Assignment #8

Scott M. Morgan

Introduction:

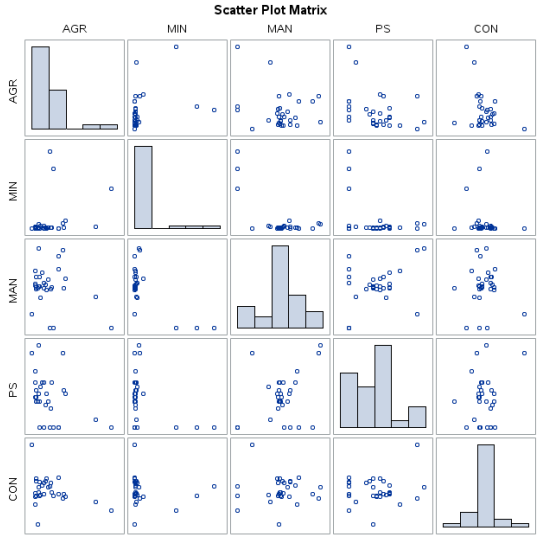
The purpose of this assignment is to perform a cluster analysis starting with a brief correlation analysis and completing with a comparison of results from raw predictor data and cluster results from transformed predictor variables using principal component analysis (PCA). To accomplish this objective, I use the European employment data set provided by the course instructor and SAS Studio to perform clustering analysis with and without PCA. PCA is not a clustering method but can sometimes be used to help reveal clusters through dimensionality reduction. I expect this assignment will be an excellent demonstration of this technique as clustering analysis has a wide range of applications from marketing and social media to city-planning and earthquake studies.

Results:

In the subsequent sections, we conduct an initial correlation analysis, principal components analysis and a cluster analysis. Clustering analysis has been shown to be an effective method to understanding the semantic information and interpreting the structure of the heterogeneous information (Wu, Meng, Deng, Huang, Wu & Badii, 2017). The industries in the data set include AGR (agriculture), MIN (mining), MAN (manufacturing), PS (power and water supply), CON (construction), SER (services), FIN (finance), SPS (social and personal services) and TC (transport and communications). Note that EU stands for European Union, EFTA stands for European Free Trade Association, and Eastern stands for Eastern European nations or the former Eastern Bloc.

Initial Correlation Analysis. In this section, we perform our initial correlation analysis. In Exhibit 1 below, we attempt to examine the Pearson correlation coefficients and accompanying scatterplot matrix.

Exhibit 1: Pearson Correlation Scatter-Plot Matrix



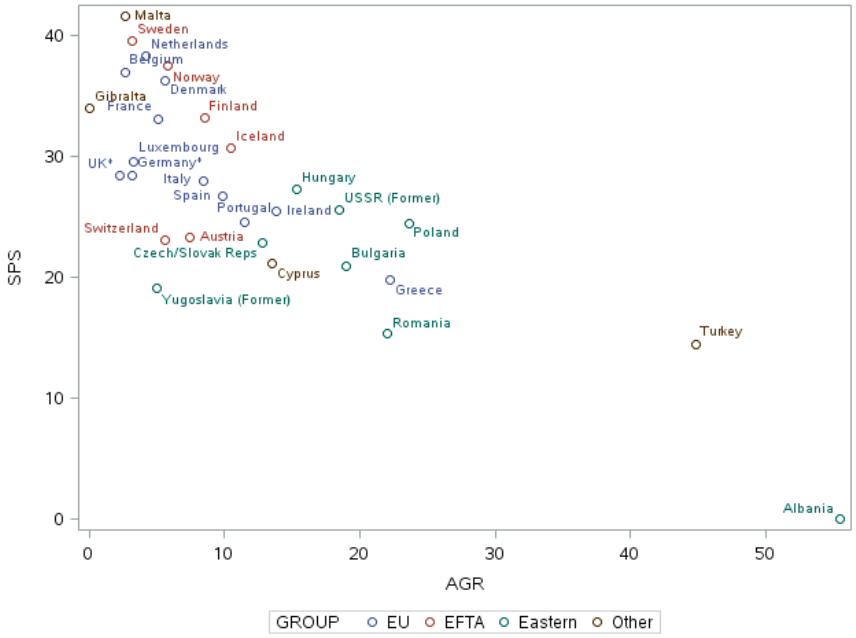
Unfortunately, Exhibit 1 does not include all of the pairwise comparisons. As such, we examine all of the correlation coefficients presented in Table 1 below.

