410-57, Assignment #8
Anamitra Bhattacharyya

INTRODUCTION

This assignment specifically deals with cluster analysis to identify and segment groups in the European employment data. The raw data set being used in this assignment comprises the employment in various industry segments reported as a percent for 30 European nations. Note that the countries are already sub-grouped in to several geopolitical clades comprising: EU, European Union; EFTA, European Free Trade Association; Eastern, Eastern European nations or the former Eastern Block; and other. The definitions of the abbreviated industries are described below.

AGR: agriculture
MIN: mining
MAN: manufacturing
PS: power and water supply
CON: construction
SER: services
FIN: finance
SPS: social and personal services
TC: transport and communications

The overall goal of the assignment is to use statistical methods in SAS (*e.g.* PROC CLUSTER and TREE, PRINCOMP) to perform segmentation analysis on the European employment data to group various countries into suitable 'bins' for association purposes.

RESULTS

1. Initial correlation analysis

The starting data table of 30 countries and 9 industry segments is shown below.

Obs	COUNTRY	GROUP	AGR	MIN	MAN	PS	CON	SER	FIN	SPS	TC
1	Belgium	EU	2.6	0.2	20.8	0.8	6.3	16.9	8.7	36.9	6.8
2	Denmark	EU	5.6	0.1	20.4	0.7	6.4	14.5	9.1	36.3	7.0
3	France	EU	5.1	0.3	20.2	0.9	7.1	16.7	10.2	33.1	6.4
4	Germany*	EU	3.2	0.7	24.8	1.0	9.4	17.2	9.6	28.4	5.6
5	Greece	EU	22.2	0.5	19.2	1.0	6.8	18.2	5.3	19.8	6.9
6	Ireland	EU	13.8	0.6	19.8	1.2	7.1	17.8	8.4	25.5	5.8
7	Italy	EU	8.4	1.1	21.9	0.0	9.1	21.6	4.6	28.0	5.3
8	Luxembourg	EU	3.3	0.1	19.6	0.7	9.9	21.2	8.7	29.6	6.8
9	Netherlands	EU	4.2	0.1	19.2	0.7	0.6	18.5	11.5	38.3	6.8
10	Portugal	EU	11.5	0.5	23.6	0.7	8.2	19.8	6.3	24.6	4.8
11	Spain	EU	9.9	0.5	21.1	0.6	9.5	20.1	5.9	26.7	5.8
12	UK*	EU	2.2	0.7	21.3	1.2	7.0	20.2	12.4	28.4	6.5
13	Austria	EFTA	7.4	0.3	26.9	1.2	8.5	19.1	6.7	23.3	6.4
14	Finland	EFTA	8.5	0.2	19.3	1.2	6.8	14.6	8.6	33.2	7.5
15	Iceland	EFTA	10.5	0.0	18.7	0.9	10.0	14.5	8.0	30.7	6.7
16	Norway	EFTA	5.8	1.1	14.6	1.1	6.5	17.6	7.6	37.5	8.1
17	Sweden	EFTA	3.2	0.3	19.0	0.8	6.4	14.2	9.4	39.5	7.2
18	Switzerland	EFTA	5.6	0.0	24.7	0.0	9.2	20.5	10.7	23.1	6.2
19	Albania	Eastern	55.5	19.4	0.0	0.0	3.4	3.3	15.3	0.0	3.0
20	Bulgaria	Eastern	19.0	0.0	35.0	0.0	6.7	9.4	1.5	20.9	7.5
21	Czech/Slovak Reps	Eastern	12.8	37.3	0.0	0.0	8.4	10.2	1.6	22.9	6.9
22	Hungary	Eastern	15.3	28.9	0.0	0.0	6.4	13.3	0.0	27.3	8.8
23	Poland	Eastern	23.6	3.9	24.1	0.9	6.3	10.3	1.3	24.5	5.2
24	Romania	Eastern	22.0	2.6	37.9	2.0	5.8	6.9	0.6	15.3	6.8
25	USSR (Former)	Eastern	18.5	0.0	28.8	0.0	10.2	7.9	0.6	25.6	8.4
26	Yugoslavia (Former)	Eastern	5.0	2.2	38.7	2.2	8.1	13.8	3.1	19.1	7.8
27	Cyprus	Other	13.5	0.3	19.0	0.5	9.1	23.7	6.7	21.2	6.0
28	Gibralta	Other	0.0	0.0	6.8	2.0	16.9	24.5	10.8	34.0	5.0
29	Malta	Other	2.6	0.6	27.9	1.5	4.6	10.2	3.9	41.6	7.2
30	Turkey	Other	44.8	0.9	15.3	0.2	5.2	12.4	2.4	14.5	4.4

The results from the initial correlation analysis of the 9 variables (industry segments) across the 30 observation sets (countries), is shown below.

The CORR Procedure

9 Variables:	AGR MIN MAN PS CON SER FIN SPS TC

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
AGR	30	12.18667	12.30690	365.60000	0	55.50000
MIN	30	3.44667	8.86573	103.40000	0	37.30000
MAN	30	20.28667	9.45679	608.60000	0	38.70000
PS	30	0.80000	0.62090	24.00000	0	2.20000
CON	30	7.53000	2.73309	225.90000	0.60000	16.90000
SER	30	15.63667	5.16016	469.10000	3.30000	24.50000
FIN	30	6.65000	3.98668	199.50000	0	15.30000
SPS	30	26.99333	8.73206	809.80000	0	41.60000
TC	30	6.45333	1.23337	193.60000	3.00000	8.80000

Pearson Correlation Coefficients, N = 30
Prob > |r| under H0: Rho=0

	AGR	MIN	MAN	PS	CON	SER	FIN	SPS	TC
AGR	1.00000	0.31607 0.0888	-0.25439 0.1749	-0.38236 0.0370	-0.34861 0.0590	-0.60471 0.0004	-0.17575 0.3529	-0.81148 <.0001	-0.48733 0.0063
MIN	0.31607 0.0888	1.00000	-0.67193 <.0001	-0.38738 0.0344	-0.12902 0.4968	-0.40655 0.0258	-0.24806 0.1863	-0.31642 0.0885	0.04470 0.8146
MAN	-0.25439 0.1749	-0.67193 <.0001	1.00000	0.38789 0.0342	-0.03446 0.8565	-0.03294 0.8628	-0.27374 0.1433	0.05028 0.7919	0.24290 0.1959
PS	-0.38236 0.0370	-0.38738 0.0344	0.38789 0.0342	1.00000	0.16480 0.3842	0.15498 0.4135	0.09431 0.6201	0.23774 0.2059	0.10537 0.5795
CON	-0.34861 0.0590	-0.12902 0.4968	-0.03446 0.8565	0.16480 0.3842	1.00000	0.47308 0.0083	-0.01802 0.9247	0.07201 0.7053	-0.05461 0.7744
SER	-0.60471 0.0004	-0.40655 0.0258	-0.03294 0.8628	0.15498 0.4135	0.47308 0.0083	1.00000	0.37928 0.0387	0.38798 0.0341	-0.08489 0.6556
FIN	-0.17575 0.3529	-0.24806 0.1863	-0.27374 0.1433	0.09431 0.6201	-0.01802 0.9247	0.37928 0.0387	1.00000	0.16602 0.3806	-0.39132 0.0325
SPS	-0.81148 <.0001	-0.31642 0.0885	0.05028 0.7919	0.23774 0.2059	0.07201 0.7053	0.38798 0.0341	0.16602 0.3806	1.00000	0.47492 0.0080
TC	-0.48733 0.0063	0.04470 0.8146	0.24290 0.1959	0.10537 0.5795	-0.05461 0.7744	-0.08489 0.6556	-0.39132 0.0325	0.47492 0.0080	1.00000

There appear to be various levels of both positive and negative correlations between the 9 industry segments. The scatter plot matrix of these segments shown below, display existence of some provisional groups that the data appear to be present within. For example, looking at the financials (FIN) row *versus* services (SER) column there appear to be 3 groups of countries that are clustered. Similarly, manufacturing (MAN) *versus* SER shows some grouping occurring. For the most part, there is a random distribution of points and no grouping.

