables might be neglected due to the structure of the PC'S as only the first few PC'S will be chosen and the rest will be neglected in terms of data reduction. Later on, after the applying of cluster analysis, it was time to perform PCA on each block individually. Then we had the most variant variables from them by using the method of cumulative percentage of total variation that was mentioned earlier and the most relevant method to our study among the other methods of deciding the number of PC'S to retain. Choosing the first few PC'A allowed us to look for the best variables in each block by looking at the loadings between the variables and the corresponding PC'S, we also made the graph that shows the relation between the variables and the corresponding PC'S to have a visual view about those loadings which can simply be considered as correlations coefficients as well as making some graphs like Biplot to extract some important features of those variables, beside making some summary tables of the chosen variables to get their characteristics and some other measurements.

Finally, the last 12 selected variables which are the best of the best of the variables have the most variation among all 251 variables, but some variations must be referred to. The chosen variables have the largest standard deviations among all other variables and that can be seen in the summary tables of the variables. However, some variables have standard deviations larger than the chosen ones which means that there are some hidden reasons and some other variations unknown beside randomness and which may be need a future study that can assign those variations according to some criteria that enables us to reveal those variations.

In the main, these 12 selected variables reflect some interesting points that should considered. In other words, it should be asked what are the reasons of the variations in these variables (courses), and why do the results differ from one semester to another in that way. There must be some reasons that can be either the level of the students' efficiency which may vary throughout the semesters or these courses were taught by different lecturers who are different in their way of teaching or it might be due to the fact that the courses are really difficult to study. Randomness makes it difficult for us to get the whole picture clearly or to see the absolute truth.

4.5 Future Investigations

Some interesting investigations can be done by means of empirical studies and simulation studies to apply some multivariate techniques instead of the techniques used in the original paper that suggested Block PCA technique. Also, trying some new methods which may discover some hidden facts and characteristics. In future work, we may use one of the better- known methods to decide the optimal number of clusters, such as The R-square method or semi-partial R-square method (SPRSq) and Variance Ration Criterion (VRC) or any other method for this objective instead of the subjective way that we used, although it is legal and allowed, but some other considerations about the optimal number of clusters might be useful. In the same context of cluster analysis which was used as classification tool to help to classify the variables into homogeneous blocks, another classification methods may be utilized like classification tress instead of Hierarchical clustering method. Alternatively, there is an intention to use another method of variable selection instead of PCA, to get an idea about the variations that PCA might not consider. Another further study with much details, is to look for the reasons of variations of the selected courses (variables) by using Regression Analysis which needs a response variables (dependent variable) to make such a study. Additionally, if the database can be given with more details, that is by the groups of the courses and the lecturers who have taught those courses throughout the duration of study which is 41 semesters, we can do an in-depth study looking for the most variant courses by means of the groups and the lecturers of the courses that can be seen as variation recourses.

References

1. Dilln, William R. & Mathew Goldsten. (1984). *Multivariate Analysis (Methods and Applications)*.John Wiley & Sons.

2. Everitt, Brian S., Sabine Landau, Morven Leese & Daniel Stahl. (2011). *Cluster Analysis*. John Wiley & Sons.

3. Greenarce , Michael. (2010). *Biplot in practice*. Foundation BBVA.

4. Gower, J., Sugnet Lubbe & Niel Le Roux. (2010). Understanding *Biplots*. John Wiley & Sons, Ltd.

5. Johnson, Richard A. & Dean W. Wichern. (2007). *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall.

6. Jolliffe, I.T. (2010). *Principle Component Analysis*. Springer – Verlag, New York, Inc.

7. Kaufman, Leonard. & Peter J. Rousseeuw. (2005). *Finding Groups in Data: An introduction to cluster analysis*. John Wiley & Sons.

8. Liu, Aiyi, Ying Zhang, Edmund Gehan & Robert Clarke. (2002). *Block principal component analysis with application to gene microarray data classification.*

9. Mardia, K.V., J.T. Kent & J.M. Bibby. (1979). *Multivariate Analysis*. Academic Press.

10. S. Joe Qin, Sergio Valle, and Michael J. Piovoso. (2001) . *On unifying multiblock analysis with application to decentralized process monitoring.*

11. B. Qiu, V. Prinet, E. Perrier and O. Monga . (2003). *Multi-Block PCA Method for Image Change Detection.*

12. Ko Nishino, Shree K. Nayar and Tony Jebara. (2005). *Clustered Blockwise PCA for Representing Visual Data.*

13. Liwei Wang, Xiao Wang, Xuerong Zhang and Jufu Feng. (2005). *The Equivalence of Two-Dimensional PCA to Line-Based PCA*

14.Tabachnick, Barbara G. & Linda S.Fidell. (2007). *Using Multivariate Statistics*. Pearson.

15. Timm, Neil H. (2002). *Applied Multivariate Analysis*. Springer – Verlag, New York, Inc.

16. Micheline Kamber, Jiawei Han.(2006).*Data mining concepts and techniques*.

17. Subhash Sharma .(1996). *Applied Multivariate Techniques*. John Wiley & Sons.

18.BOVAS ABRAHAM , JOHANNES LEDOLTER.(1983).*Statistical Methods for Forecasting.* John Wiley & Sons.

Appendix A1 (Biplots and box plots graphs)

