Cluster Analysis of European Countries According to Their Happiness Levels and Trust on their Legal System, Police Force, Parliament and Politicians

Dr. Şebnem Er

University of Cape Town
Statistical Sciences Department
Sebnem.Er@uct.ac.za

Aim

- The main aim of this research is to cluster European countries based on their
  - happiness levels (0-10:Extremely happy)
  - trust levels on their legal system (0-10:Complete trust)
  - trust levels on their police force (0-10:Complete trust)
  - trust levels on their parliament and politicians (0-10:Complete trust)

  using European Social Survey (ESS) data (2016 -ESS1-7e01).

European Social Survey

- The European Social Survey (ESS) is an academically-driven multi-country survey, which has been administered in over 30 countries to date.
- ESS collects data from European Union and some non-European countries such as:
  - Austria (AT), Belgium (BE), Bulgaria (BG), Cyprus (CY), Croatia (HR), the Czech Republic (CZ), Denmark (DK), Estonia (EE), Finland (FI), France (FR), Germany (DE), Greece (GR), Hungary (HU), Iceland (IS), Ireland (IE), Israel (IL), Italy (IT), Lithuania (LT), Luxembourg (LU), Netherlands (NL), Norway (NO), Poland (PL), Portugal (PT), the Russian Federation (RU), Slovakia (SK), Slovenia (SI), Spain (ES), Sweden (SE), Switzerland (CH), Turkey (TR), Ukraine (UA) and the United Kingdom (GB).

European Social Survey

- Its three aims are:
  - to monitor and interpret changing public attitudes and values within Europe and to investigate how they interact with Europe's changing institutions
  - to advance and consolidate improved methods of cross-national survey measurement in Europe and beyond
  - to develop a series of European social indicators, including attitudinal indicators

Data Overall

The respondents were asked to rank their

- happiness levels on a score of 0 = extremely unhappy – 10 = extremely happy
- trust levels on a score of 0 = no trust – 10 = complete trust on the following variables:
  - trstprl: "countrys parliament"
  - trstlgl: "legal system"
  - trstplc: "police"
  - trstplt: "politicians"

Data

cntry	trstprl	trstlgl	trstplc	trstplt	happy	trstprlnum	trstlglnum	trstplcnum	trstpltnum	happynum
SE	high	high	high	high	happy	10	9	10	9	7
SE	high	high	high	med	happy	8	8	9	6	8
SE	med	med	med	med	happy	6	6	6	4	8
SE	low	low	low	med	medhappy	1	3	3	4	5
SE	med	med	med	low	happy	4	4	4	3	8
SE	high	med	med	low	happy	7	4	5	3	9
SI	high	med	high	low	happy	7	4	8	3	7
SI	high	high	high	high	medhappy	8	8	8	8	5
SI	med	med	high	low	medhappy	5	4	7	2	5
SI	high	med	high	high	happy	7	4	8	7	9
SI	high	med	med	med	happy	7	4	5	4	8
SI	low	low	med	low	medhappy	3	3	5	2	5
SI	high	low	high	low	happy	8	3	9	2	8
SI	low	low	low	med	medhappy	2	2	3	4	5

Analysis involves two parts:

- Clustering countries based on their responses to trust and happiness levels using multiple correspondence analysis:

Analysis involves two parts:

- Clustering countries based on their responses to trust and happiness levels using multiple correspondence analysis:

  The responses between 0-10 were recoded into high (7-10); med (4-6) ; low (0-3) levels of trust and happiness.

Analysis involves two parts:

- Clustering countries based on their responses to trust and happiness levels using multiple correspondence analysis:

  The responses between 0-10 were recoded into high (7-10); med (4-6) ; low (0-3) levels of trust and happiness.

- Clustering countries based on their responses to trust and happiness levels using complete linkage cluster analysis:

Analysis involves two parts:

- Clustering countries based on their responses to trust and happiness levels using multiple correspondence analysis:

  The responses between 0-10 were recoded into high (7-10); med (4-6) ; low (0-3) levels of trust and happiness.

