AUSTRIAN JOURNAL OF STATISTICS
VOLUME 31 (2002), NUMBER 2&3, 123-130

Principal Components Analysis. Application to the Study of Risk-Factors for Social Dissociation on Territorial Level in Romania

Denis Enachescu[1] and Cornelia Enachescu[2]
[1] University of Bucharest
[2] Center of Mathematical Statistics of the Romanian Academy

Abstract: The aim of the paper is to study simultaneously the whole set of data and point out the risk factors of the social dissociation on territorial level. Therefore, the authors used the Principal Components Analysis – PCA-techniques taking some socio-economic indicators for 1997 as variables and the counties of Romania as individuals. Finally, using the factorial coordinates they classify the Romanian counties in nine different risk-classes.

Keywords: Principal Components Analysis, Cluster Analysis

1 Materials

Grouped into four profiles, eighteen socio-economic indicators have been taken into consideration for each Romanian county. The source of data is "The Romanian Statistical Yearbook", 1998 and the Ministry of Finance. The definitions of the indicators and some elementary statistics of the considered data are given in Table 1.

2 Methods

In our exploratory study the number of variables, i.e. the socio-economic indicators, under consideration is too large to handle. Since it is the deviations in this study which are of interest, a way of reducing the number of variables to be treated is to discard the linear combinations which have small variances and study only those with large variances. In order to do this we have used the following two steps methodology:

- the Principal Components Analysis (PCA) taking the indicators defined in Table 1 as variables and the Romanian counties as actives individuals;
- a cluster analysis applied to the Romanian counties using the above-obtained factorial coordinates.

At the first step, we have considered all the 18 indicators as continuous active variables, the 41 counties of Romania and the Bucharest Municipality as actives individuals and Romania (with the averaged values of the indicators) as illustrative individual. All the active individuals are uniformly weighted. Given the heterogeneity

of the data we do a PCA on the centered and reduced table (i.e. based on the correlation matrix) (see Lebart et al., 1995).

Table 1. Descriptive statistics of the socio-economic data

Variable			Mean	Standard deviation	Minimum	Maximum
Demographic Profile	1. URBR - Urban population (%-from the population at July 1, 1997)		50.04	15.34	7.10	100.00
	2. GROW - Average annual growing rate of the population (%-from the population at census 1992)		-0.25	0.39	-2.31	0.25
	3. YNG - Age group 0-19 years (%-from the population at July 1, 1997)		27.97	2.07	23.75	32.00
	4. OLD - Age group 65 years and over (%-from the population at July 1, 1997)		12.70	1.89	8.95	18.34
	5. FERT - General fertility rate (live-births per 1000 women)		42.30	6.46	25.20	58.40
	6. DDR - Demographic dependency ratio (Inhabitants under 15 years and over 64 years at 100 persons between 15 and 64 years)		48.09	3.97	40.17	56.57
Economic Profile	7. EMPE - Employment in economy (%-from the population between 15 and 64 years)		59.40	3.85	50.59	69.89
	8. ECDR - Economic dependency ratio (Inactive and unemployed inhabitants at 1000 employed)		1502.57	161.90	1125.00	1883.00
	9. UEPL - Unemployment rate (%)		9.09	2.92	4.00	15.00
	Structure of the employed population (%)	10. EMPP - Primary sector	41.47	12.34	0.99	61.88
		11. EMPS - Secondary sector	30.38	8.20	17.53	51.96
		12. EMPT - Tertiary sector	28.15	6.98	18.41	55.35
	13. SAL - Average net nominal monthly salary (in thousands lei per employee)		605.27	64.72	490.97	748.72
Fiscal Profile	14. FISC - Direct taxes (in thousands lei per inhabitant)		52.95	24.94	26.80	178.05
Education, Health, Culture and Justice Profile	15. EDEX - Education expenditures (in thousands lei per school population of all levels in 1997/98 school year)		227.37	40.19	170.97	343.19
	16. HEEX - Health expenditures (in thousands lei per inhabitant)		54.77	12.91	26.22	97.19
	17. CUEX - Culture expenditures (in thousands lei per inhabitant)		21.90	9.86	2.00	52.08
	18. CRIM - Criminality rate (persons definitively convicted per 100000 inhabitants)		508.71	143.53	32.00	771.00

At the second step, we have used the ascendant hierarchical cluster techniques –the Ward method- (see Lebart et al., 1995) considering the factorial coordinates of the individuals (obtained at the first step). The initial clusters are consolidated iteratively by the *k*-means method. Automatically the best three partitions are listed. The optimality criterion used to establish the partitions to be listed is the maximum of the ratio between the interclass variance and the total variance.