Table 1: Pearson Correlation Coefficients

	AGR	MIN	MAN	PS	CON	SER	FIN	SPS	TC
AGR	1	0.31607	-0.25439	-0.38236	-0.34861	-0.60471	-0.17575	-0.81148	-0.48733
		0.0888	0.1749	0.037	0.059	0.0004	0.3529	<.0001	0.0063
MIN	0.31607	1	-0.67193	-0.38738	-0.12902	-0.40655	-0.24806	-0.31642	0.0447
	0.0888		<.0001	0.0344	0.4968	0.0258	0.1863	0.0885	0.8146
MAN	-0.25439	-0.67193	1	0.38789	-0.03446	-0.03294	-0.27374	0.05028	0.2429
	0.1749	<.0001		0.0342	0.8565	0.8628	0.1433	0.7919	0.1959
PS	-0.38236	-0.38738	0.38789	1	0.1648	0.15498	0.09431	0.23774	0.10537
	0.037	0.0344	0.0342		0.3842	0.4135	0.6201	0.2059	0.5795
CON	-0.34861	-0.12902	-0.03446	0.1648	1	0.47308	-0.01802	0.07201	-0.05461
	0.059	0.4968	0.8565	0.3842		0.0083	0.9247	0.7053	0.7744
SER	-0.60471	-0.40655	-0.03294	0.15498	0.47308	1	0.37928	0.38798	-0.08489
	0.0004	0.0258	0.8628	0.4135	0.0083		0.0387	0.0341	0.6556
FIN	-0.17575	-0.24806	-0.27374	0.09431	-0.01802	0.37928	1	0.16602	-0.39132
	0.3529	0.1863	0.1433	0.6201	0.9247	0.0387		0.3806	0.0325
SPS	-0.81148	-0.31642	0.05028	0.23774	0.07201	0.38798	0.16602	1	0.47492
	<.0001	0.0885	0.7919	0.2059	0.7053	0.0341	0.3806		0.008
TC	-0.48733	0.0447	0.2429	0.10537	-0.05461	-0.08489	-0.39132	0.47492	1
	0.0063	0.8146	0.1959	0.5795	0.7744	0.6556	0.0325	0.008	

Only two pairs of variables have test statistics <0.0001; those being Mining (MIN) and Manufacturing (MAN) as well as Social and Personal Services (SPS) and Agriculture (AGR). We will be analyzing the SPS and AGR industries as they are analytically interesting in the sense that they represent contrasting economic systems of developing versus developed countires. They also have the strongest correlation between variables at -0.81148. We expect countries that are developed to have higher levels of employment in SPS as they are more service-based economies. Conversely, we expect countries that are less developed (i.e. emerging) to have a larger portion of the workforce in AGR as they are more agriculture-based economies. Exhibit 2 below provides a scatterplot between SPS and AGR.

Exhibit 2: Scatterplot of SPS to AGR by Group



More of the developed countries (UK, Germany, Norway, Iceland, etc.) that are part of the EU and EFTA likely have a greater portion of their employment in SPS or other service-based industries. For reference, Wikipedia has an extensive list of developed countries. Those countries with more agriculture employment appear to be developing countries in Eastern Europe and Other. Note that in this data there are four countries that do not belong to any of the three primary groups: Gibralta, Malta, Cyrpus and Turkey. If these had to be assigned to one of the other groups (EU, EFTA and Eastern), I would include all of them in the EU. Cyprus and Malta are already in the EU while Turkey has been in various stages of EU accession negotiations since 1987. Gibralta is a British Overseas Territory on Spain's coast. The United Kingdom has been a member of the EU since 1973 according to Europa.edu and, as such, I would group Gibralta in the EU. For reference in the referendum of 23 June 2016, a majority voted for the United Kingdom to withdraw from the European Union, but the process and date for Brexit have yet to be determined (Inman, 2017).

