Principal Component Analysis

Dataset testDF_imputed$completeObs

This dataset contains 10850 individuals and 34 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 **62.67%** of the total dataset inertia ; that means that 62.67% of the individuals (or variables) cloud total variability is explained by the plane. This percentage is relatively high and thus the first plane well represents the data variability. This value is strongly greater than the reference value that equals **6.52%**, 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 370 data tables of equivalent size on the basis of a normal distribution).

From these observations, it should be better to also interpret the dimensions greater or equal to the third one.

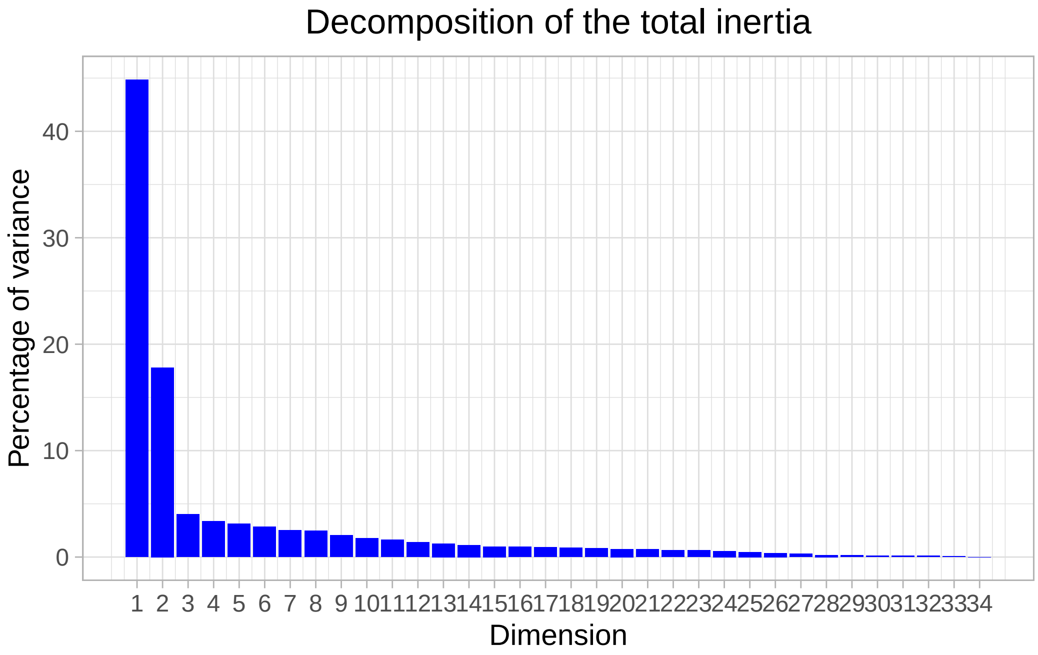

Figure 2 - Decomposition of the total inertia

An estimation of the right number of axis to interpret suggests to restrict the analysis to the description of the first 4 axis. These axis present an amount of inertia greater than those obtained by the 0.95-quantile of random distributions (70.05% against 12.89%). 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

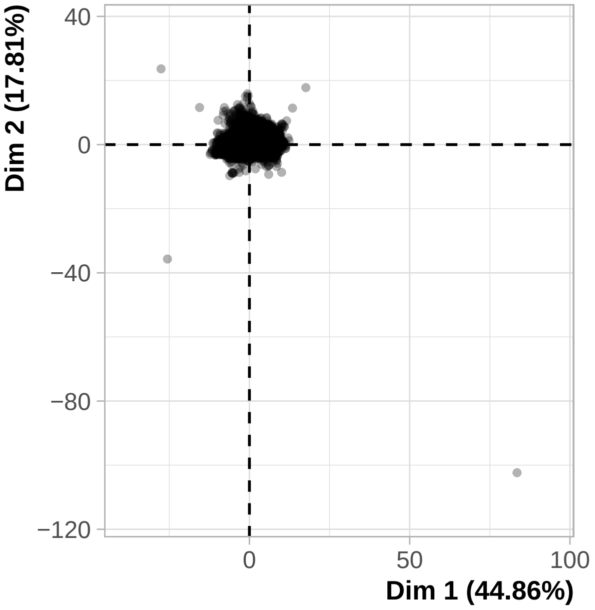

Figure 3.1 - Individuals factor map (PCA) *The labeled individuals are those with the higher contribution to the plane construction.*

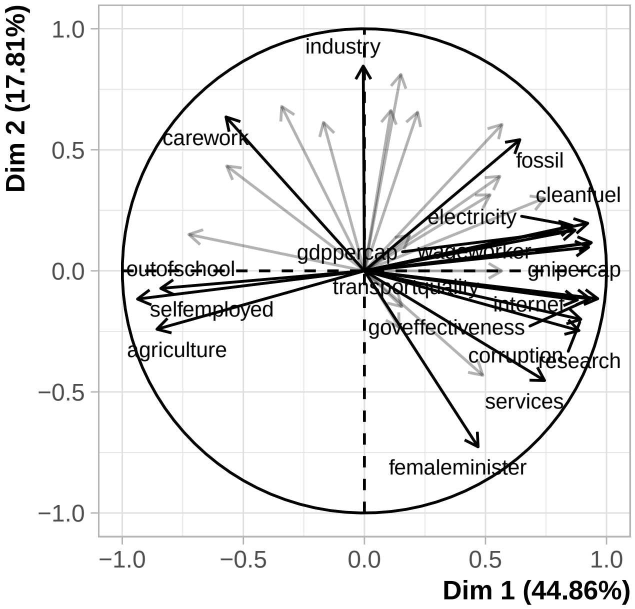

Figure 3.2 - Variables factor map (PCA) *The labeled variables are those the best shown on the plane.*

