

Symbolic Data on official statistics and visual exploration

Antonio Irpino

Symbolic Data Analysis on Official statistics

Symbolic data analysis plays an important role in the field of official statistics. Let me provide you with some relevant information:

Symbolic Data Analysis (SDA):

- SDA is a technique used to analyze complex data structures. It allows us to describe data with intricate relationships and patterns.
- In the context of official statistics, SDA provides a way to handle various forms of symbolic data efficiently.
- Symbolic data can represent individuals or groups of individuals, and it captures information beyond traditional numerical values.
- The goal of SDA is to enhance our understanding of data by considering both quantitative and qualitative aspects.

Applications in Official Statistics:

- Symbolic objects are used to describe complex data structures in official statistics.
- These objects extract information from databases and represent individuals or categories of individuals.
- By using symbolic data, statisticians can capture more nuanced information beyond simple numerical values.

In summary, symbolic data analysis provides a powerful framework for handling diverse and complex data in official statistics. It allows us to go beyond traditional numerical analysis and consider the rich context of information.

Practical Examples:

- In official statistics, symbolic objects describe complex data structures. Here are some practical examples:
 - **Symbolic Objects for Individuals:**
 - Imagine a dataset representing citizens in a country. Each citizen has attributes like age, education level, occupation, and income.
 - Instead of treating these attributes as separate numerical values, we create symbolic objects that encapsulate this information.
 - For instance, a symbolic object could represent a “young, highly educated professional with a high income.”

- In official statistics, symbolic objects describe complex data structures. Here are some practical examples:
 - **Symbolic Objects for Categories:**
 - Official statistics often deal with categories (e.g., regions, industries, products).
 - Symbolic objects can describe these categories based on various attributes.
 - For example, a symbolic object might represent a “rural agricultural region with low GDP per capita.”

Applications:

- Symbolic data analysis finds applications in:
 - **Quality Reports:** Enhancing the understanding of data quality by considering both quantitative and qualitative aspects.
 - **Exploratory Data Analysis:** Uncovering patterns and relationships in complex datasets.
 - **Policy Formulation:** Informing policy decisions by analyzing symbolic data related to social, and economic.

In summary, symbolic data analysis enriches our understanding of complex data structures in official statistics. It allows us to go beyond mere numbers and consider the context and meaning behind the data.

DATA Fusion

Certainly! Data fusion is becoming increasingly relevant in official statistics. The aim of data fusion is to combine two or more data sources using statistical methods in order to analyze different characteristics that were not jointly observed in one data source². Let's delve into this topic further.

Data fusion involves integrating multiple related datasets. It allows statisticians to manage uncertainty and conflicting data on a large scale. The goal of data fusion is to create useful representations of reality that are more complete and reliable than what a single source of data can provide⁴.

Here are some key points related to data fusion in official statistics:

1. Principles of Official Statistics:

- Official statistics play a crucial role in informing policy decisions, research, and public understanding.
- Data fusion helps enhance the quality and completeness of official statistics by combining information from various sources.

2. Scenarios for Data Fusion:

- Data fusion scenarios can be explicit or implicit.
- Explicit scenarios involve combining data from different sources explicitly (e.g., merging survey data with administrative records).
- Implicit scenarios involve indirect data fusion (e.g., imputing missing values based on related variables).

3. Imputation Approaches:

- Classical imputation approaches include:
 - **Distance Hot Deck (DHD):** Imputes missing values by matching similar records.
 - **Regression Model (RM):** Uses regression equations to predict missing values.
 - **Predictive Mean Matching (PMM):** Imputes missing values by matching observed values with similar patterns.
- Statistical learning approaches include:
 - **Decision Trees (DT):** Used for classification or regression tasks.
 - **Random Forest (RF):** Ensemble of decision trees.
 - **Predictive Value Matching (PVM):** Matches predicted values from a model.
- These approaches help handle missing data during data fusion¹.

4. Example: EU-SILC and HBS Data Fusion:

- Consider combining data from the European Union Statistics on Income and Living Conditions (EU-SILC) and the Household Budget Survey (HBS).
- Motivation: To analyze characteristics that were not jointly observed in either dataset.
- Simulation design: Create a database and perform a Monte Carlo study.
- Results: Evaluate compliance with the conditional independence assumption (CIA) and discuss the implications¹.

