Laboratory Scientific Report

NIRS & Biophysical Analyses on Cassava Cooking Properties Report

High-Throughput Phenotyping Protocols (HTPP), WP3

Saint-Pierre, Reunion, France, November 2020

Fabrice DAVRIEUX, Centre de coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), Saint Pierre, Reunion, 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, (2021). *NIRS & Biophysical Analyses on Cassava Cooking Properties Report. High-Throughput Phenotyping Protocols (HTPP), WP3.* St Pierre, Reunion, France: RTBfoods Laboratory Scientific Report, 18 p. http://doi.org/10.18167/agritrop/00717

Ethics: The activities, which led to the production of this document, 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) Xiaofei ZHANG (Alliance Bioversity-CIAT)	16/11/2020 18/11/2020 25/11/2020
Final validation by:	
Fabrice DAVRIEUX (CIRAD)	Date 09/02/2021

CONTENTS

Table of Contents

Table of Figures

ABSTRACT

This scientific report concerns data analysis of two matrices of measured data on fresh cassava 1) physico-chemical data and 2) spectral data. The data were collected on fresh cassava in CIAT, Colombia.

The analyses were performed using 87 cassava genotypes: 38 genotypes were analysed once, 37 analysed twice and 12 analysed 4 times. The total number of analyses is 160. 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 using near infrared spectroscopy. The absorption spectra were performed on ground samples of fresh roots using a FOSS 2500 spectrometer. The average dry matter is 40,4%, which is constant over months (age of the root). The mean value of cooking time is 33 min, ranging from 10 to 57 min. The wide distribution of cooking time allows to group the samples into 3 classes: less or equal to 25 min; higher than 25 min and lower or equal to 40 min; and higher than 40 min. Water absorption values at 10, 20 , 30 min are highly correlated (r >= ?). There is a non-linear relation between water absorption at 20 min and optimum cooking time: Genotypes with longer cooking time absorb less water at 20 min than “good cooking” genotypes. The values of gradient range between 225 and 2247 kg/mm with an average of 1179 kg/mm, which follows a normal distribution. Gradient is highly correlated to physical values related to Max force, Area and Linear distance. Gradient is also correlated to optimum cooking time (r = 0,735). The correlation between gradient and water absorption at 20 min of boiling is significant with r = –0,601, the highest correlation is at 40 min of boiling (r = -0,792). The relation between gradient and water absorption at 20 min of boiling is non linear (second order), genotypes with high gradient values absorb less water at 20 min of boiling than genotypes with low gradient values that showed 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 three cooking time classes was tested using different algorithms. Whatever were the pretreatments 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 3 cooking time classes failed.

The best classification method was obtained by doing a prediction of the scores of the discriminant axes calculated on six physico-chemical variables (DM, WA10, WA20, OCT, Gradient and distance at max force). The best classification was obtained for two cooking time classes: ≤ 30 min and > 30 min. The classification successful rate, for a validation set, was 80%.

The performances of the classification method which mix laboratory values and spectra values indicate that spectra contain relevant information related to cooking properties, and confirm that deep learning approaches may help for better and faster classification.

Keywords: Cassava, cooking properties, Near Infrared Spectroscopy, water absorption, classification, PLS regression

1 DATA

1.1 Material

The analyses concern 87 genotypes: 38 genotypes were analysed once, 37 analysed twice and 12 analysed 4 times. The total number of analyses is 160. Harvests took place in 2019 and 2020; the repartition of sampling is as follow:

Harvest date	November	December	January	February
2019	36	35		
2020			61	28

None of the replicates by clone were harvested at the same date.

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)
- 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

Statistique	N	Minimum	Maximum	Moyenne	Variance	Ecart-type
DM(%) fresh	160	31.55	47.27	40.40	8.72	2.95
WA10 (%)	160	-0.99	13.30	4.12	7.70	2.78
WA20 (%)	160	-0.78	32.82	7.26	26.75	5.17
WA30 (%)	124	-0.22	30.43	11.30	38.01	6.17
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)	160	10.00	56.72	32.94	203.93	14.28
Gradient (kg/mm)	125	225.23	2247.59	1179.69	205090.45	452.87
Max force (kg)	125	8472.89	39761.25	22917.08	46694305.66	6833.32
Distance at Max force (mm)	125	12.95	20.00	18.20	4.06	2.02
Area (kg.mm)	125	83472.78	489696.92	262407.18	9249361558.47	96173.60
Linear Distance (mm)	125	9502.83	53235.23	26157.84	82743381.25	9096.34
End force:Max force (%)	125	81.69	100.00	96.57	19.85	4.46

