Calibration Report

Cassava Cooking Properties Characterization using NIRS on Fresh Ground Cassava

High-Throughput Phenotyping Protocols (HTPP), WP3

Saint Pierre, La Réunion, France, 03/11/2022

Fabrice DAVRIEUX, Centre de coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), Saint Pierre, La Réunion, France

Xiaofei ZHANG, Alliance of Bioversity International & CIAT, Cali, Colombia

Luis LONDOÑO, International Center for Tropical Agriculture (CIAT), Cali, Colombia

John BELALCAZAR, CIAT, Cali, Colombia

Thierry TRAN, CIAT/CIRAD, Cali, Colombia

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

To be cited as:

Fabrice DAVRIEUX, Xiaofei ZHANG, Luis LONDOÑO, John BELALCAZAR, Thierry TRAN, (2023). *Cassava Cooking Properties Characterization using NIRS on Fresh Ground Cassava; High-Throughput Phenotyping Protocols (HTPP), WP3.* St Pierre, La Réunion, France: RTBfoods Calibration Report, 27 p. https://doi.org/10.18167/agritrop/00732

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:	
Fabrice DAVRIEUX (CIRAD) Thierry TRAN (CIRAD) Fabrice DAVRIEUX (CIRAD)	25/10/2022 15/11/2022 19/11/2022
Final validation by:	
Fabrice DAVRIEUX (CIRAD)	13/12/2022

CONTENTS

Table of Contents

1 Data 7

1.1 Material 7

1.2 Water absorption at 30 minutes of boiling 7

1.3 Near infrared spectroscopy 10

1.3.1 Exploration 10

1.3.2 Principal Components Analysis 10

2 Modelling 11

2.1 Learning and test sets 11

2.2 Indirect classification using spectra 13

2.2.1 Classification based on Ridge regression 13

2.3 Direct classification using spectra 16

2.3.1 Classification using PLSRDA 16

2.4 Direct vs indirect classification 19

2.4.1 External validation using years 19

3 Conclusion 22

4 Perspectives 22

5 Bibliographie 23

6 Appendices 24

6.1 Annex I: Genotypes analysed each year 24

6.2 Annex II: Repeated clones (10) within the false positives (Ridge Regression) 25

6.3 Annex III: List of the field trials used to produce the dataset of WA30 and NIRS in the present report 26

Table of Figures

Figure 1 : Distribution of WA30 values for the 2905 samples (left) and distribution of WA30 values per years (right) 8

Figure 2 : WA30 values for the 22 genotypes analyzed n times over 4 years 9

Figure 3 : visible and Near infrared spectra of the whole population (2905 spectra) 10

Figure 4 : Scatter plot of the PCA scores of the 2905 samples for PC_1 and PC_2; PCA on raw full spectral data. 10

Figure 5 : Scatter plot of the PCA scores of the 2905 samples for PC_1 and PC_2; PCA on raw NIR segment spectral data 11

Figure 6 : WA30 average values per set and per year 12

Figure 7 : Projection of test samples (872) on PC_1 and PC_2 of the PCA calculated on training samples (2033) 12

Figure 8 : Scatter plots of PCA scores for PC_1 and PC_2 for training set samples (left) and test samples (right), colored by classes, C1 WA30 ≤12% and C2 WA30 > 12% 13

Figure 9 : Scatter plot of WA30 measured versus predicted values (A) and scatter plot of residuals vs WA30 measured values (B) 14

Figure 10 : 10 most VIP of the Ridge regression model 14

Figure 11 :WA30 residuals distribution per class 15

Figure 12 : Confusion matrix graphic 15

Figure 13 Predicted and actual WA30 values for the 131 false positives in test set. Limit of classes = 12% 16

Figure 14 : Evolution of the classification error vs number of latent variables 17

Figure 15 : Confusion matrix graphic 18

Figure 16 : Repartition of the WA30 values of the 87 false positives from PLSRDA 18

Figure 17 : 2022 and 219-2021 Spectra (correction SNV and first derivative) 20

Figure 18 : Projections of 2022 samples onto PC_1 and PC_2 of PCA calculated on 2019-2021 samples 20

Figure 19: Projections of 2019-2021 samples onto PC_1 and PC_2 of PCA calculated on 2022 samples 20

List of Tables

Table 1 : Number of genotypes (solely this year) and samples analyzed per year 7

Table 2 : Descriptive statistics for Water absorption (WA30) values 7

Table 3 : Descriptive statistics for WA30 according to classes 8

Table 4 : Mean and standard deviation values of WA30 per year and class 8

Table 5 : Number of replicates, average and SD WA30 values and coefficient of variation 9

Table 6 : Classification rates for the 22 genotypes according to WA30 values measured 9

Table 7 : Repartition of samples per set per class per year 11

Table 8 : Repartition of samples per set per year 12

ABSTRACT

Context: This scientific report concerns data analysis of two matrices of measured data on fresh cassava 1) water absorption capacity (WA30) measured after 30 min of boiling and 2) spectral data. The data were collected on fresh ground cassava in CIAT, Colombia.

Place: Réunion and Colombia

Date: 28/10/2022

Authors: Fabrice DAVRIEUX (CIRAD), Luis LONDOÑO (CIAT), Thierry TRAN (CIAT)

Content:

This study concerns 1101 genotypes from several breeding populations harvested and analyzed at CIAT (Colombia) over 4 years between 2019 and 2022. The near-infrared spectra of the mashed fresh roots of all the genotypes were recorded according to the SOP (Belalcazar John, 2020) developed within the RTBfoods project. Water absorption at 30 minutes boiling was measured according to the SOP developed by CIAT (Escobar Salamanca Andrés Felipe, 2022).

The distribution of the WA30 values is asymmetric on the left, 1692 samples (58%) presented WA30 values lower than 12%, and 1213 samples (42%) presented WA30 >12%. The dispersion of the data increases with year, the WA30 standard deviations varies from 5,1% (2019) to 10,4% (2022), the variability of WA30 values increases by a factor 2. This may reflect the increasing diversity of the populations screened by the CIAT, with more progeny populations being screened by WA30.

The WA30 value is inversely correlated to cooking time, therefore samples with high values of WA30 correspond to candidate genotypes with low cooking time (CT) as well as softer, more mealy texture. A limit between low and long cooking time genotypes can be arbitrarily set at 12% for WA30, corresponding to ~35 minutes cooking. Samples lower or equal to 12% of WA30 can be considered as poor genotype (too long CT) and samples with WA30 values higher than 12% correspond to suitable genotypes for end user with a low CT.

Two sets of data were constituted: one for learning, tuning the parameters, and one for testing, evaluating the error of the predictive model. To do this, 70% (n = 2033) of the samples were randomly selected for learning set and the remaining 30% (n = 872) were used for testing. The repartitions per WA30 classes and year were maintained within the two sets.

Two modelling strategies were investigated: 1) an indirect classification based on a regression step to predict WA30 using spectral fingerprints and then class attribution according to WA30 predicted values; 2) a direct classification using discriminant procedure based on classes defined by WA30 laboratory values and spectral fingerprints.

The two approaches: regression (Ridge Regression) and classification (PLSRDA), based on different methods for regressor selection within the spectral data, lead to similar performances in terms of classification according to WA30 classes. Nevertheless, PLSRDA leads to better classification and is easier to implement and interpret. The classification accuracy is 81,4% when predicting test set with a sensitivity = 79,4% and a specificity of 82,9% and a false positive rate equal to 17,1% while false negative rate is 20,6%.

This model is efficient and can be implemented in a selection scheme 1) as is if the next year/generation variability remains similar to the current database 2) or with controlled update if next year/generation variability differs from current database (e.g. if some samples have WA30 values higher than the range already in the database).

Keywords: NIRS, cassava roots, water absorption, PLSR-DA, cooking ability

1 DATA

1.1 Material

This study concerns 1101 genotypes from several breeding populations harvested and analyzed at CIAT (Colombia) over 4 years between December 2019 and July 2022 (table 1) (one population was harvested at Momil in Cordoba department). The near-infrared spectra of the mashed fresh roots of all the genotypes were recorded according to the SOP (Belalcazar John, 2020) developed within the RTBfoods project. Water absorption at 30 minutes boiling was measured according to the SOP developed by CIAT (Escobar Salamanca Andrés Felipe, 2022).

