Calibration Report

NIRS & Biophysical Analyses: Tentative Prediction of Cassava Cooking Properties - Year 2

High-Throughput Phenotyping Protocols (HTPP), WP3

Saint Pierre, Réunion, France, 31/10/2021

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

John BELALCAZAR, International Center for Tropical Agriculture (CIAT), Cali, Colombia

Xiaofei ZHANG, Alliance Bioversity, 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, John BELALCAZAR, Xiaofei ZHANG, Thierry TRAN, (2022). *NIRS & Biophysical Analyses: Tentative Prediction of Cassava Cooking Properties - Year 2.* St Pierre, Réunion, France: RTBfoods Calibration Report, 23 p. http://doi.org/10.18167/agritrop/00715

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)	31/10/2021
Final validation by:	
Fabrice DAVRIEUX (CIRAD)	20/01/2022

CONTENTS

Table of Figures

ABSTRACT

The analyses concern 56 cassava genotypes harvested in 2019, 2020 and 2021: 1 genotype was analyzed 1 time, 6 analyzed 2 times, 20 analyzed 3 times, 15 analyzed 5 times, 6 six times and 9 analyzed 9 times. The total number of analyses is 250. The samples were analysed for their cooking properties (cooking time in boiling water), texture parameters (gradient, max force, distance at max force, area, linear distance and end force/ max force), dry matter content and water absorption capacity during cooking. The same genotypes were analysed in near infrared spectroscopy. The absorption spectra were performed on ground fresh roots using a FOSS 2500 spectrometer. The average Dry matter is 39,5 %, this value is constant over months (age of the root). The cooking time average value is 33,7 min, the values range from 10 to 60 minutes. The distribution of the values, allows defining 2 classes: C1 for OCT lower than 33,7 min and C2 for OCT higher or equal to 33,7 min. There is a non-linear relation between Water_Absorption at 20 min or 30 min and optimal cooking time: High time cooking genotypes absorb less water at 20 min than “good cooking” genotypes. The values of gradient range between 170 and 2489 kg/mm with an average value of 1205 kg/mm. The distribution of the values follows a normal law. Gradient is highly correlated to physical values related to Max force, Area and Linear distance. Gradient is also correlated to OCT (r = 0,719). The highest correlation between gradient and Water absorption is for WA 30 minutes (r = - 0,693). The relation between gradient and WA_30 is nonlinear (second order), genotypes with high gradient values absorb less water at 30 mn than genotypes with low gradient values which correspond to genotypes with low optimum cooking time.

Different multivariate approaches were investigated to associate spectral data and physico-chemical parameters. The direct calibrations of physico-chemical parameters were not performant. Classification according to 2 cooking time classes was tested using different algorithms. Whatever were the pre-treatments used (SNV, SNVD, first or second derivative…) and whatever the classification approach (K Nearest Neighbors, Support Vector Machine, Naive Bayesian Classifier, Random Forest, Classification Regression Trees…), the predictions of a validation set for the 2 cooking time classes failed.

The Lasso approach is encouraging and clearly improved the predictive model for OCT. The classification of the samples using predicted OCT values was 82% correct for learning set and range between 66% and 72% for validation samples depending on the validation set. The model lacks robustness, because of a relatively few number of samples and because of the variability of the samples due to harvest year, as shown by the PCA of the spectra.

These results confirms that the spectral signature contains information about textural properties and that nonlinear models or deep learning approaches good help extracting this information.

Keywords: fresh cassava, physico-chemical data, spectral data, calibrations, optimal cooking time (OCT), water absorption

1 DATA

1.1 Material

The analyses concern 56 genotypes: 1 genotype was analyzed 1 time, 6 analyzed 2 times, 20 analyzed 3 times, 15 analyzed 5 times, 6 six times and 9 analyzed 9 times. The total number of analyses is 250. Harvests took place in 2019 and 2020 and 2021; the repartition of sampling is as follow:

Harvest date	November	December	January	February	March
2019	36	35			
2020			61	28	
2021			30		60

None of the replicates by clone were harvested at the same date, except IND135 and COL2246, which were replicated in each of the 2021 harvests (IND135 in plot 7 and plot 22; COL2246 in plot 9 and plot 24).

1.2 Physical properties and wet chemistry

The physical properties estimated, are:

- Percentage of water absorption at 10, 20, 30, 40, 50, 60 minutes of boiling (WA)
- Percentage of water absorption at Optimum cooking time (WA at OCT)
- The optimum cooking time (OCT)
- Dry matter content at 30 minutes of boiling (DM30)
- Texture properties using texturometer, the retained parameters are gradient, max force, distance at max force, area, linear distance and end force/ max force.

The wet chemistry property is the dry matter content (DM) of fresh root

Table 1: Descriptive Statistics

Statistics	N	Minimum	Maximum	Mean	Variance	SD
DM(%) fresh	250	31.01	47.27	39.54	10.03	3.17
WA10 (%)	160	-0.99	13.30	4.12	7.70	2.78
WA20 (%)	250	-0.78	40.52	6.91	32.21	5.68
WA30 (%)	214	-0.22	51.13	12.98	80.99	9.00
WA40 (%)	35	0.81	26.89	14.54	40.71	6.38
WA50 (%)	35	-3.81	31.37	17.26	58.49	7.65
WA60 (%)	35	-4.00	33.91	19.22	65.85	8.12
WA at OCT (%)	92	0.62	23.48	11.90	20.42	4.52
OCT (min)	250	10.00	60.00	33.71	189.24	13.76
DM(%) 30'	90	23.25	42.71	34.35	13.77	3.71
Gradient (kg/mm)	195	170	2489	1205	219806	469
Max force (kg)	195	8473	39761	22162	39958472	6321
Distance at Max force (mm)	195	12.90	20.00	18.23	3.96	1.99
Area (kg.mm)	195	83473	489697	252090	7829955416	88487
Linear Distance (mm)	195	9503	53235	25349	68069039	8250
End force (kg)	195	8469	39422	21195	34601211	5882
End force_Max force (%)	195	74.92	100.00	96.15	25.66	5.07

A first observation confirms the previous results (T. Tran, H. Ceballos, D. Dufour, J. Belalcazar) the physical properties for a same genotype are highly dependent of date of harvest, while DM of fresh root remains almost content. As an example genotype CM7436-7, harvested 8 times:

ref	Genotype	date	DM(%) fresh	WA10 (%)	WA20 (%)	OCT (min)
M00320_001	CM7436-7	29/01/2020	39.9	3.62	6.71	37.47
M00220_04	CM7436-7	16/01/2020	41.1	1.93	5.40	55.39
M00221_01	CM7436-7	13/01/2021	36.7		3.55	51.67
M01620_01	CM7436-7	24/02/2020	38.3	2.63	3.29	49.32
M02921_01	CM7436-7	25/03/2021	37.4		2.49	47.78
M01721_01	CM7436-7	01/03/2021	36.9		3.15	28.89
M09019_04	CM7436-7	12/11/2019	41.6	9.24	15.58	21.30
M10219_04	CM7436-7	11/12/2019	42.7	1.85	3.84	55.06

2 RESULTS

2.1 Dry matter

A total of 250 quantification of DM was done, the average DM value is 39,5% with a range of: 31% to 47,3%. DM is quite constant between genotypes and over years.

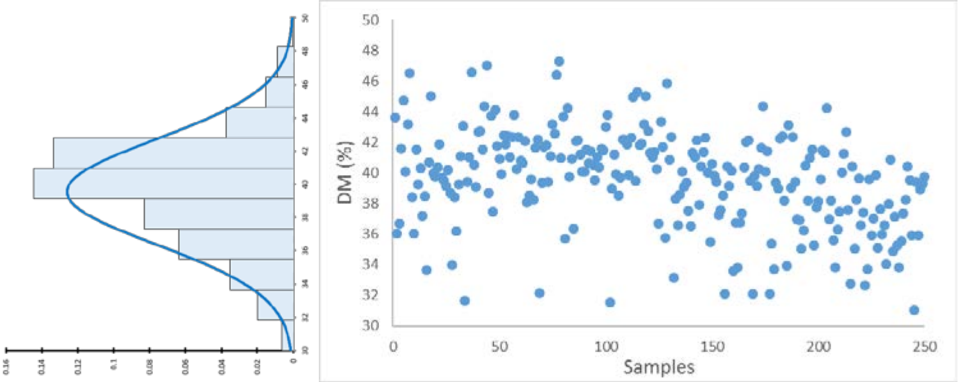


Figure I: Dry matter distribution

2.2 Optimal Cooking Time (OCT)

The average value of the 250 determinations of OCT is 33,7 min, the values range from 10 to 60 minutes. Two classes are defined C1 for OCT lower than 33,7 min and C2 for OCT higher or equal to 33,7 min.

