Methodological Report

RTBfoods Manual - Part 3 - Tutorial

Statistical Analyses (PCA and multiple regression) to Visualise the Sensory Analysis Data and Relate it to the Instrumental Data

Biophysical Characterization of Quality Traits, WP2

Montpellier, France, July 2021

Christophe BUGAUD, Centre de coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), Montpellier, France

Isabelle MARAVAL, CIRAD, Montpellier, France

Karima MEGHAR, CIRAD, Montpellier, France

This report has been written in the framework of RTBfoods project.

To be cited as:

Christophe BUGAUD, Isabelle MARAVAL, Karima MEGHAR, (2021). *RTBfoods Manual - Tutorial: Statistical Analyses (PCA and multiple regression) to Visualise the Sensory Analysis Data and Relate it to the Instrumental Data.* Montpellier, France: RTBfoods Methodological Report, 24 p.

Ethics: The activities, which led to the production of this manual, were assessed and approved by the CIRAD Ethics Committee (H2020 ethics self-assessment procedure). When relevant, samples were prepared according to good hygiene and manufacturing practices. When external participants were involved in an activity, they were priorly informed about the objective of the activity and explained that their participation was entirely voluntary, that they could stop the interview at any point and that their responses would be anonymous and securely stored by the research team for research purposes. Written consent (signature) was systematically sought from sensory panelists and from consumers participating in activities.

Acknowledgments: This work was supported by the RTBfoods project https://rtbfoods.cirad.fr, through a grant OPP1178942: Breeding RTB products for end user preferences (RTBfoods), to the French Agricultural Research Centre for International Development (CIRAD), Montpellier, France, by the Bill & Melinda Gates Foundation (BMGF).

Image cover page © LAJOUS P. for RTBfoods.

This document has been reviewed by:	
Isabelle MARAVAL (CIRAD)	15/06/2021
Karima MEGHAR (CIRAD)	04/06/2021
Final validation by:	
Christophe BUGAUD (CIRAD)	02/07/2021

CONTENTS

Table of Contents

ABSTRACT

After sensory evaluation by a trained panel, and biophysical evaluation using the instrumental measurements of the different products, statistical treatments can be used to interpret the results. The objective of this tutorial, using XLSTAT software, is to perform and interpret 2 types of statistical treatments: (1) principal component analysis (PCA) which enables rapid visualisation of the correlations between the sensory attributes, and (2) linear regression, which allows prediction of the sensory attributes based on the biophysical (textural, biochemical) parameters. The performance of the panel has previously been checked, and the sensory data were prepared for statistical analysis (see RTBfoods_F.2.4A_Tutorial for Performance Monitoring & Sensory Data Cleaning Before Statistical Analysis_2021.pdf). The present tutorial is based on an example presented in a published Excel file that goes through one step after another. The selected PCA uses sensory data to identify major trends and sensory diversity between groups of products and between individual products. The PCA also makes it possible to measure differences between repeated products that reflect the performance of the panel (if the products are indeed identical). Multiple linear regression was used to predict sensory attributes from biophysical parameters. For this purpose, in our example, the dataset was split into two datasets: a calibration set representing ¾ of the data and a validation set containing the remaining data. Three prediction indicators were calculated to assess the accuracy and robustness of the prediction: the coefficient of determination (R^2), the mean difference between observed and predicted values (RMSEC) in the calibration set, and the mean difference between the observed and predicted values in the validation set (RMSEV). The relevance of the validation and the minimum number of observations necessary to build predictive models are discussed.

Key Words: sensory profile - instrumental measurement - principal component analysis - multiple linear regression - calibration - validation - RMSE

1 OBJECTIVE

In the previous tutorial (*RTBfoods_F.2.4A_Tutorial for Performance Monitoring & Sensory Data Cleaning Before Statistical Analysis_2021*.pdf), we presented a methodology to check the performance of the panel (repeatability and homogeneity) at the end of the analyses and to prepare the data for statistical analysis. In that tutorial, we focussed on eliminating non-performing panellists or unacceptable data. The last table in that tutorial displayed the average values obtained by the panel for each product, each attribute, on a scale of 0 to 10. The table can be completed by instrumental data (textural, biochemical parameters) obtained using the same samples.

Here, we present a tutorial based on our own experience using XLSTAT software to perform and interpret 2 types of statistical treatments that are important for WP2 on sensory analysis

1- Principal component analysis, which enables rapid visualization of the correlations between sensory attributes and their products,

2- Linear regression, which allows prediction of sensory attributes based on instrumental textural, and biochemical parameters.

Please note, this tutorial does not explain the statistical basis of PCA or linear regression, readers who are not familiar with these topics are invited to consult suitable books or publications. We would also like to point out that we are not statisticians, but users with our own way of using and interpreting these statistical tools.

The data set we used for statistical treatment

To illustrate the statistical treatments, we used a data set extracted from a study of plantain boiled at different stages of ripening (Example PCA + regression.xls). The results of this study and their interpretations are the subject of a scientific publication (doi:10.1111/ijfs.14765). In our example, the data are the mean values of 35 observations (cooked products from 13 cultivars and 3 different ripening stages) with the following variables:

- Sensory variables: 6 attributes (firmness, chewiness, stickiness, mealiness, moistness and sweetness) evaluated on a scale from 0 to 10, and 2 attributes (sourness, astringency) evaluated using a binomial scale (yes/no). For the 2 binomial attributes, the frequency of "yes" responses was calculated and transformed into a scale of 0 to 10 to allow comparison with the other attributes already evaluated on an intensity scale of 0 to 10.
- Instrumental variables: 7 textural parameters measured using a double compression test (parameter_TPA) and penetrometry test (puncture force) and 3 biochemical parameters (dry matter content, soluble solids content, and titratable acidity).

2 PRINCIPAL COMPONENT ANALYSIS

2.1 Definitions - principles

Principal Component Analysis is one of the most frequently used multivariate data analyses. It investigates multidimensional datasets with quantitative variables.

Principal Component Analysis is a useful way to analyse numerical data structured in n observations / p variables. PCA projects observations from a P-dimensional space with p variables onto a K-dimensional space (where K<P) in order to conserve the maximum amount of information from the initial dimensions, making it possible to:

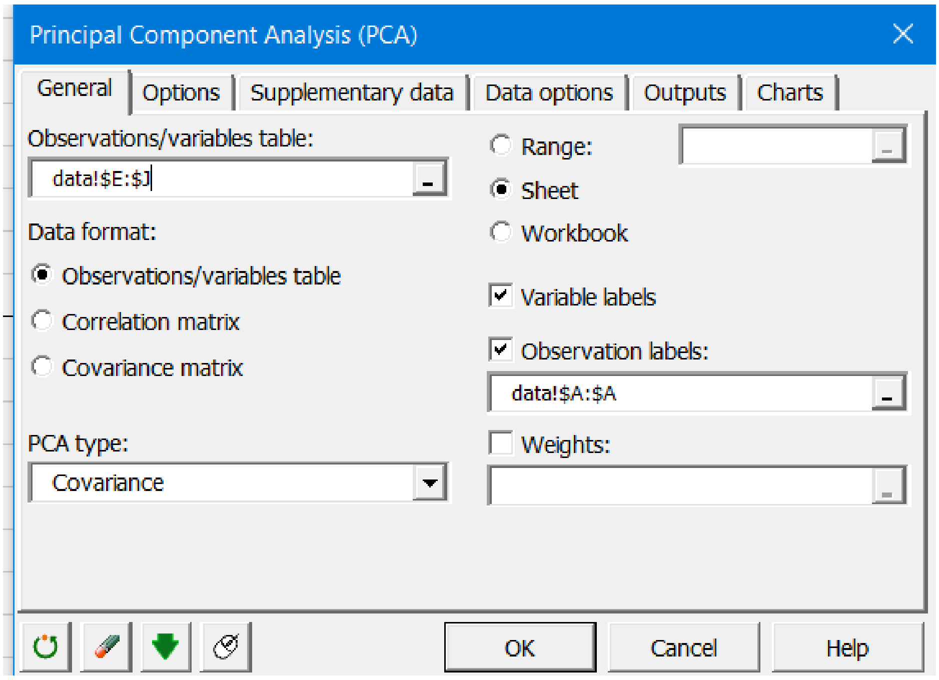
- Rapidly identify and analyse correlations between p variables,
- Indentify and analyse the n observations (originally described by the p variables) on a 2- or 3-dimensional map, in order to identify uniform or atypical groups of observations.