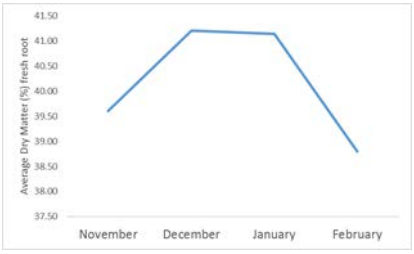
A first observation confirms the previous results (T. Tran, H. Ceballos, D. Dufour, J. Belalcazar) the physical properties for a same genotype are highly dependent of date of harvest, while DM of fresh root remains content. As an example genotype CM7436-7, harvested 4 times:

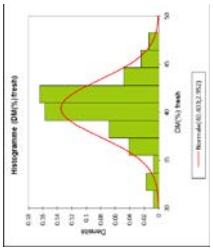
Genotype	date	DM(%) fresh	WA10 (%)	WA20 (%)	OCT (min)
CM7436-7	24/02/2020	38.28	2.63	3.29	49.32
CM7436-7	29/01/2020	39.89	3.62	6.71	37.47
CM7436-7	16/01/2020	41.06	1.93	5.40	55.39
CM7436-7	12/11/2019	41.58	9.24	15.58	21.30

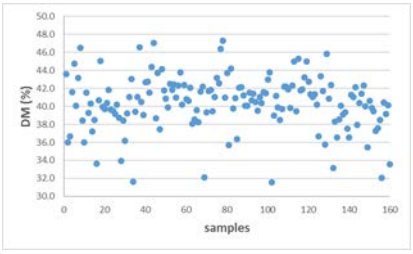
2 RESULTS

2.1 Dry matter

The average value is 40,4 %, the overall average DM is almost constant over months :

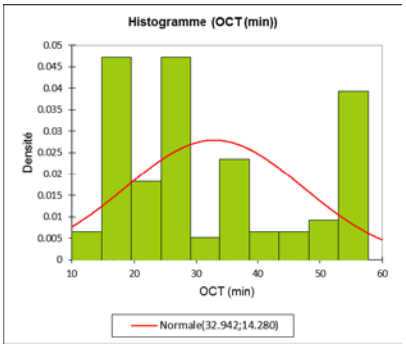


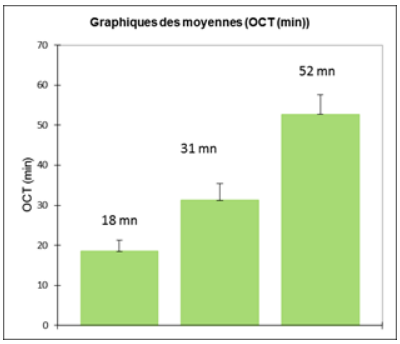




Optimal Cooking Time (OCT)

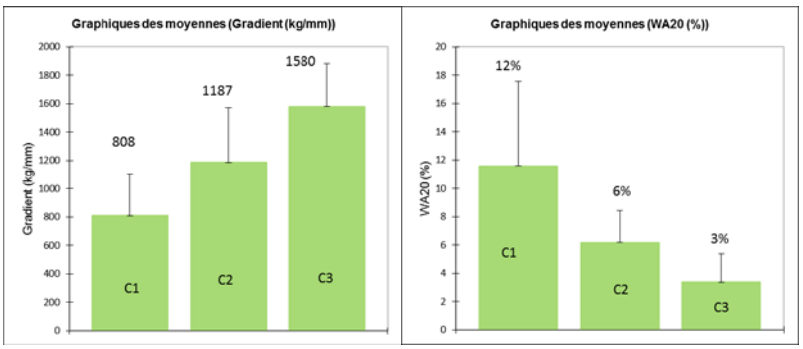
The average value is 33 min, the values range from 10 to 57 min. The distribution of the values, allows defining 3 classes: less or equal to 25 min, higher than 25 and lower or equal 40 and higher than 40 min.