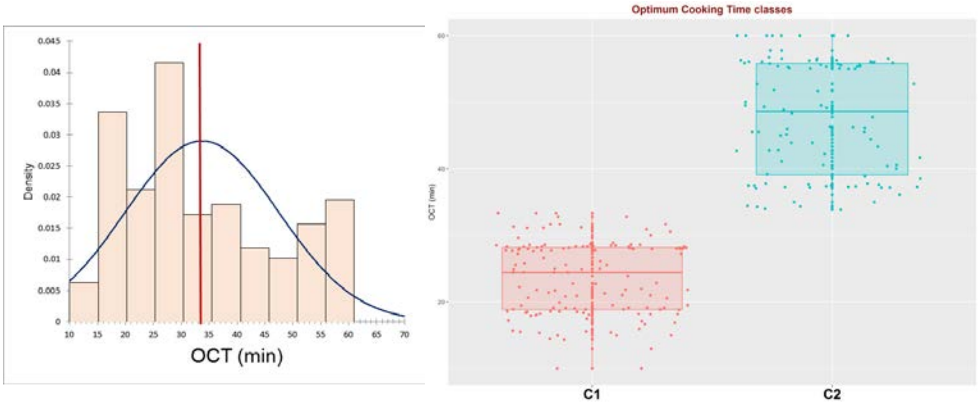


Figure II: Optimal cooking time distribution and box plots for each class

The descriptive statistics for the 2 OCT classes are:

	Statistique	N	Minimum	Maximum	Mean	SD
DM	C1	146	31.65	46.49	40.03	2.77
	C2	104	31.01	47.27	38.85	3.55
WA10	C1	95	0.88	13.30	5.22	2.88
	C2	65	-0.99	7.71	2.52	1.61
WA20	C1	146	1.45	40.52	9.29	6.23
	C2	104	-0.78	8.46	3.57	1.99
WA30	C1	114	5.27	51.13	18.11	8.95
	C2	100	-0.22	20.19	7.14	4.24
WA40	C1	25	2.16	26.89	17.32	4.73
	C2	10	0.81	14.81	7.58	4.34
WA50	C1	25	-3.81	31.37	20.08	6.61
	C2	10	2.05	17.14	10.20	5.19
WA60	C1	25	-4.00	33.91	21.90	7.36
	C2	10	3.34	20.73	12.51	5.86
WA at OCT	C1	59	1.69	20.51	11.53	3.88
	C2	33	0.62	23.48	12.56	5.48
OCT	C1	146	10.00	33.33	23.67	5.53
	C2	104	33.89	60.00	47.79	8.45
DM 30'	C1	51	23.25	41.11	33.17	3.79
	C2	39	28.51	42.71	35.90	3.01
Gradient	C1	114	170.1	1875.4	959.7	362.8
	C2	81	677	2489	1551	375
max force	C1	114	8473	39704	20292	5904
	C2	81	14075	39761	24793	5972
Distance	C1	114	13.19	20.00	19.02	1.45
	C2	81	12.90	20.00	17.11	2.12
area	C1	114	83473	455088	210318	65882
	C2	81	147567	489697	310879	82891
linear Distance	C1	114	9503	43723	22174	7102
	C2	81	14898	53235	29817	7698
end force	C1	114	8469	39422	19870	5651
	C2	81	13989	38512	23061	5724
end force at max dist	C1	114	86.73	100.00	98.25	2.86
	C2	81	74.92	100.00	93.21	5.97

DM, WA20, WA30, gradient, Area and End force:max force show different average values according to the 2 classes of OCT:

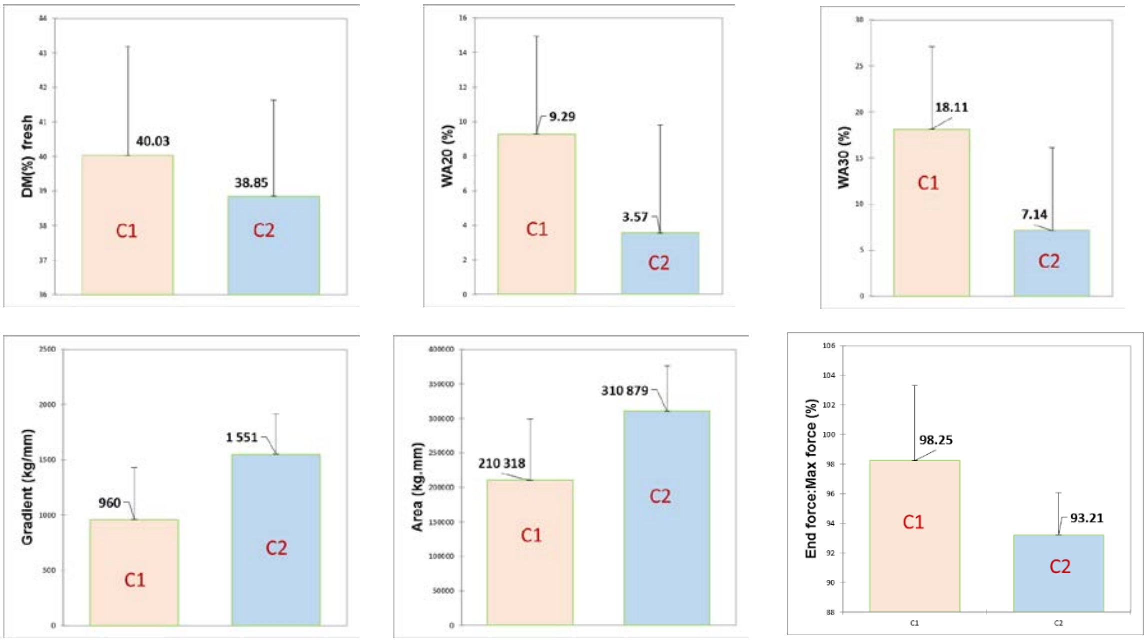


Figure III: Mean values of DM, Water absorption, Gradient, Area and End Force at max force per classes

2.3 Water Absorption (WA20 and WA30)

Water absorption values at 10, 20, 30 are highly correlated. The number of value for 40 mn to WA at OCT is too low to do good interpretation, we focus here on WA20 (n = 250) and WA30 (n = 214). Regarding WA20, the distribution of the values is narrow with an average value of 6,9 % and a SD = 5,67%, the maximum value observed was 40,52%. The distribution of value for WA30 is larger around an average value of 12,98%, with a maximum value of 51,13%.

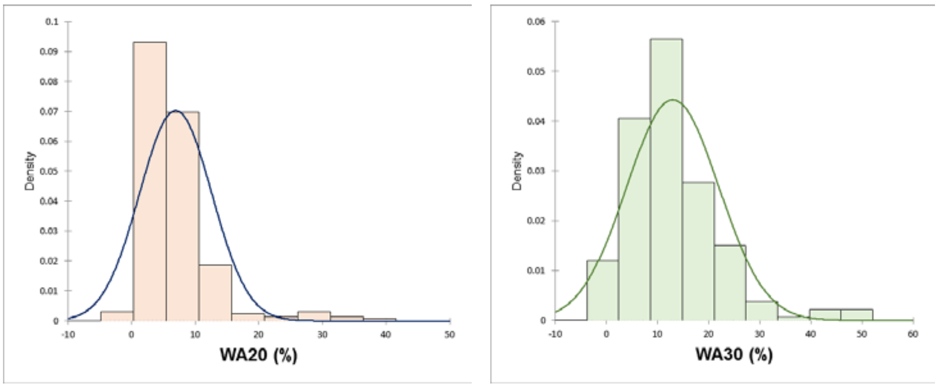


Figure IV : Histograms of Water absorption at 20 and 30 minutes

There is a relation (nonlinear, exponential or polynomial), between WA20 or WA30 and OCT, high cooking time genotypes absorb less water at 20 mn than “good cooking” genotypes. The values observed for the same genotype (here CM7436-7) from different harvest dates present a high dispersion for OCT (between 21 min to 55 min) and a quite low dispersion for WA20 and WA30 with a maximum in both cases of 15%. For WA20 this maximum corresponds to 21 min of OCT and for WA30 this maximum absorption corresponds to 37 min of OCT.