Conclusion: There are four countries categorized in the other category comprising Gibraltar, Cyprus, Malta and Turkey. Looking at the FIN *versus* SER plot, Gibraltar and Cyprus seem like they would cluster with the EU countries, while Malta and Turkey would segregate with the Eastern group based on proximity to those clusters.

2. Principal Component Analysis

This section of the assignment uses principal component analysis (PCA) to reduce the dimensionality of the data.

Principal Components Analysis using PROC PRINCOMP

The PRINCOMP Procedure

Observations	30
Variables	9

The data currently comprises 30 observations and 9 variables, see table above. Performing PCA, we obtain the eigenvalues and eigenvectors of the 9 principal components (see tables on the right).

How many components should we keep? Based on the metric of keeping sufficient variables such that >80% of the variance is explained, we would keep 4 variables. Looking at the correlation matrix cumulative (above right) and variance explained plot (right), it is noted that with 4 principal components, the cumulative variance is >80%.

Are there any other competing decision rules that you could use? Looking at the table eigenvalues, one could use a decision rule of using only the principal components with eigenvalues > 1.0. In this case, noting the first table above right top, this would still direct us to using 4 principal components.

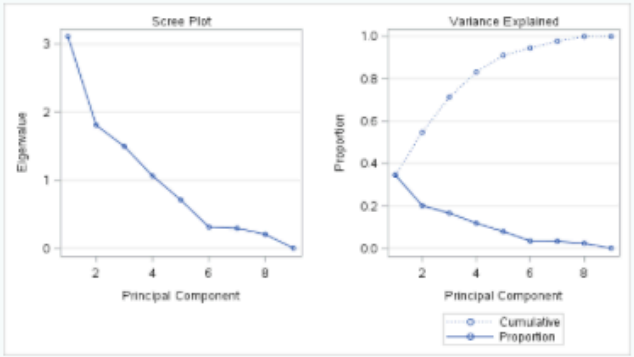
Looking at the scatter plot matrix of the principal components (right), the PCA directs us to using 4 components. Inspection of the 1st row of the matrix, shows groupings of observations in 1-4, but for principal components >5, there is more of a random scatter of points and no groupings evident.

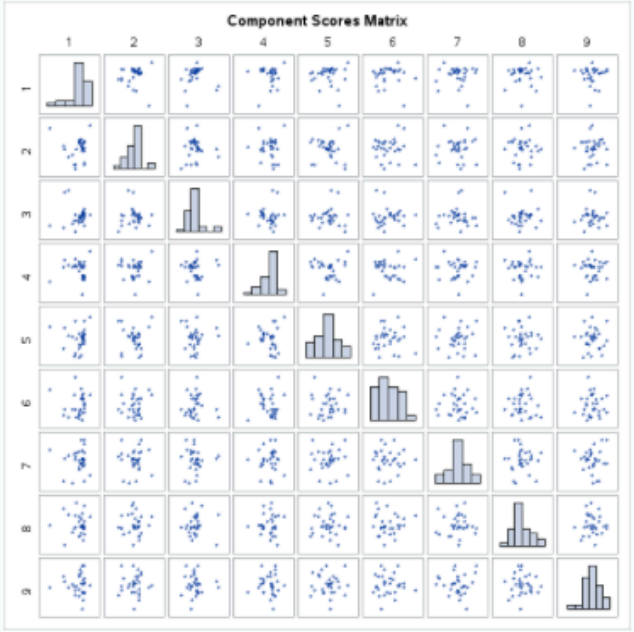
Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	3.11225795	1.30302071	0.3458	0.3458
2	1.80923724	0.31301704	0.2010	0.5468
3	1.49622020	0.43277636	0.1662	0.7131
4	1.06344384	0.35318631	0.1182	0.8312
5	0.71025753	0.39891874	0.0789	0.9102
6	0.31133879	0.01791787	0.0346	0.9448
7	0.29342091	0.08960446	0.0326	0.9774
8	0.20381645	0.20380935	0.0226	1.0000
9	0.00000710		0.0000	1.0000

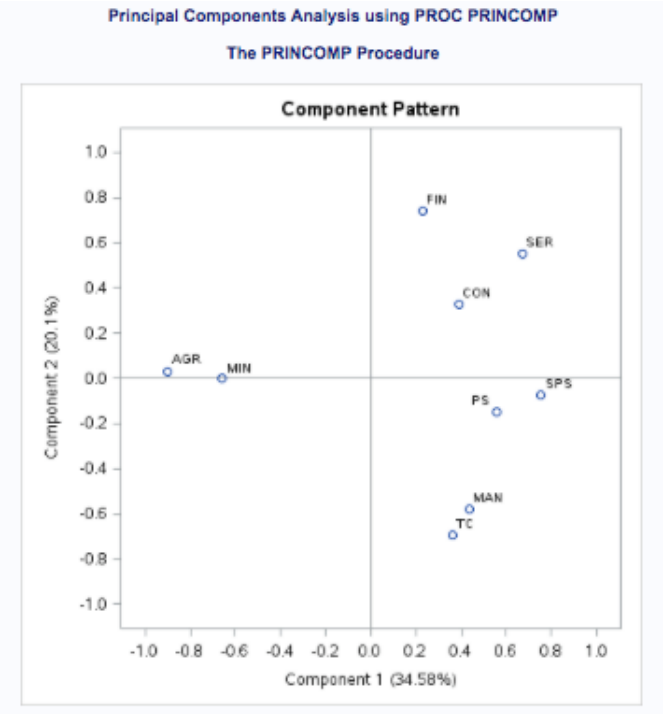
Eigenvectors

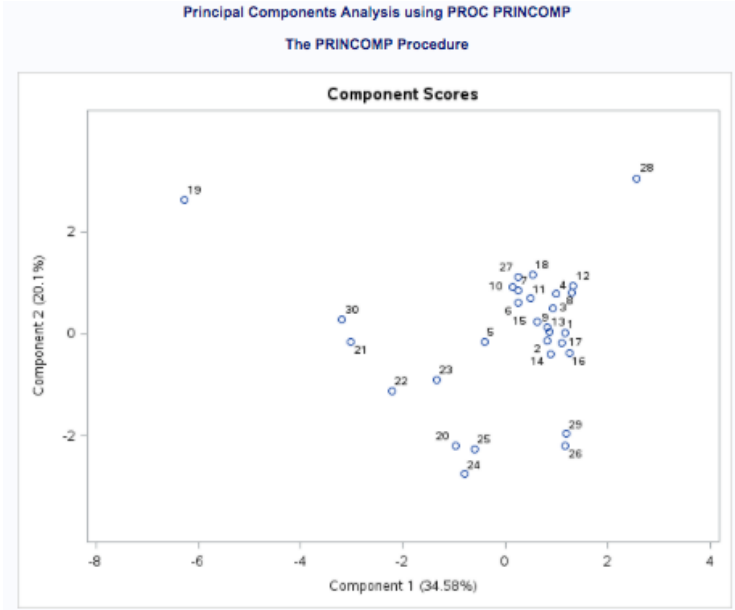
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9
AGR	-.511492	0.023475	-.278591	0.016492	-.024038	0.042397	-.163574	0.540409	0.582036
MIN	-.374983	-.000491	0.515052	0.113606	0.346313	-.198574	0.212590	-.448592	0.418818
MAN	0.246161	-.431752	-.502056	0.058270	-.233622	0.030917	0.236015	-.431757	0.447086
PS	0.316120	-.109144	-.293695	0.023245	0.854448	-.206471	-.060565	0.155122	0.030251
CON	0.221599	0.242471	0.071531	0.782666	0.062151	0.502636	-.020285	0.030823	0.128656
SER	0.381536	0.408256	0.065149	0.169038	-.266673	-.672694	0.174839	0.201753	0.245021
FIN	0.131088	0.552939	-.095654	-.489218	0.131288	0.405935	0.457645	-.027264	0.190758
SPS	0.428162	-.054706	0.360159	-.317243	-.045718	0.158453	-.621330	-.041476	0.410315
TC	0.205071	-.516650	0.412996	-.042063	-.022901	0.141898	0.492145	0.502124	0.060743





Looking at the component pattern (below) showing the factor loadings of the eigenvectors of the principal components, we can analyze one of these pairwise plots produced by PROC PRINCOMP. For example, take principal component 1 (PC1) *versus* PC2: looking at PC1 first we see AGR and MIN grouped together on the left-side (negative), while on the positive right-side we see FIN/SER/CON grouped together and TC/MAN/PS/SPS grouped together. For PC1, we interpret the values as follows: we observe the same groupings as we did with PC1 except that for PC2 AGR/MIN is basically zero (so does not play a role) and FIN/SER/CON are positive, while TC/MAN/PS/SPS is negative. So for PC2, there are two clusters the FIN/SER/CON and TC/MAN/PS/SPS clusters.