The descriptive statistics for these 3 classes are:

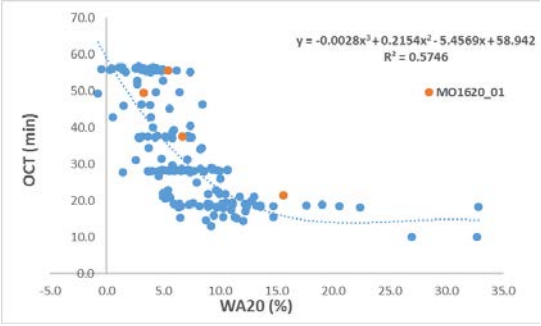
	Statistique	N	Minimum	Maximum	Moyenne	Variance (n-1)	Ecart-type (n-1)
DM	C1	56	33.96	45.28	40.63	6.32	2.51
	C2	58	31.65	47.03	40.67	8.42	2.90
	C3	46	31.55	47.27	39.78	11.84	3.44
WA10	C1	56	1.95	13.30	6.41	8.74	2.96
	C2	58	0.88	9.27	3.40	2.90	1.70
	C3	46	-0.99	5.67	2.25	2.12	1.46
WA20	C1	56	4.97	32.82	11.58	35.85	5.99
	C2	58	1.45	10.70	6.18	4.98	2.23
	C3	46	-0.78	8.46	3.36	4.15	2.04
WA30	C1	29	10.36	30.43	18.37	27.73	5.27
	C2	50	5.27	18.16	12.12	10.40	3.22
	C3	45	-0.22	14.45	5.82	12.60	3.55
OCT	C1	56	10.00	24.93	18.50	8.05	2.84
	C2	58	25.80	39.17	31.19	18.02	4.24
	C3	46	40.03	56.72	52.73	24.14	4.91
Gradient	C1	42	225.23	1626.29	808.54	88472.64	297.44
	C2	45	522.17	2179.51	1187.51	149145.12	386.19
	C3	38	1045.03	2247.59	1580.66	90366.89	300.61



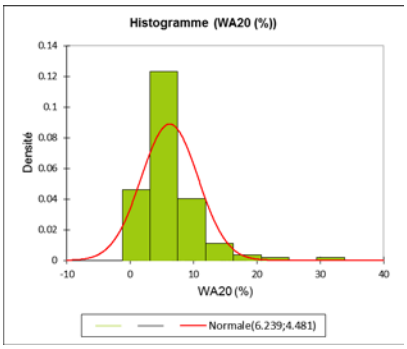
WA20 and gradient show different average values according to COT classes.

2.2 Water Absorption (WA_20)

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 WA_20 (n = 160). The distribution of the values is narrow with an average value of 7%.

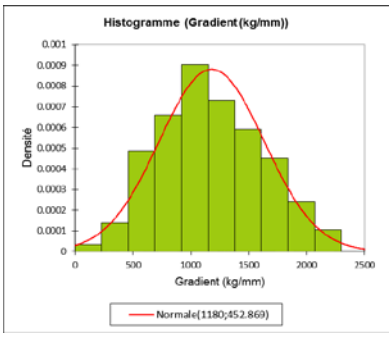


There is a relation (non linear, of order 2 or 3), between WA_20 mn and OCT, High time cooking genotypes absorb less water at 20 mn than "good cooking" genotypes.



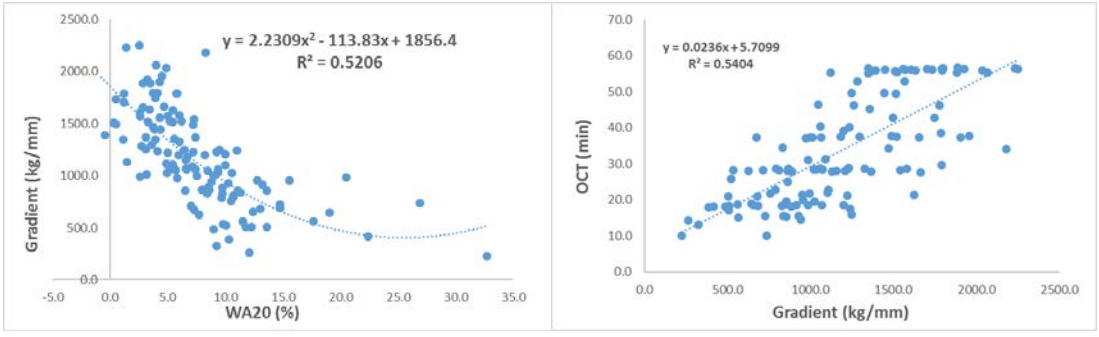
2.3 Gradient

The values of gradient range between 225 and 2247 kg/mm with an average value of 1179 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 = 0,735). The correlation between gradient and WA_20 is significant r = – 0,601, the highest correlation is at WA_40 (r = - 0,792).