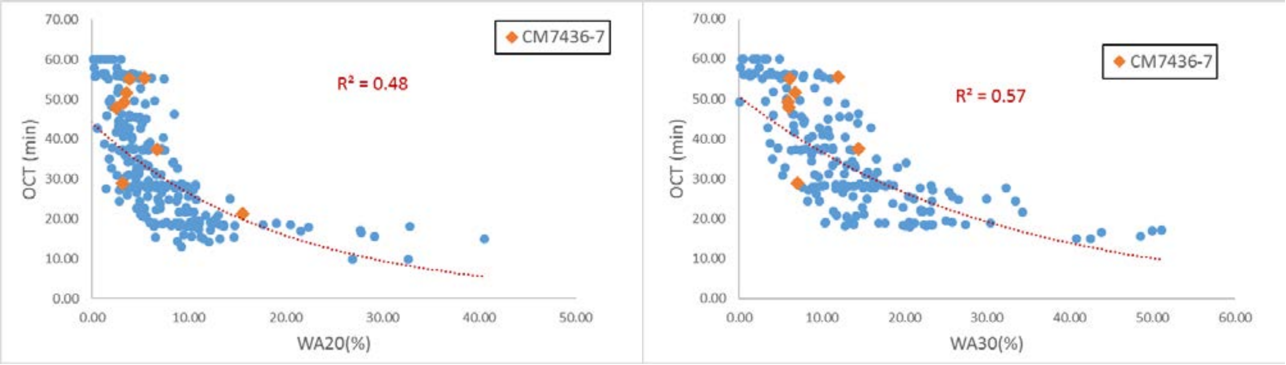


Figure V : Relation between Water absorption and Optimal cooking time

2.4 Gradient

The 195 values of gradient range between 170 and 2488 kg/mm with an average value of 1205 kg/mm. The distribution of the values follows a normal law. Gradient is highly correlated to physical values related to Max force, Area and Linear distance.

Gradient is also correlated to OCT (r = 0,72) and Water Absorption: r (WA30/Gradient) = -0,69.

The relation between gradient and WA_30 is nonlinear (second order), genotypes with high gradient values absorb less water at 30 min than genotypes with low gradient values which correspond to genotypes with low optimum cooking time.

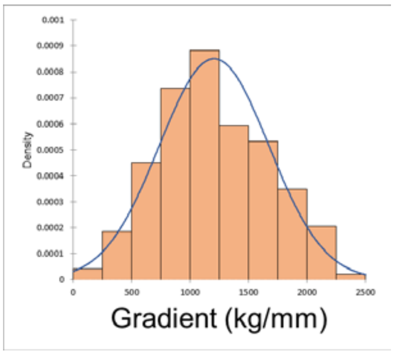


Figure VI : Histogram for Gradient

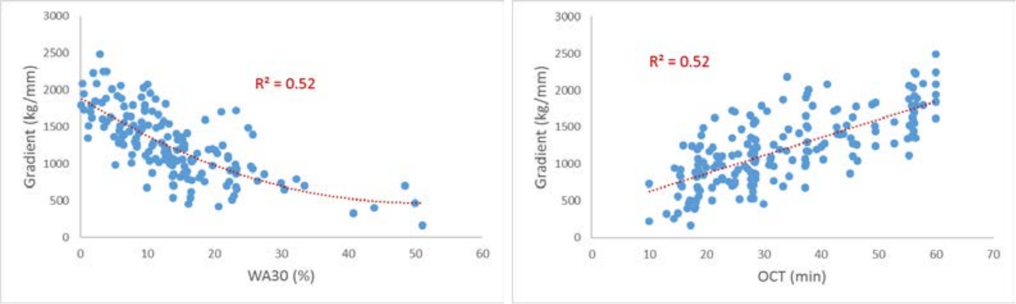


Figure VII: Relation between Gradient and Water absorption at 30 min and OCT

The relation between gradient and OCT presents a linear trend (R² = 0.52) with a high dispersion (independent of time of cooking) of gradient values for each cooking time. Globally, the value of gradient increase with the value of OCT, the genotypes with high OCT presents high values of gradient.

2.5 Area

The 195 values of area range between 83472 and 489696 kg.mm with an average value of 252089 kg.mm. The distribution of the values follows a normal law. Area is highly correlated to physical values related to gradient, max force, Linear distance (r = 0.92) and end force.

Area is also correlated to OCT (r = 0,70) and Water Absorption : r (WA30/Gradient) = -0,65.

The relation between area and WA_30 is nonlinear (second order), genotypes with high area values absorb less water at 30 min than genotypes with low area values which correspond to genotypes with low optimum cooking time.

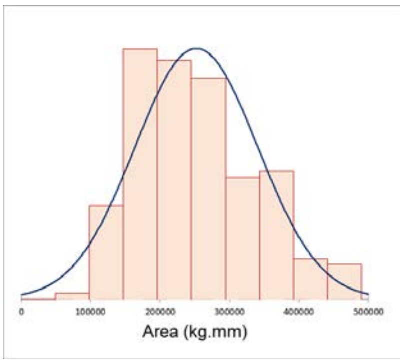


Figure VIII : Histogram for Area

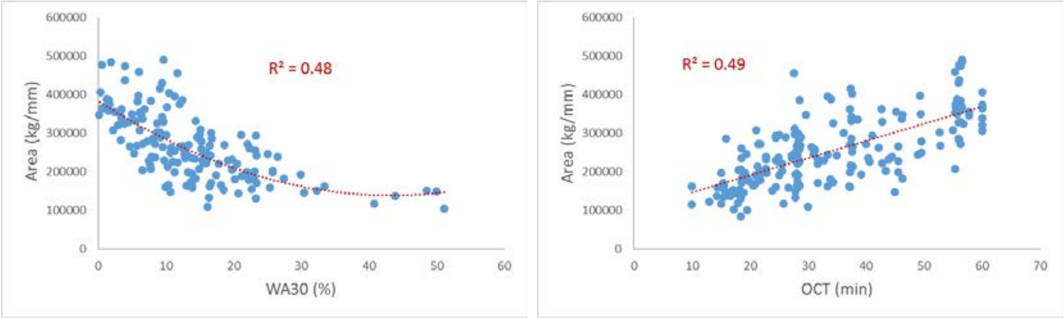


Figure IX : Relation between Area and Water absorption at 30 min and OCT

The relation between area and OCT presents a linear trend (R² = 0.49) with a high dispersion (independent of time of cooking) of area values for each cooking time. Globally, the value of area increase with the value of OCT, the genotypes with high OCT presents high values of area.

2.6 Correlations

The DM content of fresh material is correlated with WA at OCT (r = 0.44), there is no correlation with OCT (r = -0.09).

Water absorption is correlated with OCT, the highest correlation (r = -0.72) observed corresponds at 30 min of cooking. Water absorption after 20 and 30 minutes of boiling are correlated to texture parameters, with a maximum correlation (r = -0.69) between WA30 and gradient

OCT is correlated to textural parameters, the highest correlation is with gradient (r = 0.72).

