DSSR 2024

Data Science & Social Research

4th international conference

Naples (Italy) on **25th-27th March 2024**

Visualizing Multidimensional Distributional data: some new tools

Once we were satisfied

Antonio Irpino, *Ph.D.*

Dept. of Mathematics and Physics
University of Campania L. Vanvitelli
Caserta, Italy

Tuesday, the 26th of March, 2024

DSSR 2024
Data Science & Social Research
4th international conference
Naples (Italy) on 25th-27th March 2024

Motivations

(From “The grammar of graphics”, L. Wilkinsons)

Mathematics provides symbolic tools for representing abstractions.

Aesthetics, in the original Greek sense, offers principles for relating sensory attributes (color, shape, sound, etc.) to abstractions.

Data visualization and statistics

Statistics is the science of variability. Statistics aims at

- describing,
- measuring and
- modeling

variability.

Data visualization, but, in particular, **Statistical graphics** is fundamental in describing variability at the different steps of a statistical analysis,

- from the data **preprocessing** step,
- passing through the **summarization** of raw and processed data
- up to the representation of model **validation**.

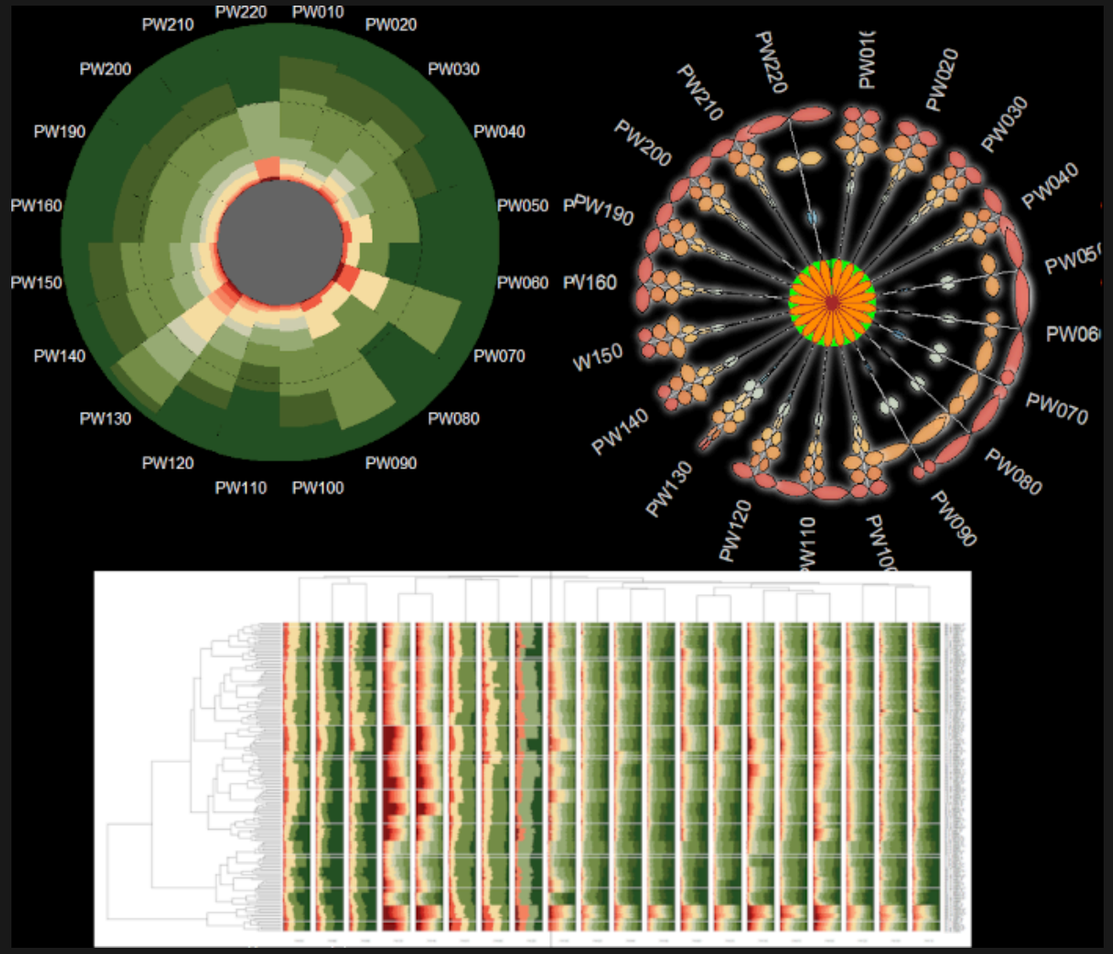
> **For the analysis of aggregate or symbolic or distributional data or macrodata, there is a lack of visualization tools because of the complexity of the information there contained.**

Aims

We propose new visualization tools for Macro data described as distributional data.

Starting from a data table where each macro unit can be described by several distributional variables, we propose

- How to visualize a single macrodata or compare few of them, through *eye iris* or *flowers* plots;
- How to visualizing large distributional data tables, extending the classic *heatmap* to the **distributional heatmap**.



An application to EUSILC data 2013 is presented for showing how to visually identify patterns in a distributional data table.

Macro data and distributional data: numeric distributional data

Let's see an example: `BLOOD` dataset from the `HistDAWass` R package.

It is a classical (in the Symbolic Data Analysis community) dataset describing

- 14 typologies of patients;
- 3 distributional variables;

after aggregating a set of raw data from a hospital. See: Billard and Diday (2006)

	Cholesterol		Hemoglobin		Hematocrit	
name	V1 bins	p1	V2 bins	p2	V3 bins	p3
u1: F-20	[80 ; 100]	0.025	[12 ; 12.9]	0.050	[35 ; 37.5]	0.025
	[100 ; 120]	0.075	[12.9 ; 13.2]	0.112	[37.5 ; 39]	0.075
	[120 ; 135]	0.175	[13.2 ; 13.5]	0.212	[39 ; 40.5]	0.188
	[135 ; 150]	0.250	[13.5 ; 13.8]	0.201	[40.5 ; 42]	0.387
	[150 ; 165]	0.200	[13.8 ; 14.1]	0.188	[42 ; 45.5]	0.287
	[165 ; 180]	0.162	[14.1 ; 14.4]	0.137	[45.5 ; 47]	0.038
	[180 ; 200]	0.088	[14.4 ; 14.7]	0.075		
	[200 ; 240]	0.025	[14.7 ; 15]	0.025		
u2: F-30	[80 ; 100]	0.013	[10.5 ; 11]	0.007	[31 ; 33]	0.046
	[100 ; 120]	0.088	[11 ; 11.3]	0.039	[33 ; 35]	0.171
	[120 ; 135]	0.154	[11.3 ; 11.6]	0.082	[35 ; 36.5]	0.295
	[135 ; 150]	0.253	[11.6 ; 11.9]	0.174	[36.5 ; 38]	0.243
	[150 ; 165]	0.210	[11.9 ; 12.2]	0.216	[38 ; 39.5]	0.170
	[165 ; 180]	0.177	[12.2 ; 12.5]	0.266	[39.5 ; 41]	0.072
	[180 ; 195]	0.066	[12.5 ; 12.8]	0.157	[41 ; 44]	0.003
	[195 ; 210]	0.026	[12.8 ; 14]	0.059		
	[210 ; 240]	0.013				
u14: M-80+	[155 ; 170]	0.067	[10.8 ; 11.2]	0.133	[33.5 ; 35.5]	0.133
	[170 ; 185]	0.133	[11.2 ; 11.6]	0.067	[35.5 ; 37.5]	0.267
	[185 ; 200]	0.200	[11.6 ; 12]	0.134	[37.5 ; 39.5]	0.267
	[200 ; 215]	0.267	[12 ; 12.4]	0.333	[39.5 ; 41.5]	0.133
	[215 ; 230]	0.200	[12.4 ; 12.8]	0.200	[41.5 ; 43]	0.200
	[230 ; 245]	0.067	[12.8 ; 13.2]	0.133		
	[245 ; 260]	0.066				

The first two and the last typology of patient in the BLOOD dataset.

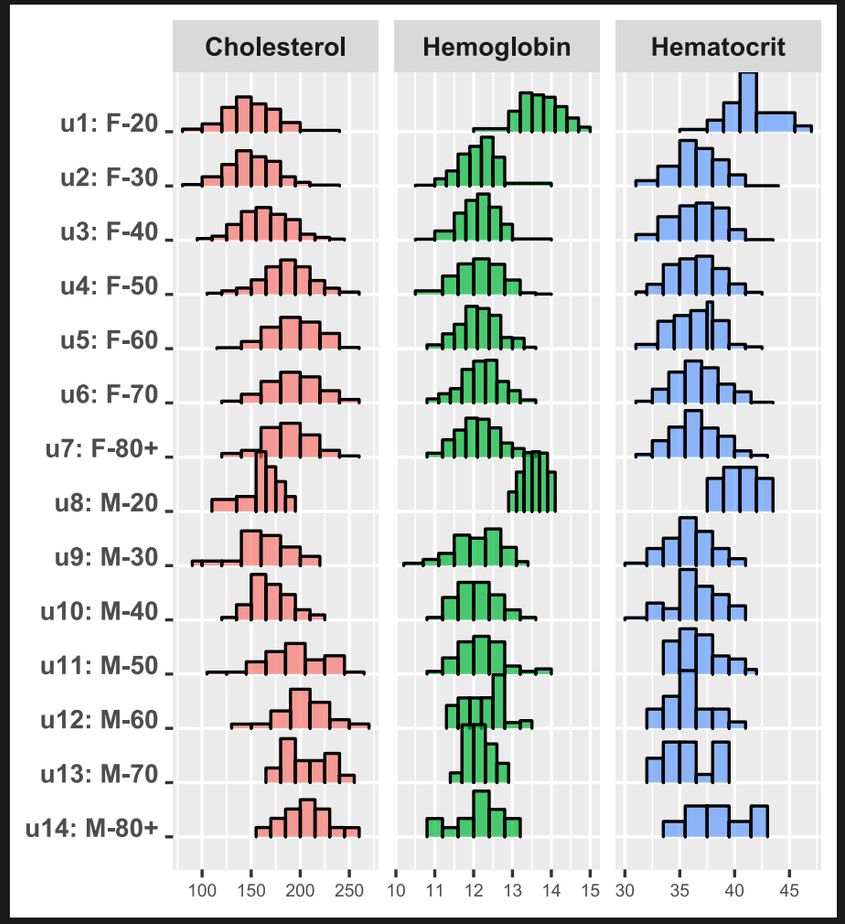
Numerical distributional dataset

A distributional dataset is a classical table with N observations on the rows and P variables, indexing the columns, such that the generic term y_{ij} is a **numerical univariate distribution**

$$y_{ij} \sim f_{ij}(x_j)$$

where $x_j \in D_j \subset \Re$ and $f_{ij}(x_j) \geq 0$,

- $\int_{x_j \subset D_j} f_{ij}(x_j)dx_j = 1$, if the distribution has a continuous support;
- $\sum_{x_j \in D_j} f_{ij}(x_j) = 1$, if the distribution has a discrete support.



The BLOOD dataset

The basic plot for the i-th observation

The i-th observation is the vector $y_i = [y_{i1}, \ldots, y_{ij}, \ldots, y_{iP}]$

Steps :

1. **Domain discretization**
   - **For continuous variables**. For each variable Y_j we consider the domain D_j and, fixing a K_j integer value we partition D_j into K_j equi-width intervals (bins) of values, such that:

$$D_j = \left\{ B_{jk} = (a_k, b_k], | \, b_k > a_k, \; k = 1, \ldots, K_j \, , \bigcup_{k=1}^{K} B_{jk} = [\min(D_j), \max(D_j)], B_{jk} \cap B_{jk'} = \emptyset, \text{ for } k \neq k' \right\}$$

   - **For discete variables**. For each variable Y_j we consider the domain D_j and, being $\#D_j = K_j$ the cardinality of D_j, we consider the elements of D_j.
2. **Choice of a divergent colour palette**_ We consider a divergent color palette with K_j levels, such that K_1 represent the lowest category and K_j the highest one.
3. **Stacked percentage barcharts** We compute the mass observed in each bin/category for each y_{ij}
   For the Y_i observation, P bars are generated. The order of the bar can be decided accordingly to the user preferences, or can be suggested by a correlation analysis for all the data in advance (one may cluster the distributional variables using a hierarchical clustering based on the Wasserstein correlation and then using the order returned by after the aggregation).
4. **Polar coordinates** Polar coordinates allow to represent the stacked barcharts as circles that mimics an Eye Iris.
   We called this plot **Eye Iris** plot (EI plot.)