The relation between gradient and WA_20 is non linear (second order), genotypes with high gradient values absorb less water at 20 mn than genotypes with low gradient values which correspond to genotypes with low optimum cooking time.



2.4 Correlations

Variables	DM(%) fresh	WA10 (%)	WA20 (%)	WA30 (%)	WA40 (%)	WA50 (%)	WA60 (%)	WA at OCT (%)	OCT (min)	Gradient (kg/mm)	Max force (kg)	Distance at Max force (mm)	Area (kg.mm)	Linear Distance (mm)	End force:Max force (%)
DM(%) fresh	1	0.254	0.272	0.324	0.137	0.097	0.102	0.444	-0.087	-0.072	0.004	0.372	-0.122	-0.063	0.305
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	0.313
WA20 (%)	0.272	0.845	1	0.829	0.105	-0.116	-0.172	0.269	-0.650	-0.673	-0.485	0.268	-0.671	-0.540	0.401
WA30 (%)	0.324	0.721	0.829	1	0.727	0.505	0.420	0.318	-0.792	-0.715	-0.478	0.561	-0.669	-0.554	0.502
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	0.212
OCT (min)	-0.087	-0.610	-0.650	-0.792	-0.687	-0.555	-0.486	0.093	1	0.735	0.572	-0.381	0.730	0.638	-0.456
Gradient (kg/mm)	-0.072	-0.601	-0.673	-0.715				0.013	0.735	1	0.733	-0.413	0.875	0.794	-0.521
Max force (kg)	0.004	-0.494	-0.485	-0.478				0.118	0.572	0.733	1	-0.247	0.889	0.970	-0.437
Distance at Max force (mm)	0.372	0.154	0.268	0.561				0.187	-0.381	-0.413	-0.247	1	-0.422	-0.390	0.841
Area (kg.mm)	-0.122	-0.643	-0.671	-0.669				-0.032	0.730	0.875	0.889	-0.422	1	0.926	-0.591
Linear Distance (mm)	-0.063	-0.534	-0.540	-0.554				0.113	0.638	0.794	0.970	-0.390	0.926	1	-0.573
End force:Max force (%)	0.305	0.313	0.401	0.502				0.212	-0.456	-0.521	-0.437	0.841	-0.591	-0.573	1

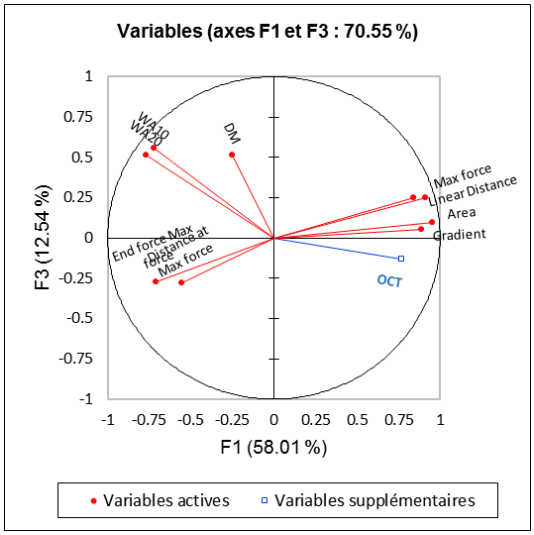
The DM content of fresh material is correlated with WA at OCT (r = 0.44), there is no correlation with OCT (r = -0.09).

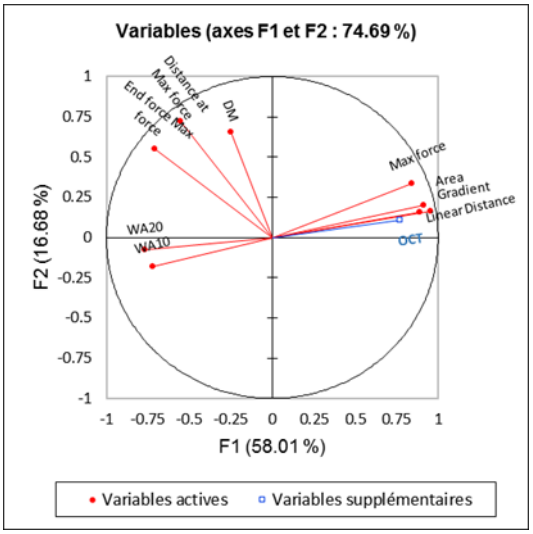
WA 10, 20 and 30 min are correlated with OCT, the highest is at 30 min with r = -0.792. And WA20 and 30 are correlated to texture parameters.