The calculus was done using the SPAD ver.3.5 software (Système Portable d'Analyse des Données) with the default setting of the parameters.

3 Results

The PCA supply the following types of results:

- the histogram of the eigenvalues of the variance-covariance matrix of the active variables in order to describe the quality of the factorial representation and render evident the most interesting axes (i.e. factors explaining the biggest percents of the whole variance). For our study this results are listed in Table 2;

Table 2. The quality of the factorial representation

Factor	Eigenvalue	% - from the total inertia	%- cumulative
1	6.4797	36.00	36.00
2	3.0997	17.22	**53.22** [1)]
3	2.1206	11.78	65.00
4	1.3640	7.58	72.58
5	1.2915	7.17	79.75
6	0.9288	5.16	84.91

[1)] The principal factorial plane explain only 53.22% from the total inertia (i.e. from the whole information)

- the correlation of the variables with the factors in order to "explain" the principal components by the initial variables. For our study this results are listed in Table 3. Following this results the interpretation of the most important axis is:
- the first axis explains 36% from the total information (quantified by the total inertia and equal with the total variance) (see Table 2). It sets off the variables specific to the urban world (i.e. a high percent of the urban population –URBR-, a direct taxes over the general mean –FISC-, a high percent of the employed population in the secondary –EMPS- and tertiary sectors –EMPT- with big health expenditures –HEEX- and salaries over the country-mean -SAL) with the variables specific to the rural world (i.e. a high percent of the employed population in the primary sector –EMPP- with a big demographic dependency ratio –DDR- and a big fertility rate -FERT);
- the second axis explains 17% from the total information. It sets off the variables specific to a young and poor world (characterized by a high percent of the group 0-19 years in the population, a high annual rate of growth of population but also with a high unemployment rate) and consequently –youth and poverty-by a high rate of the criminality with the variables specific to an old world (characterized by a high percent of the group 65 years and over in the population and high education expenditures);

Table 3. The correlation of the variables with the first three factors

Variable / Factor	Correlation variable-factor 1	2	3
1. URBR - Urban Pop.	**-0.84** [1)]	-0.27	0.00
2. GROW – Growing Pop. Rate	-0.11	-0.77 [2)]	-0.22
3. YNG – Young Group	0.30	-0.77	0.04
4. OLD – Old Group	0.49	0.67	-0.08
5. FERT – Fertility rate	**0.81**	-0.37	-0.07
6. DDR – Demo. depend. Ratio	**0.91**	0.06	-0.08
7. EMPE - Employment in Ec.	0.12	0.27	*-0.93* [3)]
8. ECDR – Ec. depend. Ratio	0.24	-0.23	*0.90*
9. UEPL - Unemployment Rate	0.26	-0.61	0.18
10. EMPP - Primary S.	**0.94**	0.05	0.05
11. EMPS - Secondary S.	**-0.75**	-0.16	-0.09
12. EMPT - Tertiary S.	**-0.79**	0.09	0.02
13. SAL – Salary	**-0.64**	-0.04	0.06
14. FISC – Direct Taxes	**-0.81**	0.26	0.26
15. EDEX – Ed. Exp.	0.09	0.40	0.11
16. HEEX - Health Exp.	**-0.67**	0.03	-0.11
17. CUEX - Culture Exp.	-0.42	-0.45	-0.27
18. CRIM – Crim. Rate	0.16	-0.55	-0.41

[1)] The most correlated variables with the first factor
[2)] The most correlated variables with the second factor
[3)] The most correlated variables with the third factor

- the third axis explains 12% from the total information. It sets off the variable specific to an employed world –EMPE- with the variables specific to an inactive and unemployed world –ECDR.
- a graphical representation of the Romanian counties in the principal factorial plane (the plane generated by the first two principal components which synthesizes the maximum information) (see Figure 1).