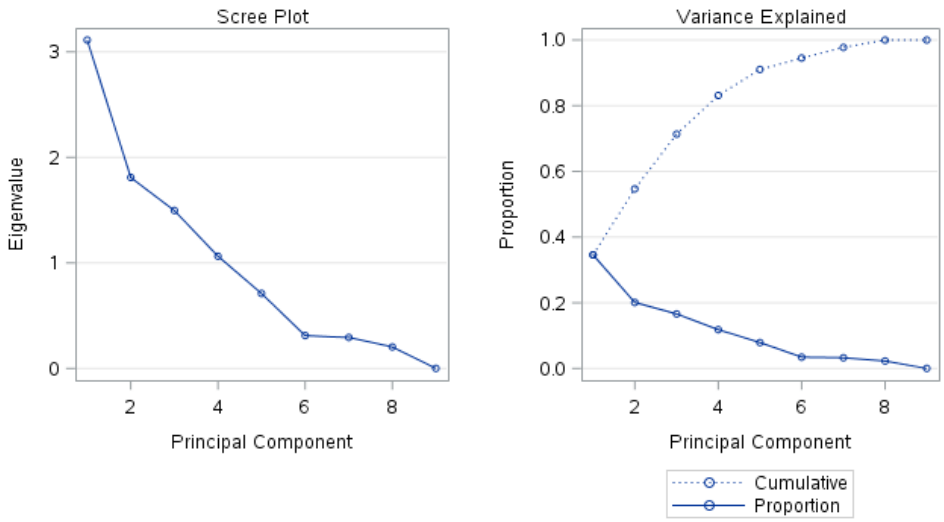
Principal Components Analysis (PCA). Next, we are interested in reducing the dimensionality of our data set which has a total of 9 variables. We will use principal components analysis (PCA) for this. Table 2 below provides the Eigenvalues of the reduced correlation matrix. There is 1 principal component for every variable in the data set. For interpretative purposes, starting with the initial observation, the first principal component is equivalent to about 3.1 of the original variables.

Table 2: Eigenvalues of the Reduced 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

Exhibit 3 below provides the scree plot by principal component as well as a line illustrating the cumulative variance explained. Scree plots are useful for finding an upper bound for the number of components that should be retained for modeling. It appears that the variance explained drops after the fifth principal component.

Exhibit 3: Scree Plot with Variance Explained

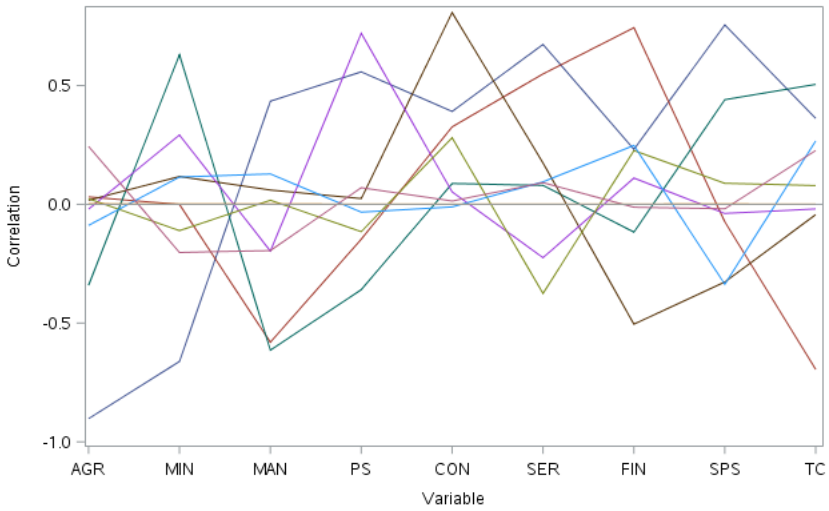


From both Table 2 and Exhibit 3, we decide which principal components to include for clustering. There are several options for deciding how many principal components to use:

- **Option 1)** Eigenvalues greater than 1.0;
- **Option 2)** Scree test (i.e. “elbow”);
- **Option 3)** Percent of total variance (~80% in this case);
- **Option 4)** A priori specification;
- **Option 5)** Parsimony.

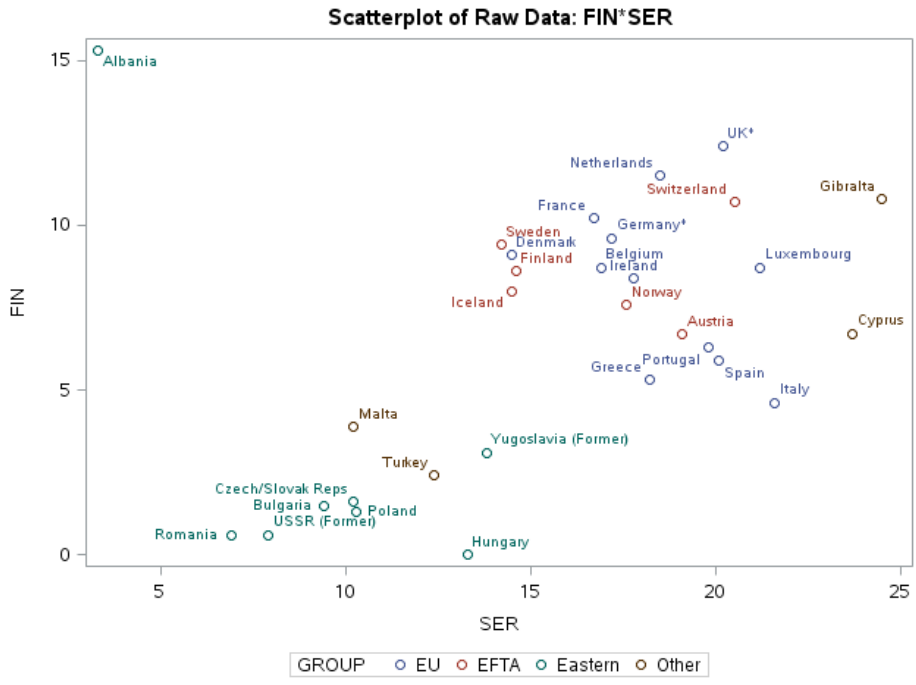
For this exercise, the recommendation is to include 4 principal components based on Options 1, and 3; the first 4 Eigenvalues are greater than 1.0 and the cumulative variance explained is greater than 80% at 4 Eigenvalues. Ideally, we would like to have fewer principal components. Lastly, Exhibit 4 below shows the Component Pattern Profile. As graph depicts, none of the components are correlated evenly across the variables.