As seen before Gradient is correlated to OCT (r = 0.735) and gradient is correlated to other texture parameters, especially to Area.

2.5 Principal Components Analysis

A PCA done on 125 individuals with full data for: DM, WA10, WA20, Gradient, Max, force, Distance at Max force, Area Linear Distance, End force Max force and OCT as supplementary variable.

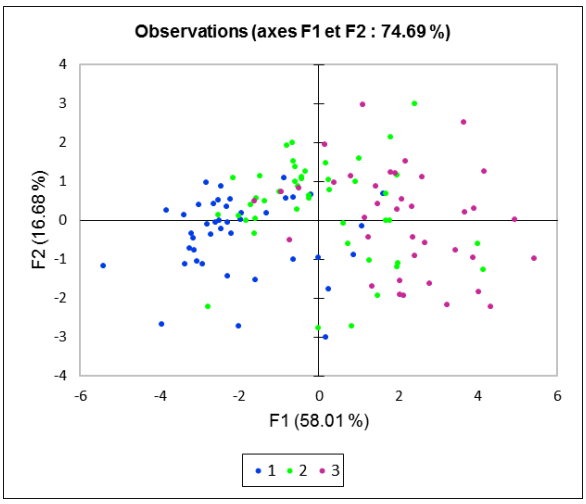


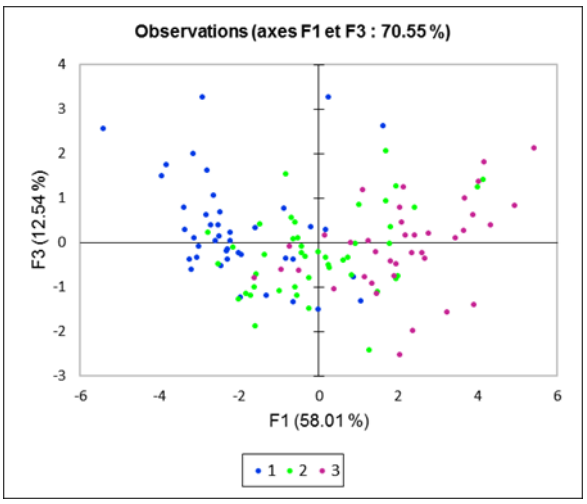


The vectors observations confirm the opposition of physical parameters and WA_20 /WA_10 and the importance of those variables in construction of PC1. The projection of the supplementary variable (OCT) shows that these factorial plans give a good representation of the variable space. Samples with low values for texture parameters (Gradient, area, max force, linear distance) and high values for WA at 10, 20 or 30 mn will have a short cooking time. And samples with intermediate values and high values for DM will present intermediate cooking time.

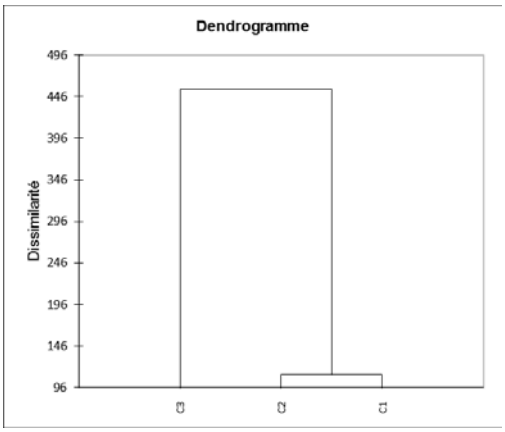
The first and second factorial map of the individuals confirm this observation, the 3 classes for cooking time are almost separated according to PC1, PC2 and PC3.

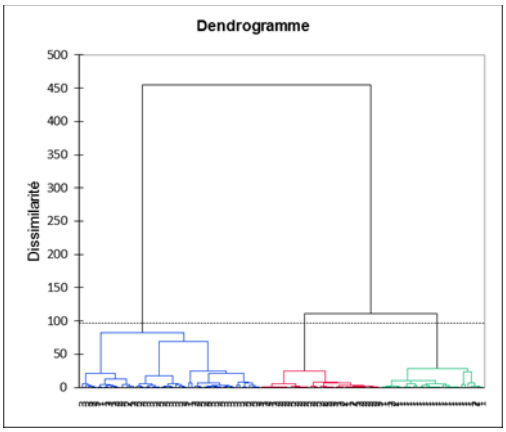
An HAC, done on PCA, even an FDA done on PC's scores confirm this result.