Correlation matrix

Variables	DM(%) fresh	WA10 (%)	WA20 (%)	WA30 (%)	WA40 (%)	WA50 (%)	WA60 (%)	WA at OCT (%)	OCT (min)	DM(%) 30'	Gradient (kg/mm)	Max force (kg)	Distance at Max force (mm)	Area (kg.mm)	Linear Distance (mm)	End force (kg)	End force:Max force (%)
DM(%) fresh	**1**	**0.254**	**0.339**	**0.262**	0.137	0.097	0.102	**0.444**	**-0.196**	0.202	**-0.217**	0.020	**0.365**	-0.129	-0.081	0.085	**0.363**
WA10 (%)	**0.254**	**1**	**0.845**	**0.721**	0.118	-0.100	-0.150	0.156	**-0.610**		**-0.601**	**-0.494**	0.154	**-0.643**	**-0.534**	**-0.474**	**0.313**
WA20 (%)	**0.339**	**0.845**	**1**	**0.827**	0.105	-0.116	-0.172	**0.269**	**-0.625**	**-0.634**	**-0.659**	**-0.394**	**0.340**	**-0.595**	**-0.475**	**-0.333**	**0.411**
WA30 (%)	**0.262**	**0.721**	**0.827**	**1**	**0.727**	**0.505**	**0.420**	**0.318**	**-0.720**	**-0.702**	**-0.693**	**-0.458**	**0.528**	**-0.649**	**-0.541**	**-0.372**	**0.458**
WA40 (%)	0.137	0.118	0.105	**0.727**	**1**	**0.940**	**0.877**		**-0.687**								
WA50 (%)	0.097	-0.100	-0.116	**0.505**	**0.940**	**1**	**0.981**		**-0.555**								
WA60 (%)	0.102	-0.150	-0.172	**0.420**	**0.877**	**0.981**	**1**		**-0.486**								
WA at OCT (%)	**0.444**	0.156	**0.269**	**0.318**				**1**	0.093		0.013	0.118	0.187	-0.032	0.113	0.156	**0.212**
OCT (min)	**-0.196**	**-0.610**	**-0.625**	**-0.720**	**-0.687**	**-0.555**	**-0.486**	0.093	**1**	**0.564**	**0.719**	**0.500**	**-0.470**	**0.698**	**0.587**	**0.416**	**-0.511**
DM(%) 30'	0.202		**-0.634**	**-0.702**					**0.564**	**1**	**0.573**	**0.384**	**-0.312**	**0.506**	**0.403**	**0.306**	**-0.266**
Gradient (kg/mm)	**-0.217**	**-0.601**	**-0.659**	**-0.693**				0.013	**0.719**	**0.573**	**1**	**0.631**	**-0.515**	**0.826**	**0.727**	**0.541**	**-0.581**
Max force (kg)	0.020	**-0.494**	**-0.394**	**-0.458**				0.118	**0.500**	**0.384**	**0.631**	**1**	**-0.194**	**0.875**	**0.960**	**0.978**	**-0.270**
Distance at Max force (mm)	**0.365**	0.154	**0.340**	**0.528**				0.187	**-0.470**	**-0.312**	**-0.515**	**-0.194**	**1**	**-0.445**	**-0.376**	-0.018	**0.860**
Area (kg.mm)	-0.129	**-0.643**	**-0.595**	**-0.649**				-0.032	**0.698**	**0.506**	**0.826**	**0.875**	**-0.445**	**1**	**0.921**	**0.802**	**-0.502**
Linear Distance (mm)	-0.081	**-0.534**	**-0.475**	**-0.541**				0.113	**0.587**	**0.403**	**0.727**	**0.960**	**-0.376**	**0.921**	**1**	**0.898**	**-0.444**
End force (kg)	0.085	**-0.474**	**-0.333**	**-0.372**				0.156	**0.416**	**0.306**	**0.541**	**0.978**	-0.018	**0.802**	**0.898**	**1**	-0.073
End force:Max force (%)	**0.363**	**0.313**	**0.411**	**0.458**				**0.212**	**-0.511**	**-0.266**	**-0.581**	**-0.270**	**0.860**	**-0.502**	**-0.444**	-0.073	**1**

The textural parameters are correlated, but not all, the highest correlations are: gradient/area (r = 0.82); gradient/linear distance (r = 0.72); max force/area (r = 0.87); max force/linear distance (r = 0.96) and max force/end force (r = 0.98)

2.7 Functional Properties: Principal Components Analysis

The total number of textural values is 195. The exploration of the data is done on the 195 individuals with values for DM, WA20, Gradient, Max, force, Distance at Max force, Area Linear Distance and End force Max force. A PCA is done on the 195 individuals with those data and OCT as supplementary variable.

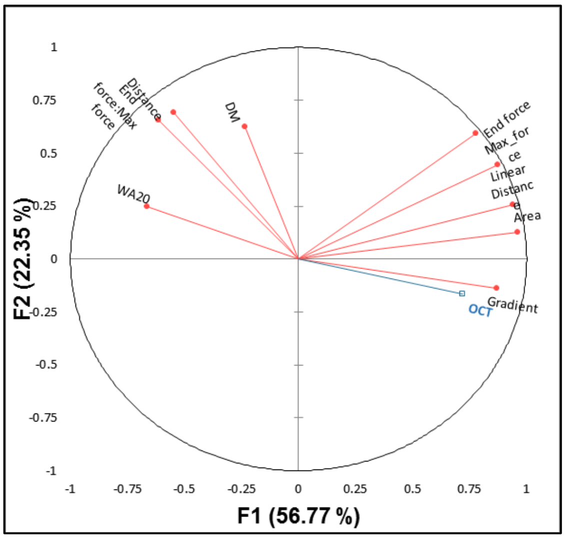


Figure X: Correlation circle for functional properties, variable OCT as supplementary

The observations of vectors confirm the opposition between textural parameters and Water absorption and the importance of those variables in construction of PC1. The projection of the supplementary variable (OCT) shows that these factorial plans give a good representation of the variable space. Samples with low values for texture parameters (Gradient, area, max force, linear distance) and high values for WA at 20 min will have a short cooking time. And samples with intermediate values and high values for DM will present intermediate cooking time. Dry matter content of fresh cassava participates to PC2 construction, however the values OCT are not explained by PC2

The repartition of individuals on the first principal plan shows a repartition according to cooking classes (C1 and C2) along PC1, but without a clear separation between the two classes

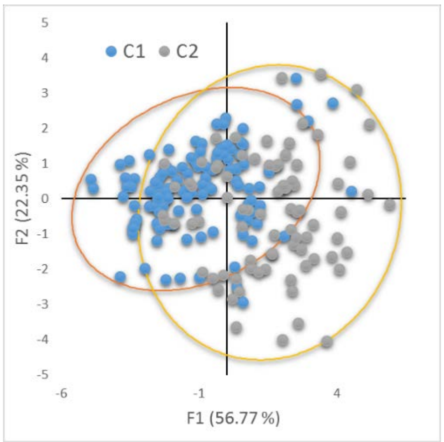


Figure XI: PCA scores plot of the 195 samples

2.8 Functional properties: Factorial Discriminant Analysis

The variables DM, WA20, WA30 and textural variables are used to run a Factorial Discriminant Analysis on the 159 individuals for the two classes of OCT. the data set is separate randomly in 2 groups: learning (112) and validation (47)

Confusion Matrix for validation.

From \To	C1	C2	Total	% correct
C1	22	1	23	95.65%
C2	6	18	24	75.00%
Total	28	19	47	85.11%

With a specificity of 95,6%, a sensibility of 75% and a correct classification of 85%, this analysis confirms that the parameters quantified or measured are relevant for classification of genotypes according to their ability to cook.

2.9 Near Infrared Spectroscopy

2.9.1 Exploration

The spectra patterns are similar for the 3 dates of harvest, there is no atypical spectrum. Somme of the genotypes presents a specific absorption at 460 nm and 480 nm due to colour.

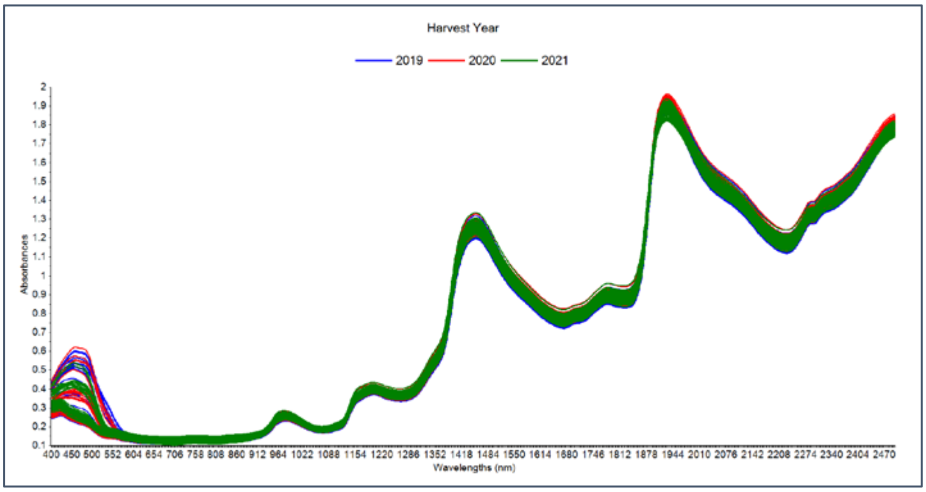


Figure XII : Absorbance spectra of the 250 samples: 400 nm to 2500 nm

2.9.2 Spectra: Principal Components Analysis

A PCA calculated on the spectra (spectral range NIR) of the samples (160) from 2019 and 2020 led to 97,5% of variance explained by the 2 first PCs. The projection of the 90 samples harvested in 2021 on the PCs scores highlights a difference between both populations. The Mahalanobis distances of these samples from 2021 are on average equal to 4 with a maximum of 13,5. There are 55 samples out of 90 that present a distance higher than the distance limit (3).