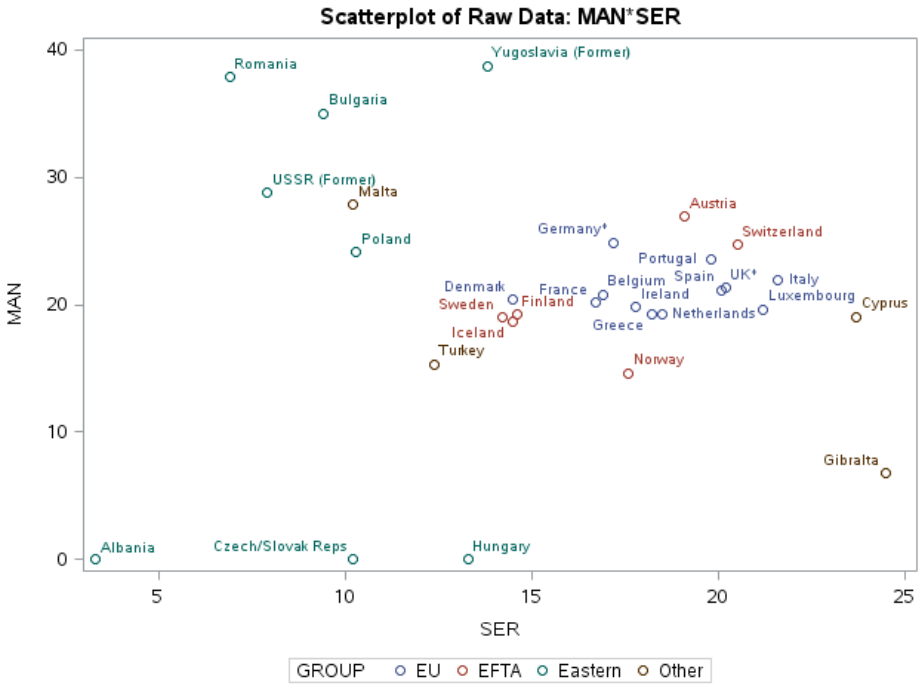
Exhibit 4: Component Pattern Profiles.



Cluster Analysis with Raw Data. We will begin our discussion of cluster analysis by generating a sample pair of scatterplots. Exhibit 5 provides scatterplots of FIN and SER in addition to MAN and SER.

Exhibit 5: Scatterplots of Raw Data





We observe 2 clusters and 1 outlier in the FIN and SER chart. In the lower graph of MAN and SER, we see 2 clusters as well as several outlying countries. Across the scatterplots, EU and EFTA countries seem to cluster together as do the Eastern countries. Of the 4 'Other' countries, Cyrpus appears to be grouping with the EU and EFTA countries while Malta and Turkey are generally with the Eastern countries. Gibralta is near the EU and EFTA countries in the top graph and an outlier in the bottom. Clearly, different projections of the data set will produce different clustering results.

We now use PROC CLUSTER in SAS to create a set of clusters algorithmically. Note that this function performs hierarchical clustering and therefore we do not need to specify the number of clusters in advance. We instead examine statistical outputs provided by SAS. Exhibit 6 provides the criteria for the number of clusters. The printed output is interpreted as follows according to the SAS Support website:

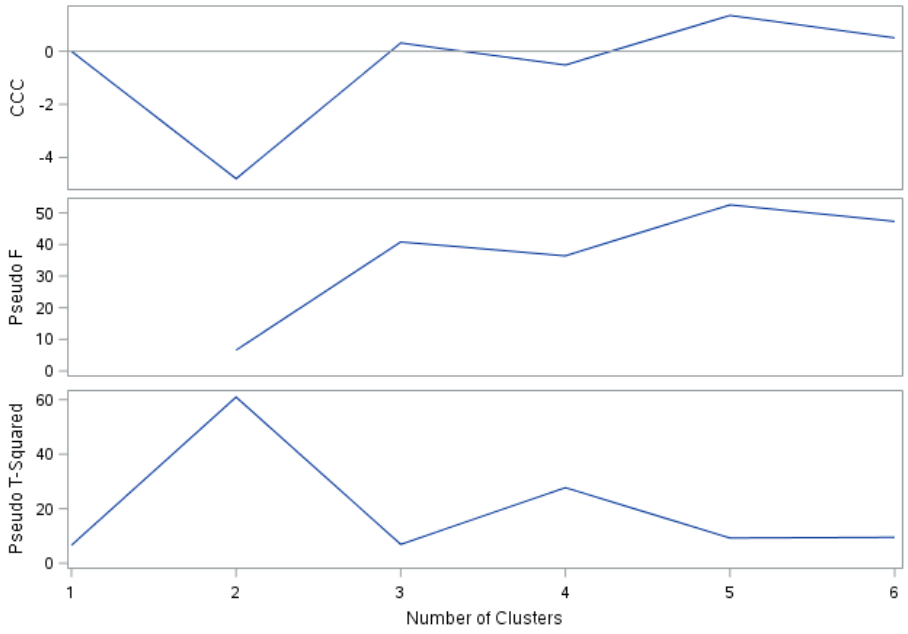
- *Cubic Clustering Criterion (CCC):*
    - Peaks on the plot with the CCC greater than 2 or 3 indicate good clusterings.
    - Peaks with the CCC between 0 and 2 indicate possible clusters but should be interpreted cautiously.
    - There may be several peaks if the data has a hierarchical structure.
    - Very distinct nonhierarchical spherical clusters usually show a sharp rise before the peak followed by a gradual decline.
    - If all values of the CCC are negative and decreasing for two or more clusters, the distribution is probably unimodal or long-tailed.
    - Very negative values of the CCC, say, -30, may be due to outliers. Outliers generally should be removed before clustering.