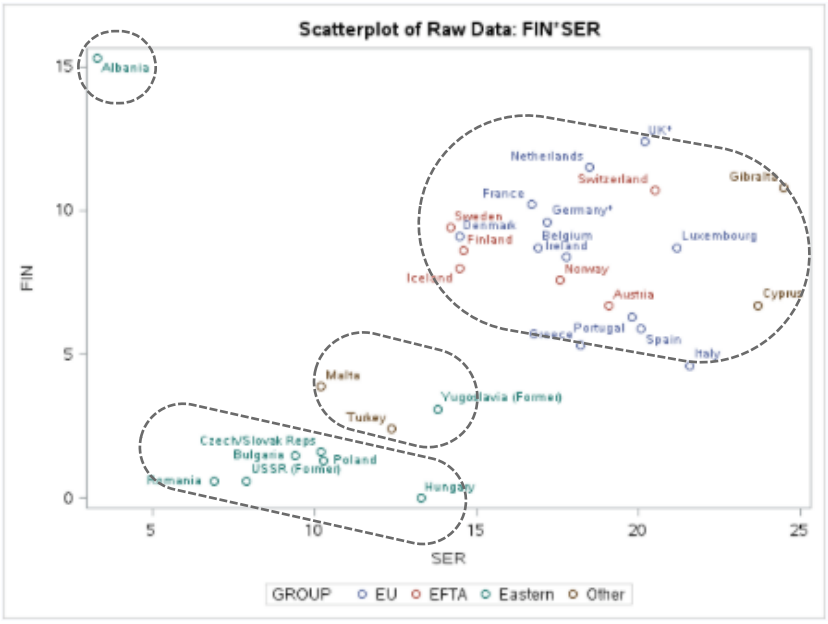
Looking at the component scores scatter plot (above right), for the 30 numbered observations one can see several distinct groupings of the data points.

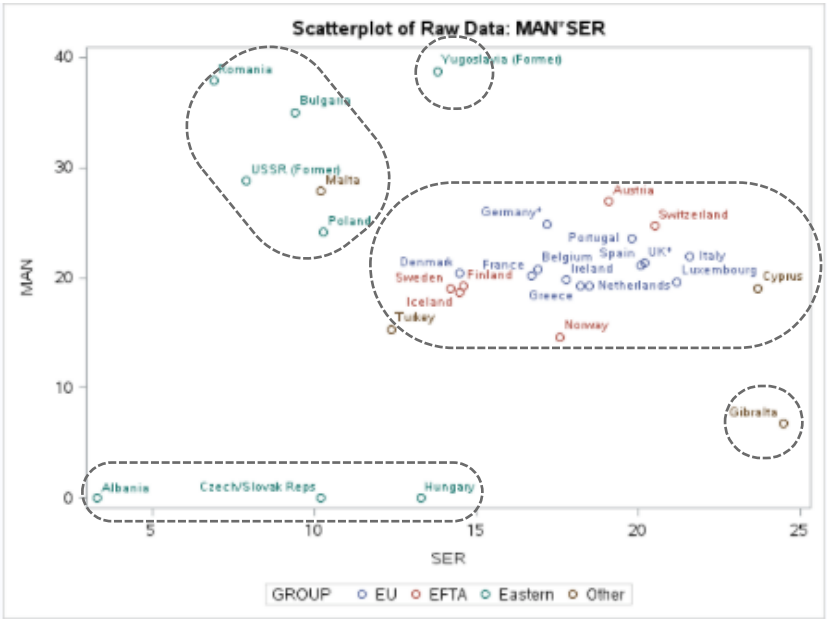
Conclusion: In this step we reduced the dimensionality of the variables from 9 to 4 using PCA, based on the metric of choosing enough components that explain >80% of the variance. Looking at the scatterplots of the principal components we can see some clustering taking place (even with zero correlations), and similarly with the component score scatter plot, suggesting together that there is evidence to continue and perform actual clustering analysis.

3. Cluster analysis

Scatterplots of financials, FIN *versus* services, SER (below top) and manufacturing, MAN *versus* services, SER (below bottom), are shown below.

*How many clusters do you see in the scatterplot of FIN*SER? How many clusters do you see in the scatterplot of MAN*SER?* For the FIN*SER one can conceptualize at least 4 clusters (dotted lines), while for the MAN*SER there could be at least 5 clusters (dotted lines).





Thus, from this it is observed that different projections produce different clustering results.

How do we interpret the measures of CCC, Pseudo F, and Pseudo T-Squared? How do we interpret the plots for these three measures?

Performing the clustering analysis, we need to interpret the optimal number of clusters to use using a combination of the three metrics described below:

a) CCC: CCC is the cubic clustering criterion, which compares the R-squared obtained with a certain number of clusters, to the R-squared you would get by clustering a uniformly distributed set of points. This value is interpreted as one would for R-squared, that is, the higher the better.

b) Pseudo-F: This is intended to capture the tightness of the clusters and is the ratio of between-cluster variance to within-cluster variance. That is, it provides a measure of how separated the clusters are.

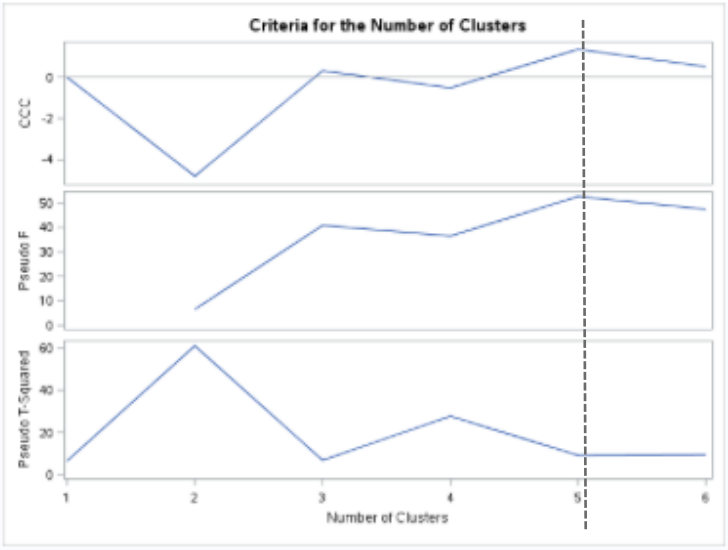
c) Pseudo T-squared: provides a measure that quantifies the difference in the ratio of between-cluster variance to within-cluster variance when clusters are merged at a given step (*i.e.* pseudo T-squared is working backwards, from right to left on the plot). If there is a distinct jump in pseudo T-squared with x number of clusters, then x+1 represents the optimal number of clusters.

The raw cluster history output from building the clusters, with no pre-conceived notion of how many clusters to use, is shown below.