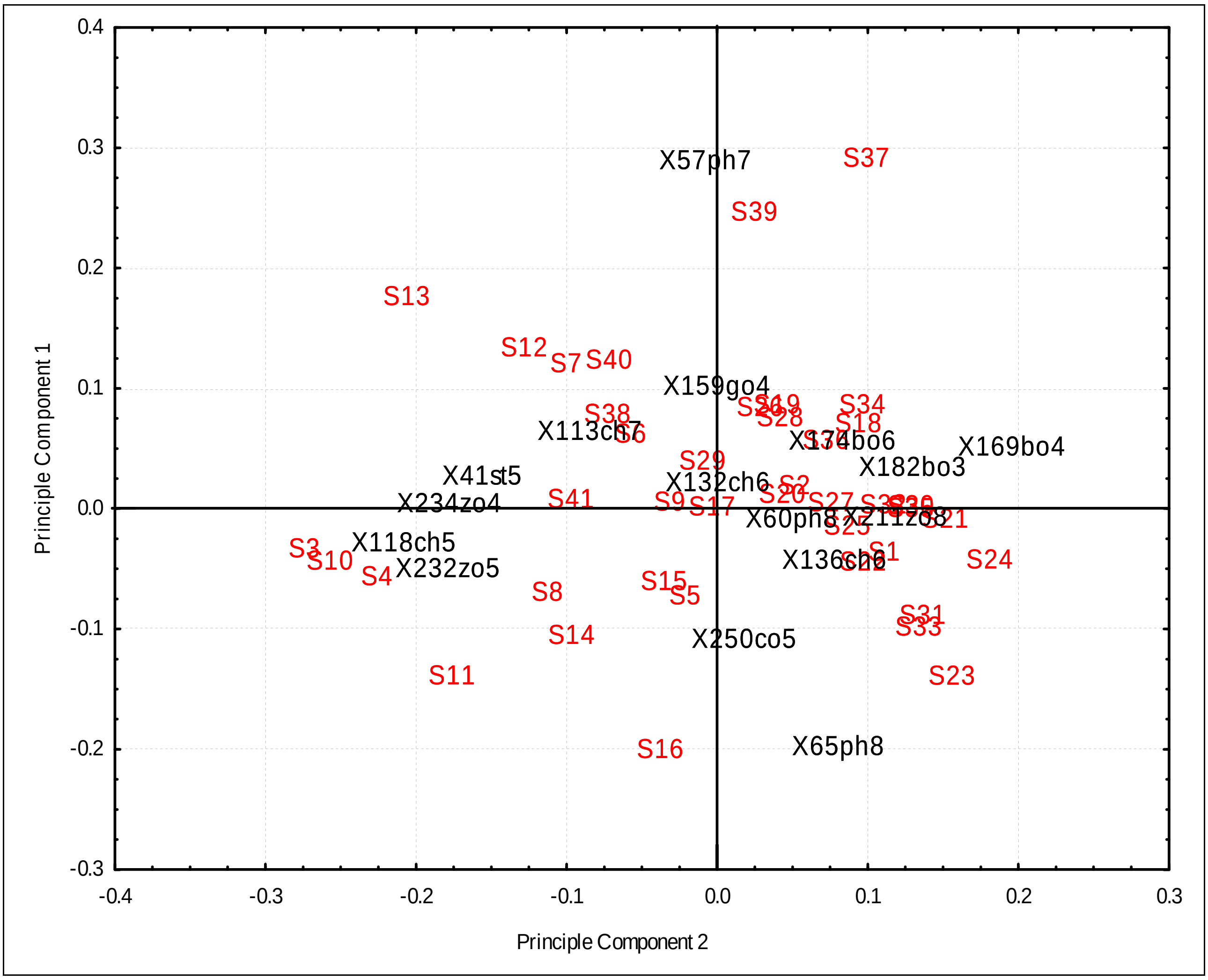


Figure (A1.1): ***Biplot of the selected variables from the first block***

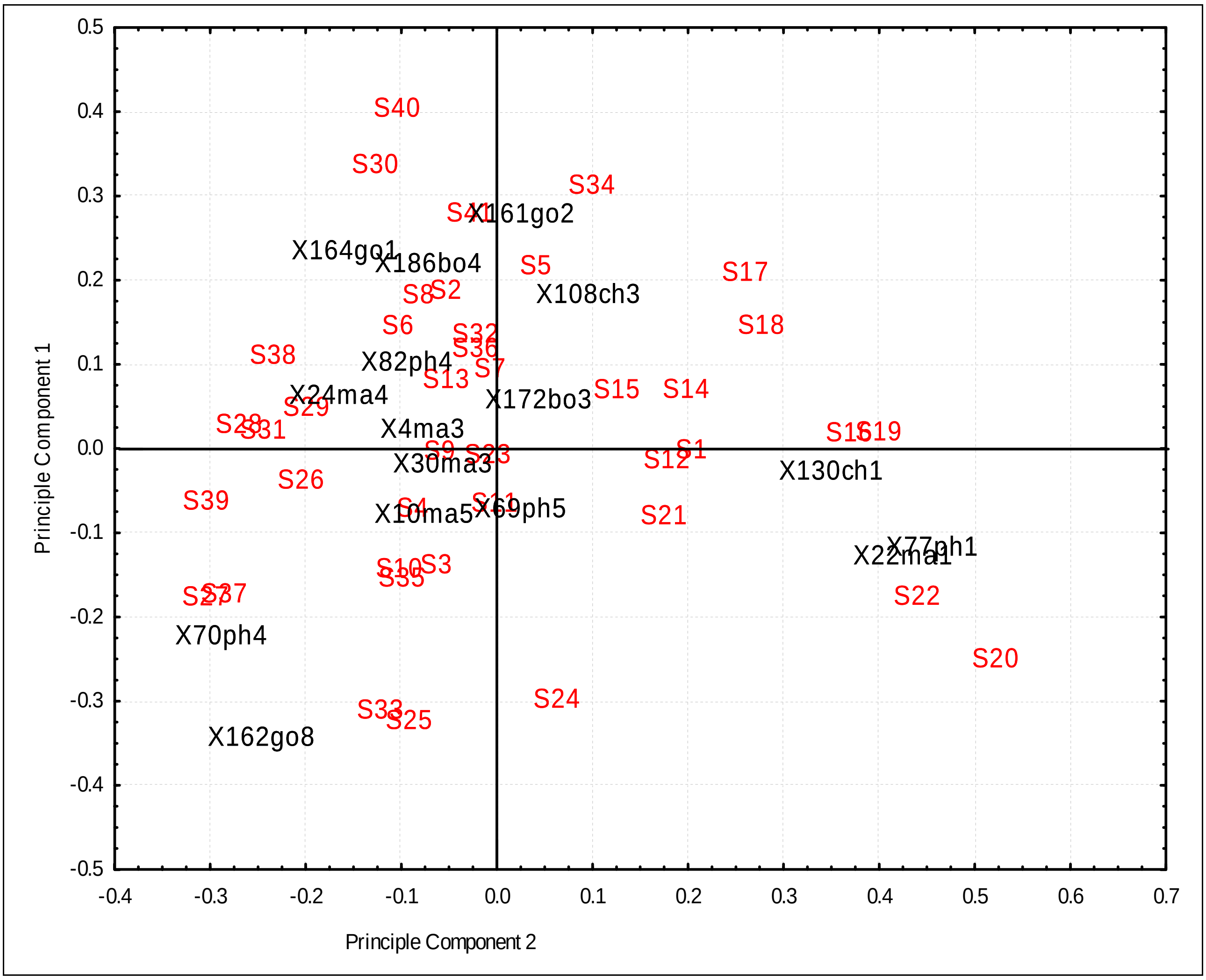


Figure (A1.2): ***Biplot of the selected variables from the second block***

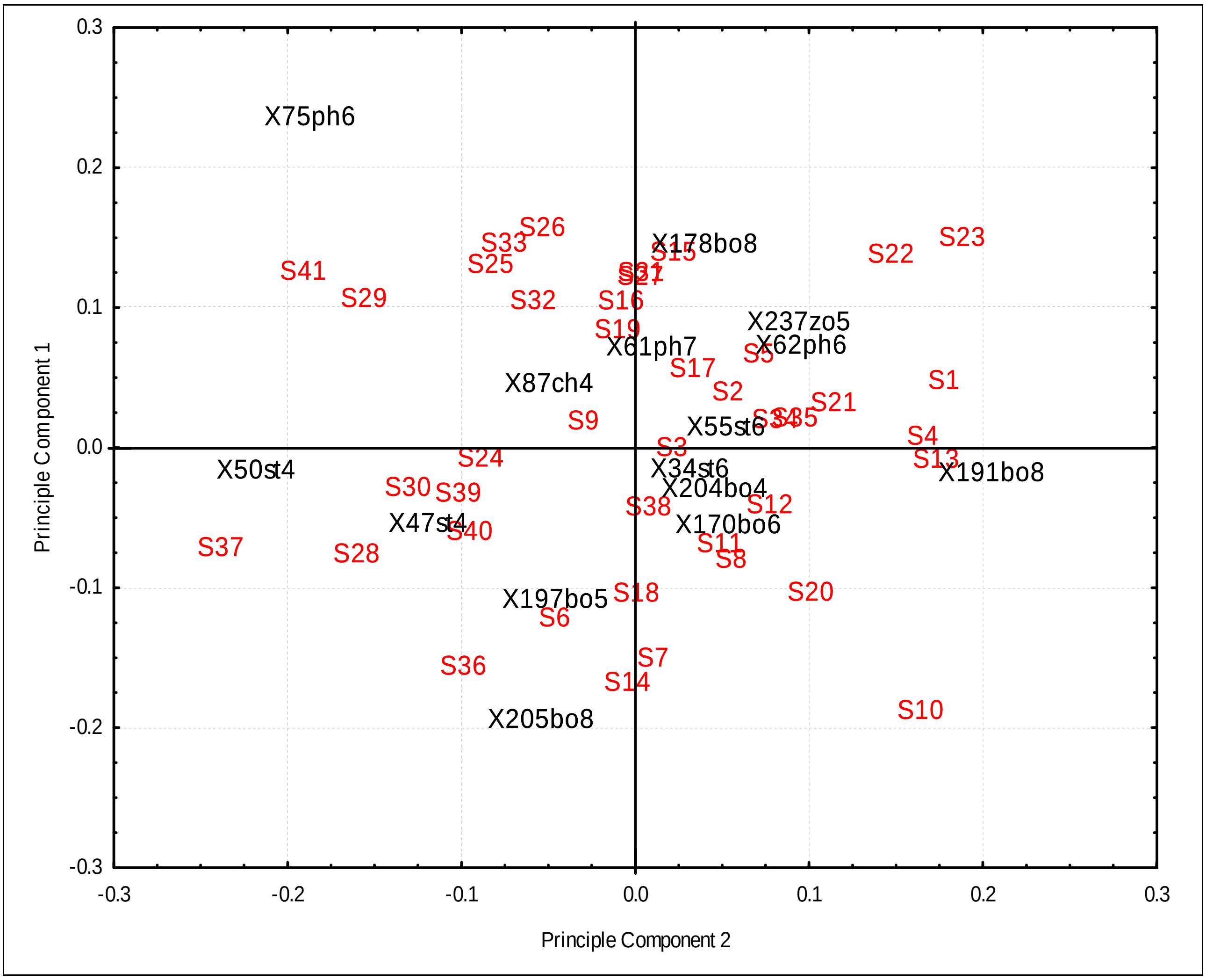


Figure (A1.3): ***Biplot of the selected variables from the third block***

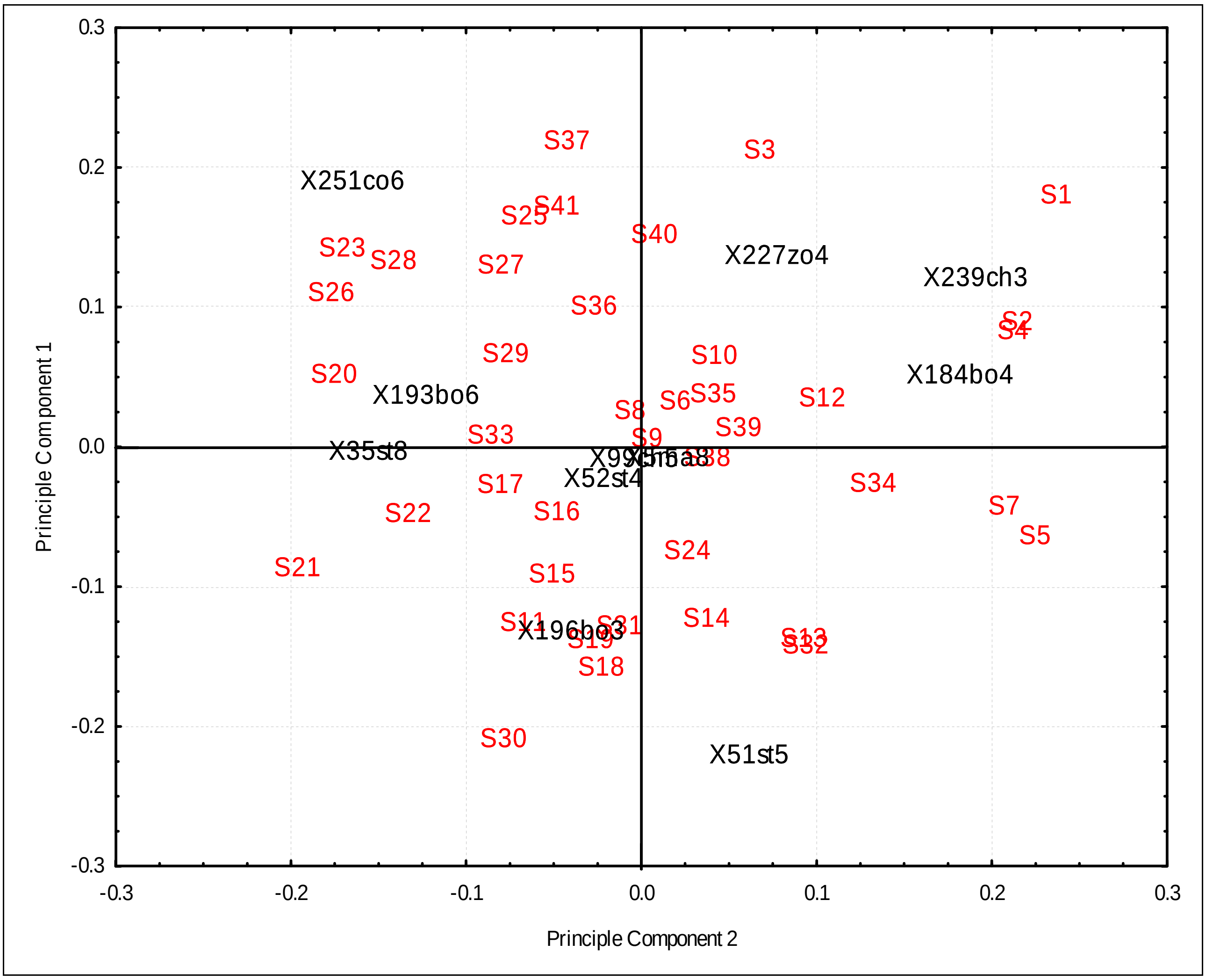


Figure (A1.4): ***Biplot of the selected variables from the fourth block***

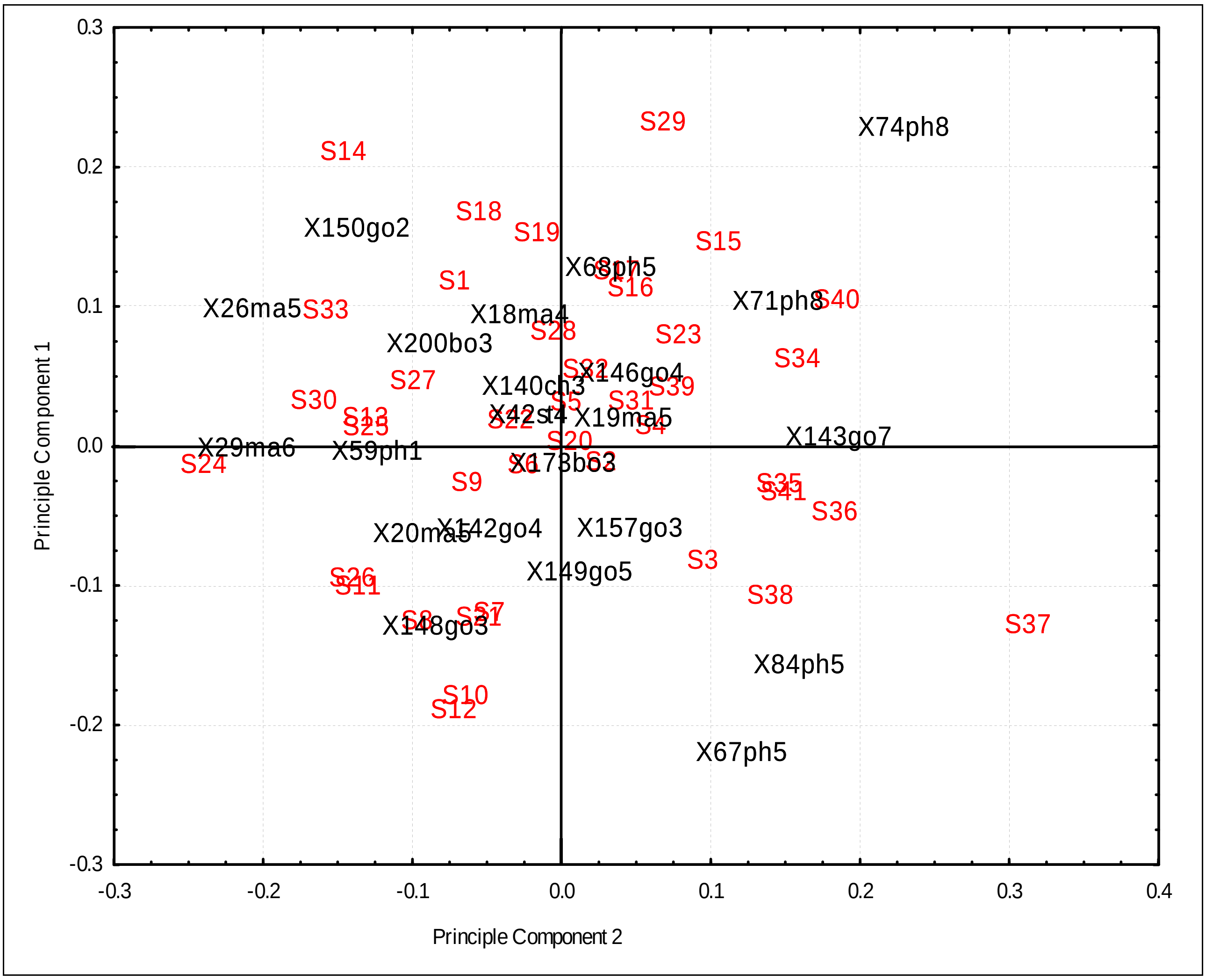


Figure (A1.5): ***Biplot of the selected variables from the fifth block***

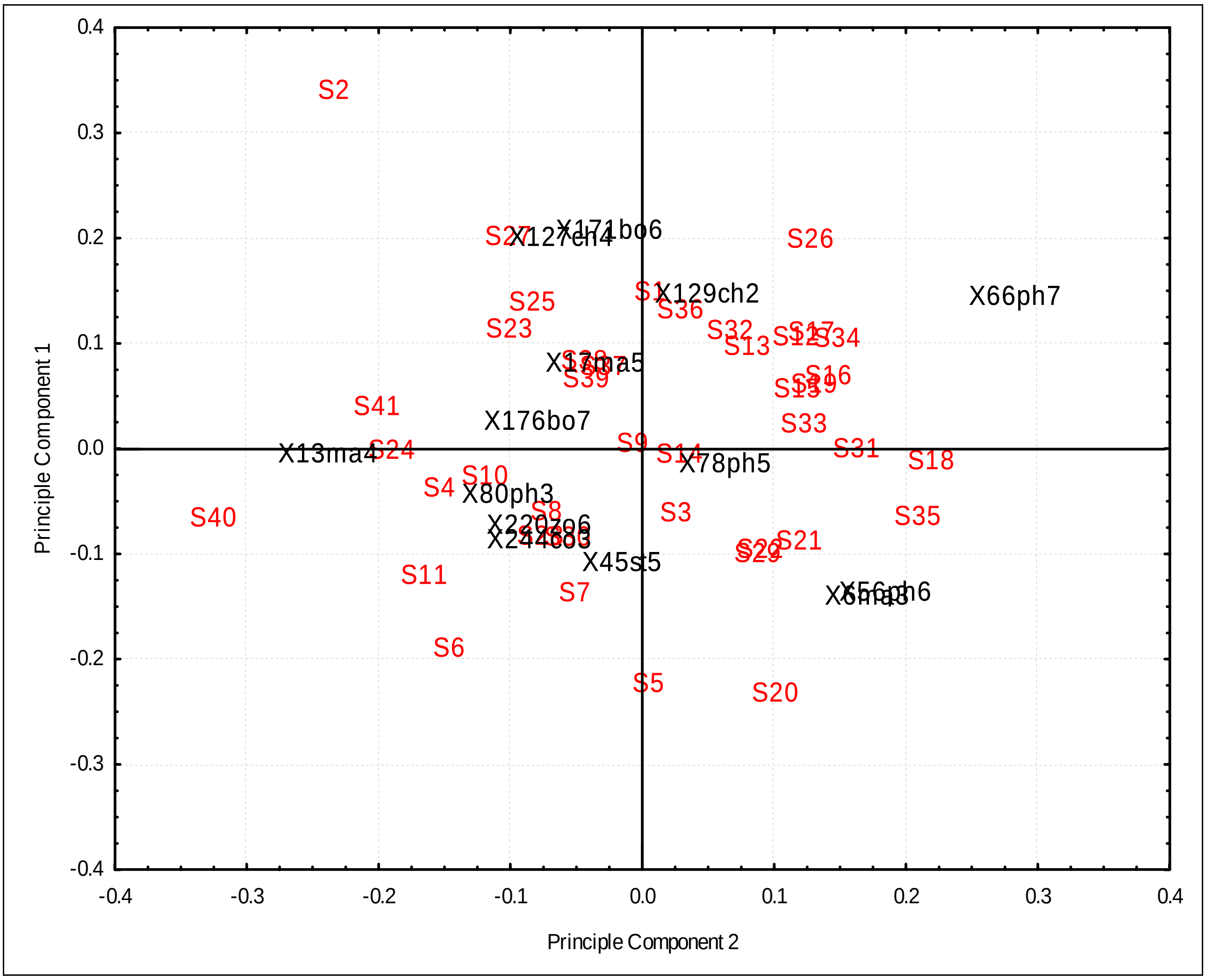


Figure (A1.6): ***Biplot of the selected variables from the sixth block***

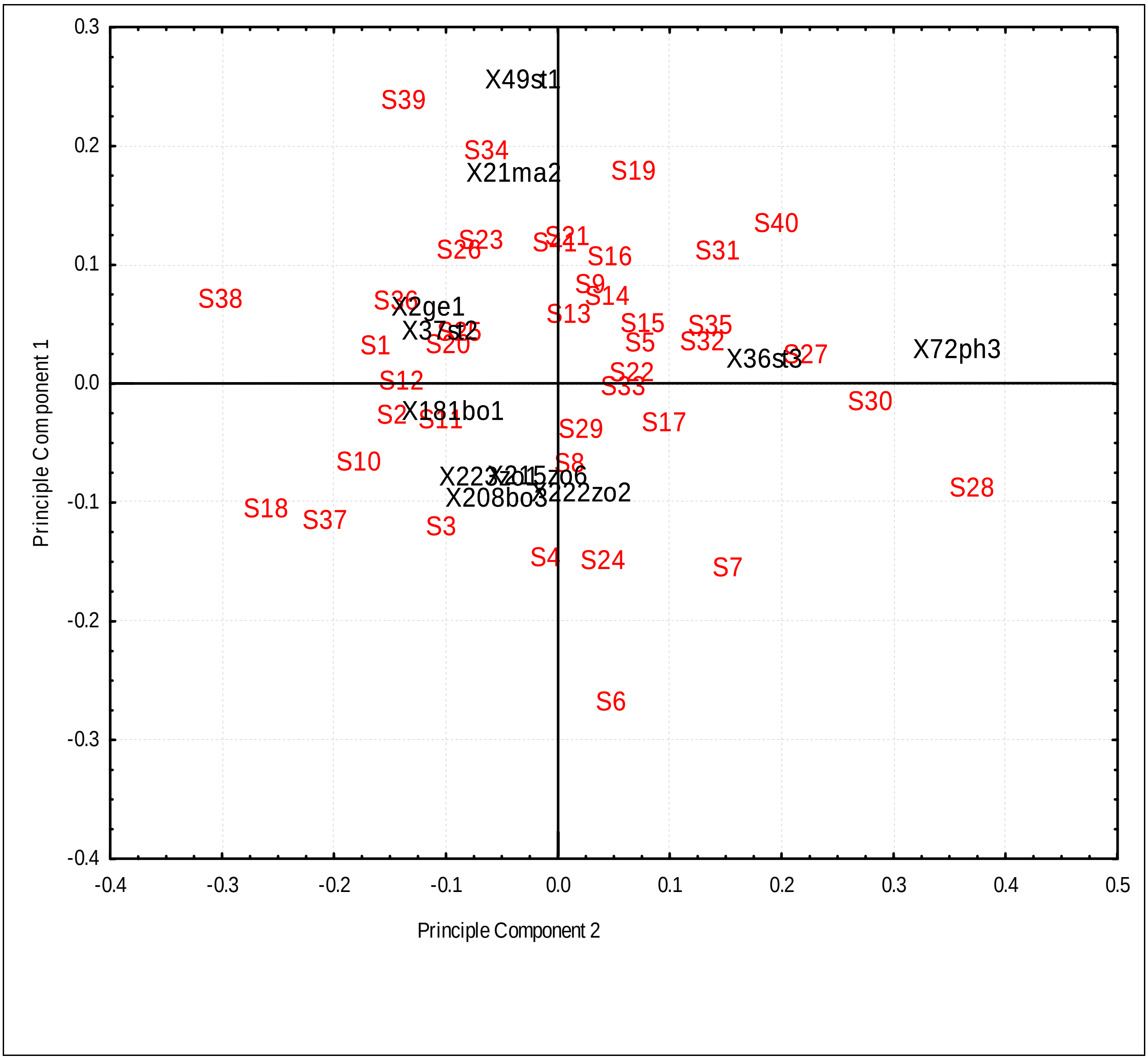