- Clustering countries based on their responses to trust and happiness levels using complete linkage cluster analysis:

  The responses (0-10) for the trust and happiness questions were averaged per country

Data for MCA

	cntry	trstprl	trstlgl	trstplc	trstplt	happycat
40833	SE	high	high	med	high	happy
40834	SE	high	high	high	high	happy
40835	SE	high	high	high	high	happy
40836	SE	high	high	high	med	happy
40837	SE	med	med	med	med	happy
40838	SE	low	low	low	med	medhappy
40839	SE	med	med	med	low	happy
40840	SE	high	med	med	low	happy
40842	SI	high	med	high	low	happy
40843	SI	high	high	high	high	medhappy
40844	SI	med	med	high	low	medhappy
40845	SI	high	med	high	high	happy
40846	SI	high	med	med	med	happy
40847	SI	low	low	med	low	medhappy

Data for Cluster Analysis

	trstprlnum	trstlglnum	trstplcnum	trstpltnum	happy
AT	5.07	6.06	6.45	3.54	7.59
BE	4.99	4.40	5.62	4.29	7.77
CH	5.75	6.17	6.81	4.95	8.00
CZ	3.66	3.78	4.98	3.22	6.76
DE	4.36	5.58	6.61	3.44	7.17
DK	6.20	7.13	7.91	5.49	8.33
ES	4.84	4.31	5.45	3.46	7.36
FI	5.80	6.74	7.95	4.80	8.03
FR	4.47	4.83	5.89	3.65	7.35
GB	4.65	5.03	6.08	3.78	7.54
GR	4.79	6.27	6.39	3.43	6.54
HU	5.00	5.08	4.89	3.89	6.36
IE	4.42	5.12	6.52	3.74	7.88
IL	4.62	6.66	6.21	3.33	7.14
IT	4.84	5.57	6.61	3.60	6.50
LU	5.60	6.23	6.47	4.90	8.00
NL	5.20	5.36	5.82	4.84	7.80
NO	5.70	6.33	6.98	4.58	7.89
PL	3.47	3.66	4.88	2.75	6.44
PT	4.32	4.25	4.98	2.84	6.91
SE	5.96	6.08	6.76	4.76	7.91
SI	4.05	4.27	4.84	3.04	6.93

What is MCA?

- Multiple correspondence analysis (MCA) is a method to obtain a graphical representation of a multivariate categorical data set.
- The most common idea is to find out whether there is association between countries and the answers for the trust variables.
- The main use of MCA is for graphical representation, so only solutions in two dimensions are really useful in this context.

Different ways of performing MCA

- Using indicator matrix:

Different ways of performing MCA

- Using indicator matrix: Dummy variables are used to indicate which level of each categorical variable the respondent chose

Different ways of performing MCA

- Using indicator matrix: Dummy variables are used to indicate which level of each categorical variable the respondent chose
- Using Burt matrix: structure of the Burt matrix is such that each off-diagonal block is the two-way frequency table of two of the categorical variables, while the diagonal blocks are diagonal matrices, with the number of times each level appears in the data on the diagonal

Different ways of performing MCA

- Using indicator matrix: Dummy variables are used to indicate which level of each categorical variable the respondent chose
- Using Burt matrix: structure of the Burt matrix is such that each off-diagonal block is the two-way frequency table of two of the categorical variables, while the diagonal blocks are diagonal matrices, with the number of times each level appears in the data on the diagonal
- Joint correspondence analysis: MCA on the Burt matrix, uses the svd to optimally approximate the whole matrix. In JCA a similar approximation is performed, but only the off-diagonal blocks are approximated, ignoring the diagonal blocks.

Indicator Matrix

Once the indicator matrix has been constructed, a PCA of the indicator matrix is performed and similar to the PCA biplot, we obtain a 2D map representing the respondents (rows) and the variables (dummy variables) in a single plot:

Indicator Matrix

Once the indicator matrix has been constructed, a PCA of the indicator matrix is performed and similar to the PCA biplot, we obtain a 2D map representing the respondents (rows) and the variables (dummy variables) in a single plot:

cntry:AT	cntry:BE	. . .	cntry:SI	trstpl:high	trstpl:med	trstpl:low	trstgl:high	. . .	trstplt:low
1	0	0	. . .	1	0	0	1	. . .	0

Burt Matrix

With the indicator matrix we would have $22 + 3x4 = 34 columns$ to analyse using the singular value decomposition.