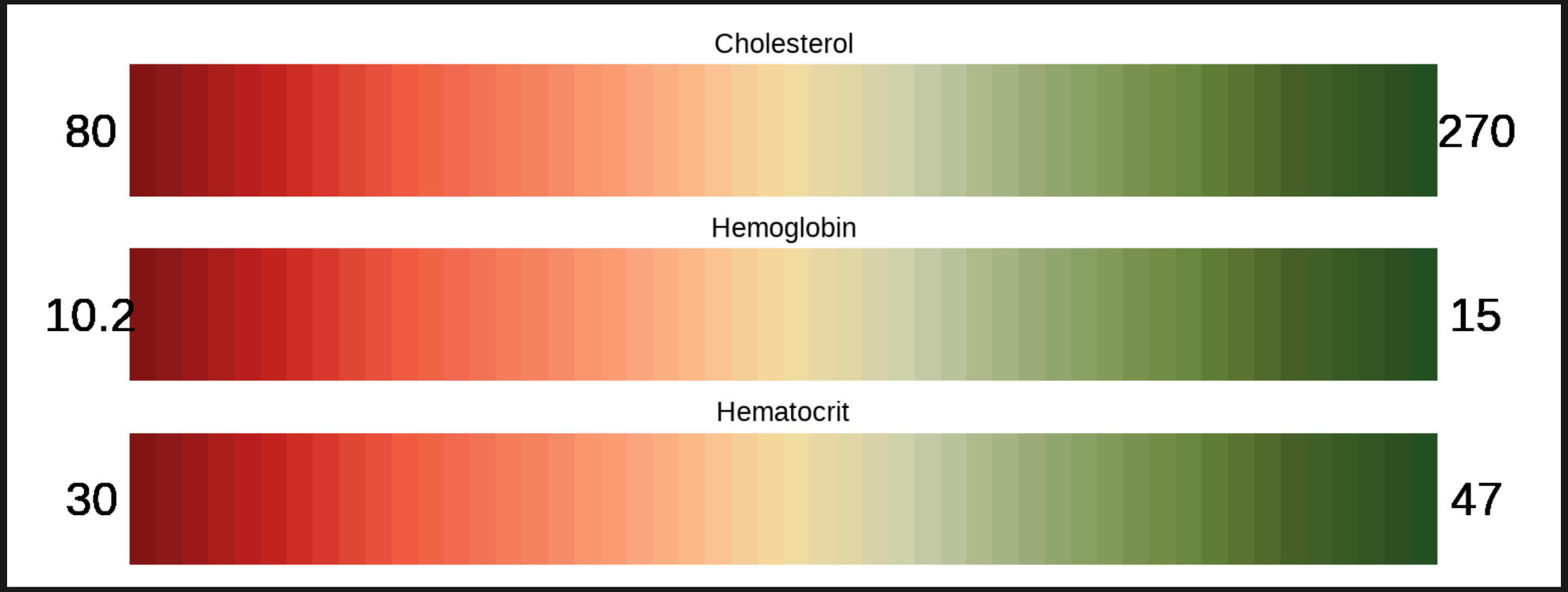
Example using BLOOD data

The extremes of the domains of the variables

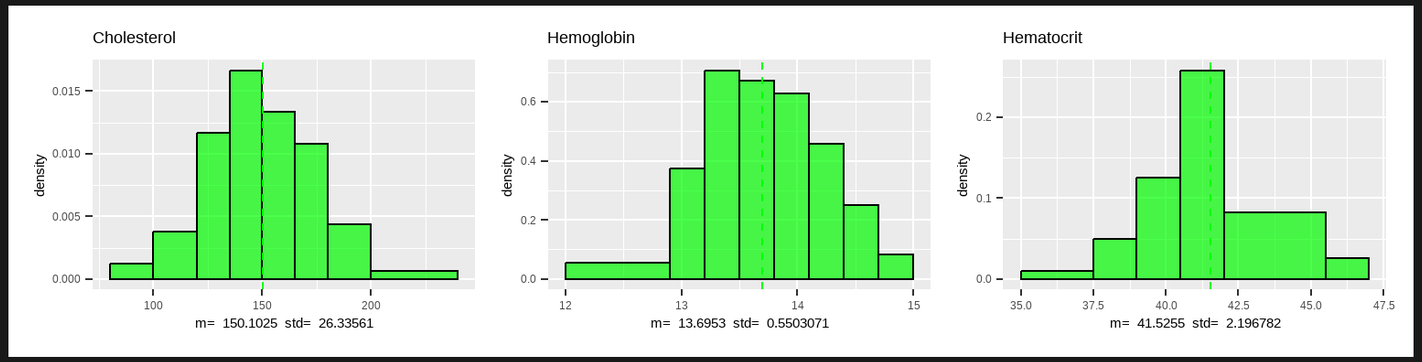
- Range of Cholesterol [80 ; 270]
- Range of Hemoglobin [10.2 ; 15]
- Range of Hematocrit [30 ; 47]

Choice of K and of a color palette

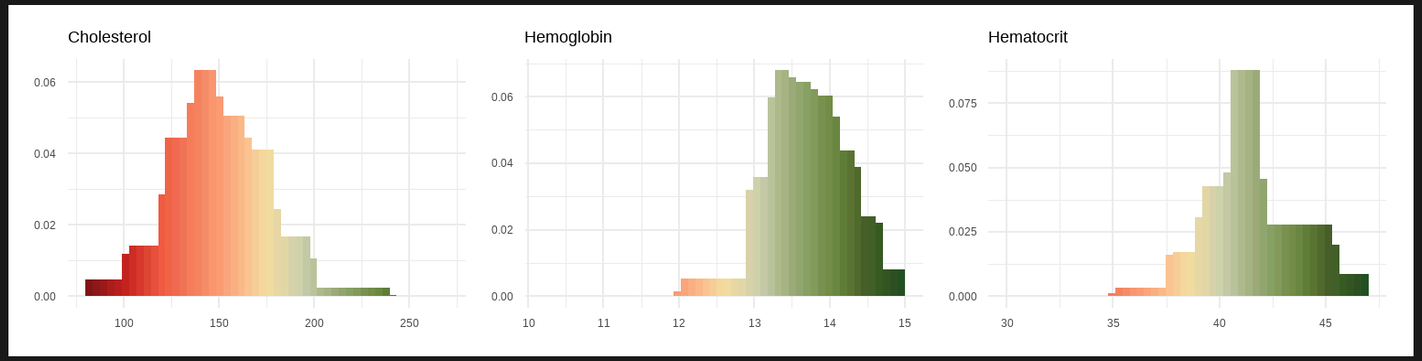
We fix $K = 50$ and we will use a color palette from Red (low values), passing through Yellow (middle values) to Green (high values).



Now, let’s take the first observation

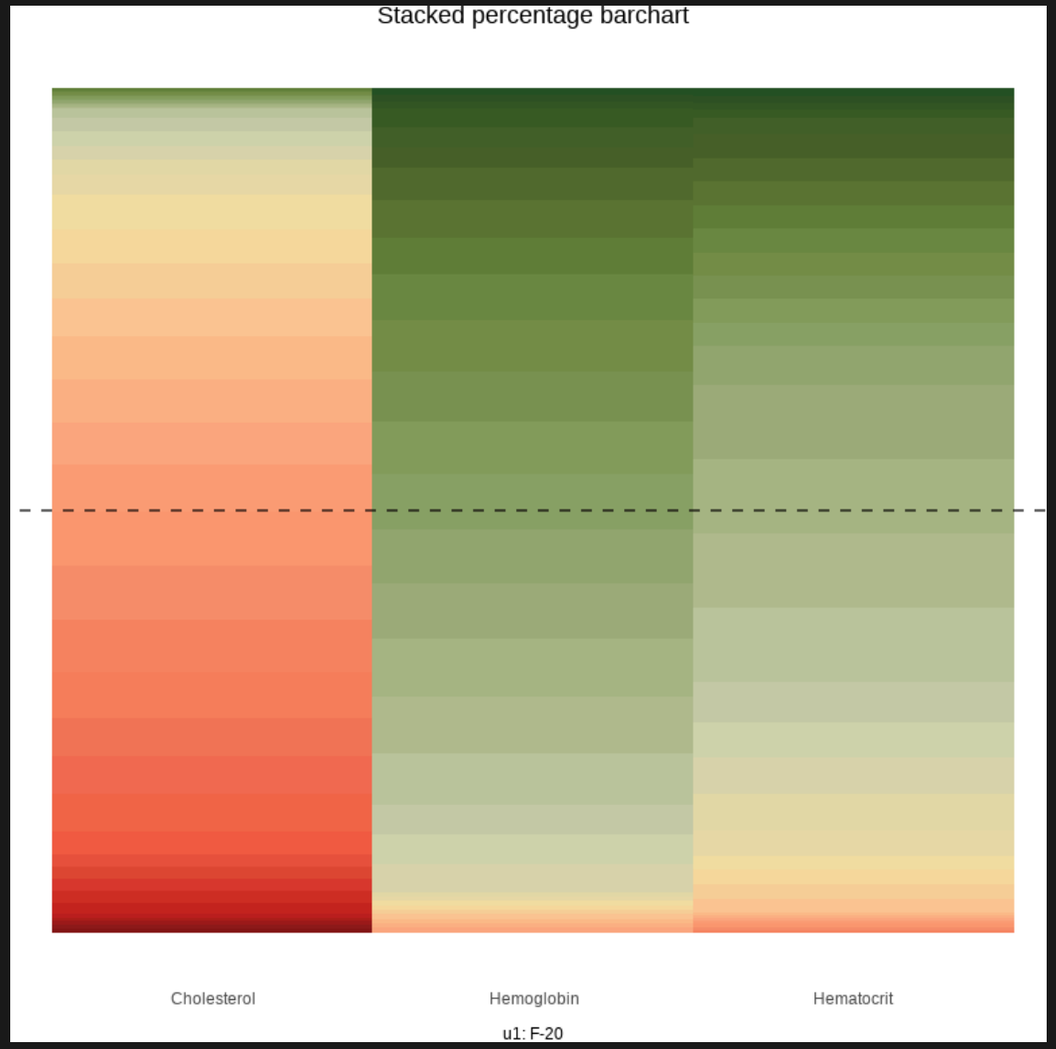


Recode the distribution according to $K = 50$ partition of the domains.



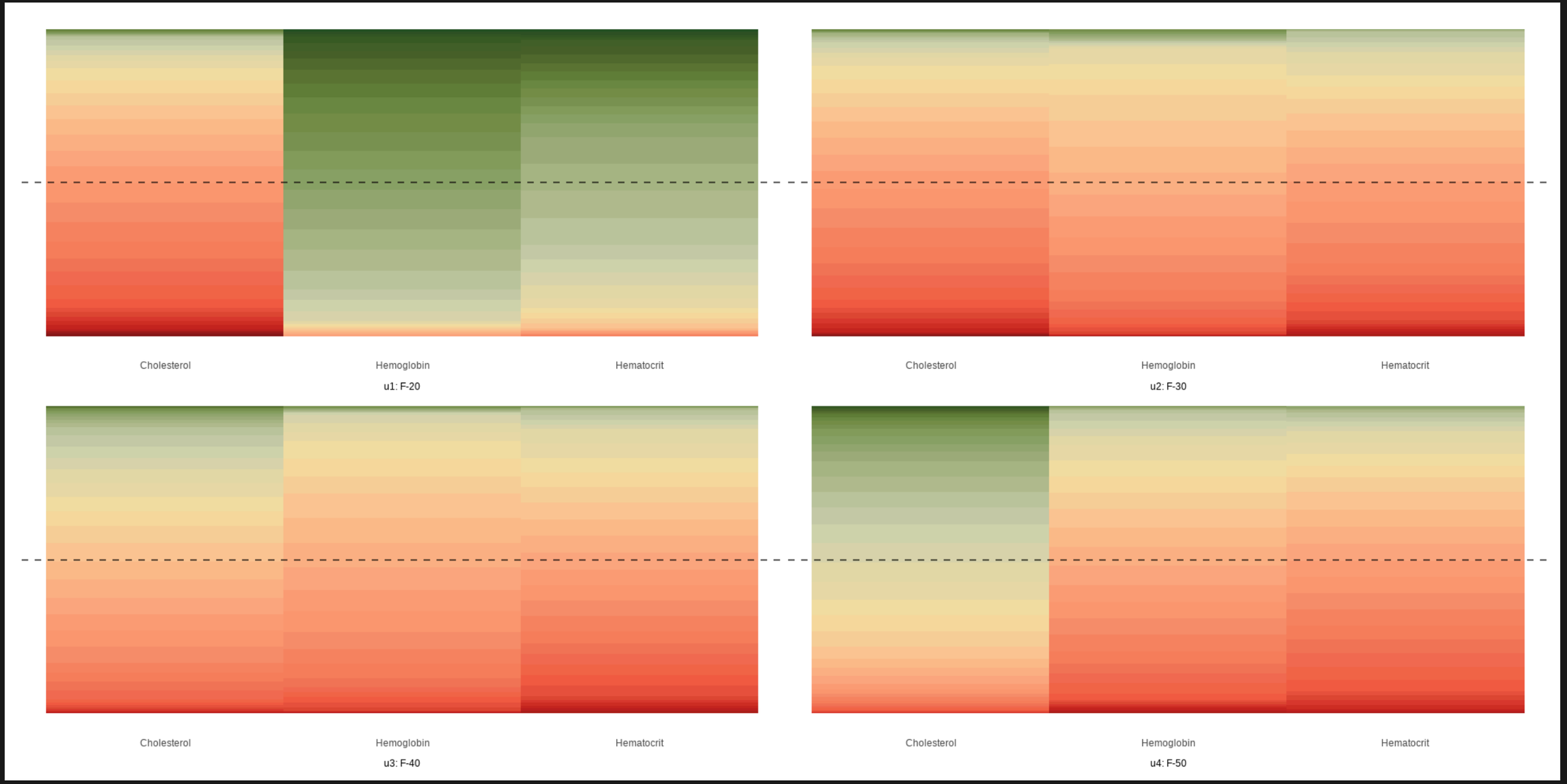
Since the bins represent classes of values, we can consider them as ranked *levels* of the domain.