In summary, data fusion in official statistics allows statisticians to leverage multiple data sources effectively, leading to more comprehensive and accurate insights.

1. OPUS 4 | Data Fusion in Official Statistics: An Evaluation of Classical <https://ubt.opus.hbz-nrw.de/opus45-ubtr/frontdoor/index/index/year/2024/docId/2214>.
2. Data Fusion | SpringerLink. https://link.springer.com/referenceworkentry/10.1007/978-3-319-32010-6_305.
3. Data Fusion in Official Statistics: An Evaluation of Classical versus https://ubt.opus.hbz-nrw.de/opus45-ubtr/files/2214/Dissertation_Schaller.pdf.
4. Data fusion using factor analysis and low-rank matrix completion - Springer. <https://link.springer.com/article/10.1007/s11222-021-10033-7>.

Data fusion: classic approach, discussion

DF is the practice by which two or more separate data sources are brought together to form a single database that contains all the previously separate information. The purpose of such integration is to obtain a reliable estimate of the true relationship between any set of statistics which are currently unavailable in single-source form.

During the fusion, **individuals** from one survey are matched to individuals in the other and the two sets of behaviours are jointly ascribed to the matched individuals. For the sake of understanding the process, it is convenient to nominate one survey the donor and the other the recipient.

An important fundamental assumption of fusion is that “hooks” or linking variables contain enough information to describe the correlations between the variables in Donor and Recipient datasets. For example: “If given two datasets, Recipient, consisting of variable sets X and Y, and Donor, consisting of variables X and Z, we can perform fusion under the assumption that Y and Z are independent given X”.

It is important to note that fusion is not a single technique – there are different approaches that might be taken depending on the objectives. The principles for the different approaches of data fusion are quite similar and follow these general steps:

Set the objectives for the data fusion

- Analyse the datasets to select the variables and metrics that are subject to data fusion, and the relationships that need to be preserved. Define the universe and common variables.
- Some common variables may be described as critical and absolutely have to be maintained Socio-demographic variables such as Gender (so that males are always fused onto males), Age and Geography are commonly used as critical variables. Other critical variables may relate to other measured behaviour such as media consumption.
- Select matching (non-critical) variables used to further predict or explain the variables being linked within each critical cell. This step should also include assigning the importance weights to each of the non-critical common variables. Various approaches for the selection and weighting of matching variables can be used (e.g. ANOVA, regression, CHAID, Principal Component Analysis).

Choose an overall data fusion technique. The main techniques used are:

- “Row wise” (The entire record of each donor is fused onto a matching recipient or group of recipients. The same set of ‘common’ variables is used to estimate the true relationship between donor and recipient for all variables.)
- “Column wise” (Datasets are fused variable by variable, or ‘column-by-column’. Each variable will have its own set of ‘common’ variables, the one that best explains the relationship for this particular variable.)
- “Hybrid” (Dataset is divided in blocks of variables with similar characteristics. Within each block, the fusion will be row-based copying all variables from a donor to a recipient for that particular group of variables. However, for each block, the fusion is independent and can take advantage of choosing the different set of common variables that best explains the relationship.)

Select a single distance metric to be used in the matching process. Usual approaches include Euclidean distance, Mahalanobis distance, or the Manhattan block distance.

Choose a matching technique – “constrained” or “un-constrained.”

Unconstrained fusion

- In approaches using “unconstrained fusion”, the donor variables are passed across and attached to the recipient’s record. Unconstrained matching simply means that a donor can be matched to any number of recipients, or none. This approach has the advantage of permitting the closest possible match of donors with recipients because it does not require for all potential donors to be used. However, there are a number of disadvantages. The most important disadvantage is that there is no guarantee that the marginal and joint distribution of the donor variables in the fused dataset will be identical to the corresponding distributions in the donor dataset. This is partly because each donor is used as often as necessary and partly because the donor’s weight is “left behind”.

Constrained fusion

- In the “constrained fusion” approach, all respondents from both surveys have to be used and their data together with their weights are transported into the new synthetic dataset. Respondents from either survey may still be used more than once, but in such cases their weight is shared out. This method has a number of advantages, but the main one is that marginal distributions from both surveys can be preserved, due to the preservation and sharing out of both sets of weights. Though in principle the matching may not be as good as for unconstrained fusion, in practice this is not an issue, especially when the two surveys with large samples are being fused.

