Principal Component Analysis

Dataset dados.ctrl

This dataset contains 18 individuals and 6 variables, 1 qualitative variable is considered as illustrative.

### 1. Study of the outliers

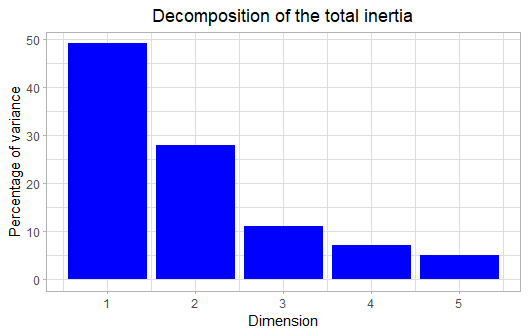
The analysis of the graphs does not detect any outlier.

### 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 **77.11%** of the total dataset inertia ; that means that 77.11% 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 greater than the reference value that equals **67.97%**, the variability explained by this plane is thus significant (the reference value is the 0.95-quantile of the inertia percentages distribution obtained by simulating 1815 data tables of equivalent size on the basis of a normal distribution).

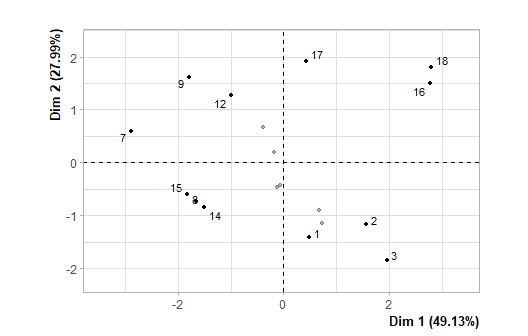
From these observations, it is probably not useful to interpret the next dimensions.



**Figure 2 - Decomposition of the total inertia**

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 (77.11% against 67.97%). 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

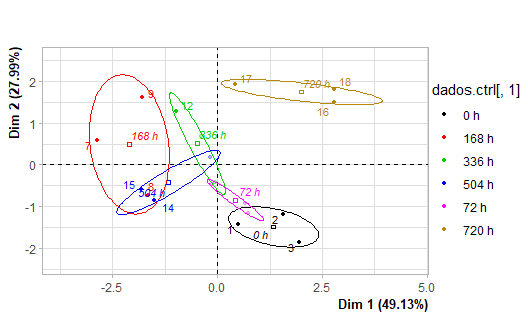


**Figure 3.1 - Individuals factor map (PCA)** *The labeled individuals are those with the higher contribution to the plane construction.*

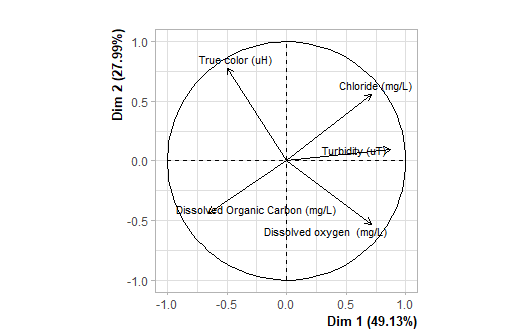
The Wilks test p-value indicates which variable factors are the best separated on the plane (i.e. which one explain the best the distance between individuals).

## dados.ctrl[, 1]   
## 6.568471e-06

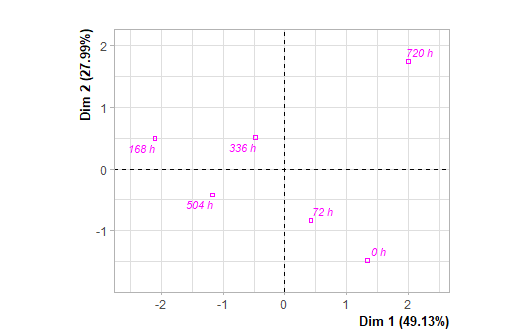
There only is one possible qualitative variable to illustrate the distance between individuals : *dados.ctrl[, 1]*.



**Figure 3.2 - Individuals factor map (PCA)** *The labeled individuals are those with the higher contribution to the plane construction.* *The individuals are coloured after their category for the variable* dados.ctrl[, 1].



**Figure 3.3 - Variables factor map (PCA)** *The labeled variables are those the best shown on the plane.*



**Figure 3.4 - Qualitative factor map (PCA)** *The labeled factors are those the best shown on the plane.*

The **dimension 1** opposes individuals such as *16* and *18* (to the right of the graph, characterized by a strongly positive coordinate on the axis) to individuals such as *15*, *9*, *14* and *7* (to the left of the graph, characterized by a strongly negative coordinate on the axis).

The group in which the individuals *16* and *18* stand (characterized by a positive coordinate on the axis) is sharing :

* high values for the variables *Chloride.(mg/L)* and *Turbidity.(uT)* (variables are sorted from the strongest).
* low values for the variable *Dissolved.Organic.Carbon.(mg/L)*.