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 variables like *industry*, *grosscapitalprivate*, *savings*, *resourcerents*, *femalemanager*, *militaryexpenditure*, *fossil*, *grosscapital*, *carework* and *urbanlevel* (variables are sorted from the strongest).
- low values for variables like *femaleminister*, *femaleparliament*, *agriculture*, *services*, *selfemployed*, *unemployed*, *outofschool*, *fdi*, *research* and *corruption* (variables are sorted from the weakest).

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

- high values for variables like *goveffectiveness*, *transportquality*, *corruption*, *research*, *internet*, *cleanfuel*, *wageworker*, *electricity*, *gnipercap* and *services* (variables are sorted from the strongest).
- low values for the variables *selfemployed*, *outofpocket*, *outofschool*, *agriculture*, *carework*, *urbanrate*, *resourcerents*, *militaryexpenditure* and *industry* (variables are sorted from the weakest).

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

- high values for the variables *selfemployed*, *agriculture*, *outofschool*, *outofpocket*, *carework*, *urbanrate* and *resourcerents* (variables are sorted from the strongest).

- low values for variables like *cleanfuel*, *wageworker*, *electricity*, *urbanlevel*, *gnipercap*, *femalemanager*, *gdppercap*, *transportquality*, *goveffectiveness* and *fossil* (variables are sorted from the weakest).

Note that the variable *transportquality* is highly correlated with this dimension (correlation of 0.93). This variable could therefore summarize itself the dimension 1.

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 variables like *industry*, *grosscapitalprivate*, *savings*, *resourcerents*, *femalemanager*, *militaryexpenditure*, *fossil*, *grosscapital*, *carework* and *urbanlevel* (variables are sorted from the strongest).
- low values for variables like *femaleminister*, *femaleparliament*, *agriculture*, *services*, *selfemployed*, *unemployed*, *outofschool*, *fdi*, *research* and *corruption* (variables are sorted from the weakest).

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

- high values for the variables *selfemployed*, *agriculture*, *outofschool*, *outofpocket*, *carework*, *urbanrate* and *resourcerents* (variables are sorted from the strongest).
- low values for variables like *cleanfuel*, *wageworker*, *electricity*, *urbanlevel*, *gnipercap*, *femalemanager*, *gdppercap*, *transportquality*, *goveffectiveness* and *fossil* (variables are sorted from the weakest).

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

- high values for variables like *goveffectiveness*, *transportquality*, *corruption*, *research*, *internet*, *cleanfuel*, *wageworker*, *electricity*, *gnipercap* and *services* (variables are sorted from the strongest).
- low values for the variables *selfemployed*, *outofpocket*, *outofschool*, *agriculture*, *carework*, *urbanrate*, *resourcerents*, *militaryexpenditure* and *industry* (variables are sorted from the weakest).

---

4. Description of the plane 3:4

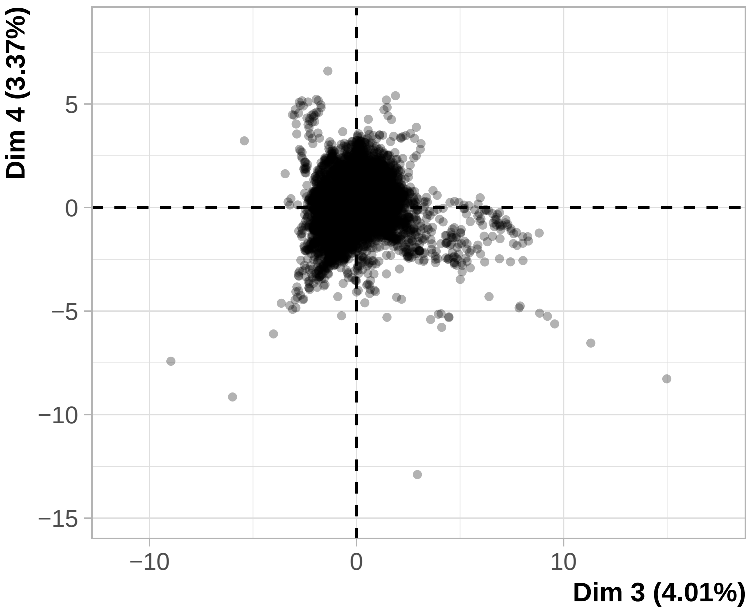

Figure 4.1 - Individuals factor map (PCA) *The labeled individuals are those with the higher contribution to the plane construction.*

