Principal Component Analysis

Dataset db_bloco_1_posn

This dataset contains 615 individuals and 5 variables.

---

1. Study of the outliers

The analysis of the graphs does not detect any outlier.

---

2. Inertia distribution

The inertia of the first dimensions shows if there are strong relationships between variables and suggests the number of dimensions that should be studied.

The first two dimensions of analyse express **73.32%** of the total dataset inertia ; that means that 73.32% of the individuals (or variables) cloud total variability is explained by the plane. This percentage is high and thus the first plane represents an important part of the data variability. This value is strongly greater than the reference value that equals **44.6%**, the variability explained by this plane is thus highly significant (the reference value is the 0.95-quantile of the inertia percentages distribution obtained by simulating 4569 data tables of equivalent size on the basis of a normal distribution).

From these observations, it is probably not useful to interpret the next dimensions.

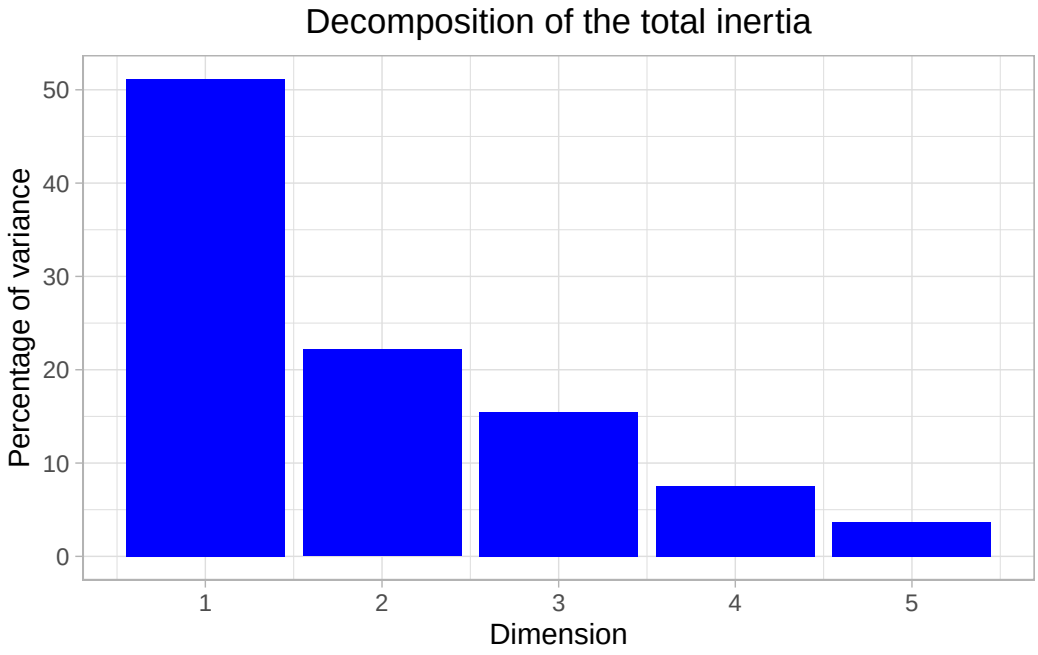

Figure 2 - Decomposition of the total inertia *The first factor is major: it expresses itself 51.17% of the data variability. Note that in such a case, the variability related to the other components might be meaningless, despite of a high percentage.*

An estimation of the right number of axis to interpret suggests to restrict the analysis to the description of the first 2 axis. These axis present an amount of inertia greater than those obtained by the 0.95-quantile of random distributions (73.32% against 44.6%). This observation suggests that only these axis are carrying a real information. As a consequence, the description will stand to these axis.

