ACADEMIA DE STUDII ECONIMICE DIN BUCUREȘTI

MASTERATUL DE CIBERNETICĂ ȘI ECONOMIE CANTITATIVĂ

DATA MINING

-PROIECT-

CORNEA COSTEL-CRISTIAN

București

2016

Cuprins

Introducere

Pentru aplicarea tehnicilor de analiza dobandite la disciplina Data Mining, am ales utilizarea unor date la nivel macroeconomic, ale unor țări care au fost selectate fără un criteriu prestabilit, aceste date privind indicatori precum:

- Produsul Intern Brut per capita, în prețuri comparabile, în dolari americani
- Consumul de energie electrică, in kWh per capita
- Cheltuielile cu sănătatea, în dolari per capita
- Cheltuielile cu educația, în milioane dolari americani
- Importurile de bunuri și servicii, in milioane dolari americani
- Exporturile in bunuri si servicii, in milioane dolari americani
- Rata natalității, raportată la 1000 locuitori

Sursa datelor este reprezentata de baza de date a site-ului worldbank.org.

I. Analiza descriptivă a datelor

Pentru inceput am realizat o analiza descriptiva a datelor selectate pentru cele 61 de observatii (tari).

Utilizand procedura univariate implementata in SAS, vom obtine pentru fiecare variabila tabele privind momentele, masurile statistice, quantilele dar si valorile extreme.

Astfel, pentru prima variabila, PIB-ul, codificata GDP_PCAP, avem urmatorul output:

Moments			
N	61	Sum Weights	61
Mean	18783.8386	Sum Observations	1145814.16
Std Deviation	19966.7564	Variance	398671359
Skewness	1.08086558	Kurtosis	0.40450104
Uncorrected SS	4.54431E10	Corrected SS	2.39203E10
Coeff Variation	106.297529	Std Error Mean	2556.48119

Astfel, in tabelul de mai sus observam ca avem numarul de observatii 61, media este 18783,8386 , abaterea standar (sau abaterea medie patratica) 19966,7564 , varianta 398671359, suma observatiilor, coeficientii de aplatizare, skewness=1,08 arata asimetria distributiei datelor, fiind alungita catre stanga, majoritatea valorilor extreme fiind situate la dreapta, in zona valorilor foarte mari. Kurtosis=0,40 arata gradul de aplatizare, aceasta fiind platikurtica.

In tabelul de mai jos avem media, mediana si modul, observam ca valoare modala nu avem,iar media este mai mare ca mediana, acest aspect fiind anticipat din interpretarea coeficientului de asimetrie (skewness). De asemenea se regasesc si valorile abaterii standard, a a variantei, dar si a amplitudinii si a dimensiunii intervalului interquartilic.

Basic Statistical Measures			
Location		Variability	
Mean	18783.84	Std Deviation	19967
Median	9029.73	Variance	398671359
Mode	.	Range	81591
		Interquartile Range	33376

Interpretand quantilele, in functie de nivel, putem spune pentru Q2, care este reprezentata de mediana ca 50% din date sunt situate sub aceasta valoare, iar restul de 50% deasupra acestei valori. Altfel spus, aceasta valoare imparte setul de date in doua parti, in functie de pozitia valorilor fata de acest punct.

In cazul lui Q3, 75% din valori sunt situate sub aceasta valoare, diferenta de 25% fiind valori mai mari ca 36203,43.

De asemenea putem observa si valorile minime, respectiv maxima din setul de date.

Quantiles (Definition 5)	
Level	Quantile
100% Max	81852.976
99%	81852.976
95%	57279.668
90%	45727.098
75% Q3	36203.430
50% Median	9029.734
25% Q1	2827.126
10%	1217.746
5%	942.524
1%	261.946
0% Min	261.946

In tabelul privind valorile extreme se regasesc valorile atat din categoria celor mai mici, cat si a celor mai mari, alaturi de numarul de ordine (al observatiei), din cadrul reprezentarii tabelata a datelor.

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
261.946	17	48143.8	23
384.772	45	57279.7	32
563.435	5	58534.4	57
942.524	12	64612.6	47
1050.201	43	81853.0	42

Pentru restul de 6 variabile am sintetizat informatiile in cadrul urmatoarelor tabele, interpretarile valorilor ramanand aceleasi.

USE_ELEC					
Extreme Observations				Mean	STD
Lowest		Highest			
Value	Obs	Value	Obs		
105.316	17	15530.1	42	5778.100	7604
105.500	45	15738.4	26		
171.583	18	16473.2	13		
255.527	12	23173.6	47		
258.618	5	52373.9	32		

HEALTH_XPD					
Extreme Observations				Mean	STD
Lowest		Highest			
Value	Obs	Value	Obs		
24.0148	17	5117.94	46	4628.635	21119
66.7874	5	5672.87	57		
85.4111	45	6020.02	42		
106.7800	18	6105.88	47		
120.0849	12	166300.61	30		

ADJ_AEDU					
Extreme Observations				Mean	STD
Lowest		Highest			
Value	Obs	Value	Obs		
237.526	1	135224	10	30463.06	47324
270.106	18	143497	61		
550.217	17	149591	27		
559.784	43	179249	28		
742.835	12	193961	38		

EXP_GNFS					
Extreme Observations				Mean	STD
Lowest		Highest			
Value	Obs	Value	Obs		
1063.36	45	654937	27	197500.7	323305
1425.81	1	714898	61		
2222.18	43	744943	38		
2776.75	17	1432977	28		
3362.89	12	1748396	15		

IMP_GNFS					
Extreme Observations				Mean	STD
Lowest		Highest			
Value	Obs	Value	Obs		
2395.89	1	627835	38	188191.6	281219
3742.15	43	699722	27		
3913.01	45	725047	61		
4294.58	8	1234095	28		
5238.54	32	1413668	15		

CBRT_RATE					
Extreme Observations				Mean	STD
Lowest		Highest			
Value	Obs	Value	Obs		
8.1	28	26.201	8	15.45384	7.70389
8.3	38	28.317	39		
8.8	31	38.186	12		
9.0	52	38.373	18		
9.0	37	43.585	17		

II. Analiza Componentelor Principale

Inainte de a trece la analiza componentelor principale, deoarece unitatea de masura a indicatorilor selectati in analiza ca variabile nu este aceeasi si avem dolari, milioane dolari si o rata, iar acest tip de analiza este foarte sensibil la unitatea de masura, am ales utilizarea procedurii de standardizare a datelor, pentru care am setat media 0 si abterea standard 1, conform legii distributiilor normale.

In continuare toate analizele si operatiile au fost realizate pe setul de date standardizat.

Selectia de mai jos din output-ul procedurii PRINCOMP cuprinde numarul de observatii si numarul de variabile, media si abaterea standard pentru fiecare variabila, ale caror valori sunt 0 si respectiv 1, din standardizare, iar mai apoi observam matricea de varianta-covarianta, unde pe diagonala principala se afla dispersia fiecarei variabile, iar restul elementelor aratand covarianta dintre fiecare 2 variabile. Deoarece am transformat datele, matricea de covarianta este identica, cu matricea de corelatie a setului de date. Astfel, valorile din matrice reprezinta totodata si valorile coeficientului de corelatie Pearson.

Varianta totala de asemenea este egala cu 7, numarul variabilelor considerate in analiza.

Rezultate Analiza Componentelor Principale

The PRINCOMP Procedure

Observations	61
Variables	7

Simple Statistics

	GDP_PCAP	USE_ELEC	HEALTH_XPD	ADJ_AEDU	EXP_GNFS	IMP_GNFS	CBRT_RATE
Mean	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
StD	1.000000000	1.000000000	1.000000000	1.000000000	1.000000000	1.000000000	1.000000000

Covariance Matrix

		GDP_PCAP	USE_ELEC	HEALTH_XPD	ADJ_AEDU	EXP_GNFS	IMP_GNFS	CBRT_RATE
GDP_PCAP	GDP_PCAP	1.000000000	0.684060312	0.169655839	0.310863879	0.277828555	0.288143756	-0.434289886
USE_ELEC	USE_ELEC	0.684060312	1.000000000	0.051522234	0.090072989	0.066316514	0.066453671	-0.304927780
HEALTH_XPD	HEALTH_XPD	0.169655839	0.051522234	1.000000000	-0.031469809	0.136011434	0.149694471	-0.075921565
ADJ_AEDU	ADJ_AEDU	0.310863879	0.090072989	-0.031469809	1.000000000	0.807464954	0.848641506	-0.301529175
EXP_GNFS	EXP_GNFS	0.277828555	0.066316514	0.136011434	0.807464954	1.000000000	0.990792531	-0.308278130
IMP_GNFS	IMP_GNFS	0.288143756	0.066453671	0.149694471	0.848641506	0.990792531	1.000000000	-0.315506189
CBRT_RATE	CBRT_RATE	-0.434289886	-0.304927780	-0.075921565	-0.301529175	-0.308278130	-0.315506189	1.000000000

Total Variance	7

In figura de mai jos, selectie de output, se regasesc 2 tabele, unul ce cuprinde valorile proprii, iar altul vectorii proprii asociati fiecarei componente principale.