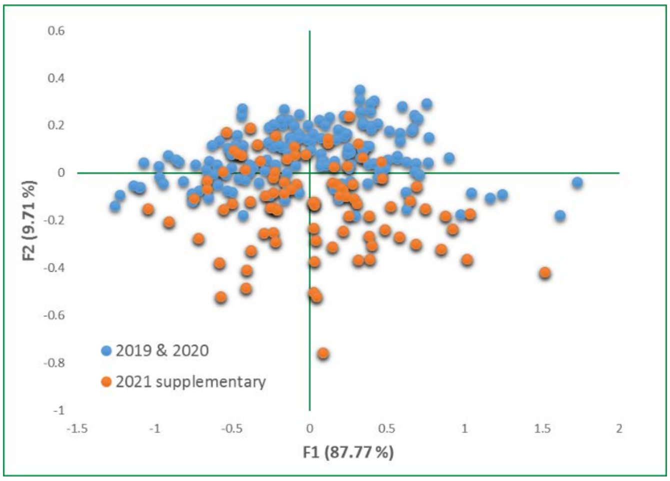


Figure XIII : Scatter plot of the PCA scores of the 160 samples (2019 & 2020) and projection of the 2021 samples (90).

This observation is confirmed by the results of PCA done all the samples (250) for the NIR range: there is a trend among PC2 scores according to date of harvest (year).

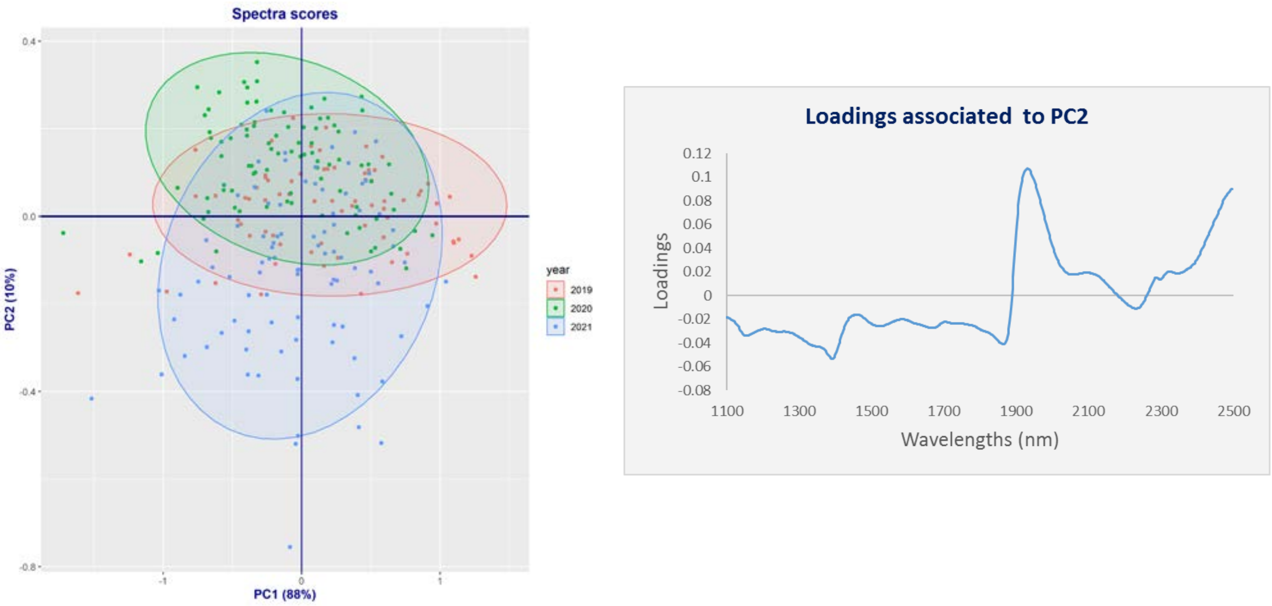


Figure XIV : Scatter plot of the PCA scores of the 250 samples and loadings for PC2

According to the loadings, this trend among PC2 is mainly due to dry matter contents which differs from years to years: the average DM in 2019 and 2020 was respectively equal to 40,39% and 40,41%, the average DM was 38,00% in 2021.

2.9.3 Quantitative analysis

The different parameters were calibrated using classical linear regression such as PLS regression, different pre-treatments were tested and best models with highest R², lowest SEC and SECV, minimum PLS factors and highest ratio SEC/SECV were retained.

2.9.4 Statistics parameters for calibrations:

The statistic parameters show that the only relevant calibration is for DM content (R² = 0,96, SEP = 0,71%), for others parameters calibrations are very weak with a cross validation error (SECV) close to SD. The linear approach whatever the pre-treatments will not allow efficient calibrations for WA, OCT and physical parameters. Even if some trends are observed for WA20, WA30, OCT and Gradient parameters.

Constituent	N	Mean	SD	SEC	R²cal	SECV	R²cv	#PLS	#outliers	SEP	maths	segment
DM	243	39.5911	3.1642	0.6288	0.9605	0.7365	0.9456	8	7	0.711	1,4,4	1100-2500
WA20	236	5.8429	3.2542	2.5703	0.3762	2.8016	0.2557	5	14	5.084	1,4,4	400-2500
WA30	205	11.8842	6.7903	5.1307	0.4291	6.0031	0.2146	5	9	7.201	1,4,4	400-2500
OCT	245	33.2761	13.5563	9.4341	0.5157	10.5716	0.3894	7	5	10.075	1,4,4	400-2500
Gradient	192	1195.5142	464.079	329.3516	0.4963	354.1216	0.4147	7	3	345	1,4,4	400-2500

As illustrated for the water absorption at 30 minutes, the best fitting is nonlinear (polynomial order 2 or superior).

2.10 Classification using spectra

Whatever the pretreatments used (SNV, SNVD, first or second derivative…) and whatever the classification approaches, supervised or not, (K Nearest Neighbors, Support Vector Machine, Naive Bayesian Classifier, Random Forest, Classification Regression Trees…), the predictions of a validation set, for the 2 cooking time classes failed.

Models were able to find patterns within the learning sets, but were unable to predict new independent samples.