HAC on PC's scores :





AFD: Confusion Matrix for cross validation.

de \ Vers	1	2	3	Total	% correct
1	33	9	0	42	78.57%
2	7	29	9	45	64.44%
3	0	10	28	38	73.68%
Total	40	48	37	125	72.00%

These results confirm that WA at 10, 20 or 30 min, DM fresh material, and texture parameters are relevant for classifying genotypes according to OCT, and to do 3 classes make sense.

2.6 Near Infrared Spectroscopy

2.6.1 Quantitative analysis

The different parameters were calibrated using classical linear regression such as PLS regression, different pre-treatments were tested and best models (higher R^2, lower SEC and SECV, minimum PLS factors) were retained.

2.6.2 Statistics parameters for calibrations:

Constituent	N	Mean	SD	SEC	R^2	SECV	1-VR	#	SEP	math	correction	segment
DM	155	40.4887	2.8956	0.6935	0.9426	0.7714	0.9286	693	0.78	1,4,4	none	1100-2500
WA20	153	6.5601	3.5708	2.7058	0.4258	3.1439	0.2197	1036	4.36	1,4,4	none	400-2500
WA30	121	10.9463	5.7916	3.8522	0.5576	4.9494	0.2636	1036	4.26	1,4,4	none	400-2500
OCT	155	32.2186	13.9171	9.3915	0.5446	10.678	0.4075	1036	10.63	1,4,4	none	400-2500
Gradient	123	1165.5137	442.17	330.8668	0.4401	386.2239	0.2308	1036	346.68	1,4,4	none	400-2500

The statistic parameters show that the only relevant calibration is for DM content with good performances, for others parameters calibrations are very weak, the error (SECV) is similar to SD. The linear approach whatever the pre-treatments will not allow good calibrations for WA and OCT and physical parameters.

There is 125 samples with references values for DM, WA10, WA20, OCT, Gradient and Distance at Max force, the FDA done on these data spitted in two groups: learning (90) and validation (35) leads to:

Confusion matrix for learning set

de \ Vers	1	2	3	Total	% correct
1	25	1	0	26	96.15%
2	0	33	0	33	100.00%
3	0	1	30	31	96.77%
Total	25	35	30	90	97.78%

Confusion matrix for validation set

de \ Vers	1	2	3	Total	% correct
1	16	0	0	16	100.00%
2	0	12	0	12	100.00%
3	0	0	7	7	100.00%
Total	16	12	7	35	100.00%

Based on lab values of these 6 parameters the discrimination in 3 cooking time classes is efficient.

The same approach is realized on predicted values using calibrations developed from spectra. To do this, the calibrations were fitted on learning set (90) for the six parameters and then the whole database (125 samples) was predicted using the “learning calibrations”. Then, the AFD, same as previous is done on predicted values. The results are:

Confusion matrix for learning set

de \ Vers	1	2	3	Total	% correct
1	20	6	0	26	76.92%
2	4	24	5	33	72.73%
3	2	4	25	31	80.65%
Total	26	34	30	90	76.67%

Confusion matrix for validation set

de \ Vers	1	2	3	Total	% correct
1	7	6	3	16	43.75%
2	1	5	6	12	41.67%
3	1	3	3	7	42.86%
Total	9	14	12	35	42.86%

When using NIRS predicted values the discrimination for learning set still presents good results, but the performances for validation are very bad. This means that the calibration for WA, OCT, Gradient and distance are not reliable.

2.7 Classification using spectra

Whatever were the pretreatments 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 3 cooking time classes failed.

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

2.7.1 Classification Regression Trees (CART)

Learning and Validation

From \ To	2	1	3	Total	% correct
2	38	3	0	41	92.7
1	0	33	0	33	100.0
3	0	0	38	38	100.0
Total	38	36	38	112	97.3

From \ To	1	2	3	Total	% correct
1	10	5	8	23	43.5
2	4	6	7	17	35.3
3	2	3	3	8	37.5
Total	16	14	18	48	39.6