Valorile proprii sunt ordonate descrescator, acestea indicand de altfel si cat de mare este nivelul informatiei sintetizat de catre fiecare componenta principala. In coloana Proportion regasim procentual cantitatea de informatie din date explicata de ficare componenta in parte. Astfel, de exemplu, componenta 1 explica aproximativ 45.65% din varianta datelor selectionate.

Pentru a stabili numarul de componente principale retinute in analiza, ne bazam pe 2 criterii si anume pe criteriul suprafetei de acoperire si criteriul lui Kaiser, deoarece avem datele

standardizate. Criteriul suprafetei de acoperire presupune sa avem explicata aproximativ 70-77% minim din varianta datelor si pentru aceasta ne uitam in coloana Cumulative din tabelul cu valori proprii si observam in dreptul celei de-a treia valori un coeficient de 0,8336 (83,36%), cee ace inseamna ca vom pastra 3 componente principale.

De asemenea, criterial lui Kaiser spune ca vom pastra in analiza componentele principale pentru care valorile proprii au o valoare mai mare sau egala cu 1, acest lucru pe datele noastre fiind valabil in cazul celor 3 componente selectionate si pe baza primului criteriu.

In cel de-al doilea tabel din output se afla vectorii proprii asociati fiecarei valori proprii ai fiecarei component principale. Astfel, euațiile componentelor principale sunt urmatoarele(combinatii liniare ale variabilelor initiale, formula de baza fiind $w_i=\alpha^{(j)}*x$):

w_1=0.329*GDP_PCAP + 0.197*USE_ELEC + 0.1*HEALTH_XPD + 0.479*ADJ_AEDU + 0.504*EXP_GNFS +0.513*IMP_GNFS -0.306*CBRT_RATE

w_2=0.536*GDP_PCAP + 0.63*USE_ELEC + 0.114*HEALTH_XPD - 0.244*ADJ_AEDU - 0.279*EXP_GNFS -0.28*IMP_GNFS -0.29*CBRT_RATE

w_3=-0.018*GDP_PCAP - 0.142*USE_ELEC + 0.968*HEALTH_XPD - 0.169*ADJ_AEDU + 0.043*EXP_GNFS +0.049*IMP_GNFS +0.093*CBRT_RATE

Dacă ar fi să analizăm importanța pe care o are fiecare variabila in cadrul fiecărei componente principale, am observa in primul rand ca a treia componenta principal are o valoare de 0,96 pentru coeficientul privind cheltuielile cu sanatatea, fiind si singurul coeficient mai mare de 0,4, ca valoare. Aceasta component ar putea arata important ape care o acorda statele selectionate cheltuielilor cu sanatatea, sau altfel spus cat investeste fiecare in sanatatea populatiei sale, de la bugetul de stat.

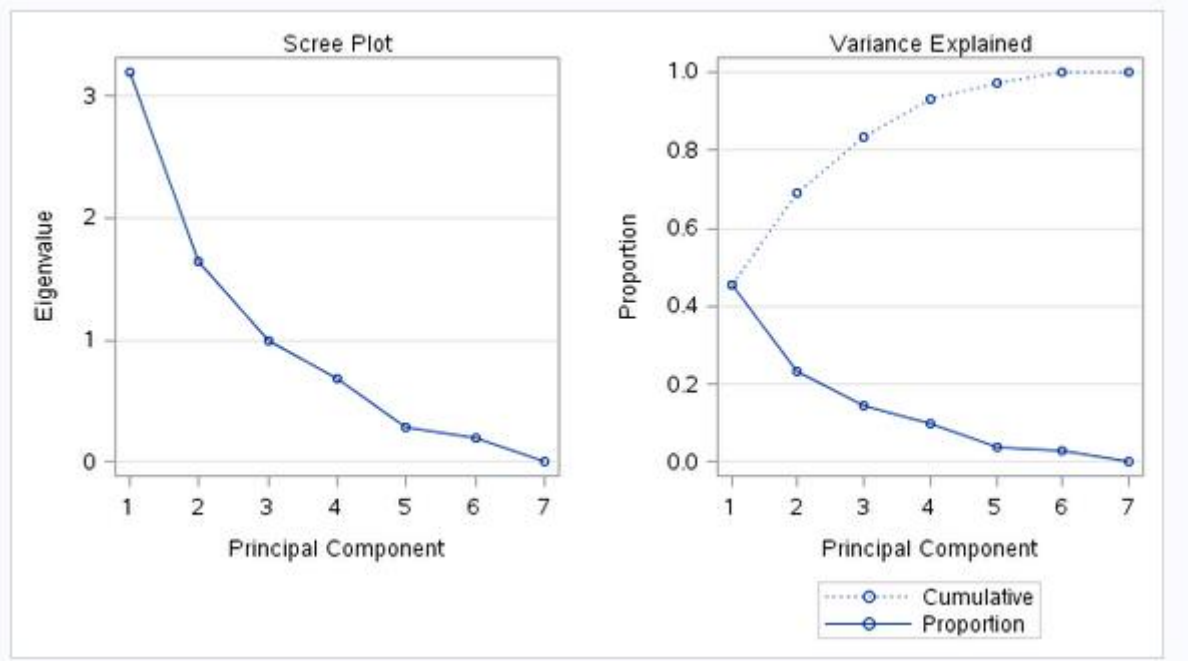
In cazul celei de-a doua componente principale, valorile mai mari decat valoarea de referinta 0.4 sunt pentru PIB-ul per capita si consumul de energie electrica, cee ace ar putea indica un anumit nivel al dezvoltarii si de asemenea al utilizarii tehnologiei.

Iar in cazul primei componente principale, indicatorii ai caror coeficienti sunt mai mari decat 0.4 sunt PIB per capita, cheltuielile cu educatia si importurile respective exporturile de bunuri si servicii. Astfel aceasta componenta ar putea avea legatura cu balanta de plati cu exteriorul (curba BP din modelul IS-LM-BP), dar de asemenea observam acel indicator al cheltuielilor cu educatia, a carui legatura inca nu o putem justifica in mod clar, ci doar putem lansa anumite ipoteze.

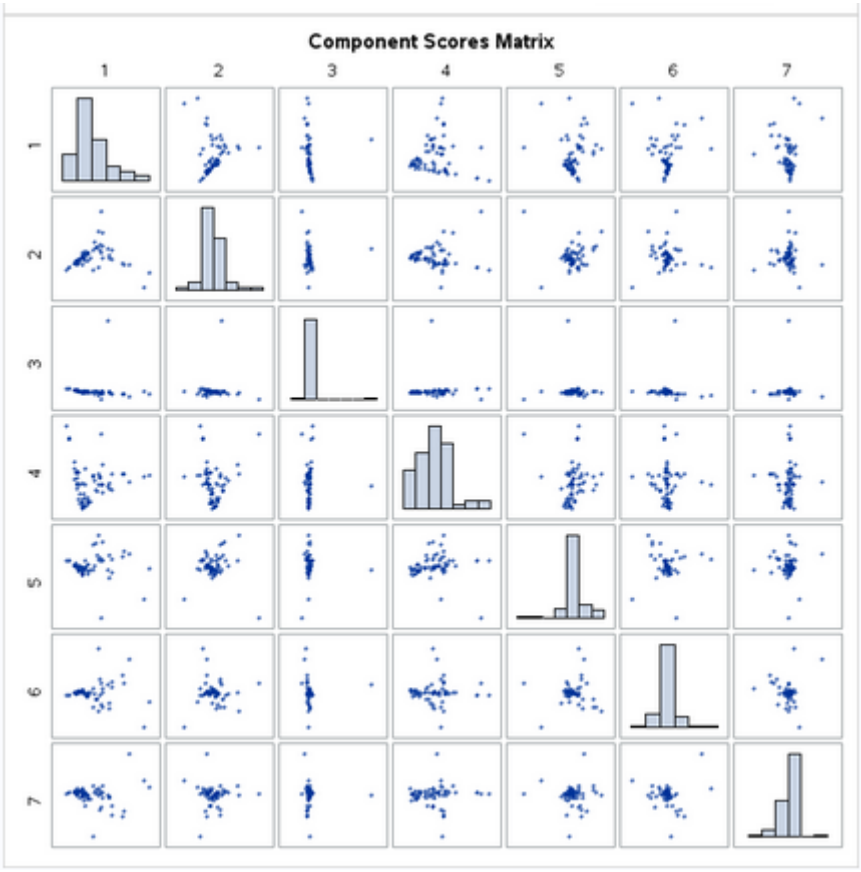
Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	3.19547158	1.55692793	0.4565	0.4565
2	1.63854364	0.63740050	0.2341	0.6906
3	1.00114314	0.31640607	0.1430	0.8336
4	0.68473707	0.40298712	0.0978	0.9314
5	0.28174995	0.08913138	0.0402	0.9717
6	0.19261858	0.18688254	0.0275	0.9992
7	0.00573604		0.0008	1.0000