Run the matching (or modelling) process.

- Create a new fused dataset.
- Validate the results and provide fusion diagnostics. For example:
 - Evaluation of the matching algorithm to measure the success in finding fused pairs of individuals with similar profiles.
 - Comparison of the incidences for key fused variables between the original and fused dataset.
- Assess the effect of “regression-to-the-mean.” Analyse the fused dataset, combining variables from the donor and the recipient datasets.

For example, in marketing, marketers want to know everything they can about their target consumer in order to maximise the return on their research investment:

- Who they are (demographics, geo-demographics, psychographics etc.)
- What they think about brands in the category they are asking about
- How they behave (purchasing levels, brand choice etc.)
- What they intend to purchase in the future

Those planning and buying advertising campaigns need to uncover the best ways of reaching and influencing their target audiences:

- Which media do they come into contact with at different times of day (TV programmes, newspapers, magazines, radio stations, web sites, apps, poster panels...)?
- Which media are they more or less attentive to or engaged with at different times of day?
- When is the best time to reach people with an advertising message (message receptiveness, when are they in the market to buy...)?

But no individual respondent will agree to answer such a large number of questions. And many of the questions will be impossible to answer accurately.

This conundrum of needing more information while finding it harder to collect from surveys alone is likely to get harder rather than easier as time passes. One of the statistical techniques used to help address this is data fusion.

The practical application of data science demands a high level of skill and expertise, as well as experience – many of the decisions and choices made in building fusions, for example, are not black and white, demanding judgement and a deep knowledge of the context.

Going back to surveys of Official Statistics Institutes

We match **concepts/classes** (not the single individuals), which represents aggregation of microdata, not always of the same nature (income is measured on citizens, health services are measured on hospitals, land cover use is measured on the geographical space,...) from different database/surveys e.g.

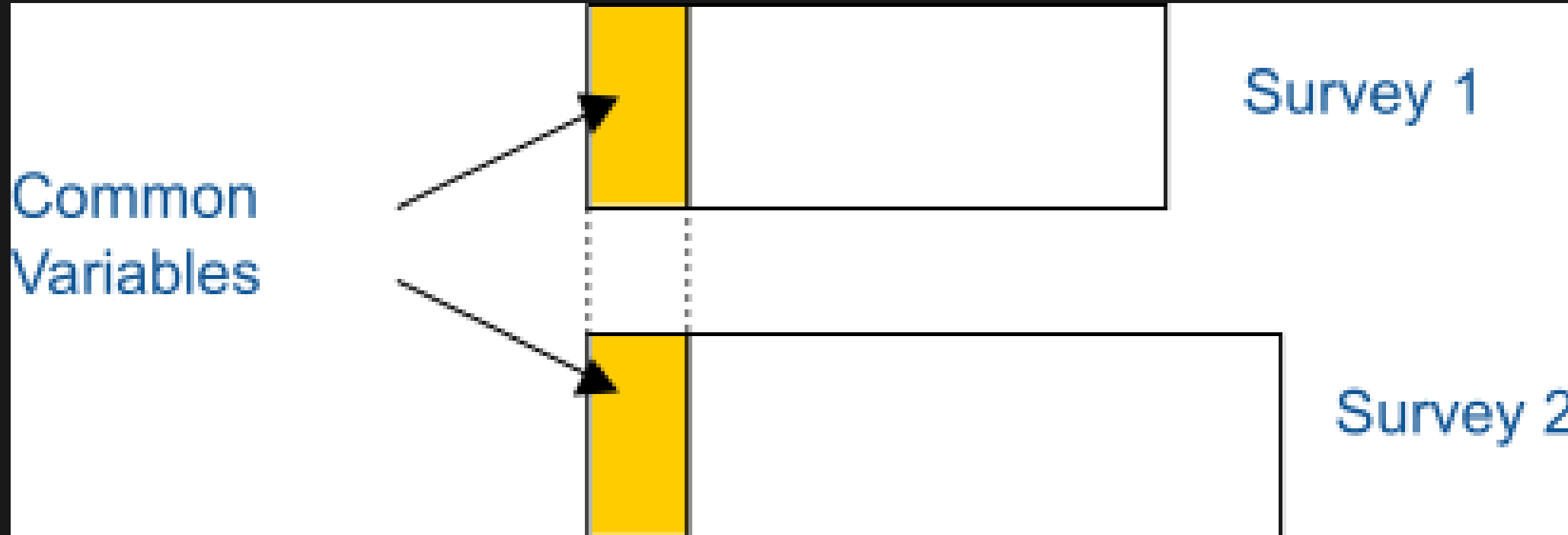
- Incomes
- Time use
- Behaviours
- ...

