

Data and Spatial Analysis for Economy and Education of 28 E.U. Member-States for 2014

Alexiou Dimitra, Fragkaki Maria

Abstract—The objective of the paper is the study of geographic, economic and educational variables and their contribution to determine the position of each member-state among the EU-28 countries based on the values of seven variables as given by Eurostat. The Data Analysis methods of Multiple Factorial Correspondence Analysis (MFCA) Principal Component Analysis and Factor Analysis have been used. The cross tabulation tables of data consist of the values of seven variables for the 28 countries for 2014. The data are manipulated using the CHIC Analysis V 1.1 software package. The results of this program using MFCA and Ascending Hierarchical Classification are given in arithmetic and graphical form. For comparison reasons with the same data the Factor procedure of Statistical package IBM SPSS 20 has been used. The numerical and graphical results presented with tables and graphs, demonstrate the agreement between the two methods. The most important result is the study of the relation between the 28 countries and the position of each country in groups or clouds, which are formed according to the values of the corresponding variables.

Keywords—Multiple factorial correspondence analysis, principal component analysis, factor analysis, E.U.-28 countries, statistical package IBM SPSS 20, CHIC Analysis V 1.1 Software, Eurostat.eu statistics.

I. INTRODUCTION

THE analysis of the data, which are presented in two dimension matrix, can be performed using the known reduction methods of Principal Component Analysis, Factor Analysis and Correspondence Analysis and Hierarchical Analysis.

The main purpose of this paper is to analyze data from Eurostat Statistics concerning important socio-economic and educational parameters during specific period (2014). We are using data of seven variables for EE-28 countries. The coding of names for the seven variables and 28 countries are presented with capital letters using three letters for the variables and the official E.E. coding for the countries.

II. DATA FOR 7 VARIABLES AND EU 28 COUNTRIES

A. The Data Matrix

The data have collected from the base of eu.eurostat statistics [2]. From the vast amount of data, we have selected seven variables that were considered most relevant to the intended purpose of the research aims to study the relationship

Dimitra Alexiou is with the Dept. of Spatial Planning and Development Engineering, School of Engineering, Aristotle University of Thessaloniki, Greece (phone:00306937163380, e-mail: dimitraalexiou@plandevl.auth.gr).

Fragkaki Maria (PhD) is with the University of Macedonia Thessaloniki Greece (e-mail: mariafra@otenet.gr, phone: 00306945377070).

of education and economic fundamentals in the EU-28 countries for a certain period. The data used are:

TABLE I
 VARIABLE VALUES FOR EU-28 COUNTRIES

CNTR	GDP	UNE	HICP	TGE	DEB	DEF	GRD
BE	119	8.5	0.5	55.1	106.7	-3.1	46.6
BG	45	11.4	-1.6	42.1	27.1	-5.8	32.8
CZ	84	6.1	0.4	42.6	42.7	-1.9	35.9
DK	124	6.6	0.3	56.9	45.1	1.5	57.9
DE	124	5.1	0.8	44.3	74.9	0.3	46.6
EE	73	7.4	0.5	38.1	10.4	0.7	37.1
IE	132	11.3	0.3	38.2	107.5	-3.9	53.6
EL	72	26.5	-1.4	49.9	178.6	-3.6	30.1
ES	93	24.5	-0.2	44.5	99.3	-5.9	46.6
FR	107	10.3	0.6	57.5	95.6	-3.9	51.5
CR	59	17.3	0.2	48.2	85.1	-5.6	42.9
IT	97	12.7	0.2	51.2	132.3	-3.1	46.6
CY	85	16.1	-0.3	49.3	108.2	-8.9	40.6
LV	64	10.8	0.7	37.1	40.6	-1.5	27.8
LT	74	10.7	0.2	34.8	40.7	-0.7	31.7
LU	263	6.1	0.7	42.4	23.1	1.4	50.6
HU	68	7.7	0	49.9	76.2	-2.5	48.3
MT	85	5.9	0.8	44.1	68.3	-2.1	50.1
NL	130	7.4	0.3	46.2	68.2	-2.4	51.5
AU	128	5.6	1.5	52.7	84.2	-2.7	46.6
PL	68	9.1	0.1	42.1	50.4	-3.3	39.1
PT	78	14.1	-0.2	51.7	130.2	-7.2	40.6
RO	54	6.8	1.4	34.9	39.9	-1.4	32.9
SI	83	9.7	0.4	49.8	80.8	-5	68.4
SK	76	13.2	-0.1	41.6	53.5	-2.8	32.2
FI	110	8.7	1.2	58.3	59.3	-3.3	53.5
SE	124	7.9	0.2	51.8	44.9	-1.7	51.6
UK	108	6.1	1.5	43.9	88.2	-5.7	46.5