Eigenvectors								
		Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7
GDP_PCAP	GDP_PCAP	0.329235	0.536760	-.018341	0.204082	0.706840	-.248745	-.003273
USE_ELEC	USE_ELEC	0.197978	0.630541	-.142372	0.359908	-.623749	0.156033	-.003262
HEALTH_XPD	HEALTH_XPD	0.100096	0.114827	0.968310	-.018172	-.035803	0.191865	0.027320
ADJ_AEDU	ADJ_AEDU	0.479533	-.244265	-.169707	0.109370	0.202951	0.785123	0.109602
EXP_GNFS	EXP_GNFS	0.504915	-.279779	0.043631	0.084095	-.196993	-.438030	0.653554
IMP_GNFS	IMP_GNFS	0.513312	-.280245	0.049055	0.083998	-.144238	-.260045	-.748387
CBRT_RATE	CBRT_RATE	-.306610	-.290935	0.093050	0.895766	0.096089	-.032482	-.001949

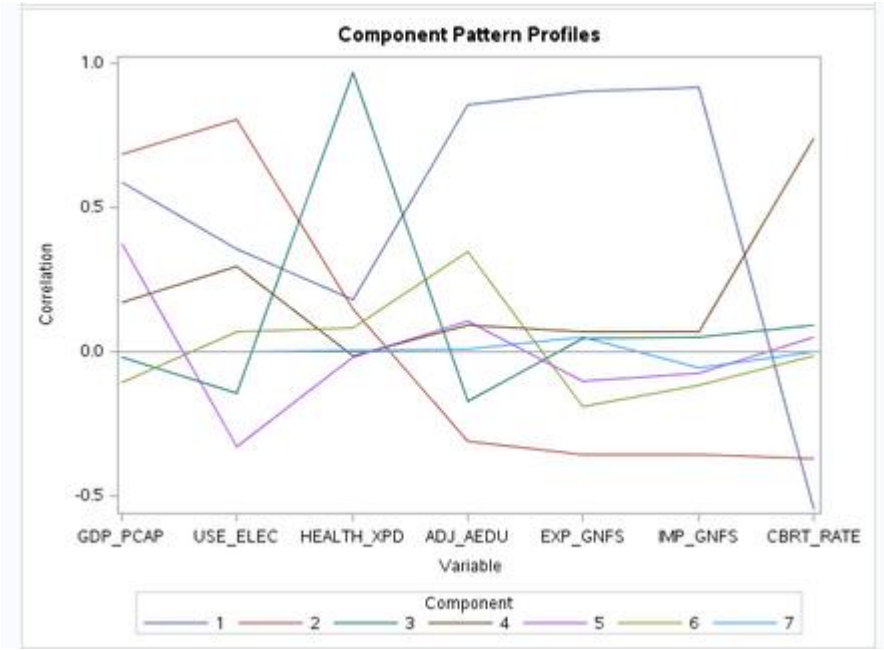
In figura urmatoare se regasesc 2 screeploturi, in care in primul sunt reprezentate valorile proprii in functie de fiecare componenta principala careia ii este asociata, iar in cel de-al doilea explicarea variantei, se regaseste proportia fiecarei componente principale, dar si proportia cumulata. Astfel, de exemplu, prima componenta principala are valoarea proprie mai mare putin fata de 3 si singura explica aproximativ 40-45% din varianta datelor; luand inc alcul si cea de-a doua componenta, impreuna cu prima explica 65-70% din varianta totala. De asemenea si prin metoda grafica se poate stabili numarul componentelor principale.



In figura de mai jos sunt reprezentate scorurile asociate tuturor componentelor principale, nu doar celor 3 pe care le consideram retinute in analiza.

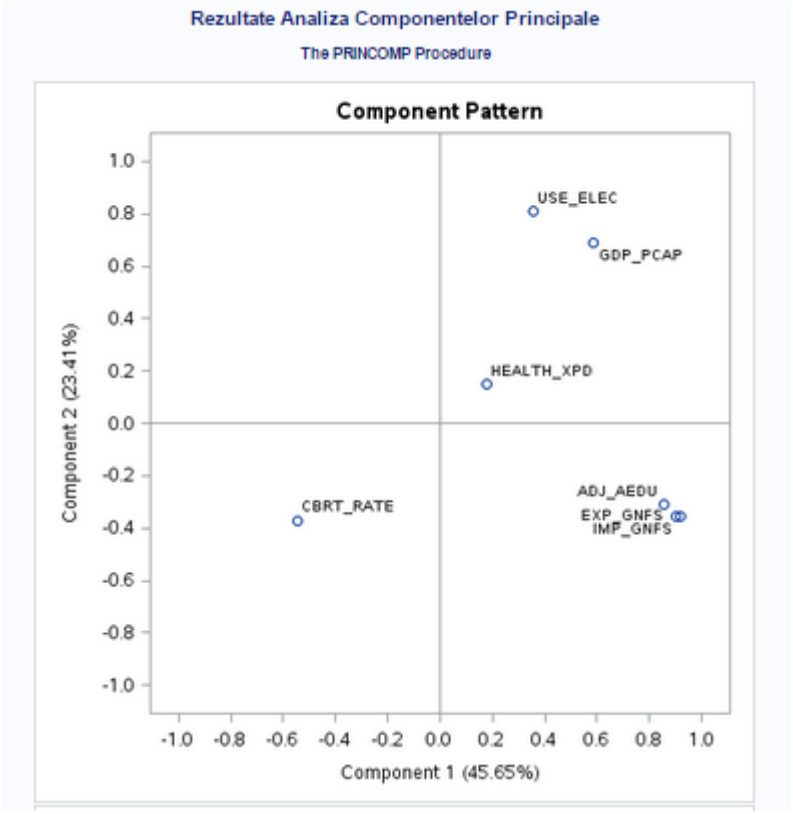


Elementele fiecarui vector propriu, deoarece reprezinta cate un coeficient pentru fiecare dintre cele 7 variabile pot fi reprezentate intr-un sistem de axe si astfel sa se obtina reprezentarea din figura urmatoare.

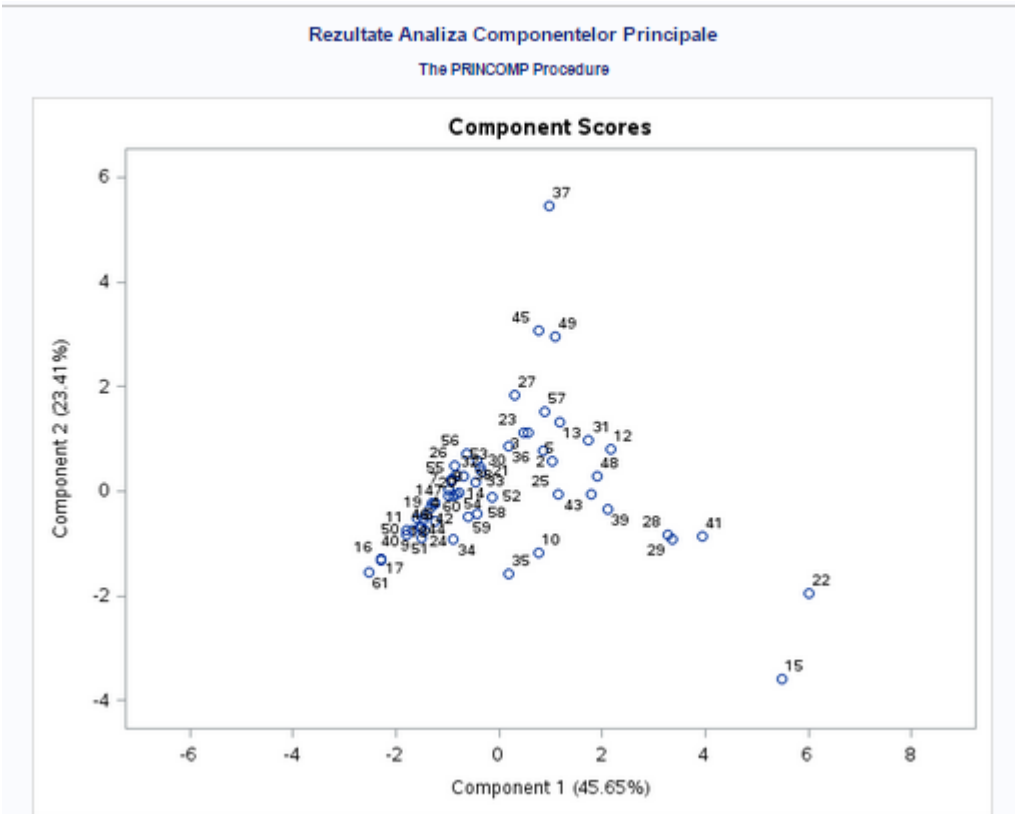


In graficul urmator, un grafic de tip scatterplot sunt reprezentate variabilele in functie de primele doua component principale. Astfel observam ca in raport cu prima component principal sunt correlate pozitiv toate variabilele, mai putin rata natalitatii, codificata cu CBRT_RATE.