Data Fusion: the Symbolic advanced approach

Uses of Joining Objects in Official Statistics: Data Fusion We find a new application of joining symbolic objects that consists of joining assertions coming from different surveys. This joining enables us to obtain additional information, data imputation, to obtain conclusions about causes and possible effects,...

Fusion using symbolic objects differs from the traditional data fusion in the way of matching. Instead of joining record by record of common variables, we join by symbolic objects each one describing a group.

The fusion allows us to relate independent surveys to some common items. In the SODAS project framework, this comparison will be between surveys of different countries of the European Union.



Example

An example showing this new use in EUSTAT, is the fusion of the two independent surveys Use of Time (EPT) and Living Conditions (ECV). They have common variables (socio-demographic) and it is probable that there is a relation between them.

- The first step is to define the common socio-demographic variables and to create assertions for each survey separately. The group attribute for these assertions will be the concatenation of common variables.

The common variables chosen for this study could be: Sex, Marital status, Age, Relation to Activity and Level of Education.

- The second step is to join assertions describing the same group. Then, for the same group we will have the description in the specific variables of each survey.

We consider the following data arrays:

X_1 is a symbolic data array that describes socio-demographic groups by the following variables of Use of Time:

$Y_{11}(limp) = \text{Participation in Cleaning}$

$Y_{12}(prpc) = \text{Participation in Preparing Meals}$

$Y_{13}(prac) = \text{Sport Practice}$

$Y_{14}(cuip) = \text{Time used in Personal Care}$

One of the objects of the array is:

```
os "Woman Married < 35 years Employed Secondary"(54) =  
[limp = {"Null Particip."(0.347273), "Low Particip."(0.188186), "Average  
Particip."(0.346782), "High Particip."(0.117759)}] &  
[prpc = {"Null Particip."(0.0719004), "Low Particip."(0.400066), "Average  
Particip."(0.436589), "High Particip."(0.0914451)}] &  
[prac = {"Null Particip."(0.877218), "Low Particip."(0.122782)}] &  
[cuip = [0:170]]
```


X_2 is a symbolic data array that describes the same socio-demographic groups by the following variables of Living Conditions:

$Y_{21}(jorna) = \text{Length of Working Day}$

$Y_{22}(comt) = \text{Return home to have lunch}$

$Y_{23}(ractp) = \text{Branch of Economic Activity}$

Symbolic description from the X_2

```
os "Woman Married < 35 years Employed Secondary"(34) =  
[jorna = {"SPLIT SHIFT"(0.394297), "CONTINUOUS"(0.434047), "NOT APPLICABLE"  
(0.171656)}] &  
[comt = {"RETURN HOME TO LUNCH"(0.637714), "NOT RETURN HOME TO LUNCH"  
(0.345755), "NOT APPLICABLE"(0.0165312)}] &  
[ractp2 = {"PAPER-GRAPHIC ART"(0.0165312), "CONSTRUCTION AND CIVIL WORKS"  
(0.0235133), "COMMERCE-HOSTELRY-REPARING-RECOVERY"(0.337456), "TRANSPORTS AND  
COMMUNICATION"(0.0317708), "BANK AND INSURANCES"(0.048266), "NON-COMMERCIAL  
SERVICES"(0.078488), "PUBLIC ADMINISTRATION- TEACHING"(0.137167), "VEHICLES AND  
TRANSPORT MATERIAL"(0.0167604), "CHEMISTRY"(0.0331782), "COMMERCIAL SERVICES"  
(0.140833), "RUBBER AND PLASTIC TRANSFORMATIONS"(0.0199348), "AGRICULTURE-  
CATTLE-FORESTRY- FISHING"(0.0201623), "METALLIC CONSTRUCTION"(0.0246369),  
"ELECTRIC MATERIAL AND MACHINERY"(0.0497522), "WOOD-FURNITURE"(0.0215493)}]
```