We propose to see all the three distributions using a stacked percentage barchart as follows. Note that each level of color has a area that is proportional to the mass associated with each bin.



The dashed line is positioned at level 0.5 suggesting where the median of each distribution is positioned taking into consideration the level of color associated with the bin of the respective domain.

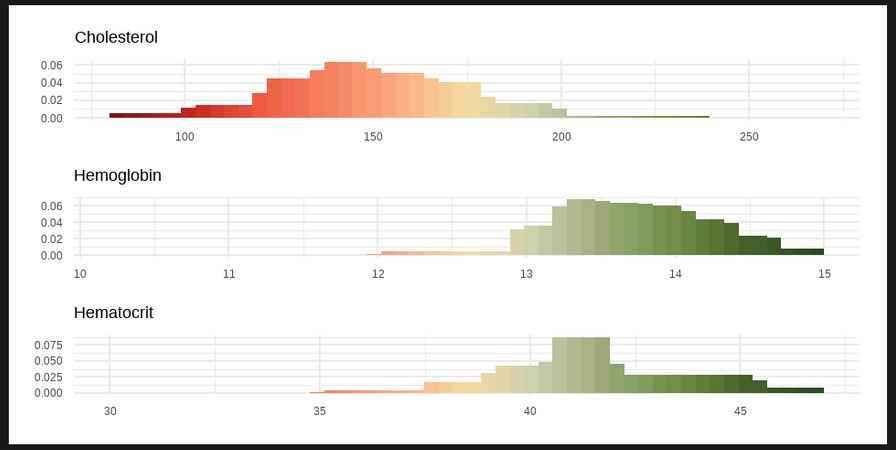
But, this kind of visualization is not so immediate for comparing several observations. Let’s see an example:

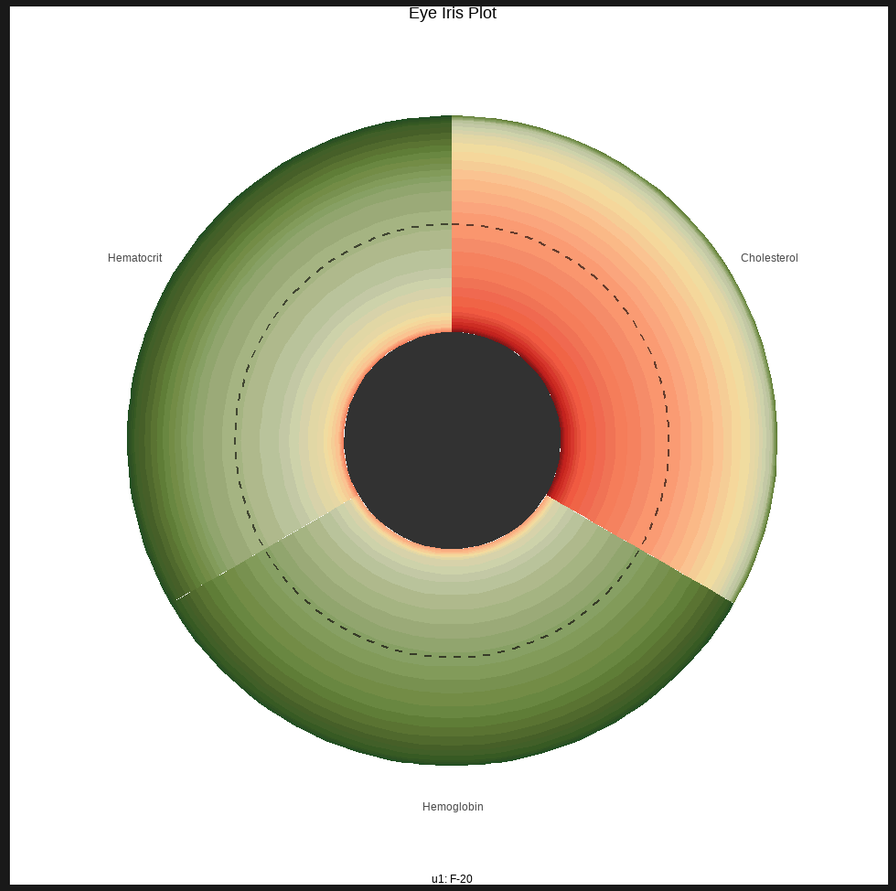


We propose to use a plot based on polar coordinates. We add a *pupil* for reducing the distortion due to the polar transformation

Eye iris plot grammar

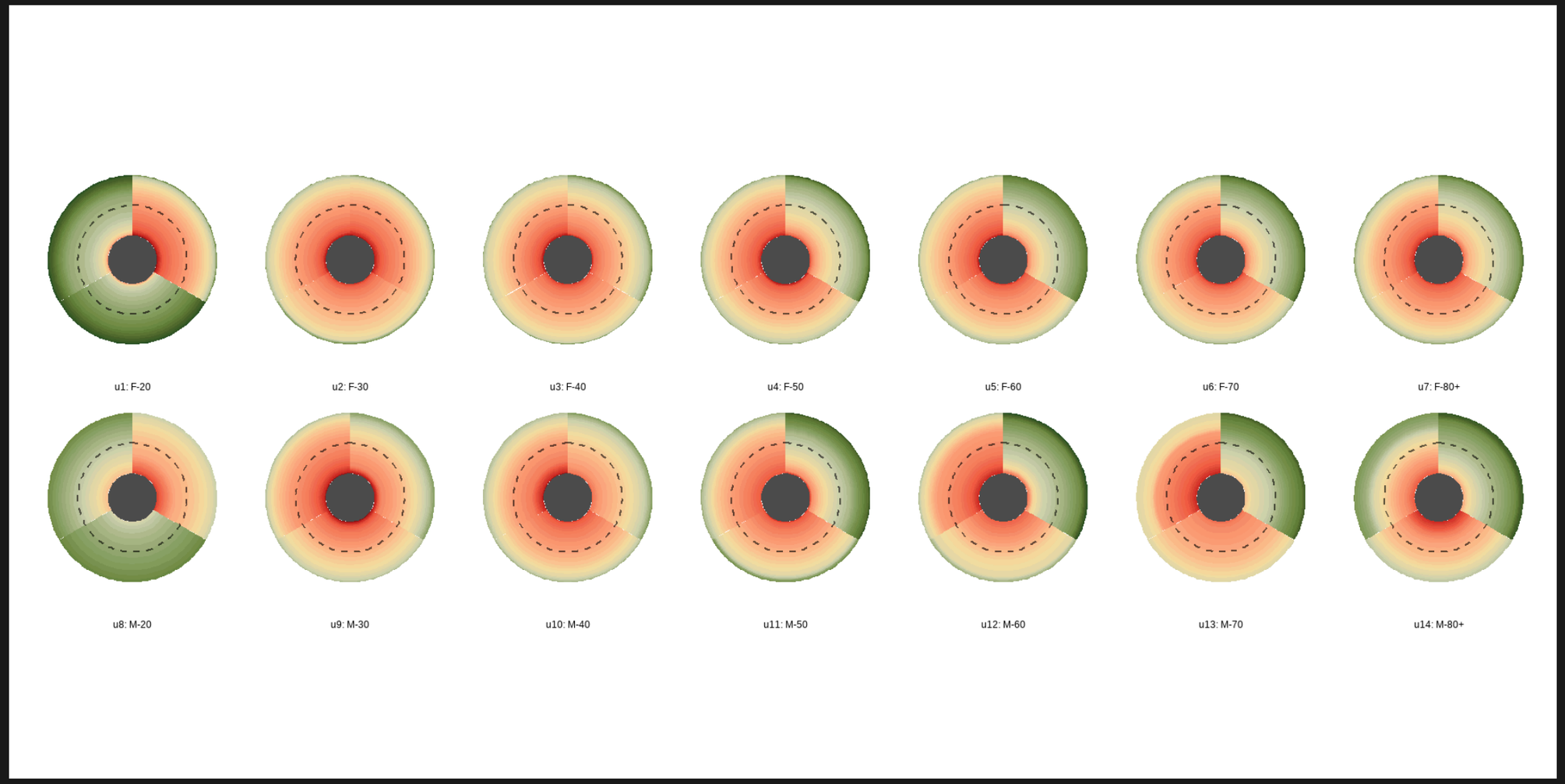
1. Each sector represents a distribution
2. Abundance of color
   - towards the red = the distribution is concentrated on low values of the domain
   - towards the green = the distribution is concentrated on high values of the domain





3. the dotted circle represents the level 0.5 of the CDFs, then the color close to it represents the median value of the distribution.
4. The more the colors of a sector are uniformly distributed the more the distribution is flat.

Since a human is able to catch eyes shapes and color, we believe that this kind of visualization can be more interpretable. For example, let's see all the 14 observations together.



Interpretation

According to the filling colours we can compare both observations and distributional values.

The Enriched plot

We propose to add information about dispersion and skewness.

The dispersion

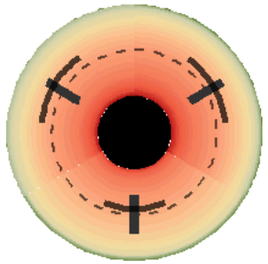
Each variable in the dataset may have a different dispersion. Each distributional variable has its dispersion accounted by its proper standard deviation σ_{ij}. We normalize each standard deviation σ_{ij} by the maximum standard deviation of observed for the the j-th variable $\max(\sigma_{ij})$ where $i = 1, \ldots, N$. A segment, centered in the respective sector, allow to perceive the dispersion associated with each distribution.

The skewness

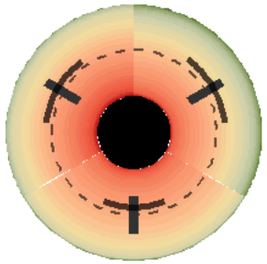
Each y_{ij} has its skewness value computed via the **Third standardized moment** γ_{ij}.

We represent the skewness of y_{ij} external to the dashed circle if it is positive, while it is positioned internally if it is negative. The distance from the dashed circle represent the absolute value of the skewness index. If the segment is very close to the dashed circle, it means that the distribution is almost symmetric.