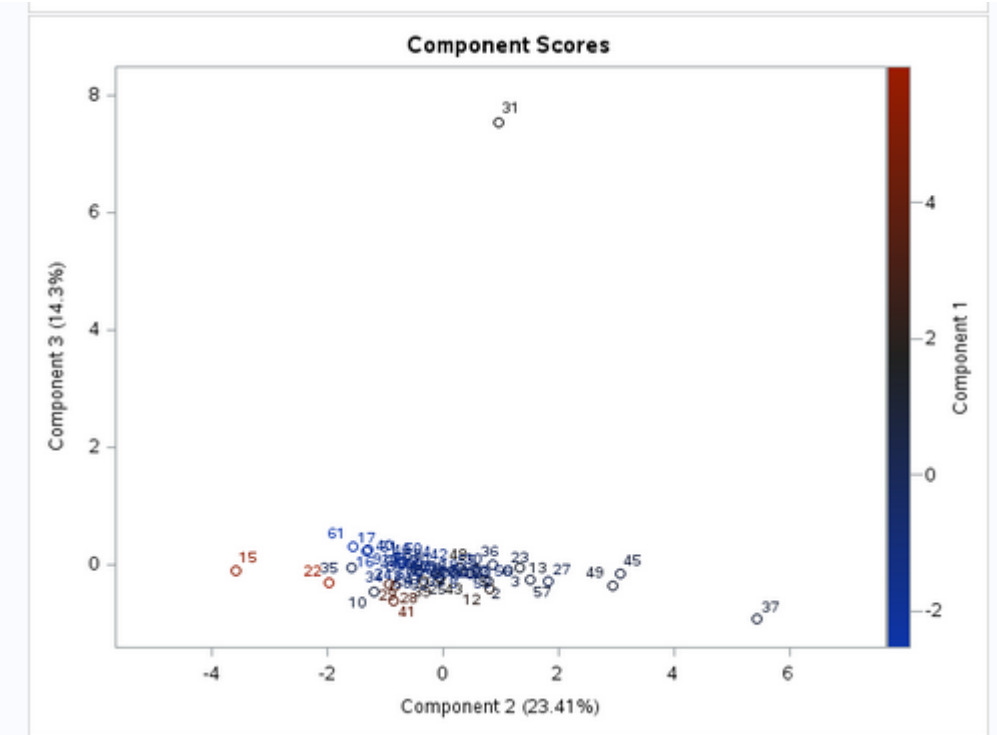
Astfel de reprezentari pot fi realizate pt fiecare component cu fiecare, fiind o reprezentare bidimensionala a variabilelor.



In continuare formele (tarile) au fost reprezentate infunctie de scorurile associate in planul alcatuit din primele 2 componente. Observam ca inregistrarile 37,22 si 15 par a fi outliars in raport cu prima componenta (22 - Danemarca si 15 - China) si respectiv cu cea de-a doua (37 - Japonia).



Pentru a putea reprezenta obiectele in functie de scorurile obtinute pentru cele 3 componente retinute, pentru componenta 1 se va utiliza un gradient, iar pe cele 2 axe se vor situa celelalte doua componente.



Utilizand procedura CORR se realizeaza cateva statistici „simple”, cum ar fi media, abaterea standard, suma elementelor, minimul si maximul. Mai apoi este calculata matricea de corelatie dintre cele 3 componente principale si variabilele setului de date. Pe primul rand al fiecarei celule se afla valoarea coeficientului Pearson si observam in cazul primei componente ca exista o legatura foarte puternica intre aceasta si variabila importurilor, iar pe cel de-ai doilea rand este probabilitatea ca rezultatul sa fie eronat.

The CORR Procedure

3 With Variables:	Prin1 Prin2 Prin3
7 Variables:	GDP_PCAP USE_ELEC HEALTH_XPD ADJ_AEDU EXP_GNFS IMP_GNFS CBRT_RATE

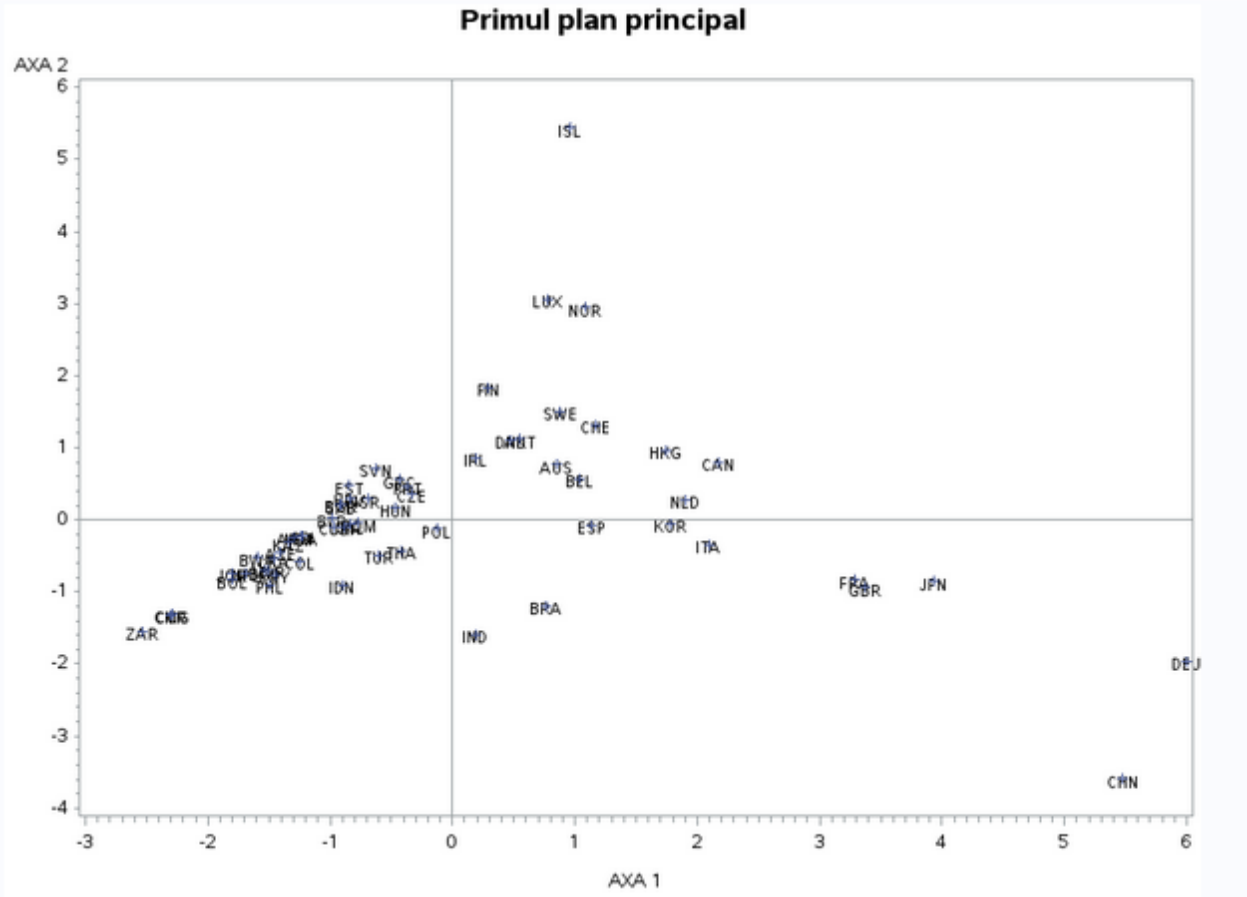
Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
Prin1	61	0	1.78759	0	-2.53400	5.99670	
Prin2	61	0	1.28006	0	-3.59677	5.44170	
Prin3	61	0	1.00057	0	-0.93430	7.53922	
GDP_PCAP	61	0	1.00000	0	-0.92764	3.15871	GDP_PCAP
USE_ELEC	61	0	1.00000	0	-0.74598	6.12745	USE_ELEC
HEALTH_XPD	61	0	1.00000	0	-0.21804	7.65541	HEALTH_XPD
ADJ_AEDU	61	0	1.00000	0	-0.63869	3.45486	ADJ_AEDU
EXP_GNFS	61	0	1.00000	0	-0.60759	4.79700	EXP_GNFS
IMP_GNFS	61	0	1.00000	0	-0.66068	4.35774	IMP_GNFS
CBRT_RATE	61	0	1.00000	0	-0.95456	3.65155	CBRT_RATE