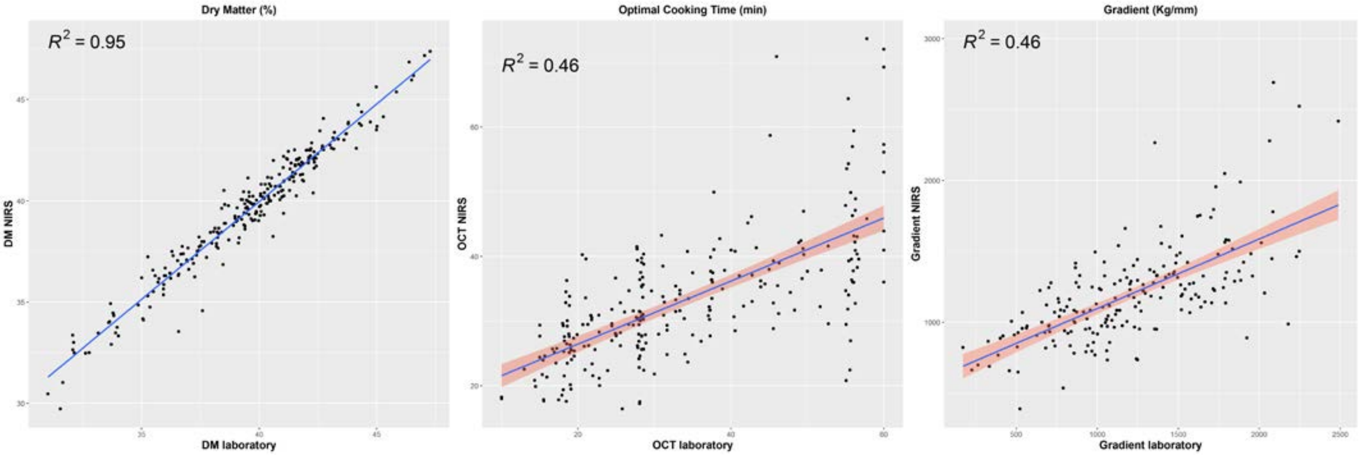


Figure XV : Scatter plots between laboratory and predicted values for DM, OCT and Gradient

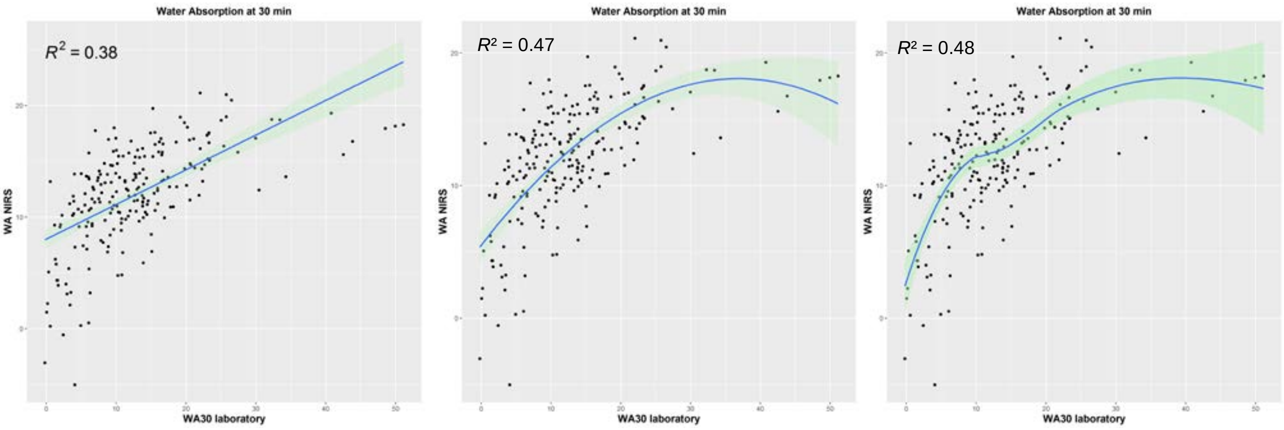


Figure XVI : Scatter plot between laboratory and predicted values for WA at 30 min: linear and polynomials fittings

2.10.1 Classification based on Lasso regression

LASSO stands for Least Absolute Shrinkage and Selection Operator. The LASSO regression was proposed by Robert Tibshirani[1] in 1996. It is an estimation method which forces its coefficients not to explode, unlike the standard linear regression in large dimensions. The large-scale context covers all the situations where there are a very large number of variables in relation to the number of individuals.

LASSO regression is one of the methods which overcomes the shortcomings (instability of the estimate and unreliability of the forecast) of linear regression in a large-scale context. The main advantage of LASSO regression is its ability to perform variable selection, which can be valuable when there are a large number of variables.

The Lasso regression was ran using Xlstat v. 2021.4.1(Addinsoft (2021). XLSTAT statistical and data analysis solution. Paris, France. https://www.xlstat.com/fr) using OCT as dependent variable and first derivative (Norris gap derivative) of spectral data corrected for baseline shift using SNVD correction.

First the lasso parameters were optimized on the full dataset (N = 250) using 5 blocks for cross validation with 100 values tested. The R² of the regression was 0.58 with a RMSE (Root of the Mean of the Squares of the Errors) equal to 9,08 min.

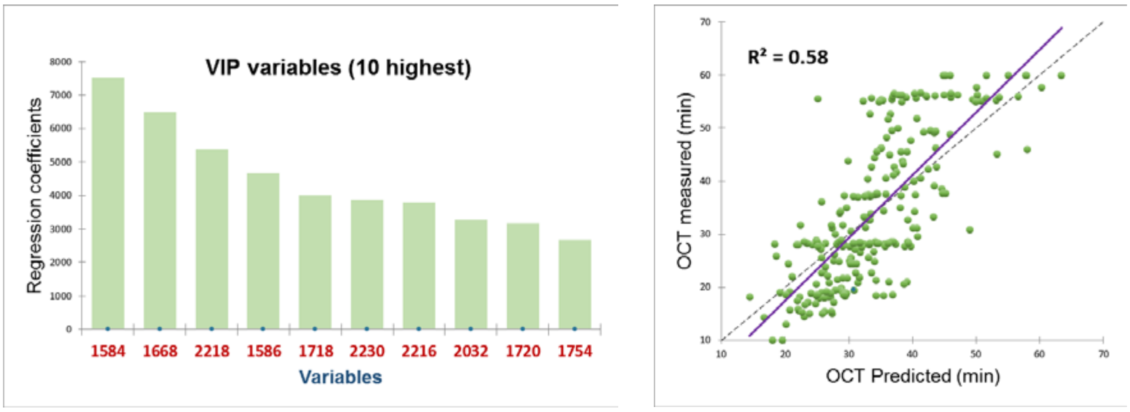


Figure XVII : Ten highest relevant variables of the model and scatter plot of predicted vs laboratory values

Based on this regression and according to the rule: class C1 corresponds to OCT < 33,7 min a class C2 corresponds to OCT ≥ 33,7 min, the classification rate based on OCT predicted values is 81%.

[1] Jerome Friedman, Trevor Hastie et Rob Tibshirani (2008). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Volume 2.

Validation using random selection

In order to validate the Lasso approach 2 sets of data were set up using a stratified random selection (In each stratum, the number of sampled observations is proportional to the frequency of the stratum). The selection was fixed to 70% of the samples which led to 175 samples in the learning set and 75 samples in validation set. The classification rate for learning set was 82% with a R² = 0,58 and a RMSE = 9.08 min for the Lasso regression.

The classification rate is 72% for the validation set with the following repartition:

from\to	*C1*	*C2*	*N*	*Rate*
C1	**33**	**11**	**44**	**75%**
C2	**10**	**21**	**31**	**68%**
			75	**72%**

The R² for validation set was equal to 0,33 with a RMSE = 11,3 minutes.

