SDA 2012 ~ Madrid
Workshop in Symbolic Data Analysis

SYROKKO
Roissy-Aéropôle
+ 33 1 74 37 26 55

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Analyzing European social survey data using symbolic data methods and *Syrokko* software

Filipe Afonso[1] , Seppo Laaksonen[2]

1. SYROKKO, Paris
afonso@syrokko.com
2. University of Helsinki
Seppo.Laaksonen@Helsinki.Fi

November 2012

We apply a two-stage data mining strategy to handle and analyze big statistical micro data sets.

- The first stage consists of smart aggregation of such micro data, and

- the second stage data continues to analyze and visualize the smartly aggregated data, further.

The smart aggregation here requires the three steps:

- One is to decide and to create the appropriate aggregates themselves, called also ‘concepts.’
- Second, the characteristics for the concepts need to be implemented.
- The third step in smart aggregation is to operationalize the first two steps by creating the new data set. This operation is performed by the *SYR* software for SDA from *Syrokko* company.

CONCEPTS in our application

26 countries by two age groups:

Y = under 50 years (YOUNG)

O = 50 years or more (Experienced)

(the maximum = 100)

The 26x2 = 52 concepts are now

with their frequencies from the sample on the right side

The data are from the fifth round of the European Social Survey, that is, from the years 2010-2011.

concept	Frequency
BE_O	761
BE_Y	943
BG_O	1469
BG_Y	965
CH_O	698
CH_Y	808
CY_O	513
CY_Y	570
CZ_O	1069
CZ_Y	1317
DE_O	1420
DE_Y	1611
DK_O	761
DK_Y	815
ES_O	874
ES_Y	919
FI_O	944
FI_Y	934
FR_O	859
FR_Y	869
GB_O	1205
GB_Y	1217
GR_O	1180
GR_Y	1535
HR_O	920
HR_Y	729
HU_O	733
HU_Y	828
IE_O	1068
IE_Y	1508
IL_O	984
IL_Y	1310
NL_O	923
NL_Y	906
NO_O	672
NO_Y	876
PL_O	737
PL_Y	1014
PT_O	1283
PT_Y	867
SE_O	733
SE_Y	764
SI_O	661
SI_Y	742
SK_O	1010
SK_Y	846
UA_O	1025
UA_Y	906

Variables in our application

AS YOU KNOW:

The symbolic data analysis approach offers good tools for this, since we can broaden the description of the variables available for classic aggregation. In this paper, our symbolic variables are of the two kinds: frequencies of categorical variables, and the intervals of continues or ordinal initial variables. We do not lose any information in the previous case if we do not collapse the initial categories. We, naturally, lose some information in the case of interval variables but if the intervals are well designed, our data loss is limited, but our research focus will be clearer, respectively.

Variables in our application

The weighting variable DWEIGTH is applied for categorical (histogram) variables, but not for the life value variables that are intervals.

Micro variables

Values and categories for the following three ones

1 All of the time,
2 Most of the time,
3 More than half of the time,
4 Less than half of the time,
5 Some of the time,
6 At no time

ACTVGRS ‘Have felt active and vigorous last 2 weeks.’
CLMRLX ‘Have felt calm and relaxed last 2 weeks.’
GDSPRT ‘Have felt cheerful and in good spirits last 2 weeks.’

These three variables are thus concerned on ‘**feelings**’ of recent times, so that higher values mean ‘non-good feelings’.

Variables in our application

The weighting variable DWEIGTH is applied for categorical (histogram) variables, but not for the life value variables that are intervals.

Micro variables

Values and categories for the following 12 ones are from 0 (=most negative) till 10 (most positive)

HAPPY ‘How happy are you.’

IMBGECO ‘Immigration bad or good for country's economy.’

IMWBCNT ‘Immigrants make country worse or better place to live.’

PPLFAIR ‘Most people try to take advantage of you, or try to be fair.’

PPLHLP ‘Most of the time people helpful or mostly looking out for themselves.’

PPLTRST ‘Most people can be trusted or you can't be too careful.’

STFECO ‘How satisfied with present state of economy in country.’

STFHLTH ‘State of health services in country nowadays.’

TRSTLGL ‘Trust in the legal system.’

TRSTPLC ‘Trust in the police.’

TRSTPLT ‘Trust in politicians.’

TRSTPRL ‘Trust in country's parliament.’

The last four variables are relating to trusting in political things, and can be influenced by the politics of the country over recent years.