---

3. Description of the plane 1:2

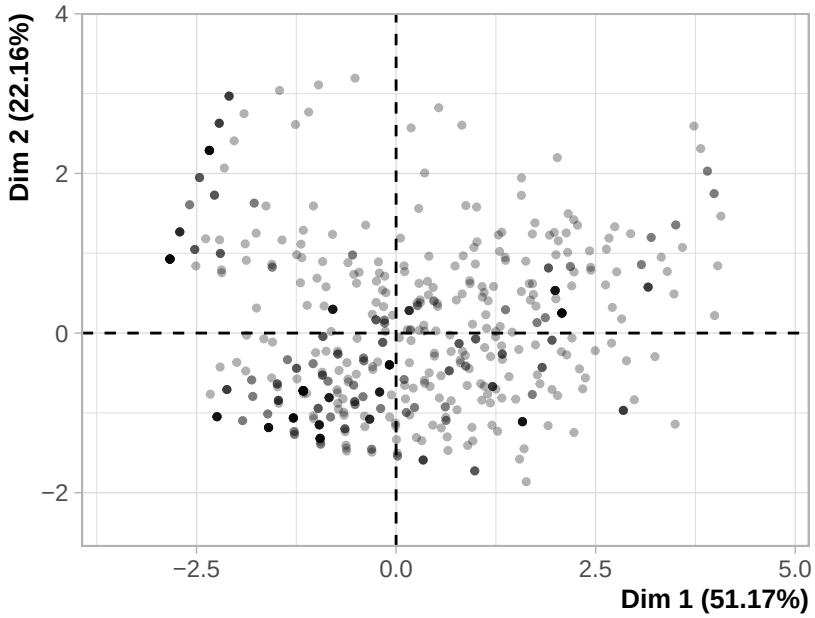

Figure 3.1 - Individuals factor map (PCA) *The labeled individuals are those with the higher contribution to the plane construction.*

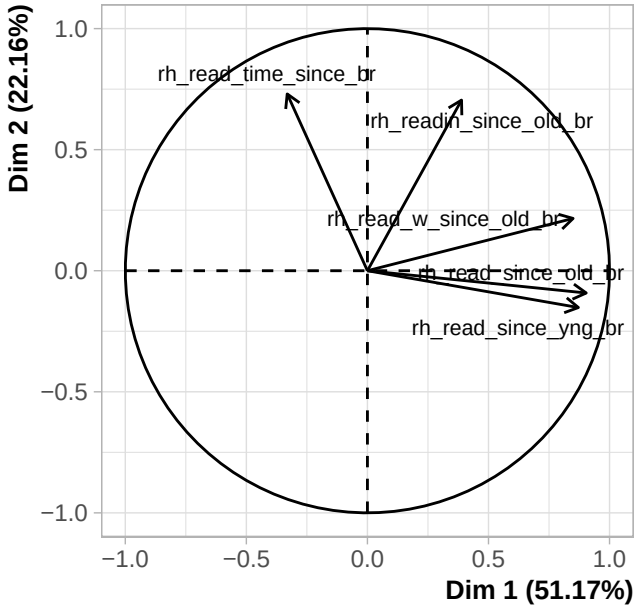

Figure 3.2 - Variables factor map (PCA) *The labeled variables are those the best shown on the plane.*

The **dimension 1** opposes individuals characterized by a strongly positive coordinate on the axis (to the right of the graph) to individuals characterized by a strongly negative coordinate on the axis (to the left of the graph).

The group 1 (characterized by a positive coordinate on the axis) is sharing :

- high values for the variables *rh_read_since_old_br*, *rh_read_w_since_old_br*, *rh_read_since_yng_br* and *rh_readin_since_old_br* (variables are sorted from the strongest).
- low values for the variable *rh_read_time_since_br*.

The group 2 (characterized by a negative coordinate on the axis) is sharing :

- high values for the variables *rh_read_time_since_br* and *rh_readin_since_old_br* (variables are sorted from the strongest).
- low values for the variables *rh_read_since_yng_br*, *rh_read_since_old_br* and *rh_read_w_since_old_br* (variables are sorted from the weakest).

The group 3 (characterized by a negative coordinate on the axis) is sharing :

- low values for the variables *rh_readin_since_old_br*, *rh_read_w_since_old_br*, *rh_read_time_since_br*, *rh_read_since_old_br* and *rh_read_since_yng_br* (variables are sorted from the weakest).