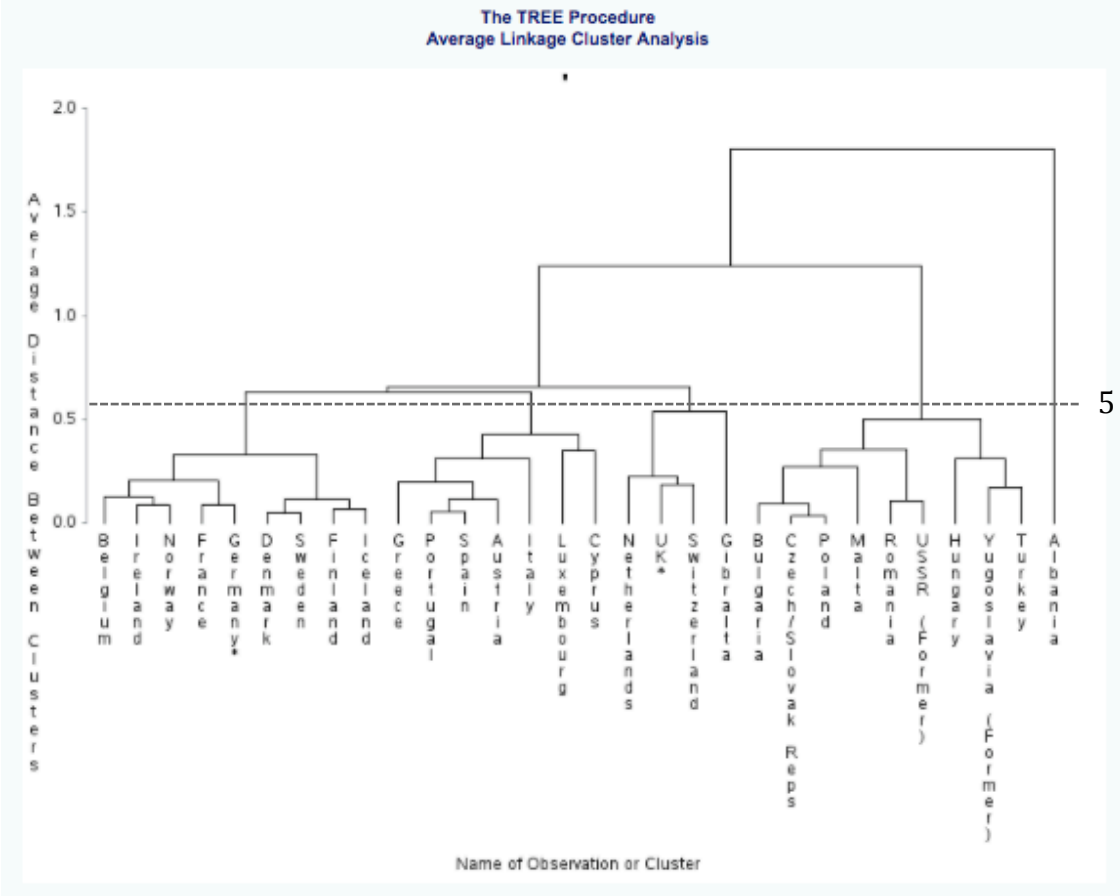
Cluster History

Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Approximate Expected R-Square	Cubic Clustering Criterion	Pseudo F Statistic	Pseudo t-Squared	Norm RMS Distance	Tie
29	Czech/Slovak Reps	Poland	2	0.0000	1.00	.	.	881	.	0.0343	
28	Denmark	Sweden	2	0.0001	1.00	.	.	652	.	0.046	
27	Portugal	Spain	2	0.0001	1.00	.	.	537	.	0.0542	
26	Finland	Iceland	2	0.0002	1.00	.	.	438	.	0.066	
25	France	Germany*	2	0.0002	.999	.	.	340	.	0.0847	
24	Ireland	Norway	2	0.0003	.999	.	.	294	.	0.0894	
23	Bulgaria	CL29	3	0.0004	.999	.	.	248	9.7	0.0939	
22	Romania	USSR (Former)	2	0.0004	.998	.	.	226	.	0.1084	
21	CL28	CL26	4	0.0008	.998	.	.	183	6.9	0.1127	
20	CL27	Austria	3	0.0006	.997	.	.	173	5.8	0.116	
19	Belgium	CL24	3	0.0006	.996	.	.	167	2.2	0.1236	
18	Yugoslavia (Former)	Turkey	2	0.0010	.995	.	.	151	.	0.1697	
17	UK*	Switzerland	2	0.0012	.994	.	.	138	.	0.1872	
16	Greece	CL20	4	0.0019	.992	.	.	119	5.6	0.2009	
15	CL19	CL25	5	0.0029	.989	.	.	99.1	7.8	0.2037	
14	Netherlands	CL17	3	0.0019	.987	.	.	96.9	1.5	0.2214	
13	CL23	Malta	4	0.0036	.984	.	.	86.1	16.8	0.269	
12	CL16	Italy	5	0.0048	.979	.	.	76.4	5.5	0.3104	
11	Hungary	CL18	3	0.0041	.975	.	.	73.8	4.1	0.3108	
10	CL15	CL21	9	0.0141	.961	.	.	54.5	19.5	0.3272	
9	Luxembourg	Cyprus	2	0.0042	.957	.	.	58.0	.	0.3472	
8	CL13	CL22	6	0.0098	.947	.	.	56.0	8.8	0.3526	
7	CL12	CL9	7	0.0127	.934	.	.	54.4	5.5	0.4254	
6	CL8	CL11	9	0.0263	.908	.900	0.50	47.3	9.5	0.4996	
5	CL14	Gibralta	4	0.0141	.894	.869	1.35	52.6	9.2	0.5369	
4	CL10	CL7	16	0.0859	.808	.822	-.52	36.4	27.7	0.6305	
3	CL4	CL5	20	0.0564	.751	.741	0.31	40.8	6.9	0.6595	
2	CL3	CL6	29	0.5610	.190	.570	-4.8	6.6	60.9	1.2374	
1	CL2	Albania	30	0.1904	.000	.000	0.00	.	6.6	1.806	

The criteria for the number of clusters to use, derived from CCC, pseudo-F and pseudo t-squared is shown below:



Using these 3 metrics we can choose 5 clusters (shown by dotted line) as the optimal number of clusters to select for the tree drawing aspect of the assignment.

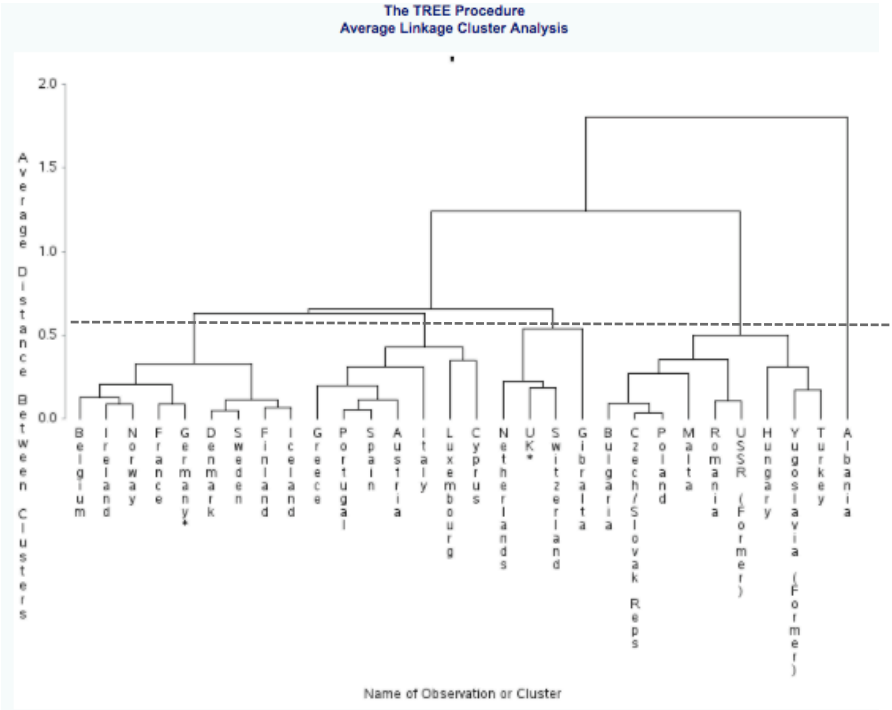


Invoking the SAS TREE procedure draws out the tree (see above); the dotted line shows that if one cuts the tree at the point in the tree indicated, five clusters will be released from the tree. The raw output table for an n=5-cluster tree is shown below for the 30 observations. For each observation the values of the FIN and SER variables are shown as is the cluster number (1-5) and arbitrary name of the cluster.