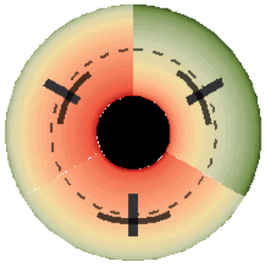
u1: F-20



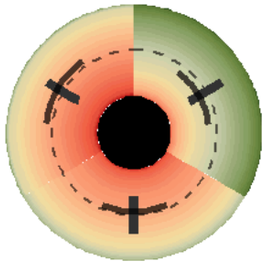
u2: F-30



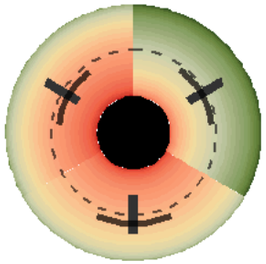
u3: F-40



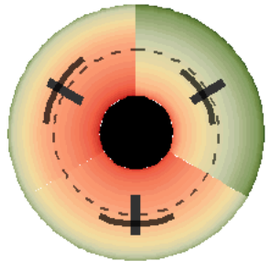
u4: F-50



u5: F-60

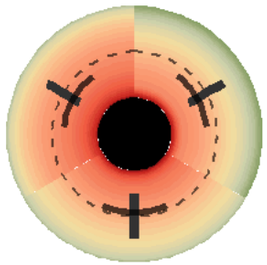


u6: F-70

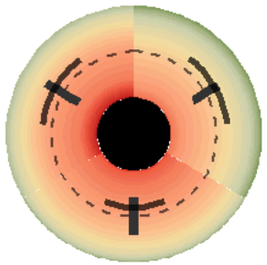


u7: F-80+

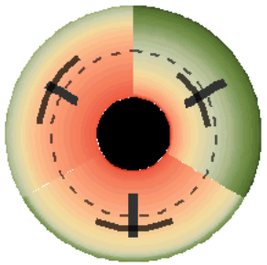
u8: M-20



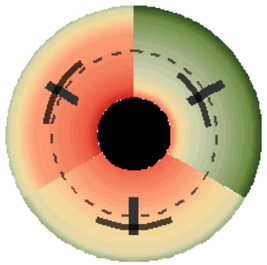
u9: M-30



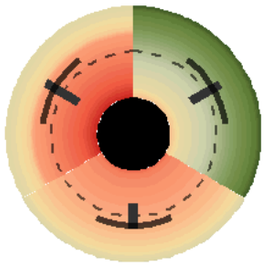
u10: M-40



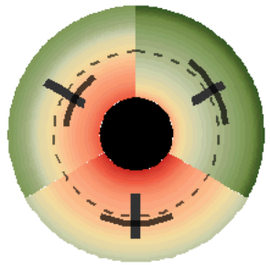
u11: M-50



u12: M-60



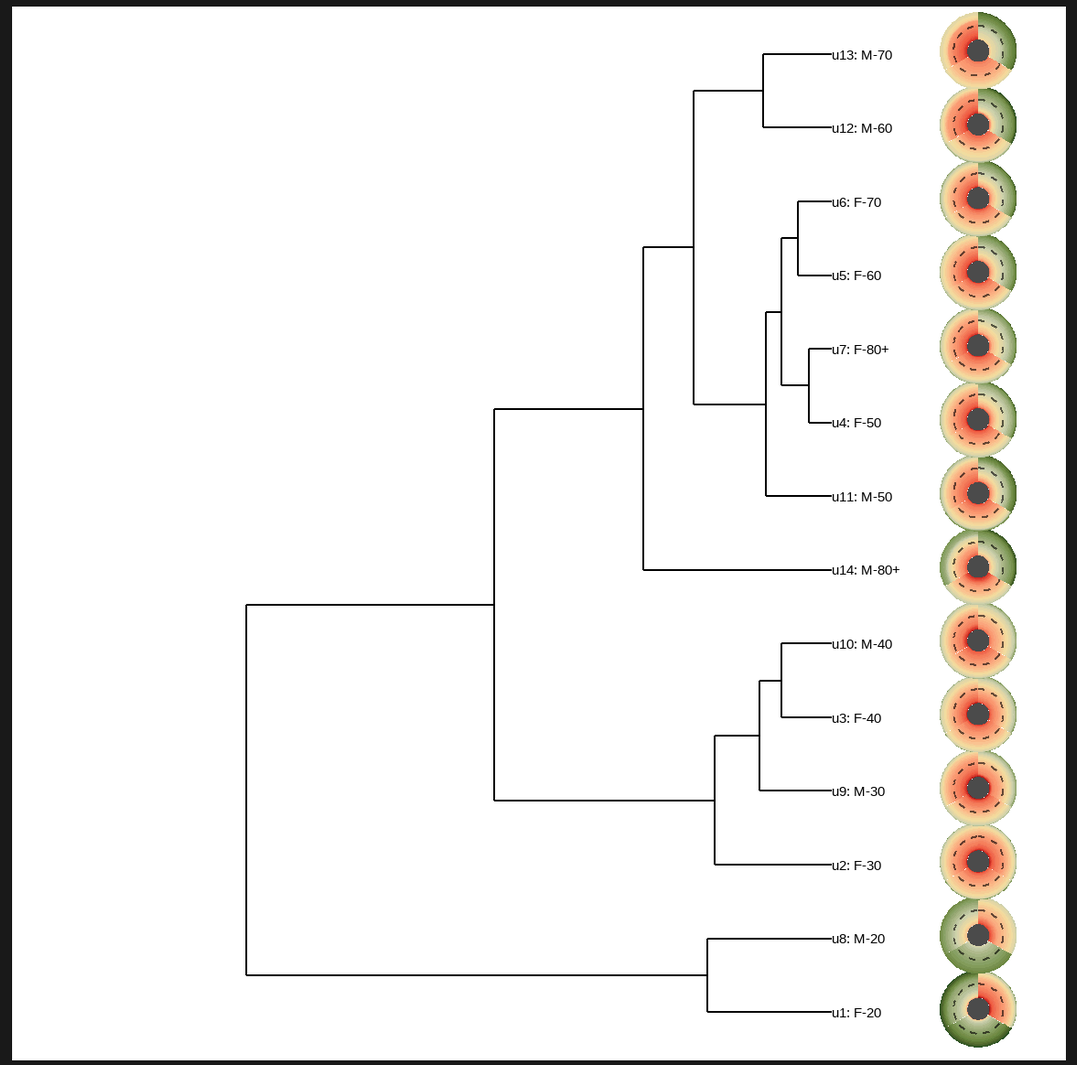
u13: M-70



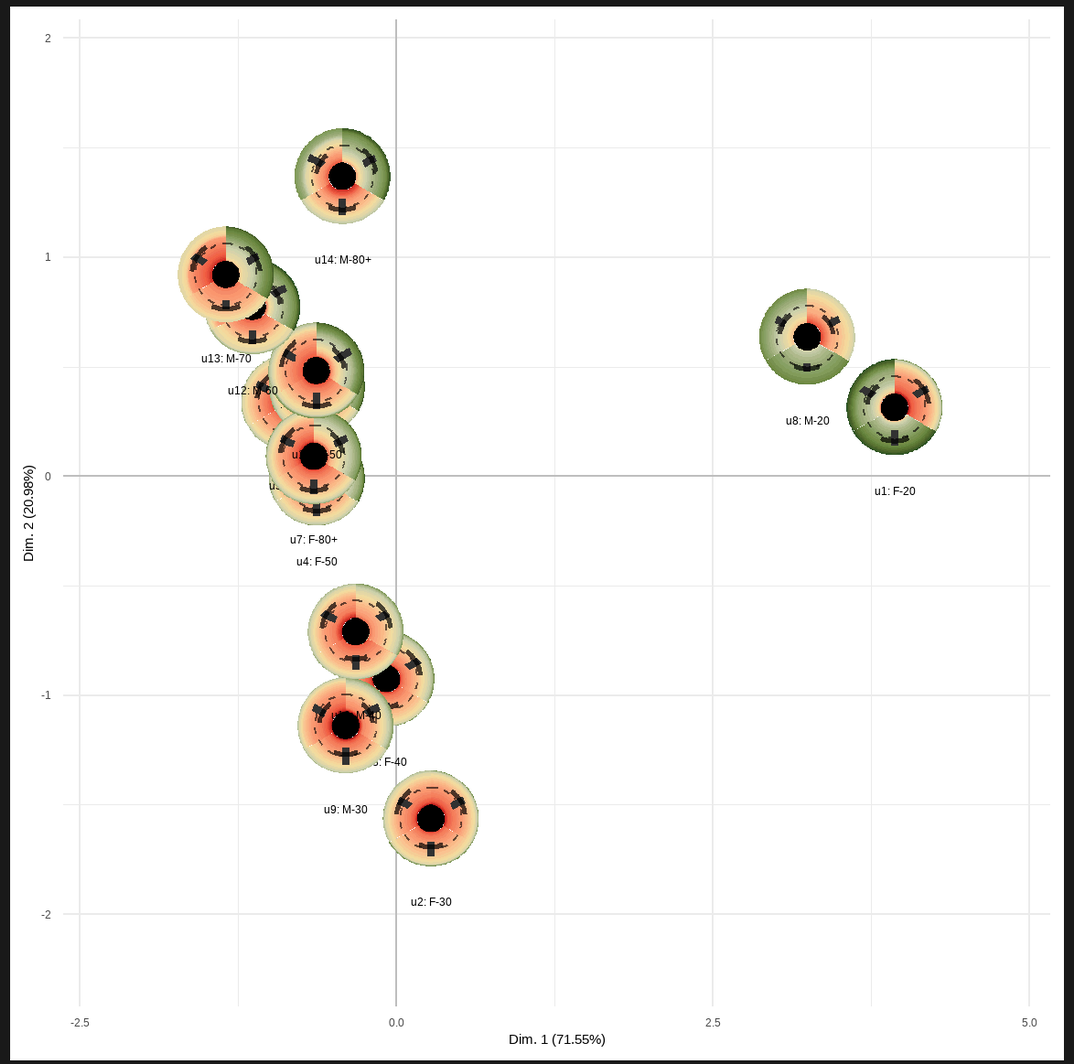
u14: M-80+

An example applied to

Hierarchical clustering



Principal Component Analysis



Visualizing all the data table

A distributional heatmap

The EI plots can be useful for medium-sized data tables: less than 20 units or variables.

In any case, in this slides we present an alternative represantation of the BLOOD dataset by adapting the classical heatmap plot to distributional data.

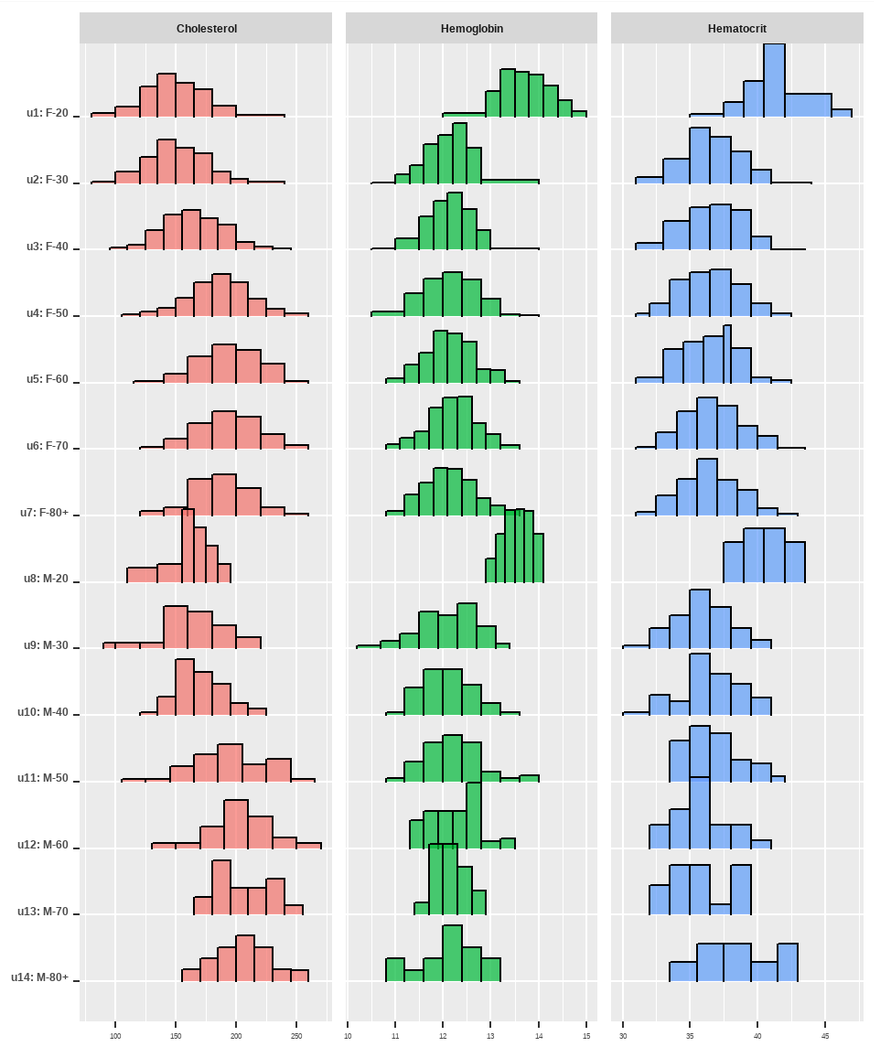
> In distributional heatmaps each tile contains a visualization of a row-column distribution.