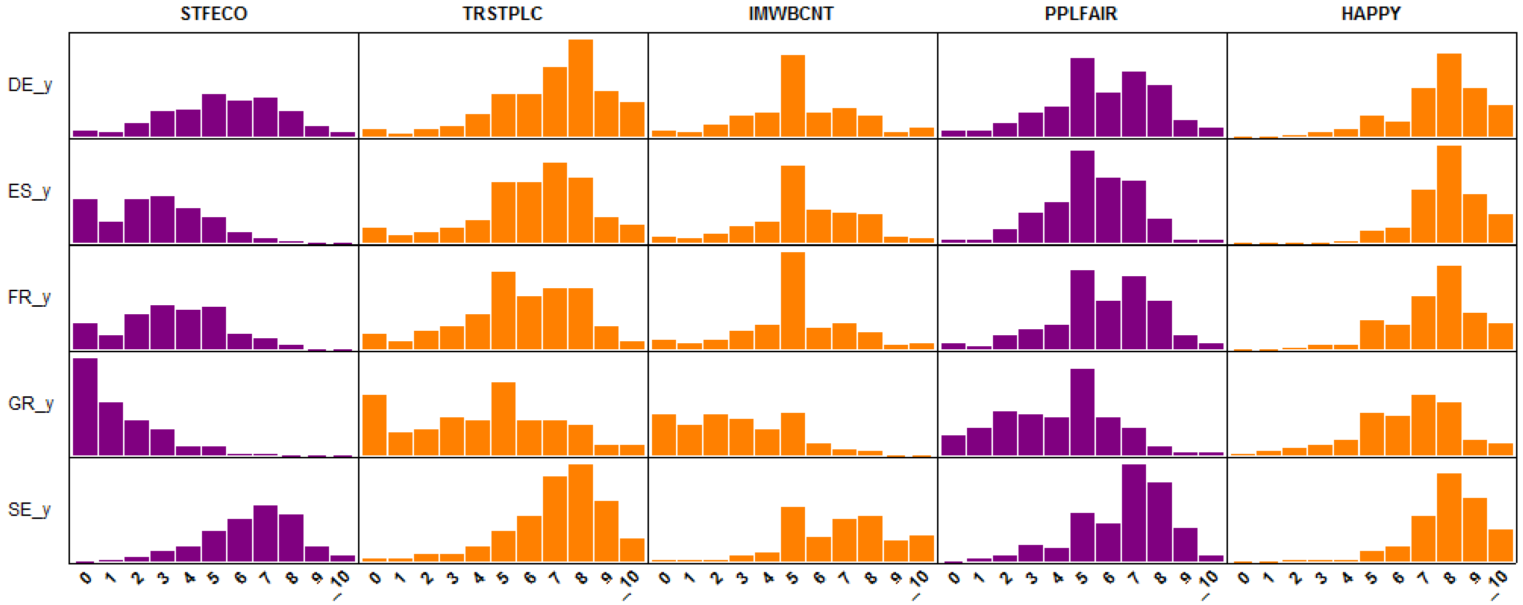
Variables in our application

Finally, we have the four life value variables created from the 21 questions by exploratory factor analysis.
Thus, the mean over all 52 symbolic objects is equal to 0, and the standard deviation equal to 1, respectively. Our aggregation for each object is such that the minimum= 25% quartile
and the maximum = 75% quartile, respectively.

TRADITION ‘ Traditions, formal rules, safe, etc are more important if the factor score is higher.’
EQUALITY ‘Equality, caring the nature, understanding different people, etc are important.’
ENJOY ‘Enjoying, adventures, seeking fun, etc are important.’
SUCCESS ‘Success, riches, thinking new ideas, etc are important.’

Comparison between young people on histogram variables

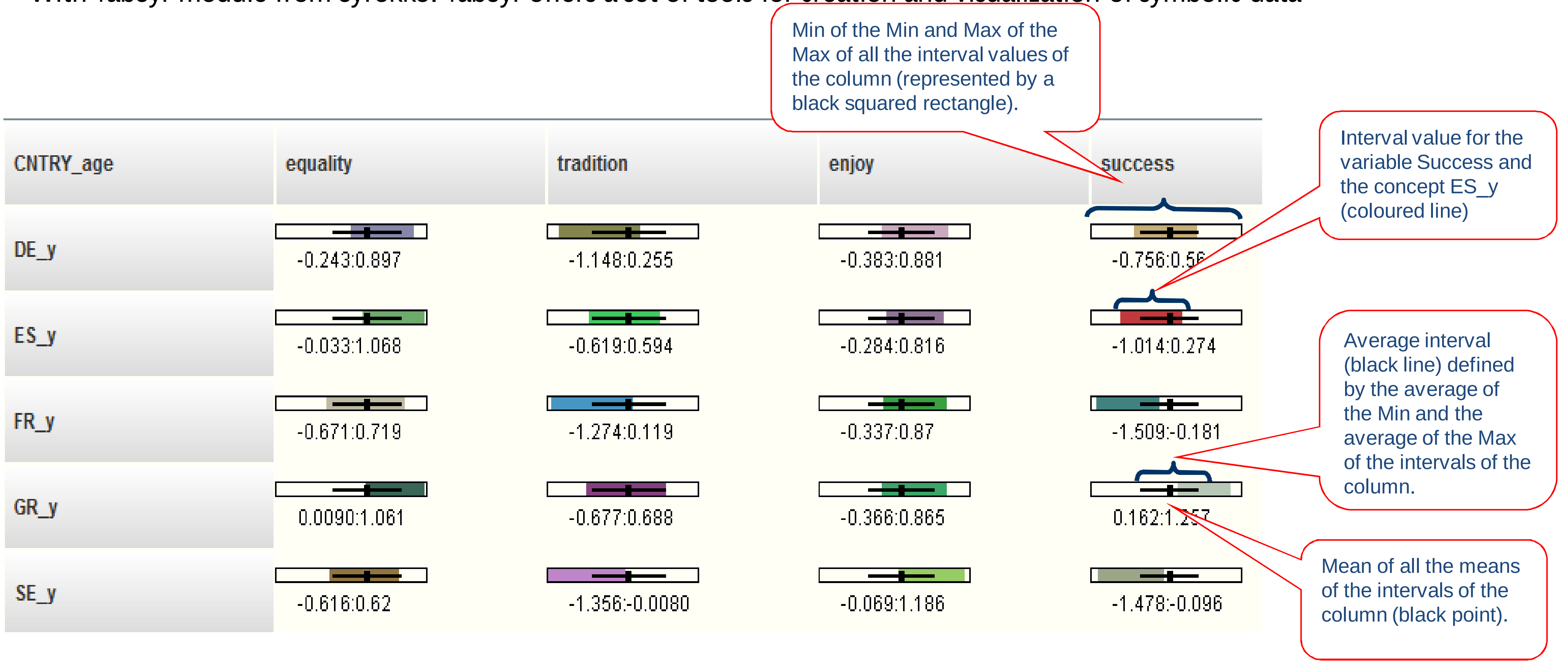
•from Germany (DE), Spain (ES), France (FR), Greece (GR) and Sweden (SE).
• Usin StatSyr module from Syrokko. StatSyr characterizes the concepts and shows their variation. It offers various methods for symbolic variables as the following matrix of histograms