Figure (A1.7): ***Biplot of the selected variables from the seventh block***

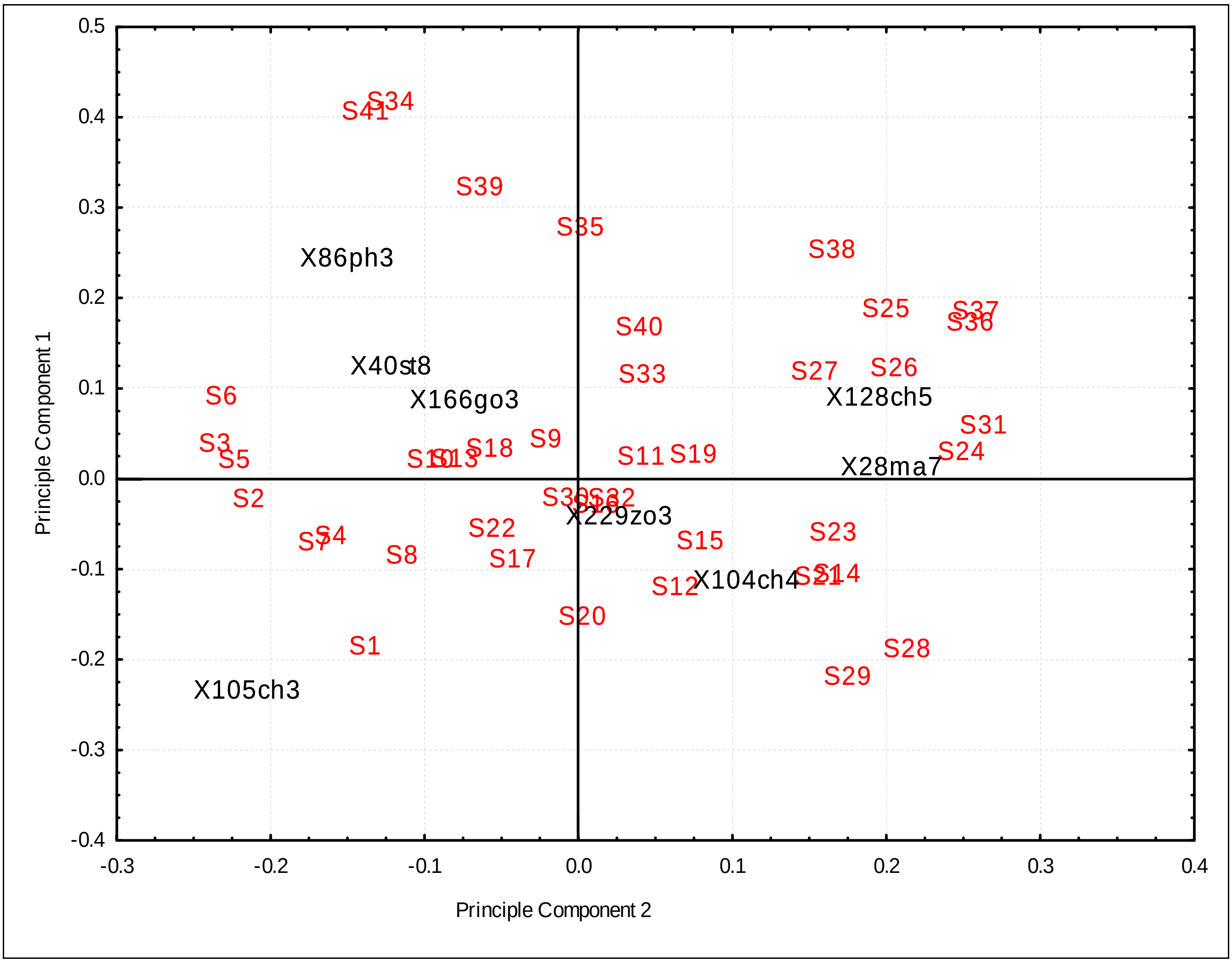


Figure (A1.8): ***Biplot of the selected variables from the eighth block***

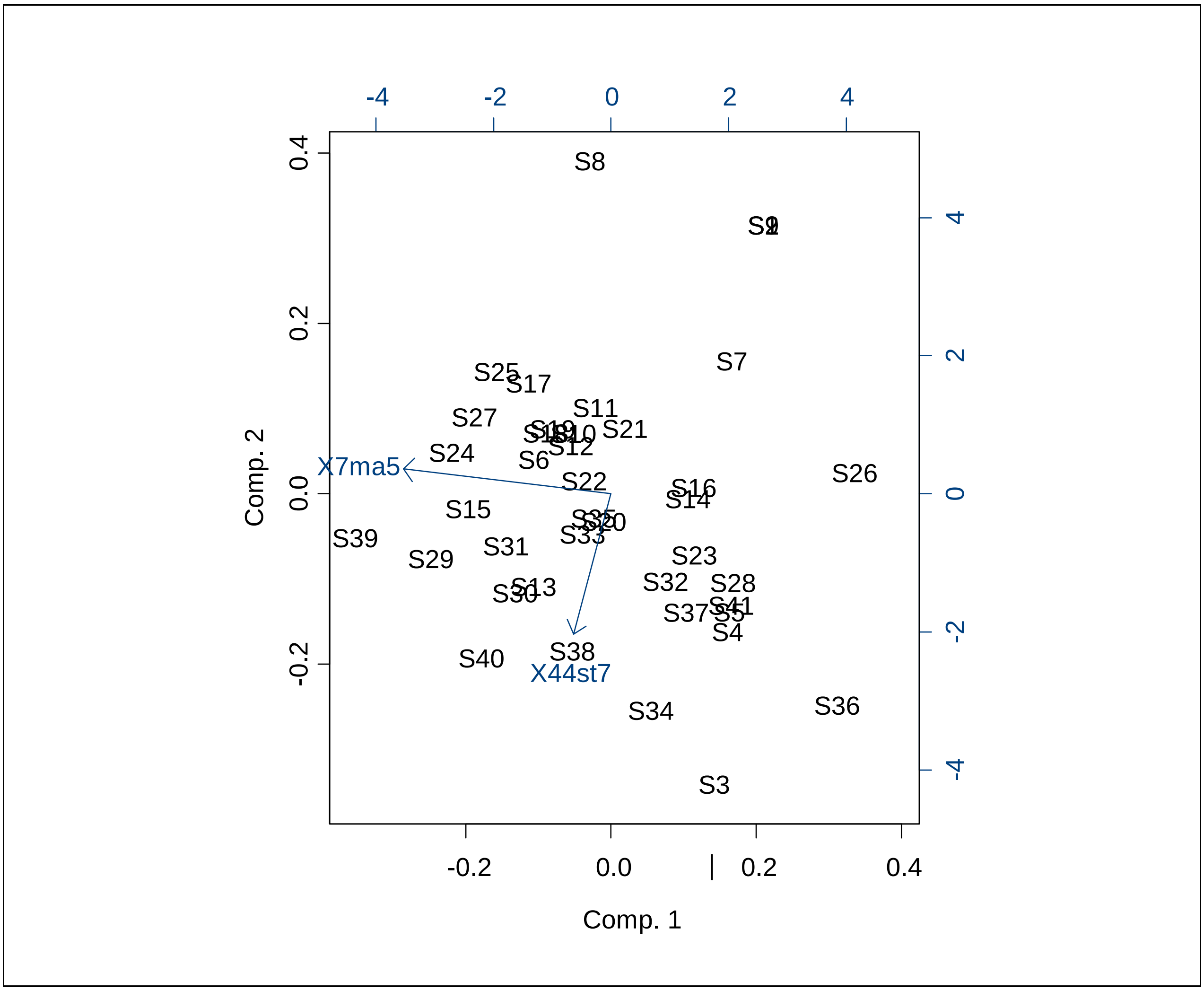


Figure (A1.9): ***Biplot of the selected variables from the ninth block***

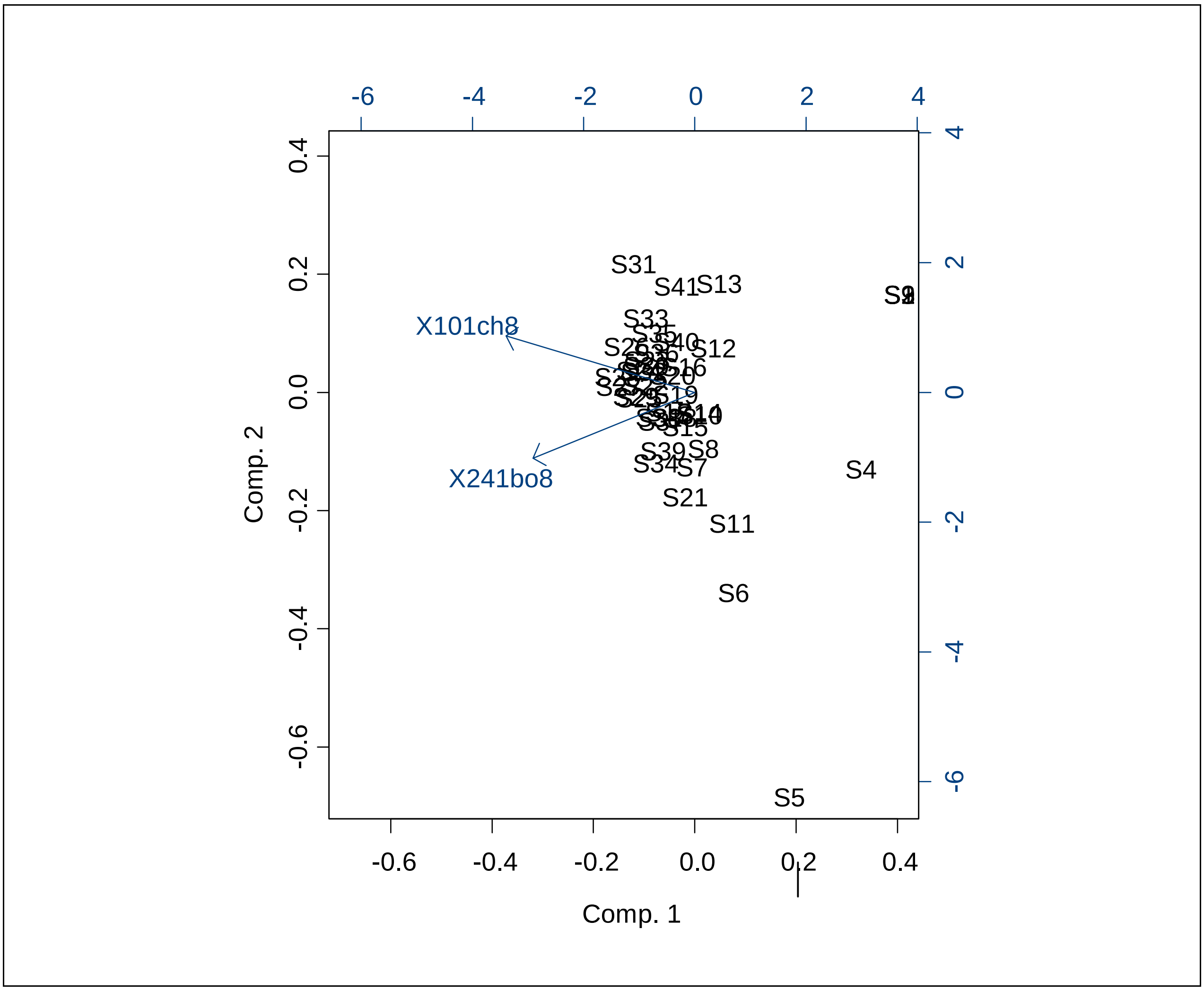


Figure (A1.10): ***Biplot of the selected variables from the tenth block***

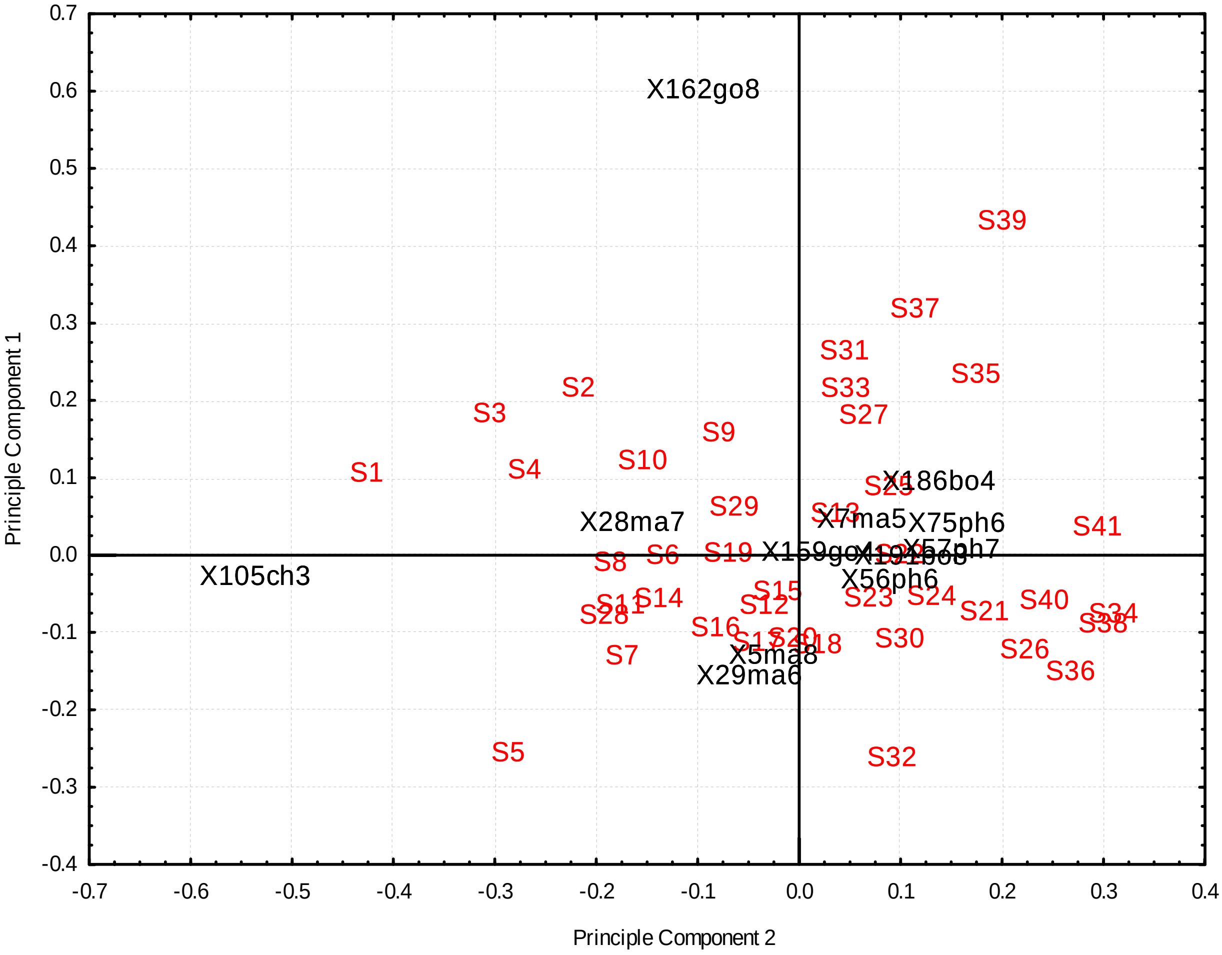


Figure (A1.11): ***Biplot of the selected variables from the last PCA***

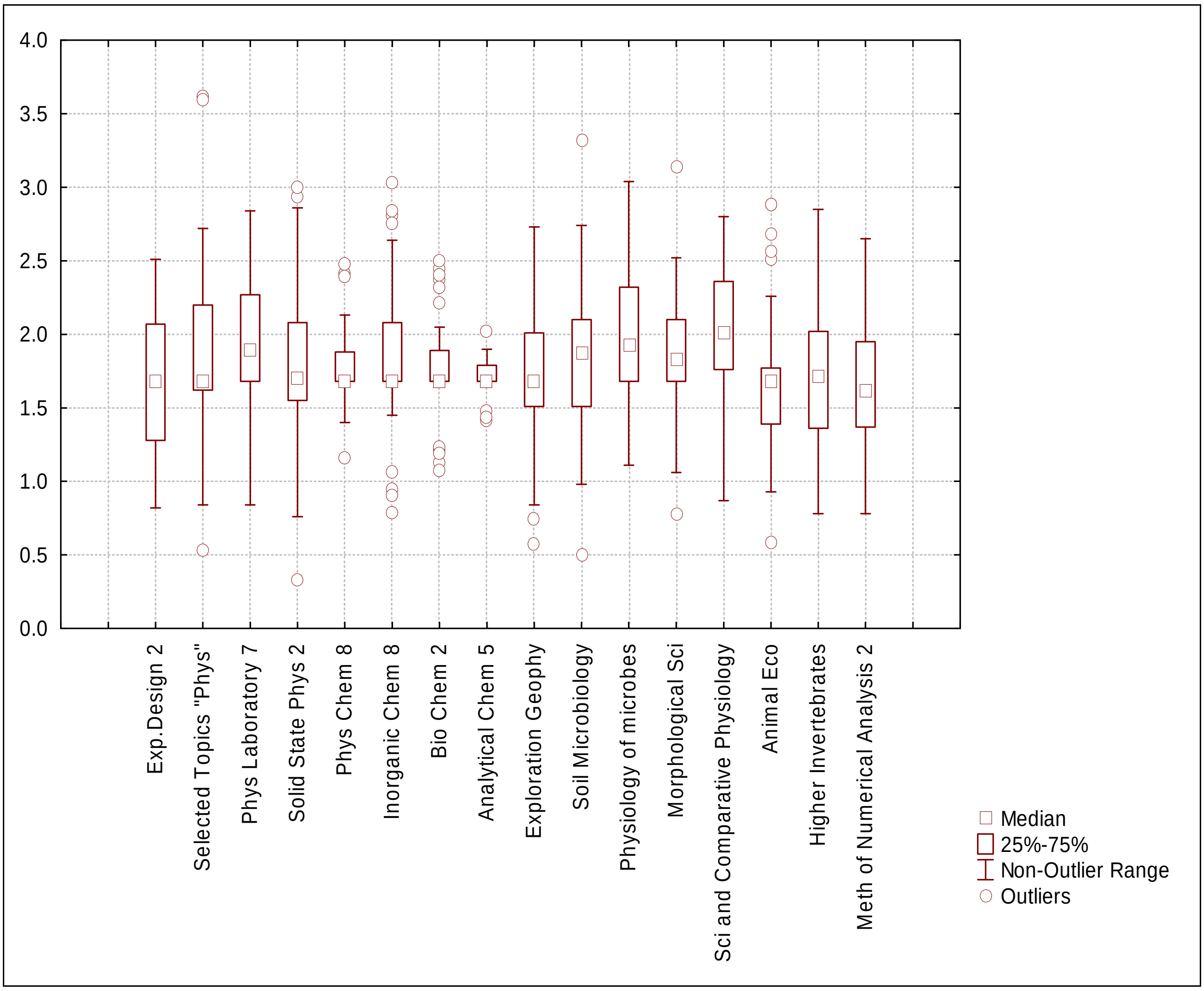


Figure (A1.12): ***Box plot of the selected variables from first block***

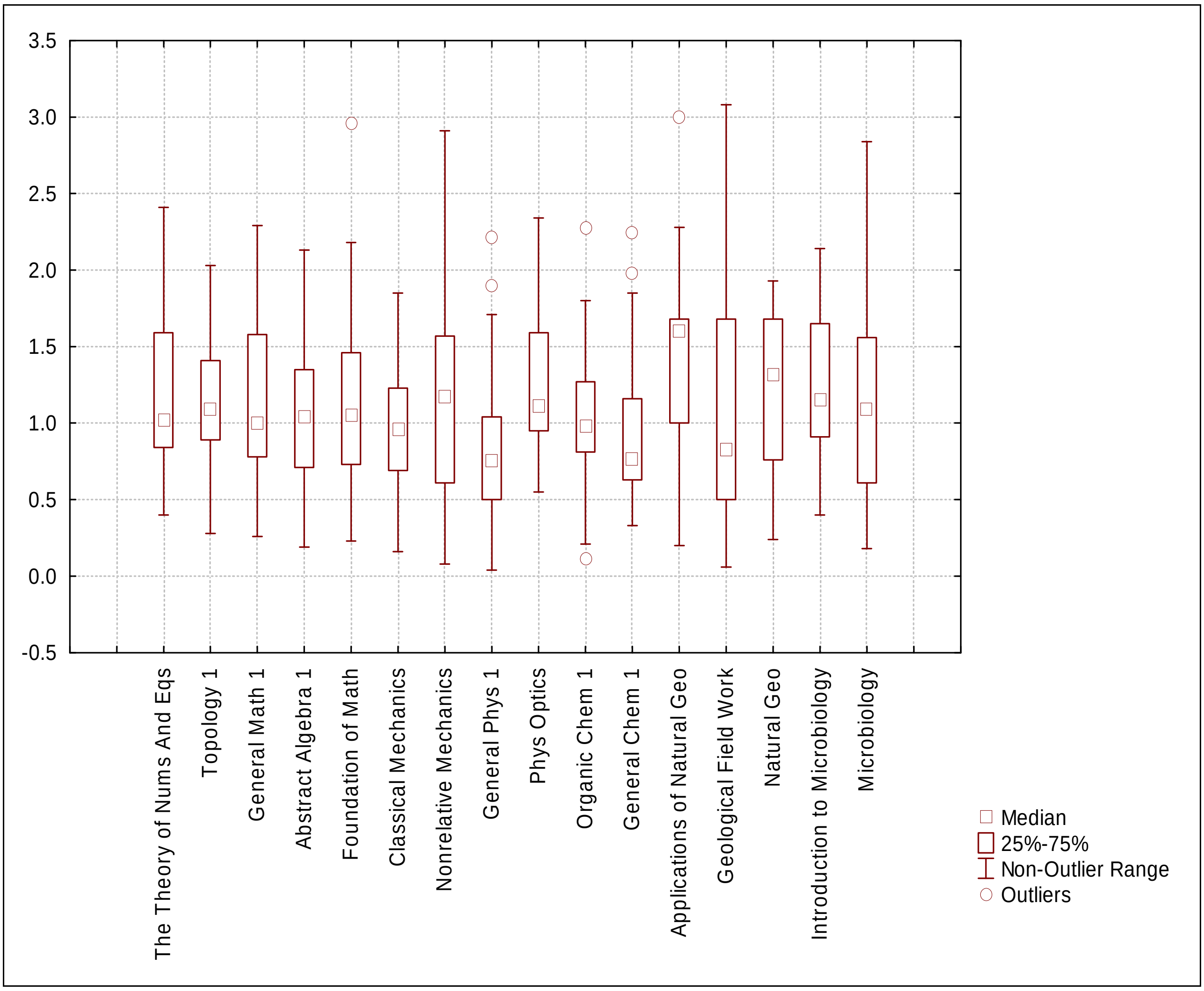


Figure (A1.13): ***Box plot of the selected variables from second block***