The factorial plane is an approximative euclidian image of the cloud of individuals because of the deformation caused by the varimax projection without rotated factors (involvd by the PCA), consequently we are interested not only of the general quality representation (given, in Table 2, by the cumulated percent of the explained variance) but also the individual quality representation. This variable is mesured by the squared cosinus of the angle between the individual and its projection on the factorial plane; so if the cosinus is near zero then the angle is important and the point is far off the projection plane, otherwise if the cosinus is near one the angle is unimportant and the point is good represented by its projection. Hence 20 of the most representatives counties (bolded in Table 4) are good represented by the projection (squared cosinus between 0.51 and 0.82).

Correlating the above results, Figure 2 can be used to explain the factorial axis from the counties point of view. So, the first axis sets off the developed counties (Bucharest-B, Brasov-BV, Cluj-CJ, Constanta-CT, Ilfov-SAI) with the less developed counties (Botosani-BT, Giurgiu-GR, Teleorman-TR, Vaslui-VS), the second axis sets off

counties with aged population (Arad-AR, Giurgiu-GR, Ilfov-SAI, Teleorman-TR) with counties with young population (Bacau-BC, Galati-GL, Iasi-IS, Vaslui-VS).

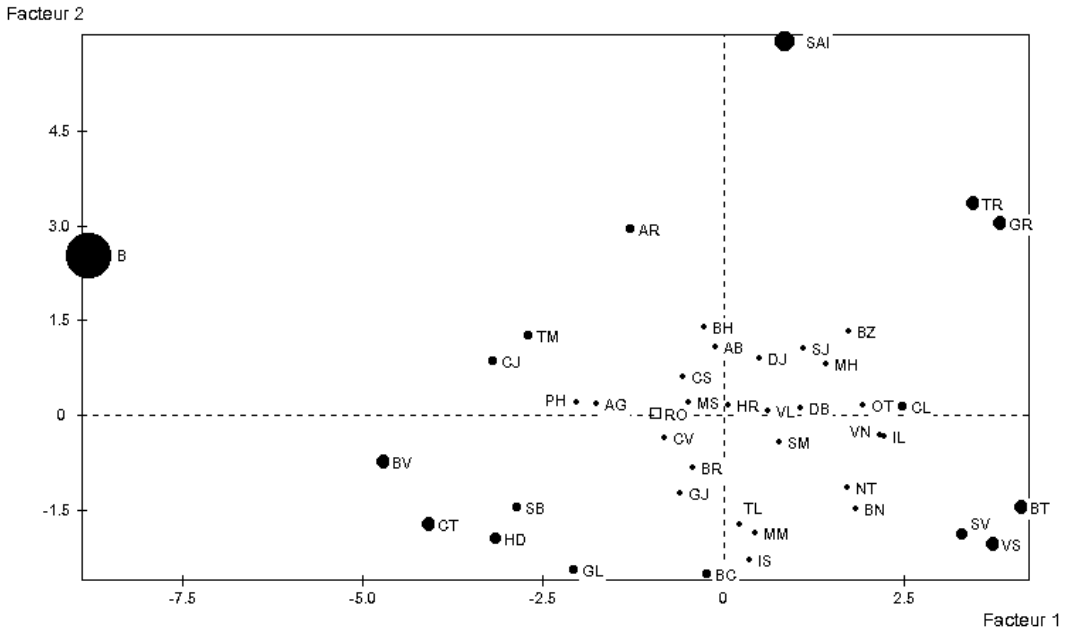

Figure 1. Representation of the counties in the principal factorial plane (proportionally with the contributions to the factors)