B. Variables

- (GDP) General Domestic Product, GDP per capita in PPS Index.
- (UNE) Unemployment rates of the population aged 25 to 64 Annual average.
- (HICP) Harmonized Indices of Consumer Prices (HICPs) inflation rate.
- (TGE) Total general government expenditure % of GDP
- (DEB) General government gross debt % of GDP and million EUR.
- (DEF) General government deficit/surplus % of GDP and million EUR.
- (GRD) General expenditure in Research and Development.

C. 28-EU Countries

Belgium (BE), Bulgaria (BG), Czech Republic (CZ), Denmark (DK), Germany (DE), Estonia (EE), Ireland (IE), Greece (EL), Spain (ES), France (FR), Croatia (CR), Italy (IT), Cyprus (CY), Latvia (LV), Lithuania (LT), Luxembourg (LU), Hungary (HU), Malta (MT), Netherlands (NL), Austria (AU), Poland (PL), Portugal (PT), Romania (RO), Slovenia (SI), Slovakia (SK), Finland (FI), Sweden (SE), United Kingdom (UK).

The data table consists from the matrix $A(28,7) \in \mathbb{R}^{28 \times 7}$ objects (lines) and 7 variables (columns) for 2014. With these data we have applied the methods of Data Analysis: MFCA Factor Analysis with Principal Component Analysis (PCA) and Ascending Hierarchical Classification.

III. METHODS OF DATA PROCESSING

A. The Method of MFCA

The method of Multiple Correspondence Analysis requires categorical variables [4]. It also can accept nominal variables, ordinal variables, and/or discretized interval - ratio variables (e.g. quartiles). Since our variable values are continuous, the software package CHIC Analysis is creating discrete categories from continuous variables by transforming each continuous variable to categorical. This approach is a subject of research which needs more study.

Correspondence Analysis is used which was developed by [1] and extended by [7], and many of his colleagues is used for mining these data.

In this paper, the computations of the MFCA method with data the matrix $A(k,n)$ has been done with the software package CHIC Analysis v.1.1.

The results are consisted from tables and graphics of first and second factorial axes, the factorial plane of two axes 1×2 and the corresponding scree diagram showing the factors and the computed eigenvalues.

In the graphical diagrams of the factorial plane of two axes 1×2 , the positions of countries are marked with small triangle and the positions of variables with small square.

B. PCA

The PCA is a variable reduction technique which is used when variables are highly correlated. This method reduces the number of observed variables to a smaller number of principal components which account to most of the variance of the observed variables.

Factor analysis is giving similar results with PCA and the computations are executed with the factor procedure of IBM SPSS 20 software package.

C. Factor Analysis (FA)

The FA is a variable technique which identifies the number of eigenvalues constructs and the underlying factor structure of a set of variables and estimates factors which influence responses on observe variables. Also allows the description and the identification of the number of factors. The purpose of FA is to study and verify patterns in a set of correlation coefficients. If the analysis is considering all the variance

including the correlation coefficients and error then is called PCA.

D. Hierarchical Cluster Analysis (HCA)

HCA is a method of cluster analysis which is used to construct a hierarchy of clusters. The most important result is the graphical picture of the hierarchical tree.

With this analysis, the spatial analysis of the relative position of the 28 EU countries can be studied for 2014. By using similar data of other years, the change of the country position can be studied.

IV. RESULTS

A. Results from CHIC Analysis V 1.1 (MFCA)

The basic theories established by [1] and developed by [7] were applied. The run of the software CHIC Analysis v.1.1 with the data of from 2014 gave the arithmetic and graphical results which are shown in Tables II and III, and in Figs. 1-3. The software package has been produced by [6] and is free for research and educational use.