The **dimension 2** opposes individuals characterized by a strongly positive coordinate on the axis (to the top of the graph) to individuals characterized by a strongly negative coordinate on the axis (to the bottom of the graph).

The group 1 (characterized by a positive coordinate on the axis) is sharing :

- high values for the variables *rh_read_time_since_br* and *rh_readin_since_old_br* (variables are sorted from the strongest).
- low values for the variables *rh_read_since_yng_br*, *rh_read_since_old_br* and *rh_read_w_since_old_br* (variables are sorted from the weakest).

The group 2 (characterized by a positive coordinate on the axis) is sharing :

- high values for the variables *rh_read_since_old_br*, *rh_read_w_since_old_br*, *rh_read_since_yng_br* and *rh_readin_since_old_br* (variables are sorted from the strongest).
- low values for the variable *rh_read_time_since_br*.

The group 3 (characterized by a negative coordinate on the axis) is sharing :

- low values for the variables *rh_readin_since_old_br*, *rh_read_w_since_old_br*, *rh_read_time_since_br*, *rh_read_since_old_br* and *rh_read_since_yng_br* (variables are sorted from the weakest).

---

4. Classification

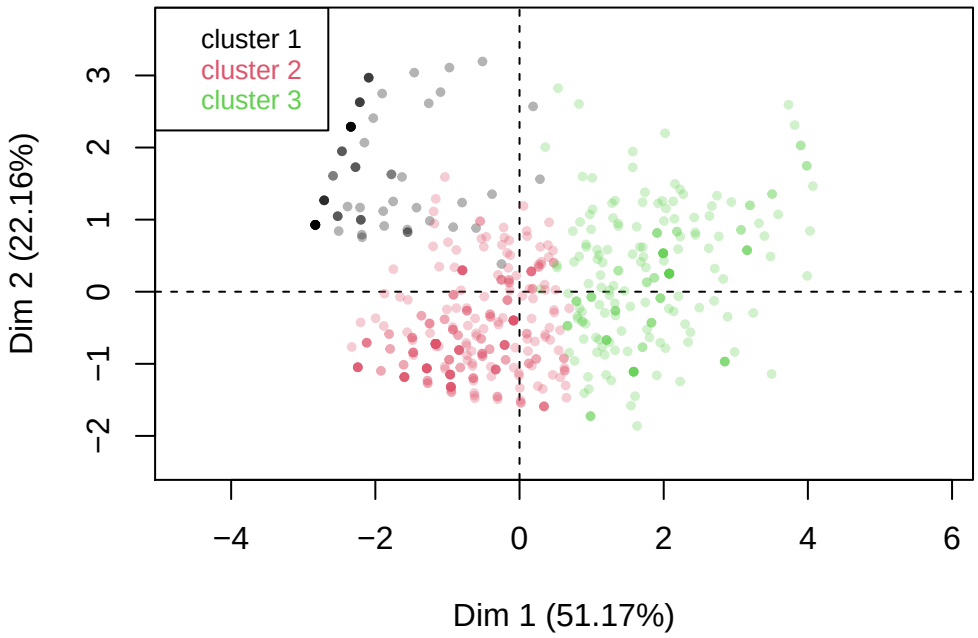

Figure 4 - Ascending Hierarchical Classification of the individuals. *The classification made on individuals reveals 3 clusters.*

The **cluster 1** is made of individuals sharing :

- high values for the variable *rh_read_time_since_br*.
- low values for the variables *rh_read_since_yng_br*, *rh_read_since_old_br* and *rh_read_w_since_old_br* (variables are sorted from the weakest).

The **cluster 2** is made of individuals sharing :

- low values for the variables *rh_read_time_since_br*, *rh_read_w_since_old_br*, *rh_read_since_old_br*, *rh_read_since_yng_br* and *rh_readin_since_old_br* (variables are sorted from the weakest).

The **cluster 3** is made of individuals sharing :

- high values for the variables *rh_read_since_old_br*, *rh_read_since_yng_br*, *rh_read_w_since_old_br* and *rh_readin_since_old_br* (variables are sorted from the strongest).
- low values for the variable *rh_read_time_since_br*.

Annexes