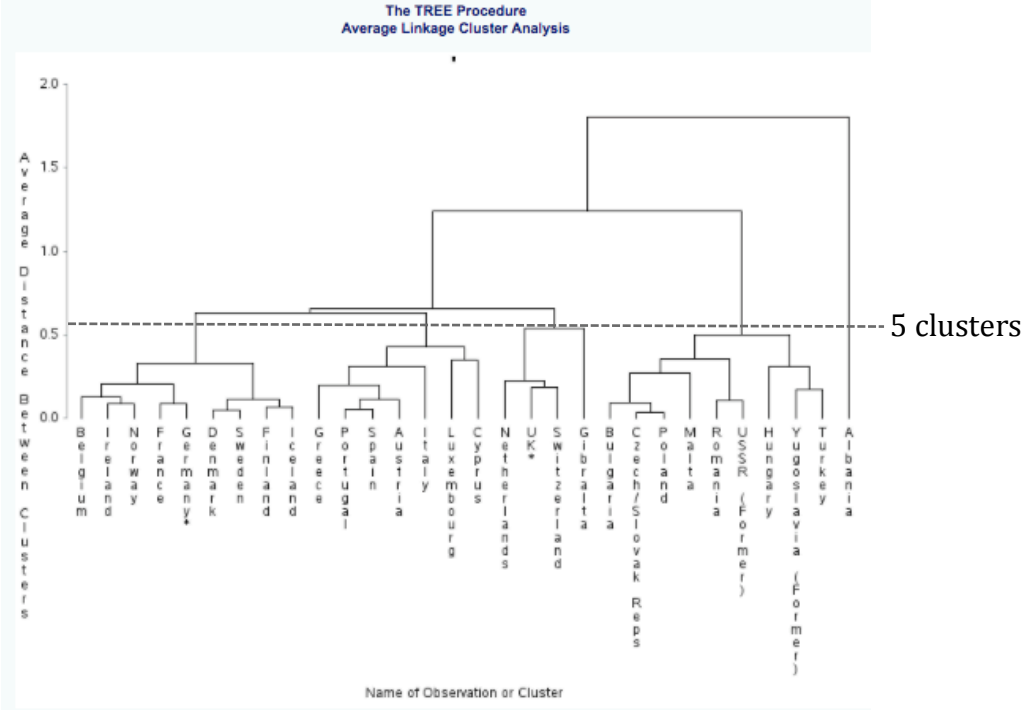
Obs	_NAME_	FIN	SER	CLUSTER	CLUSNAME
1	Czech/Slovak Reps	1.6	10.2	1	CL6
2	Poland	1.3	10.3	1	CL6
3	Denmark	9.1	14.5	2	CL10
4	Sweden	9.4	14.2	2	CL10
5	Portugal	6.3	19.8	3	CL7
6	Spain	5.9	20.1	3	CL7
7	Finland	8.6	14.6	2	CL10
8	Iceland	8.0	14.5	2	CL10
9	France	10.2	16.7	2	CL10
10	Germany*	9.6	17.2	2	CL10
11	Ireland	8.4	17.8	2	CL10
12	Norway	7.6	17.6	2	CL10
13	Bulgaria	1.5	9.4	1	CL6
14	Romania	0.6	6.9	1	CL6
15	USSR (Former)	0.6	7.9	1	CL6
16	Austria	6.7	19.1	3	CL7
17	Belgium	8.7	16.9	2	CL10
18	Yugoslavia (Former)	3.1	13.8	1	CL6
19	Turkey	2.4	12.4	1	CL6
20	UK*	12.4	20.2	4	CL5
21	Switzerland	10.7	20.5	4	CL5
22	Greece	5.3	18.2	3	CL7
23	Netherlands	11.5	18.5	4	CL5
24	Malta	3.9	10.2	1	CL6
25	Italy	4.6	21.6	3	CL7
26	Hungary	0.0	13.3	1	CL6
27	Luxembourg	8.7	21.2	3	CL7
28	Cyprus	6.7	23.7	3	CL7
29	Gibralta	10.8	24.5	4	CL5
30	Albania	15.3	3.3	5	Albania

This tree was based on the assumption of a cluster number of n=5. However, if we wanted to draw trees for n=4 and n=3 we would generate the trees shown below, which are the same topology as the one above. Note the dotted line shows if the tree is cut at the specific location 5 cluster groups will fall out.

Cluster size (n)=4; dotted line is n=5 clusters



Cluster size (n)=3; dotted line is n=5 clusters

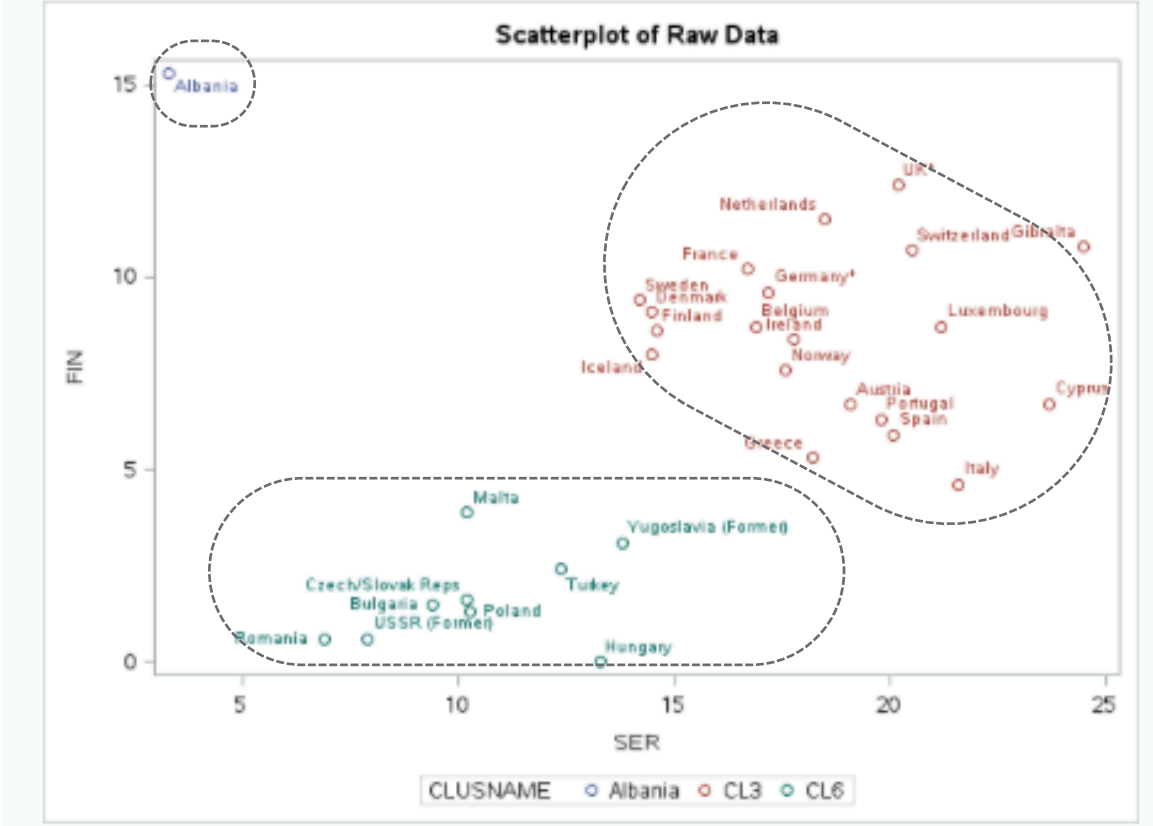


Drawing out the scatter plot of n=3, 4 and 5 cluster size separately for the 30 country observations, we see the output below. Each cluster group has ≥ 1 member.