Burt Matrix

With the indicator matrix we would have $22 + 3x4 = 34$*columns* to analyse using the singular value decomposition.

	cntry:IE	cntry:IL	cntry:IT	cntry:LU	cntry:NL	cntry:NO	cntry:PL	cntry:PT	cntry:SE	cntry:SI	trstprl:high	trstprl:low
cntry:AT	0	0	0	0	0	0	0	0	0	0	645	544
cntry:BE	0	0	0	0	0	0	0	0	0	0	467	413
cntry:CH	0	0	0	0	0	0	0	0	0	0	740	237
cntry:CZ	0	0	0	0	0	0	0	0	0	0	151	657
cntry:DE	0	0	0	0	0	0	0	0	0	0	516	1024
cntry:DK	0	0	0	0	0	0	0	0	0	0	738	166
cntry:ES	0	0	0	0	0	0	0	0	0	0	384	417
cntry:FI	0	0	0	0	0	0	0	0	0	0	862	300
cntry:FR	0	0	0	0	0	0	0	0	0	0	277	457
cntry:GB	0	0	0	0	0	0	0	0	0	0	469	627
cntry:GR	0	0	0	0	0	0	0	0	0	0	708	778
cntry:HU	0	0	0	0	0	0	0	0	0	0	461	459
cntry:IE	2046	0	0	0	0	0	0	0	0	0	440	696
cntry:IL	0	2499	0	0	0	0	0	0	0	0	666	852
cntry:IT	0	0	1207	0	0	0	0	0	0	0	244	285
cntry:LU	0	0	0	1552	0	0	0	0	0	0	457	194

Joint Correspondence Analysis - JCA

- When MCA is performed on the Burt matrix, a large part of the inertia is attributed to the diagonal blocks.
- In JCA, only the off-diagonal blocks are used since these show relationships.

Cluster Analysis

Multiple correspondence analysis (MCA) for the graphical representation of categorical variables and Cluster Analysis for the grouping of observations into homogenous clusters.

Cluster Analysis

Multiple correspondence analysis (MCA) for the graphical representation of categorical variables and Cluster Analysis for the grouping of observations into homogenous clusters.

- Cluster analysis is the name given to a group of techniques whose main aim is to group objects that are similar in some way together.
- The aim is to produce a set of groups or clusters such that each object is similar to the others in its own cluster, and different to objects in other clusters.

Types of Cluster Analysis

As with MCA, Cluster Analysis has a vast number of different algorithms to perform the analysis:

- Hierarchical
  - Agglomerative: Single, complete centroid, average, Ward's method
  - Divisive
- Non-hierarchical
  - K-means
  - Model based clustering (MND)

Apart from K-means and the model based clustering, which assumes each cluster has a multivariate normal distribution, all other methods need a matrix of dissimilarities.