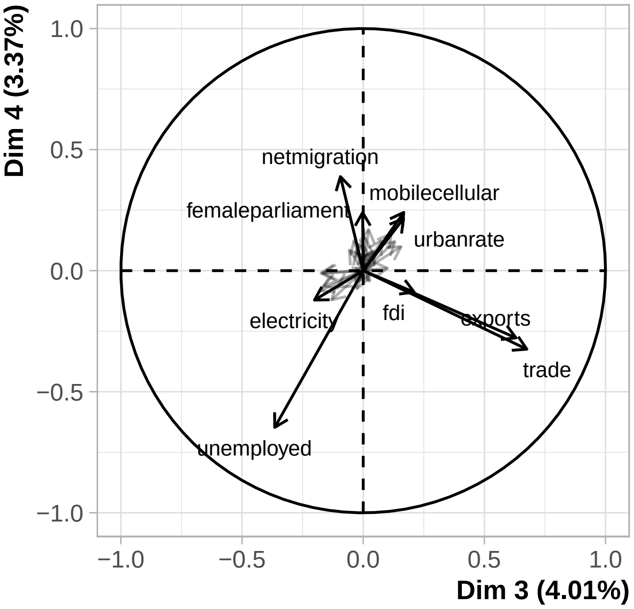

Figure 4.2 - Variables factor map (PCA) *The labeled variables are those the best shown on the plane.*

The **dimension 3** 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 *selfemployed*, *agriculture*, *urbanrate*, *trade*, *exports*, *outofschool*, *carework*, *resourcerents*, *mobilecellular* and *fdi* (variables are sorted from the strongest).
- low values for variables like *urbanlevel*, *unemployed*, *electricity*, *fossil*, *wageworker*, *cleanfuel*, *industry*, *research*, *transportquality* and *militaryexpenditure* (variables are sorted from the weakest).

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

- high values for variables like *exports*, *trade*, *gdppercap*, *femalemanager*, *gnipercap*, *transportquality*, *goveffectiveness*, *corruption*, *fdi* and *services* (variables are sorted from the strongest).
- low values for the variables *outofschool*, *selfemployed*, *agriculture*, *outofpocket*, *carework*, *resourcerents*, *industry*, *unemployed* and *urbanrate* (variables are sorted from the weakest).

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

- high values for variables like *urbanlevel*, *netmigration*, *internet*, *fossil*, *research*, *industry*, *femaleparliament*, *electricity*, *cleanfuel* and *gdppercap* (variables are sorted from the strongest).
- low values for variables like *trade*, *exports*, *unemployed*, *agriculture*, *outofschool*, *fdi*, *selfemployed*, *outofpocket*, *carework* and *resourcerents* (variables are sorted from the weakest).

The group 4 (characterized by a negative coordinate on the axis) is sharing :

- high values for the variables *unemployed*, *outofpocket*, *wageworker*, *electricity*, *outofschool*, *industry*, *militaryexpenditure*, *cleanfuel* and *carework* (variables are sorted from the strongest).
- low values for variables like *urbanrate*, *gdppercap*, *internet*, *netmigration*, *gnipercap*, *femaleparliament*, *grosscapital*, *savings*, *mobilecellular* and *grosscapitalprivate* (variables are sorted from the weakest).

The **dimension 4** 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 variables like *urbanlevel*, *netmigration*, *internet*, *fossil*, *research*, *industry*, *femaleparliament*, *electricity*, *cleanfuel* and *gdppercap* (variables are sorted from the strongest).
- low values for variables like *trade*, *exports*, *unemployed*, *agriculture*, *outofschool*, *fdi*, *selfemployed*, *outofpocket*, *carework* and *resourcerents* (variables are sorted from the weakest).

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

- high values for the variables *selfemployed*, *agriculture*, *urbanrate*, *trade*, *exports*, *outofschool*, *carework*, *resourcerents*, *mobilecellular* and *fdi* (variables are sorted from the strongest).
- low values for variables like *urbanlevel*, *unemployed*, *electricity*, *fossil*, *wageworker*, *cleanfuel*, *industry*, *research*, *transportquality* and *militaryexpenditure* (variables are sorted from the weakest).

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

- high values for the variables *unemployed*, *outofpocket*, *wageworker*, *electricity*, *outofschool*, *industry*, *militaryexpenditure*, *cleanfuel* and *carework* (variables are sorted from the strongest).
- low values for variables like *urbanrate*, *gdppercap*, *internet*, *netmigration*, *gnipercap*, *femaleparliament*, *grosscapital*, *savings*, *mobilecellular* and *grosscapitalprivate* (variables are sorted from the weakest).

The group 4 (characterized by a negative coordinate on the axis) is sharing :

- high values for variables like *exports*, *trade*, *gdppercap*, *femalemanager*, *gnipercap*, *transportquality*, *goveffectiveness*, *corruption*, *fdi* and *services* (variables are sorted from the strongest).
- low values for the variables *outofschool*, *selfemployed*, *agriculture*, *outofpocket*, *carework*, *resourcerents*, *industry*, *unemployed* and *urbanrate* (variables are sorted from the weakest).

---

5. Classification

The dataset is too large to perform the classification.

Annexes