Some of the genotypes were analyzed more than once over the 4 years, in these cases the year of harvest and the site of origin were different. Thus, 582 genotypes were analyzed one time and for the 519 remaining the number of replicates varies between 2 (177) and 68 (1: genotype Ven25). 22 genotypes were analyzed each year (392 spectra) (**Annex I**). None of the replicates by clone were harvested at the same date;

Table 1 : Number of genotypes (solely this year) and samples analyzed per year

Harvest Year	2019	2020	2021	2022	
N (genotypes)	35 (2)	49 (4)	821 (368)	716 (268)	Total
N samples	35	89	1642	1139	2905

1.2 Water absorption at 30 minutes of boiling

The water absorption capacity measurements were done following the SOP developed in CIAT, Columbia (Escobar Salamanca Andrés Felipe, 2022) and the results were expressed as % change in weight of the raw fresh roots: WA30 (%). The overall average value was 11,9%

Table 2 : Descriptive statistics for Water absorption (WA30) values

	Year	N	Minimum	Maximum	Mean	SD
WA30	overall	2905	-3.90*	52.83	11.90	9.97
	2019	35	0.04	20.56	10.71	5.12
	2020	89	-0.22	30.43	11.53	6.54
	2021	1642	-3.19	52.83	10.57	9.69
	2022	1139	-3.90	50.88	13.88	10.38

*The negatives values correspond to roots which loose material during boiling, especially for long time cooking root.

The distribution of the WA30 values is asymmetric on the left (fig. 1) with 1692 samples (58%) of the set) presented WA30 values lower than 12%, and 1213 samples (42%) with WA30 >12%.

The dispersion of the data increases with year (tab.1 and fig.1), the standard deviations varies from 5,1% (2019) to 10,4% (2022), the variability of WA30 values increases by a factor 2. This may reflect the increasing diversity of the populations screened by the CIAT team using the WA30 method, with more progeny populations being integrated in the screening.

The WA30 value is inversely correlated to cooking time (Thierry Tran, 2021), therefore samples with high values of WA30 correspond to candidate genotypes with low cooking time (CT) as well as softer, more mealy texture. A limit between low and long cooking time genotypes can be arbitrarily set at 12% for WA30. Samples lower or equal to 12% of WA30 can be considered as poor genotype (too long CT) and samples with WA30 values higher than 12% correspond to suitable genotypes for end user with a low CT (lower than 35 min; equation with 2020-2021 dataset: Cooking time (min) = -1.46 x WA30(%) + 53.98).

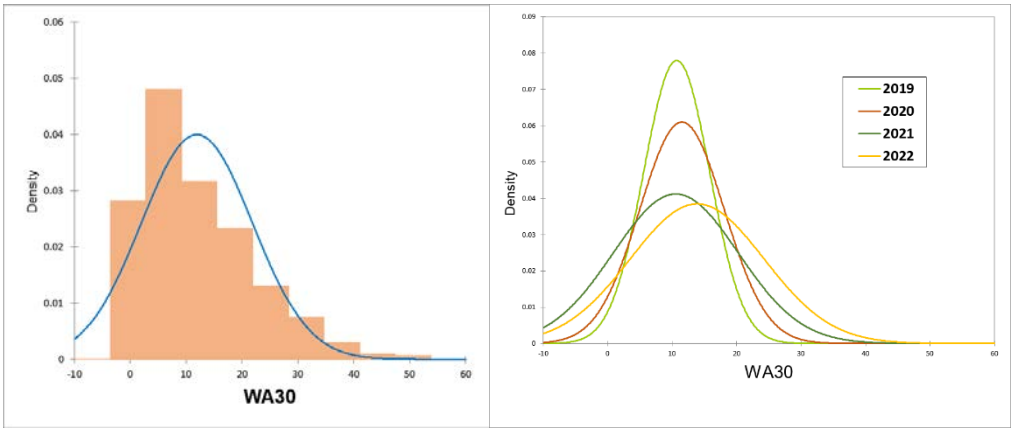


Figure 1 : Distribution of WA30 values for the 2905 samples (left) and distribution of WA30 values per years (right)

Obviously, the 2 classes (C1 and C2) defined on the basis of WA30 (≤ 12% and >12%) present different variability for WA30 (tab.3) with mean values respectively equal to 5% and 21;5%

Table 3 : Descriptive statistics for WA30 according to classes

Statistics	C1	C2
N	1692	1213
Minimum	-3.90	12.00
Maximum	12.00	52.83
Mean	5.00	21.53
SD	3.47	7.89

The average values and standard deviations of WA30 remain constant for class1 over years, and increase regularly with years for class2 (tab.4)

Table 4 : Mean and standard deviation values of WA30 per year and class

	C1		C2	
Year	**Mean**	**SD**	**Mean**	**SD**
2019	7.71	3.42	15.79	3.05
2020	6.57	3.24	17.07	4.52
2021	4.67	3.33	21.43	8.04
2022	5.38	3.64	22.08	7.86

The water absorption capability observed within genotypes replicates over years is highly variable (tab.5 and fig. 2), with CV ranged between 28% and 137% (Tab.5).

For some genotypes this high variability observed as no impact on the characterization of the clone classification as good (C2) or bad (C1) cooking time, (tab.6) due to the range of values (e.g Ven25, analyzed 27 times presents an average value of 2% with a SD of 1,4% and is assigned as C1 for 100% of the measurements, tab. 6)

Actually, only 6 genotypes over 22 evaluated every year were assigned to the same class for all the replicates (Tab.6).

Table 5 : Number of replicates, average and SD WA30 values and coefficient of variation for the 22 genotypes analyzed each year

#	Genotype	N	Mean (WA30)	SD (WA30)	CV
1	**BRA325**	8	3.55	1.74	49%
2	**BRA512**	13	2.65	1.81	68%
3	**CM7436-7**	10	11.04	5.61	51%
4	**COL1505**	18	15.94	9.35	59%
5	**COL1722**	45	14.78	6.00	41%
6	**COL1736**	8	13.68	4.46	33%
7	**COL1910**	28	1.97	2.70	137%
8	**COL2215**	10	15.48	7.23	47%
9	**COL2246**	14	8.65	6.32	73%
10	**CUB46**	9	17.93	6.17	34%
11	**GUA24**	10	12.12	6.42	53%
12	**IND135**	6	29.20	10.59	36%
13	**MAL3**	12	26.39	12.03	46%
14	**MEX2**	4	18.08	7.39	41%
15	**PAN70**	4	16.47	6.16	37%
16	**PAR98**	6	18.99	11.15	59%
17	**PER183**	54	16.61	9.37	56%
18	**PER496**	12	23.91	6.66	28%
19	**SM1127-8**	14	16.98	5.35	32%
20	**VEN208**	27	23.95	13.83	58%
21	**VEN25**	68	2.00	1.38	69%
22	**VEN77**	12	16.36	5.90	36%

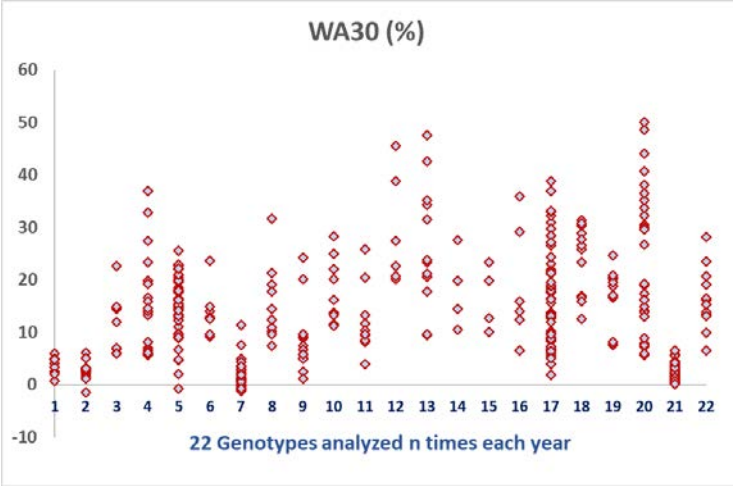


Figure 2 : WA30 values for the 22 genotypes analyzed n times over 4 years

Table 6 : Classification rates for the 22 genotypes according to WA30 values measured

Genotype	C1	C2
BRA325	100%	
BRA512	100%	
CM7436-7	50%	50%
COL1505	33%	67%
COL1722	29%	71%
COL1736	25%	75%
COL1910	100%	
COL2215	40%	60%
COL2246	86%	14%
CUB46	22%	78%
GUA24	70%	30%
IND135		100%
MAL3	17%	83%
MEX2	25%	75%
PAN70	25%	75%
PAR98	17%	83%
PER183	37%	63%
PER496		100%
SM1127-8	21%	79%
VEN208	22%	78%
VEN25	100%	
VEN77	17%	83%

1.3 Near infrared spectroscopy

1.3.1 Exploration

The near infrared spectra were acquired on the 2905 ground (puree) samples according to the RTBfoods SOP (Belalcazar John, 2020). The spectra present characteristics absorption bands for water, proteins, starch, cellulose and color (fig.2). All spectra present same patterns with variability in response (absorbances), no atypical spectrum was present. No clear clustering can be highlighted according to classes (C1 and C2).