PCA dimensions are also called axes or factors.

In our example, n = 35, p = 6 (only sensory attributes rated on a 0 – 10 scale) or p = 8 (all sensory attributes).

2.2 How to set up a principal component analysis using XLSTAT

- Open the file « Example ACP + regression_EN.xls».
- Once XLSTAT is activated, select **Analysing data / Principal Component Analysis**.
- In the window “General”, click on “Observations/variables table” and select the 6 sensory attributes (columns E to J) from “data" sheet.
- Click on « Observation labels » and select column A on the “data” sheet.
- Activate the option “sheet” to display the results in a new worksheet.
- In « PCA type », select “Covariance”.



Observations/variables table: data!$E:$J

Data format:

● Observations/variables table

○ Correlation matrix

○ Covariance matrix

PCA type: Covariance

○ Range:

● Sheet

○ Workbook

☑ Variable labels

☑ Observation labels: data!$A:$A

□ Weights:

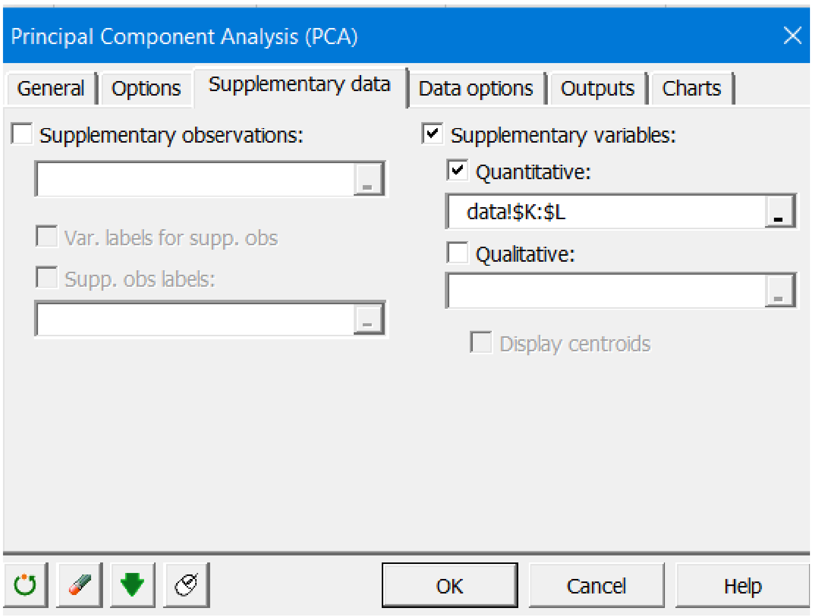
Why choose the 'Covariance' option?

When the variables are on identical scales (which is the case in our example), or when we want the variance of the variables to influence the construction of the factors, we use covariance. With this option, a sensory attribute using the full intensity scale (0 - 10) will influence the construction of the factors much more than an attribute limited to one part of the scale (for example between 0 and 3).

The "Pearson correlation" option is used when the variables are not all on the same scale (especially with instrumental data), making it possible to eliminate scale effects: thus a variable varying between 0 and 1 000 does not weigh more in the projection than a variable varying between 0 and 1. In this case, the data for each variable were already centered and reduced (by Xlstat).

- In the window “Options”, “n” is chosen by default. You can choose the "Rotation" option if you want to apply a rotation to the factorial coordinate matrix. In exceptional cases, this would make it possible to better position the variables on axes 1 and 2 thereby improving the interpretation of the results. Personally, we have never used the "Rotation" option.

- In the window “Supplementary data”, XLSTAT allows you to add additional observations or variables. However these observations or variables are not taken into account when you calculate the factors. In our example, we selected two attributes rated on a binomial scale (y/n), which were less important than the 6 attributes (active variables) rated on a scale from 0 to 10, as supplementary variables (columns K and L).



□ Supplementary observations:

□ Var. labels for supp. obs

□ Supp. obs labels:

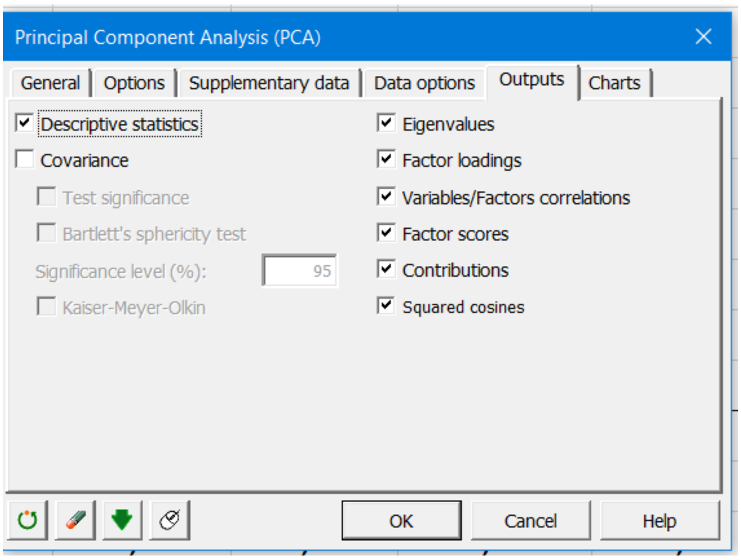
☑ Supplementary variables:

☑ Quantitative: data!$K:$L

□ Qualitative:

□ Display centroids

- In the window “Data options“, click on « remove the observations”.
- In the window « Outputs », select “descriptive statistics”, and select all options on the right.



☑ Descriptive statistics

□ Covariance

□ Test significance

□ Bartlett's sphericity test

Significance level (%): 95

□ Kaiser-Meyer-Olkin

☑ Eigenvalues

☑ Factor loadings

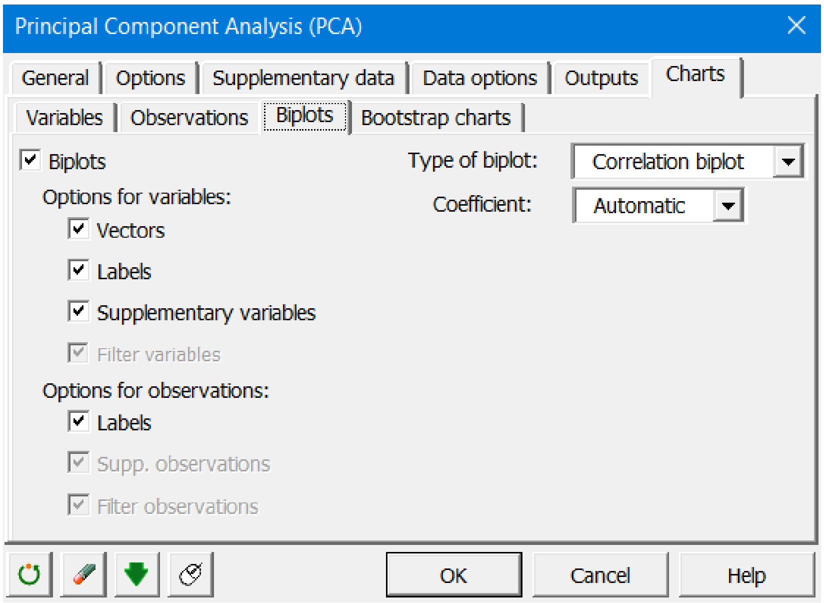
☑ Variables/Factors correlations

☑ Factor scores

☑ Contributions

☑ Squared cosines

- Configure charts in the window “Charts”:
    - In the window “Variables”, click on “correlation charts” and “vectors”,
    - In the window “Observations”, click on “observations charts” and “labels”,
    - In the window “Biplots”, select all options and « correlation biplot» in « type of biplot ».