In the diagram of the factorial plane 1×2 the clusters of the countries separated according to the categorized values of continuous variables are presented. The index 1, 2, 3 in each variable name shows the low, median and high value of the corresponding variables which have been used for the categorization of continuous variables.

The different values of the categorized variables are the reason for the position of each country in the 1×2 factorial plane. In Fig. 1, the formation of groups (clusters) of countries, with near the similar values of the categorized variables, is clearly observed. The countries in the three main clusters which are formed are:

- Cluster A: Low value of categorized variables (nine countries). SK, BG, ES, CR, PT, EL, HU, IT, CY.
- Cluster B: Median value categorized variables (six countries). LT, RO, LV, EE, LU, CZ.
- Cluster C: High value categorized variables (13 countries). DE, NL, DK, BE, UK, SE, FR, FI, SI, AU, PL, MT, IE.

The horizontal axes represent factor 1 and the vertical factor 2. In the factorial plane, the 1×2 positions of the 21 categories of seven variables (7×3) blue squares and 28 countries red triangles are marked. In the neighborhood of each country's mark are present the marks of the categories of related variables. With this approach the clusters are formed.

The selection and presentation of the results by the software CHIC Analysis V.1.1 are given in arithmetic form in Table II. The factors and the corresponding Inertia, % Inertia and the eigenvalues in graphical form are given. From these results, the graphical form of the eigenvalues is given with the scree diagram.

In the Scree plot from CHIC V 1.1 program the X coordinate of Fig. 2 shows the number and the Y coordinate the value of each eigenvalue. In this diagram 14 (2×7) eigenvalues are presented because of the categorization of variables.

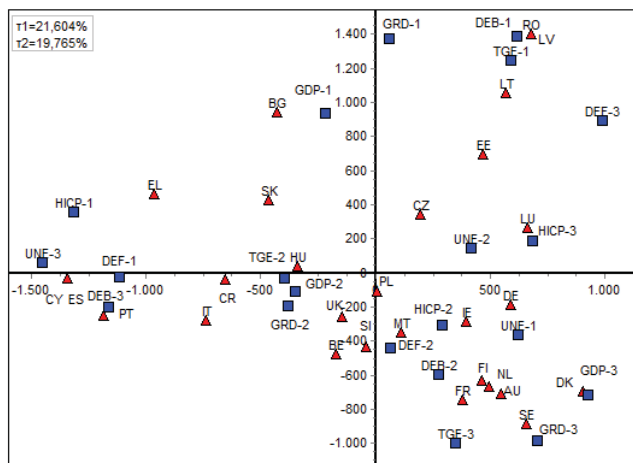


Fig. 1 Factorial plane 1x2 factors for Variables and Countries

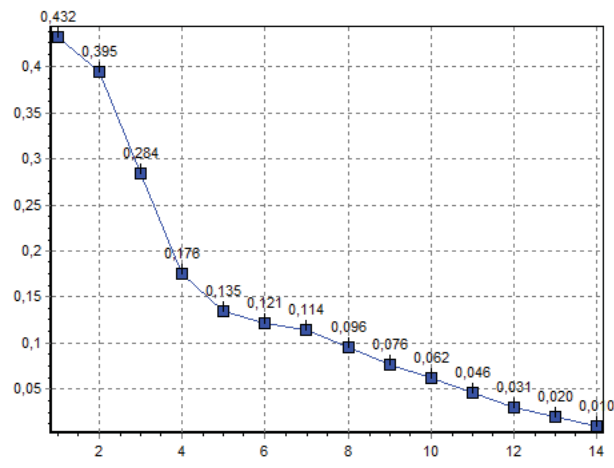


Fig. 2 Scree plot from CHIC Analysis V 1.1. Note (read full stop (.) instead comma (,))

TABLE II
THE FACTORS LOADING AND THE EIGENVALUES

Factor	Inertia	% Inertia	Cum %	
1	0.432	21.604	21.604	*****
2	0.395	19.765	41.370	*****
3	0.284	14.216	55.586	*****
4	0.176	8.796	64.382	****
5	0.135	6.750	71.132	***
6	0.121	6.044	77.176	***
7	0.114	5.715	82.891	***
8	0.096	4.818	87.709	**
9	0.076	3.784	91.492	**
10	0.062	3.089	94.582	**
11	0.046	2.316	96.898	*
12	0.031	1.571	98.469	*
13	0.020	1.008	99.477	*
14	0.010	0.523	100.000	

The scree plot graphs the eigenvalues against the factor number. The four (2x2) only factors are considered and the graph goes flat which means that each successive factor accounts for smaller of the total variance.