The JOINT Symbolic object

Then, the joint symbolic objects is:

os “Woman Married < 35 years Employed Secondary”(88) =

```
[limp= {"Null Particip."(0.347273), "Low Particip."(0.188186), "Average Particip."(0.346782), "High Particip."(0.117759)}] &  
[prpc= {"Null Particip."(0.0719004), "Low Particip."(0.400066), "Average Particip."(0.436589), "High Particip."(0.0914451)}] &  
[prac= {"Null Particip."(0.877218), "Low Particip."(0.122782)}] &  
[cuip= [0:170]] &  
[jorna= {"SPLIT SHIFT"(0.394297), "CONTINUOUS"(0.434047), "NOT APPLICABLE"(0.171656)}] &  
[comt= {"RETURN HOME TO LUNCH"(0.637714), "NOT RETURN HOME TO LUNCH"(0.345755), "NOT APPLICABLE"(0.0165312)}] &  
[ractp2= {"BANK AND INSURANCES"(0.048266), "NON-COMMERCIAL SERVICES"(0.078488), "PAPER-GRAPHIC ART"(0.0165312), "COMMERCIAL SERVICES"(0.140833), "CHEMISTRY"(0.0331782), "VEHICLES AND TRANSPORT MATERIAL"(0.0167604), "COMMERCE-HOSTELRY-REPARING-RECOVERY"(0.337456), "METALLIC CONSTRUCTION"(0.0246369), "ELECTRIC MATERIAL AND MACHINERY"(0.0497522), "TRANSPORTS AND COMMUNICATION"(0.0317708), "WOOD-FURNITURE"(0.0215493), "AGRICULTURE-CATTLE-FORESTRY-FISHING"(0.0201623), "PUBLIC ADMINISTRATION- TEACHING"(0.137167), "CONSTRUCTION AND CIVIL WORKS"(0.0235133), "RUBBER AND PLASTIC TRANSFORMATIONS"(0.0199348)}]
```

Visualization of Symbolic Objects.

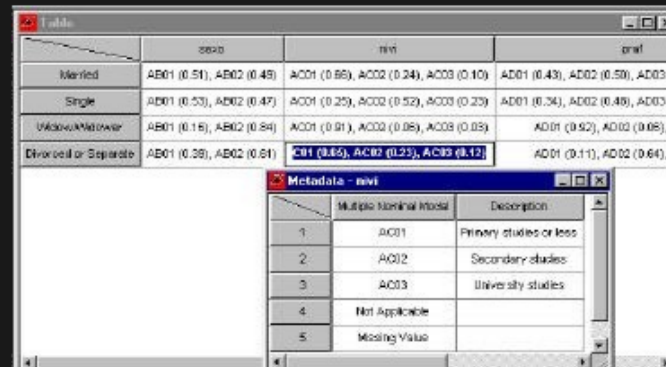
Symbolic Objects may be visualized in three different ways:

- In a symbolic table,
- By star graphs and
- By the specific language of symbolic objects, SOL (Symbolic Object Language).
- By other plots that are specific for interval or distributional descriptions.

Symbolic tables

A symbolic data table is like a classical datatable but each cell contains a multi-valued description

	sex0	pref
Married	Man (0.51), Woman (0.49)	Inactive (0.43), Employed (0.50), Unemployed (0.07), Unemployed (0.00)
Single	Man (0.53), Woman (0.47)	Inactive (0.34), Employed (0.48), Unemployed (0.12), Unemployed (0.07)
Widow/Widower	Man (0.16), Woman (0.84)	Inactive (0.92), Employed (0.06), Unemployed (0.02)
Divorced or Separate	Man (0.39), Woman (0.61)	Inactive (0.11), Employed (0.64), Unemployed (0.25)



The screenshot shows a software window titled 'Table' containing the symbolic data table from the previous block. Below the table, there is a 'Metadata - nvi' window. The 'Table' window has a header row with 'sex0' and 'pref' columns. The 'Metadata - nvi' window has a header row with 'Multiple Nominal Interval' and 'Description' columns. The 'Metadata - nvi' window contains the following data:

	Multiple Nominal Interval	Description
1	AC01	Primary studies or less
2	AC02	Secondary studies
3	AC03	University studies
4	Not Applicable	
5	Missing Value	

Zoom star

A Zoom star is a radial plot where each radius represents a variable. There are two types of zoom star visualization, 2D and 3D, which provide different levels of detail. The 2D representation provides a global impression of the symbolic object, whereas 3D representation provides much more detailed information.