Principal Component Analysis (PCA)
General | Options | Supplementary data | Data options | Outputs | Charts
Variables | Observations | Biplots | Bootstrap charts
OK
Cancel
Help

☑ Biplots

Options for variables:

☑ Vectors

☑ Labels

☑ Supplementary variables

☑ Filter variables

Options for observations:

☑ Labels

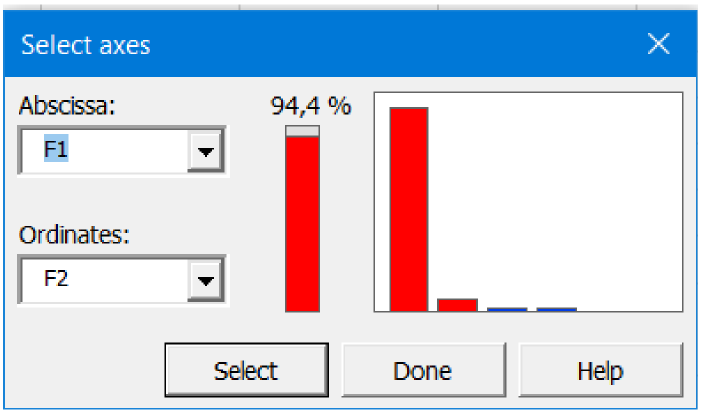
☑ Supp. observations

☑ Filter observations

Type of biplot: Correlation biplot

Coefficient: Automatic

- The computations begin once you have clicked on OK. You are asked to select the factors for which you want to display plots. By default, XLSTAT proposes the first 2 factors. In this example, the percentage of variability represented by the first two factors is 94.4%, so it will not be necessary to look at the following factors. Click on « Finish» to finish calculations.



Abscissa: F1

Ordinates: F2

2.3 Interpreting the results of the PCA

2.3.1 How to interpret eigenvalues in PCA

This table below concern a mathematical object, the eigenvalue, which reflects the quality of the projection of the data. Each eigenvalue corresponds to a factor. The eigenvalues and the corresponding factors are sorted by descending order according to how much of the initial variability they represent (converted into %).

In our example, the first eigenvalue represents 87.8% of total variability. This means that if we plot the data on a single axis, we will still be able to see 87.6% of the total variability of the data. With two axes, we will see 94.4% of the total variability, which is very good. The fewer variables and the more correlations between variables, more representative the first two axes. In our example, we don’t need to account for the 3rd axis, which only represents 2.5% of total variability. But (in our personal opinion), you may have to, especially when it represents more than 10% of total variability.

Eigenvalues

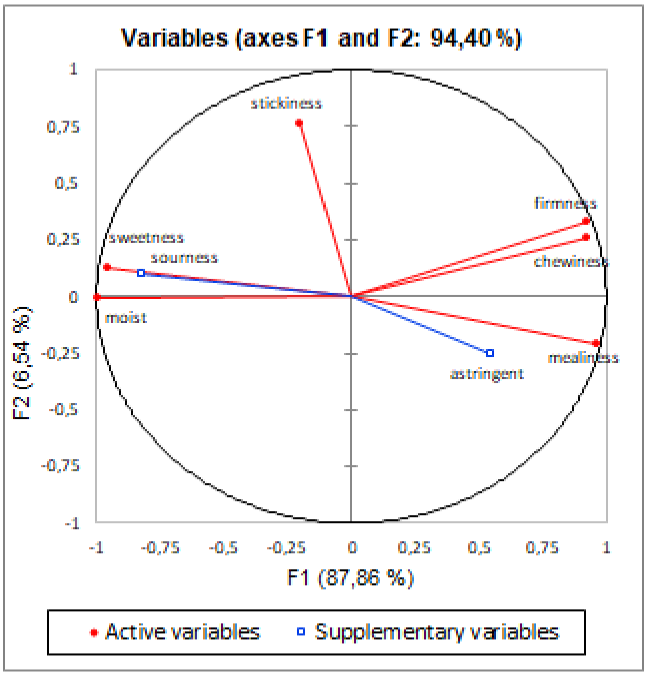
	F1	F2	F3
Eigen values	16.658	1.241	0.473
Variability (%)	87.857	6.543	2.496
Cumulative %	87.857	94.400	96.895

2.3.2 How to interpret the variables chart

The first map presented in the results is called the **correlation circle** or variable chart. It shows a projection of the initial variables in the factor space. When two variables are located far from the centre:

- If they are close to one another, they are significantly positively correlated (R close to 1). This was the case for firmness and chewiness or sweetness and moistness.
- If they are orthogonal, they are not correlated (R close to 0). This was the case for stickiness and firmness.
- If they are located on opposite sides of the centre, they are significantly negatively correlated (R close to -1). This was the case for firmness / chewiness / mealiness opposed to moistness / sweetness.

Supplementary variables can also be displayed in the shape of vectors. Curiously, sourness is close to sweetness, indicating a possible positive correlation between the two attributes.



When variables are close to the circle, it means that these variables had the most influence on the corresponding PCA axes. When the variables are close to the centre, some information is carried on other axes by looking at the correlation circle on axes F1 and F3.

To confirm that a variable is linked with an axis, look at the table of squared cosines: the higher the squared cosine, the greater the link with the corresponding axis. The closer the squared cosine of a given variable is to zero, the more careful you need to be when interpreting the results in terms of trends on the corresponding axis. Firmness, chewiness, mealiness, moistness and sweetness contributed strongly to axis 1, and stickiness to axis 2. No attribute contributed to axis 3.

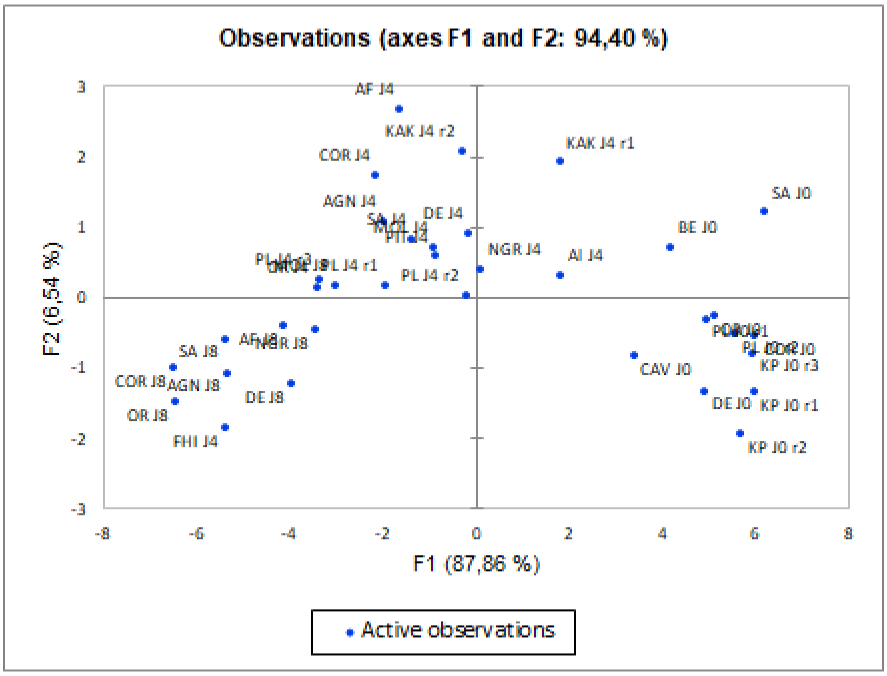
Squared cosines of the variables:

	F1	F2	F3
firmness	**0.850**	0.111	0.022
chewiness	**0.848**	0.067	0.003
stickiness	0.040	**0.587**	0.263
mealiness	**0.925**	0.045	0.001
moist	**0.983**	0.000	0.000
sweetness	**0.909**	0.016	0.052
sourness	**0.680**	0.010	0.000
astringent	**0.302**	0.065	0.071

Squared cosines reflect the representation quality of a variable on a PCA axis. Squared cosine analysis is used to avoid interpretation errors due to projection effects. If the squared cosines of a variable associated with an axis is low, the position of the variable on this axis should not be interpreted.