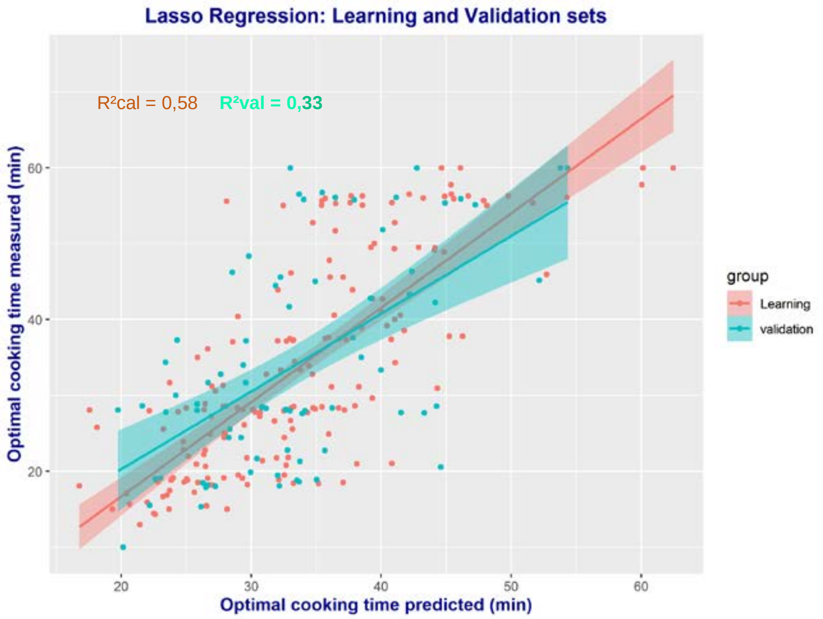
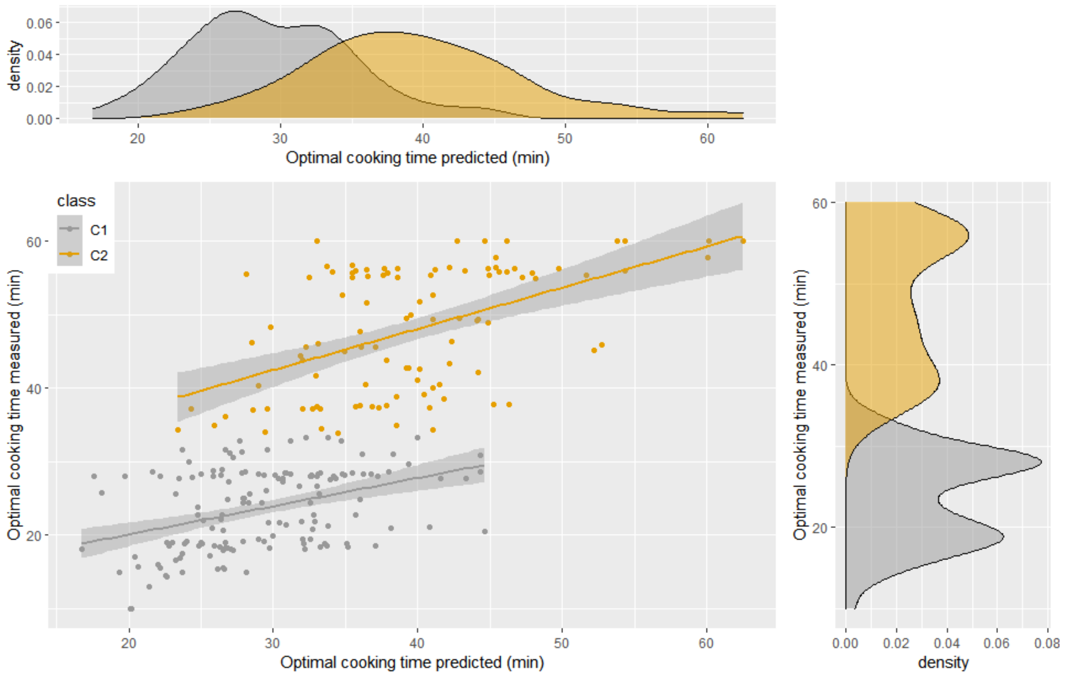


Figure XVIII : Scatter plot of predicted vs laboratory values of OCT per set.



The scatter plot of measured values of OCT vs predicted values from Lasso regression per class with the marginal distributions highlights that 1) the fit is similar for both classes, 2) The distribution of OCT predicted values is quite different from original distribution, (especially for C2) and 3) The average predicted values of OCT for C1 and C2 are significantly different.

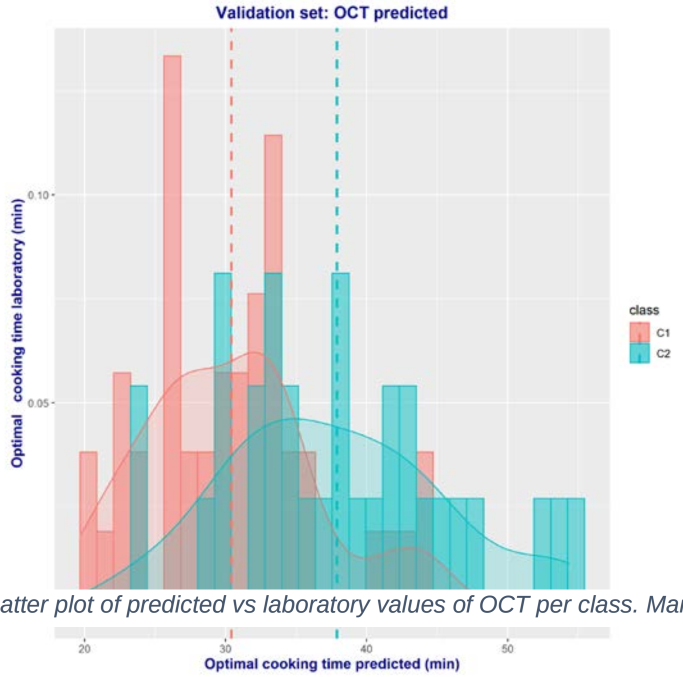


Figure XIX : Scatter plot of predicted vs laboratory values of OCT per class. Marginal distributions

Figure XX : Distributions of OCT predicted values for the two classes

One factor ANOVA done on OCT predicted values with the factor class led for the 75 validation data to an R^2 =0.22 and a p-value 0.00002. The mean values for each class are significantly different at level α =5%.

A false positive sample corresponds to a predicted class C1 (OCT < 33,7 min) when the measured class is C2 (OCT ≥ 33,7 min). Which means that we select a genotype which is actually “bad”.

According to this, 10 samples out of the 75 validation samples are false positives, the minimum OCT predicted values for these 10 samples is 23,4 minutes (corresponding to a sample with a measured OCT of 34 min). The average OCT predicted value for these 10 samples is 29,5 minutes.

False positives samples

Genotype	Ref	OCT	Pred(OCT)	Résidu	class	Class_predicted
CM2600-2	M09019_01	34.00	29.39	4.61	C2	C1
GM8413-1	M10219_28	37.18	29.61	7.56	C2	C1
PAN139	M00220_16	46.23	28.55	17.68	C2	C1
SM3759-36	M00220_29	34.36	23.40	10.96	C2	C1
COL1722	M00220_33	37.28	24.30	12.98	C2	C1
CM6370-2	M00221_18	41.67	32.93	8.73	C2	C1
IND27	M00221_19	48.33	29.82	18.51	C2	C1
VEN25	M00221_30	60.00	33.00	27.00	C2	C1
VEN208	M02921_05	44.44	31.88	12.56	C2	C1
BRA158	M02921_23	45.56	32.25	13.31	C2	C1

The classification rate of the validation set is 72%, with 10 false positives samples which are quite close to the limit of the class (in terms of OCT). This result is encouraging for a selection of samples with short cooking times, and the “bad cooking” samples wrongly selected on predicted values present relatively short cooking times (maximum of 44 minutes) except for reference M00221_30 (genotype: *VEN25*) with an OCT = 60 minutes and a prediction at 33 min.

At the opposite, 11 false negative samples (rejected even though they are “good”) were not selected according to the regression model. This led to miss 11 interesting genotypes out of the 75 validation samples. The range of OCT for these 11 samples was: 21,3 min – 33,3 min, while the range of the predictions for the same samples was: 33,7 min – 44,6 min.

Validation using 2021 samples

The learning set is based on 2019 & 2020 samples (N = 160), samples harvested and analyzed in 2021 (n = 90) are kept as external validation set. The classification rate for learning set was 81% with a R² = 0,55 and a RMSE = 9.85 min for the Lasso regression

The classification rate is 66% for the validation set with the following repartition:

from\to	***C1***	***C2***	***N***	***Rate***
C1	**37**	**14**	**51**	**73%**
C2	**17**	**22**	**39**	**56%**
			90	**66%**

The R² for validation set was equal to 0,26 with a RMSE = 11,4 minutes.