We note the much better results of Sweden for all the variables. On the contrary, we see the very bad results of Greece for all the variables. The other countries are between Greece and Sweden with better results for Germany than for France and Spain for the variable STTFECO (How satisfied with present state of economy in country). We also note poorer results for IMWBCNT (Immigrants make country worse or better place to live).

Comparison between young people on interval variables

- With TabSyr module from Syrokko. TabSyr offers a set of tools for creation and visualization of symbolic data



CNTRY_age	equality	tradition	enjoy	success
DE_y	-0.243:0.897	-1.148:0.255	-0.383:0.881	-0.756:0.56
ES_y	-0.033:1.068	-0.619:0.594	-0.284:0.816	-1.014:0.274
FR_y	-0.671:0.719	-1.274:0.119	-0.337:0.87	-1.509:-0.181
GR_y	0.0090:1.061	-0.677:0.688	-0.366:0.865	0.162:1.257
SE_y	-0.616:0.62	-1.356:-0.0080	-0.069:1.186	-1.478:-0.096

Min of the Min and Max of the Max of all the interval values of the column (represented by a black squared rectangle).

Interval value for the variable Success and the concept ES_y (coloured line)

Average interval (black line) defined by the average of the Min and the average of the Max of the intervals of the column.

Mean of all the means of the intervals of the column (black point).

For young people, Equality is more important in Spain and Greece. Tradition is less important in Sweden, France, and Germany than for other countries. Spain and Greece are the average of the other countries for tradition. The 5 countries are similar according to the enjoy variable. Finally, we note that success is much more important in Greece than in Sweden or France.

Principal Component Analysis (PCA) of the concept “country x age”

▪In the previous pictures, we can compare a few concepts for several variables at the same time.

▪In the following pictures, we use NetSyr module to visualize all the concepts in a same biplot and to compare them.

▪NetSyr offers a set of tools for extending standard Principal Component Analysis (PCA) to symbolic data where the units are concepts described by symbolic variables of modal (histogram) and/or interval type, and/or continuous variables.

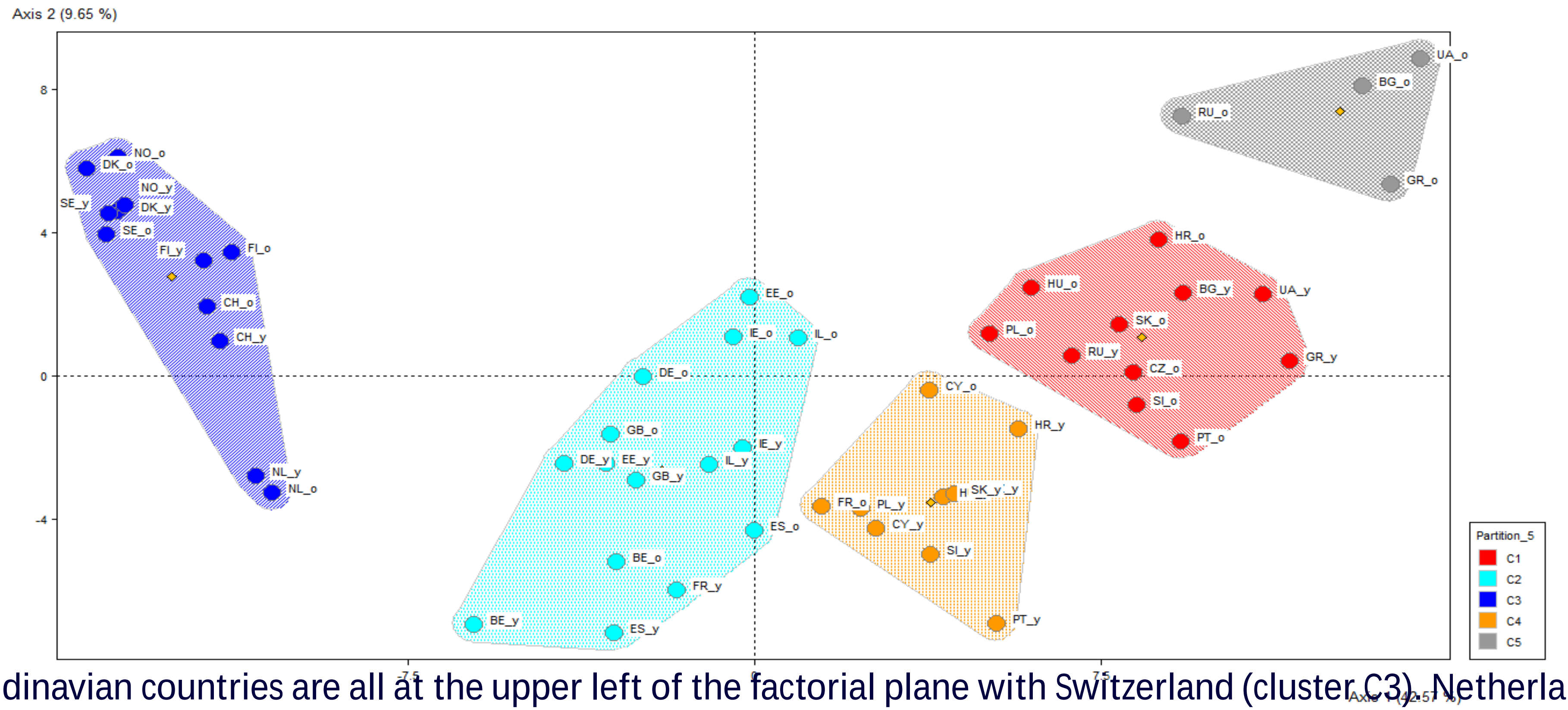
▪All the different types of variables can be mixed in a same PCA.

