Analyzing European Social Survey data using symbolic data methods and *Syrokko* software

Filipe Afonso*, Seppo Laaksonen**

*SYROKKO, Aéropôle, Paris - Roissy CDG Cedex, France
afonso@syrokko.com
**University of Helsinki, Finland
Seppo.Laaksonen@helsinki.fi

Abstract. The paper presents an application of Symbolic Data Analysis (SDA) with SYR software of Syrokko company to the fifth round of the European Social Survey (ESS) carried out among European inhabitants. In this study, we are not interested in studying the people themselves but by the comparison of different European countries, or different regions of Europe (Western Europe, Eastern, Nothern, Southern), or some groups of inhabitants by age, gender, region, etc. Here, we study, however, the 52 European countries by age groups. We describe each of them by all the results of its inhabitants using symbolic data. Symbolic Data Analysis (SDA) is proving so useful to aggregate up micro data (at the level of the inhabitants) to higher level units called concepts (the countries or European regions, for instance), using symbolic barchart or interval-valued variables. Using this aggregation we lose less information than occurs in classic analysis because symbolic data allow keeping the variation within the concepts. It allows keeping the variation of the results at the level of the inhabitants when they are aggregated up to their country. Hence, this could be called as "smart aggregation".

1. Introduction

The Sodas software was developed in two EU research projects, that is, Sodas and Asso. The newest version of the software is from 2004. The book, edited by Diday and Noirhomme (2008) was much based on this development. Since this first software for SDA such technology has been considered to be highly appreciated by data miners. As far as data mining with big data are concerned, the symbolic approach basically includes the two steps, one for aggregating smartly big micro data sets, and then for analyzing such aggregated data. This paper follows this approach so that our big micro data are the multinational survey data, but the aggregation is made by the new symbolic data software, SYR.

Our paper is next organized so that the principles of the SYR have been explained in Section 2. Our big data are the fifth round of the European Social Survey (ESS), collected late 2010 and the first half of 2011. The ESS is an academically-driven social survey designed to chart and explain the interaction between Europe's changing institutions and the attitudes, beliefs and behaviour patterns of its diverse populations. In annex, the details of our initial data and principles of construction are given. SYR, naturally, has performed the aggregation itself. The remaining sections present examples on SYR techniques to analyze the aggregated

data. We do not concentrate on subject-matter analysis but a reader will find some interesting results when comparing 26 European countries.

2. The Syrokko software for symbolic data analysis (SDA)

The SYR software is a SYROKKO company product (See Afonso et al., 2013). Its aim is to extract from one or several micro data files a reduced number of units called "concepts" which summarize the initial data. These units are described by standard categorical or numerical variables, as well as by interval variables, multi-valued variables and by bar-chart and histogram-valued variables. These new kinds of variables allow keeping the internal variation of each concept. In the following sections, we will use different solutions of the software.

TabSyr solution is for data fusion of several multisource and heterogeneous files into a unique symbolic data file. It allows creating a symbolic data file, visualizing a symbolic data table thanks to a user-friendly graphical output, handling a symbolic data table (select, cut, move…) and using statistics methods to score the symbolic variables from the most discriminant to the least discriminant (and conversely) of the different concepts.

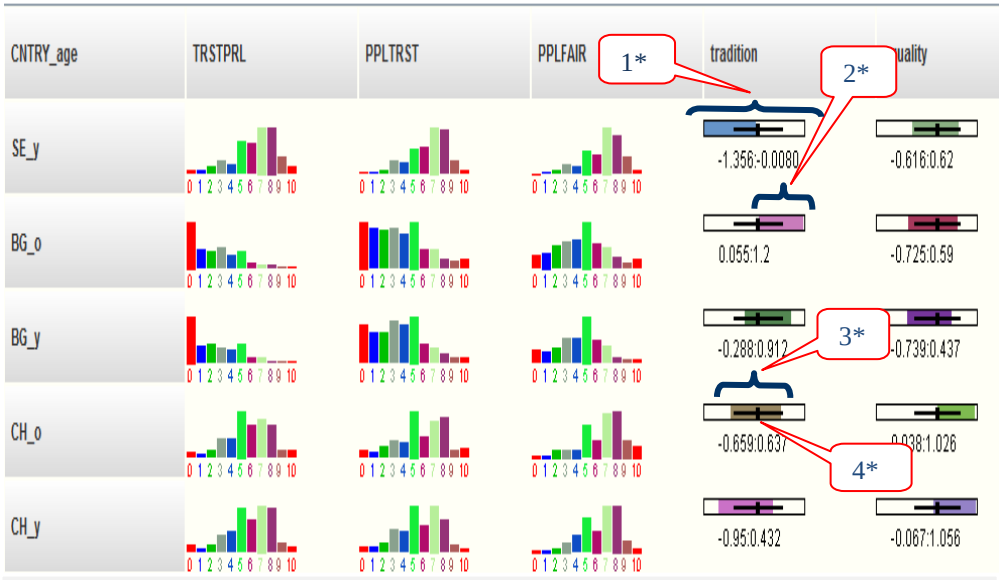
ClustSyr solution is for k-means clustering of the concepts extended to symbolic data (histogram-valued, bar-chart, interval-valued variables). The different types can be mixed in a same clustering and also with classical data (continuous, nominal variables). It allows performing a partition or an overlap clustering. Moreover, this solution can communicate with the NetSyr solution (see below) in order to display a clustering on a factorial plane. This clustering can be performed on the concepts described by the symbolic variables as well as on the coordinates of the concepts on the factorial axes.