The Zoom Star representation is derived from Kiviat Diagrams where each axis corresponds to a variable. In the same graph we can represent categorical variables, intervals, weighted values, taxonomies,... without overloading the graph.

The following table summarizes the representation of each variable depending on its type.

Variable Type	Axis Description
Quantitative	Graduated axis
Categorical	Dots equally distributed on the axis
Categorical not weighted	Axis drawn in black
Categorical weighted	Axis drawn in claret
Not applicable	Axis drawn in grey

The limit for variables to be represented is 24 and for categories is 15.

Selecting an axis with the mouse, we can display the distribution of the chosen variable (histogram). Moreover, we can also display taxonomies and dependencies of a variable by clicking on the icon that appears in the corresponding axis.

Graphics can be moved right, left, up and down for a better visualization.

2D Zoom stars

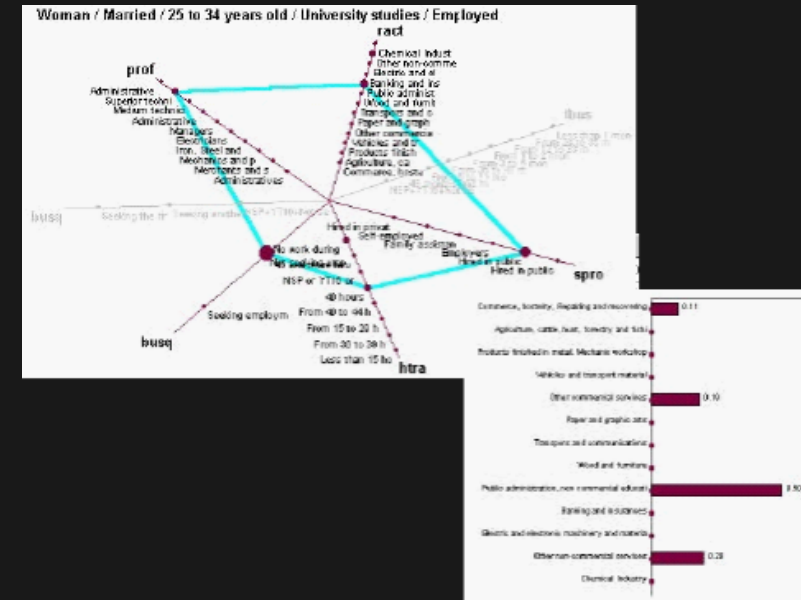
In the 2D Zoom Star, axes are linked by a line that connects most frequent values of each variable. If there were a tie of the most frequent value in several categories, the line would link all of them.

In the presence of an interval variable, the line is linked to the minimum and maximum limits and the entire area is filled.

For instance, we have defined symbolic objects as groups of population defined by sex, age, marital status, level of education and relation to activity in the P.R.A. survey. We have obtained 314 symbolic objects, which are the combination of the modalities of these variables.

In this case, as we use a survey, the distribution has been calculated taking into account sampling weights.

In the following graph, we can see two mother-daughter variables. Daughter variables that are N.A. appear in the graph as a grey axis. On the right, we can see the distribution of one of the variables.