Pearson Correlation Coefficients, N = 61
Prob > |r| under H0: Rho=0

	GDP_PCAP	USE_ELEC	HEALTH_XPD	ADJ_AEDU	EXP_GNFS	IMP_GNFS	CBRT_RATE
Prin1	0.58854 <.0001	0.35390 0.0051	0.17893 0.1677	0.85721 <.0001	0.90258 <.0001	0.91759 <.0001	-0.54809 <.0001
Prin2	0.68708 <.0001	0.80713 <.0001	0.14699 0.2583	-0.31267 0.0142	-0.35813 0.0046	-0.35873 0.0045	-0.37241 0.0031
Prin3	-0.01835 0.8884	-0.14245 0.2734	0.96886 <.0001	-0.16980 0.1908	0.04366 0.7383	0.04908 0.7072	0.09310 0.4754

In functie de scorurile fiecarei inregistrari pentru o componenta principala, acestea pot fi reprezentate intr-un plan, numit primul plan principal, ale carui axe sa fie reprezentate de primele doua componente principale. In acest caz, pentru valorile extreme se vad si codurile asociate.



Mai jos se regaseste si matricea scorurilor principale, pentru primele observatii, scorurile principale (valorile din dreptul coloanelor asociate componentelor principale) sut obtinute prin inlocuirea calorilor pentru fiecare variabila in ecuatiile prezentate mai sus, w1, w2 si w3.

Matricea scorurilor

Obs	CCode	GDP_PCAP	USE_ELEC	HEALTH_XPD	ADJ_AEDU	EXP_GNFS	IMP_GNFS	CBRT_RATE	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7
1	ARM	-0.836058634	-0.529092074	-0.208496747	-0.638693972	-0.606469831	-0.660681031	-0.184431038	-1.29596	-0.24182	-0.07887	-0.69882	-0.18605	0.02740	0.02721
2	AUS	0.8874919391	0.6488416894	-0.031181739	0.7906411346	-0.149332974	0.1196160008	-0.240636424	0.85645	0.76705	-0.29606	0.28362	0.37322	0.53737	-0.10586
3	AUT	1.1186787127	0.3413288335	0.0078996789	-0.154189657	-0.040186438	-0.070135137	-0.798796146	0.55136	1.11755	-0.11482	-0.39066	0.48752	-0.28276	0.00632
4	AZE	-0.786084466	-0.535565621	-0.195360917	-0.607958147	-0.53425162	-0.622130075	0.4862692607	-1.41412	-0.45122	-0.00391	-0.07772	-0.09627	-0.02283	0.04783
5	BEL	0.9584184799	0.2949944186	-0.026035298	0.0417459745	0.4455409283	0.508213202	-0.500245597	1.03057	0.56572	-0.09405	-0.06114	0.29371	-0.47566	-0.08842
6	BGD	-0.912536994	-0.725823817	-0.216010421	-0.597001545	-0.549428216	-0.585064716	0.6698129461	-1.53515	-0.70364	0.02188	-0.00418	-0.04873	-0.02538	0.01149
7	BGR	-0.706361975	-0.120246671	-0.168027256	-0.59997183	-0.551052418	-0.58891869	-0.75985477	-0.90845	0.21257	-0.15445	-1.02646	-0.41955	0.07286	0.01443
8	BLR	-0.701180748	-0.282785038	-0.184149222	-0.587451256	-0.526549914	-0.552577546	-0.513226056	-0.97912	0.01917	-0.12334	-0.85620	-0.29775	0.02475	0.00421
9	BOL	-0.879766948	-0.677857511	-0.207434706	-0.612192907	-0.598098361	-0.653929399	1.3950311712	-1.80357	-0.82919	0.08731	0.65770	0.03033	-0.02065	0.02811
10	BRA	-0.65421488	-0.439235614	-0.170156858	2.213698541	-0.120602633	0.1461888314	-0.01581488	0.76115	-1.19100	-0.46547	-0.05842	0.26807	1.81490	0.05336
11	BWA	-0.616413968	-0.549079805	-0.18062378	-0.613695392	-0.594519596	-0.64831125	1.1164705284	-1.59931	-0.52472	0.06487	0.50839	0.10661	-0.05608	0.02607
12	CAN	0.9133924316	1.4064232555	-0.004132685	1.2440472322	0.6499156476	0.9479769988	-0.578128349	2.16733	0.79343	-0.41105	0.44514	-0.29932	0.45576	-0.15491
13	CHE	1.9908381899	0.2827582573	0.049446086	-0.005501521	0.2799953811	0.2104726736	-0.681972018	1.17226	1.31500	-0.06887	-0.06310	1.07690	-0.60115	0.02012
14	CHL	-0.488517252	-0.290621865	-0.149171552	-0.437407095	-0.433462353	-0.426962802	-0.163921913	-0.83082	-0.06713	-0.07499	-0.46857	-0.11624	0.01035	-0.01291
15	CHN	-0.784397255	-0.326142387	-0.199140201	2.1246138118	4.7969992735	4.3577363051	-0.457410084	5.47527	-3.59877	-0.11206	0.31824	-1.53017	-1.44548	0.10577
16	CMR	-0.893550997	-0.726230303	-0.213486704	-0.628016323	-0.600478355	-0.649302328	2.950739141	-2.30170	-1.31716	0.23616	2.02959	0.19705	-0.08905	0.01836
17	COG	-0.844494015	-0.737269122	-0.214116712	-0.638005545	-0.588837312	-0.646375662	2.9750125987	-2.29265	-1.30656	0.24082	2.05752	0.23623	-0.11759	0.02250
18	COL	-0.73336947	-0.612190595	-0.188062855	-0.423723317	-0.509040552	-0.48821745	0.5184607981	-1.25126	-0.56935	-0.00750	-0.03232	0.00473	0.05124	-0.01550
19	CRI	-0.664357886	-0.517350484	-0.160336094	-0.590392967	-0.573679008	-0.621880559	0.0062518996	-1.23111	-0.22405	-0.02418	-0.47832	-0.05767	0.00304	0.02524
20	CUB	-0.687855902	-0.585391352	-0.197399609	-0.459970074	-0.570849183	-0.632230358	-0.74116291	-0.96820	-0.09612	-0.14201	-1.16281	-0.07492	0.11929	0.04987
21	CZE	-0.194665206	0.0671231436	-0.125998497	-0.469004333	-0.218369897	-0.258640982	-0.656011101	-0.33020	0.36236	-0.13166	-0.69230	-0.25285	-0.14929	-0.00230

III. Analiza Cluster

Setul de date utilizat pentru aplicarea analizei cluster este identic cu cel utilizat in cadrul aplicatiei anterioare, datele utilizate fiind standardizate.

Vom utiliza procedura PROC CLUSTER, iar metoda aleasa pentru evaluarea distanțelor este metoda lui Ward.

In sectiunea de output de mai jos observam valorile proprii, dar si faptul ca varianța totală medie patratică a datelor este 1, iar distanța medie pătratică dintre observații este aproximativ 3,74.

Analiza claselor - Pregatirea datelor

The CLUSTER Procedure
Ward's Minimum Variance Cluster Analysis

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	3.19547158	1.55692793	0.4565	0.4565
2	1.63854364	0.63740050	0.2341	0.6906
3	1.00114314	0.31640607	0.1430	0.8336
4	0.68473707	0.40298712	0.0978	0.9314
5	0.28174995	0.08913138	0.0402	0.9717
6	0.19261858	0.18688254	0.0275	0.9992
7	0.00573604		0.0008	1.0000