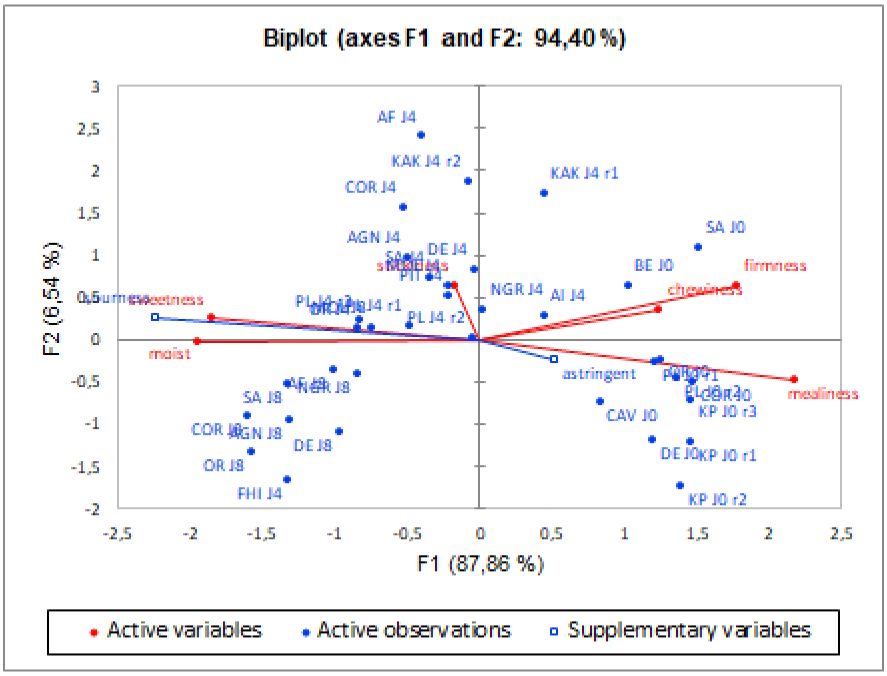
2.3.3 How to interpret the observations chart

The **observations chart** represents the observations in the PCA space. It shows you the observations on a two-dimensional map, allowing you to identify trends.



Axis 1 clearly separated the varieties according to their ripening stage (J0 – J4 – J8). The more mature the plantains, the less firm, and the mealier the associated boiled products, but the sweeter, moister and chewier the boiled products. Boiled plantains were perceived to be the stickiest at the intermediate stage of ripening (J4). The variability among the three ripening stages was greater than the variability between cultivars. You can see that the replicates of the samples are close (for KP J0, PL J0, PL J4, and KAK J4), suggesting similar products and good panel performance in terms of repeatability. Some varieties like FHI (plantain hybrid) had the sensory characteristics of a mature plantain (J8) at the intermediate ripening stage (J4).

It is also possible to show biplots, which are simultaneous representations of variables and observations in the PCA space. Biplots can be useful when giving a lecture or in written articles to show the sensory characteristics of all the products in a single graph.



3 LINEAR REGRESSIONS

One of the objectives of the RTBfoods project is to provide easy-to-use and rapid tools to assess the sensory quality of products. It is clear that organising sensory analyses with a trained panel (QDA) is complicated as it requires organising the sessions, defining number of samples that have to be analysed, keeping the panellists motivated and getting them to perform well over time, etc. To help avoid these problems, we want to develop instrumental methods which could replace the QDA and that are easier to use and more reliable over time. But to do so, we need the most reliable and relevant models to predict sensory quality.

Our objective here is to use multiple linear regression to predict sensory attributes from instrumental parameters. As you will see, this goes beyond simply relating variables.

3.1 Definitions - principles

The principle of linear regression is modelling a quantitative dependent variable Y (in our example a sensory attribute) using a linear combination of p quantitative explanatory variables X1, X2, X3, ... (here, biophysical parameters). For observation i, the model is written as follows:

$Y_i = a_0 + a_1X_{i,1} + a_2X_{i,2} + \ldots..a_pX_{i,p} + \varepsilon_i,$

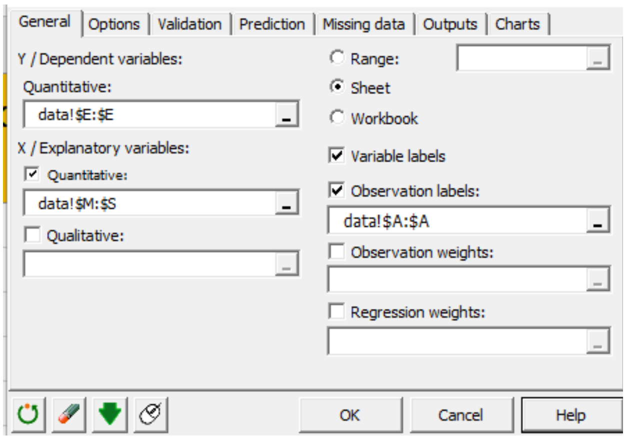
where Y_i is the value observed for the dependent variable for observation i, $X_{i,j}$ is the value taken by variable j for observation i, and ε_i is the error of the model.

Before building and interpreting the model, we need to mention several important concepts.

In our example, we will predict the sensory attributes that describe boiled plantain (used in Antonin Kouassi's thesis) among the 7 instrumental texture parameters (Puncture force, Hardness_TPA, etc.) and the 3 biochemical parameters (dry matter content, soluble solids content, and titratable acidity).

3.2 Setting up a linear regression by using XLSTAT

- Open the file « Example ACP + regression_EN.xls».
- Once XLSTAT is activated, select **Modeling data / Linear regression**.
- In the window "General", click on "Y / dependent variables" and select firmness (column E) in the "data" sheet.
- Click on "X / Explanatory variables" and select textural parameters (column M - S) or biochemical parameters (columns T – V) in the "data" sheet.
- Click on « Observation labels » and select column A in the "data" sheet.
- Activate the option "sheet" to display the results in a new worksheet.

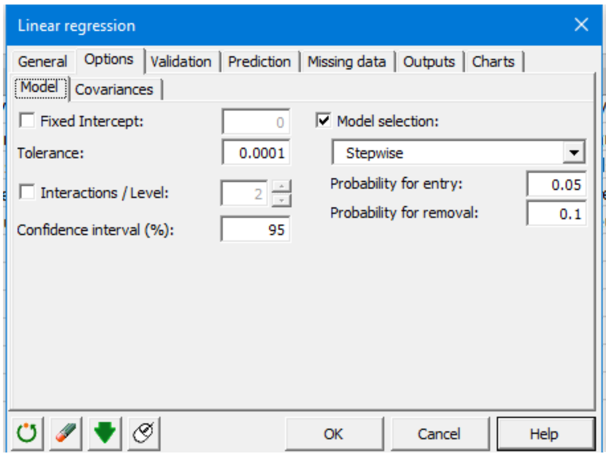


Y / Dependent variables:
Quantitative:
data!$E:$E

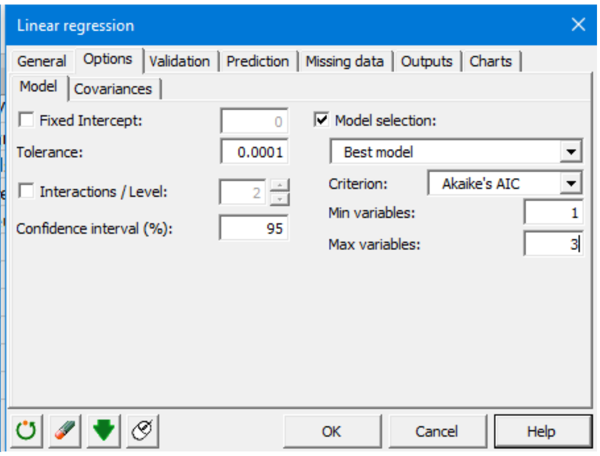
X / Explanatory variables:
☑ Quantitative:
data!$M:$S
☐ Qualitative:

○ Range:
● Sheet
○ Workbook
☑ Variable labels
☑ Observation labels:
data!$A:$A
☐ Observation weights:
☐ Regression weights:

- In the window "Option", default tolerance is set at 0.0001 and the confidence interval at 95%. Among the proposed models (best model, stepwise, forward and backward), we propose 2 options: Stepwise and best model.
- If you choose the Stepwise model, set a probability of 0.05 for entering a variable and 0.1 for removing a variable. If you choose the best model, set 1 minimum variable, 3 maximum variables, and the AIC index to determine the best model.