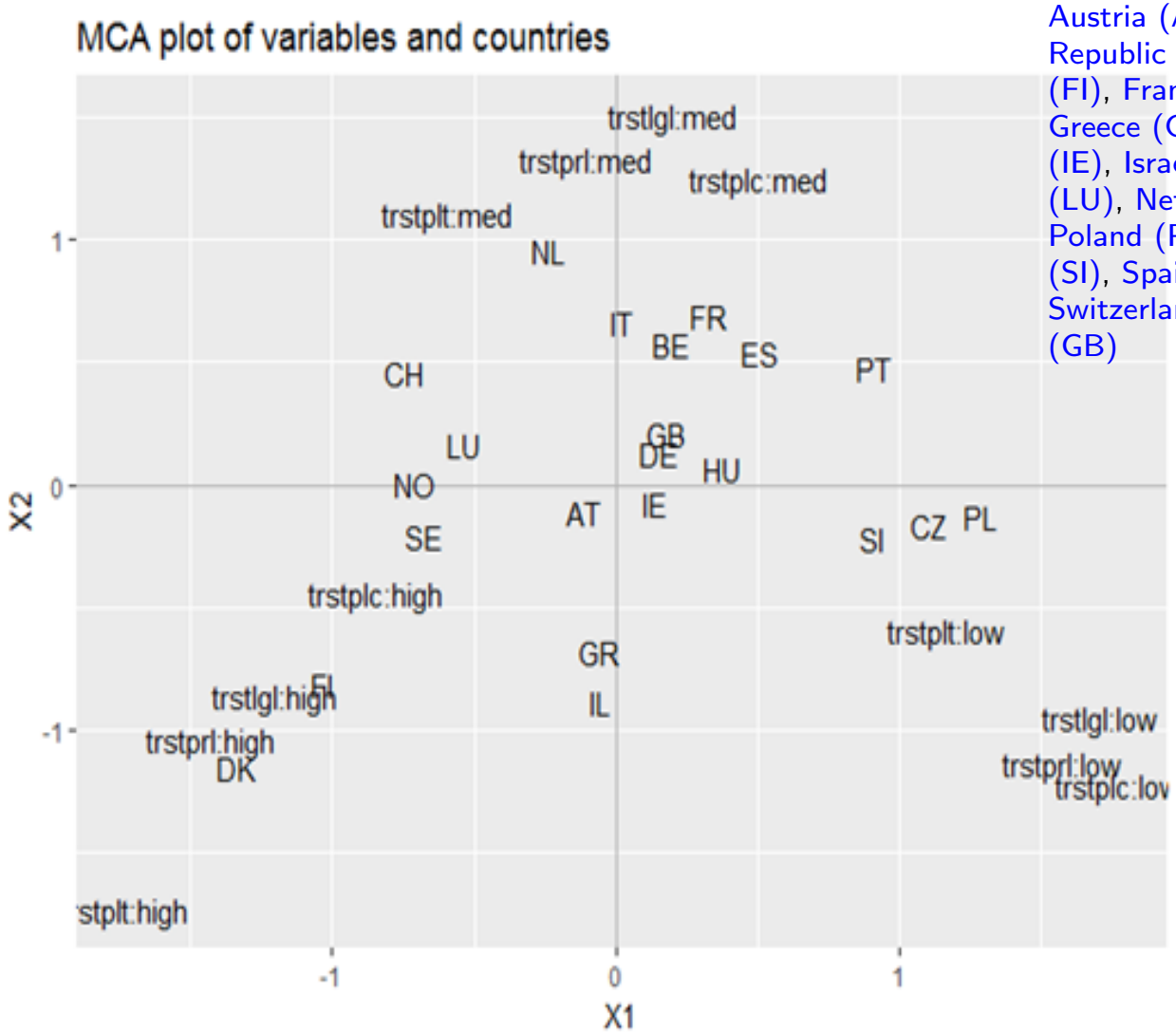
Different Distances

- Continuous Variables
  - Euclidean: $d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$
  - City-Block (Manhattan): $d_{ij} = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$
  - Correlation $d_{ij} = 1 - r_{ij}$
- Categorical Variables
  - Proportional Difference $d_{ij} = \frac{1}{p} \sum_{k=1}^{p} I(x_{ik} \neq x_{jk})$
  - Correlation (Spearman Rank)
- Mixture of Variables: Gower distance

Complete Linkage

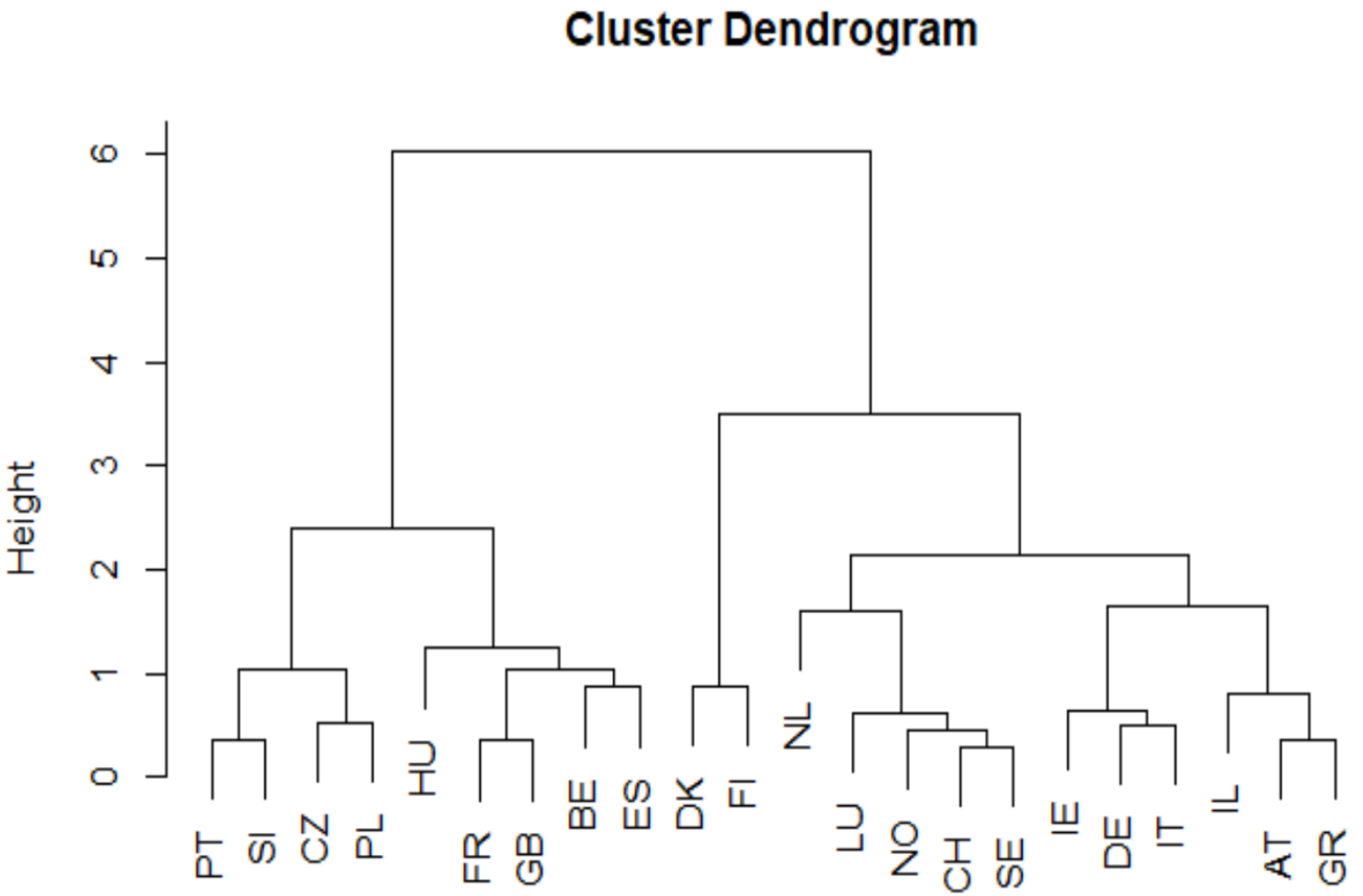
- Start with each object in its own cluster
- Continue to merge two clusters together until all objects belong to the same cluster. At each stage the two closest clusters are merged, to form a single new cluster.
- Complete linkage defines the dissimilarity between two clusters as the maximum of the dissimilarities between individual objects in each of the two clusters.