The novelty

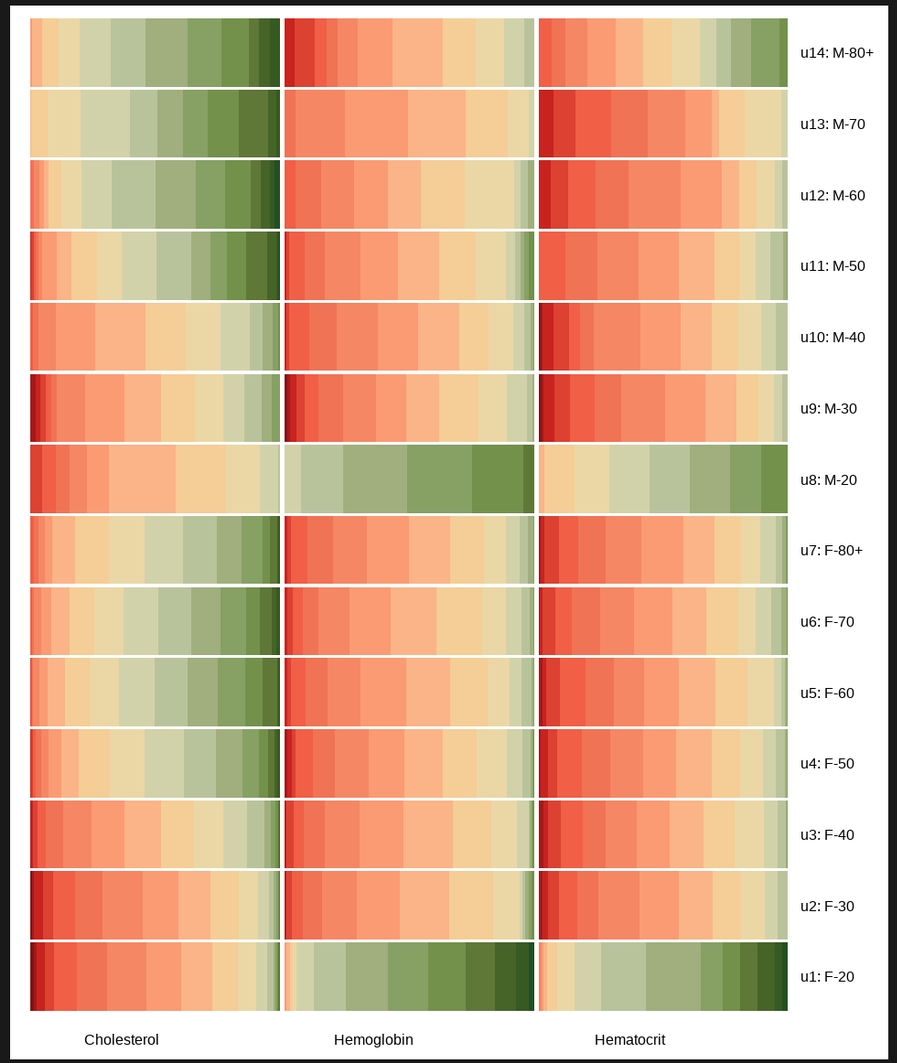
We substitute a single tile using horizontal stacked percentage bars that are constructed after binning each variable's domain and using a diverging color palette, like for the EI plot.

A comparison between

Classical visualization

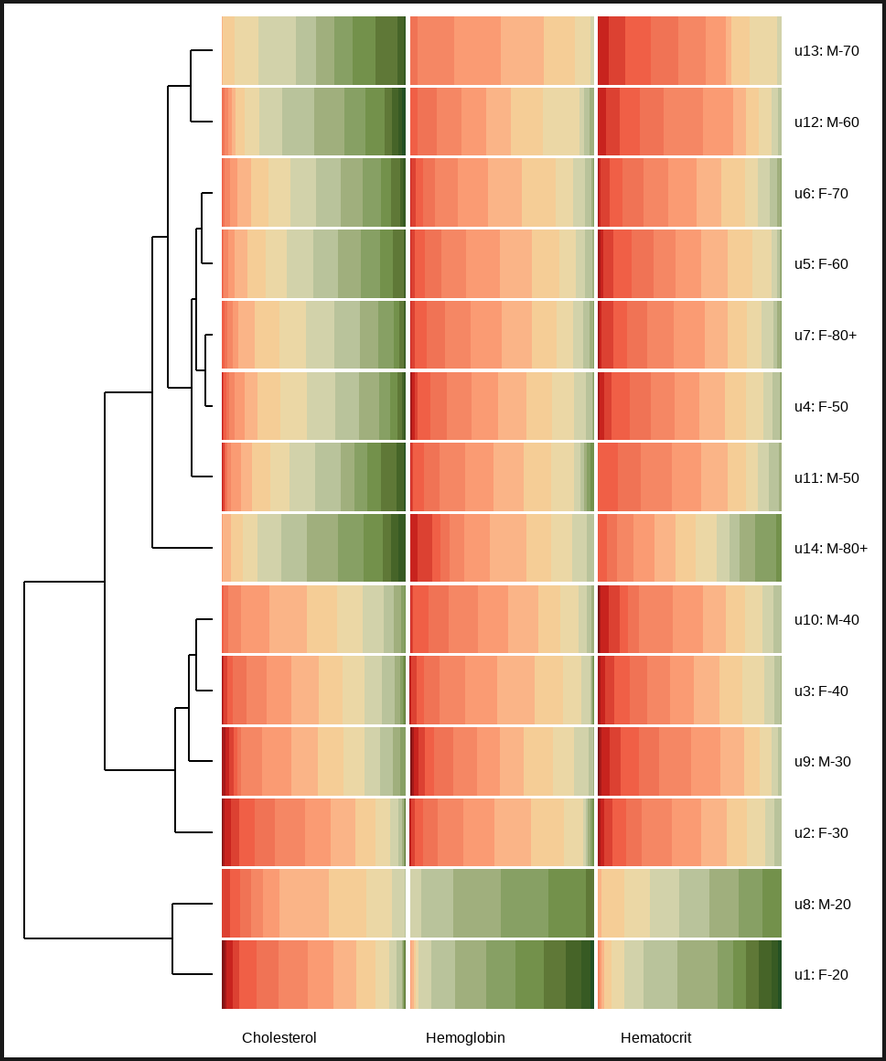


Distributional heatmap (new!!)