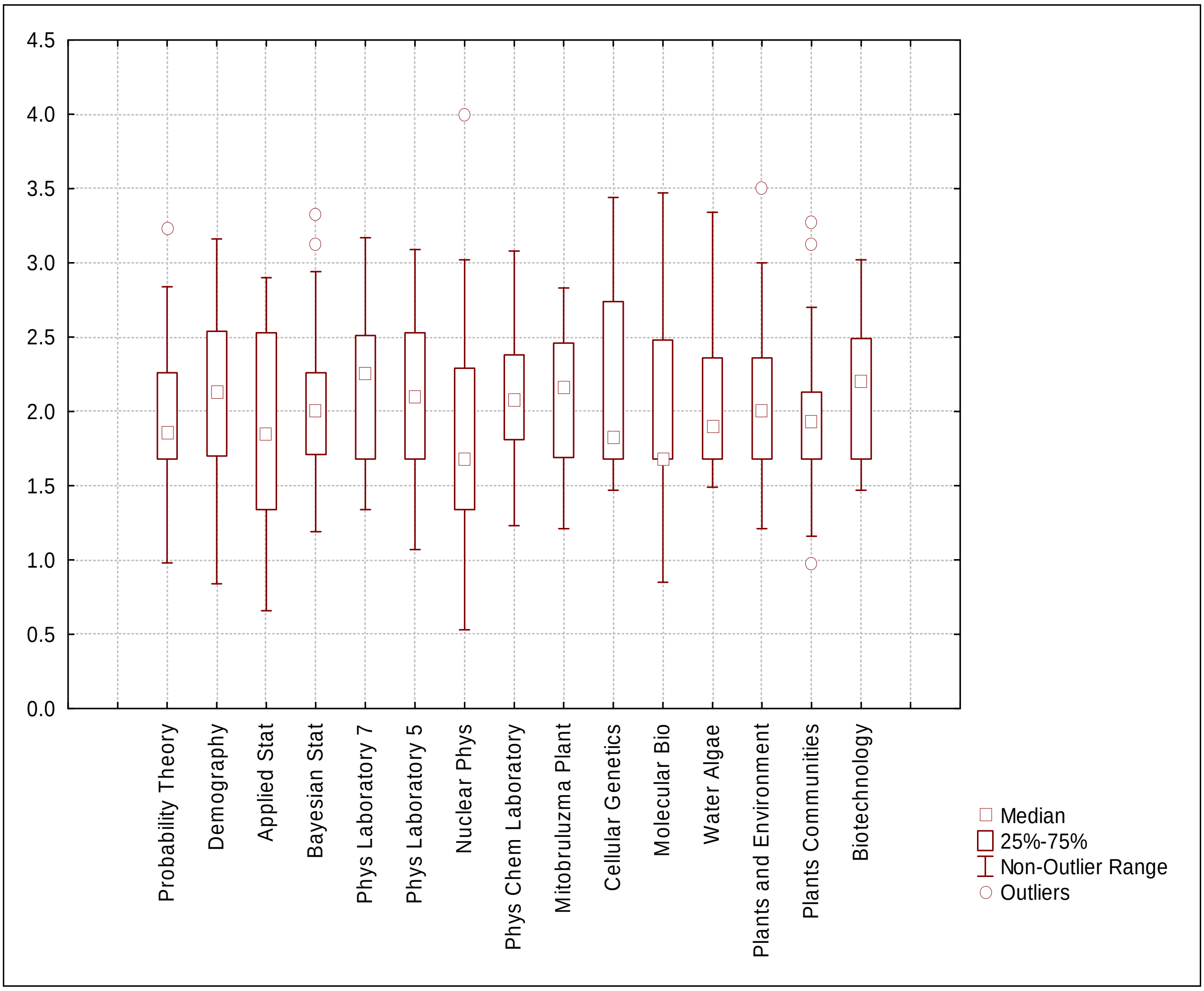


Figure (A1.14): ***Box plot of the selected variables from third block***

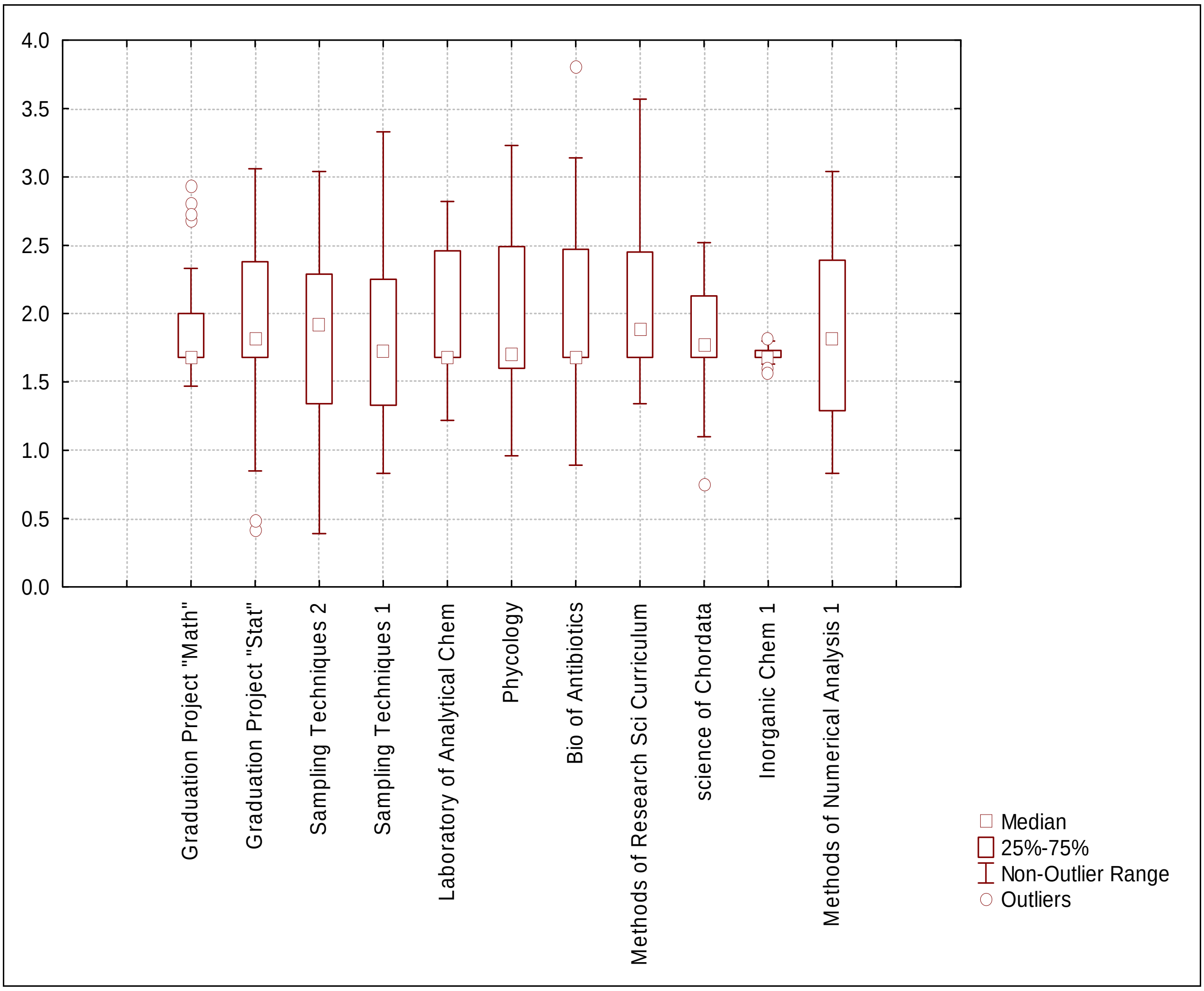


Figure (A1.15): ***Box plot of the selected variables from fourth block***

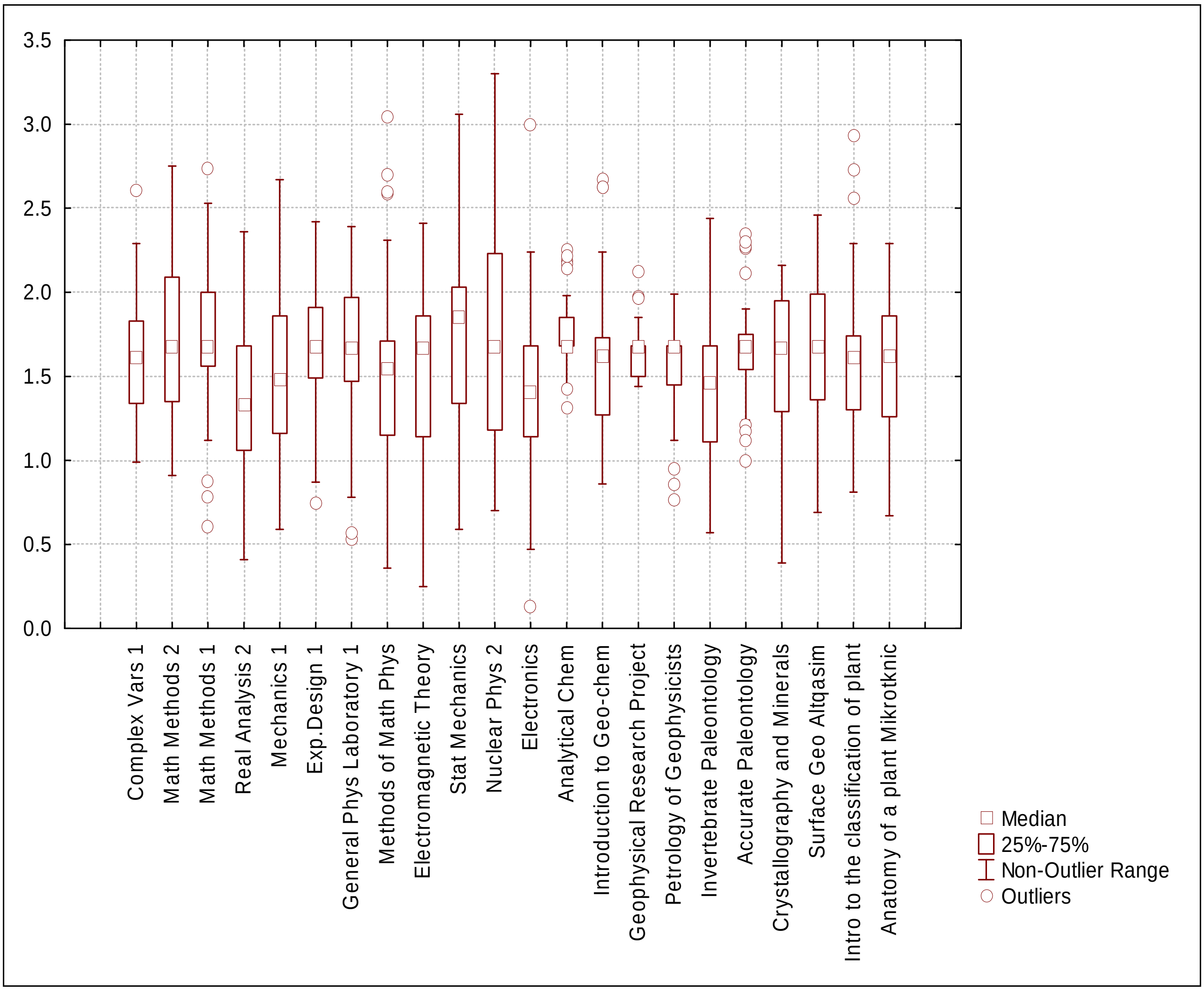


Figure (A1.16): ***Box plot of the selected variables from fifth block***

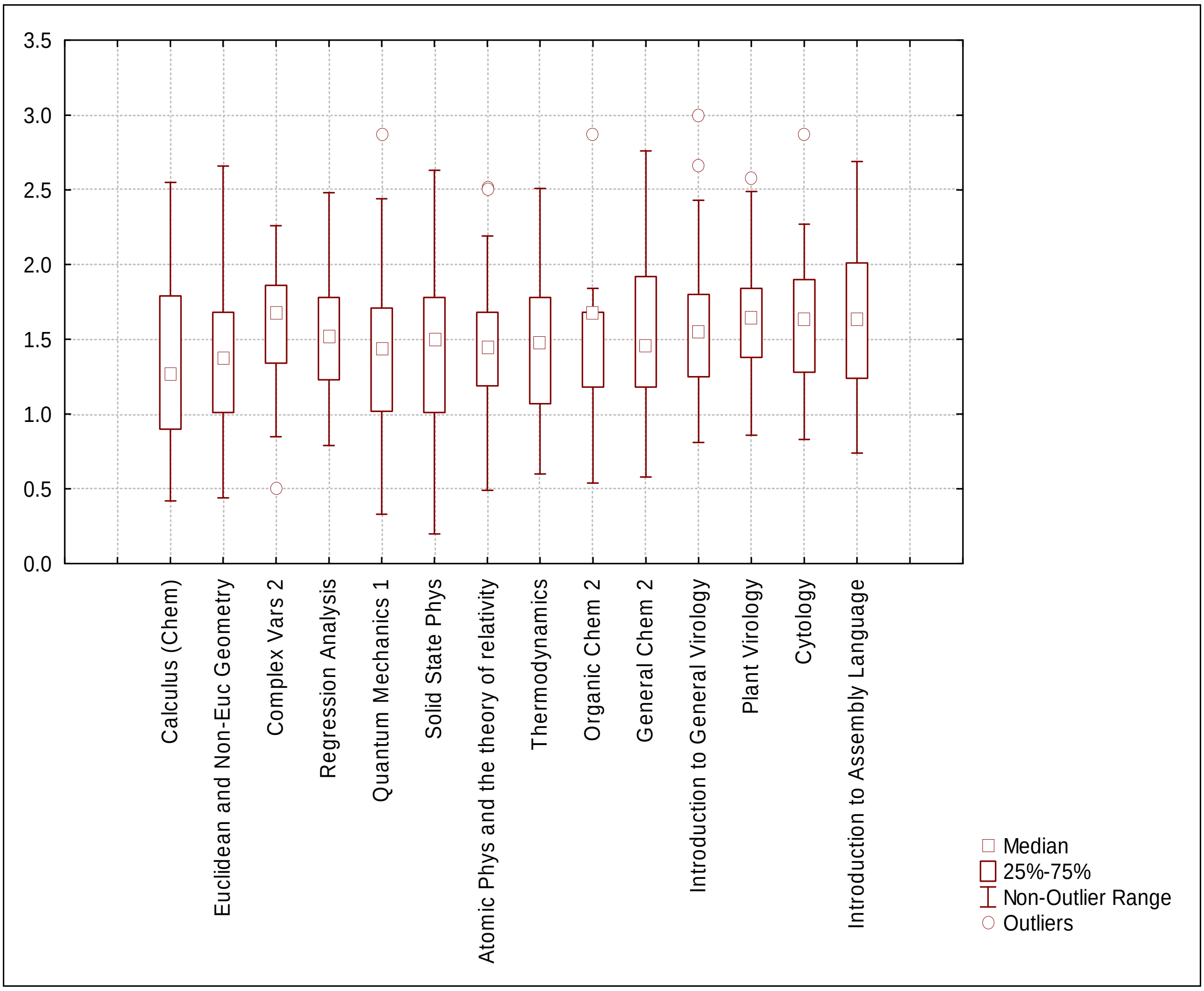


Figure (A1.17): ***Box plot of the selected variables from sixth block***

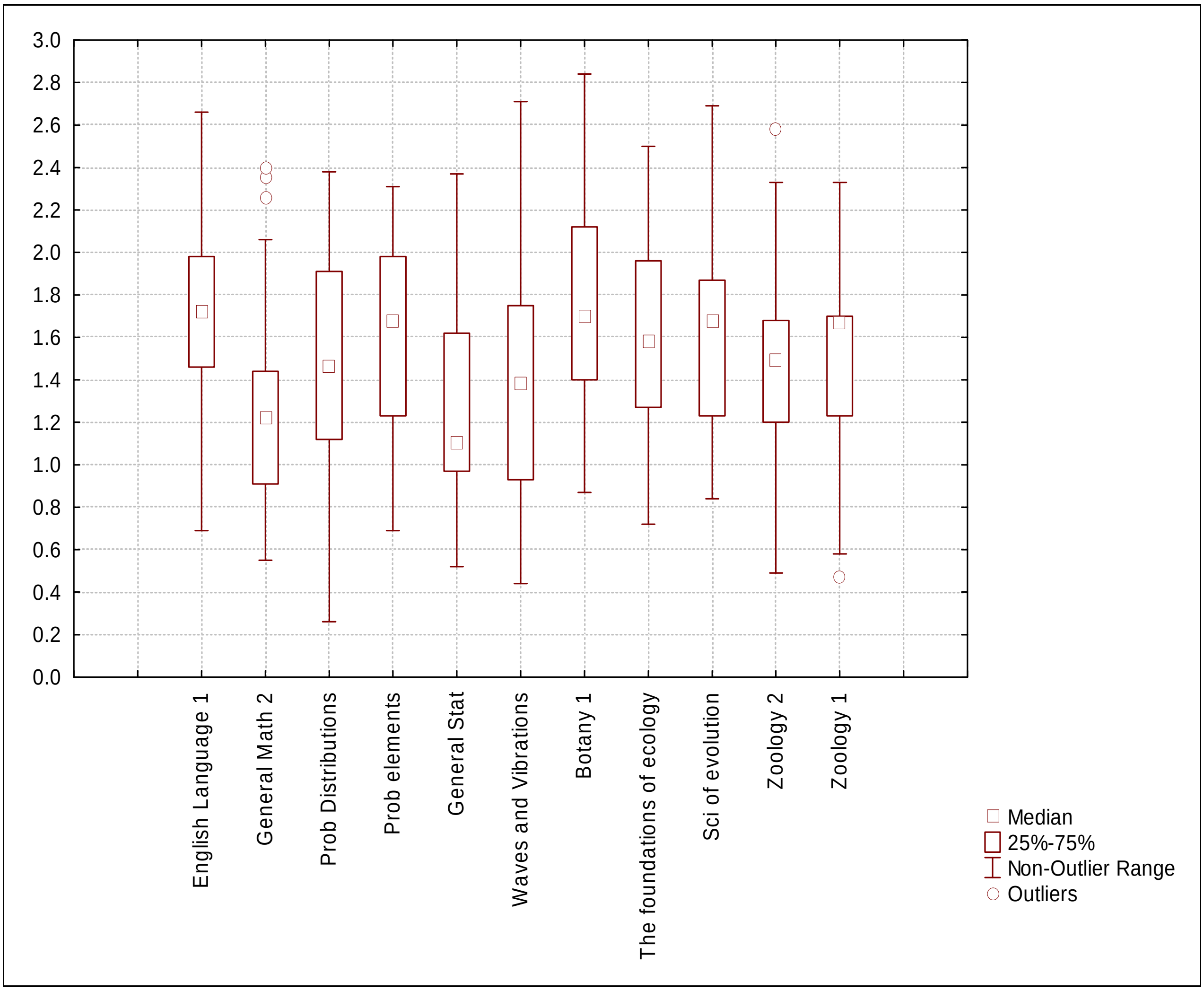


Figure (A1.18): ***Box plot of the selected variables from seventh block***

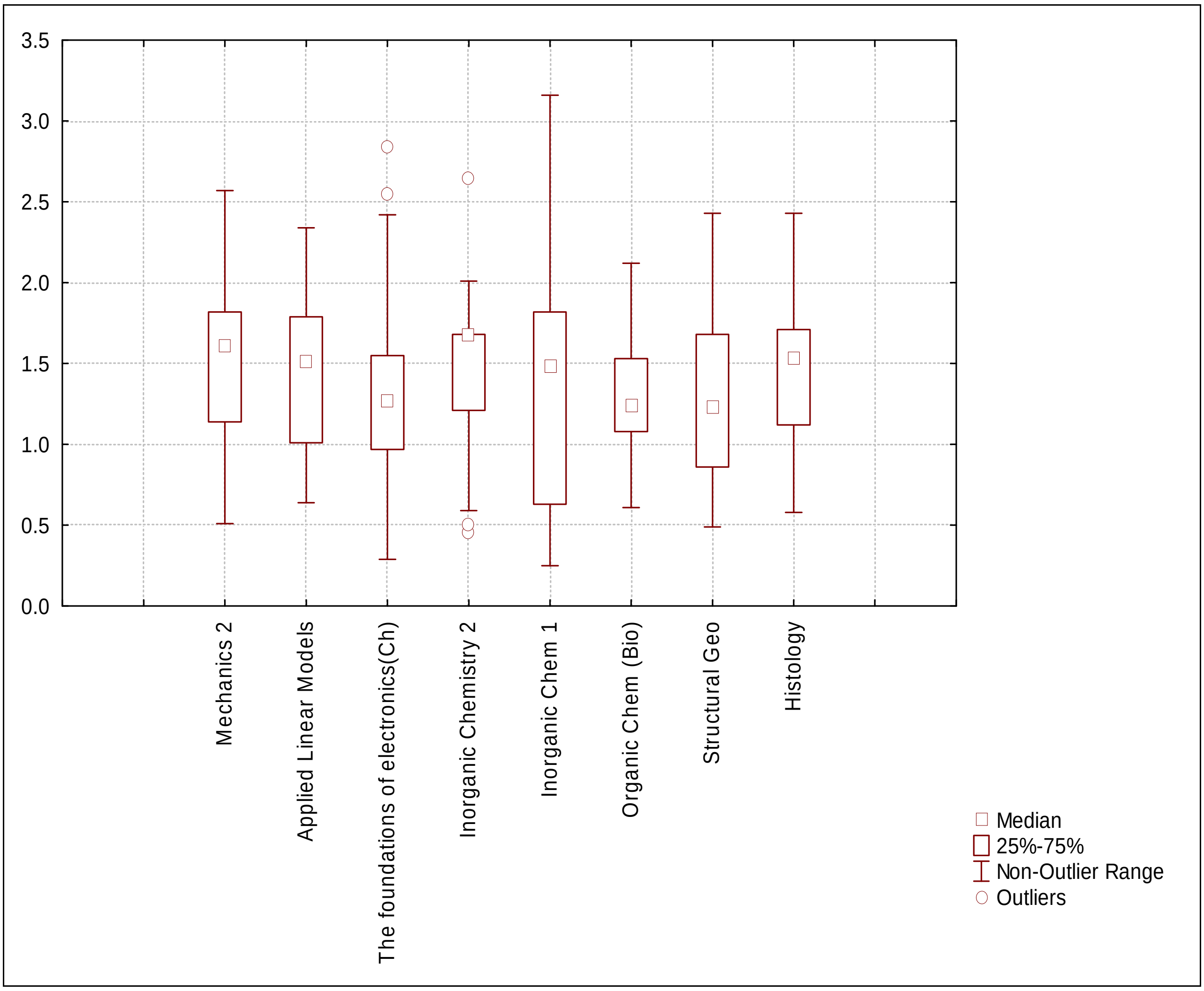


Figure (A1.19): ***Box plot of the selected variables from eighth block***

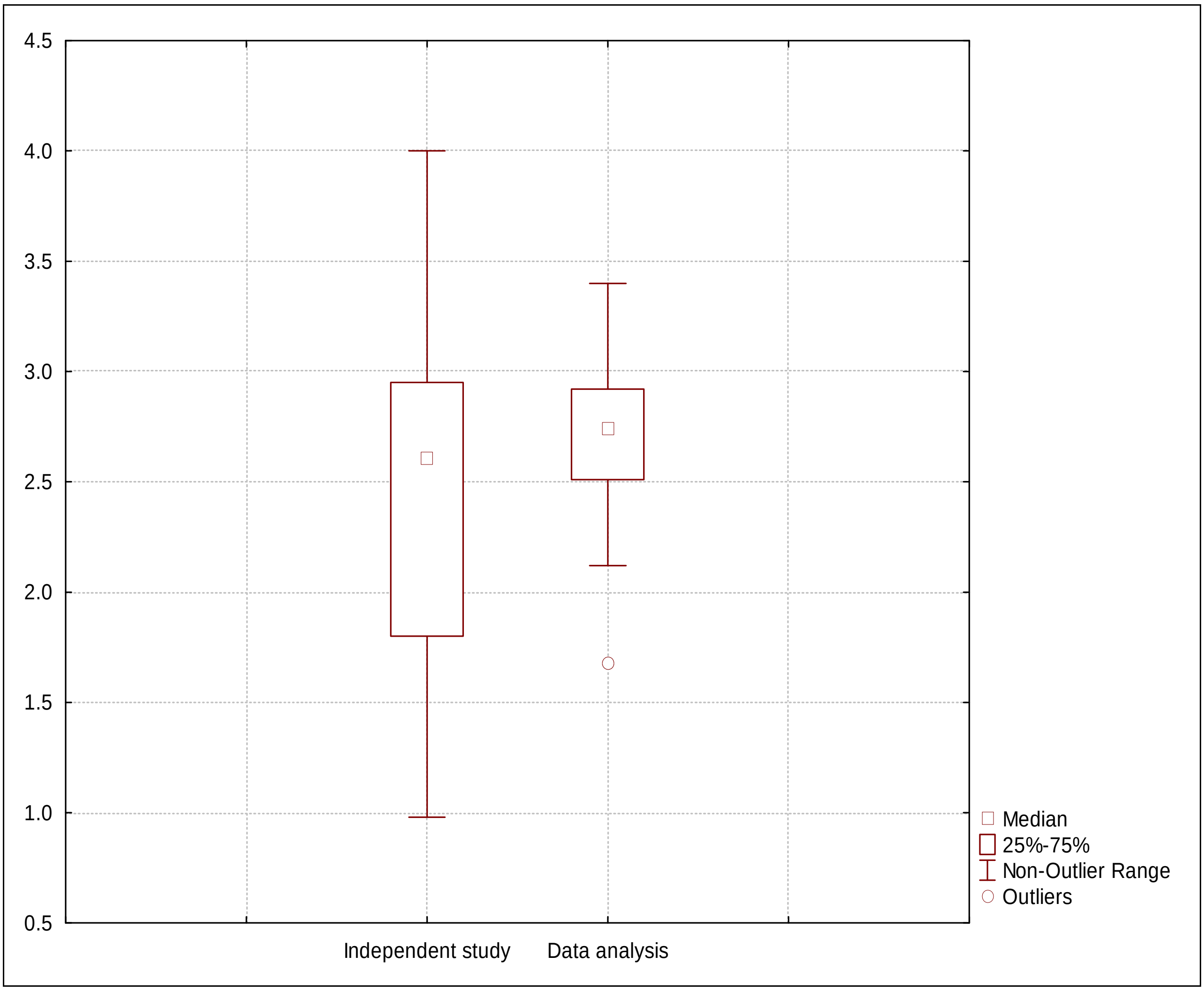