- *Pseudo F:*
    - Look for a relatively large value.

- *Pseudo T-Squared:*
    - Start at the top of the printed output and look for the first relatively large value, then move back up one cluster.
    - Look for a relatively large value.

The CCC method exhibits a sharp incline between 2 to 3 clusters, 2 peaks at 3 and 5 clusters followed by a gradual decline. Under the Pseudo F measure, between 4 or 5 clusters would be acceptable given the relatively large value. Lastly, the Pseudo T-Squared implies between 2 to 4 clusters. It is recommended to look for consensus when possible among the 3 statistics. Specifically, look for local peaks of the CCC and Pseudo F statistics combined with a small value of the Pseudo T-Squared statistic followed by a larger Pseudo T-Squared. Given this, the recommendation would be to use between 3 to 4 clusters. According to the SAS support website, it is important to note that these criteria are appropriate only for compact or slightly elongated clusters, preferably clusters that are roughly multivariate normal.

Exhibit 6: Criteria for the Number of Clusters



We use PROC TREE to assign our data to a set number of clusters. Tables 3 and 4 below provide the output for the 3 cluster tree and 4 cluster tree, respectively. A large portion of the 3 cluster table is concentrated into a single cluster (CL3). With the 4 cluster table, CL3 is broken up into 2 separate clusters. All the EFTA and EU countries are either in CL3 (Table 3) or distributed across CL4 and CL5 (Table 4). For simplicity, we prefer to use 3 clusters.

Table 3: 3 Cluster Tree Frequency by Group

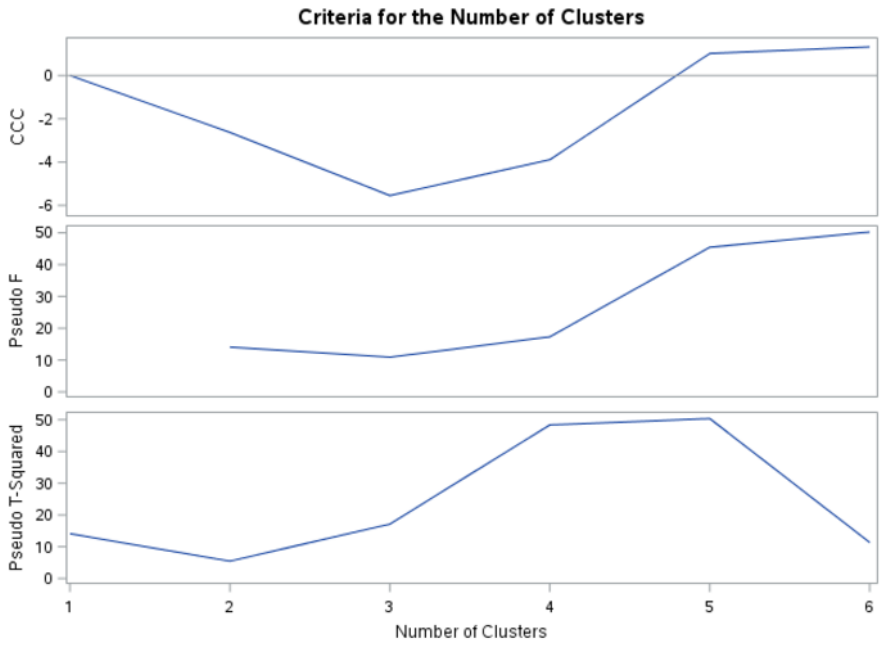
	Albania	CL3	CL6	Total
EFTA	0	6	0	**6**
EU	0	12	0	**12**
Eastern	1	0	7	**8**
Other	0	2	2	**4**
Total	**1**	**20**	**9**	**30**