StatSyr is for statistics on concepts and clusters. It allows finding the categories which characterize the concepts / clusters and showing their variation. Moreover, it allows analyzing correlations between symbolic variables (for example, between interval or bar-chart variables).

Finally, NetSyr solution is for extending PCA to symbolic data (see method in Diday, 2013). It 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 histogram, bar-chart and interval type mixed with continuous variables. The different types 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: concepts with their variation, clusters of concepts (from a partition or an overlap clustering performed with ClustSyr), proximities between concepts thanks to networks, etc.

3. Data descriptions

We examine the 52 (26 x 2) concepts country x age. The country symbols are standard international codes whereas we use the two age groups, 'younger' and 'older'. For example, the acronym SE_y means younger than 50 years old (<50) in Sweden and CH_o, older or equal to 50 years old (>=50) in Switzerland. The interpretation of all the variables of our data is given in the appendix . We visualize, in Figure 1, an extract of the symbolic data matrix with TabSyr.



***Interpretation of the graphical representation of the interval values.**

1* Min of the Min and Max of the Max of all the interval values of the column (represented by a black box).

2* The colored box represents the interval value for the variable "tradition" and the concept BG_o. The interval value changes from one concept to another so that the color also changes (as well as length and position). The interval value is given below the graphical representation.

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

4* Mean of all the means of the intervals of the column (black vertical line).

FIG. 1. *An illustration of the symbolic data table for three bar-chart and two interval variables using TabSyr.*

Then, we use StatSyr module. StatSyr characterizes the concepts and shows their variation. It offers several methods for symbolic variables as the matrix of bar-chart valued variables shown in Figure 2. In this figure, we see an example of comparison between young people from Germany (DE), Spain (ES), France (FR), Greece (GR) and Sweden (SE).

We note the much better results of Sweden for all the variables since the frequencies of the categories are higher at the right of the bar-charts (good scores 5 to 10). On the contrary, we see the very bad results of Greece for all the variables (the frequencies are more important for the bad scores 0 to 4). The other countries are between Greece and Sweden with better results for Germany than for France and Spain for the variable STFECO (How satisfied with present state of economy in country). We also note poorer results for IMWBCNT (Immigrants make country better place to live) where the most frequent score is 5 for the 5 countries. Finally, it is interesting to note that the results are similar for the variable "happy" among all the countries in the exception of Greece.

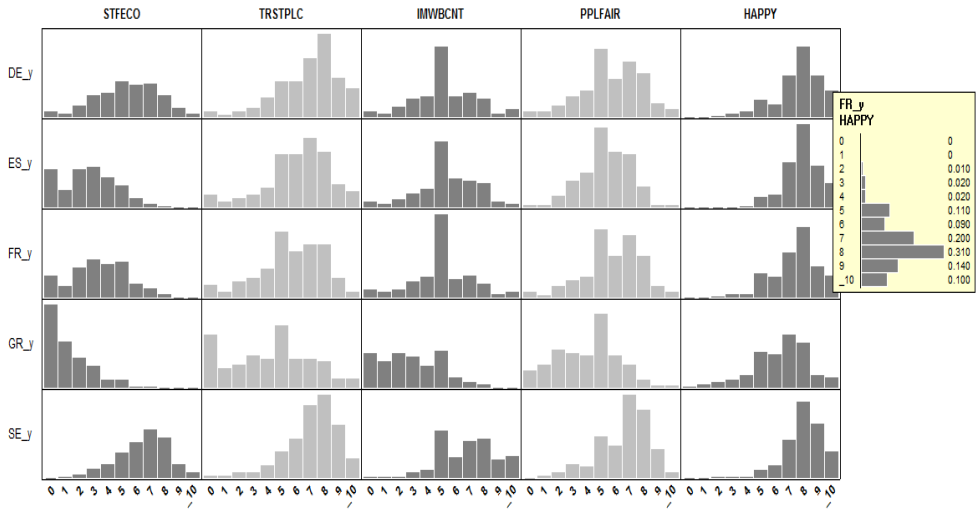


FIG. 2. *An illustration of the symbolic data table for five bar-chart variables using StatSyr.*

Moreover, we looked at our four interval-values variables concerning people's life values. We find that 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. These 5 countries are, interestingly, quite similar, according to the "enjoy" variable. Moreover, we find that success is much more important in Greece than in Sweden or France.