Results - MCA without happiness variable



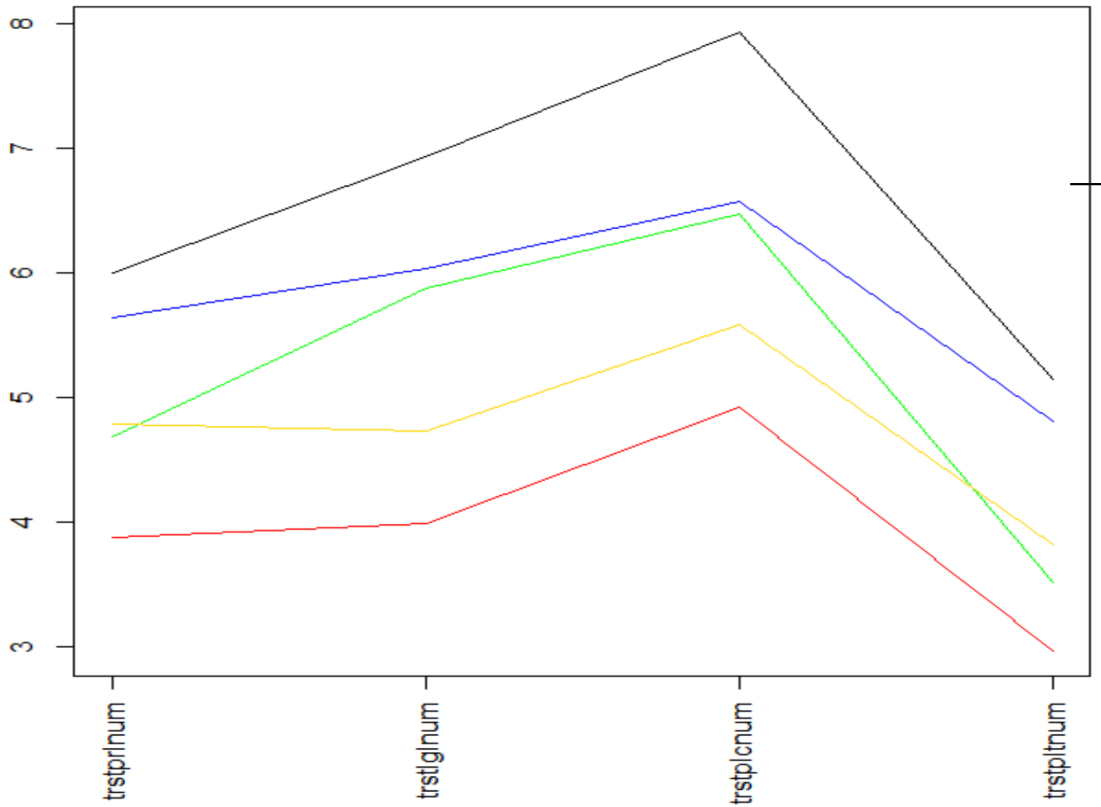
Austria (AT), Belgium (BE), the Czech Republic (CZ), Denmark (DK), Finland (FI), France (FR), Germany (DE), Greece (GR), Hungary (HU), Ireland (IE), Israel (IL), Italy (IT), Luxembourg (LU), Netherlands (NL), Norway (NO), Poland (PL), Portugal (PT), Slovenia (SI), Spain (ES), Sweden (SE), Switzerland (CH), the United Kingdom (GB)