Table 4: 4 Cluster Tree Frequency by 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**

Cluster Analysis with PCA Data. Lastly, we conduct a hierarchical clustering with the principal component data set. Exhibit 7 provides the criteria for the number of clusters using PCA.

Exhibit 7: Criteria for the Number of Clusters (PCA)



Following the aforementioned criteria and seeking a consensus between the 3 statistics, the recommendation would be to use at least 5 clusters. We once again use PROC TREE to assign our data to a set number of clusters. Tables 5 and 6 below provide the output for the 3 cluster tree and 4 cluster tree, respectively. Using PCA appears to have highlighted the outlying countries in the data set more.

Intuitively, PCA should help clustering, but as Chang (1983) showed the set of principal components with the largest eigenvalues do not always necessarily capture the cluster structure information.

Table 5: 3 Cluster Tree Frequency by Group

	Albania	CL3	CL6	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**

Table 6: 4 Cluster Tree Frequency by Group

	Albania	CL4	CL5	CL6	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**

Overall, I prefer the cluster analysis using the raw data over the principal component data given the simplicity of the output. Bias and/or clustering illusion, however, is always a concern with this type of analysis, especially when the initial clustering confirms structural elements already present in the data set.

References:

Chang, Wei-Chien. "On Using Principal Components Before Separating a Mixture of Two Multivariate Normal Distributions." Applied Statistics 32.3 (1983): 267. Web.

Inman, Phillip (5 January 2017). "Chief Economist of Bank of England Admits Errors in Brexit Forecasting". *The Guardian*. Retrieved 18 January 2017.

"The Number of Clusters." *SAS/STAT(R) 9.2 User's Guide, Second Edition*. SAS, 30 Apr. 2010. Web. 21 May 2017.

Wu, Jibing, Qinggang Meng, Su Deng, Hongbin Huang, Yahui Wu, and Atta Badii. "Generic, network schema agnostic sparse tensor factorization for single-pass clustering of heterogeneous information networks." Plos One 12.2 (2017): n. pag. Web.