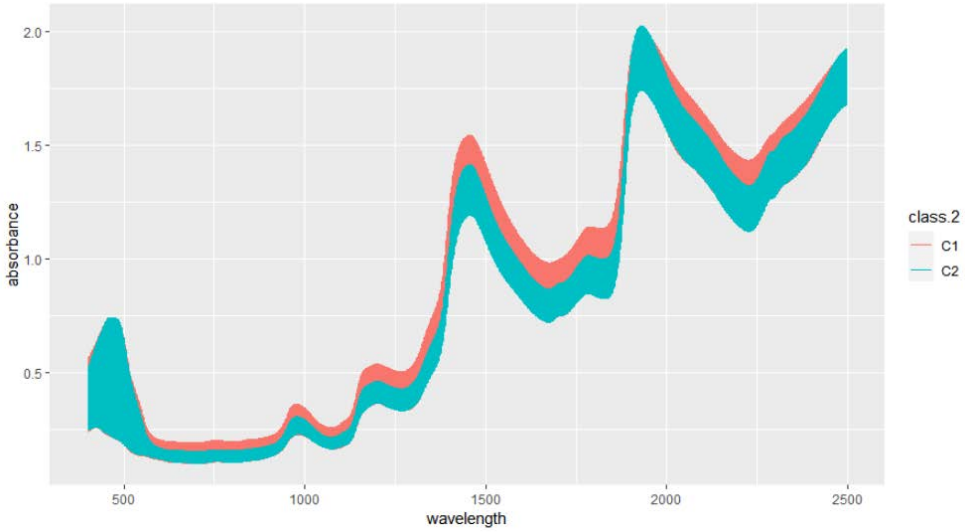


Figure 3 : visible and Near infrared spectra of the whole population (2905 spectra)

1.3.2 Principal Components Analysis

A PCA calculated on raw spectra led to 95,4% of variance explained by the first two PCs. The representation of the scores for the first two PCs shows the clustering among PC_1 due to color, as highlighted by loading associated to PC_1 (fig.3). There is no clustering according to year except artificial one due to over representation of samples in 2021 and 2022.

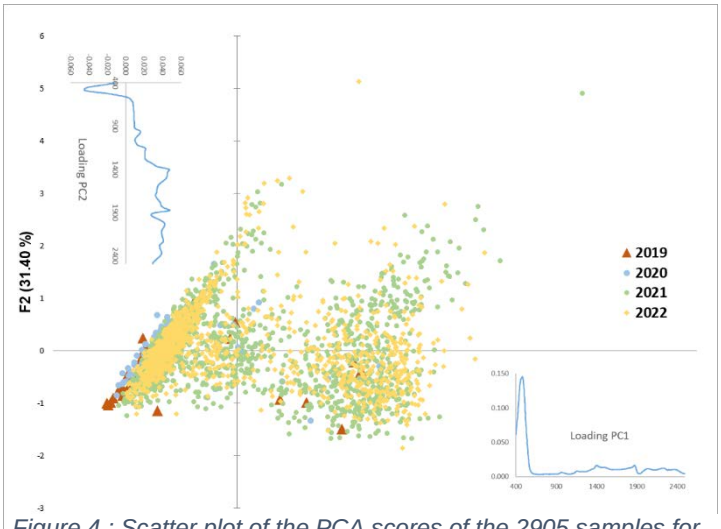


Figure 4 : Scatter plot of the PCA scores of the 2905 samples for PC_1 and PC_2; PCA on raw full spectral data.

A PCA calculated on raw spectra for NIR segment only (800-2500 nm) led to 98,2% of variance explained by the first two PCs. The effect of colour was removed, and the repartition of the spectra among PC_1 and PC_2 (fig.4), then the scores distribution on the factorial map highlighted the higher variability for 2021 and 2022 samples. Two samples presented high Mahalanobis distances (GH) from average spectrum: M01422_025 a.k.a. 202109DVGN6_momi_rep1_22-06_COL941_25 (GH= 14,8) and M06021_075 a.k.a. 202050CQPRG_ciat_rep1_SM4864-33_75 (GH= 21,5), these samples presented WA30 values respectively equal to -2,5% and -2,9%, obviously they are atypical samples.

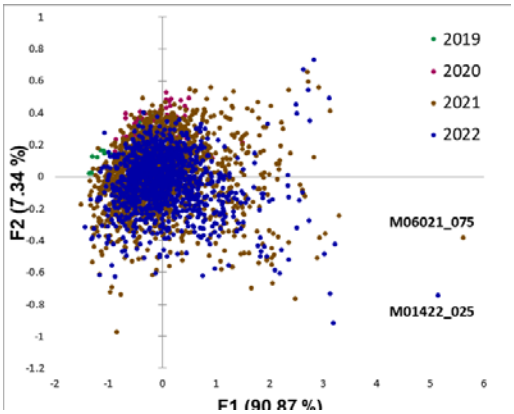


Figure 5 : Scatter plot of the PCA scores of the 2905 samples for PC_1 and PC_2; PCA on raw NIR segment spectral data

2 MODELLING

2.1 Learning and test sets

In order to develop optimal model for WA30 quantification and/or classes discrimination two sets of data were constituted: one for learning, tuning the parameters, and one for testing, evaluating the error of the predictive model. To do this, 70% of the samples were randomly selected for learning set and the 30% remaining were used for testing. The selection was based on stratified random procedure (Addinsoft, 2022; R. A. Sugden, 2000): Rows were randomly chosen within N strata. In each stratum, the number of sampled observations was proportional to the frequency of the stratum. The stratums corresponded to year and classes, this selection ensures to keep in each set the same repartition as in the whole data set.

The two sets contained respectively 2033 samples (learning) and 872 samples (test) with an equivalent repartition for year and classes (tab. 5 & 6)

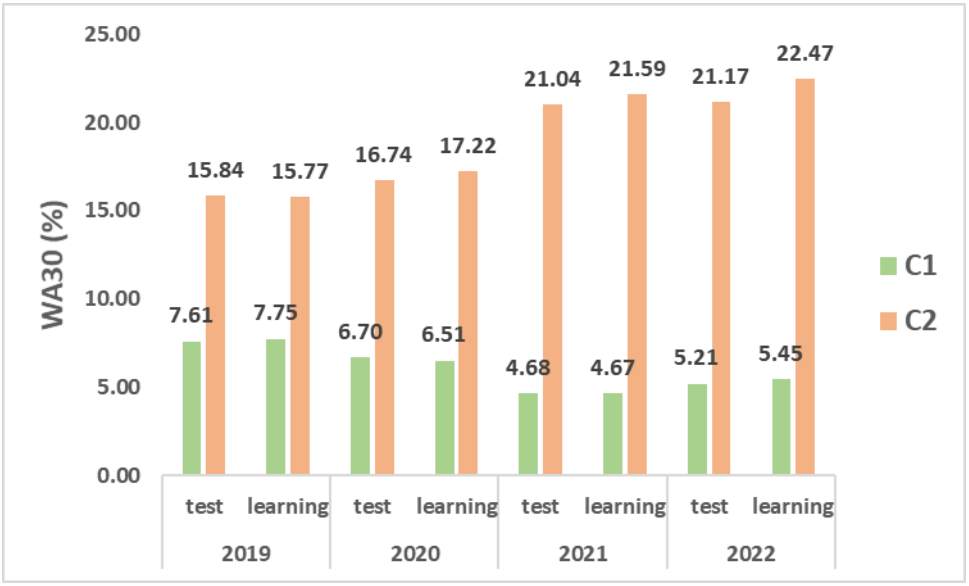
Table 7 : Repartition of samples per set per class per year

	Data		test		learning	
	C1	C2	C1	C2	C1	C2
2019	63%	37%	64%	36%	63%	38%
2020	53%	47%	52%	48%	53%	47%
2021	65%	35%	65%	35%	65%	35%
2022	49%	51%	49%	51%	49%	51%

Table 8 : Repartition of samples per set per year

	Data	Test	Learning
2019	1.2%	1.3%	1.2%
2020	3.1%	3.1%	3.0%
2021	56.5%	56.4%	56.6%
2022	39.2%	39.2%	39.2%

The average values for WA30 per set were respectively equal to 11,61% for test set and 12,03 % for learning set; the corresponding SD were 9,73% and 10,07 % respectively. The average WA30 values per year and classes are reported in figure 5, the average values per classes are similar for the 2 sets for each year.