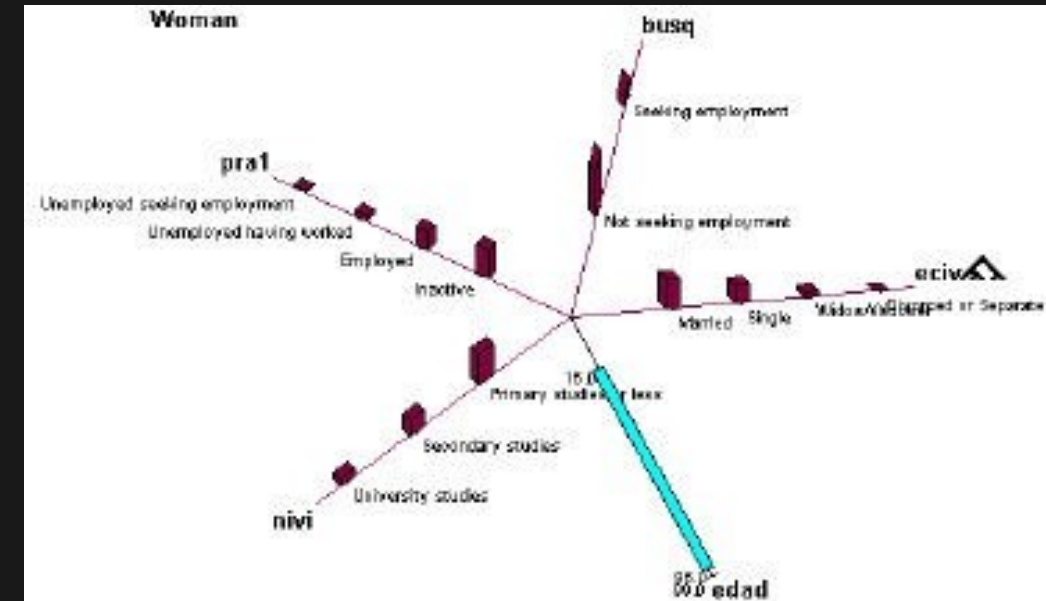


2d Zoom star

3D Zoom stars

In the 3D representation, we can see distributions corresponding to each variable with weighted values. Numerical variables are represented by rectangles from the minimum to the maximum value.

For example, the distribution of the symbolic object “Woman” in the P.R.A survey corresponding to a quarter in Alava is the following,



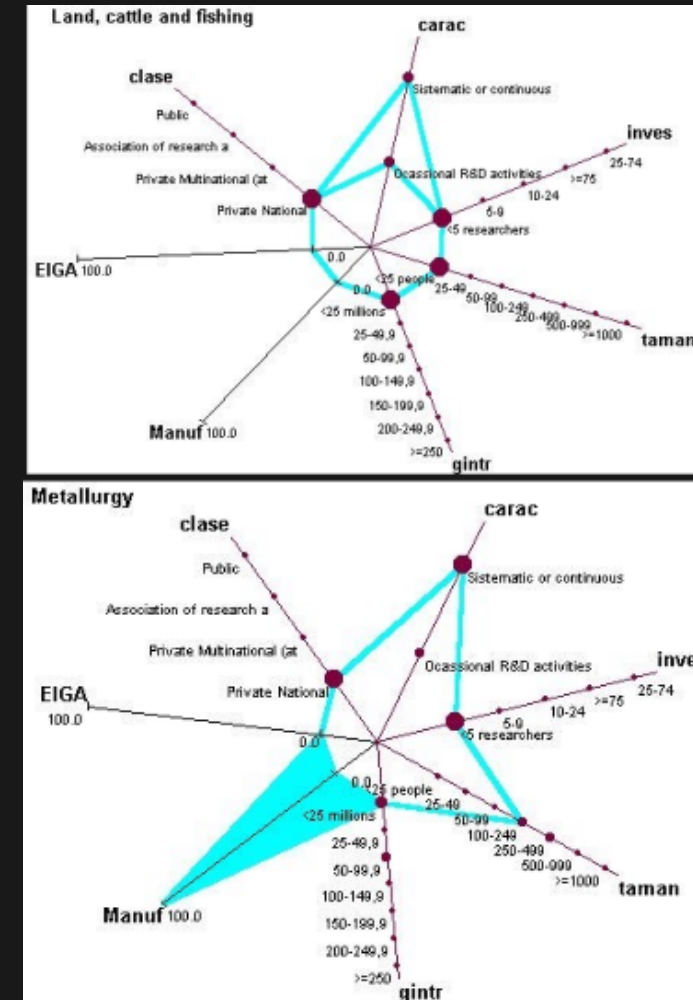
2d Zoom star

Comparison of several Symbolic Objects

The comparison of several symbolic objects is easier using the 2D representation. We compare if the shapes of the lines that link the axes are similar.