The group in which the individuals *15*, *9*, *14* and *7* stand (characterized by a negative coordinate on the axis) is sharing :

* high values for the variable *True.color.(uH)*.
* low values for the variables *Dissolved.oxygen..(mg/L)* and *Turbidity.(uT)* (variables are sorted from the weakest).

Note that the variable *168 h* is highly correlated with this dimension (correlation of 0.94). This variable could therefore summarize itself the dimension 1.

The **dimension 2** opposes individuals such as *16* and *18* (to the top of the graph, characterized by a strongly positive coordinate on the axis) to individuals such as *3*, *2* and *1* (to the bottom of the graph, characterized by a strongly negative coordinate on the axis).

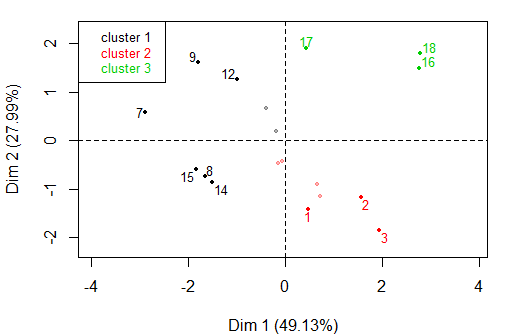
The group in which the individuals *16* and *18* stand (characterized by a positive coordinate on the axis) is sharing :

* high values for the variables *Chloride.(mg/L)* and *Turbidity.(uT)* (variables are sorted from the strongest).
* low values for the variable *Dissolved.Organic.Carbon.(mg/L)*.

The group in which the individuals *3*, *2* and *1* stand (characterized by a negative coordinate on the axis) is sharing :

* high values for the variable *Dissolved.oxygen..(mg/L)*.
* low values for the variable *True.color.(uH)*.

### 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 such as *7*, *8*, *9*, *12*, *14* and *15*. This group is characterized by :

* high values for the variable *True.color.(uH)*.
* low values for the variables *Dissolved.oxygen..(mg/L)* and *Turbidity.(uT)* (variables are sorted from the weakest).

The **cluster 2** is made of individuals such as *1*, *2* and *3*. This group is characterized by :

* high values for the variable *Dissolved.oxygen..(mg/L)*.
* low values for the variable *True.color.(uH)*.

The **cluster 3** is made of individuals such as *16*, *17* and *18*. This group is characterized by :

* high values for the variables *Chloride.(mg/L)* and *Turbidity.(uT)* (variables are sorted from the strongest).
* low values for the variable *Dissolved.Organic.Carbon.(mg/L)*.

## Annexes

dimdesc(res, axes = 1:2)

$Dim.1  
$quanti  
 correlation p.value  
Turbidity (uT) 0.8732020 2.240333e-06  
Dissolved oxygen (mg/L) 0.7122125 9.128607e-04  
Chloride (mg/L) 0.7117999 9.220448e-04  
True color (uH) -0.4963556 3.615441e-02  
Dissolved Organic Carbon (mg/L) -0.6584955 2.964083e-03  
  
$quali  
 R2 p.value  
dados.ctrl[, 1] 0.8148127 0.0004564716  
  
$category  
 Estimate p.value  
dados.ctrl[, 1]=720 h 2.001628 0.013292918  
dados.ctrl[, 1]=168 h -2.106269 0.008340847  
  
attr(,"class")  
[1] "condes" "list "   
  
$Dim.2  
$quanti  
 correlation p.value  
True color (uH) 0.7760078 0.000153407  
Chloride (mg/L) 0.5551292 0.016782360  
Dissolved oxygen (mg/L) -0.5355601 0.021985990  
  
$quali  
 R2 p.value  
dados.ctrl[, 1] 0.7858152 0.001044702  
  
$category  
 Estimate p.value  
dados.ctrl[, 1]=720 h 1.740790 0.002985824  
dados.ctrl[, 1]=0 h -1.479202 0.015827164  
  
attr(,"class")  
[1] "condes" "list "   
  