TABLE III
THE FIRST 4 FACTORS AND THEIR LOADING

Variable	F1	F2	F3	F4
GDP	0.288	0.353	0.037	0.528
UNE	0.712	0.044	0.697	0.063
HICP	0.607	0.085	0.471	0.013
TGE	0.191	0.584	0.228	0.227
DEB	0.469	0.669	0.039	0.269
DEF	0.559	0.297	0.128	0.111
GRD	0.199	0.734	0.390	0.020

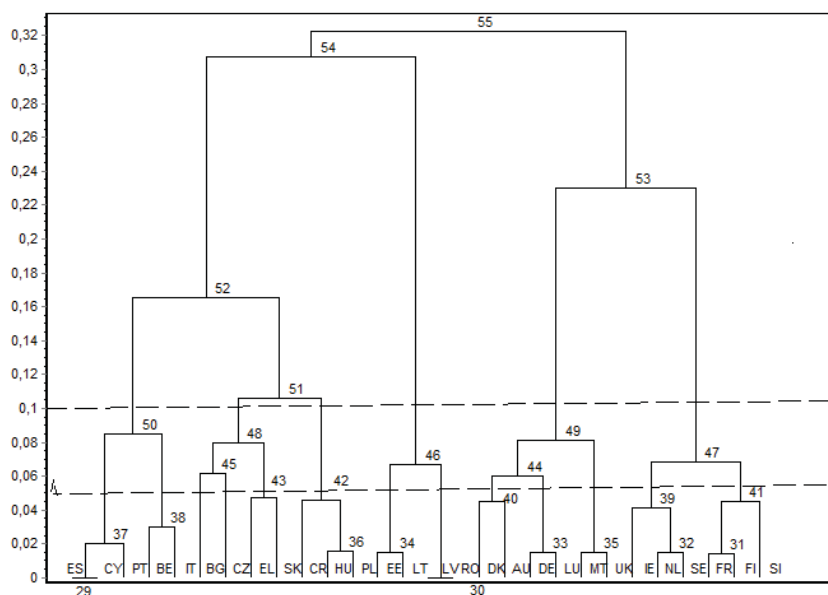


Fig. 3 Tree diagram of HCA for EU 28 countries

B. HCA Results

The results of this procedure give the position of each country in the cluster which is formed according to the values of the variables of the country. The CHIC Analysis program V 1.1, which has used by [3], contains the choice of this procedure, which is applied with the data A(m,k).

In the base of tree diagram, the initial nodes (countries) are shown, as they have classified according to the values of variables, while the higher nodes are presented by numbers. This graphical result can be compared with the result of MFCA method in Fig. 1.

By examining the base line of the tree diagram the classification of countries can be observed and found. On the left the countries ES, CY, PT, BE and on the right the countries SE, FR, FI, SI are present.

C. Results IBM SPSS 20 Statistical Package

The run of the software IBM SPSS 20 Statistical Package [5] with the data of the year 2014 gave with FA and PCA procedures, arithmetic and graphical results. These results are given and are shown for comparison with the results obtained with CHIC V 1.1

Table IV contains the unrotated factor (component) loadings, which are the correlations between the variable and the component. The loadings have possible values between -1 to 1. If the values are less than 0.35 then are not considered, as that means low correlations without meaning.

Table V contains the rotated factor (component), which represent how the variables are weighted for each factor and also the correlation between the variables and the factor. The

loadings have possible values between -1 to 1. If the values are less than 0.35 then are not considered.

The first three columns of the Table VI are:

Component: The number of Components (Factors) is the same as the number of variables used.

TABLE IV
 COMPONENT MATRIX^A

	Component	
	1	2
CDP	0.569	0.469
UNE	-0.882	0.023
HICP	0.734	0.161
TGE	-0.123	0.849
DBT	-0.678	0.539
DEF	0.753	-0.249
GRD	0.360	0.811

^A Extraction Method: PCA. Two components extracted.