Results - Cluster Analysis



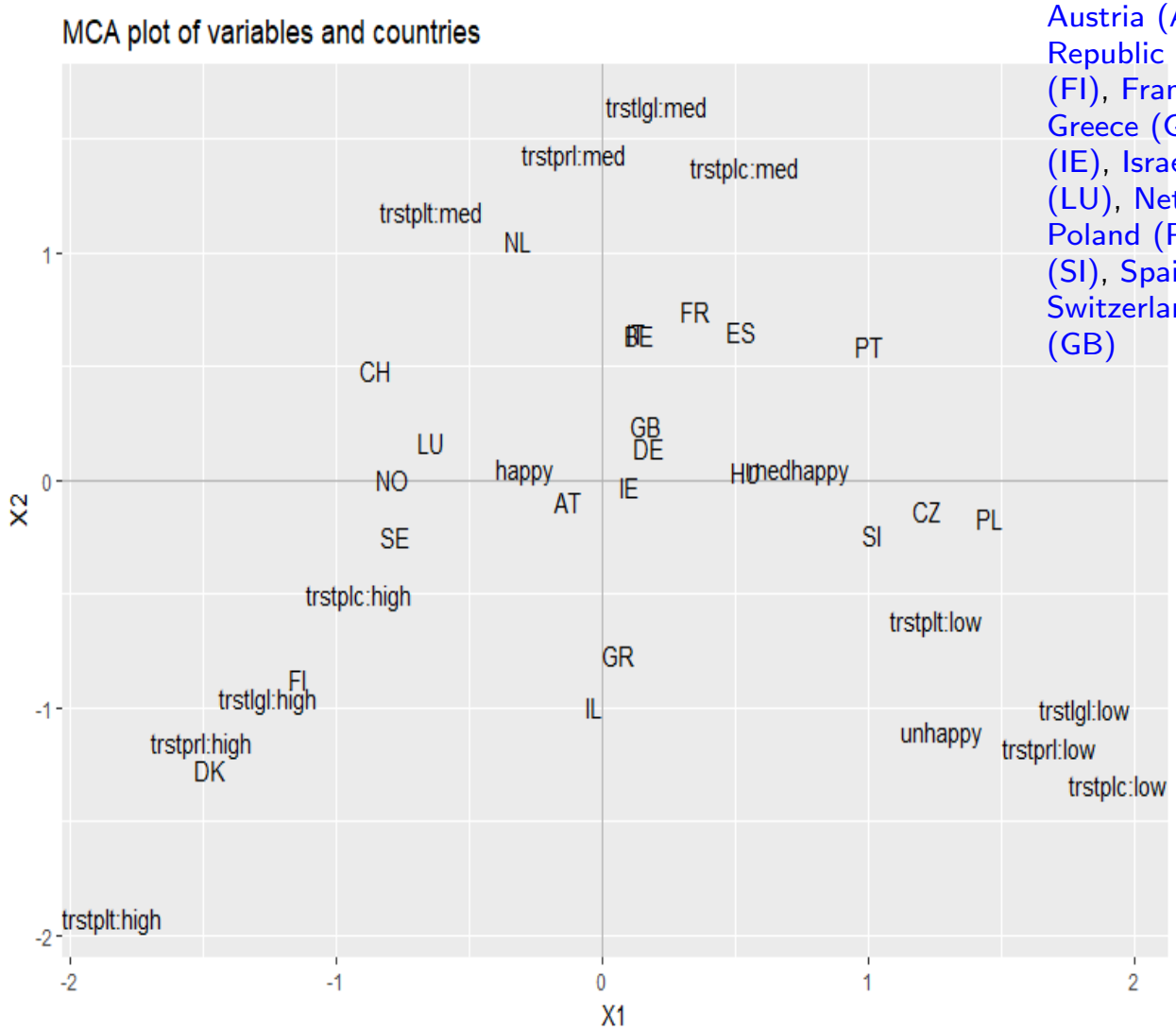
Results - Cluster Profiles

After the cluster analysis, we can create clusters appropriately and here 5 clusters seems appropriate. The cluster profiles are as follows:



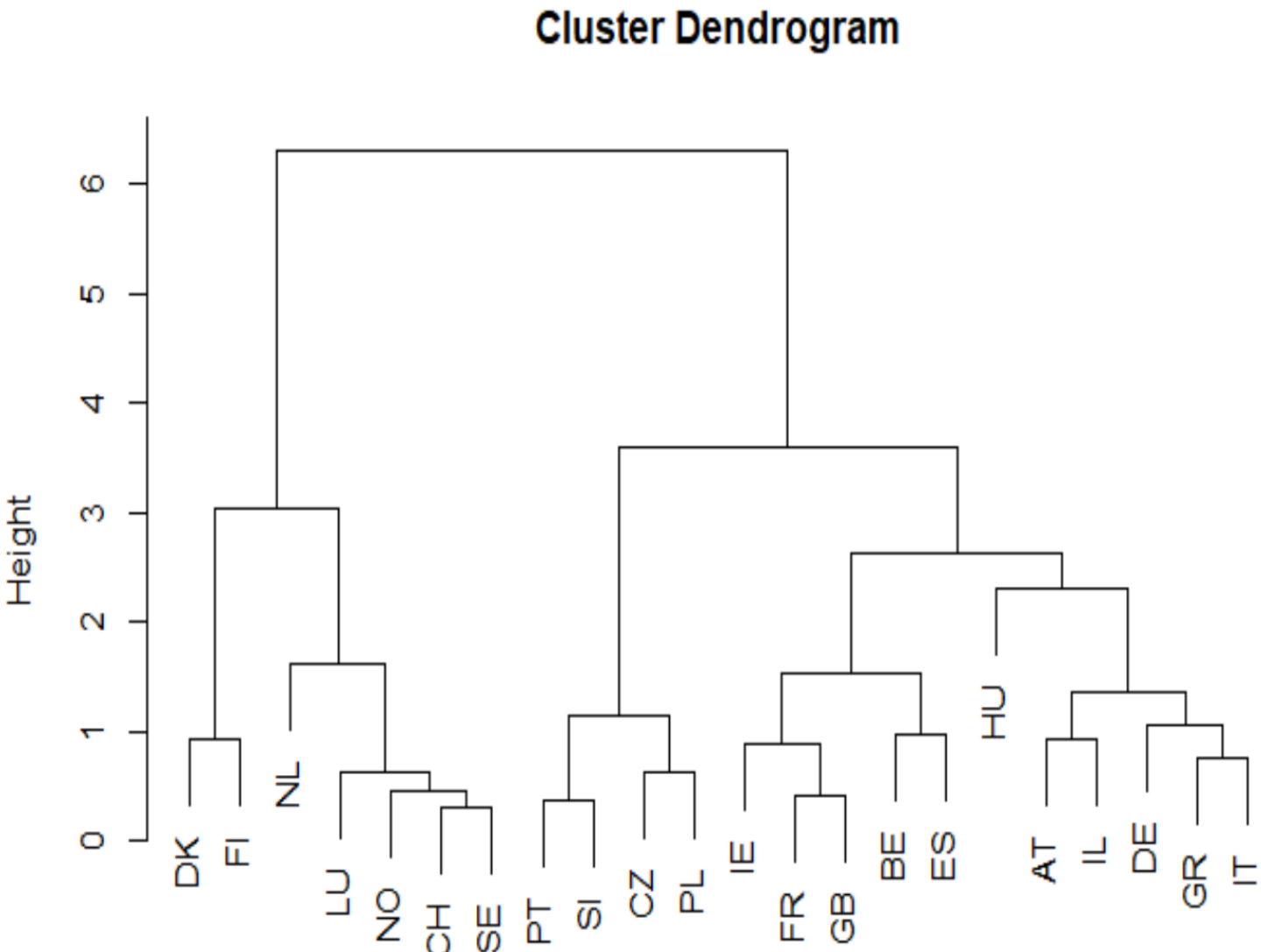
cluster	countries	trstprlnum
1	AT DE GR IE IL IT	4.68
2	BE ES FR GB HU	4.79
3	CH LU NL NO SE	5.64
4	CZ PL PT SI	3.88
5	DK FI	6.00