For n=3 clusters:

The FREQ Procedure

Frequency	Table of GROUP by CLUSNAME			
	CLUSNAME			
GROUP	Albania	CL3	CL6	Total
EFTA	0	6	0	6
EU	0	12	0	12
Eastern	1	0	7	8
Other	0	2	2	4
Total	1	20	9	30



With three clusters, there is not very nuance in the segmentation. The CL3 cluster can be broken up into at least two clusters. The frequency table above shows each cluster has a predominant geopolitical contingent that predominates, so for CL3 it is the EU and for CL6 it is Eastern.

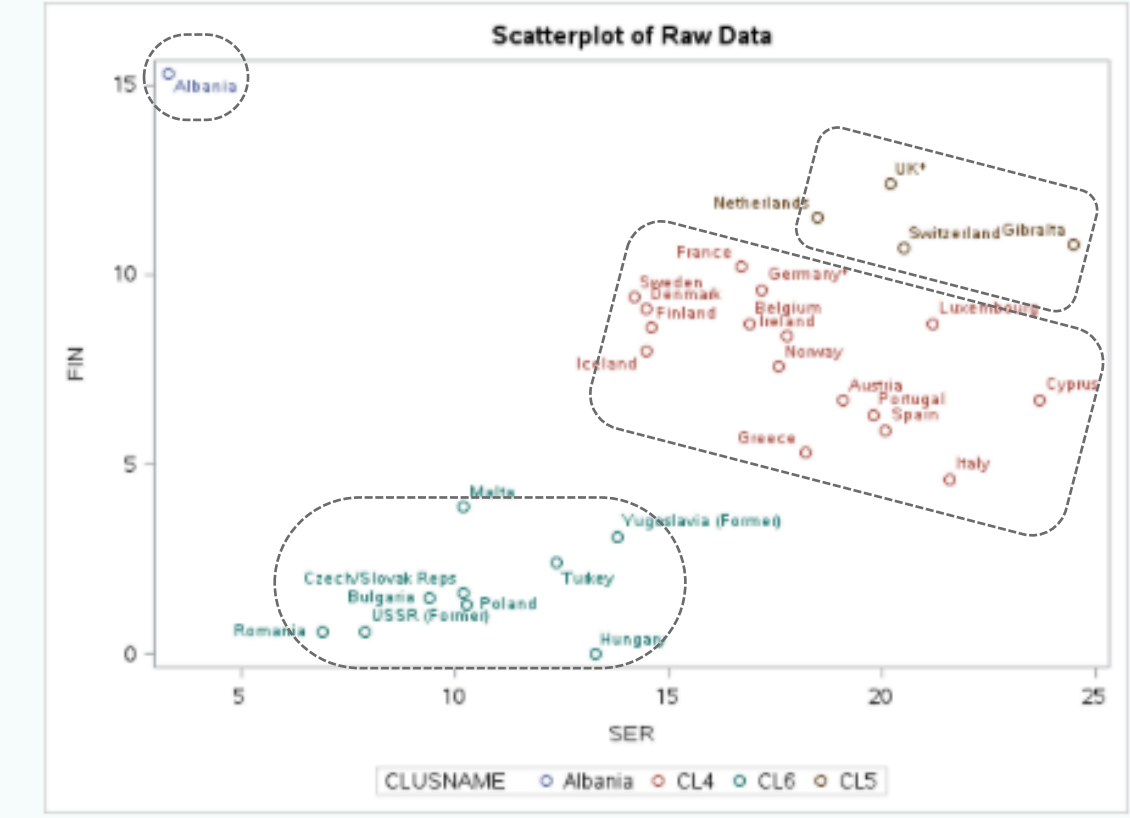
For n=4 clusters:

Scatterplot of Raw Data

The FREQ Procedure

Frequency | Table of GROUP by CLUSNAME

	CLUSNAME				
GROUP	Albania	CL4	CL5	CL6	Total
EFTA	0	5	1	0	6
EU	0	10	2	0	12
Eastern	1	0	0	7	8
Other	0	1	1	2	4
Total	1	16	4	9	30



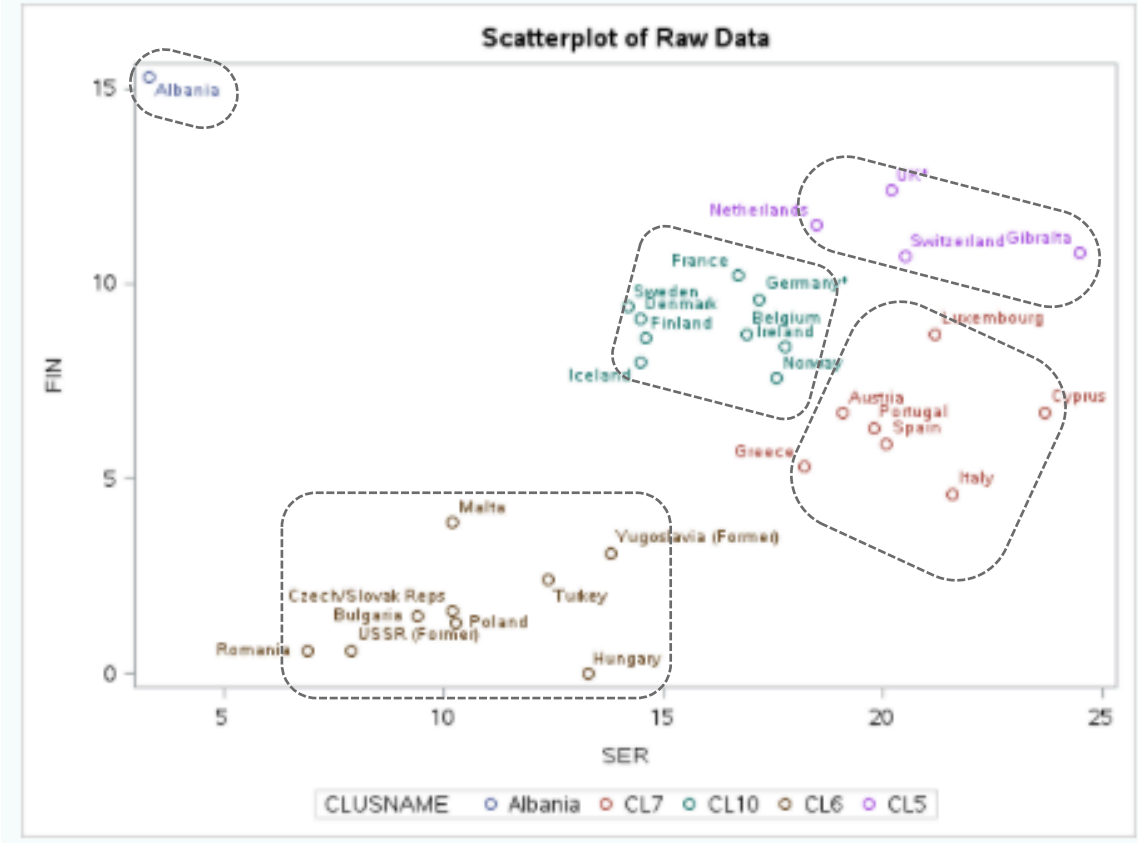
With four clusters, the CL6 and Albania (single) cluster are unchanged from above, but the CL3 cluster is split into two clusters – CL4 (largest member cluster) and CL5 (UK, Gibraltar, Switzerland and the Netherlands). The frequency table above again shows each cluster has a predominant geopolitical contingent that predominates, so for CL4 and CL5 it is the EU. For CL6 it is Eastern and Albania is unchanged.