TABLE V
 ROTATED COMPONENT MATRIX^A

	Component	
	1	2
CDP	0.430	0.599
UNE	-0.858	-0.204
HICP	0.668	0.344
TGE	-0.337	0.789
DBT	-0.794	0.347
DEF	0.792	-0.047
GRD	0.139	0.877

^AExtraction Method: PCA. Rotation Method: Varimax with Kaiser Normalization. Rotation converged in three iterations.

TABLE VI
 TOTAL VARIANCE EXPLAINED

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.81	40.171	40.171	2.81	40.171	40.171	2.75	39.385	39.385
2	1.97	28.251	68.422	1.97	28.251	68.422	2.03	29.037	68.422
3	.763	10.905	79.327						
4	.557	7.956	87.282						
5	.434	6.206	93.488						
6	.264	3.770	97.258						
7	.192	2.742	100.00						

Extraction Method: PCA

Total: The eigenvalues are the variances of the factors. Their sum equals the number of variables. The value of eigenvalue is considered for the selection of significant factor. A criterion is the value must be greater or equal to 1. In this case, the first two factors are considered significant as they explain the 68.422 % of total variance.

% of Variance: shows the percentage of total variance from each factor.

The eigenvalues shown are computed in the table Extraction Method: PCA. In the diagram is shown that the first two eigenvalues have value greater than 1. The third eigenvalue with value 0.763 is not considered.

The scree plot graphs the eigenvalues against the factor number. The two only factors are considered and the graph

goes flat, which means that each successive factor accounts for less than the total variance.

V. CONCLUSIONS

The conclusions which can be derived are:

- MFCA and PCA methods of variable reduction applied, give interest results concerning the factors, the principal components and the percentage of the total variance which is imposed by each variable.
- The eigenvalues calculated are used in deciding how many factors to extract in overall FA. Usually eigenvalues are taken in consideration if have value greater than 1.

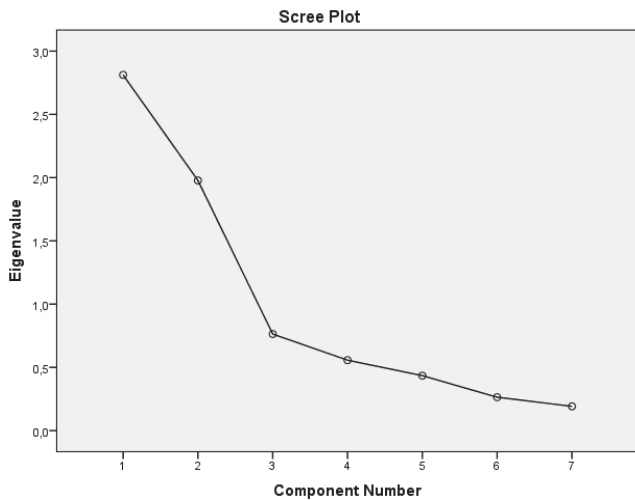


Fig. 4 Scree plot from IBM SPSS 20

- c. Most of the computed results between the two methods agree. Some differences are due to the method of categorization in the application of the software CHIC Analysis, because the correspondence analysis uses only categorical data.
- d. The tree diagram produced from the run of the software CHIC Analysis, using the method of Ascending Hierarchical Classification in good agreement with the results from MFCA method.

REFERENCES

- [1] Benzecri, Jean Paul. (1992). Correspondence Analysis Handbook. New York, Marcel Dekker, Inc.
- [2] <http://ec.europa.eu/eurostat/data/browse-statistics-by-theme> Data for 28 EU countries by theme Feb.2016.
- [3] Fragkaki, M. (2009). Statistical Analysis of Economic and Educational data for EU – 15 countries PhD Thesis University of Macedonia Thessaloniki.
- [4] Greenacre, M. J. (1984). Theory and Applications of Correspondence Analysis, London, Academic Press.
- [5] IBM SPSS 20 Statistical Package
- [6] Markos, Agelos (2010) The CHIC Analysis V.1. 1 Software Program.
- [7] Papadimitriou, Ioannis (2007) in Greek, Data Analysis Edition Typothito, Athens.