The spectral projections (n = 872) of test samples onto the first factorial map of the PCA calculated from the training spectra (n = 2033) highlights that the spectral variability of the randomized test samples covers the whole population variability (fig. 7)

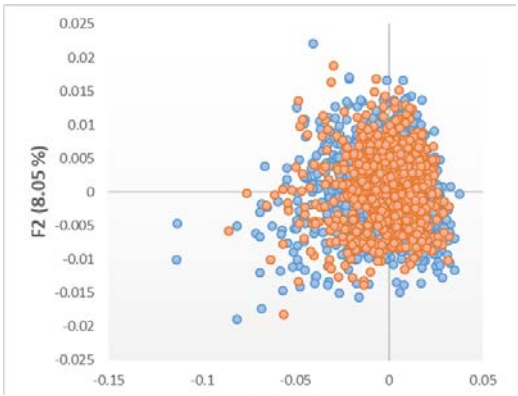


Figure 7 : Projection of test samples (872) on PC_1 and PC_2 of the PCA calculated on training samples (2033)

The scatter plots of the PCA scores of the samples for the two sets, training and test, with different color for classes show similar repartition among PC_1 and PC_2 (fig 8), with higher dispersion for C2 samples.

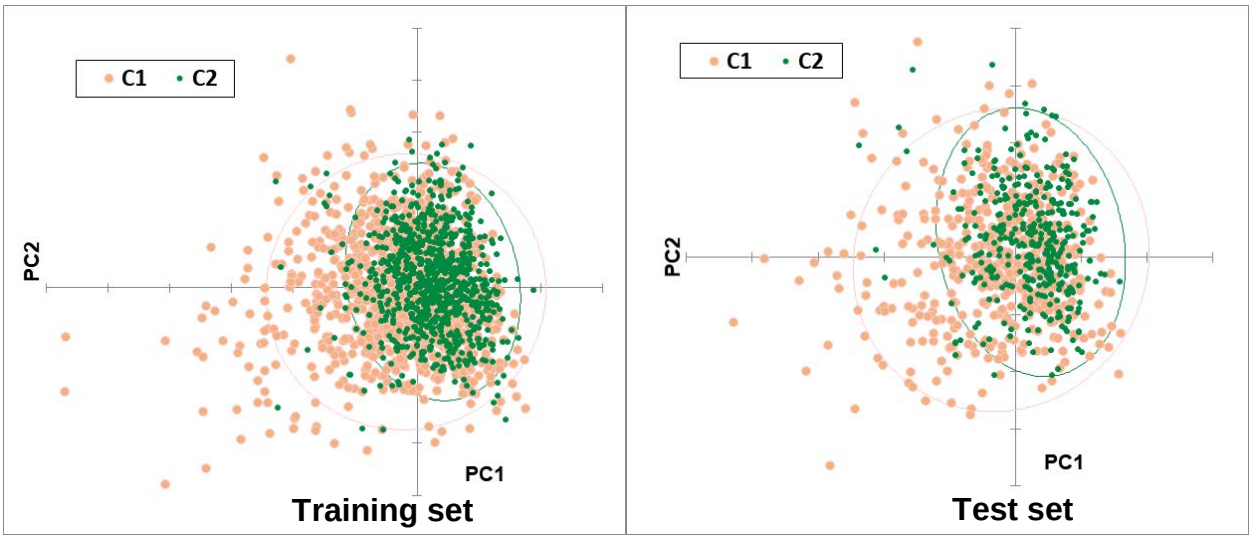


Figure 8 : Scatter plots of PCA scores for PC_1 and PC_2 for training set samples (left) and test samples (right), colored by classes, C1 WA30 ≤12% and C2 WA30 > 12%

Two modelling strategies were investigated:

- an indirect classification based on a regression step to predict WA30 using spectral fingerprints and then class attribution according to WA30 predicted values
- a direct classification using discriminant procedure based on classes defined by WA30 laboratory values and spectral fingerprints

Different methods were investigated for both approaches: PLSR, Lasso, Ridge, Elastic Net, Local PLSR, SVM for regression methods and PLSRDA, K Nearest Neighbors, Naive Bayesian Classifier, Random Forest, Classification Regression Trees, SVM for classification methods.

For all the investigations, different pre-treatments (SNV, SNVD, smoothing, first or second derivative…) of the spectra were tested in order to decide which one leads to lower error in prediction and classification rates and corresponds to higher parsimonious model.

The two best models, one per approaches, are reported here according to the previous criteria.

2.2 Indirect classification using spectra

Three regression methods gave similar results for WA30 modelling: Ridge regression, Lasso regression and Local PLSR regression. The regression error (RMSEP) and the indirect classification error (rate %) evaluated onto the test set were lower for the Ridge regression with in particular the lowest rate of false positives.

2.2.1 Classification based on Ridge regression

The Ridge regression, a method derived from Tikhonov regularization, was proposed by Hoerl and Kennard in 1970 (Hoerl, 1970). This factorial method is an estimation method that constrains its coefficients not to explode, unlike standard high-dimensional linear regression. The high-dimensional context covers all situations where a very large number of variables is available in relation to the number of individuals.

Ridge Regression is suitable for data that suffer from multicollinearity, such as spectral data. Due to multicollinearity the unbiased least squares estimates present too large variances which results in less accurate results. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors.

Ridge regression is one of the methods that overcomes the shortcomings (instability of the estimate and lack of reliability of the forecast) of linear regression in a high-dimensional context (Jerome

Friedman, 2008). Ridge regression stands out from LASSO regression in its greater robustness against datasets with strong multicollinearity.

The Ridge regression is applied to the learning set (N = 2033) on spectra pre-treated as follow: a) wavelength range retained 800 nm to 2498 nm with a 2 nm step, b) correction for light scattering using Standard Normal Variate procedure (SNV, (Barnes, 1989)), c) Savitzky-Golay (Savitzky, 1964) Derivative (Order: 1, Polynomial Order: 2, Smoothing Points: 9, Left Points: 4, Right Points:4).

The performances of the regression are relatively poor with a R^2 = 0,543 and RMSECV = 6.80%, the scatter plot of predicted values versus measured values of WA30 (fig. 9A) illustrate the poor efficiency of the model especially for high WA30 values (> 30 %) as illustrated by the scatter plot of residuals versus measured values (fig.9B).

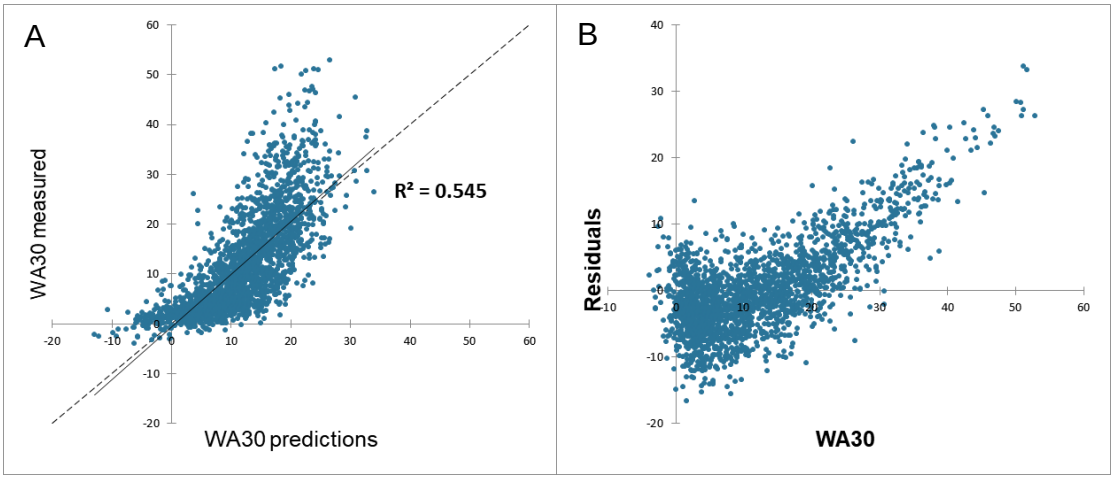


Figure 9 : Scatter plot of WA30 measured versus predicted values (A) and scatter plot of residuals vs WA30 measured values (B)