▪NetSyr also offers a set of user-friendly graphical tools for visualizing symbolic data in the factorial planes:

- ▪ individuals with their variation (visualization of the histogram and interval values in the factorial plane),
- ▪clusters of individuals (from a partition or an overlap clustering),
- ▪ bins and symbolic variables
- ▪proximities between individuals thanks to networks, etc.

Clustering of the concepts and visualization of 5 clusters in the factorial plane

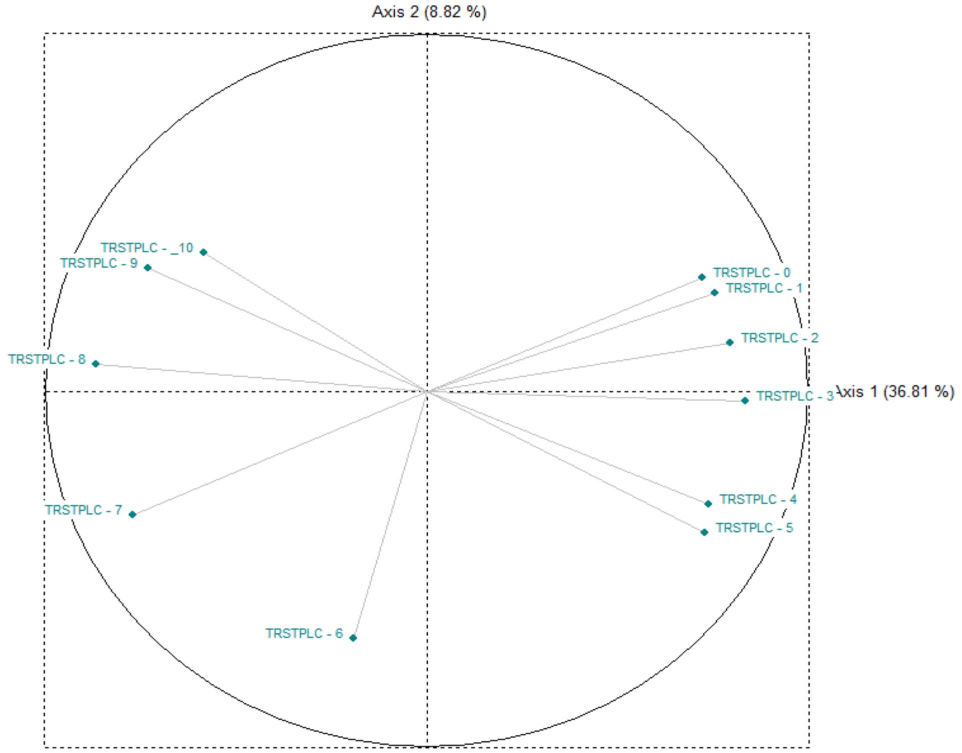
- Clusters calculated with k-means extended to symbolic data (ClustSyr) applied to the coordinates of the points in the factorial plane.



Scandinavian countries are all at the upper left of the factorial plane with Switzerland (cluster C3). Netherland is also in the same cluster but at the lower left of the factorial plane. There is a cluster at the upper right with only people over 50 years experienced from Russia (RU), Ukraine (UA), Bulgaria (BG) and Greece (GR) (Cluster C5). Western countries (except Portugal and Greece) are at the middle of the plane (cluster C2). Portugal is at the right with eastern countries (Clusters C1, C4). Younger people and experienced people from France or Portugal are not in the same clusters.