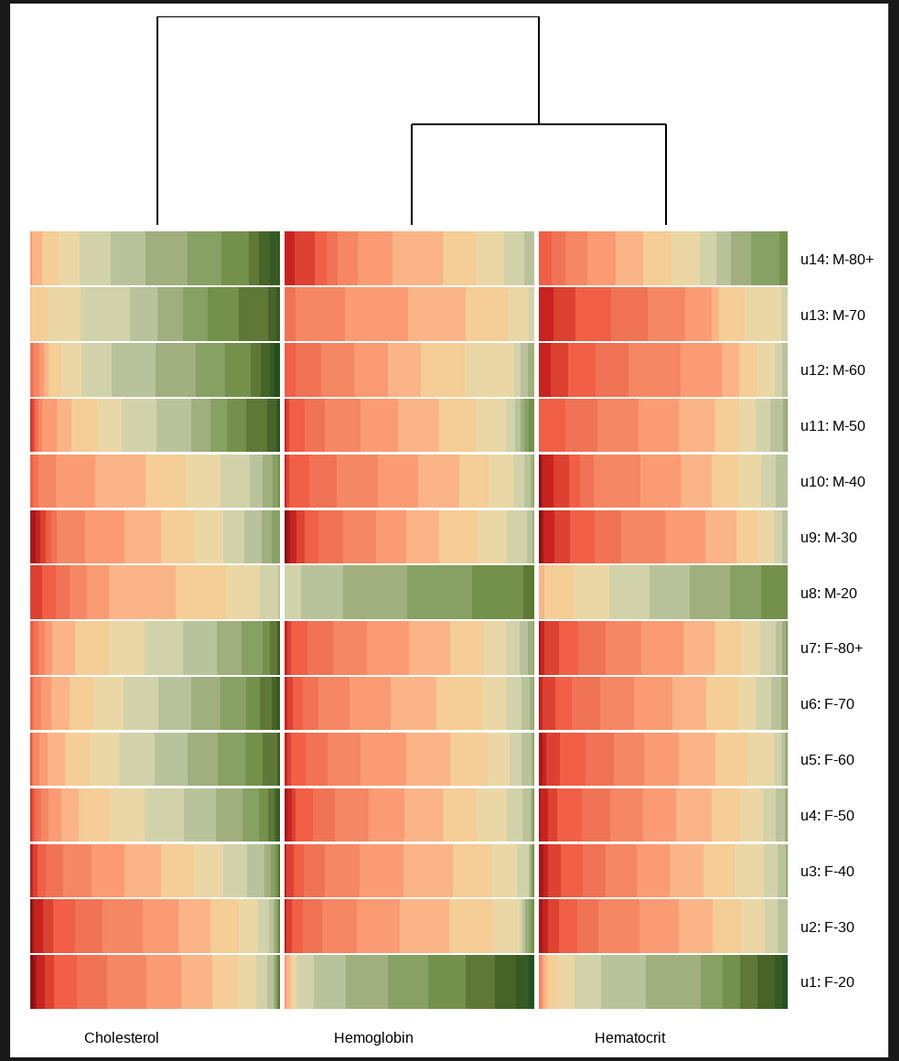


Advantages of heatmaps

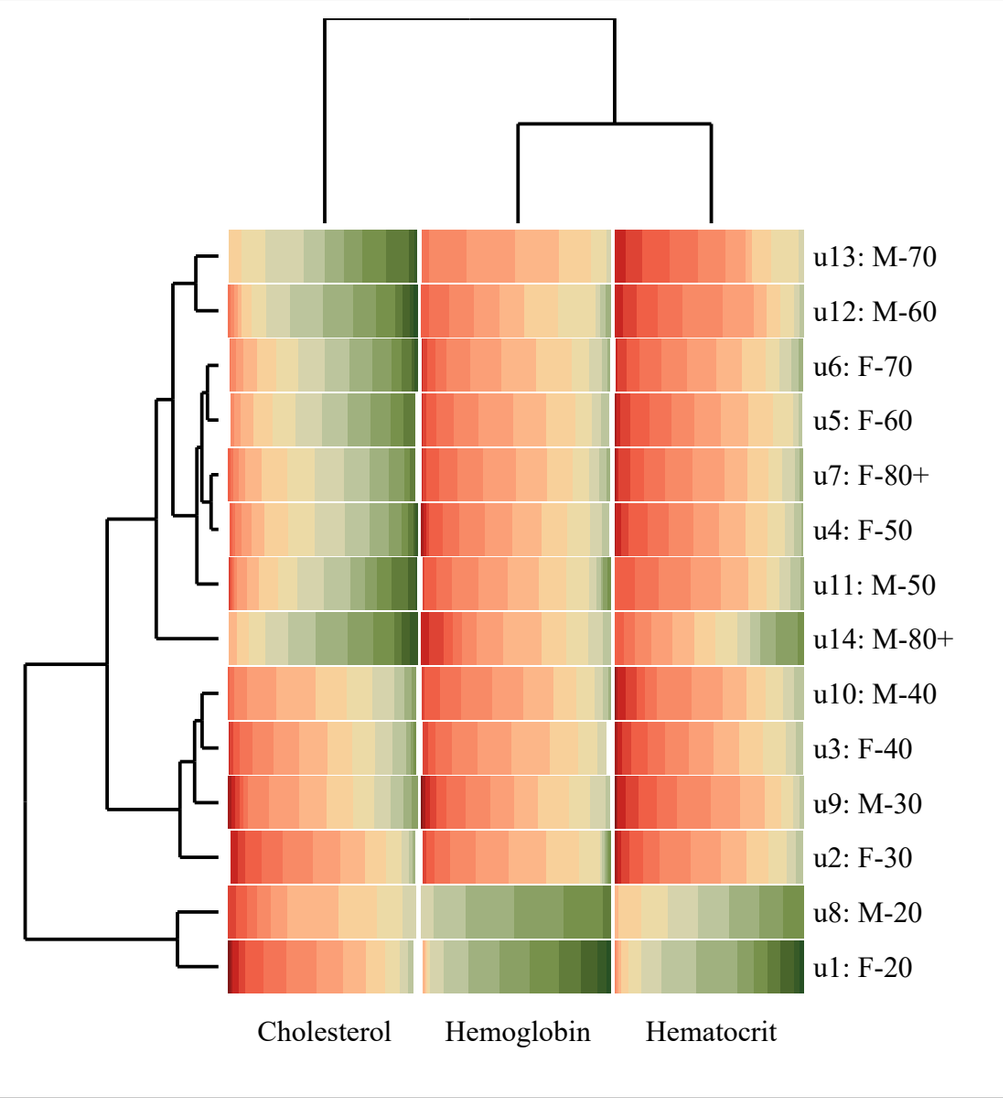
Clustering rows

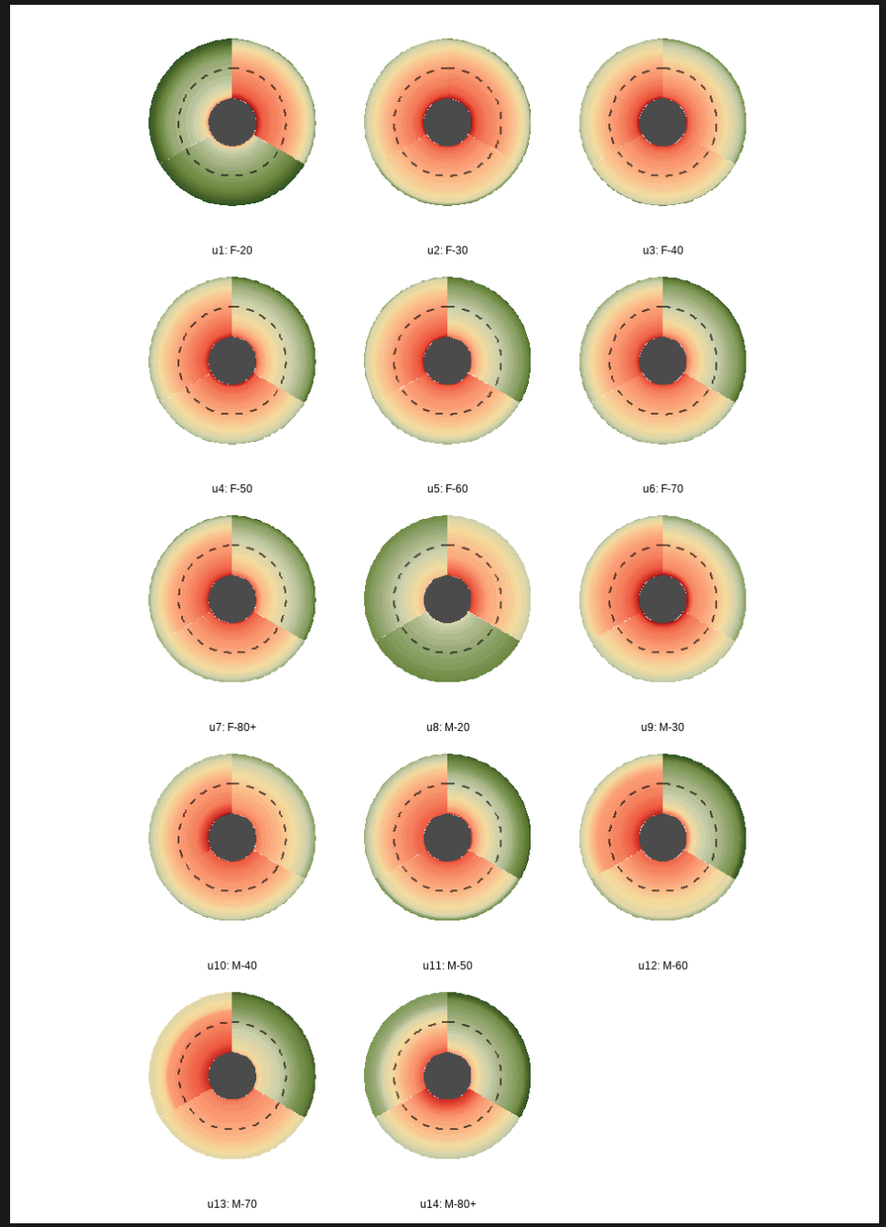


Clustering columns



Advantages of heatmaps, clustering both





An application to EUSILC 2013 survey

Data come from EUSILC survey (EU Statistics on Income and Living Conditions)

We downloaded the raw public available microdata from https://ec.europa.eu/eurostat/web/microdata/public-microdata/statistics-on-income-and-living-conditions

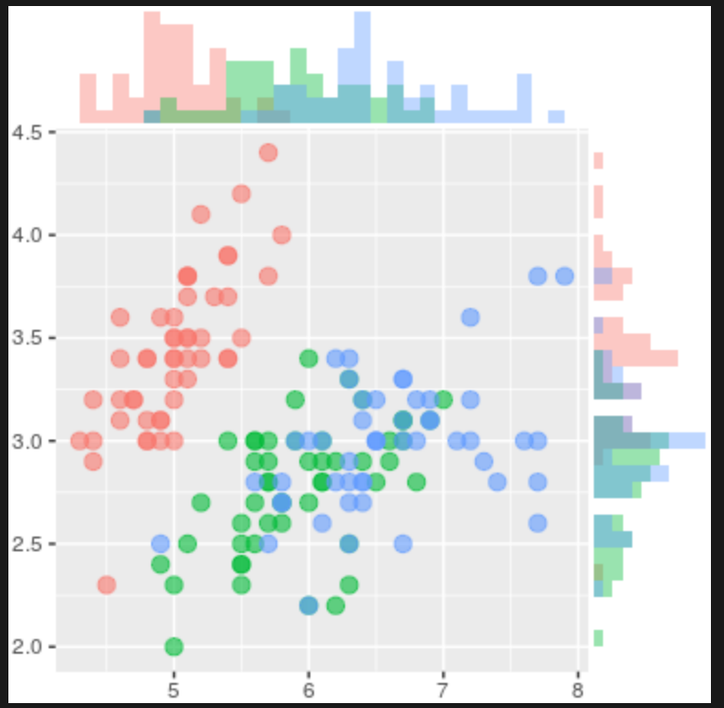
The website contains the EUSILC data for 2013 about 24 European countries.

- number of cases **390,658** individuals
- number of variables **285**

We grouped the cases according to the

- **Country** (24 countries),
- **Gender** (Male and Female) and
- **Age class** (4 classes)
  - 15-24 y.o. (EARLY working age),
  - 25-54 years (PRIME working age),
  - 55-64 years (MATURE working age),
  - 65 years and over (ELDER)

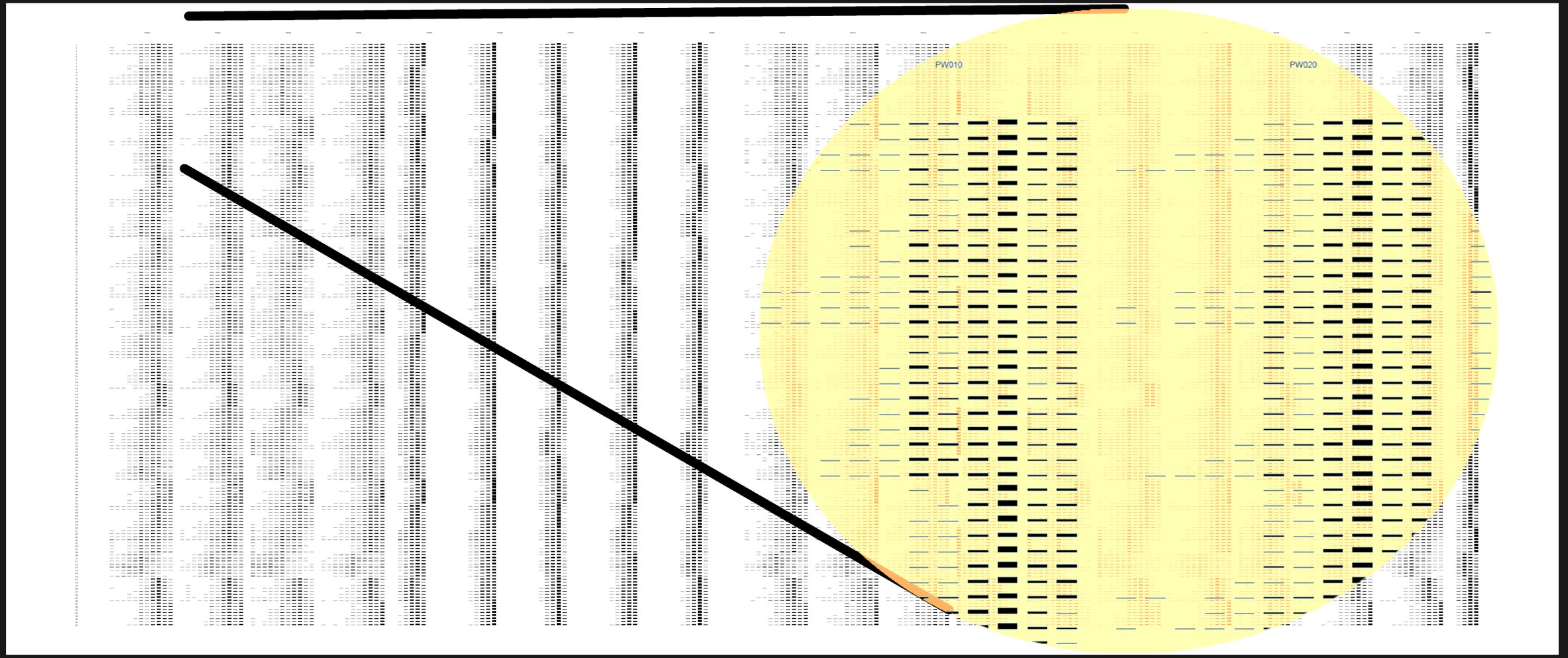
obtaining **192** groups (Macrounits). Each group is described by **20** distributional variables obtained from (Likert-type) items related to “satisfaction”.



Dictionary of variables

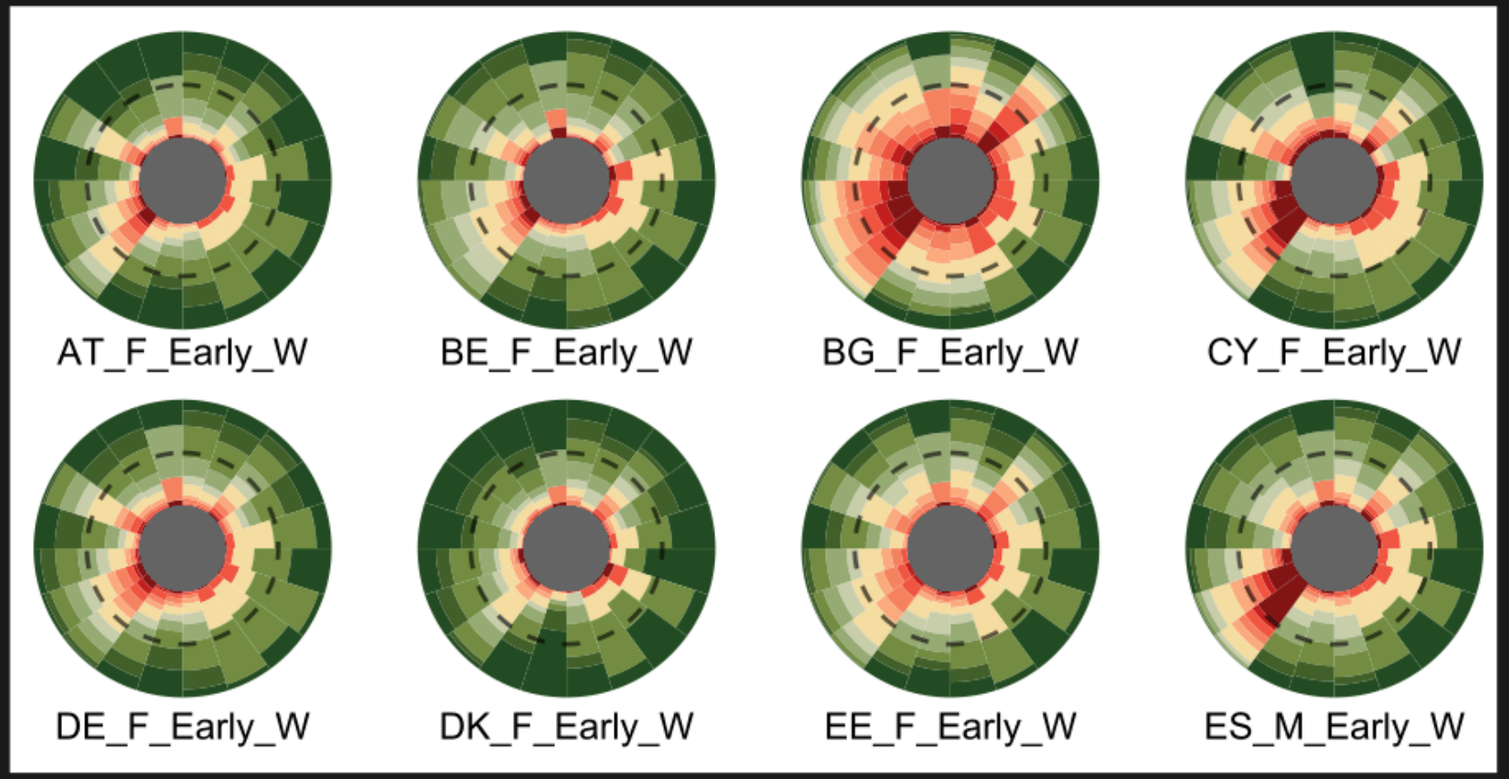
- PW010: OVERALL LIFE SATISFACTION (0-10)
- PW020: MEANING OF LIFE (0-10)
- PW030: SATISFACTION WITH FINANCIAL SITUATION (0-10)
- PW040: SATISFACTION WITH ACCOMMODATION (0-10)
- PW050: BEING VERY NERVOUS (1-5 rev)
- PW060: FEELING DOWN IN THE DUMPS (1-5)
- PW070: FEELING CALM AND PEACEFUL (1-5 rev)
- PW080: FEELING DOWNHEARTED OR DEPRESSED (1-5)
- PW090: BEING HAPPY (1-5)
- PW100: JOB SATISFACTION (0-10)
- PW110: SATISFACTION WITH COMMUTING TIME (0-10)
- PW120: SATISFACTION WITH TIME USE (0-10)
- PW130: TRUST IN THE POLITICAL SYSTEM (0-10)
- PW140: TRUST IN THE LEGAL SYSTEM (0-10)
- PW150: TRUST IN THE POLICE (0-10)
- PW160: SATISFACTION WITH PERSONAL RELATIONSHIPS (0-10)
- PW190: TRUST IN OTHERS (0-10)
- PW200: SATISFACTION WITH RECREATIONAL OR GREEN AREAS (0-10)
- PW210: SATISFACTION WITH LIVING ENVIRONMENT (0-10)
- PW220: PHYSICAL SECURITY (0-10)