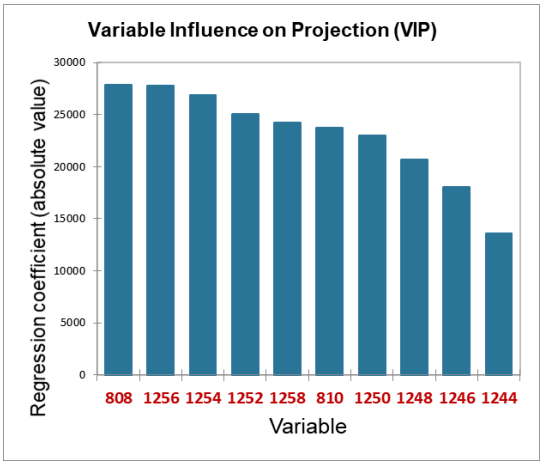


Figure 10 : 10 most VIP of the Ridge regression model

The 10 most relevant variables important on projection (VIP) of the model (fig.10) are related to 2 spectral zones in short-wave NIR (SWNIR) region, 808 – 810 nm corresponding to N-H (third overtone) and 1244 -1258 nm corresponding to -CH2; -CH3 and C=C (third overtone) according to P. Williams and K. Norris ((Williams, 1987).

The ridge regression model applied to the prediction of the test set gives an error of prediction RMSEP = 6.84% similar to RMSECV, and a coefficient of determination R^2_p = 0,50 also similar to the one observed for training set. Similarly to learning step, higher residuals were observed for high WA30 values. This can be illustrated by the distribution on the residuals according to corresponding classes (fig. 11), the dispersion of residuals is higher for C2 class (WA30 >12%).

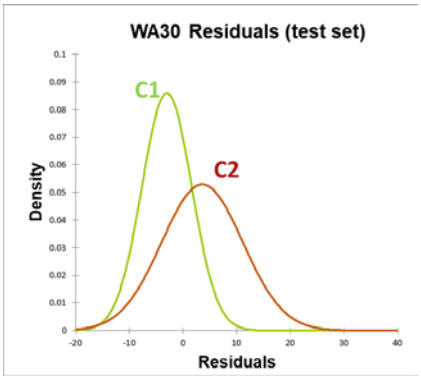
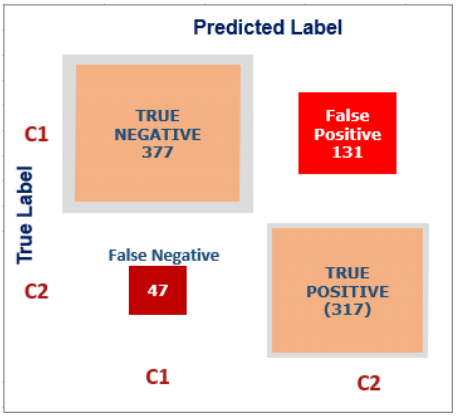


Figure 11 :WA30 residuals distribution per class

Based on the predicted values for WA30, the samples are assigned to predicted classes C1 and C2 using the same rule: C1 for $WA30_{pred}$ ≤ 12% and C2 for $WA30_{pred}$ > 12%. Then a confusion matrix is calculated between actual classes and predicted classes (tab. 9).

Table 9 : Confusion matrix for test set (N = 872)

From\To	C1	C2	N	Rate
C1	377	131	508	74.2%
C2	47	317	364	87.1%
				79.6%



The number of True Positive is equal to TP = 317; the n number of True Negative TN = 377, the number of False positive FP = 131 and the number of False Negative FN = 47 (fig. 12).

The overall **classification (accuracy) rate is 79,6%**, the **true positive rate (sensitivity) is 74,2%** and the **true negative rate (specificity) is 74,2 %.**

The precision or Positive Predictive value (PPV) corresponds to the proportion of positive predictions actually correct, PPV = 0,889. In other words, the model accurately predicts the positive class in 88,9 % of cases.

The False Positive Rate is FPR = 25,8% while the False Negative Rate FNR = 12,9%.

From this confusion matrix and parameters, we can calculate the Matthew's correlation coefficient (MCC) which is MCC = 0.605 and the F1 score (harmonic mean of precision and recall) F1 = 0,781. When those parameters are close to 1, the model is considered as efficient and balanced.

False positives

The number of false positives is 131 samples (spectra) corresponding to 120 clones, nine clones are present 2 times and one 3 times (Annex II).

The descriptive statistics (actual and predicted values) for these 131 false positives are:

Variable	Type	N	Minimum	Maximum	Mean	SD
WA30	Laboratory	131	-0.63	12.00	7.79	2.77
	Predicted	131	12.06	26.84	15.11	2.47

The distributions for both population (actual and predicted values) are similar (fig. 13), the maximum predicted value is 26,8%.

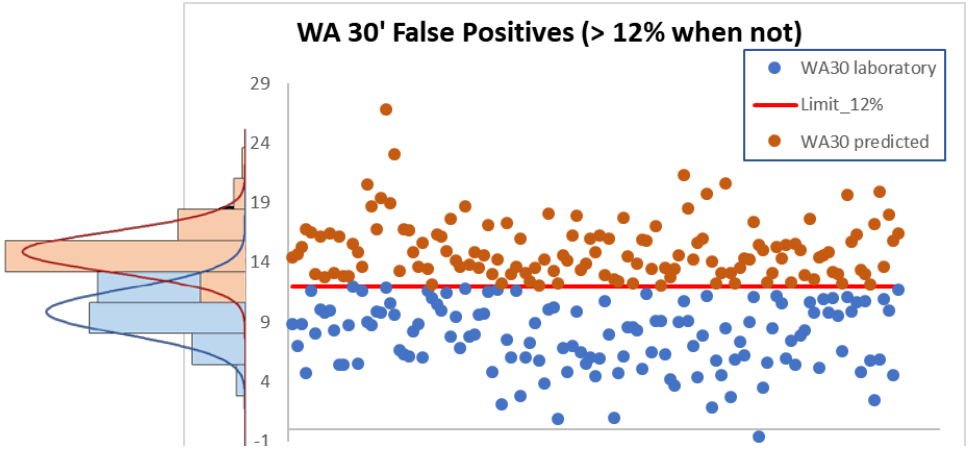


Figure 13 Predicted and actual WA30 values for the 131 false positives in test set. Limit of classes = 12%

Within the 131 false positives samples: 53 samples (40%) presents actual values between 9% and 12%, these samples coming from class 1 are close to the limit of classes.

2.3 Direct classification using spectra

Among all methods investigated for direct classification according to spectra fingerprint and classes defined by WA30 values: C1 for WA30 ≤ 12% and C2 for WA30 > 12%, two methods led to efficient models: PLSRDA and PCA-LDA. The overall classification rate (accuracy) was better using PLSRDA, with a lower false positive rate.

2.3.1 Classification using PLSRDA

Unlike traditional multiple regression models, the partial least squares (PLS) algorithm is based on multivariate projection and is not limited to uncorrelated variables. One of the many advantages of PLS is that it can handle high-dimensional, noisy, collinear and missing data while simultaneously modelling several response variables. PLS is built from a matrix of descriptive variables (quantitative values) X (N, P) to which is associated a matrix of response variables (qualitative values) Y (N, Q), with N observations, P variables, and Q classes. The matrix Y is encoded by dummy values so that each observation corresponds to an indicator variable for each class ([1 0] for class 1 and [0 1] for class 2). PLS aims at maximizing the covariance of the X and Y matrices through latent variables (LVs), which are linear combinations of original variables.

The PLSRDA was run using the software R© (https://www.R-project.org/, 2022) with the package rchemo (Lesnoff, 2022). The PLSRDA is applied on the training set (N = 2033) in order to tune the

model parameters, especially the number of latent variables (LVs). The number of LVs used to establish the dimension of the model was defined using cross-validation per block with repetition, the number of block was 4 and 10 repetitions were run. The maximum number of LVs was fixed at 50 during model construction, and the number of LVs providing the minimum value of prediction error was selected. Then, the model was applied to the test set (n = 872).

Prior to model construction, the spectra were pre-treated as follow: a) wavelength range retained 800 nm to 2498 nm with a 2 nm step, b) correction for light scattering using Standard Normal Variate procedure, c) Savitzky-Golay Derivative (Order: 1, Polynomial Order: 2, Smoothing Points: 9, Left Points: 4, Right Points:4).