Interpretation of the factorial plane thanks to the correlation circles

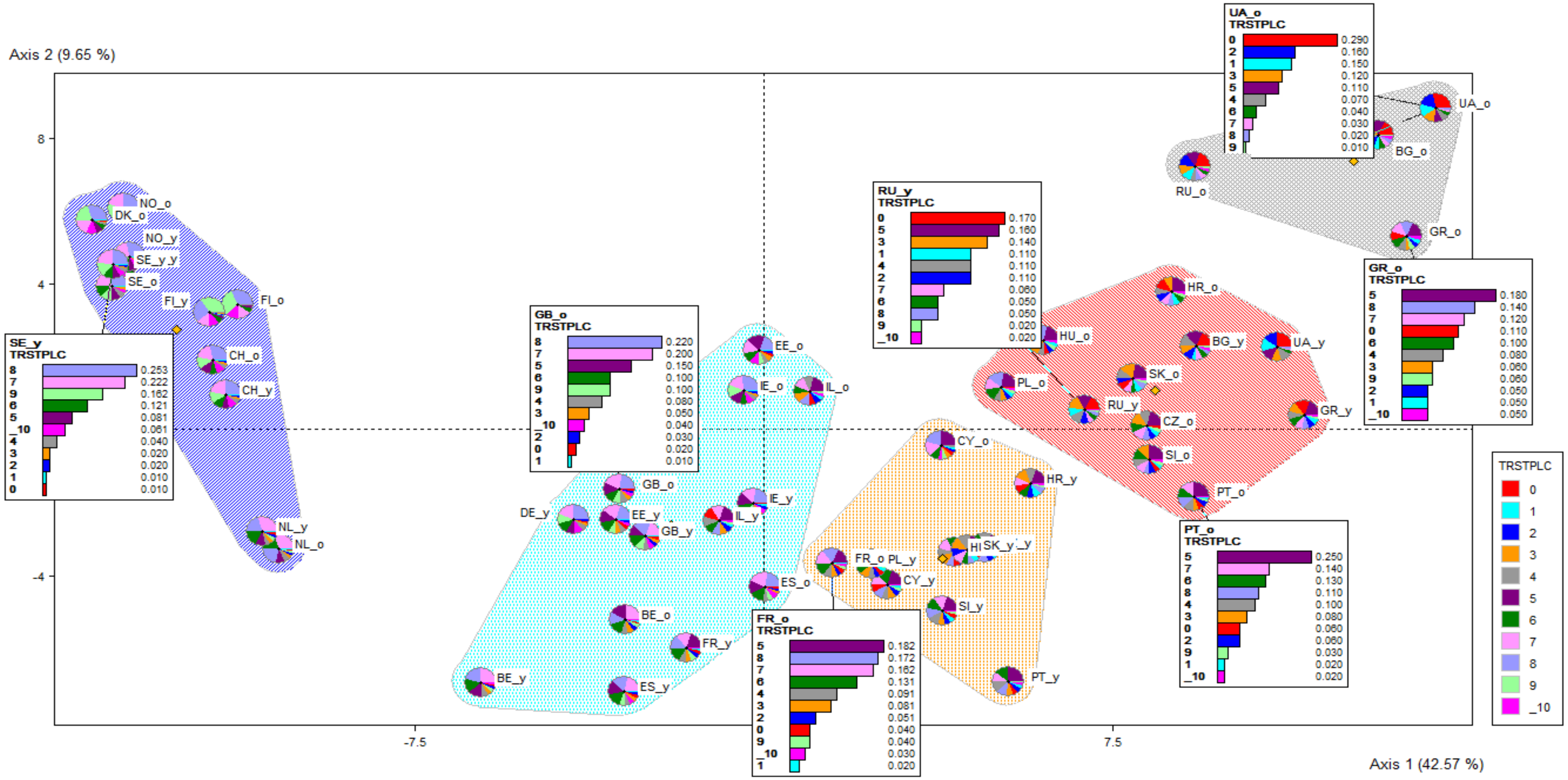
Example of variable TRSTPLC (Trust in police) variable. is very correlated with the first axis (also with the second one).



The variable TRSTPLC (Trust in police) is very correlated with the first axis (also with the second one). In the following picture, we visualize the scores of this variable (between 0 and 10) in the correlation circle. We can see the scores increasing from the upper right quadrant until the upper left quadrant passing successively through the lower right quadrant and the lower left quadrant. We have the same results for TRSTLGL (Trust in Legal System) and TRSTPRT (parliament)