The classical (useless) visualization (each cell is a barchart)



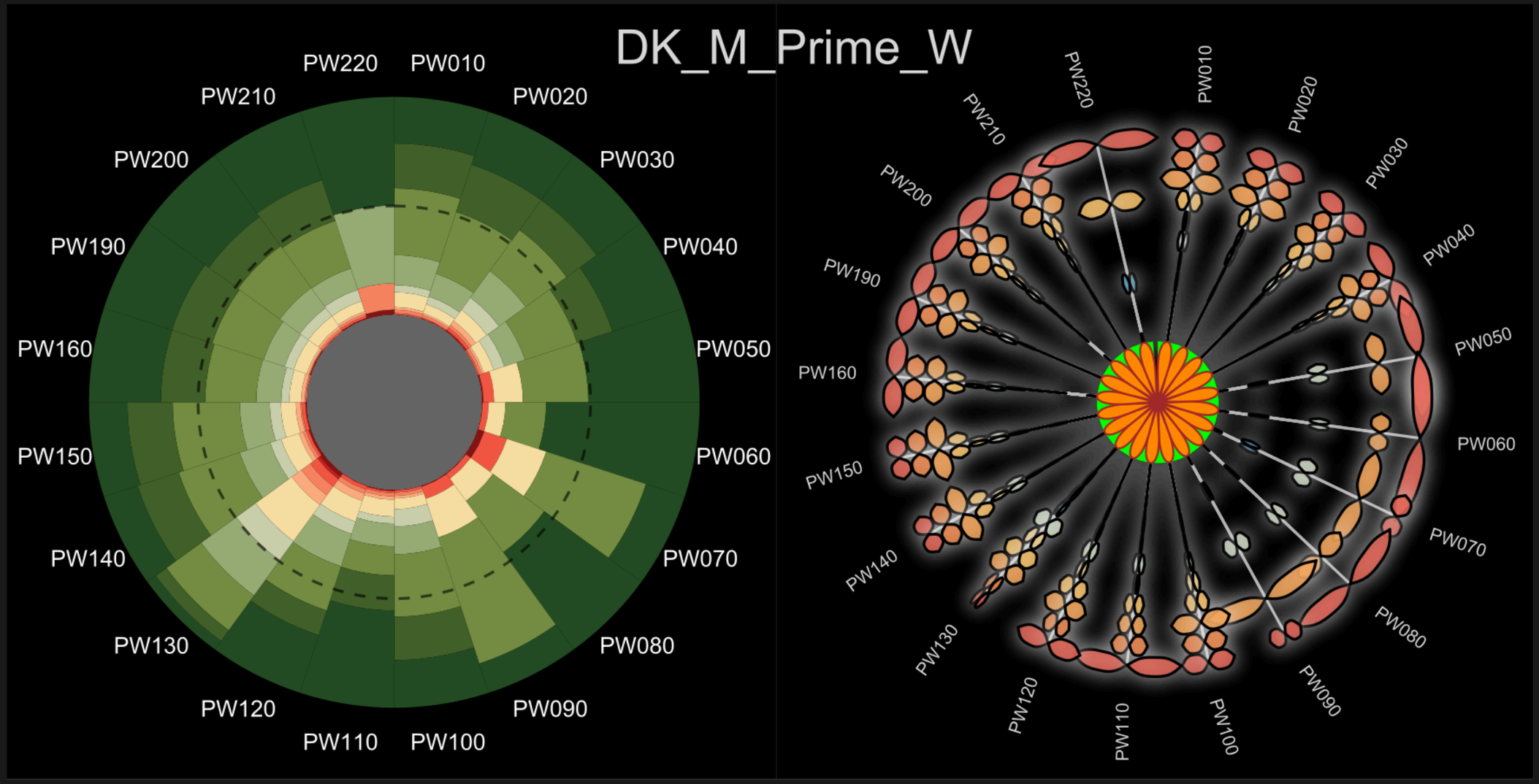
Comparing units: the Eye iris plot

Let’s observe eight eyes:



A new plot the Flower plot

In this plot each main direction is a variable and each petal is proportional to the relative frequency, the filling color is related to the value of the domain.

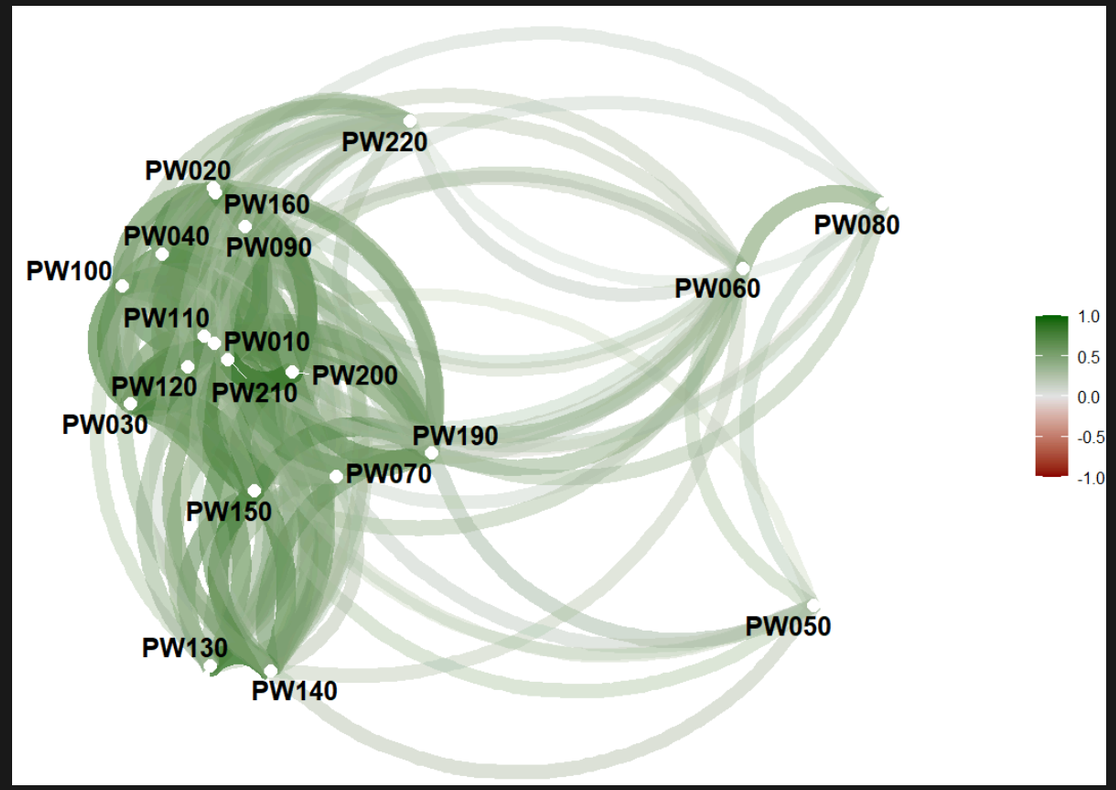


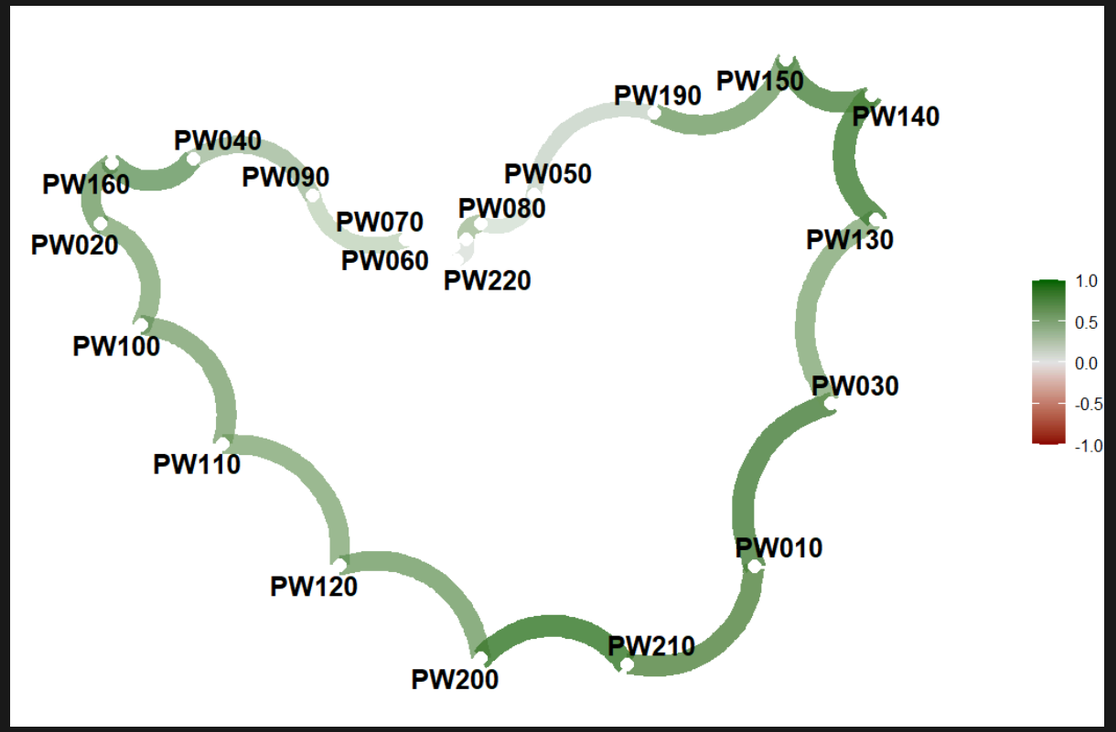
A problem: How to arrange the sectors for a better view?

We propose to sort variables according to an Hamiltonian cycle and to a the Traveling Salesman Problem.