Results - MCA with happiness variable

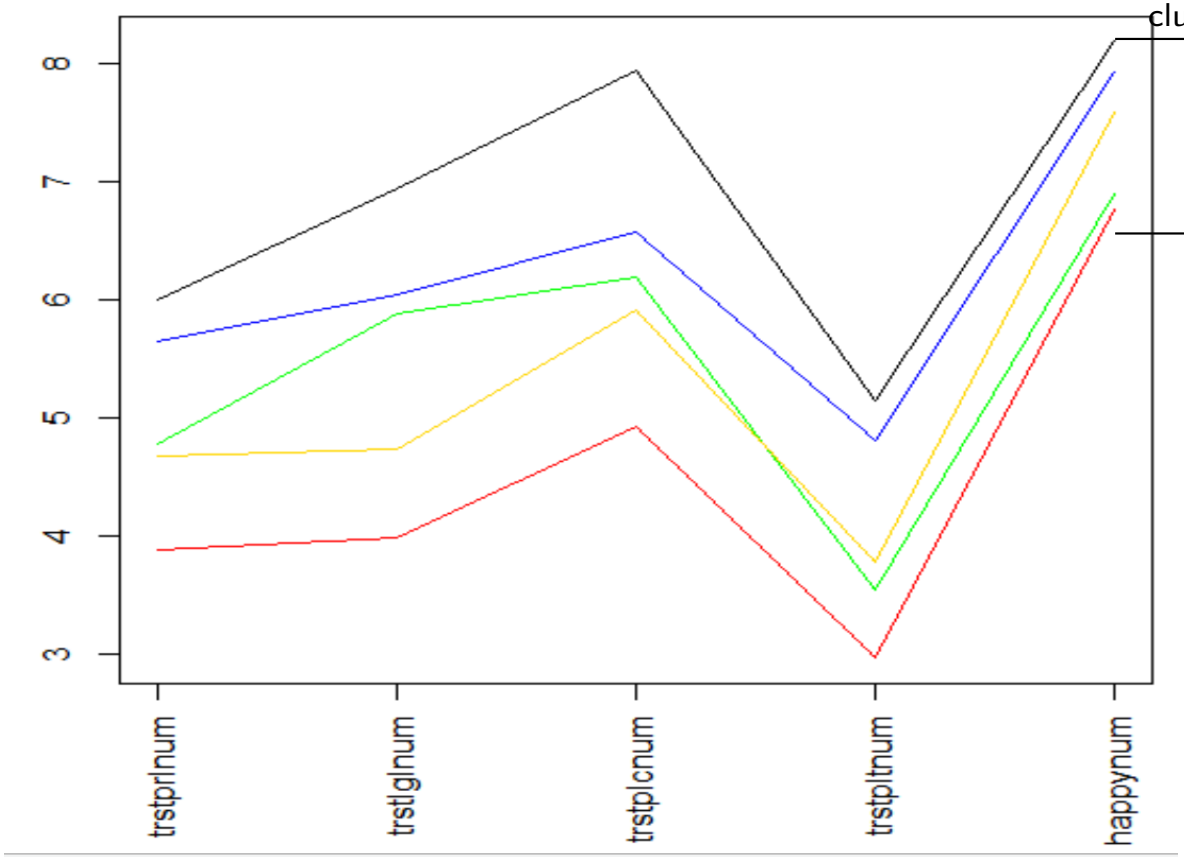


Austria (AT), Belgium (BE), the Czech Republic (CZ), Denmark (DK), Finland (FI), France (FR), Germany (DE), Greece (GR), Hungary (HU), Ireland (IE), Israel (IL), Italy (IT), Luxembourg (LU), Netherlands (NL), Norway (NO), Poland (PL), Portugal (PT), Slovenia (SI), Spain (ES), Sweden (SE), Switzerland (CH), the United Kingdom (GB)

Results - Cluster Analysis



Results - Cluster Profiles



cluster	countries	trstprlnum
1	AT DE GR HU IL IT	4.78
2	BE ES FR GB IE	4.67
3	CH LU NL NO SE	5.64
4	CZ PL PT SI	3.88
5	DK FI	6.00

Thank you!

Hvala!

Teşekkurler