The data have been standardized to mean 0 and variance 1

Root-Mean-Square Total-Sample Standard Deviation	1

Root-Mean-Square Distance Between Observations	3.741657

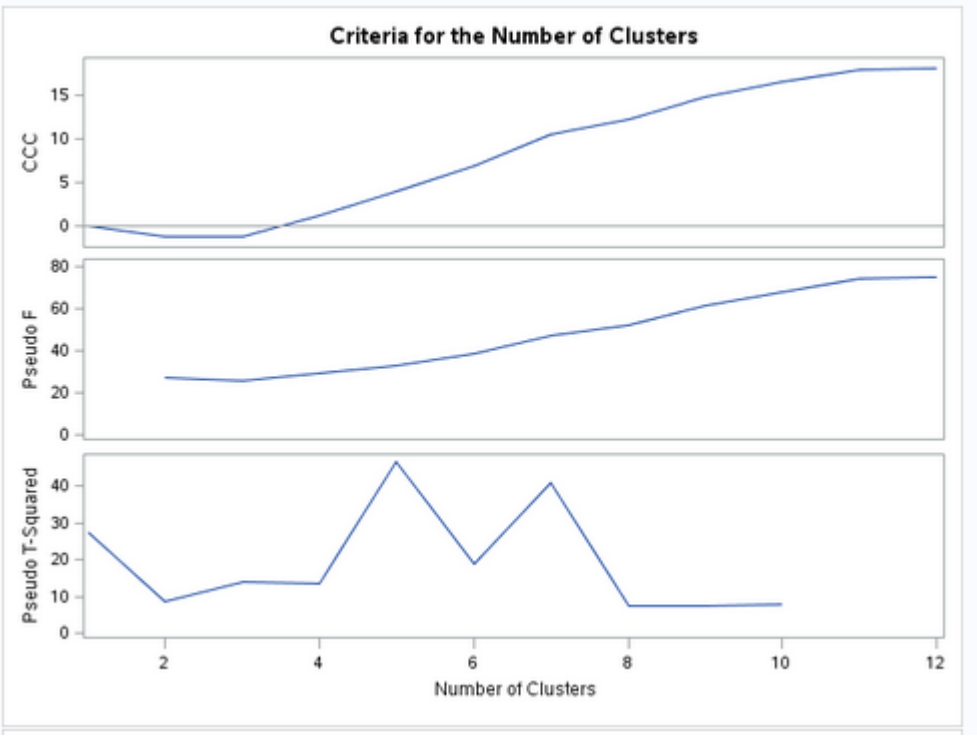
In tabelul următor se află ultimele 20 de clustere (mai exact rezultatul ultimelor 20 de interatii/ alipiri dintre clustere și obiecte). Astfel, analizand valorile din coloana Cubic Clustering Criterion, vom considera acceptabile 4 clustere, dar optime 5. Pe baza testului Pseudo F Statistic putem considera 4, dar mai curand 5 clustere, deoarece diferenta dintre valorile pentru 5 - 6 clustere este mai mare fata de diferenta dintre valorile pentru 4 si respectiv 5 clustere. Daca ar fi sa analizam pe baza Pseudo t Statistic, ar fi indicat sa luam 4 sau 6 clustere, dar in niciun caz 5.

De asemenea acceptabil este sa luam si doar 3 clustere, tinand cont ca observatiile sunt țări (state) și acestea ar putea fi clasificate în Dezvoltate, Emergente și Sărace.

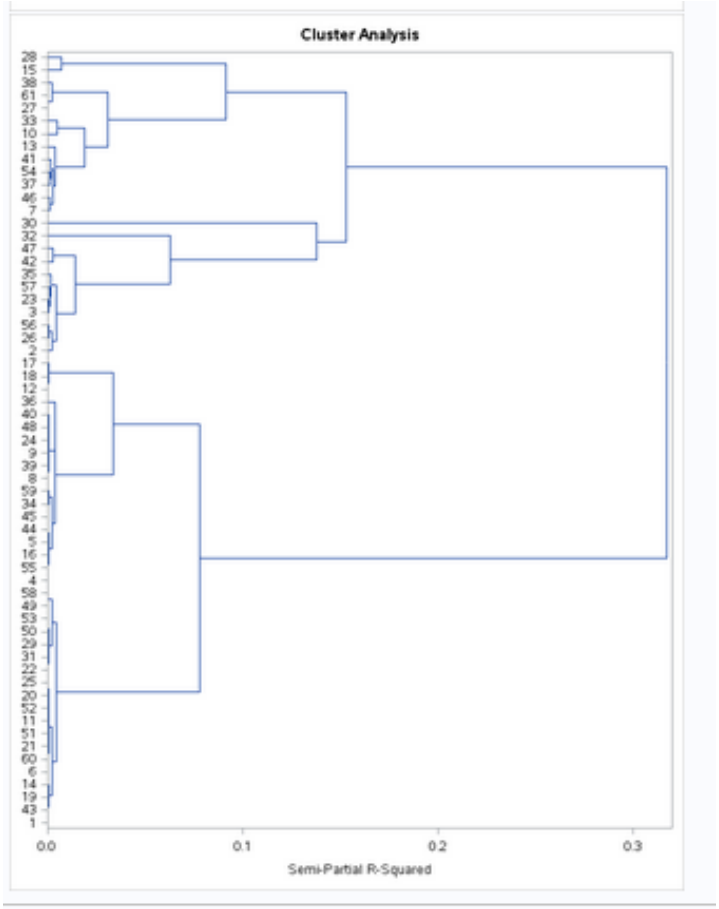
Cluster History

Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Approximate Expected R-Square	Cubic Clustering Criterion	Pseudo F Statistic	Pseudo t-Squared	Tie
20	CL47	OB38	3	0.0024	.974	.	.	81.6	12.5	
19	OB42	OB47	2	0.0026	.972	.	.	79.8	.	
18	CL29	CL25	5	0.0028	.969	.	.	78.6	2.0	
17	CL30	OB36	7	0.0037	.965	.	.	76.1	10.7	
16	CL24	CL17	15	0.0038	.961	.	.	74.5	5.8	
15	CL18	OB13	6	0.0040	.957	.	.	73.7	2.3	
14	CL21	CL26	7	0.0042	.953	.	.	73.5	3.6	
13	CL23	CL22	19	0.0044	.949	.	.	74.0	9.8	
12	OB10	OB33	2	0.0048	.944	.802	18.1	75.0	.	
11	OB15	OB28	2	0.0072	.937	.787	17.9	74.0	.	
10	CL14	CL19	9	0.0140	.923	.770	16.5	67.7	7.7	
9	CL15	CL12	8	0.0188	.904	.750	14.9	61.2	7.2	
8	CL9	CL20	11	0.0308	.873	.728	12.3	52.1	7.5	
7	CL16	CL31	18	0.0333	.840	.702	10.5	47.2	40.6	
6	CL10	OB32	10	0.0628	.777	.670	6.98	38.4	18.8	
5	CL13	CL7	37	0.0778	.699	.631	3.92	32.6	46.6	
4	CL8	CL11	13	0.0911	.608	.580	1.30	29.5	13.4	
3	CL6	OB30	11	0.1380	.470	.502	-1.1	25.7	13.9	
2	CL3	CL4	24	0.1530	.317	.358	-1.2	27.4	8.5	
1	CL5	CL2	61	0.3172	.000	.000	0.00	.	27.4	

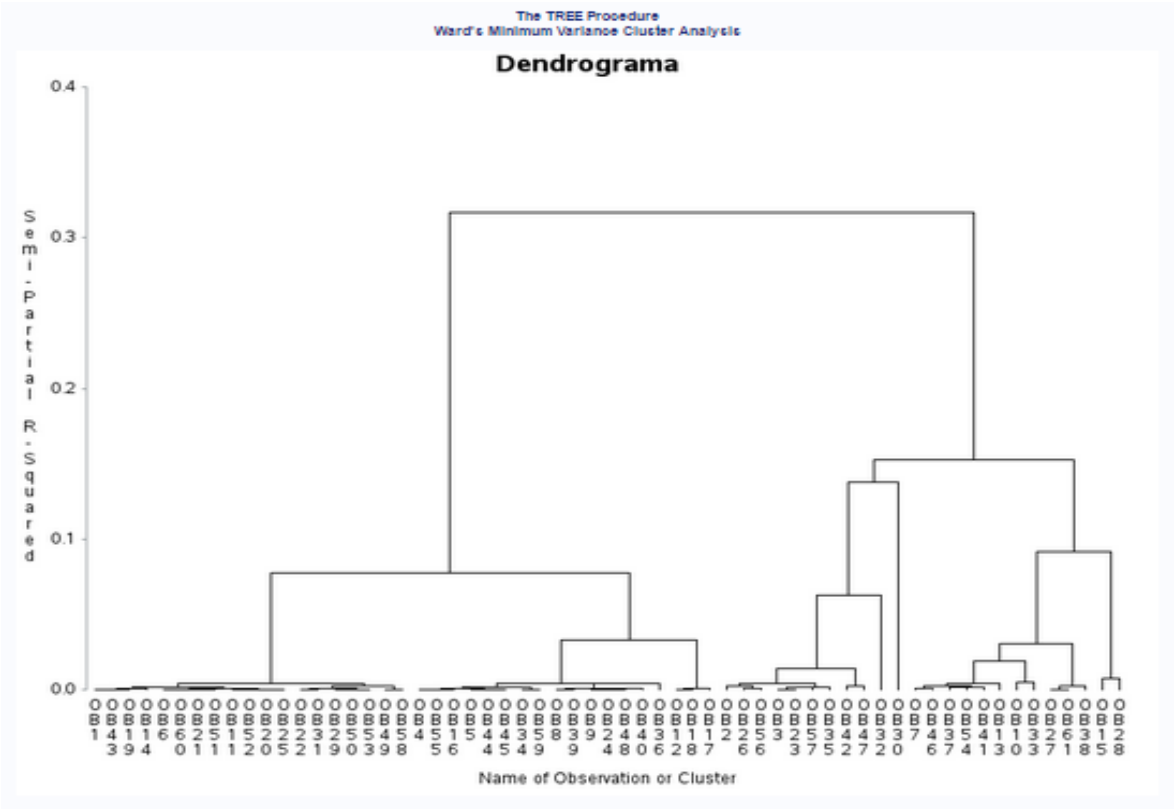
In urmatoarea figura sunt reprezentate grafic criteriile pe baza carora se realizeaza stabilirea numărului de clustere, astfel că modificarile bruște pot fi identificate și grafic, iar mai apoi pe datele din tabel sa se verifice respectivele ipoteze.