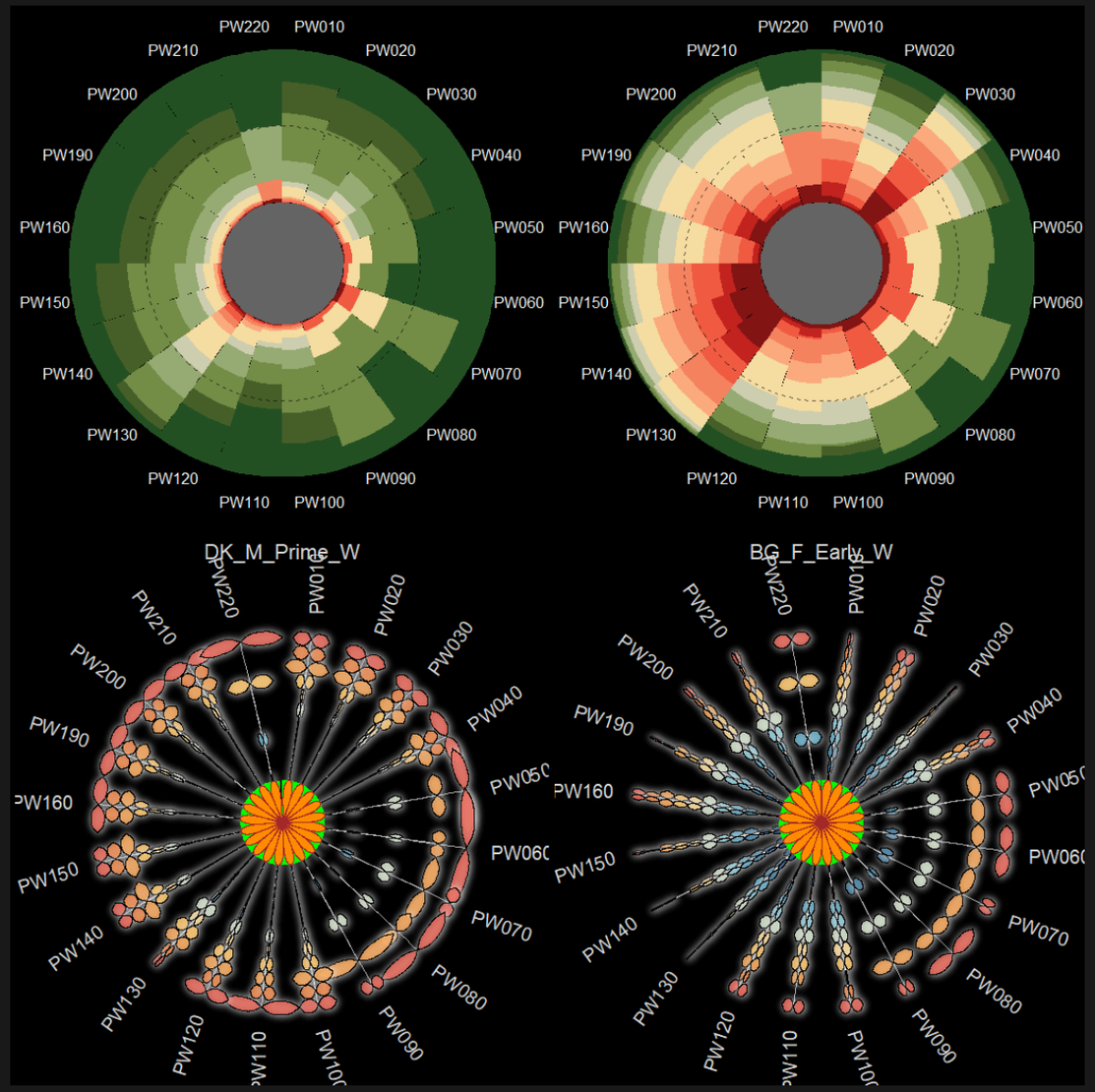
The problem assumes a network where nodes are the variables and the links are weighted by the correlation distance:

$$d(X_1, X_2) = \sqrt{2\left[1 - Corr(X_1, X_2)\right]}$$

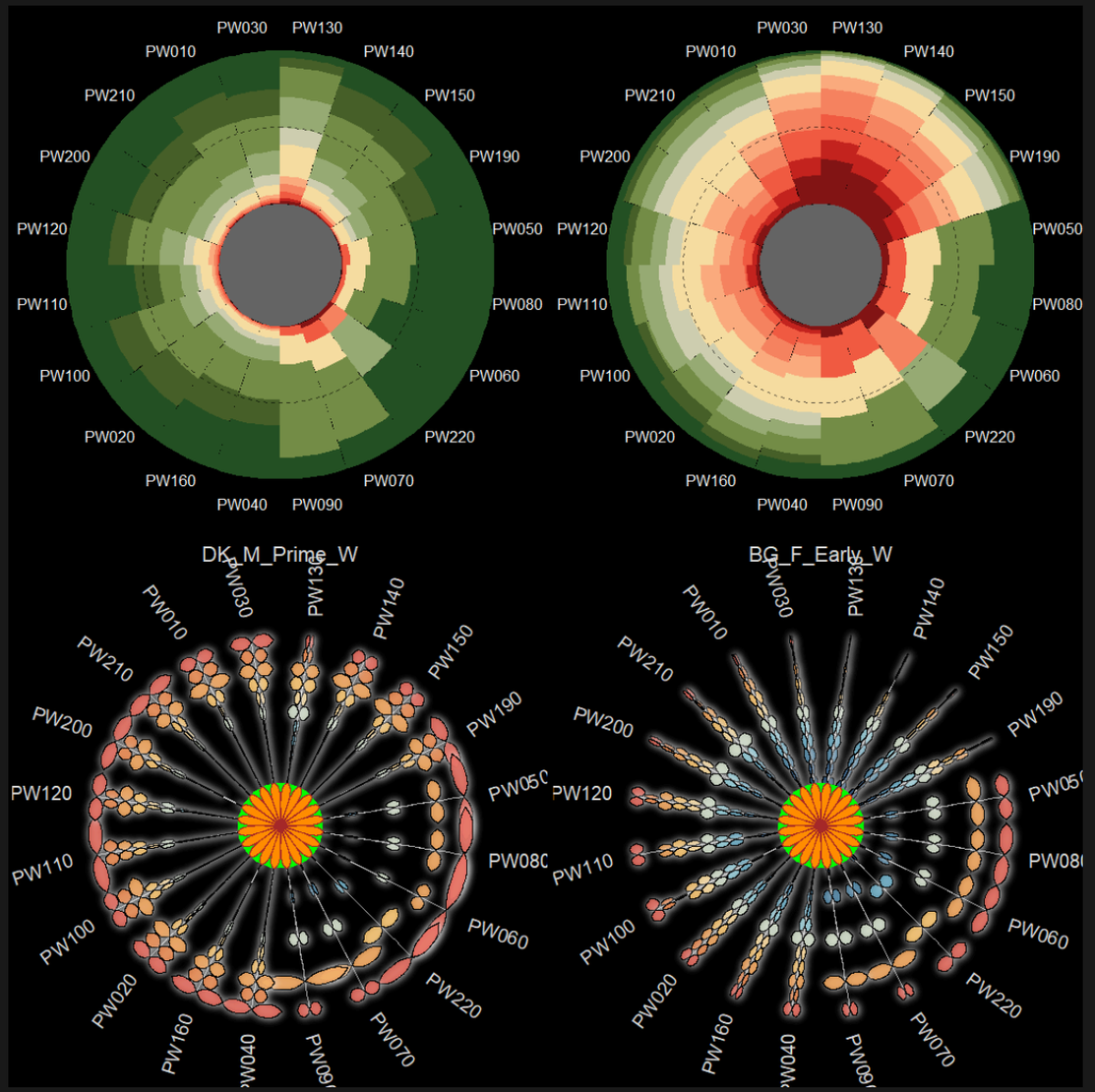




Comparing them

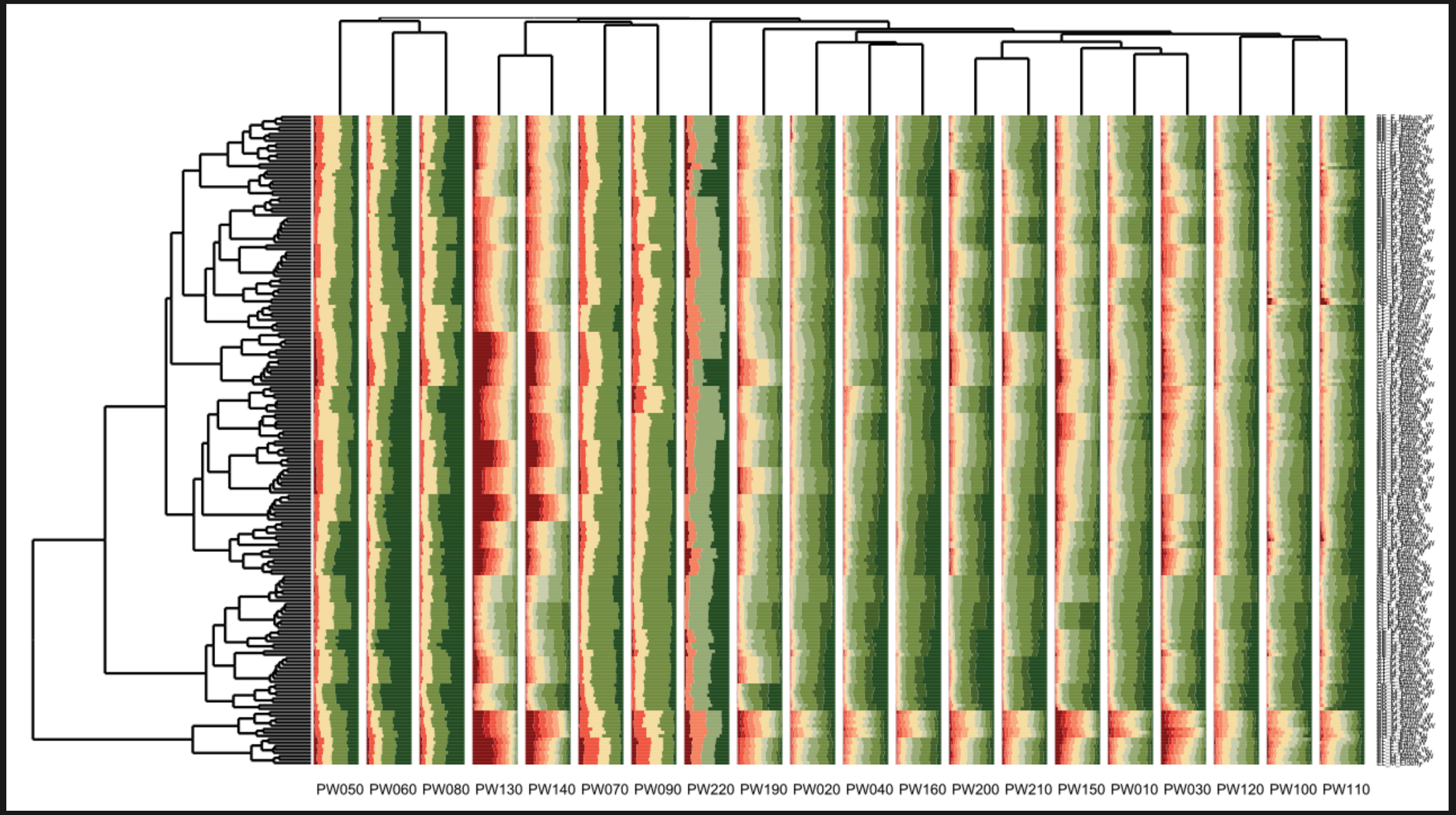


Unsorted variables



Sorted variables

Looking all the units and all the variables: The distributional heatmap



Conclusions

- We introduced new visualization tools for data tables of macrodata described by numeric distributions.
- The **Green Eye Iris plot** and the **Flower plot** allow to compare few macrodata.
- The **Distributional Heatmap** is a useful tool for moderate to big sized table.

… for the future …

- Macrodata can be described by categoracal distributions, and, as for the classic tables, effective visualization tools needs to be developed.
- The perception of differences between distributions is, indeed, difficult to visualize when the size of data increases both either in the number of individuals or the variables.
- Visualizations tools able to translate statistical measures and concepts needs to be developed.

References

Billard, L., and E. Diday. 2006. *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. Wiley.

Brito, P., and S. Dias. 2022. *Analysis of Distributional Data*. Chapman; Hall/CRC. https://doi.org/https://doi.org/10.1201/9781315370545.

Irpino, A. 2018. *HistDAWass: Histogram Data Analysis Using Wasserstein Distance*.

Irpino, A., and E. Romano. 2007. “Optimal Histogram Representation of Large Data Sets: Fisher Vs Piecewise Linear Approximation.” *Revue Des Nouvelles Technologies de l’Information* RNTI-E-9: 99–110.

Irpino, A., and R. Verde. 2006. “A New Wasserstein Based Distance for the Hierarchical Clustering of Histogram Symbolic Data.” In *Data Science and Classification*, edited by V. et al. Batagelj, 185–92. Berlin: Springer.

———. 2015. “Basic Statistics for Distributional Symbolic Variables: A New Metric-Based Approach.” *Advances in Data Analysis and Classification* 9 (2): 143–75.

Verde, R., and A. Irpino. 2018. “Multiple Factor Analysis of Distributional Data.” *Statistica Applicata, Italin Journal of Applied Statistics*.

Thank you !

- antonio.irpino@unicampania.it