K Nearest Neighbors (KNN): validation

	C1	C2	C3	N	% correct
C1	11	9	3	23	48%
C2	8	6	3	17	35%
C3	2	2	4	8	50%

Naive Bayesian Classifier: validation

	C1	C2	C3	N	% correct
C1	12	3	8	23	52%
C2	4	2	11	17	12%
C3	3	2	3	8	38%

2.7.2 Support Vector Machine (SVM)

SVM with linear kernel and one against all strategy

Learning

	1	2	3	Total	% correct
1	16	6	1	23	69.6%
2	6	5	6	17	29.4%
3	1	5	2	8	25.0%
Total	23	16	9	48	47.9%

Validation

	1	2	3	Total	% correct
1	29	3	1	33	87.9%
2	6	31	4	41	75.6%
3	0	0	38	38	100.0%
Total	35	34	43	112	87.5%

SVM with Sigmoid kernel and one against all strategy

Learning

	1	2	3	Total	% correct
1	23	0	0	23	100.0%
2	17	0	0	17	0.0%
3	8	0	0	8	0.0%
Total	48	0	0	48	47.9%

Validation

	1	2	3	Total	% correct
1	33	0	0	33	100.0%
2	41	0	0	41	0.0%
3	38	0	0	38	0.0%
Total	112	0	0	112	29.5%

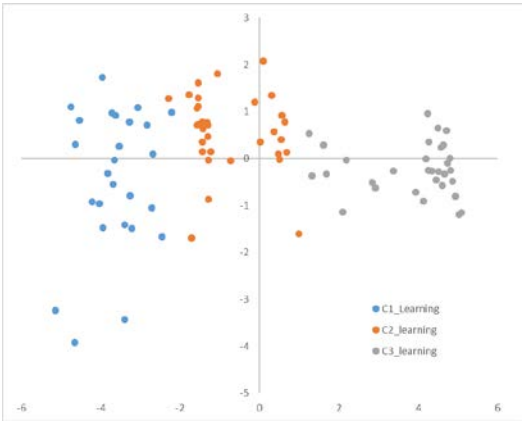
Random Forest: method bagging, random sampling with replacing, 100 trees.

Confusion matrix learning set (n =160)

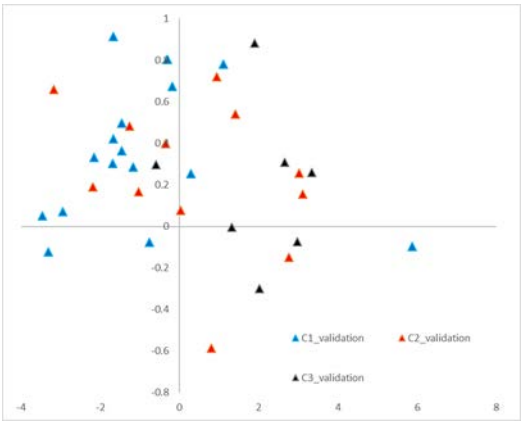
	1	2	3	Total	% correct
1	32	24	11	67	47.8
2	16	21	19	56	37.5
3	8	13	16	37	43.2
Total	56	58	46	160	43.1

2.7.3 Method mix FDA_MPLS_KNN: 3 classes

1. An FDA is done on learning set (n = 90), with 6 laboratory parameters (DM, WA10, WA20, OCT, Gradient and distance at max force) and 3 cooking time classes (<=25 min, 25< <=40 min, >40 min). The scores of the two factors are allocated to corresponding nirs spectra.
2. A PLS regression is calculated for the scores of the 2 factors, based on first derivative spectra corrected snvd.
3. The 2 factors scores for the 35 validation samples are predicted using the PLS model
4. A KNN analyze is done on learning and validation scores.



Factorial map: learning



Factorial map: validation

KNN: Confusion matrix for validation set

	C1	C2	C3	N	% correct
C1	3	12	1	16	19%
C2	1	7	4	12	58%
C3	0	1	6	7	86%
				35	46%

This approach do not led to good classification for all classes, but the classification rate observed for C3 is 86%, is quite promising and allows to imagine a strategy of step by step discrimination “one against all”

2.7.4 Method mix AFD_MPLS_KNN: 2 classes

Same methodology is applied on 2 classes of OCT: ≤ 30 min and > 30 min, using the same learning set (90 samples) and validation set (n = 35).