Figure (A1.20): ***Box plot of the selected variables from ninth block***

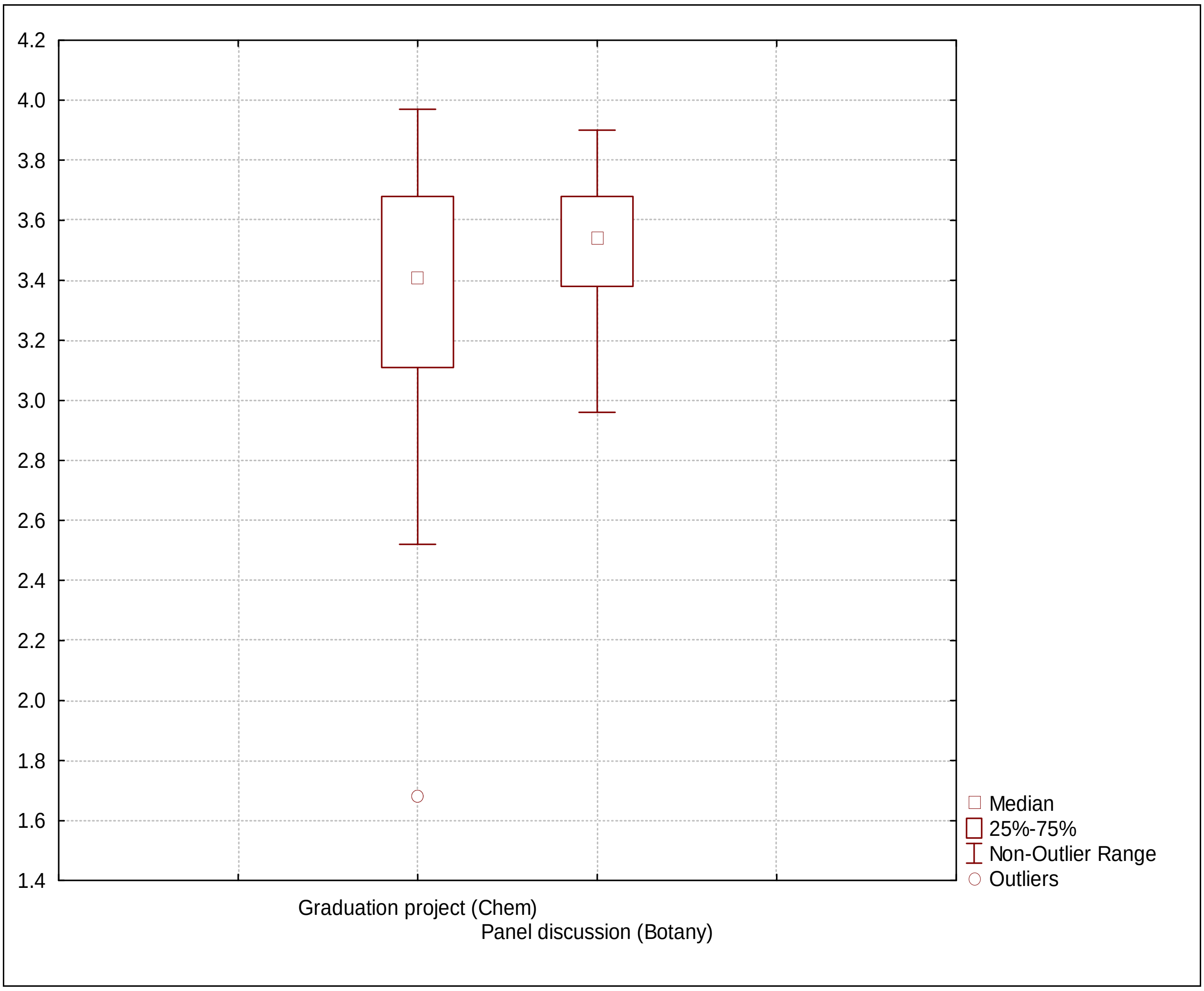


Figure (A1.21): ***Box plot of the selected variables from tenth block***

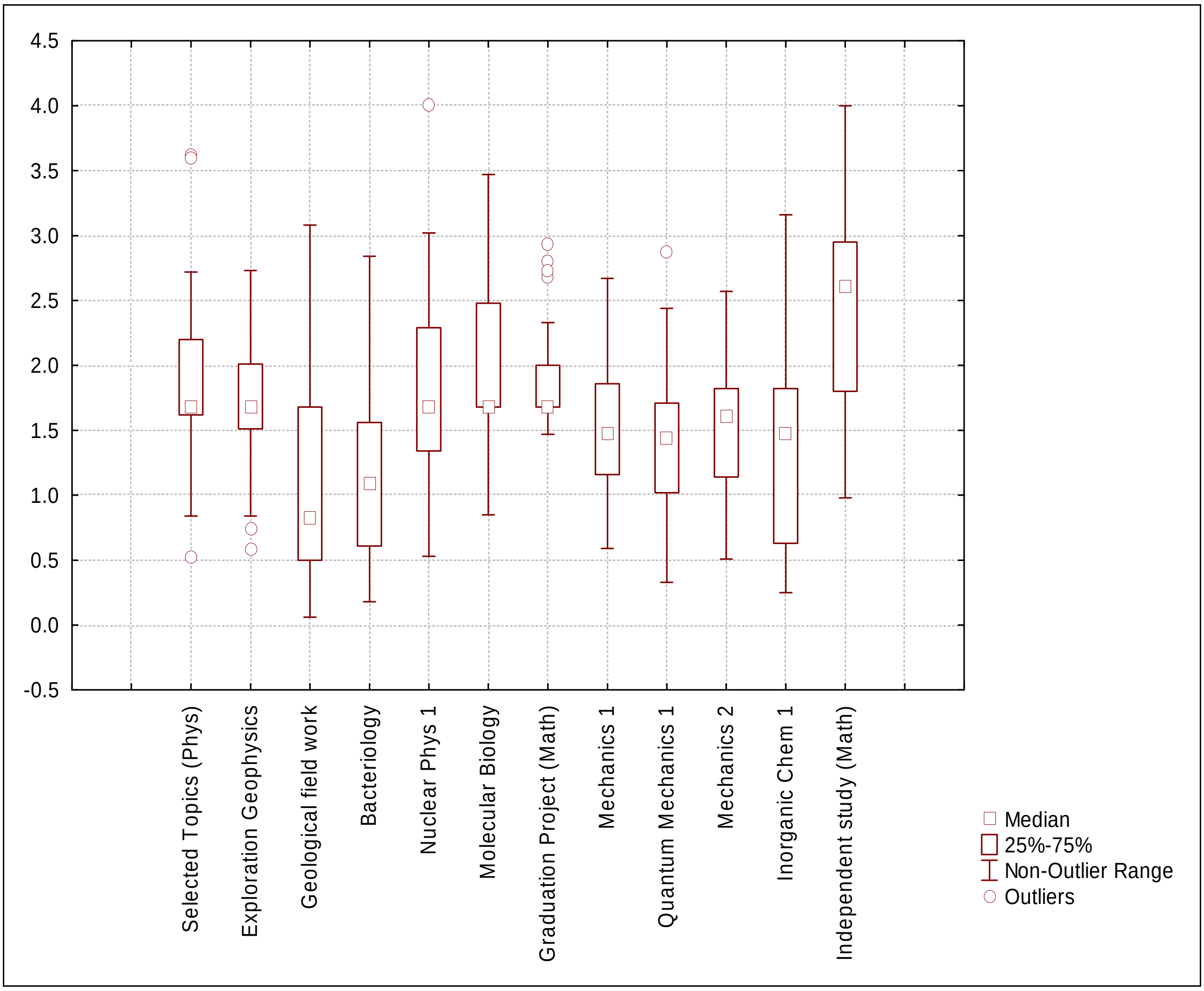


Figure (A1.22): ***Box plot of the selected variables from the final analysis***

Appendix A2 (Additional Tables)

Table (A2.1): ***Summary of the selected variables from the first block***

Variables	Min	Max	Mean	Median	Standard Deviation	C.V
1.Exp.Design 2.	0.82	2.51	1.68	1.68	0.48	28.90
2.Selected topics "physics".	0.53	3.62	1.84	1.68	0.68	37.06
3.Phys laboratory 7.	0.84	2.84	2.00	1.89	0.44	21.75
4.Solid State Phys 2.	0.33	3.00	1.79	1.70	0.58	32.58
5.Phys chemistry 8.	0.49	2.97	1.73	1.68	0.44	25.15
6.Inorganic Chemistry 8.	0.27	3.03	1.83	1.68	0.59	31.92