$call  
$call$num.var  
[1] 1  
  
$call$proba  
[1] 0.05  
  
$call$weights  
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
  
$call$X  
 Dim.1 dados.ctrl[, 1] Turbidity (uT) True color (uH) Dissolved oxygen (mg/L)  
6 -0.12947047 dados.ctrl[, 1]=72 h 5.97 80.7657 6.30  
7 -2.88041991 dados.ctrl[, 1]=168 h 4.59 95.8061 3.22  
10 -0.06378557 dados.ctrl[, 1]=336 h 5.99 73.2455 4.85  
12 -0.99209035 dados.ctrl[, 1]=336 h 5.31 84.5258 2.89  
14 -1.50985201 dados.ctrl[, 1]=504 h 4.50 73.2455 3.91  
16 2.77793910 dados.ctrl[, 1]=720 h 7.27 77.0056 6.30  
15 -1.82531916 dados.ctrl[, 1]=504 h 4.26 77.0056 3.80  
5 0.66775092 dados.ctrl[, 1]=72 h 5.91 69.4854 6.26  
9 -1.78715347 dados.ctrl[, 1]=168 h 3.83 92.0460 3.06  
18 2.79537125 dados.ctrl[, 1]=720 h 10.10 77.0056 4.20  
2 1.56870969 dados.ctrl[, 1]=0 h 6.50 65.7253 7.20  
1 0.48542717 dados.ctrl[, 1]=0 h 6.55 65.7253 5.85  
11 -0.38775291 dados.ctrl[, 1]=336 h 5.51 84.5258 4.70  
8 -1.65123354 dados.ctrl[, 1]=168 h 3.59 69.4854 3.24  
17 0.43157356 dados.ctrl[, 1]=720 h 5.17 88.2859 4.73  
4 0.73096815 dados.ctrl[, 1]=72 h 6.25 65.7253 5.96  
13 -0.18098909 dados.ctrl[, 1]=504 h 5.77 77.0056 4.34  
3 1.95032665 dados.ctrl[, 1]=0 h 6.95 65.7253 8.90  
 Chloride (mg/L) Dissolved Organic Carbon (mg/L)  
6 61.5443 18.8  
7 48.6389 18.7  
10 59.0873 16.6  
12 69.3126 18.8  
14 50.4481 16.9  
16 76.5410 11.0  
15 52.9002 18.2  
5 58.4088 15.0  
9 63.3046 16.6  
18 72.2762 12.1  
2 58.4847 13.4  
1 58.3756 17.6  
11 59.4973 15.0  
8 51.9858 15.9  
17 64.3679 10.5  
4 60.8607 16.6  
13 57.9144 14.7  
3 62.2655 16.5

**Figure 5 - List of variables characterizing the dimensions of the analysis.**

res.hcpc$desc.var

Link between the cluster variable and the categorical variables (chi-square test)  
=================================================================================  
 p.value df  
dados.ctrl[,.1] 0.0002959125 10  
  
Description of each cluster by the categories  
=============================================  
$`1`  
NULL  
  
$`2`  
 Cla/Mod Mod/Cla Global p.value v.test  
dados.ctrl[,.1]=72 h 100 42.85714 16.66667 0.04289216 2.024759  
dados.ctrl[,.1]=0 h 100 42.85714 16.66667 0.04289216 2.024759  
  
$`3`  
 Cla/Mod Mod/Cla Global p.value v.test  
dados.ctrl[,.1]=720 h 100 100 16.66667 0.00122549 3.23288  
  
  
Link between the cluster variable and the quantitative variables  
================================================================  
 Eta2 P-value  
Dissolved.oxygen..(mg/L) 0.6643845 0.0002778532  
Dissolved.Organic.Carbon.(mg/L) 0.6639091 0.0002808187  
Turbidity.(uT) 0.5425130 0.0028368796  
Chloride.(mg/L) 0.5023053 0.0053360847  
True.color.(uH) 0.4139756 0.0181705249  
  
Description of each cluster by quantitative variables  
=====================================================  
$`1`  
 v.test Mean in category Overall mean sd in category Overall sd  
True.color.(uH) 1.993982 81.70573 76.796706 8.5640353 9.079101  
Turbidity.(uT) -2.803088 4.67000 5.778889 0.7416704 1.458884  
Dissolved.oxygen..(mg/L) -3.121834 3.64500 4.983889 0.6069596 1.581628  
 p.value  
True.color.(uH) 0.046153969  
Turbidity.(uT) 0.005061593  
Dissolved.oxygen..(mg/L) 0.001797285  
  
$`2`  
 v.test Mean in category Overall mean sd in category Overall sd  
Dissolved.oxygen..(mg/L) 3.099380 6.474286 4.983889 1.181982 1.581628  
True.color.(uH) -2.648678 69.485400 76.796706 5.317584 9.079101  
 p.value  
Dissolved.oxygen..(mg/L) 0.001939262  
True.color.(uH) 0.008080723  
  
$`3`  
 v.test Mean in category Overall mean sd in category Overall sd  
Chloride.(mg/L) 2.802318 71.061700 60.345217 5.0433024 7.051384  
Turbidity.(uT) 2.192194 7.513333 5.778889 2.0200055 1.458884  
Dissolved.Organic.Carbon.(mg/L) -3.338980 11.200000 15.716667 0.6683313 2.494271  
 p.value  
Chloride.(mg/L) 0.005073688  
Turbidity.(uT) 0.028365488  
Dissolved.Organic.Carbon.(mg/L) 0.000840866

**Figure 6 - List of variables characterizing the clusters of the classification.**