Code:

```
libname mydata "/scs/wtm926/" access=readonly;


data temp;
        set mydata.european_employment;
run;


data employ;
  set mydata.european_employment;


proc contents data=employ;



*************************************************************************************
*******;
* Part 1 - An Initial Correlation Analysis;
*************************************************************************************
*******;



ods graphics on;
proc corr data=employ nomiss plots=matrix(histogram);
   var AGR MIN MAN PS CON SER FIN SPS TC;


proc sgplot data=employ;
  scatter y=SPS x=AGR/ datalabel=country group=group;
```

```
run; quit; ods graphics off;



*************************************************************************************
*******;
* Part 2 - Principal Components Analysis;
*************************************************************************************
*******;

title 'Part 2 - Principal Components Analysis';

PROC PRINCOMP; proc princomp data=employ out=pca_9components outstat=eigenvectors plots=all;
 run;
 ods graphics off;



*************************************************************************************
*******;
* Part 3 - Cluster Analysis;
*************************************************************************************
*******;



ods graphics on;

proc sgplot data=temp;
```

```
title 'Scatterplot of Raw Data: FIN*SER';
scatter y=fin x=ser / datalabel=country group=group;
run;
quit;

ods graphics off;

ods graphics on;

proc sgplot data=temp;
title 'Scatterplot of Raw Data: MAN*SER';
scatter y=man x=ser / datalabel=country group=group;
run; quit;

ods graphics off;



ods graphics on;

 proc cluster data=temp method=average outtree=tree1 pseudo ccc plots=all;
  var fin ser;
  id country;
 run; quit;

ods graphics off;
```

```
****************************************;
**Assign data to a set number Clusters  *;
****************************************;

proc tree data=tree1 ncl=4 out=_4_clusters;
 title 'Four Cluster Tree';
 copy fin ser;

proc tree data=tree1 ncl=3 out=_3_clusters;
 title 'Three Cluster Tree';
 copy fin ser;

****************************************;
**Macro to make tables                 *;
****************************************;

%macro makeTable(treeout,group,outdata);
data tree_data;
        set &treeout.(rename=(_name_=country));
run;

proc sort data=tree_data; by country; run; quit;

data group_affiliation;
```

```
        set &group.(keep=group country);
run;

proc sort data=group_affiliation; by country; run; quit;

data &outdata.;
        merge tree_data group_affiliation;
        by country;
run;

proc freq data=&outdata.;
        table group*clusname / nopercent norow nocol;
run;
%mend makeTable;

****************************************;
**Call Macro Function                *;
****************************************;

%makeTable(treeout=_3_clusters,group=temp,outdata=_3_clusters_with_labels);

****************************************;
**Plot the clusters for a visual display*;
****************************************;
```

```
ods graphics on;
proc sgplot data=_3_clusters_with_labels;
title 'Scatterplot with 3 Clusters';
scatter y=fin x=ser / datalabel=country group=clusname;
run;
quit;
ods graphics off;

%makeTable(treeout=_4_clusters,group=temp,outdata=_4_clusters_with_labels);

****************************************;
**Plot the clusters for a visual display*;
****************************************;

ods graphics on;
proc sgplot data=_4_clusters_with_labels;
title 'Scatterplot with 4 Clusters';
scatter y=fin x=ser / datalabel=country group=clusname;
run;
quit;
ods graphics off;

****************************************;
**Using the first 2 principapl components*;
****************************************;
```

```
ods graphics on;
proc cluster data=pca_9components method=average outtree=tree3 pseudo ccc plots=all;
var prin1 prin2;
id country;
run; quit;
ods graphics off;

ods graphics on;
proc tree data=tree3 ncl=4 out=_4_clusters;
copy prin1 prin2; run; quit;



proc tree data=tree3 ncl=3 out=_3_clusters;
copy prin1 prin2;
run; quit;
ods graphics off;

%makeTable(treeout=_3_clusters,group=temp,outdata=_3_clusters_with_labels);
%makeTable(treeout=_4_clusters,group=temp,outdata=_4_clusters_with_labels);

****************************************;
**Plot the clusters for a visual display*;
****************************************;
```

```
ods graphics on;
proc sgplot data=_3_clusters_with_labels;
title 'Scatterplot of Raw Data';
scatter y=prin2 x=prin1 / datalabel=country group=clusname; run;
quit;
ods graphics off;



****************************************;
**Plot the clusters for a visual display*;
****************************************;

ods graphics on;
proc sgplot data=_4_clusters_with_labels;
title 'Scatterplot of Raw Data';
scatter y=prin2 x=prin1 / datalabel=country group=clusname;
run; quit;
ods graphics off;

%makeTable(treeout=_3_clusters,group=temp,outdata=_3_clusters_with_labels);
%makeTable(treeout=_4_clusters,group=temp,outdata=_4_clusters_with_labels);
```