7.Bio Chemistry 2.	0.50	2.50	1.71	1.68	0.41	23.82
8.Analytical Chemistry 5.	0.45	2.90	1.82	1.68	0.47	25.78
9.Exploration Geophysics.	0.58	2.73	1.71	1.68	0.47	27.82
10.Soil Microbiology.	0.50	3.32	1.82	1.87	0.52	28.71
11.Physiology of microbes.	1.11	3.04	2.01	1.93	0.43	21.14
12.Morphological Science.	0.78	3.14	1.84	1.83	0.42	22.92
13.Sci and Comparative Physiology.	0.87	2.80	2.04	2.01	0.41	19.99
14.Animal Eco.	0.59	3.00	1.73	1.68	0.55	31.88
15.Higher Invertebrates.	0.78	2.85	1.72	1.71	0.51	29.88
16.Meth of Numerical Analysis 2.	0.78	2.65	1.64	1.62	0.46	27.82

Table (A2.2): ***Summary of the selected variables from the second block***

Variable	Min	Max	Mean	Median	Standard Deviation	C.V
1.The theory of Numbers and Equations.	0.40	2.41	1.18	1.02	0.48	41.08
2.Topology 1.	0.28	2.03	1.14	1.09	0.42	37.11
3.General Math 1.	0.26	2.29	1.09	1.00	0.57	52.44
4.Abstract Algebra 1.	0.19	2.13	1.08	1.04	0.48	44.75
5.Foundation of Math.	0.23	2.96	1.13	1.05	0.56	49.37
6.Classical Mechanics.	0.16	1.85	0.97	0.96	0.42	43.19
7.Nonrelative Mechanics.	0.08	2.91	1.17	1.17	0.62	53.50
8.General Phys 1.	0.04	2.21	0.85	0.76	0.48	56.13
9.Phys Optics.	0.55	2.34	1.24	1.11	0.42	33.73
10.Organic Chemistry 1.	0.11	3.10	1.07	0.98	0.54	50.56
11.General Chemistry 1.	0.33	2.25	0.95	0.77	0.48	51.00
12.Applications of Natural Geo.	0.20	3.00	1.40	1.60	0.57	40.74
13.Geological field work.	0.06	3.08	1.04	0.83	0.71	67.87
14.Natural Geo.	0.24	1.93	1.20	1.32	0.51	42.74
15.Introduction to Microbiology.	0.40	2.14	1.25	1.15	0.44	35.14
16.Microbiology.	0.18	2.84	1.12	1.09	0.58	51.58