☐ Fixed Intercept: 0
Tolerance: 0.0001
☐ Interactions / Level: 2
Confidence interval (%): 95
☑ Model selection: Stepwise
Probability for entry: 0.05
Probability for removal: 0.1



☐ Fixed Intercept: 0
Tolerance: 0.0001
☐ Interactions / Level: 2
Confidence interval (%): 95
☑ Model selection: Best model
Criterion: Akaike's AIC
Min variables: 1
Max variables: 3

Details on model selection

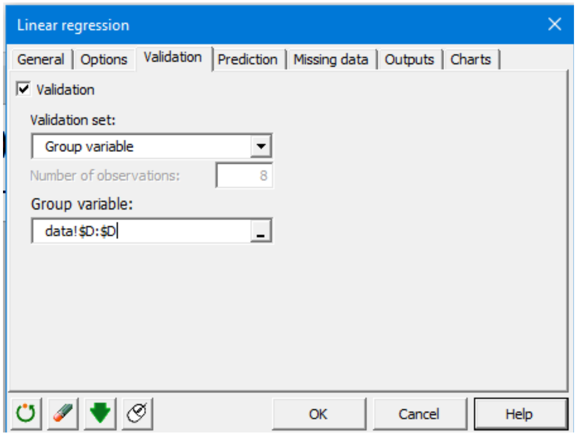
Stepwise: The selection process starts by adding the variable with the largest contribution to the model (the criterion used is Student's t statistic). If a second variable is such that the probability associated with its t is less than the "**Probability for entry** ", it is added to the model. The same goes for a third variable. After the third variable is added, the impact of removing each variable in the model after it has been added is evaluated (still using the t statistic). If the probability is greater than the "**Probability of removal** ", the variable is removed. The procedure continues until no more variables can be added or removed.

Forward: The procedure is the same as for stepwise selection except that variables are only added and never removed.

Backward: The procedure starts by simultaneously adding all variables. The variables are then removed from the model following the procedure used for stepwise selection.

Best model: This method lets you choose the best model among all the models that can handle a number of variables varying from "Min variables" to "Max Variables". In addition, you can choose several "criteria" to determine the best model. We generally use the criterion Akaike's AIC: the weaker the criterion, the better the model (it can take negative values).

- In the "Validation" window, select "random" or "Group variables" in « validation » command. Then define the number of observations you will have in your validation dataset.
  - When you choose the "group variables" option, you then have to select an indicator variable 0 for the calibration observations, and 1 for the validation observations. In our example, column D indicates our calibration and validation observations. In our example, we decided that the 27 samples of plantains from Côte d'Ivoire would be the calibration observations and the 8 samples of bananas from other countries would be the validation observations, but we could also have used the samples from the last two tasting sessions as our validation set.



☑ Validation

Validation set: Group variable

Number of observations: 8

Group variable: data!$D:$D

  - By choosing the "random" option, the observations are selected randomly. If you run the linear regression again with this option, the new validation observations will obviously be different and the results will also differ slightly. Using this option, it is consequently difficult to decide on the final model.

Why is it important to have calibration and validation data?

For the prediction to be robust, the model to be implemented needs to be validated. The objective is to prove that the model produces good estimates of the values of the variable under study. To do so, you need at least one calibration data set (also called training set in XLSTAT), and one validation data set. To put it simply, the calibration data are used to build the model and the validation data are used to show that the model is reliable/robust.

Independence of calibration and validation data - percentage of calibration / validation data

To be as objective as possible, i.e., to be sure the result is not biased, the calibration and validation data sets should come from independent populations. What is generally done is to divide an initial baseline sample set into a set of calibration data and a set of validation data. If you can collect two different datasets, one for calibration and one for model validation, this is warmly recommended! It is often recommended to use between 60 and 80% of the initial data set as a calibration set and the remaining 20 to 40% as a validation set, but these percentages are not fixed.

Note: in our example, 27 observations were used for calibration and 8 observations for validation, i.e., 23% of the total observations.

- In the window « missing data », click on “remove the observations », then « check for each Y separately ». The observation will be deleted even if the missing data correspond to an explanatory variable that was not taken into account by the model. You will then have to redo the linear regression by removing this variable so that all observations are taken into account.
- In the window « Outputs », click on “descriptive statistics”, “correlations”, “multicollinearity statistics”, “analysis of variance”, « Type I/III SS », « standardized coefficients », « Predictions and residuals », « confidence intervals ».
- In the window « test assumptions », click on « normality test ».
- In the window « charts », click on all options.

The computations begin once you have clicked on OK.

3.3 Interpreting results of linear regression

The results on the prediction of firmness from the textural measurements are presented in different tabs. In the sheets LR0 to LR2, the validation set was chosen from the option "group variables".

3.3.1 Descriptive statistics

The descriptive statistics make it possible to visualize information on the data from the calibration and validation sets, to compare the values (mean, standard deviation, etc.) between calibration and validation data, and to identify outliers.

The correlation matrix makes it possible to identify simple correlations between the sensory descriptor and the explanatory variables, as well as between explanatory variables.

3.3.2 How to interpret multicollinearity statistics

After reviewing the descriptive statistics and the correlation matrix, it is important to focus on the multicollinearity statistics.

What is multicollinearity?

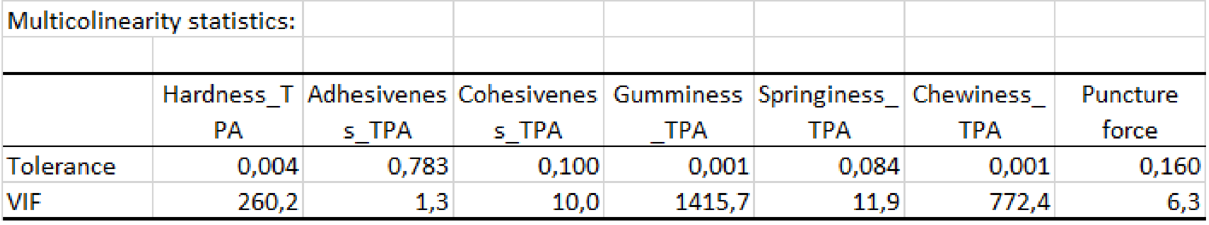
Collinearity refers to a linear relationship between 2 explanatory variables (R > 0.90 indicates high multicollinearity, although it is not an absolute rule). Collinearity creates several problems:

- the values/signs of the coefficients are contradictory, they do not agree with the knowledge of the domain

- the results are very unstable; the addition or deletion of a few observations alters the values and signs of the coefficients.

The variance inflation factor (VIF) is an indicator of collinearity. It is accepted that a VIF > 7 indicates a problematic amount of collinearity.

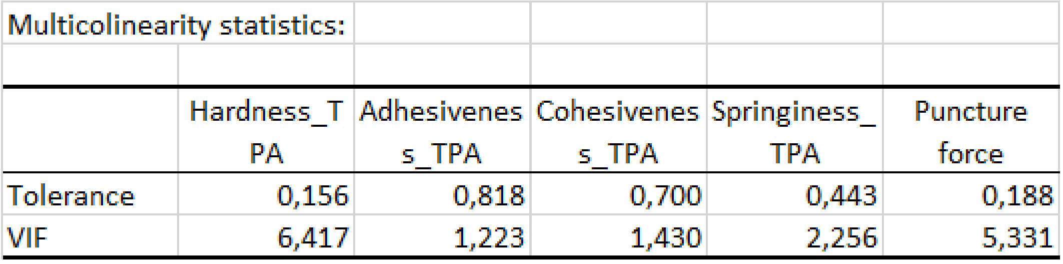
In our example, a very high VIF was found for the following variables: hardness, gumminess and chewiness (sheet LR0). The correlation matrix shows that these variables are very strongly correlated (R > 0.95). This is understandable since gumminess is the product of hardness and cohesiveness and chewiness is the product of gumminess and springiness.



Multicolinearity statistics:

	Hardness_TPA	Adhesiveness_TPA	Cohesiveness_TPA	Gumminess_TPA	Springiness_TPA	Chewiness_TPA	Puncture force
Tolerance	0,004	0,783	0,100	0,001	0,084	0,001	0,160
VIF	260,2	1,3	10,0	1415,7	11,9	772,4	6,3

By rerunning the linear regression without gumminess and chewiness variables, the VIFs were less than 7, indicating the absence of collinearity (LR1 sheet).



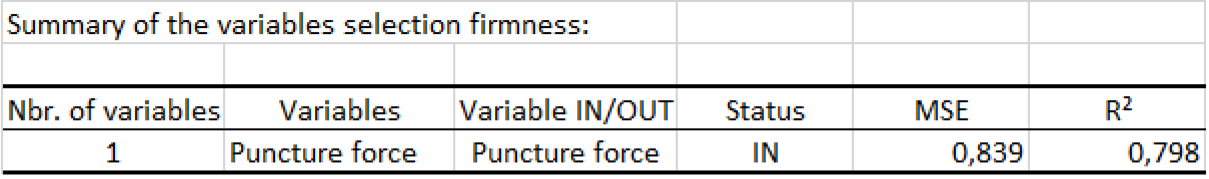
Multicolinearity statistics:

	Hardness_TPA	Adhesiveness_TPA	Cohesiveness_TPA	Springiness_TPA	Puncture force
Tolerance	0,156	0,818	0,700	0,443	0,188
VIF	6,417	1,223	1,430	2,256	5,331

In the following, the results are those obtained from these 5 explanatory variables.