Bar-chart variables allow to compare and discriminate the countries and the clusters of countries as we clearly note the differences between their desciptions. For instance, in Figure 2, when we look at the variable STFECO, we clearly note that the bar-chart variable is inclined to the left for Greece (bad results), and inclined to the right for Sweeden (good results). Several variables are correlated as we obtain the same results for the variables TRSTPLC, IMWBCNT, PPLFAIR as for STFECO. Life values variables (equality, tradition, enjoy, success), however, have a different "behavior" compared to all the other variables.

People younger and older than 50 years old of a same country are not always in the same cluster. In particular, for Greece, where they are separated on immigration issues. People older than 50 years old are very strongly against immigration. For the countries where older and younger people are not in the same cluster, people older than 50 years are always the most pessimistics and unhappy.There is a cluster with only people older than 50 years old from Russia (RU), Ukraine (UA), Bulgaria (BG) and Greece (GR). They are the most pessimistics and unhappy in the analysis. On the contrary, there is a cluster with people older and younger than 50 years old with very good results for quite all the variables. It is the cluster with scandinavian countries, Switzerland and The Netherlands.

To perform the analysis, we use symbolic data, a clustering method, a PCA method and several visualization tools applied to the concepts "countries x age". At the end of this paper, we see that we can apply the same methods to the 5 prototypes describing 5 clusters of concepts. These prototypes are the symbolic representations of the clusters resulting form the generalization of the characteristics of the concepts belonging to the cluster (see Diday and

Noirhomme, 2008). Here, the prototypes are described, in each column, by the mid-values, of the bar-charts or the intervals, calculated among all the concepts in a given cluster. We reach the same conclusions when we study the 5 prototypes as when we study all the 52 concepts "countries x age."

4. Principal component analysis (PCA) and other SDA analysis

In the previous pictures, we can compare a few concepts for several variables at the same time. Next, we use NetSyr module to visualize all the concepts in a same biplot (factorial plane resulting from a PCA method on symbolic data, see section 2) and to compare them. In Figure 3, we visualise 5 clusters of concepts "country x age". These clusters are visualized in different colours in the first factorial plane. They are obtained with k-means method 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). The Netherlands are 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 old from Russia (RU), Ukraine (UA), Bulgaria (BG) and Greece (GR) (Cluster C4). Western countries (except Portugal and Greece) are at the middle of the plane (cluster C1). Curiously, Estonia is also in the "Western" cluster C1 whereas Portugal is at the right with the eastern countries (Clusters C2, C5). Younger people and older people from France or Portugal are not in the same clusters. We give the full description of the five clusters of concepts "age x group":

- Cluster C1 is the cluster with Belgium (BE), Germany (DE), Estonia (EE), Spain (ES), Great Britain (GB), Ireland (IE), and Israel (IL) and young people from France (FR).
- Cluster C2 is the cluster with young people from Bulgaria (BG), Greece (GR), Russia (RU) and Ukraine (UA) and people>50y from Czech Republic (CZ), Croatia (HR), Hungary (HU), Poland (PL), Portugal (PT), Slovenia (SI), Slovakia (SK).
- C3 is the cluster with old and young people from Switzerland (CH), Denmark (DK), Finland (FI), The Netherlands (NL), Norway (NO), Sweden (SE).
- C4 is the cluster with people >50y from Bulgaria (BG), Greece (GR), Russia (RU) and Ukraine (UA).
- C5 is the cluster of young people from Cyprus (CY), Czech (CZ), Croatia (HR), Hungry (HU), Poland (PL), Portugal (PT), Slovenia (SI), Slovakia (SK) and people>50y from Cyprus (CY) and France (FR).

Correlation circles allow to interpret the factorial axes. The variable TRSTPLC (Trust in police) is highly correlated with the first axis and also correlated with the second one. In figure 4, 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 except that they are very bad for Greece. The general tendency is also similar for the variables STFECO, STFHLTH, PPLTRST and PPLFAIR.