Table(A2.3): ***Summary of the selected variables from the third block***

Variables	Min	Max	Mean	Median	Standard Deviation	C.V
1.Probability Theory.	0.98	3.23	2.00	1.86	0.45	22.40
2.Demography.	0.84	3.16	2.08	2.13	0.57	27.44
3.Applied Stat.	0.66	2.90	1.91	1.85	0.65	34.14
4.Bayesian Stat.	1.19	3.33	2.06	2.00	0.44	21.49
5.Phys Laboratory 7.	1.34	3.17	2.16	2.25	0.49	22.74
6.Phys Laboratory 5.	1.07	3.09	2.13	2.10	0.44	20.50
7.Nuclear Phys.	0.53	4.00	1.76	1.68	0.70	39.82
8.Phys Chemistry Laboratory.	1.23	3.08	2.11	2.08	0.44	20.70
9.Mitobruluzma Plant.	1.21	2.83	2.09	2.16	0.44	21.26
10.Cellular Genetics.	1.47	3.44	2.13	1.82	0.57	26.89
11.Molecular Bio.	0.85	3.47	2.02	1.68	0.65	32.17
12.Water Algae.	1.49	3.34	2.09	1.90	0.51	24.49
13.Plants and Environment.	1.21	3.50	2.05	2.00	0.48	23.40
14.Plant Communities.	0.98	3.60	1.97	1.93	0.56	28.36
15.Biotechnology.	1.47	3.02	2.14	2.20	0.41	19.22

Table (A2.4): ***Summary of the selected variables from the fourth block***

Variables	Min	Max	Mean	Median	Standard Deviation	C.V
1.Graduation Project "Math".	1.47	3.78	1.98	1.68	0.60	30.20
2.Graduation Project "Stat".	0.41	3.06	1.94	1.82	0.61	31.55
3.Sampling Techniques 2.	0.39	3.04	1.82	1.92	0.62	34.03
4.Sampling Techniques 1.	0.83	3.33	1.81	1.72	0.60	32.99
5.Laboratory of Analytical Chem.	1.22	2.82	2.00	1.68	0.45	22.68
6.Phycology.	0.96	3.23	2.01	1.70	0.59	29.26
7.Bio of Antibiotics.	0.89	3.80	2.03	1.68	0.59	28.87
8.Methods of research Sci Curriculum.	1.34	3.57	2.11	1.89	0.57	27.19
9.Science of Chordata.	0.75	3.60	1.83	1.77	0.49	26.91
10.Inorganic Chemistry 1.	0.47	3.16	1.81	1.68	0.48	26.22
11.Methods of Numerical Analysis 1.	0.83	3.04	1.88	1.82	0.61	32.60

Table (A2.5): ***Summary of the selected variables from the fifth block***

Variables	Min	Max	Mean	Median	Standard Deviation	C.V
1.Complex Vars 1.	0.99	2.61	1.62	1.61	0.34	21.27
2.Math Methods 2.	0.91	2.75	1.71	1.68	0.47	27.57
3.Math Methods 1.	0.61	2.74	1.73	1.68	0.46	26.38
4.Real Analysis 2.	0.41	2.36	1.41	1.33	0.52	36.64
5.Mechanics 1.	0.59	2.67	1.53	1.48	0.49	32.21
6.Exp. Design 1.	0.75	2.42	1.67	1.68	0.38	22.56
7.General Phys Laboratory 1.	0.53	2.39	1.65	1.67	0.45	27.15
8.Methods of Math Phys.	0.36	3.04	1.53	1.55	0.59	38.58
9.Electromagnetic Theory.	0.25	2.41	1.53	1.67	0.54	35.51
10.Stat Mechanics.	0.59	3.06	1.72	1.85	0.52	30.33
11.Nuclear Phys 2.	0.70	3.30	1.70	1.68	0.64	37.64
12.Electronics.	0.13	3.00	1.43	1.41	0.51	35.68
13.Analytical Chemistry.	0.47	2.37	1.71	1.68	0.35	20.34
14.Introduction to Geo-Chemistry.	0.86	2.67	1.59	1.62	0.46	28.67
15.Geophysical Research Project.	0.13	2.60	1.57	1.68	0.52	33.01
16.Petrology of Geophysicists.	0.25	2.64	1.55	1.68	0.41	26.29
17.Invertebrate Paleontology.	0.57	2.44	1.46	1.46	0.44	29.92
18.Accurate Paleontology.	0.80	2.80	1.71	1.68	0.43	25.24
19.Crystallography and Minerals.	0.39	2.16	1.54	1.67	0.47	30.17
20.Surface Geo Altqasim.	0.69	2.46	1.66	1.68	0.45	27.39
21.Introduction to the Classification of Plant.	0.81	2.93	1.62	1.61	0.46	28.28
22.Anatomy of a Plant Mikrotknic.	0.67	2.29	1.59	1.62	0.41	25.46

Table (A2.6): *Summary of the selected variables from the sixth block*

Variables	Min	Max	Mean	Median	Standard Deviation	C.V
1.Calculus "Chem".	0.42	2.55	1.34	1.27	0.53	39.96
2.Euclidean and Non-Euc Geometry.	0.44	2.66	1.38	1.37	0.49	35.53
3.Complex Vars 2.	0.50	2.26	1.57	1.68	0.39	24.68
4.Regression Analysis.	0.79	2.48	1.56	1.52	0.45	29.12
5.Quantum Mechanics 1.	0.33	2.87	1.39	1.44	0.55	39.66
6.Solid State Physics.	0.20	2.63	1.47	1.50	0.56	38.03
7.Atomic Phys and the Theory of Relativity.	0.49	2.51	1.48	1.45	0.42	28.32
8.Thermodynamics.	0.60	2.51	1.45	1.48	0.49	33.59
9.Organic Chemistry 2.	0.54	2.87	1.43	1.68	0.42	29.50
10.General Chemistry 2.	0.58	2.76	1.54	1.46	0.51	33.09
11.Introduction to General Virology.	0.81	3.00	1.60	1.55	0.48	30.00
12.Plant Virology.	0.86	2.58	1.66	1.65	0.42	25.63
13.Cytology.	0.83	2.87	1.62	1.63	0.42	25.97
14.Introduction to Assembly Language.	0.74	2.69	1.64	1.64	0.45	27.77

Table (A2.7): ***Summary of the selected variables from the seventh block***

Variables	Min	Max	Mean	Median	Standard Deviation	C.V
1.English Language 1.	0.69	2.66	1.71	1.72	0.41	24.31
2.General Mathematics 2.	0.55	2.40	1.28	1.22	0.46	36.16
3.Probability Distributions.	0.26	2.38	1.52	1.46	0.53	35.00
4.Probability Elements.	0.69	2.31	1.58	1.68	0.48	30.66
5.General Statistics.	0.52	2.37	1.27	1.10	0.49	38.52
6.Waves and Vibrations.	0.44	2.71	1.42	1.38	0.61	42.64
7.Botany 1.	0.87	2.84	1.73	1.70	0.52	30.22
8.The foundations of Ecology.	0.72	2.50	1.60	1.58	0.48	29.84
9.Science of Evolution.	0.84	2.69	1.64	1.68	0.50	30.40
10.Zoology 2.	0.49	2.58	1.49	1.49	0.46	30.78
11.Zoology 1.	0.47	2.33	1.45	1.67	0.48	33.16

Table (A2.8): ***Summary of the selected variables from the eighth block***

Variables	Min	Max	Mean	Median	Standard Deviation	C.V
1.Mechanics 2.	0.51	2.57	1.57	1.61	0.52	33.16
2.Applied Linear Models.	0.64	2.34	1.46	1.51	0.51	35.26
3.The foundations of electronics "Chem".	0.29	2.84	1.31	1.27	0.54	41.15
4.Inorganic Chemistry 2.	0.46	2.65	1.46	1.68	0.44	30.26
5.Inorganic Chemistry 1.	0.25	3.16	1.41	1.48	0.80	57.17
6.Organic Chemistry "Bio".	0.61	2.12	1.30	1.24	0.33	25.51
7.Structural Geology.	0.49	2.43	1.28	1.23	0.50	38.84
8.Histology.	0.58	2.43	1.41	1.53	0.45	31.91

Table (A2.9): ***Summary of the selected variables from the ninth block***

Variables	Min.	Max.	Mean	Median	Standard Deviation	C.V
1.Independent Study.	0.98	4.00	2.45	2.61	0.71	29.03
2.Data Analysis.	1.68	3.40	2.66	2.74	0.43	16.02

Table (A2.10): ***Summary of the selected variables from the tenth block***

Variables	Min	Max	Mean	Median	Standard Deviation	C.V
1.Graduation Project "Chemistry".	1.68	3.97	3.19	3.41	0.70	22.00
2.Panel Discussion "Botany".	1.68	3.90	3.35	3.54	0.62	18.41

Table (A2.11): ***Summary of the selected variables from the final PCA***

Variables	Min	Max	Mean	Median	Standard Deviation	C.V
1.Selected Topics "Phys".	0.53	3.62	1.84	1.68	0.68	37.06
2.Exploration Geophysics.	0.58	2.73	1.71	1.68	0.47	27.82
3.Geological field work.	0.06	3.08	1.04	0.83	0.71	67.87
4.Bacteriology.	0.18	2.84	1.12	1.09	0.58	51.58
5.Nuclear Physics 1.	0.53	4.00	1.76	1.68	0.70	39.82
6.Molecular Biology.	0.85	3.47	2.02	1.68	0.65	32.17
7.Graduation Project "Math".	1.47	3.78	1.98	1.68	0.60	30.20
8.Mechanics 1.	0.59	2.67	1.53	1.48	0.49	32.21
9.Quantum Mechanics 1.	0.33	2.87	1.39	1.44	0.55	39.66
10.Mechanics 2.	0.51	2.57	1.57	1.61	0.52	33.16
11.Inorganic Chemistry 1.	0.25	3.16	1.41	1.48	0.80	57.17
12.Independent study "Math".	0.98	4.00	2.45	2.61	0.71	29.03

الخلاصة

في العديد من الدراسات وجد أن تطبيق اغلب أساليب الاستدلال الإحصائي على قواعد البيانات الضخمة يكون صعب،غير ملائم و غير موثوق به. احد الحلول الحديثة المتبناة للتعامل مع الإعداد الكبيرة للمتغيرات هو تحليل قطاعات المركبات الأساسية (Block PCA). هذا الأسلوب المعدل يستخدم لتخفيض البيانات عن طريق اختيار المتغيرات التي تحتوي على اكبر كمية ممكنة من المعلومات (الاختلاف). هذه المتغيرات المختارة يمكن ان تستخدم في دراسات وبحوث مستقبلية.

في هذه الدراسة تحليل قطاعات المركبات الأساسية تم استخدامه لتخفيض حجم قاعدة البيانات الضخمة لمتوسطات معدلات الطلبة في جميع المواد الخاصة بكلية العلوم بجامعة بنغازي.

بمعنى أخر تم إنشاء قاعدة بيانات بديلة لمتوسطات معدلات الطلبة بأقل عدد من المتغيرات التي تحتوي على اكبر كمية من الاختلاف في قاعدة البيانات الأصلية.

تحليل قطاعات المركبات الأساسية هو تحليل متعدد المراحل ملائم لتحليل المكونات الرئيسية الأصلي. حيث انه يتضمن تطبيق التحليل العنقودي و كذلك اختيار المتغيرات من خلال المكونات الرئيسية المختارة. تطبيق تحليل قطاعات المركبات الأساسية في هذه الدراسة هو عبارة عن نسخة معدلة من العمل الأصلي الذي قام به ليو وآخرون (2002) وجبريل (2005) .الهدف الرئيسي كان تطبيق تحليل المركبات الرئيسية على مجموعات اقل من المتغيرات بدلا من تطبيقه على مجموعة المتغيرات الكلية والذي ثبت انه غير موثوق به.

في هذه الدراسة عدد المتغيرات والتي تمتلك اكبر كمية ممكنة من الاختلاف من بين 251 متغير المتضمنة في هذه الدراسة تكون 12 متغير. المتغيرات الاثنى عشر(المقررات) والمختارة بشكل نهائي تم اختيارها باستخدام أسلوب اختيار المتغيرات وتخفيض البيانات (تحليل قطاعات المركبات الأساسية) وبهذا تم الحصول على قاعدة بيانات لمتوسطات معدلات الطلبة لكلية العلوم بحجم اقل واكبر كمية ممكنة من الاختلاف.

جامعة بنغازي
كلية العلوم
قسم الإحصاء

تطبيق تحليل قطاعات المركبات الأساسية على قاعدة بيانات معدلات درجات الطلبة بكلية العلوم بجامعة بنغازي

إعداد

أسامة حمزة الضراط

إشراف

د.رامي صلاح جبريل

هذه الأطروحة قدمت كمتطلب جزئي لنيل درجة الإجازة العالية (الماجستير) في علم الإحصاء

بنغازي- ليبيا

2012-2011