In the following tables, XLSTAT summarises the selection of variables for the model.

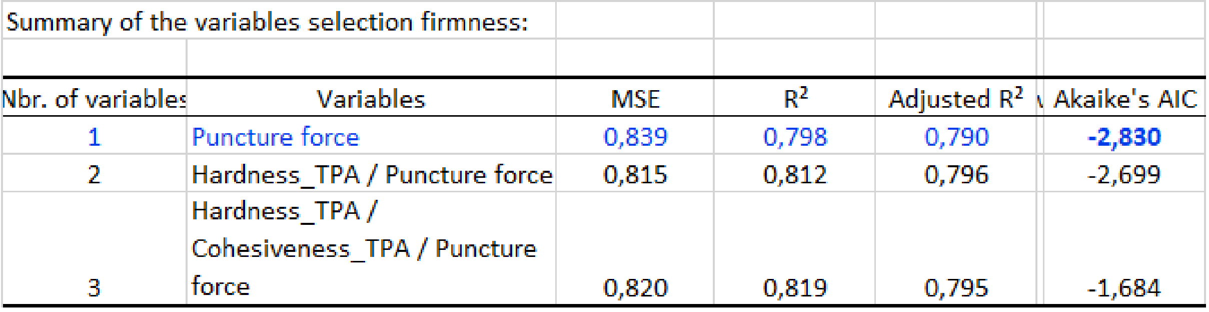
In LR1 sheet which offers Stepwise selection of the variables, only one variable has been selected.



Summary of the variables selection firmness:

Nbr. of variables	Variables	Variable IN/OUT	Status	MSE	R²
1	Puncture force	Puncture force	IN	0,839	0,798

In the LR2 sheet which offers a selection of the variables made in the best model (AIC), only one variable has been selected. Note that 2 other models with 2 and 3 variables are proposed which show a slightly higher R², but because the AIC for the first model is lower (with only one variable), this variable was selected. In any case, the 2 models with 2 and 3 variables do not improve the prediction very much: as proof, the adjusted R² in the models with 2 or 3 variables is almost equal to that in the model with only one variable (the adjusted R² is a correction of the R² which accounts for the number of variables used in the model).

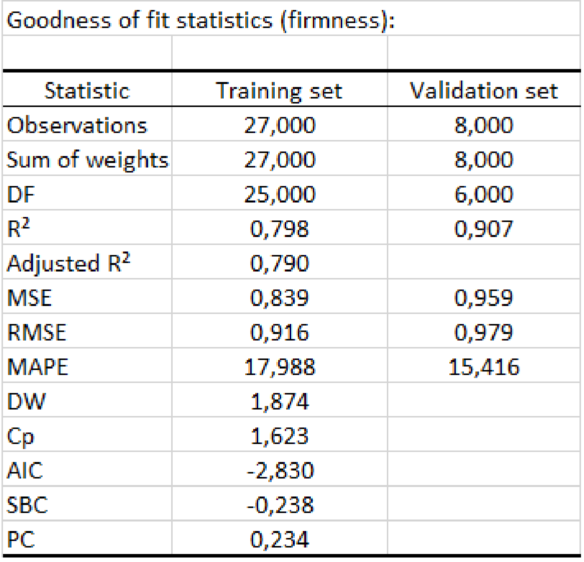


Summary of the variables selection firmness:

Nbr. of variables	Variables	MSE	R²	Adjusted R²	Akaike's AIC
1	Puncture force	0,839	0,798	0,790	**-2,830**
2	Hardness_TPA / Puncture force	0,815	0,812	0,796	-2,699
3	Hardness_TPA / Cohesiveness_TPA / Puncture force	0,820	0,819	0,795	-1,684

The above example shows that the selection of variables using Stepwise or the best model (AIC) leads to the same result: a prediction of firmness by Puncture force with a coefficient of determination (R²) of 0.798. **This R² means that 79.8% of firmness variability is explained by puncture force. This coefficient is one of the 3 important criteria (R², RMSEC, RMSEV), you will have to calculate the accuracy and robustness of the prediction.**

The table "Goodness of fit statistics (firmness)" shows the statistics related to the fit of the regression model. You can look up the meaning of the coefficients in the Help directory by clicking on the "Help" icon in the general window.

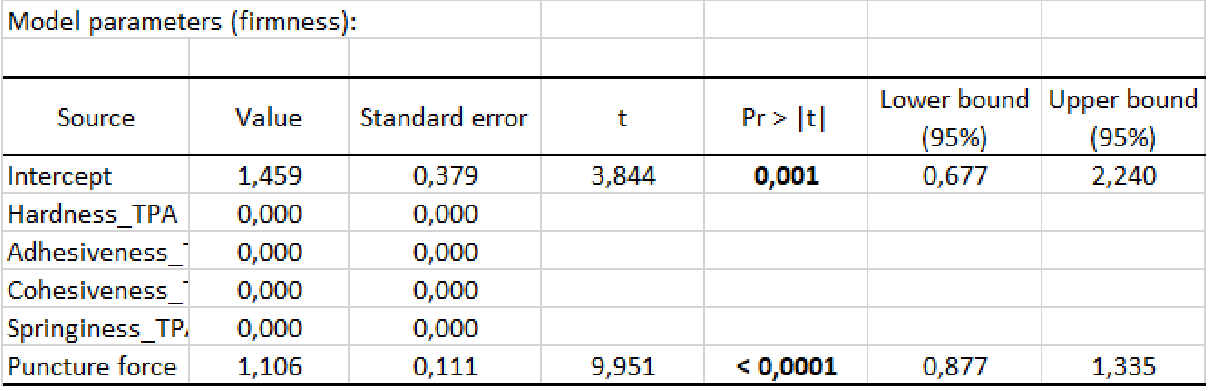


Goodness of fit statistics (firmness):

Statistic	Training set	Validation set
Observations	27,000	8,000
Sum of weights	27,000	8,000
DF	25,000	6,000
R²	0,798	0,907
Adjusted R²	0,790	
MSE	0,839	0,959
RMSE	0,916	0,979
MAPE	17,988	15,416
DW	1,874	
Cp	1,623	
AIC	-2,830	
SBC	-0,238	
PC	0,234	

The **analysis of variance table** is used to evaluate the explanatory power of the explanatory variables. Since the Type I/III SS (SS: Sum of Squares) option has been activated, the Type I SS and Type III SS tables are displayed. You only need to look at the Type III SS table where the order of selection of the variables in the model does not influence the values obtained. The lower the probability, the larger the contribution of the variable to the model, all the other variables already being in the model. As here there is only one variable, there is no difference between Type I and III SS.

The model parameters are shown. The model equation is then given with the significant parameters of the model (in bold in the table).

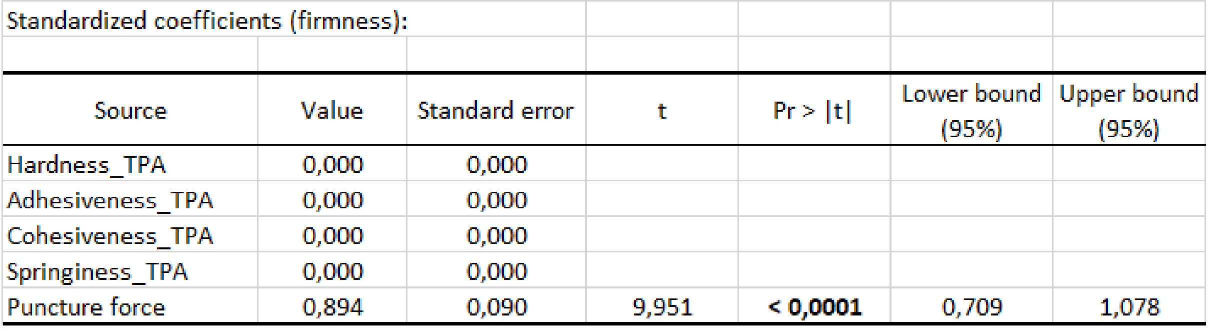


Model parameters (firmness):

Source	Value	Standard error	t	Pr > |t|	Lower bound (95%)	Upper bound (95%)
Intercept	1,459	0,379	3,844	**0,001**	0,677	2,240
Hardness_TPA	0,000	0,000				
Adhesiveness_	0,000	0,000				
Cohesiveness_	0,000	0,000				
Springiness_TP	0,000	0,000				
Puncture force	1,106	0,111	9,951	**< 0,0001**	0,877	1,335

Firmness is predicted by puncture force according to the following equation: Firmness = 1.459 + 1.106*Puncture force.