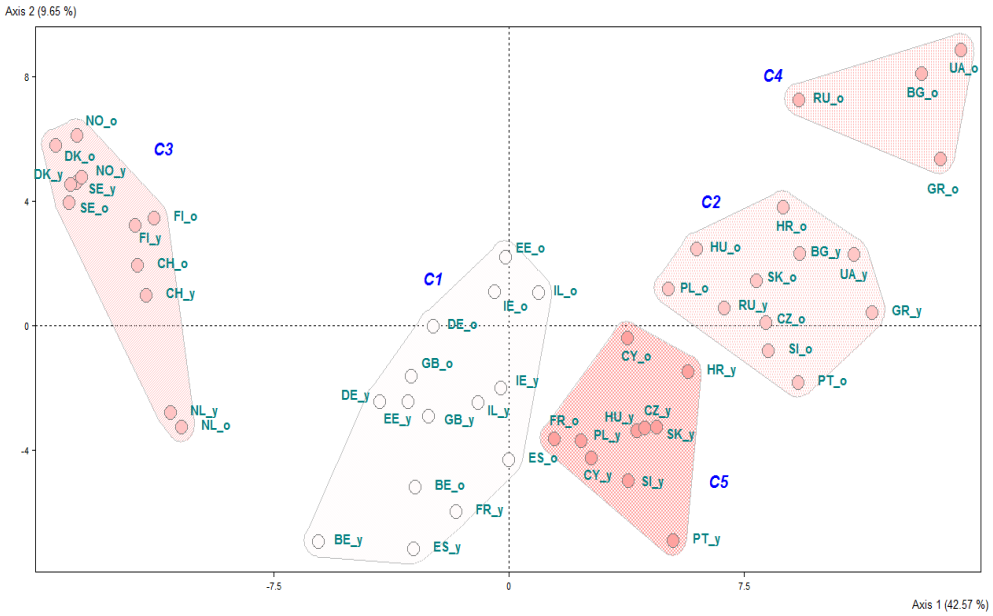


FIG. 3 - *Visualization of the concepts "Country x age" in the first factorial plane from a PCA analysis. Display of a clustering of the concepts into 5 clusters (k-means method).*

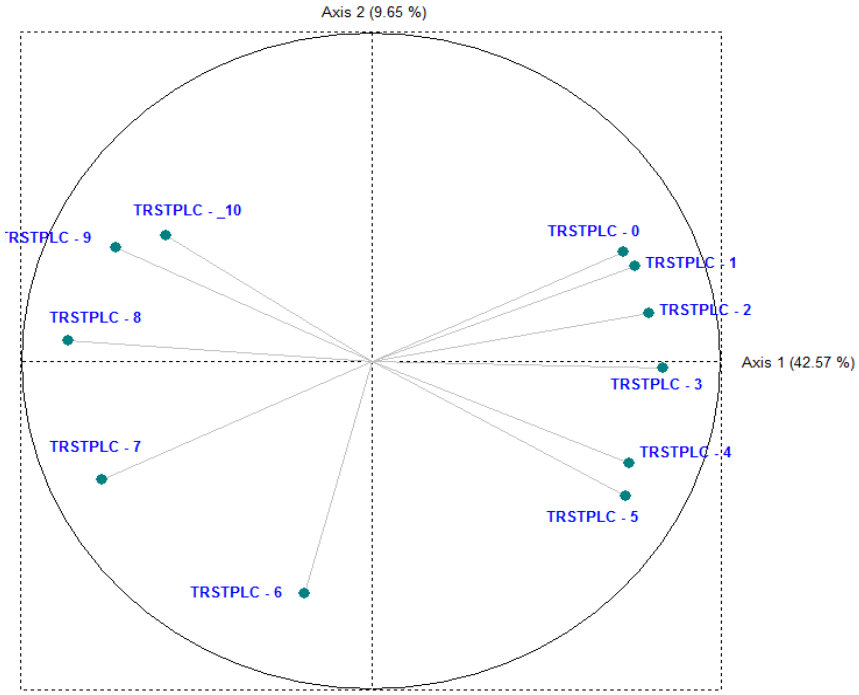


FIG. 4 - *Correlation circle : display of the categories of the bar-chart variable TRSTPLC.*