Example

From the survey of Enterprises doing R&D in the Basque Country, we have built some symbolic objects describing branches of economic activity. From the 18 available branches, we have chosen 2 to compare them, “Land, Cattle and Fishing” and “Metallurgy”. The chosen variables to describe both branches are: Type of Enterprise, Type of R&D activity, number of researchers in the activity, size of the enterprise in staff, intramural expenses, percentage of researching dedicated to manufacturing products and energy.



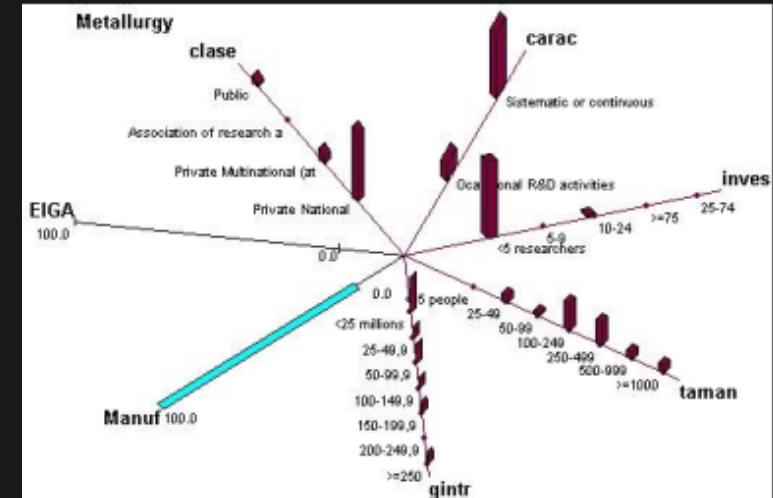
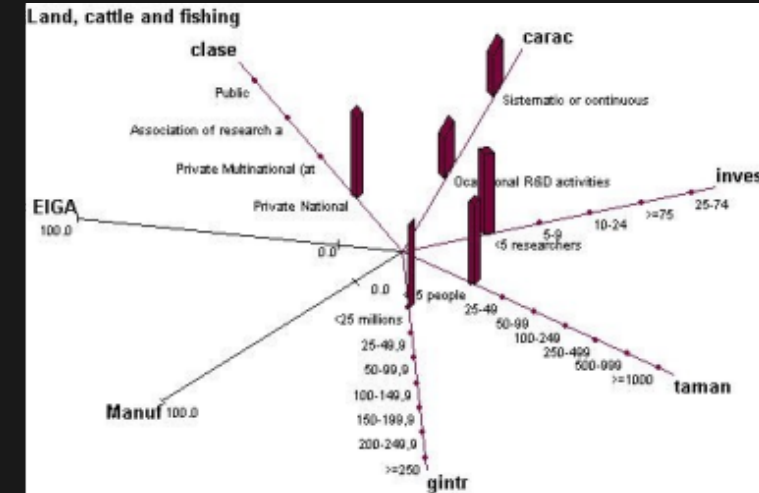
From the graphs we can draw the following conclusions:

The two branches differ in the character of the R&D activities, in “Land, Cattle and Fishing” the activities can be both systematic or occasional, whereas in “Metallurgy” the activities are mostly systematic. Another difference is the size of the enterprises, in the metallurgic industry they are larger than in the “Land, Cattle and Fishing” branch. Moreover, the metallurgic industry uses 100% of intramural expenses for researching manufacturing products.

The comparison with histograms (3D representation) also provides relevant information about the distributions.

We represent the same 3D graphs of branches of economic activity as in the previous example, to obtain more information.

Now, we can observe better the differences between distributions in the two branches. In “Metallurgy”, the distributions of the variables “size of the enterprise” (taman) and “intramural expenses” (gintr) are much more dispersed among all categories, whereas in “Land, Cattle and Fishing” the distributions of these two variables are centred in a unique value.



Basic statistics of Symbolic Objects

Basic Statistics of Symbolic Objects consists of a set of graphs and summary measures depending on the type of variable.

If the variables are multinomial, we can draw frequency graphs such as bar graphs and pie charts.

If the variables are interval, we can draw frequency graphs with central tendency and dispersion measures. Moreover, we can represent biplots.

Finally, if the variables are probabilistic, we can draw graphs of capacities.