The table of standardised coefficients makes it possible to compare the relative weight of the variables. The higher the absolute value of a coefficient, the greater the weight of the corresponding variable. In the case of a multivariate model, the table below shows which variable has the greatest influence on the variable to be explained.



Standardized coefficients (firmness):

Source	Value	Standard error	t	Pr > |t|	Lower bound (95%)	Upper bound (95%)
Hardness_TPA	0,000	0,000				
Adhesiveness_TPA	0,000	0,000				
Cohesiveness_TPA	0,000	0,000				
Springiness_TPA	0,000	0,000				
Puncture force	0,894	0,090	9,951	**< 0,0001**	0,709	1,078

In the table, predictions and residuals are provided for each observation, in particular for the observed and the predicted value, as well as the difference between the 2 (called residual).

Using this table, **you will be able to compute two very important criteria for the prediction, the root mean square error (RMSE) on the calibration dataset (RMSEC) and on the validation dataset (RMSEV).**

What is the root mean square error (RMSE)?

This criterion shows the dispersion of the observed values compared to the predicted values. It can be related to the variance of the model.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (Yi_p - Yi_o)^2}{n}}$$

where Yip and Yio are respectively, the values predicted and observed for observation i and n is the number of observations.

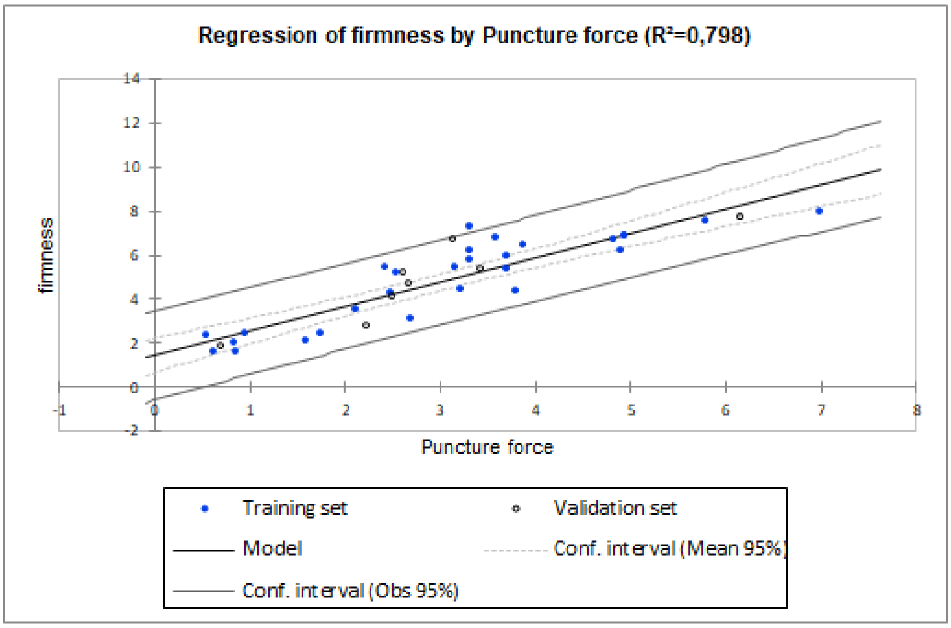
The unit of this criterion is that of the variable to be explained. In the case of sensory attributes, the RMSE expresses the dispersion of the observed values compared to the predicted values on a scale of 0 to 10. An RMSE close to 1 means that the model is able to predict a sensory attribute with approximately 1 point of deviation on a scale of 0 to 10, which is generally a good prediction because it is better than the level of performance repeatability of a well-trained panel (between 1 and 2).

RMSEC was calculated using the 27 observations in the calibration dataset and RMSEV using the 8 observations in validation dataset according to the formula below.

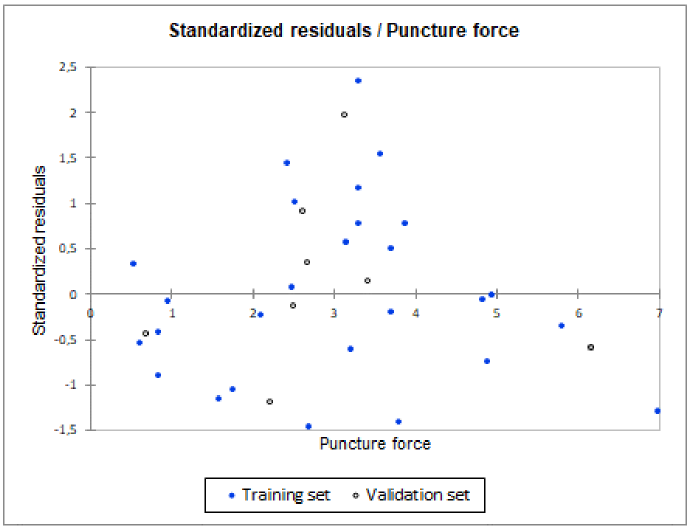
RMSEC = 0.88 (cell O157) and RMSEV = 0.85 (cell O193).

These values are lower than 1, meaning the prediction is accurate (with R² of 0.798 and RMSEC of 0.88) and robust (with RMSEV of 0.85).

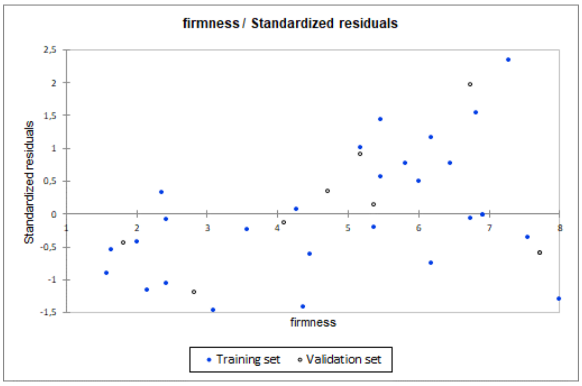
The following graph shows the above-mentioned results. Since there is only one explanatory variable in the model, the first graph displayed shows the observed values, the regression line and the two types of confidence intervals around the predictions.

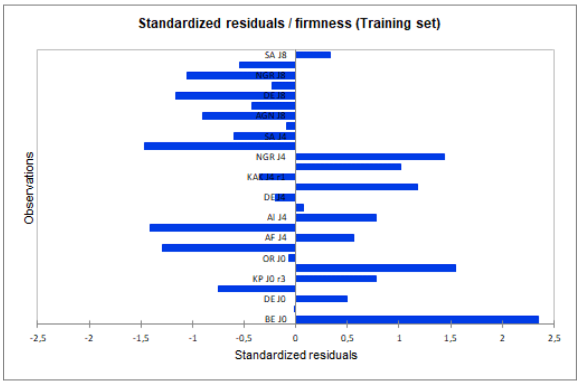


The second graph shows the standardised residuals as a function of the explanatory variable. In theory, the residuals should be randomly distributed around the x-axis. A trend or a shape rather than random distribution, would reveal a problem in the model, namely heteroscedasticity (or funnel shape).

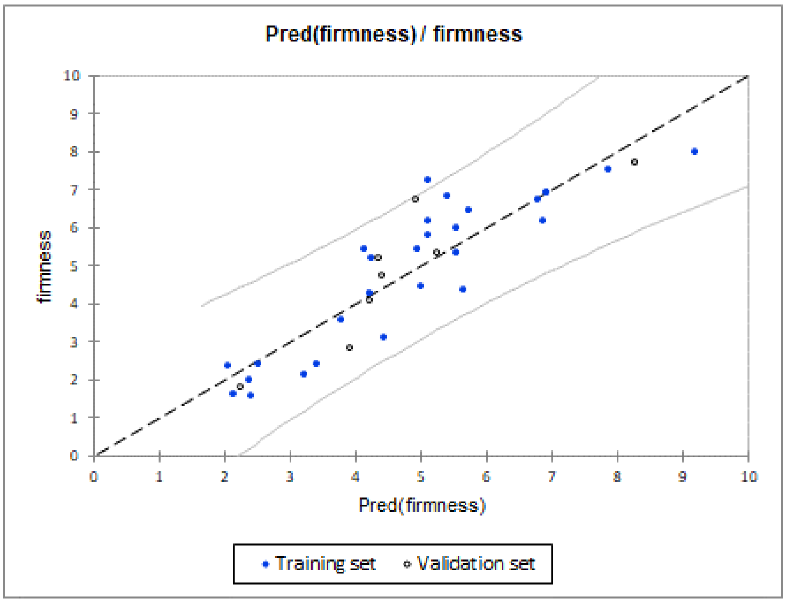


The third graph shows changes in the normalized residuals as a function of the dependent variable, the distance between the predictions and the observations (in an ideal model, the points would all be on the bisector), and the normalized residuals in the form of a bar graph (the fourth graph) for the calibration set and the validation set. The bar graph makes it possible to quickly see if an abnormal number of data fall outside the interval [-2, 2], knowing that, under the assumption of normality, this interval should contain about 95% of the data.