NetSyr also allows the display of symbolic variables on the factorial plane. In Figure 5, we visualize the bar-chart variable TRSTPLC (trust in police) and the interval variable "enjoy" simultaneously in the same factorial plane. For each concept "country x age", we visualize the bar-chart variable thanks to pie charts. By clicking on each pie chart, we can obtain the details of each bar-chart. Light colors are given to "bad" scores from 0 to 5 and dark colors are given to "good" scores from 6 to 10. We note the very bad results of Russia, Bulgaria and Ukraine at the upper right (with more "light colors") and the very good results of Finland, Switzerland, Denmark, Sweden, Norway at the left (with more "dark colors"). Concerning the "enjoy" interval-valued variable, the representation is the same as for TabSyr. Each value is indicated by a blue box in full within an empty black box indicating the min of the min and the max of the max among all the concepts. We note, for instance, top right in cluster C4, that the blue box, for Ukrainian older than 50 years old, is located at the extreme left of the black box. This means they have the poorest results. On the contrary, top left in cluster C3, young Swedens have the best results since the blue box is at the extreme right of the black box. We had the same results for quite all the variables that means the best scores are in the cluster at the upper left of the factorial plane and these scores decrease from the upper left to the cluster at the upper right through the clusters at the lower left and at the lower right. "enjoy" is an exception to this conclusion and the other "life variables" also (tradition, equality and success) as the clusters are not very characterized by the values of "life variables".

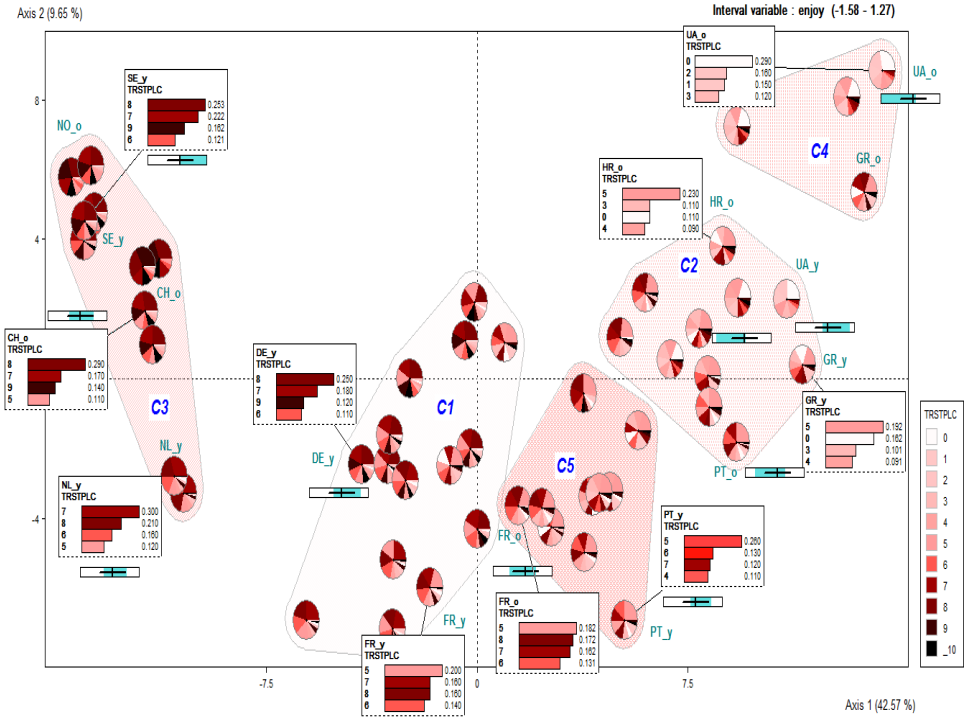


FIG.5. *Display of variables TRSTPLC (bar-chart) and enjoy (interval) in the factorial plane.*

In Figure 6, we visualize the interval-valued variable “enjoy” in the correlation circle. For interval-valued variables, the program calculates four different correlations, for the min of the intervals, the max, the mid-value, and the range of the intervals (called extent in the graphic). We note that this variable is slightly correlated with the axis 2 (more than with axis 1). Especially, we see that it is in the upper right quadrant that we have the countries with the highest disparities (extent) for the variable “enjoy”. In the lower left quadrant, we have the minimum value as well as the maximum value. This means that, for this variable, we have very different scores within the same clusters.

We displayed many other variables like in figure 5. The variable HAPPY shows, especially, the separation between East Europeans (less happy) and the rest of Europe (happier). 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 Nothern countries and very bad results for Greece where 0 has the highest frequency for people over 50 years old. For the variable IMWBCNT (Immigrants make country better place to live), the results are the same than IMBGECO but with lower values for all the countries. The category 5 has the highest frequency for all the concepts except people over 50 years old from Greece where the highest frequency is 0.