For n=5 clusters:

Scatterplot of Raw Data

The FREQ Procedure

Frequency	Table of GROUP by CLUSNAME					
	CLUSNAME					
GROUP	Albania	CL10	CL5	CL6	CL7	Total
EFTA	0	4	1	0	1	6
EU	0	5	2	0	5	12
Eastern	1	0	0	7	0	8
Other	0	0	1	2	1	4
Total	1	9	4	9	7	30



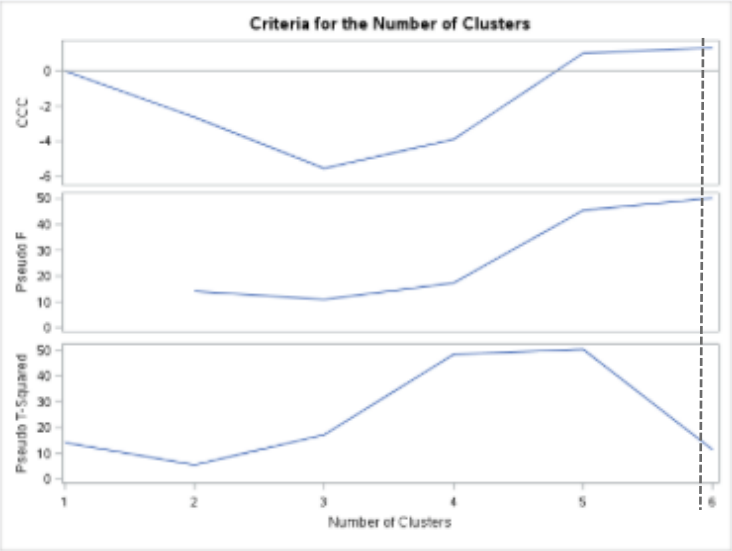
With five clusters, there is greater nuance in the segmentation of the former CL3 cluster. Albania and the Eastern (CL6) cluster members remained the same. The CL3 cluster (from the n=3 cluster version) has now become broken up into three clusters – CL5, CL7 and CL10. The frequency table for n=5 above shows each cluster has members from different geopolitical

contingents within the cluster, so for CL10 it is the EU and EFTA equally split, and similarly for CL5 while for CL7 EU member countries predominate. With the n=5 clusters since there is greater nuance in the segmentation of the large n=3 CL3 cluster, this is preferred over n=3 and n=4 cluster sizes.

Using principal component analysis to perform a dimensionality reduction of variables (principal components 1 and 2), we can re-evaluate the cluster analysis. The overall raw cluster history results table is shown below.

Cluster History

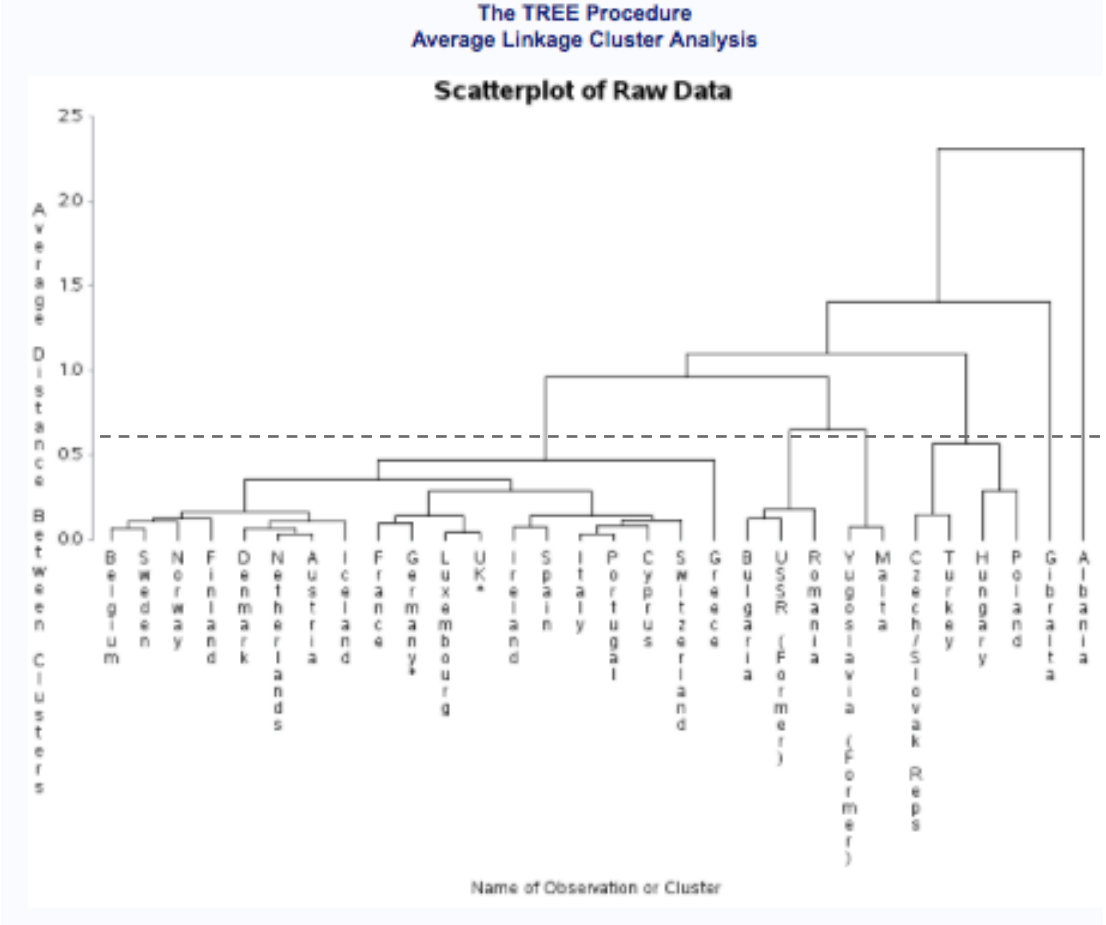
Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Approximate Expected R-Square	Cubic Clustering Criterion	Pseudo F Statistic	Pseudo t-Squared	Norm RMS Distance	Tie
29	Netherlands	Austria	2	0.0000	1.00	.	.	1079	.	0.031	
28	Italy	Portugal	2	0.0000	1.00	.	.	940	.	0.0364	
27	Luxembourg	UK*	2	0.0001	1.00	.	.	761	.	0.046	
26	Belgium	Sweden	2	0.0002	1.00	.	.	512	.	0.0683	
25	Denmark	CL29	3	0.0002	.999	.	.	396	6.5	0.07	
24	Yugoslavia (Former)	Malta	2	0.0002	.999	.	.	360	.	0.0758	
23	Ireland	Spain	2	0.0002	.999	.	.	338	.	0.0793	
22	CL28	Cyprus	3	0.0003	.999	.	.	312	6.1	0.0798	
21	France	Germany*	2	0.0003	.998	.	.	292	.	0.0965	
20	CL26	Norway	3	0.0005	.998	.	.	263	2.8	0.1053	
19	CL25	Iceland	4	0.0006	.997	.	.	239	4.5	0.1089	
18	CL22	Switzerland	4	0.0007	.997	.	.	219	4.1	0.1201	
17	Bulgaria	USSR (Former)	2	0.0005	.996	.	.	217	.	0.1234	
16	CL20	Finland	4	0.0007	.996	.	.	210	2.2	0.1279	
15	CL23	CL18	6	0.0013	.994	.	.	187	4.2	0.137	
14	CL21	CL27	4	0.0012	.993	.	.	177	6.2	0.1435	
13	Czech/Slovak Reps	Turkey	2	0.0008	.992	.	.	183	.	0.1497	
12	CL16	CL19	8	0.0026	.990	.	.	157	7.5	0.1634	
11	CL17	Romania	3	0.0013	.988	.	.	162	2.4	0.178	
10	CL14	CL15	10	0.0116	.977	.	.	93.6	22.8	0.2862	
9	Hungary	Poland	2	0.0028	.974	.	.	98.2	.	0.2874	
8	CL12	CL10	18	0.0290	.945	.	.	53.9	22.8	0.355	
7	CL8	Greece	19	0.0116	.933	.	.	53.6	4.0	0.4664	
6	CL13	CL9	4	0.0205	.913	.892	1.32	50.2	11.3	0.5691	
5	CL11	CL24	5	0.0337	.879	.859	1.02	45.5	50.4	0.6457	
4	CL7	CL5	24	0.2127	.666	.808	-3.9	17.3	48.4	0.9639	
3	CL4	CL6	28	0.2198	.447	.721	-5.5	10.9	17.1	1.0973	
2	CL3	Gibralta	29	0.1118	.335	.508	-2.6	14.1	5.5	1.4022	
1	CL2	Albania	30	0.3348	.000	.000	0.00	.	14.1	2.3141	



