Principal Component Analysis

Dataset dados.ctrl

This dataset contains 18 individuals and 7 variables, 1 qualitative variable is considered as illustrative.

### 1. Study of the outliers

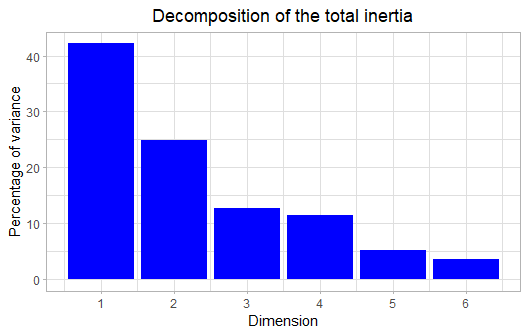
The analysis of the graphs does not detect any outlier.

### 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 **67.09%** of the total dataset inertia ; that means that 67.09% 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 greater than the reference value that equals **61.93%**, 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 2212 data tables of equivalent size on the basis of a normal distribution).

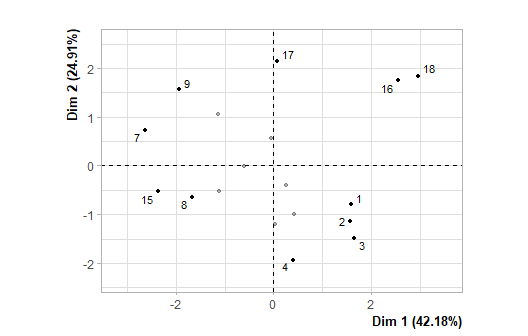
From these observations, it should be better to also interpret the dimensions greater or equal to the third one.



**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 2 axis. These axis present an amount of inertia greater than those obtained by the 0.95-quantile of random distributions (67.09% against 61.93%). 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

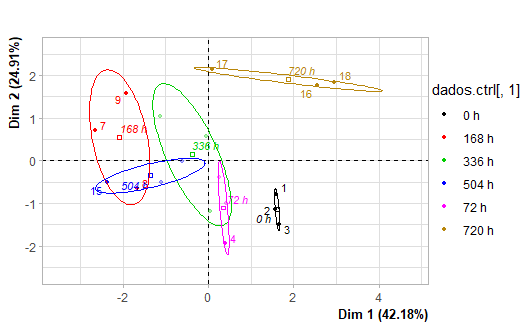


**Figure 3.1 - Individuals factor map (PCA)** *The labeled individuals are those with the higher contribution to the plane construction.*

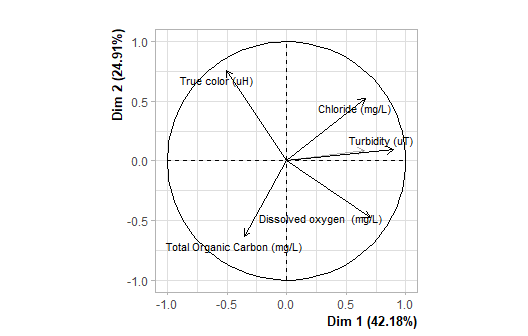
The Wilks test p-value indicates which variable factors are the best separated on the plane (i.e. which one explain the best the distance between individuals).

## dados.ctrl[, 1]   
## 1.054609e-05

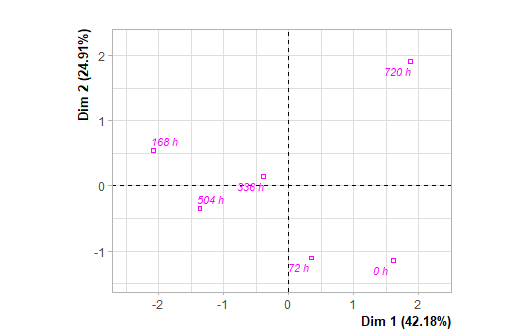
There only is one possible qualitative variable to illustrate the distance between individuals : *dados.ctrl[, 1]*.



**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*, *2*, *18*, *17*, *4*, *1* and *3* (to the right of the graph, characterized by a strongly positive coordinate on the axis) to individuals such as *9*, *7*, *15* and *8* (to the left of the graph, characterized by a strongly negative coordinate on the axis).

The group in which the individuals *2*, *4*, *1* and *3* stand (characterized by a positive coordinate on the axis) is sharing :

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

The group in which the individuals *16*, *18* and *17* 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 *Total.Organic.Carbon.(mg/L)*.

The group in which the individuals *9*, *7*, *15* and *8* 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)*, *Turbidity.(uT)* and *Nitrate.(mg/L)* (variables are sorted from the weakest).

Note that the variable *168 h* is highly correlated with this dimension (correlation of 0.91). This variable could therefore summarize itself the dimension 1.

The **dimension 2** opposes individuals such as *16*, *18* and *17* (to the top of the graph, characterized by a strongly positive coordinate on the axis) to individuals such as *2*, *4*, *1* and *3* (to the bottom of the graph, characterized by a strongly negative coordinate on the axis).

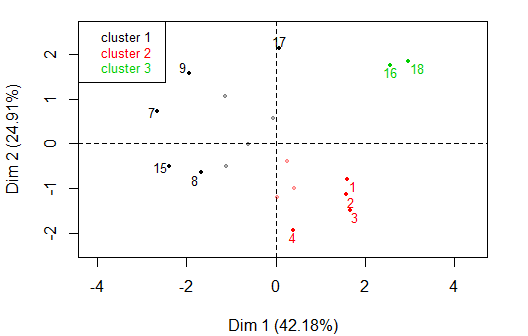
The group in which the individuals *16*, *18* and *17* stand (characterized by a positive coordinate on the axis) is sharing :

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

The group in which the individuals *2*, *4*, *1* and *3* 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*, *15* and *17*. 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*, *3* and *4*. 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* 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 *Total.Organic.Carbon.(mg/L)*.