Table 4. The cluster analysis results

Distance to the gravity center of the class	County	Class	Characteristic variable	Mean		Standard deviation	
				of the class	of the whole population	of the class	of the whole population
1	2	3	4	5	6	7	8
2.530	AB	CLASS 1 (effectives = 7)	EMPE - Employment in Ec.	63.82	59.40	1.36	3.85
3.145	HR		EDEX - Ed. Exp	269.76	227.37	42.13	40.19
4.422	SJ		ECDR - Ec. depend. Ratio	1331.14	1505.57	56.98	161.90
5.956	BH						
6.777	VL						
9.999	GJ						
13.702	CV						
4.040	**CJ**	CLASS 2 (effectives = 6)	HEEX - Health Exp.	67.79	54.77	10.30	12.91
4.385	MS		YNG - Young Group	26.10	27.97	0.64	2.07
5.220	CS						
6.173	**PH**						
6.748	**TM**						
14.967	AR						

4.596	**GL**	CLASS 3 (effectives = 6)	SAL – Salary	694.90	605.27	52.36	64.72
5.102	**SB**		CUEX – Culture Exp.	34.46	21.90	9.03	9.86
6.955	**HD**		EMPS - Secondary S.	40.56	30.38	8.87	8.20
7.738	**BV**		URBR - Urban Pop.	67.22	50.04	10.12	15.34
8.579	AG						
14.651	**CT**		EDEX - Ed. Exp.	188.14	227.37	13.74	40.19
			EMPP – Primary S.	27.21	41.47	5.87	12.34
			OLD - Old Group.	10.46	12.70	0.92	1.89
			DDR - Demo. depend. Ratio	43.10	48.09	1.72	3.97
0	**B**	CLASS 4 (effectives = 1)					
0	**SAI**	CLASS 5 (effectives = 1)					
1.970	OT	CLASS 6 (effectives = 9)	CRIM - Crim. Rate	632.67	508.71	87.97	143.53
2.614	VN		EMPP – Primary S.	51.64	41.47	4.56	12.34
3.317	DJ						
3.816	DB						
3.890	MH						
4.171	**IL**						
4.690	**BZ**						
6.676	SM						
8.147	CL						
3.335	**GR**	CLASS 7 (effectives = 2)	OLD - Old Group	17.99	12.70	0.35	1.89
3.335	**TR**		DDR - Demo. depend. Ratio	55.24	48.09	1.03	3.97
			YNG - Young Group.	24.51	27.97	0.74	2.07
2.804	**MM**	CLASS 8 (effectives = 6)	ECDR - Ec. depend. Ratio	1704.50	1502.57	127.88	161.90
4.067	**BC**		UEPL - Unemployment Rate	12.00	9.09	1.70	2.92
4.808	TL						
5.307	NT		EMPE - Employment in Ec.	54.51	59.40	2.83	3.85
5.415	IS						
9.391	BR						
0.670	**VS**	CLASS 9 (effevtives = 4)	FERT – Fertility Rate	55.47	42.30	2.37	6.46
0.823	**SV**		YNG - Young Group	31.44	27.97	0.48	2.07
2.342	**BT**		DDR - Demo. depend. Ratio	54.37	48.09	2.05	3.97
3.071	**BN**		UEPL - Unemployment Rate	12.58	9.09	0.95	2.92
			SAL – Salary	530.00	605.27	23.68	64.72

It is easy to see in Figure 2 that Bucharest and SAI have an important contributions to the first and second factor, respectively. Applying the distance to the centroid test (see Enachescu & Enachescu, 2000) Bucharest and SAI may be considered as outliers. Hence we have carry out the PCA without Bucharest and SAI and have obtained, exception a rotation, the same configuration of the principal factorial plane as in Figure 1.

The cluster analysis supply two partitions of the counties: with 6, respectively 9 classes. The most homogeneous partition is the 9-class partition with a final ratio between the interclass variance and the total variance equal with 0.6992. This partition is listed in Table 4 and graphical represented in Figure 2. Each class is specified by two modalities: by the constitutives counties ordered ascendently with the distance to the gravity center of the class and by the characteristic variables ordered ascendently with the 'test-values' (see Lebart et al., 1995) (first the variables with the mean of the class

greater then the general mean, afterthat the variables with the mean of the class less then the general mean).