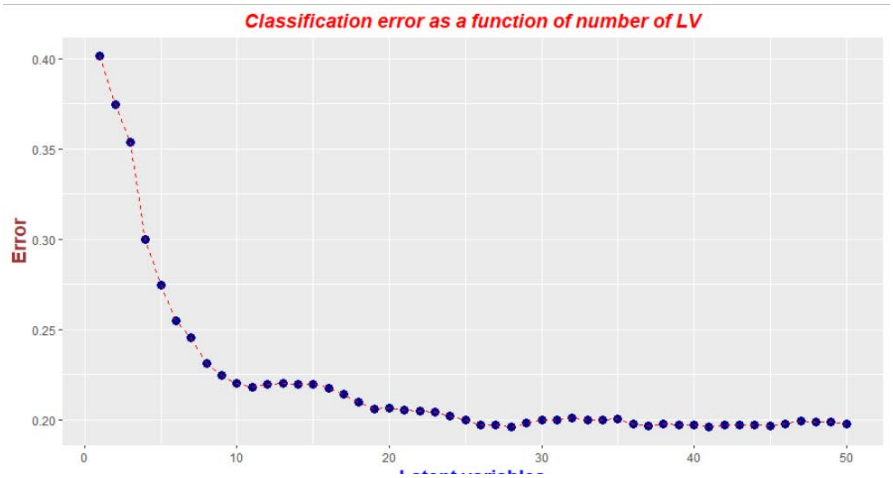


Figure 14 : Evolution of the classification error vs number of latent variables

The cross-validation process led to select 41 LVs which corresponds to a minimum classification error of 18,2%, the profile of errors versus the number of LVs (fig.14) shows that with 26 LVs the error is 18,6%. According to the rule of parsimonious model, i.e minimum of parameters, the number of LVs retained for final model is 26, even if considering the high number of dimensions of the data a number of 41 LVs is acceptable and do not leads to overfitting models.

The model with 26 LVs applied to the test set (n = 872) led to a classification rate (accuracy) of 81,4%, the sensitivity is 79,4% and a specificity of 82,9% (tab.10).

Table 10: Confusion matrix for test set (model based on 26 LVs)

From / To	C1	C2	N	Rates	
C1	421	87	508	82.9%	specificity
C2	75	289	364	79.4%	sensitivity
			872	81.4%	Accuracy

The precision or Positive Predictive value (PPV) corresponds to the proportion of positive predictions actually correct, PPV = 0,849. In other words, the model accurately predicts the positive class in 84,9 % of cases (tab.10 & fig. 15).

The False Positive Rate is FPR = 17,1% while the False Negative Rate FNR = 20,6%.

From this confusion matrix and parameters, we can calculate the Matthew's correlation coefficient (MCC) which is MCC = 0.62 and the F1 score (harmonic mean of precision and recall) F1 = 0,84. The model is considered as efficient and balanced.

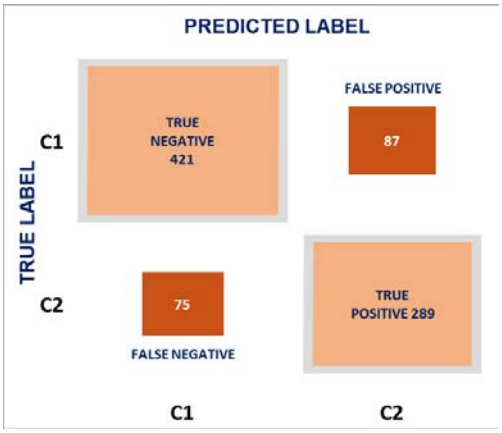


Figure 15 : Confusion matrix graphic

False positives

The number of false positives is 87 samples (spectra) corresponding to 81 clones, six clones are present 2 times (CM5948-1, GM13169-10, GM13240-28, SM4921-9, SM4954-33 and VEN208).

The descriptive statistics for WA30 values for these 87 false positives are:

Statistics	N	Minimum	Maximum	Mean	SD
WA30	87	-0.63	11.89	8.27	2.74

The distribution (fig. 16) of the WA30 values for the 87 false positives is clearly asymmetric to the left, within the 87 false positives samples: 51 samples (61%) presents WA30 values ≥ 8%, among these 24 samples (28% of the 87) present WA30 > 10% these samples coming from class 1 are close to the limit of classes. The WA30 values are < 8% for 36 samples within the 87 false positives.

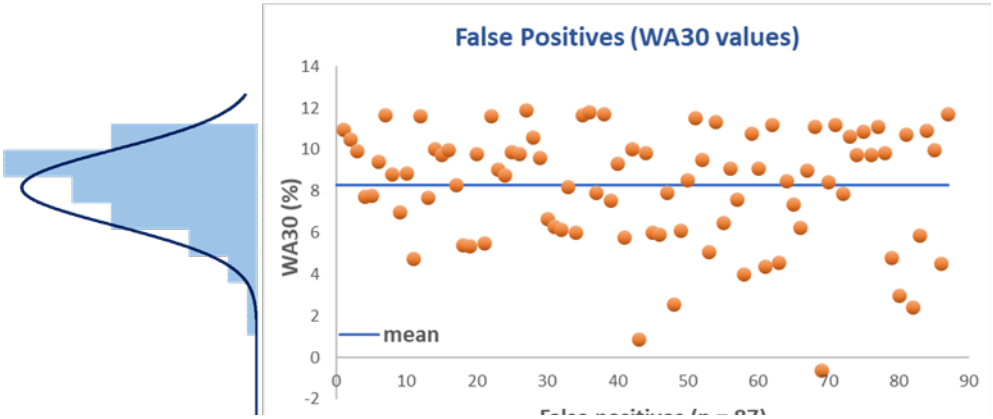


Figure 16 : Repartition of the WA30 values of the 87 false positives from PLSRDA

2.4 Direct vs indirect classification

Both approaches, direct classification based on classes and indirect classification using WA30 predictions led to similar results in terms of accuracy of classification (respectively 81,4 % and 79,2%) with a false positive rate slightly higher for indirect classification 17,1% and 25,8%.

Obviously, both approaches are based on same “references” as classes are defined by the WA30 values (≤ 12% and > 12%), but the chemometrics approaches are different especially in the way to define predictors and so to select latent variables within the spectra data. **The fact that both methods led to similar results confirms that the spectral fingerprints of fresh cassava contain relevant information linked to water absorption capability**. The direct classification approach, based on PLSRDA, led to better results for accuracy (successful classified rate), specificity (true negative rate) and sensitivity (True positive rate) and this for both sets: learning (cross validation) and test sets (prediction).

Moreover, the discriminant approach presents the advantage to be easier to set up and run in routine analysis, with an easier interpretation of coefficients and loadings.

Evaluating the performances using a test set selected within the whole population implied that this test set is representative of the variability that will be faced in the next years. If the variability of the genotypes for both WA capacity and spectral fingerprints is stable after 4 years, the model will be applicable as is, otherwise it will have to be updated to include the sources of variability.

2.4.1 External validation using years

The variability due to the date (year) of harvesting and analysis and its effect on prediction accuracy can be tested by developing models per year. As the number of samples in 2019 and 2020 is low (n = 124, respectively 35 and 89) solely 2 data sets are constituted: one gathering data from 2019 until 2021 (n = 1766) and the other one with 2022 samples (n = 1139). This external validation is tested for both approaches (direct and indirect).

In paragraph 1.2 we have seen that variability of WA30 increases over years. The repartition of samples per classes for both sets according to classes is:

	N		%	
Year	**C1**	**C2**	C1	C2
[2019-2021]	1133	633	64%	36%
[2022]	559	580	49%	51%

The spectra of the two sets present similar patterns (fig.17). Nevertheless, the projections of 2022 samples onto PCA space of samples from 2019-2021 (fig.18), or the projections of 2019-2021 samples onto PCA space of samples from 2022 (fig. 19) highlight some difference in terms of variability of the two datasets.

Thereby, the first plan of PCA done on 2019-2021 samples explains 89% of the variability of 2022 samples, and first plan of PCA done on 2022 explains 90% of 2019-2021variability.