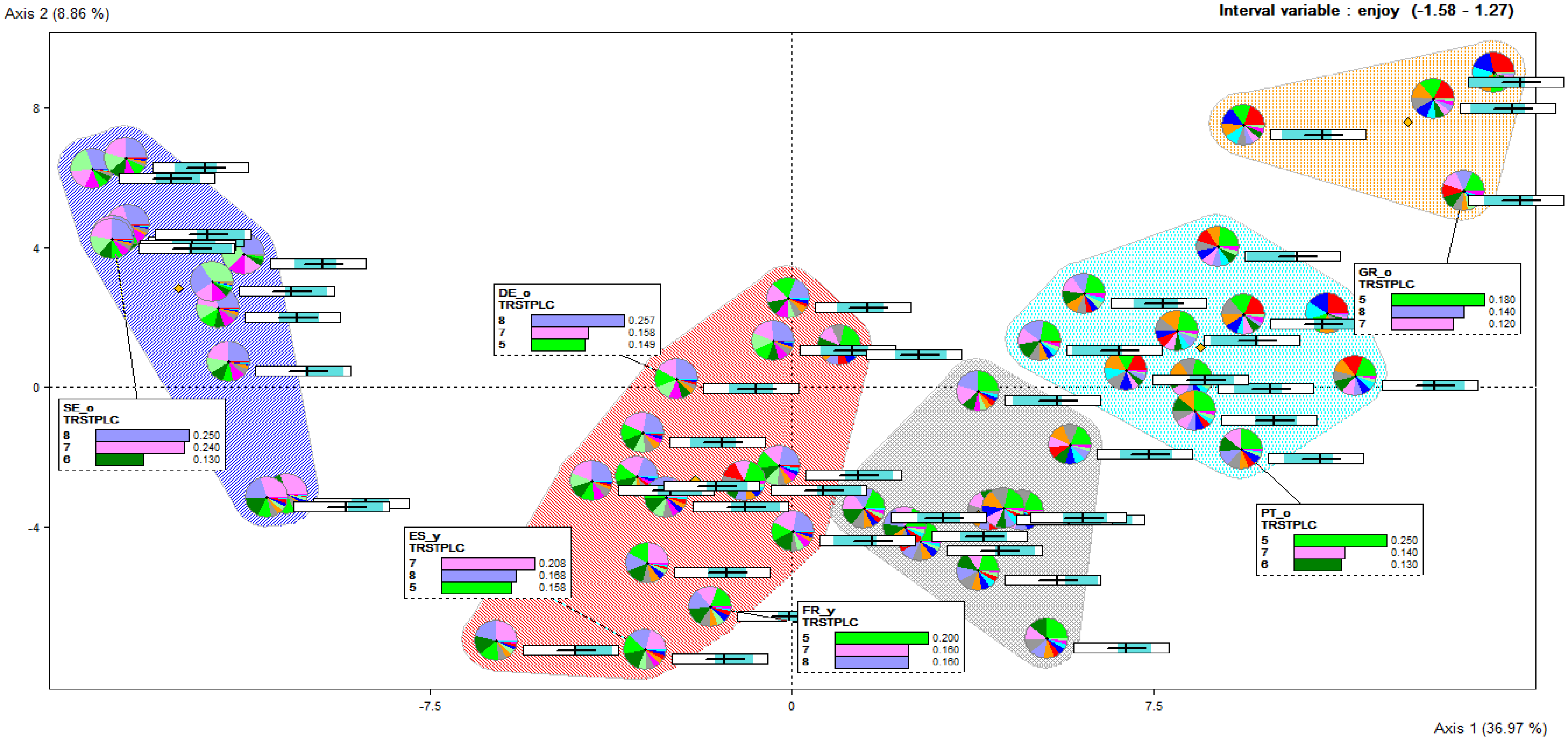
Interpretation of the clusters by visualizing the symbolic variables in the factorial plane

In the following picture, we visualize the histogram variable TRSTPLC (trust in Police) in the factorial plane. For each concept “country x age”, we visualize this variable thanks to pie charts. By clicking on each pie chart, we obtain the details of each histogram.



We note the very bad results of Russia, Bulgaria and Ukraine at the upper right and the very good results of Finland, Switzerland, Denmark, Sweden, Norway at the left.

Interpretation of the clusters by visualizing histogram-valued variable and an interval-valued variable simultaneously in the factorial plane



Main results

- The variable TRSTPLC (Trust in police) is highly correlated with the first axis and also correlated with the second one. Below we visualize the scores of this variable (between 0 and 10) in the correlation circle. We can see the scores increasing from the upper right quadrant until the upper left quadrant passing successively through the lower right quadrant and the lower left quadrant.

- The results for TRSTLGL (Trust in Legal System) and TRSTPRT (parliament) are quite similar.

- We find, for example, the very bad results of Russia, Bulgaria and Ukraine and the very good results of Finland, Switzerland, Denmark, Sweden and Norway.

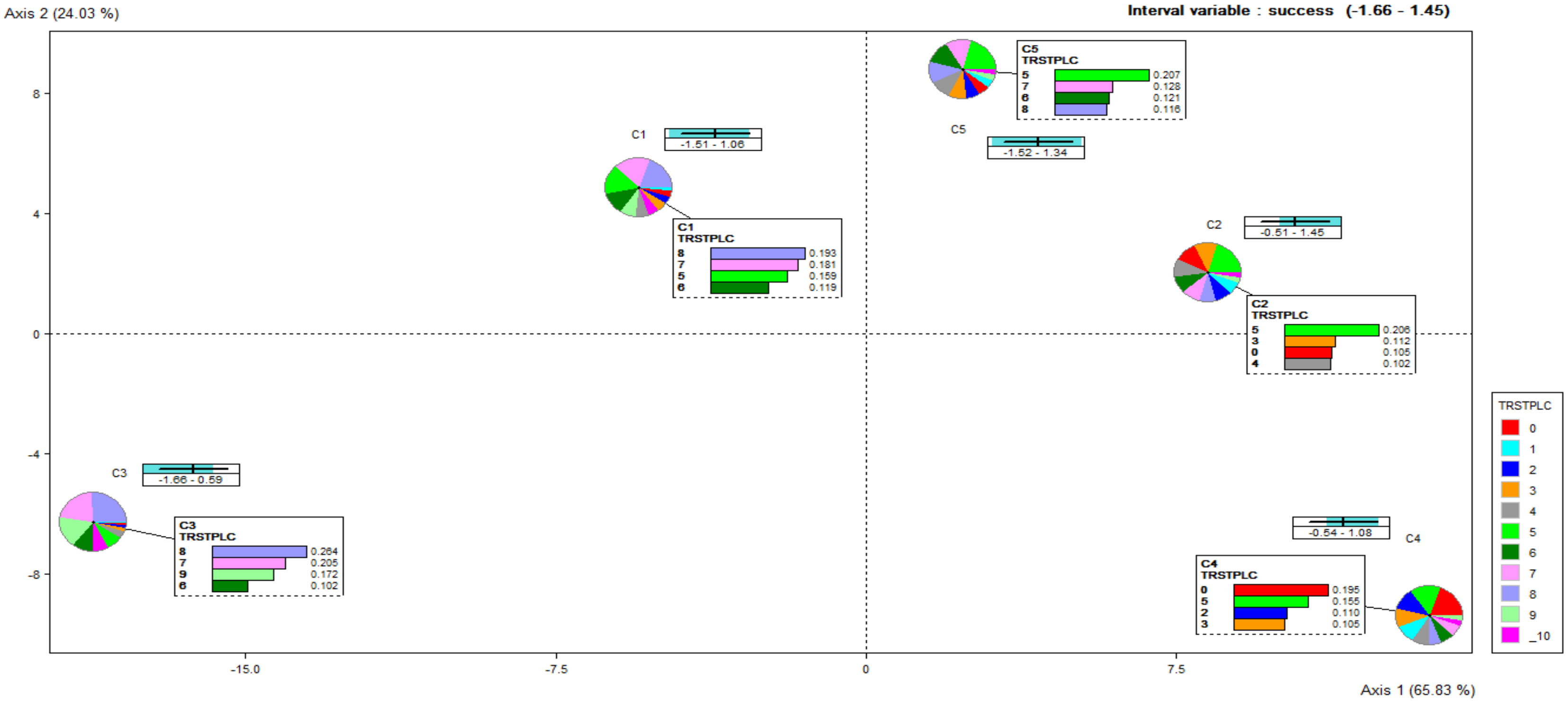
- The results with the variable TRSTPLT (trust in politicians) do differ much from the previous ones except that they are very bad for Greece. The general tendency is similar for the variables STFECO, STFHLTH, PPLTRST and PPLFAIR.