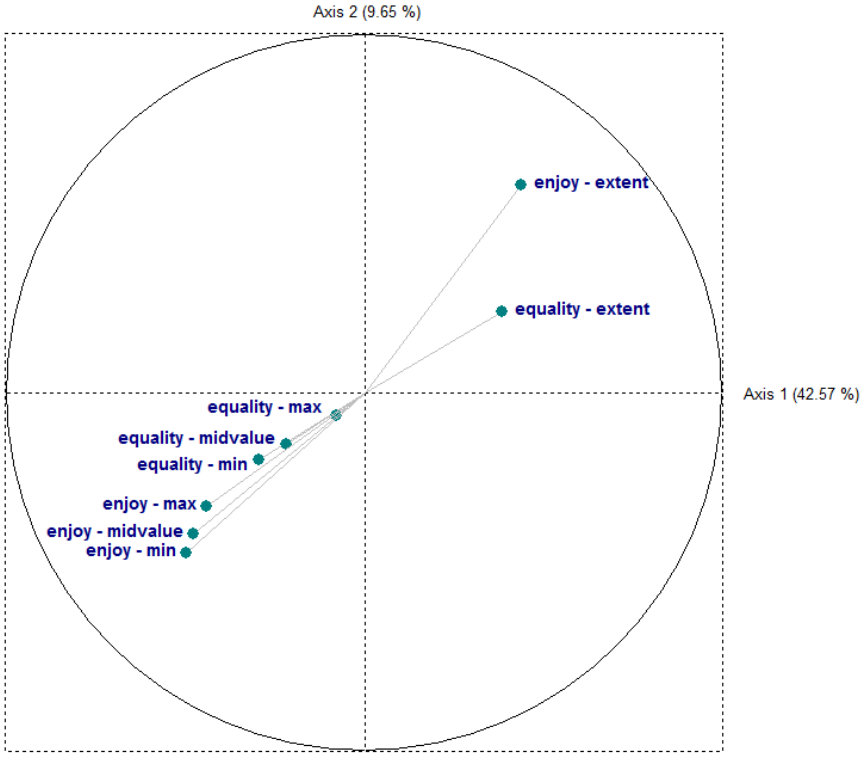


FIG. 6. *Correlation circle : display of the interval-valued variables “enjoy” and “equality”.*

In order to achieve this demonstration, we would like to show that it is possible to analyze directly some clusters of concepts in the same way we analyze the concepts themselves. Thus, after executing ClustSyr from the factorial plane of NetSyr, we obtain 5 clusters of concepts "country x age". The ClustSyr module provided the 5 prototypes describing these clusters. These prototypes can be visualized in a new factorial plane. We perform a new PCA on the prototypes in order to show that we can apply the symbolic data methods directly to them. In Figure 7, we see the factorial plane calculated on the table of the 5 prototypes. We also visualize, for each prototype, the bar-chart variable "TRSTPLC" and the interval-valued variable "success."

It is interesting to note that we obtain the same general conclusions when analyzing a reduced number of clusters (five clusters) instead of analyzing all the concepts. This is because we use symbolic data preserving the data variation during the aggregation from the initial units (inhabitants) to the concepts (country x age) and then to the clusters of concepts.

C4 has the worst results for the variable TRSTPLC (more "light colors", scores 0 to 5) whereas C3 has the best results (more "dark colors", scores 6 to 10). We do not have similar conclusions with the interval-valued variable "success". C3 has the worst results (the blue box is at the extreme left of the black box) and C2 has the best results. Thus, success is more important in cluster C2 than in cluster C3.We also note that the extent (range) of the interval-values is important for all the clusters. This means that we have very different scores, for success, within the same clusters.