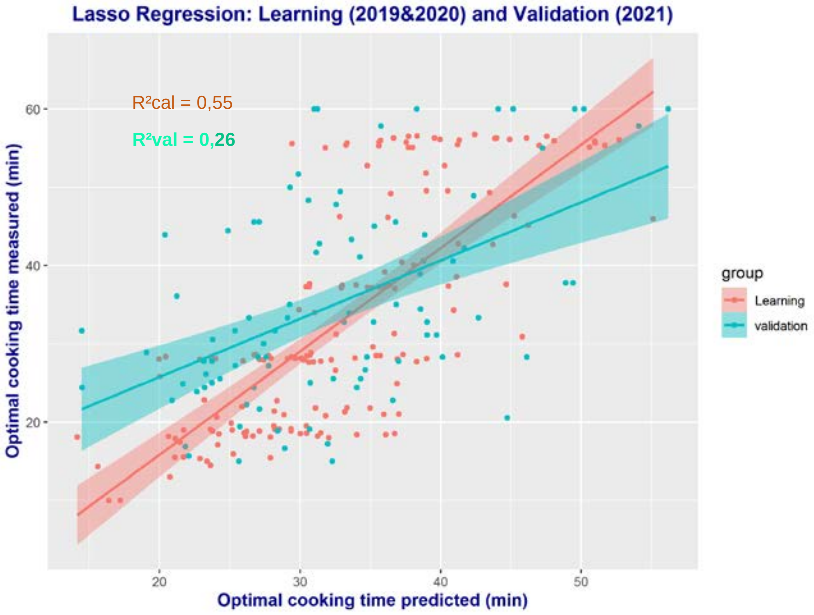


Figure XXI : Scatter plot of predicted vs laboratory values of OCT per set

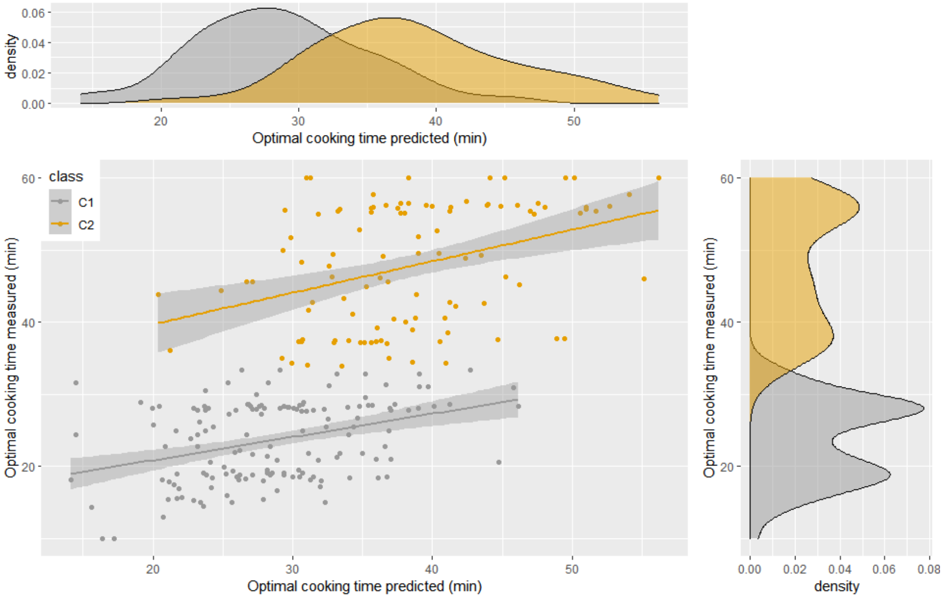


Figure XXII : Scatter plot of predicted vs laboratory values of OCT per class. Marginal distributions

Here using 2021 samples for validation led to lower performances for discrimination, this is probably due to the differences seen in spectra between 2019, 2020 and 2021. And the distribution of OCT predicted values according to classes are slightly different from original distributions.

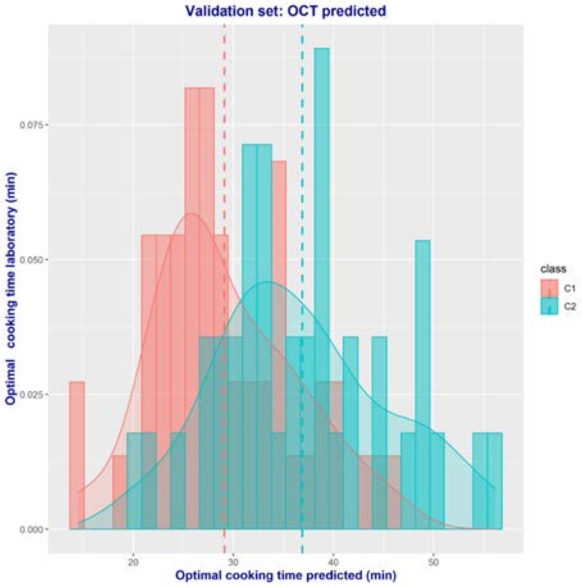


Figure XXIII : Distributions of OCT predicted values for the two classes

The mean OCT average value for C1 is 29,1 min and the for C2 the mean average value is 36,9 minute. There is an important overlapping between both distributions. However a one factor ANOVA led for the 90 validation data to an R² =0.66 and a p-value 0.0001. The mean values for each class are significantly different at level α =5%.

17 samples out of the 90 validation samples are false positives, the minimum OCT predicted values for these 10 samples is 20,5 minutes (corresponding to a sample with a measured OCT of 43,9 min). The average OCT predicted value for these 10 samples is 29,3 minutes. At the opposite, 14 false negative samples (rejected even though they are “good”) were not selected according to the regression model. This led to miss 14 interesting genotypes out of the 90 validation samples. The range of OCT for these 14 samples was: 24,4 min – 33,3 min, while the range of the predictions for the same samples was: 34,0 min – 46,1 min.

False positives samples

Reference	Genotype	OCT	Pred(OCT)	class	class_lasso
M02921_23	BRA158	45.6	27.1	C2	C1
M02921_06	CM5948-1	50.0	29.3	C2	C1
M00221_18	CM6370-2	41.7	31.2	C2	C1
M00221_01	CM7436-7	51.7	29.9	C2	C1
M02921_01	CM7436-7	47.8	32.6	C2	C1
M00221_08	COL1505	35.0	29.3	C2	C1
M01721_08	COL1505	42.8	31.4	C2	C1
M02921_08	COL1505	43.3	33.7	C2	C1
M00221_25	COL2089	49.4	32.9	C2	C1
M02921_25	COL2089	60.0	31.3	C2	C1
M01721_09	COL2246_9	43.9	20.4	C2	C1
M02921_29	CUB46	33.9	33.5	C2	C1
M01721_02	GUA24	45.6	26.7	C2	C1
M00221_19	IND27	48.3	30.6	C2	C1
M02921_19	IND27	36.1	21.2	C2	C1
M02921_05	VEN208	44.4	24.9	C2	C1
M00221_30	VEN25	60.0	31.0	C2	C1

The classification rate of the validation set is 66%, with 17 false positives samples which are quite close to the limit of the class. This independent validation result is encouraging for developing a robust and reliable predicted model based on spectra and Lasso Regression.

Among these 17 false positive samples, 4 genotypes are replicated: CM7436-7 (2); COL1505 (3); COL2089 (2); IND27 (2). A closer look to predicted OCT versus measured OCT indicates that predicted values for replicates of same genotype are less spread than measured values e.g. *genotype COL2089 OCT predicted 32,9 and 31, 2 min vs OCT measured 49,4 and 60 min*.

3 CONCLUSION

The high performances of the calibration for dry matter quantification confirms the representativeness of the spectra and the perfect link between spectra and laboratory analysis.

The parameters quantified in the laboratory are relevant and clearly linked to genotype cooking ability.

Linear approaches (Global or Local) do not help for developing efficient models for WA, OCT and physical properties quantification using spectral information.

Classification methods (supervised or unsupervised) using spectral fingerprints did not applied.

The Lasso regression applied to spectral data as explicative variables and OCT as dependent variable improves the fitting, with a $R^2 = 0,58$ and a RMSE equal to 9,08 min.

The classification rate of the samples into 2 cooking time classes (C1 < 33,7 min and C2 ≥ 33,7 min) was 82 % using OCT predicted values by Lasso regression for the learning set.

The classification rate was 72 % for validation samples randomly selected among the 250 samples (3 harvest year).

The classification rate was 66% when samples from 2021 (n = 90) were predicted using model based on 2019 and 2020 samples (n = 160).

The Lasso approach is encouraging, the model lacks robustness, because of a relatively few numbers of samples. And, the PCA exploration done on spectra highlighted the differences among spectra according to year, especially for 2021 samples.

These results confirm that the spectral signature contains information about textural properties and that nonlinear models or deep learning approaches could help extracting this information

.

Institute:	Cirad – UMR QualiSud
Address:	C/O Cathy Méjean, TA-B95/15 - 73 rue Jean-François Breton - 34398 Montpellier Cedex 5 - France
Contact Tel:	+33 4 67 61 44 31
Email:	rtbfoodspmu@cirad.fr
Website:	https://rtbfoods.cirad.fr/