- The variable HAPPY shows, especially, the separation between East Europeans (less happy) and the rest of Europe (more happy). For the variable IMBGECO (Immigration bad or good for country's economy), the results are not very high for all the countries. Nevertheless, we note much better results for the Northern countries and very bad results for Greece where 0 has the highest frequency for people over 50 years experienced.

- For the variable IMWBCNT (Immigrants make country worse or better place to live), the results are the same than IMBGECO but with lower values for all the countries. 5 is the highest frequency for all the concepts except people over 50 years experienced from Greece where the highest frequency is 0.

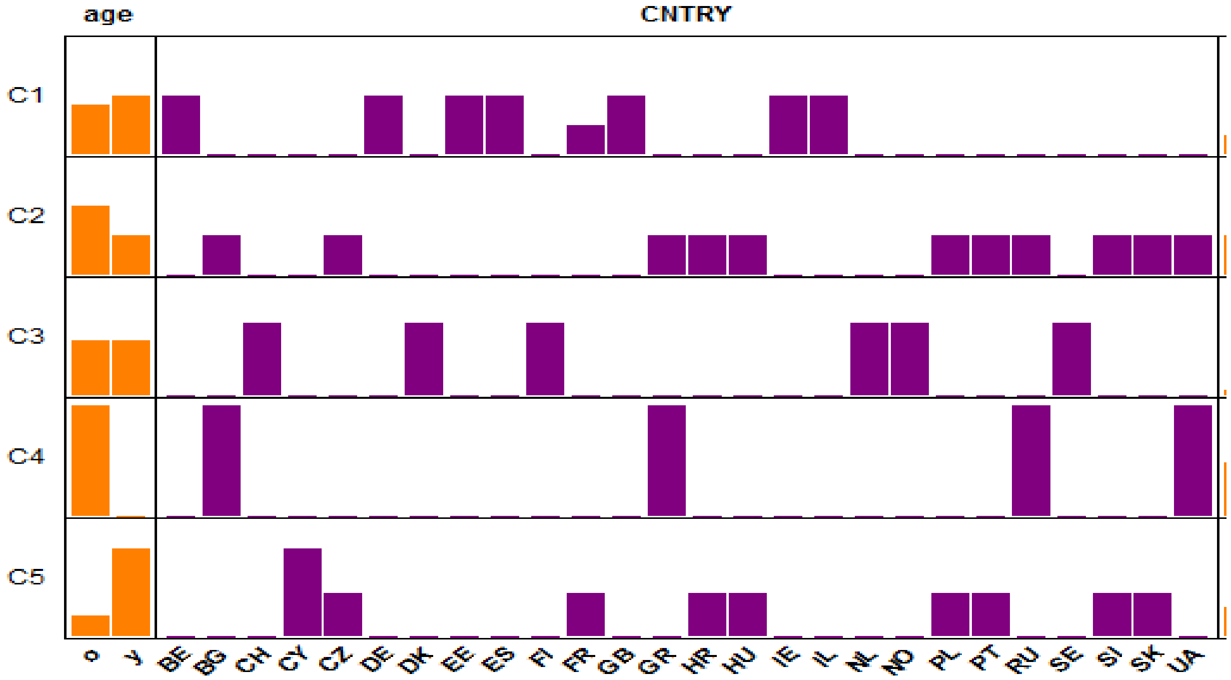
Analysis of the 5 prototypes describing the 5 clusters instead of analyzing all the concepts “countries x age”

After executing ClustSyr from the factorial plane of NetSyr in order to visualize the clustering results of the concepts "country x age", we obtain 5 clusters. Moreover, the ClustSyr module has provided the 5 prototypes describing these clusters. These prototypes can be visualized in a new factorial plane.