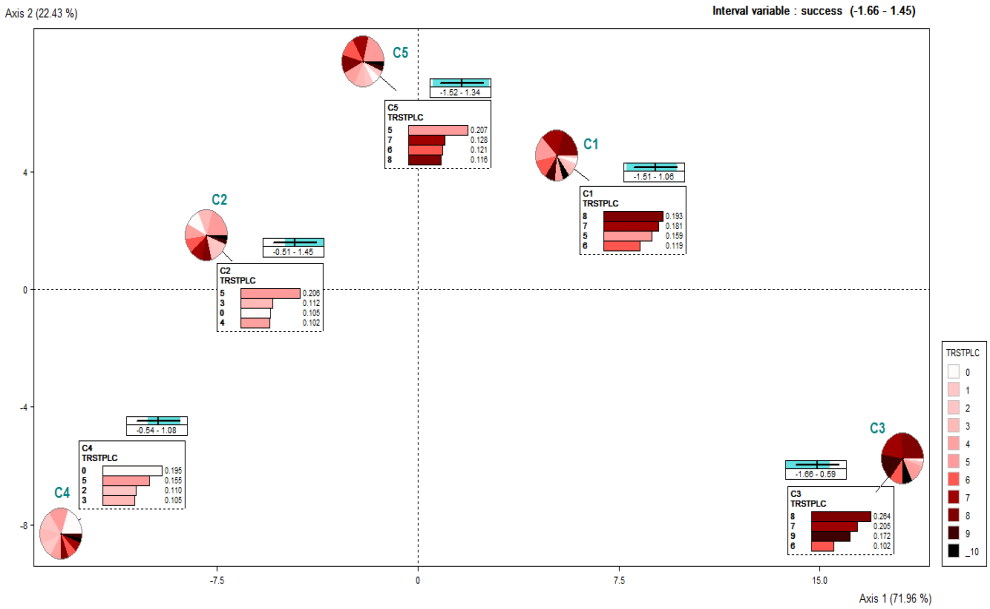


FIG. *7 – Analysis of the 5 prototypes describing the 5 clusters instead of analyzing all the concepts "countries x age". Display of the symbolic variables TRSTPLC and "success" in the first factorial plane of a symbolic PCA.*

We next use STATSYR in order to have a complete description of the 5 prototypes. In figure 8, we first present the variable age and country used for the construction of the con-

cepts. We note, for the variable “Country”, that the bars are not of the same size from C1 to C5. It is because the frequencies depend on the number of concepts in the cluster, and also, on the presence or not of the two concepts (country x age>=50) and (country x age<50) in the same cluster.

Then, in Figure 9, we display the results for some other variables. We note that, for cluster C3, the bar-charts are clearly inclined towards the good scores (8, 9, 10). That means that 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). On the contrary, clusters C2 and C4 have bar-charts inclined toward the bad scores. Cluster C4 also gets scores much worse than other clusters for variables IMWBCNT (Immigrants make country better place to live) and HAPPY (How happy you are). We note good scores for the clusters C1, C3 and C5 for the variable HAPPY.

5. Conclusion

This paper is an illustration of symbolic data analysis (SDA) using a rather new software SYR applied on data that have been aggregated up from a big survey micro data file. There is a high need for such kind of data mining for many reasons. One is to continue the analysis from the standard micro data analysis. This is helpful since it is hard to compare efficiently some aggregates with each other with classical techniques. Here, our aggregates are very concrete, that is, 26 European countries divided in two age groups. Another advantage for SDA against classic aggregate analysis is to lose less information. Moreover, symbolic data are quite comfortable to analyze when compared with huge data sets. Finally, symbolic data are easily made confidential and can be released to anyone without legal problems. We hope that this two-stage strategy will become more common in future, amongst survey data analysts.

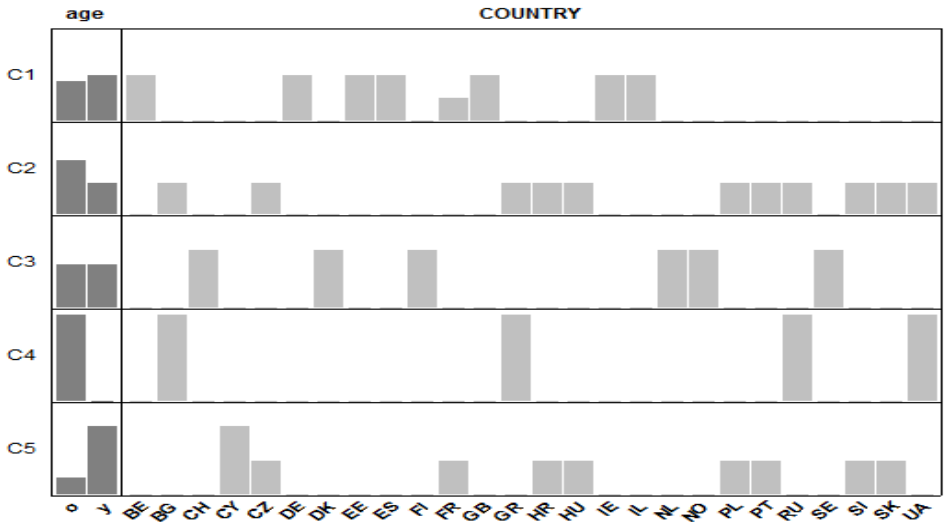


FIG. 8 - *An illustration of the symbolic data table for the five prototypes describing the five clusters using StatSyr. Variables : age and COUNTRY.*