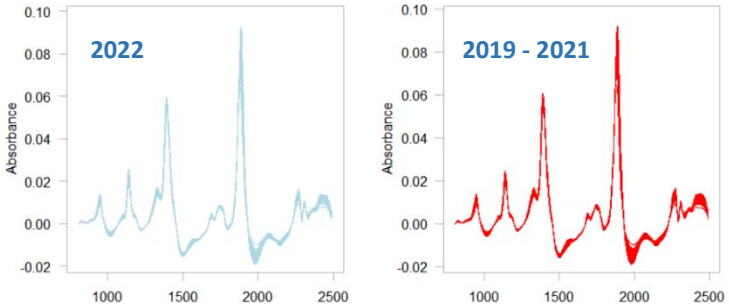


Figure 17 : 2022 and 219-2021 Spectra (correction SNV and first derivative)

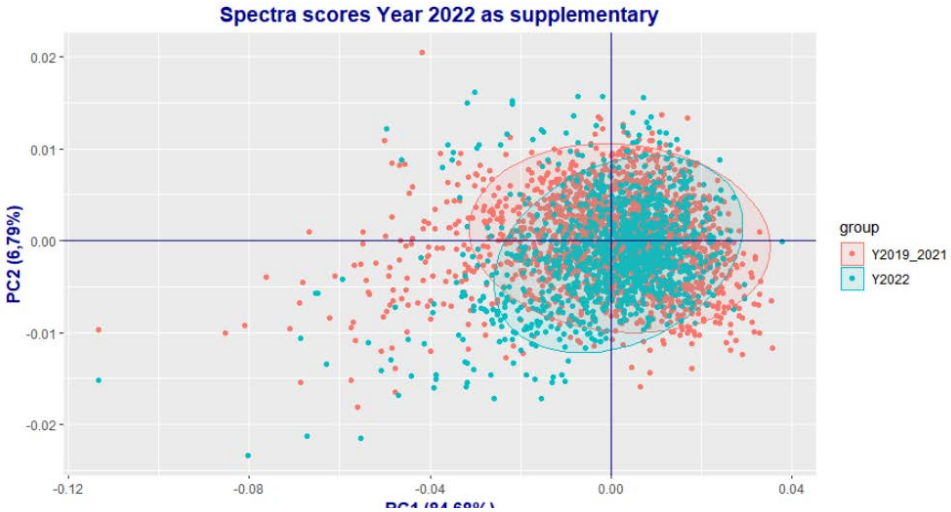


Figure 18 : Projections of 2022 samples onto PC_1 and PC_2 of PCA calculated on 2019-2021 samples

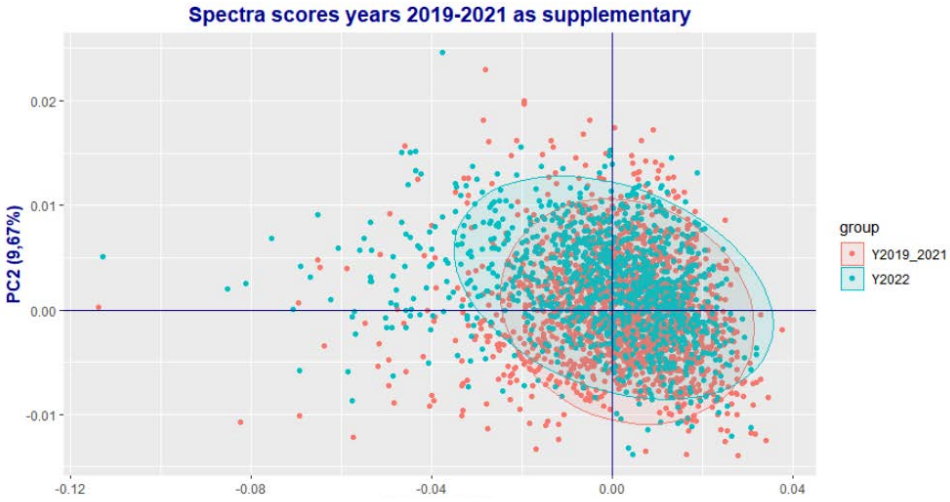


Figure 19: Projections of 2019-2021 samples onto PC_1 and PC_2 of PCA calculated on 2022 samples

PLSRDA, samples [2019-2021] as learning set

The PLSRDA model developed on 2019-2021 samples (n=1766) retains 29 LVs (cross-validation), the prediction of 2022 samples (n = 1139) leads to a classification rate of 72,6% (accuracy) with a rate of true negatives of 61,5% (specificity) and a rate of true positives of 83,3% (sensitivity). The rate of false positives (FP) = 38% with 215 samples misclassified. The average value of WA30 for FP samples is equal to 6,87% and 58% of these false positives present WA30 values < 8%.

From \ To	C1	C2	N	Rates	
C1	344	215	559	61.5%	Specificity
C2	97	483	580	83.3%	Sensitivity
			1139	72.6%	Accuracy

PLSRDA, samples [2022] as learning set

The PLSRDA model developed on 2022 samples (n = 1139) retains 21 LVs (cross-validation), the prediction of 2019-2021 (n = 1766) samples leads to a classification rate of 75,4% (accuracy) with a rate of true negatives of 85,4% (specificity) and a rate of true positives of 57,3% (sensitivity). The rate of false positives (FP) = 15% with 165 samples misclassified and the rate of false negatives (FN) is 42,7%. The average value of WA30 for FP samples is equal to 7,33 % and 53% of these false positives present WA30 values < 8%.

From \ TO	C1	C2	N	Rates	
C1	968	165	1133	85.4%	Specificity
C2	270	363	633	57.3%	Sensitivity
			1766	75.4%	Accuracy

Clearly these 2 external validations highlight the importance of the representativeness of the learning database in terms of variability, after 3 years (2019 -2021, with mainly samples from 2021) the variability of the real situation wasn't covered. Moreover, the samples from 2019 present a narrowest distribution for WA30, which leads to 2 classes close to their means (fig. 1). The classes are unbalanced with C1 accounting for 64%, in this situation the model loses specificity (True Negative rate)

Using the 2022 samples as training and predicting 2019-2021 samples leads to similar accuracy 75,4% but in this situation the sensitivity (True positives rate) is only 57,3%.

The 10% of unexplained variability observed on spectra (PCA) in both situations is one of the reasons of these less efficient classifications.

Ridge Regression, samples [2019-2021] as learning set

The Ridge regression model developed on the WA30 values of the 2019-2021 samples (n = 1766) present à R^2= 0,554 and RMSEC = 6,33%. The prediction of 2022 (n = 1139) samples presents a R^2_p = 0,357 and RMSEP = 8,31%.

Test samples are assigned to classes based on their WA30 predicted values. The classification rate is accuracy = 75,4% with a rate of true negatives of 92,9%% (specificity) and a rate of true positives of 48,3% (sensitivity). The rate of false positives (FP) = 52% with 289 samples misclassified and the rate of false negatives (FN) is 7,1%. The average value of predicted WA30 for FP samples is equal to 15.82% when mean actual value for same samples is 6,60%. Within this 289 FP, 111 (38,4%) present WA30 actual values ≥8% and 178 present WA30 actual values < 8% (61,5%). In this situation the FP rate indicates that variability of 2022 samples in terms of WA30 values is not covered by 2019-2021 samples.

Ridge Regression, samples [2022] as learning set

The Ridge regression model developed on the WA30 values of the 2022 samples (n = 1139) present à R^2= 0,577 and RMSEC = 6,75%. The prediction of 2019-2021 (n = 1766) samples presents a R^2_p = 0,280 and RMSEP = 8,05%.

Test samples are assigned to classes based on their WA30 predicted values. The classification rate is accuracy = 67,5% with a rate of true negatives of 82,9%% (specificity) and a rate of true positives

of 58,9% (sensitivity). The rate of false positives (FP) = 41,1% with 466 samples misclassified and the rate of false negatives (FN) is 17,1%. The average value of predicted WA30 for FP samples is equal to 15,57% when mean actual value for same samples is 6,34%. Within this 466 FP, 158 (33,9%) present WA30 actual values ≥8% and 308 present WA30 actual values < 8% (66,1%). This validation indicates that a part of the variability of cassava genotypes is not fully cover by 2022 samples.

These external validation procedures confirm that variability spectral and variability regarding cooking behaviour of cassava genotypes need to be investigated more in order to understand and identify their different sources. This will be mandatory in order to develop a robust model.

3 CONCLUSION

The number of samples is high with 2905 spectra and corresponding values for water absorption at 30' (WA30), this sampling covers 4 years of harvesting and contains 1101 different genotypes. The variability over years in terms of cooking behaviour expressed as WA30 is high, this variability is also present, at a lower level of amplitude, within spectral fingerprints.