We visualize for each prototype the histogram-valued variable “TRSTPLC” and the interval valued variable “Success.”

Description of the prototypes with StatSyr : Age and Country



Cluster C1 is the cluster with Belgium (BE), Germany (DE), Spain (SP), Great Britain (GB), Ireland (IE), and Israel (IL)

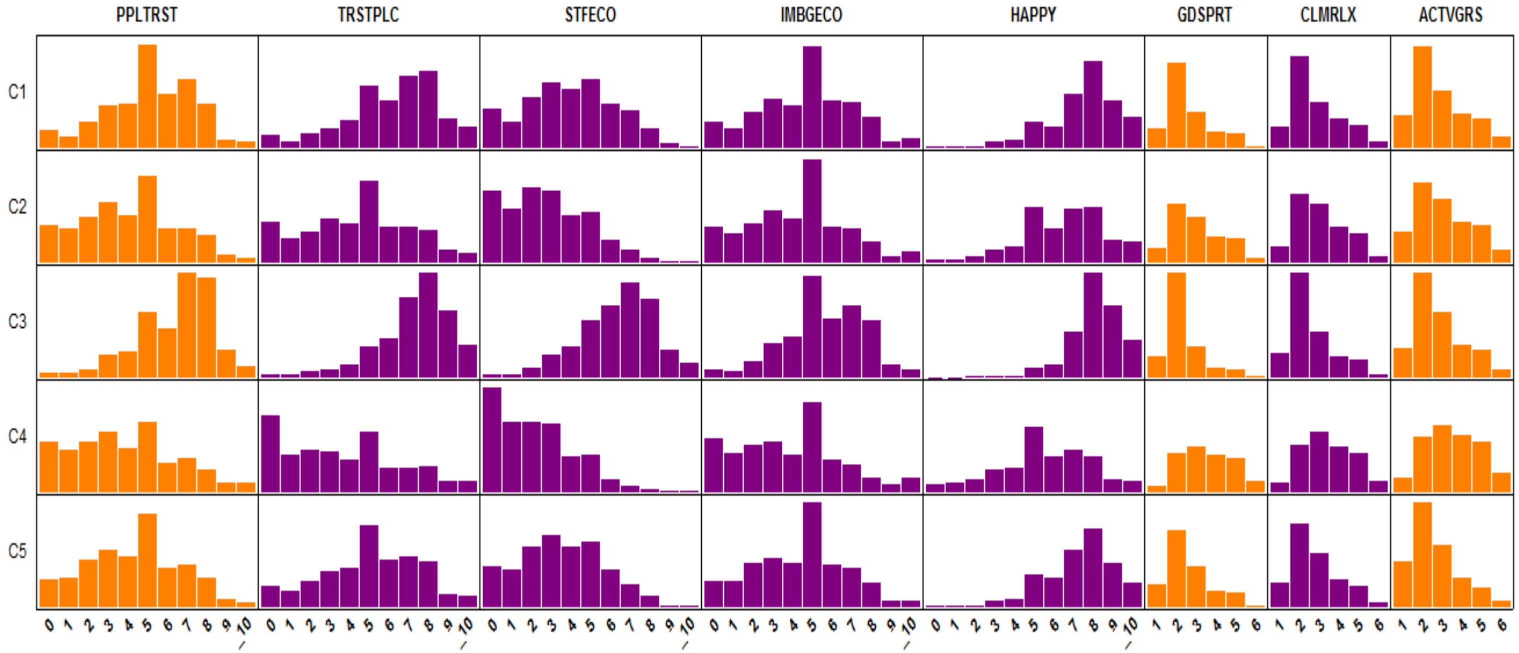
Cluster C2 is the cluster with young people from Bulgaria (BG), Greece (GR), Russia (RU) and Ukraine (UK) And people>50y from Czech Republic (CZ), Croatia (HR), Hungary (HU), Poland (PL), Portugal (PT), Slovenia (SI), Slovakia (SK)

C3 is the cluster with experienced and young people Switzerland (CH), Denmark (DK), Finland (FI), Netherland (NL), Norway (NO), Sweden (SE)

C4 is the cluster with people >50y from Bulgaria (BG), Greece (GR), Russia (RU) and Ukraine (UK)

C5 is the cluster of young people from Cyprus (CY), Czech (CZ), France (FR), Croatia (HR), Hungry (HU), Poland (PL), Portugal (PT), Slovenia (SI), Slovakia (SK) and people>50y from Cyprus (CY)

Description of the prototypes with StatSyr : social variables



We note that the cluster C3 obtains the best results for quite all the variables and mainly for the variables PPLTRST (Most people can be trusted or you can't be too careful), STFECO (Satisfaction with present state of economy in country) and TRSTPLC (Trust in the police). For this cluster, the histograms are clearly inclined towards the good scores (8, 9, 10). On the contrary, clusters C2 and C4 have histograms inclined toward the bad scores. Cluster C4 also gets scores much worse than other clusters for variables IMWBCNT (Immigrants make country worse or better place to live) and HAPPY (How happy are you). We note good scores for the clusters C1, C3 and C5 for the variable HAPPY. Interestingly, the last three variables on 'feelings' are focused on right (non-good 'feelings') in cluster C4 (experienced peoples in those countries.