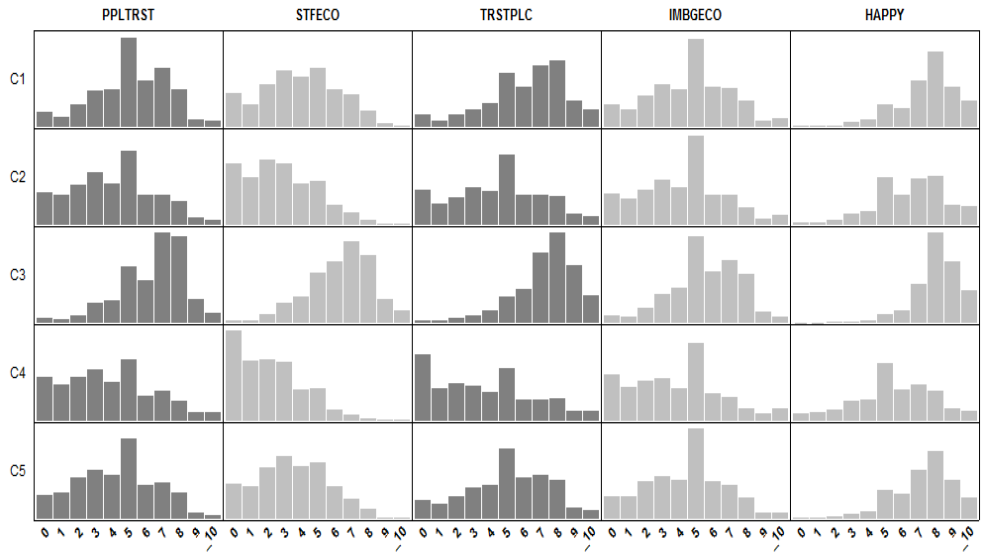


FIG. 9 - *An illustration of the symbolic data table for the five prototypes describing the five clusters using StatSyr. Variables : PPLTRST, STFECO, TRSTPLC, IMBGECO and HAPPY.*

References

Afonso, F., Haddad, R., Toque, C., Eliezer E.-S., Diday, E. (2013). *User Manual of the SYR Software*, Syrokko internal publication, 70pp., internal document

Billard, L. and Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. 321 pages. Wiley series in computational statistics. Wiley, Chichester, ISBN 0-470-09016-2.

Diday (2013) E. "Principal Component Analysis for bar charts and Metabins tables". *Statistical Analysis and Data Mining*. Article first published online: 20 May 2013. DOI: 10.1002/sam.11188. 2013 Wiley. Statistical Analysis and Data Mining,6,5, 403-430.

Diday, E. and Noirhomme-Fraiture, M. (eds and co-authors) (2008). *Symbolic Data Analysis and the SODAS software*. Wiley, Chichester, ISBN 978-0-470-01883-5.

Laaksonen, Seppo (2008). People's Life Values and Trust Components in Europe - Symbolic Data Analysis for 20-22 Countries. In. *Edwin Diday and Monique Noirhomme-Fraiture, "Symbolic Data Analysis and the SODAS Software", Chapter 22*, pp.405-419. Wiley and Sons: Chichester, UK.

Laaksonen, Seppo (2010). The Survey as a Basis for Symbolic Data Analysis. In: *Official Statistics , Methodology and Applications in Honour of Daniel Thorburn* (Eds. Michael Carlson, Hans Nyquist and Mattias Villani), 93-106.

The ESS data archive: *Online;accessed 5 -Sept-2012]*. http://ess.nsd.uib.no/ess/round5/

APPENDIX – micro variables

The micro survey data of the fifth round of the European Social Survey (ESS) are the source of the SYR file. The weighting variable DWEIGTH is for individuals and applied to categorical variables when calculating the frequencies of the bar-charts, but not for the life value variables that are interval.

Values and categories for the following three ones :

6 All of the time, 5 Most of the time, 4 More than half of the time, 3 Less than half of the time, 2 some of the time, 1 At no time
ACTVGRS 'Have felt active and vigourous last 2 weeks.'
CLMRLX 'Have felt calm and relaxed last 2 weeks.'
GDSPRT 'Have felt cheerful and in good spirits last 2 weeks.'

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

HAPPY 'How happy you are.'
IMBGECO 'Immigration is good for country's economy.'
IMWBCNT 'Immigrants make country better place to live.'
PPLFAIR 'Most people try to be fair.'
PPLHLP 'Most of the time people are helpful.'
PPLTRST 'Most people can be trusted.'
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.'

Finally, we have the four **life value variables** created from 21 questions (micro-data) by exploratory factor analysis. They have been created with classical methods before this study. In this study, we directly use these 4 variables. For these variables, the mean of the micro-values is equal to 0, and the standard deviation is equal to 1 and we only use the values in the inter-quartile interval.

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.'