The criteria for the number of clusters to use, derived from CCC, pseudo-F and pseudo t-squared is shown on the left. Using these 3 metrics we now choose 6 clusters (shown by dotted line) as the optimal number of clusters to select to draw the dendrogram shown below.

Invoking the SAS TREE procedure draws out the tree (see below); the dotted line shows that if one cuts the tree at the point in the tree

indicated, six (6) clusters will be released from the tree.



The cluster tables for n=3, n=4, n=5 and n=6 cluster sizes are shown below.

n=3 clusters

Scatterplot of Raw Data

The FREQ Procedure

Frequency	Table of GROUP by CLUSNAME			
	CLUSNAME			
GROUP	Albania	CL3	Gibralta	Total
EFTA	0	6	0	6
EU	0	12	0	12
Eastern	1	7	0	8
Other	0	3	1	4
Total	1	28	1	30

n=5 clusters

Scatterplot of Raw Data

The FREQ Procedure

Frequency	Table of GROUP by CLUSNAME					
	CLUSNAME					
GROUP	Albania	CL5	CL6	CL7	Gibralta	Total
EFTA	0	0	0	6	0	6
EU	0	0	0	12	0	12
Eastern	1	4	3	0	0	8
Other	0	1	1	1	1	4
Total	1	5	4	19	1	30

n=4 clusters

Scatterplot of Raw Data

The FREQ Procedure

Frequency | Table of GROUP by CLUSNAME

GROUP	CLUSNAME				
	Albania	CL4	CL6	Gibralta	Total
EFTA	0	6	0	0	6
EU	0	12	0	0	12
Eastern	1	4	3	0	8
Other	0	2	1	1	4
Total	1	24	4	1	30

n=6 clusters

Scatterplot of Raw Data

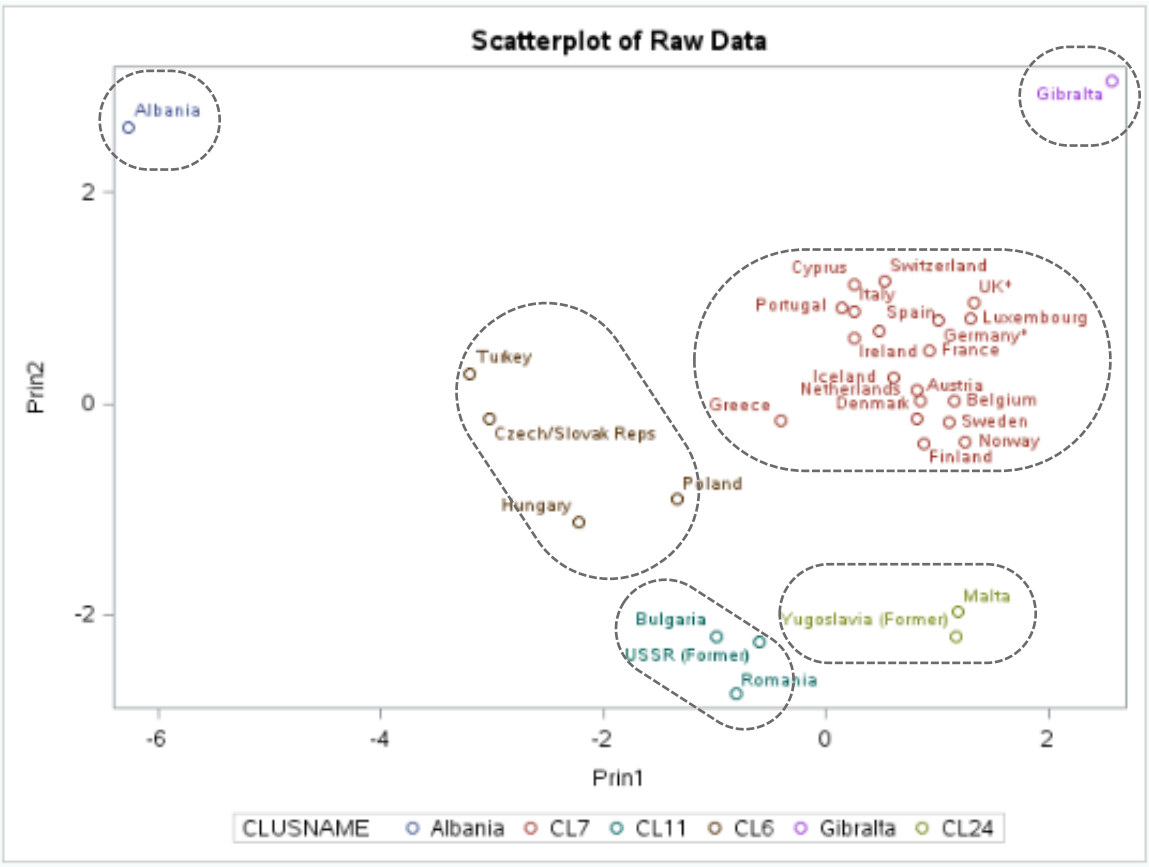
The FREQ Procedure

Frequency | Table of GROUP by CLUSNAME

GROUP	CLUSNAME						
	Albania	CL11	CL24	CL6	CL7	Gibralta	Total
EFTA	0	0	0	0	6	0	6
EU	0	0	0	0	12	0	12
Eastern	1	3	1	3	0	0	8
Other	0	0	1	1	1	1	4
Total	1	3	2	4	19	1	30

Looking at the n=6 cluster frequency table we can see that there are some single country clusters such as Albania and Gibraltar, a 2 member cluster (CL24) and the other clusters contain a predominant geopolitical grouping. An example of the latter includes CL6 and CL11 where Eastern predominates while for cluster CL7 the EU dominates its member participation.

The scatter plot output for n=6 cluster size showing data for all the 30 country observations, is displayed below. Each cluster group has ≥ 1 member.



Conclusion: The 2 principal component view of the tree and scatter plot provides a much more nuanced and detailed segmentation of the 30 countries, and I prefer this view since it provides a reasonable separation between clusters. In addition, looking at the contents of the clusters makes some sense from a political standpoint since for example, Albania and Gibraltar are somewhat isolated entities, USSR/Bulgaria/Romania are former Soviet bloc countries, Yugoslavia/Malta and standalone countries while the largest cluster contains western and northern European countries.

CONCLUSIONS

This assignment deals with cluster analysis to identify and segment groups using European employment data from various industry segments reported from 30 European nations. The use of statistical methods in SAS (*e.g.* PROC CLUSTER and TREE, PRINCOMP) supported performing segmentation analysis on the European employment data to group various countries into suitable 'bins' for association purposes. Specifically, use of clustering methods to choose the optimal number of clusters in conjunction with building dendrograms to partition the data, facilitated the development of a segmentation model from the primary data. Further, the use of principal component analysis of the 9 original variables together with clustering analysis added better detail to the segmentation scatter plot displays. The results from latter analysis were easier to understand and made more sense from a geopolitical standpoint.