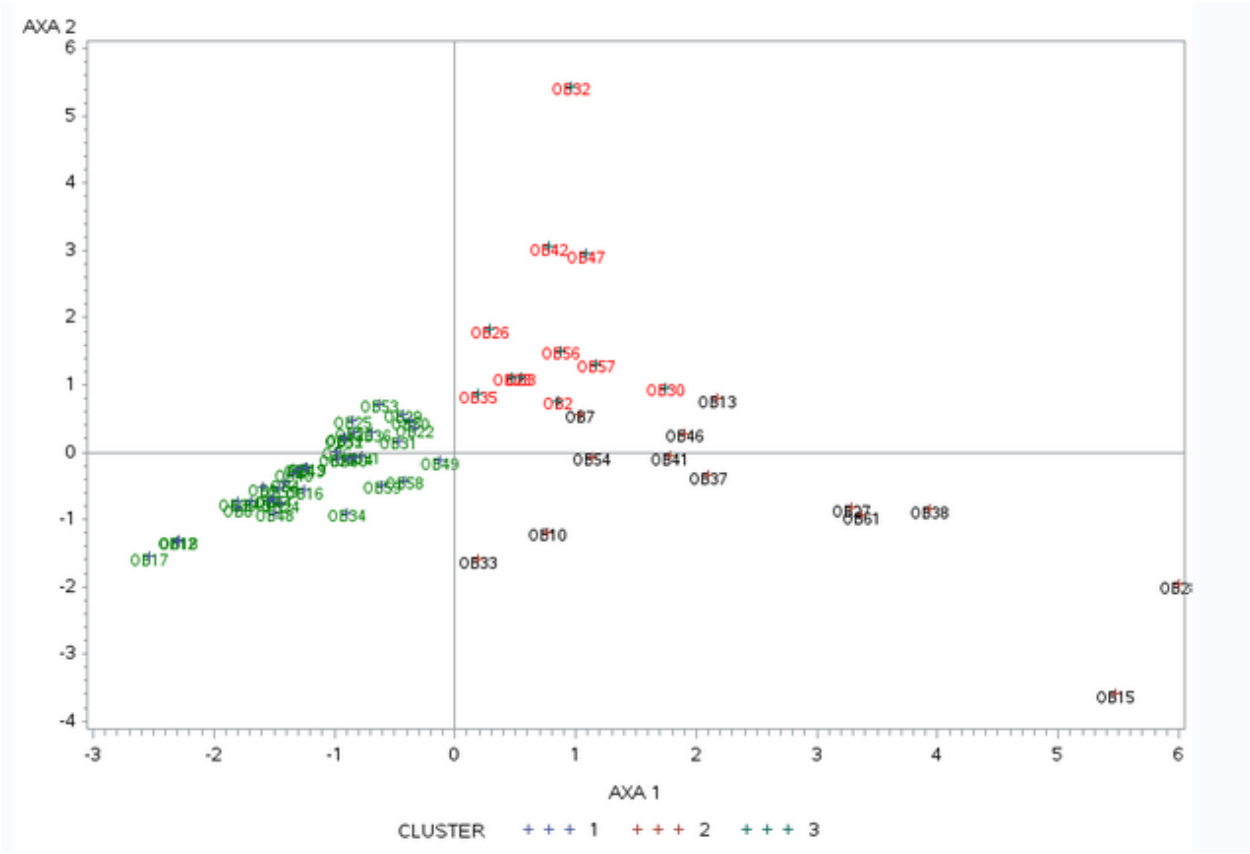
In figura urmatoare este reprezentata alipirea (concatenarea) obiectelor la fiecare pas si in final alcatuirea clusterelor.



Sau utilizand procedura TREE se construieste dendograma, astfel încât să observăm mai ușor distanțele dintre pașii concatenărilor realizate.



Repetând analiza componentelor principale, obținem următorul grafic în care de aceasta dată identificăm și apartenența fiecărei țări la unul dintre primele 3 clustere, alese pentru a fi reținute în analiză.



IV. Analiza Discriminant

Acest tip de analiză face parte din clasa metodelor si tehnicilor de recunoastere a formelor în mod supervizat, deoarece se cunoaste apartenenta fiecarui obiect la o anumită clasă.

Pe baza aceluiași set de date, vom efectua analiza discriminant, iar in prima secțiune de output, prezenta mai jos, observam ca in primul tabel sunt expuse numarul observatiilor, numarul de variabile si numarul de clase, dar si numarul de grade de libertate pentru total, in interiorul claselor, dar si interclasa.

De asemenea apar numarul de observarii citite si respectiv utilizate, 61, reprezentand toate observatiile.

In tabelul urmatorul se afla informatiile legate de nivelul claselor, pe coloana Frequency fiind prezent numarul de obiecte care se afla in cadrul fiecarei clase, si anume: 37 pentru clusterul 1, 13 pentru cel de-al doilea si respectiv 11 pentru cel de-al treilea cluster. Apoi urmeaza o coloana numita Weight, insa deoarece obiectele sunt identice, observam ca valorile sunt si ele identice cu cele din coloana anterioara. In coloana Proportion observam procentual cantitatea de observatii din total retinuta in fiecare cluster. Astfel in primul cluster se afla 60.55% din observatii, in cel de-al doilea 21.31%, iar in cel de-al treilea 18.03%. Ultima coloana cuprinde Probabilitatea fiecarui cluster de a „accepta" observatii si aceastea sunt echiprobabile.

Analiza discriminant

The DISCRIM Procedure

Total Sample Size	61	DF Total	60
Variables	7	DF Within Classes	58
Classes	3	DF Between Classes	2

Number of Observations Read	61
Number of Observations Used	61

Class Level Information

CLUSTER	Variable Name	Frequency	Weight	Proportion	Prior Probability
1	1	37	37.0000	0.606557	0.333333
2	2	13	13.0000	0.213115	0.333333
3	3	11	11.0000	0.180328	0.333333

In cea de-a doua sectiune a output-ului in primul tabel este prezentata distanta patratica generalizata dintre clustere, de exemplu intre clusterul 1 si 2 diind o distanta de 27.81, iar intre clustere 1 si 3 de 23.72.

In cel de-al doilea tabel al sectiunii regasim functiile liniare discriminant, functii de tip Fisher.

Pe baza acestui tabel putem identifica o relatie a formelor cuprinse in fiecare cluster, ca functie de variabilele incluse in analiza.

Pentru clusterul 1 vom avea:

w_1=-1.59 – 2.64 *GDP_PCAP – 0.37*USE_ELEC – 0.21*HEALTH_XPD – 0.76*ADJ_AEDU + 5.74*EXP_GNFS -0.31*CBRT_RATE

Analog si pentru celelalte clustere, in cazul primului cluster singura variabila al carei coeficient este pozitiv fiind cel al variabilei exporturi.

Analiza discriminant

The DISCRIM Procedure

Generalized Squared Distance to CLUSTER			
From CLUSTER	1	2	3
1	0	27.81662	23.72118
2	27.81662	0	21.46161
3	23.72118	21.46161	0

Linear Discriminant Function for CLUSTER				
Variable	Label	1	2	3
Constant		-1.59529	-5.86368	-5.97345
GDP_PCAP	GDP_PCAP	-2.64391	2.01222	6.51508
USE_ELEC	USE_ELEC	-0.37571	0.49632	0.67721
HEALTH_XPD	HEALTH_XPD	-0.21513	-0.68537	1.53359
ADJ_AEDU	ADJ_AEDU	-0.76608	2.38018	-0.23612
EXP_GNFS	EXP_GNFS	5.74999	-17.03348	0.78958
IMP_GNFS	IMP_GNFS	-7.50784	22.00270	-0.74956
CBRT_RATE	CBRT_RATE	-0.31110	0.28263	0.71240

In ultima sectiune, elementele matricei de confuzie sunt cuprinse in primul tabel, pe prima linie fiind numarul de elemente, iar pe cea de-a doua proportia.pe linie sunt clasele reale, iar pe coloane clasele predictate. Astfel, pentru clusterul 1 din 37 de forme, toate au fost corect clasificate, in cazul clusterului 2 din 13 forme predictate, 11 au fost corect clasificare, 2 incorect, iar in cazul celui de-al treilea cluster din 11 forme clasificare, 10 previzionate corect, iar una incorect.