The two approaches: regression (Ridge Regression) and classification (PLSRDA), based on different methods for regressor (LVs) selection within the spectral data, lead to similar performances in terms of classification according to WA30 classes. Nevertheless, PLSRDA leads to better classification and is easier to implement and interpret. The classification accuracy is 81,4% when predicting test set with a sensitivity = 79,4% and a specificity of 82,9% and a false positive rate equal to 17,1% while false negative rate is 20,6%.

This model is efficient and can be implemented in a selection scheme 1) as is if the next year/generation variability remains similar the current database 2) or with controlled update if next year/generation variability differs from current database.

The external validations using samples from 2019 to 2021 against samples from 2022 and inversely, demonstrated that the part of unexplained variability can have an important impact onto classification rates (accuracy, specificity and sensitivity), especially on the level of false positives.

4 PERSPECTIVES

The main activity should focus on further investigation of the variability of cooking behaviour within and between genotypes and between locations.

The 2023 sampling should be designed to evaluate the robustness and efficiency of the 4 years (n = 2905) models and to identify the sources of variations in order to integrated them.

This should be coupled with a clear determination of the laboratory error for WA30 measurement. Better is known the noise within reference data easier is the choice of modelling strategy.

An evaluation of moving the limit of the 2 classes should be tested in regards to the ratio benefit/risk for breeders.

5 BIBLIOGRAPHY

Addinsoft Addinsoft (2022). XLSTAT statistical and data analysis solution. Paris, France // https://www.xlstat.com. - 2022.

Barnes R. J., Dhanoa, M. S., & Lister, S. J. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra [Revue]. - [s.l.] : Applied spectroscopy, 1989. - (5), p. 772-777 : Vol. 43.

Belalcazar John Tran Thierry, Meghar Karima, Davrieux Fabrice. NIRS measurement on fresh ground cassava. High-throughput phenotyping protocols (HTPP) [Rapport]. - Cali, Columbia : WP3. Cali : RTBfoods Project-CIRAD, 2020. - https://doi.org/10.18167/agritrop/00676.

computing R Core Team (2022). R: A language and environment for statistical R Foundation for Statistical Computing, Vienna, Austria.. - [s.l.] : URL https://www.R-project.org/..

Escobar Salamanca Andrés Felipe Tran Thierry, Arufe Vilas Santiago. characterization of water absorption, cooking time and closing angle of boiled cassava. Biophysical characterization of quality traits. [Rapport]. - Cali, Columbia : RTBfoods Project-CIRAD, 2022. - https://doi.org/10.18167/agritrop/00683.

Hoerl A. E., & Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems [Revue]. - [s.l.] : Technometrics, 1970. - (1) 55-67 : Vol. 12.

https://www.R-project.org/ R Core Team (2022). R: A language and environment for statistical computing. - [s.l.] : R Foundation for Statistical Computing, Vienna, Austria., 2022.

Jerome Friedman Trevor Hastie et Rob Tibshirani The elements of statistical learning: data mining, inference, and prediction [Revue]. - New York : Springer, 2008. - Vol. 2, pp 1-758.

Lesnoff M. rchemo: Dimension reduction, Regression and Discrimination for chemometrics. - https://github.com/mlesnoff/rchemo : R package version 0.0-17 ; R package version 0.0-17, 2022.

R. A. Sugden T. M. F. Smith, R. P. Jones Cochran's rule for simple random sampling [Revue]. - [s.l.] : Journal of the Royal Statistical Society: Series B (Statistical Methodology),https://doi.org/10.1111/1467-9868.00264, 2000. - 787-793 : Vol. 62.

Savitzky A., & Golay, M. J. Smoothing and differentiation of data by simplified least squares procedures. [Revue]. - [s.l.] : Analytical chemistry, 1964. - (8), 1627-1639 : Vol. 36.

Thierry Tran Xiaofei Zhang, Hernan Ceballos, Jhon L. Moreno, Jorge Luna, Andrés Escobar, Nelson Morante, John Belalcazar, Luis A. Becerra, Dominique Dufour Correlation of cooking time with water absorption and changes in relative density during boiling of cassava roots [Revue]. - [s.l.] : Int. J. Food Sci. Technol., 56: 1193-1205., 2021. - https://doi.org/10.1111/ijfs.14769.

Williams P., & Norris, K. Near-infrared technology in the agricultural and food industries [Livre]. - Saint Paul, Minesota, USA : American Association of Cereal Chemists, Inc., 1987.

6 APPENDICES

6.1 Annex I: Genotypes analysed each year

Genotype	2019	2020	2021	2022	N
BRA325	1	1	4	2	8
BRA512	1	1	7	4	13
CM7436-7	1	3	4	2	10
COL1505	1	3	10	4	18
COL1722	1	3	21	20	45
COL1736	1	1	4	2	8
COL1910	1	1	14	12	28
COL2215	1	1	2	6	10
COL2246	1	3	6	4	14
CUB46	1	1	5	2	9
GUA24	1	3	4	2	10
IND135	1	3	1	1	6
MAL3	1	3	5	3	12
MEX2	1	1	1	1	4
PAN70	1	1	1	1	4
PAR98	1	3	1	1	6
PER183	1	3	32	18	54
PER496	1	3	5	3	12
SM1127-8	1	1	1	11	14
VEN208	1	3	16	7	27
VEN25	1	1	42	24	68
VEN77	1	3	5	3	12

6.2 Annex II: Repeated clones (10) within the false positives (Ridge Regression)

	ID	year	Month	parcelle	WA30	WA30_pred
BRA1177	M01422_194	2022	March	194	5.59	12.30
	M01422_035	2022	February	35	7.37	13.51
CM5948-1	M01522_06P01	2022	March	6	11.53	17.13
	M00221_06	2021	January	6	11.82	18.74
COL1722	M01322_118	2022	February	118	9.10	18.54
	M01921_286	2021	March	286	8.92	13.58
GM13174-49	M01322_247	2022	February	247	4.58	13.13
	M04121_097	2021	April	97	4.71	12.41
GM13240-28	M01322_109	2022	February	109	10.79	21.32
	M02321_299	2021	March	299	6.00	16.03
GM13240-29	M01322_203	2022	February	203	1.81	14.04
	M00422_203	2022	January	203	6.28	13.52
PER183	M02321_331	2021	March	331	5.93	16.29
	M01921_260	2021	March	260	7.23	12.44
SM4921-9	M03622_215	2022	May	215	4.79	13.39
	M03421_012	2021	April	12	10.03	16.13
SM4954-33	M03622_156	2022	April	156	11.12	19.64
	M03421_061	2021	April	61	9.99	16.45
VEN208	M01921_138	2021	March	138	6.02	13.14
	M01221_282	2021	February	282	7.56	17.27
	M02921_05	2021	March	5	7.92	14.86

6.3 Annex III: List of the field trials used to produce the dataset of WA30 and NIRS in the present report

202025DVGN1_ciat (21-02) Diversidad GWAS

202028BCCOB_ciat (21-11) Campo observacion carotenos

202029DVGN2_ciat (21-18) Familias carotenos

202030DVGN2_ciat (21-14) Familias carotenos

202032DVGN2_ciat (21-15) Multi Familias carotenos

202031DVGN1_ciat (21-16) Diversidad genetica

202050DVPRG_ciat (21-19) Parentales mejoramiento

RTBfoods progenitors harvests:

- 201903CQQU1_ciat (19-64; 20-02)
- 2019111CQQU1_ciat (20-03; 20-11)
- 202022CQQU1_ciat (21-01, 21-05, 21-09)
- 202102CQQU1_ciat (21-33, 22-01, 22-09)

RTBfoods progeny harvests:

- 202023CQQU2_ciat (21-03, 21-06)
- 202002CQQU1_ciat (21-07, 21-12)
- 202108CQQU2_ciat (22-02, 22-05)

202109DVGN6_momi (E22-06) Diversidad Genetica GWAS

202127BCCOB_ciat (E22-14) Campo de Observacion Betacarotenos

202128BCMUL_ciat (E22-15) Campo de Observacion Betacarotenos 2020

202136DVPRG_ciat (E22-19) Parentales Mejoramiento

202132BCF1C_ciat (E22-20) F1C1 Betacarotenos

202175CQCOB_ciat (E22-24) Campo Observacion Calidad Culinaria

Detailed information of these field trials is available on Cassavabase (cassavabase.org), using the field codes above.

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/