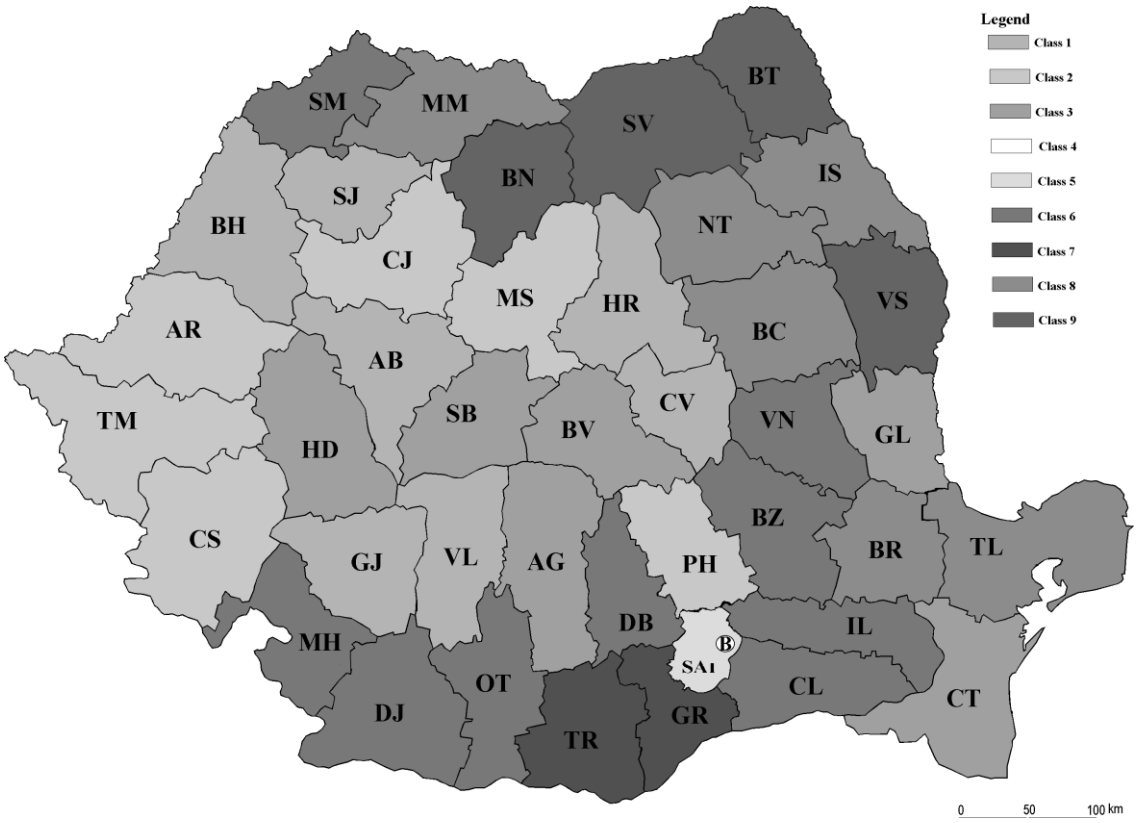

Figure 2. The classification of the Romanian counties following the socio-economic characteristics, in 1997

In "The Green Charter. The Regional Development Policy in Romania" published by the Romanian Government & European Commission, in 1997, are proposed eight regions. These regions, clusters of neighbor counties, are obtained using only a few macro-economic variables and observing the traditional relationships between the historical regions of Romania. In our demarche, if we merge class 4 and 5 (Bucharest-B and SAI), as in the above-cited charter, we obtain eight regions too. The both classifications coincide in proportions that vary between 33% and 50%. Our classification point out better the regional disparities in Romania but it is difficult to apply it in a territorial development program because not all the regions contain only neighbor counties.

Acknowledgement

The authors thank the referee for his helpful comments that led to an improvement of the original manuscript.

References

C. Enachescu and D. Enachescu. Some Simple Rules for Interpreting Outputs of Principal Components and Correspondence Analysis. *Anal. Univ. Buc., Informatics*, XLIX:3-8, 2000.

L. Lebart, A. Morineau and M. Piron. *Statistique exploratoire multidimensionnelle*. Dunod, Paris, 1995.

Author's address:

Assoc.-Prof. Dr. Denis Enachescu
University of Bucharest
Faculty of Mathematics
Str. Academiei 14,
Bucharest
Romania

Tel. +40 1 211 21 84
E-mail: denaches@k.ro