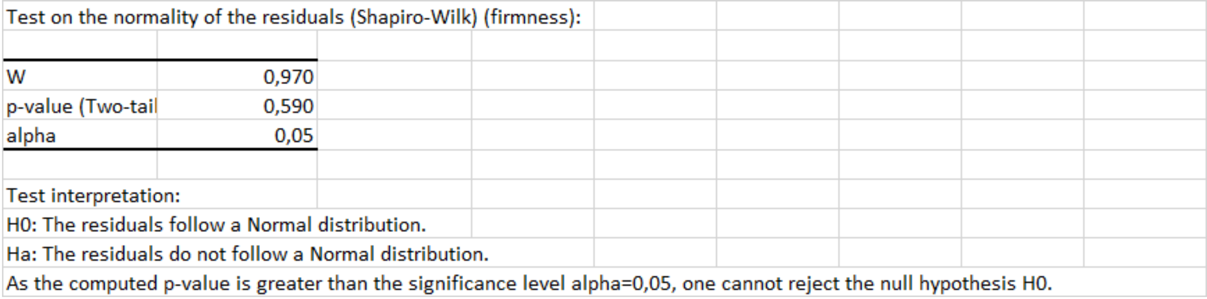




The fifth graph below shows observed versus predicted firmness with the confidence interval. Usually the opposite is presented in articles or in lectures, i.e., predicted firmness according to observed firmness. To do the same, by clicking on the dots in the figure, you can change the letters in the series to show the predicted firmness on the ordinate and the observed firmness on the abscissa.



Finally, there should be a test of normality of the residuals. This is important because if it is negative (i.e., the hypothesis of normality of the residuals is rejected), it means that the residuals do not follow a normal distribution, and therefore that the prediction does not make sense. Here the p-value is greater than 0.05, meaning the residuals follow a normal distribution and therefore the prediction is acceptable.



Test on the normality of the residuals (Shapiro-Wilk) (firmness):

W	0,970
p-value (Two-tail	0,590
alpha	0,05

Test interpretation:

H0: The residuals follow a Normal distribution.

Ha: The residuals do not follow a Normal distribution.

As the computed p-value is greater than the significance level alpha=0,05, one cannot reject the null hypothesis H0.

3.4 Some questions

3.4.1 Can one manage without validation?

In many published works, the relationships between sensory descriptors and instrumental measures are established using only one data set with no validation.

We did the same for the variable Firmness (LR3 sheet) based on the 35 observations with no validation data set. Firmness is predicted by both puncture force and hardness with an R² of 0.83 and an RMSE of 0.79. These results show that firmness can be predicted by 2 complementary texture variables (penetrometry and a compression test) with an error of less than 0.80 points on a scale of 0 to 10. This result is better than the one found previously with 27 observations (R² = 0.80, RMSEC = 0.88) but requires an additional measure (hardness measured with a compression test). Note that the prediction was only slightly improved (R² changed from 0.80 to 0.83, RMSE from 0.88 to 0.80) despite the additional variable. The absence of a validation set prevents us from concluding on the robustness of the prediction. To validate the prediction, we would need another data set using new varieties or varieties studied previously in a different production season (of course judged by the same panel and using exactly the same conditions).

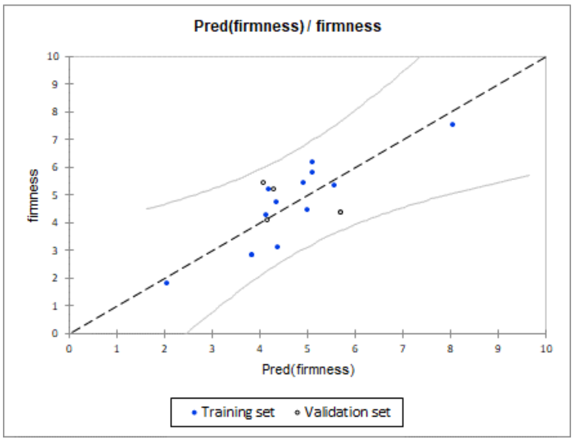
3.4.2 How many observations are needed to make the predictions?

First, no statistical tool is available to define the minimum number of observations required to make predictions. We can thus only give advice.

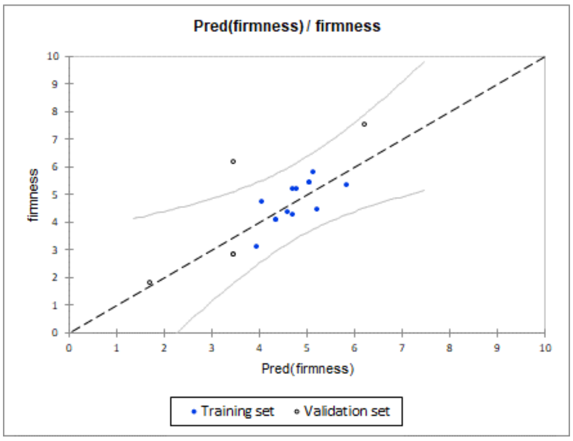
1- Given the heavy workload involved in sensory analysis, it is very difficult to exceed 50 observations (which would require at least 10 sessions with the same panel), contrary to instrumental measurements (more than 100 to predict chemical parameters from spectral measurements).

2- We recommend observations that are representative of the whole sensory diversity, i.e., observations that cover the range of intensity scales for the major attributes. Below we give 2 examples of firmness built with only 16 observations extracted from plantains at the J4 ripening stage. In both cases, 12 observations were chosen for calibration and 4 for validation. In the first case (sheet LR4), the calibration observations covered a wide range of firmness (from 1.8 to 7.5), whereas in the second case (sheet LR5), the calibration observations covered a smaller range of firmness (from 3.0 to 5.8). The table below and the figures summarise the prediction results.

Calibration data	firmness	Model equation	R^2	RMSEC	RMSEV
LR4	1.8 – 7.5	firmness = 1.2 + 1.2*Pf	0.77	0.73	1.06
LR5	3.0 – 5.8	firmness = 0.9 + 8.0*Hardness	0.47	0.54	1.55
LR1	*1.6 – 8.0*	*Firmness = 1.4 + 1.1*Pf*	*0.80*	*0.85*	*0.88*



With LR4 dataset



with LR5 dataset

The prediction obtained with LR4 data is very close to the prediction obtained with all the data (LR1), even though the number of observations was 16 versus 35 for LR1. The fact of having firmness values that cover the whole intensity scale enabled good prediction despite the limited number of observations. On the other hand, the prediction of firmness with LR5 data was less accurate and differed from previous ones. This is because the observations used for the calibration only partially covered the intensity scale. Note that one observation in the validation set (KAK J4 r2) was poorly predicted by the LR5 equation, which calls the robustness of this prediction into question (and hence the relevance of having a set of observations for validation, see first question "can we manage without validation? ").

3. The accuracy and robustness of a prediction depends not only on the number of observations but also on the nature of the links between the variable to be explained and the explanatory variables. In our example, stickiness was very poorly predicted by the textural measurements, whatever the number of observations. It is clear that other types of instrumental measures are needed to identify predictors of this attribute.

Institute: Cirad – UMR QualiSud

Address: C/O Cathy Méjean, TA-B95/15 - 73 rue Jean-François Breton - 34398 Montpellier Cedex 5 - France

Tel: +33 4 67 61 44 31

Email: rtbfoodspmu@cirad.fr

Website: https://rtbfoods.cirad.fr/