In cel de-al doilea tabel regasim eroarea previzionarii in cazul fiecarui cluster, dar si totalul, astfel modelul obtinut ar eo eroare de previzionare de aproximativ 8.16%.

Analiza discriminant

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.COUNTRIES_CLASIF
Resubstitution Summary using Linear Discriminant Function

Number of Observations and Percent Classified into CLUSTER				
From CLUSTER	1	2	3	Total
1	37 100.00	0 0.00	0 0.00	37 100.00
2	0 0.00	11 84.62	2 15.38	13 100.00
3	0 0.00	1 9.09	10 90.91	11 100.00
Total	37 60.66	12 19.67	12 19.67	61 100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for CLUSTER				
	1	2	3	Total
Rate	0.0000	0.1538	0.0909	0.0816
Priors	0.3333	0.3333	0.3333	

Anexa 1

CCode	GDP_PCAP	USE_ELEC	HEALTH_XPD	ADJ_AEDU	EXP_GNFS	IMP_GNFS	CBRT_RATE
ARM	2090.4595692	1754.6523083	225.4660529	237.52644353	1425.810823	2395.8922641	14.033
AUS	36504.173941	10712.17547	3970.1189011	67879.326651	149220.54269	221829.85644	13.6
AUT	41120.223922	8373.7133138	4795.4655241	23166.190238	184508.18758	168468.32525	9.3
AZE	3088.2816192	1705.4246326	502.87699889	1692.0674654	24774.335415	13237.134353	19.2
BGD	563.43479297	258.61815582	66.787423116	2210.5772856	19867.66281	23660.599257	20.614
BLR	4783.5334701	3627.6786657	739.65280241	2662.534789	27264.337173	32796.59271	11.5
BEL	37920.346899	8021.3661721	4078.8047593	32438.640586	341546.38052	331110.58014	11.6
BOL	1217.746318	623.37491265	247.89491112	1491.6618624	4132.3506572	4294.5760404	26.201
BWA	6476.0511132	1602.6568668	814.1053764	1420.5582879	5289.3842838	5874.5036258	24.055
BRA	5721.2895041	2437.9601052	1035.1525863	135224.04121	158509.21217	229302.62812	15.332
BGR	4680.0811741	4863.6914411	1080.1268979	2070.011629	19342.549683	22576.79056	9.6
CMR	942.52356692	255.52706072	120.08491416	742.83511765	3362.886313	5595.7940637	38.186
CAN	37021.322763	16473.156803	4541.3582933	89336.299563	407621.79405	454780.29731	11
CHL	9029.7336809	3568.0840034	1478.3338253	9763.2189673	57360.03147	68121.782931	14.191
CHN	3121.9697442	3297.9704344	423.06361956	131008.19887	1748395.577	1413667.7094	11.93
COL	4140.8290903	1122.7348399	657.00217435	10410.789567	32925.205383	50895.842612	19.448
ZAR	261.9458575	105.3158261	24.014765433	550.21686734	2776.7541874	6162.9392596	43.585
COG	1922.0323837	171.58305365	106.78000588	270.10555201	7126.4961933	6418.8267171	38.373
CRI	5518.7665798	1843.9404941	1242.5538154	2523.3213521	12027.254886	13307.30297	15.502
HRV	10830.628574	3900.6013612	1361.7144038	2467.4396752	18320.61742	19602.953396	9.6
CUB	5049.587428	1326.527997	459.82256814	8695.4493823	12942.152073	10396.747999	9.744
CZE	14897.005874	6288.5334041	1967.717376	8267.9124599	126900.54512	115456.99351	10.4
DNK	48143.833504	6121.9924992	4456.1780719	27929.228092	145596.29516	131886.5126	10.6
EGY	1551.2539097	1742.9109709	307.83312587	10120.401729	43988.053824	49938.785226	23.823
EST	11256.109498	6314.4144477	1293.6922986	1194.9604568	13882.44267	13552.20589	11
FIN	40530.050871	15738.440821	3381.9962055	16684.809778	91166.641173	86338.40969	11.1
FRA	35771.54755	7292.146247	4128.4260287	149590.79994	654936.73444	699721.9849	12.6
DEU	38469.884495	7080.9593813	4473.8368165	179249.49473	1432977.4379	1234095.346	8.1
GRC	20007.269818	5380.4597646	2322.4347643	9140.0234024	54267.714893	67881.557106	9.6
HKG	32607.962458	5948.8658861	166300.60855	7181.0155414	468862.86088	450649.10813	13.5
HUN	11341.686181	3895.2129355	1689.9362983	6054.5522127	110195.80875	100207.69356	8.8
ISL	57279.667842	52373.877009	3361.2614803	1041.2446618	7058.6572812	5238.540316	14.1
IND	1086.0485973	684.10570129	145.7234594	57443.296466	309632.5286	413387.68048	20.999
IDN	1650.5550051	679.70010653	132.0833729	20310.287921	149862.20115	125993.48302	19.633
IRL	47538.22741	5701.1523787	3702.923401	12977.597724	201111.07496	147823.91733	16.2
ISR	23754.071279	6925.6226017	2185.7951881	14658.027892	76912.896179	74814.342458	21.4
ITA	30915.230947	5514.7867741	3016.8390848	93246.710042	490051.33139	494174.91307	9
JPN	36203.430066	7847.8044873	3403.6466968	193960.63384	744943.13379	627835.44528	8.3
JOR	2827.1258595	2289.4353664	494.25216837	1603.1460606	8018.555182	11767.408168	28.317
KAZ	5015.4450792	4892.9127961	533.6334626	7075.7849533	32578.954614	28528.141686	22.5
KOR	22883.756211	10161.946378	2198.4942679	50994.806755	580360.31997	499597.63456	9.4
LUX	81852.975981	15530.137148	6020.0234518	1408.4505743	77370.044272	66330.025369	10.9
MDA	1050.2014089	1470.2305009	436.03370645	559.78357624	2222.1792717	3742.1521368	12.317
MAR	2432.8244168	826.40286987	321.24169708	4999.172564	24590.108202	32203.891669	22.322
NPL	384.77187235	105.50012084	85.411101462	796.98925263	1063.355871	3913.0131051	22.27
NLD	44195.21374	7035.6723992	5117.9427367	50596.145957	544638.44837	475426.53261	10.8
NOR	64612.647783	23173.624212	6105.881218	30749.481844	128905.70431	104571.05161	12.2
PHL	1430.0385896	646.96240284	182.24262937	6540.7681054	60448.111127	63857.467659	24.79
POL	10420.336131	3832.1326194	1445.2912409	25049.440295	165384.0674	172100.75371	10.1
PRT	18916.520711	4848.2793373	2615.0132202	12997.213063	66737.364946	75125.165771	9.2
ROM	5793.4251829	2639.0334338	863.88704918	6723.5381334	49537.340573	70497.917357	9.7
SRB	4197.1271885	4489.5708505	1172.1403456	2033.979705	10950.710329	16732.262042	9
SVN	19404.395582	6806.1707593	2423.2892828	2682.7643851	28748.160769	27100.911	10.7
ESP	25937.240669	5529.7622215	2984.4692997	64709.442146	336200.61869	329695.4365	10.1
LKA	1724.81198	490.24869424	182.80973497	982.37853644	9348.0903975	14202.468583	18.332
SWE	45727.097716	14030.163147	3938.0235148	35481.479848	209875.15049	189349.16186	11.8
CHE	58534.419707	7928.3170538	5672.8693825	30202.701888	288024.641	247380.43407	10.2
THA	3158.0666713	2315.9882103	372.2957884	13529.525941	176310.97498	167839.83227	10.748

CCode	GDP_PCAP	USE_ELEC	HEALTH_XPD	ADJ_AEDU	EXP_GNFS	IMP_GNFS	CBRT_RATE
TUR	8413.3183808	2709.2621135	1047.1723771	20281.34493	131424.05129	158967.62694	17.439
UKR	2084.7824065	3662.4433063	527.88119279	9439.1283614	38820.50205	52218.449262	11
GBR	39808.815449	5472.1454452	3364.3235435	143496.92129	714897.70436	725046.7082	12.8

Bibliografie

Prof. Univ. Dr. Ruxanda Gheorghe- Suport de curs

Prof. Univ. Dr. Ruxanda Gheorghe- DATA MINING, Bucuresti, 2013

Asist. Drd. Alexandru Alexa- Suport de seminar