## Annexes

dimdesc(res, axes = 1:2)

$Dim.1  
$quanti  
 correlation p.value  
Turbidity (uT) 0.8933708 5.984634e-07  
Dissolved oxygen (mg/L) 0.7001668 1.214518e-03  
Chloride (mg/L) 0.6590502 2.931643e-03  
Nitrate (mg/L) 0.6519646 3.368710e-03  
True color (uH) -0.5049208 3.258274e-02  
  
$quali  
 R2 p.value  
dados.ctrl[, 1] 0.8268597 0.0003105033  
  
$category  
 Estimate p.value  
dados.ctrl[, 1]=720 h 1.868066 0.02522035  
dados.ctrl[, 1]=168 h -2.082811 0.01067576  
  
attr(,"class")  
[1] "condes" "list "   
  
$Dim.2  
$quanti  
 correlation p.value  
True color (uH) 0.7573067 0.0002732362  
Chloride (mg/L) 0.5248115 0.0253388255  
Dissolved oxygen (mg/L) -0.4769044 0.0453788312  
Total Organic Carbon (mg/L) -0.6333442 0.0047772299  
  
$quali  
 R2 p.value  
dados.ctrl[, 1] 0.7401155 0.00309755  
  
$category  
 Estimate p.value  
dados.ctrl[, 1]=720 h 1.912919 0.001227338  
  
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)  
13 -0.60877226 dados.ctrl[, 1]=504 h 5.77 77.0056 4.34  
6 0.24966683 dados.ctrl[, 1]=72 h 5.97 80.7657 6.30  
1 1.60020668 dados.ctrl[, 1]=0 h 6.55 65.7253 5.85  
4 0.39579023 dados.ctrl[, 1]=72 h 6.25 65.7253 5.96  
7 -2.64152054 dados.ctrl[, 1]=168 h 4.59 95.8061 3.22  
15 -2.37277360 dados.ctrl[, 1]=504 h 4.26 77.0056 3.80  
5 0.41758056 dados.ctrl[, 1]=72 h 5.91 69.4854 6.26  
14 -1.11342067 dados.ctrl[, 1]=504 h 4.50 73.2455 3.91  
17 0.08295795 dados.ctrl[, 1]=720 h 5.17 88.2859 4.73  
2 1.57307608 dados.ctrl[, 1]=0 h 6.50 65.7253 7.20  
9 -1.93496640 dados.ctrl[, 1]=168 h 3.83 92.0460 3.06  
12 -1.13585584 dados.ctrl[, 1]=336 h 5.31 84.5258 2.89  
3 1.65507370 dados.ctrl[, 1]=0 h 6.95 65.7253 8.90  
16 2.55911179 dados.ctrl[, 1]=720 h 7.27 77.0056 6.30  
10 0.03533946 dados.ctrl[, 1]=336 h 5.99 73.2455 4.85  
8 -1.67194555 dados.ctrl[, 1]=168 h 3.59 69.4854 3.24  
11 -0.05167776 dados.ctrl[, 1]=336 h 5.51 84.5258 4.70  
18 2.96212935 dados.ctrl[, 1]=720 h 10.10 77.0056 4.20  
 Nitrate (mg/L) Chloride (mg/L) Total Organic Carbon (mg/L)  
13 0.5947 57.9144 18.3  
6 0.7532 61.5443 21.2  
1 1.0615 58.3756 17.6  
4 0.7011 60.8607 24.8  
7 0.6046 48.6389 19.6  
15 0.2815 52.9002 19.1  
5 0.6737 58.4088 18.7  
14 0.7615 50.4481 17.6  
17 0.7647 64.3679 11.4  
2 0.8710 58.4847 17.2  
9 0.5822 63.3046 18.1  
12 0.5211 69.3126 21.2  
3 0.5204 62.2655 17.5  
16 0.8414 76.5410 12.1  
10 0.8912 59.0873 24.8  
8 0.6592 51.9858 17.9  
11 0.9393 59.4973 18.7  
18 0.9852 72.2762 14.6

**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] 0.004287781 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 66.66667 100 16.66667 0.01960784 2.333769  
  
  
Link between the cluster variable and the quantitative variables  
================================================================  
 Eta2 P-value  
Turbidity.(uT) 0.7517157 2.898139e-05  
Dissolved.oxygen..(mg/L) 0.6451485 4.220366e-04  
Chloride.(mg/L) 0.5197683 4.082080e-03  
True.color.(uH) 0.4452091 1.204953e-02  
Total.Organic.Carbon.(mg/L) 0.3707749 3.097741e-02  
  
Description of each cluster by quantitative variables  
=====================================================  
$`1`  
 v.test Mean in category Overall mean sd in category Overall sd  
True.color.(uH) 2.561370 82.436856 76.796706 8.3348622 9.079101  
Turbidity.(uT) -2.976936 4.725556 5.778889 0.7166916 1.458884  
Dissolved.oxygen..(mg/L) -3.176043 3.765556 4.983889 0.6661350 1.581628  
 p.value  
True.color.(uH) 0.010426038  
Turbidity.(uT) 0.002911450  
Dissolved.oxygen..(mg/L) 0.001492989  
  
$`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) 2.907334 74.4086 60.345217 2.1324 7.051384  
Turbidity.(uT) 2.903828 8.6850 5.778889 1.4150 1.458884  
Total.Organic.Carbon.(mg/L) -2.153347 13.3500 18.355556 1.2500 3.388579  
 p.value  
Chloride.(mg/L) 0.003645240  
Turbidity.(uT) 0.003686309  
Total.Organic.Carbon.(mg/L) 0.031291383

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