The FDA (two classes /one factor) led to a classification rate in validation equal to 91.4%, the confusion matrix for validation set is:

From \ to	C1	C2	Total	% correct
C1	22	1	23	95.65%
C2	2	10	12	83.33%
Total	24	11	35	91.43%

The MPLS model parameters for F1 scores are:

Constituent	N	Mean	SD	SEC	R²	SECV	1-VR	Correction	Maths	segment	# variable
F1_C2	90	0	2.03	1.34	0.56	1.77	0.23	SNVD	first derivative	Vis and IR	1036

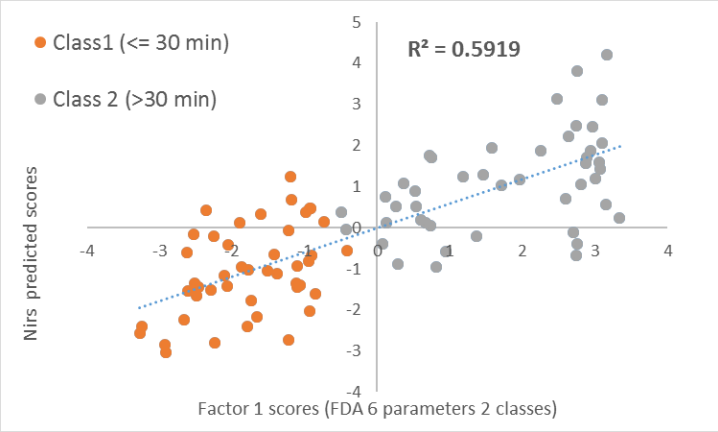


Figure 1 Scatter plot of the F1 scores versus predicted scores for the learning set

The scatter plot of predicted scores from the spectra versus calculated scores through FDA (fig.1) shows that the predictive model caught the trend between class 1 and 2, even if the regression coefficient of determination (R^2) is equal to 0.59.

The scores for F1 are predicted for the 35 validation samples and predicted values are compared to calculated values (fig. 2), same trend as learning set is observed with R^2 equal to 0.25.

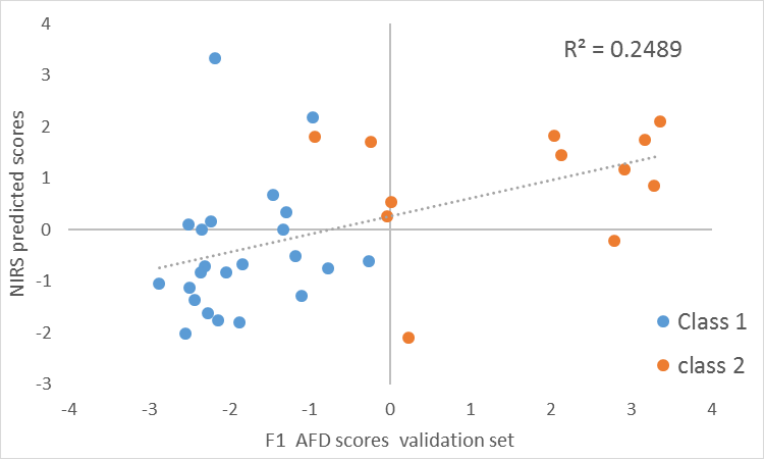


Figure 2 Scatter plot of the F1 scores versus predicted scores for the validation set

A KKN procedure is applied to the two sets, with the following settings:

Number of neighbors : 9
Metric / Distance : Euclidian Distance
Equality management: smallest index
Cross validation / Number of blocs : 2
Size of learning set : 90
Size of prediction set : 35
Seed (random figures) : 65782946

The confusion matrix for validation samples is:

From \ To	C1	C2	Total	% correct
C1	16	7	23	69.57%
C2	2	10	12	83.33%
Total	24	11	35	74.29%

The Kmeans approach was also tested on validation set with the following settings:

Classification criterion: Trace(W)
Iterations stop condition = 500 / Convergence = 0.00001
Number of classes: de :2 à :4
Center : Non
Reduce : Non
Starting partition: Random
Repetitions : 10
Seed (random figures) : 4495227

The confusion matrix is:

From \ To	C1	C2	Total	% correct
C1	19	4	23	82.61%
C2	3	9	12	75.00%
Total	24	11	35	80.00%

3 CONCLUSION

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

Linear approaches were not relevant for calibrations of WA, OCT and physical properties

The high performances for DM calibration reflects the good quality of spectra and laboratory data (no mismatch with references)

Classification using spectral fingerprints (using linear methods or not) did not applied

The solution to tackle this problem could be deep learning approaches combined with hierarchical (or step-by-step) discrimination. The classes (boundaries) can be discussed as well.

The performances of the classification method which mix laboratory values and spectra values indicate that spectra contain relevant information, and confirm that deep learning approaches may help for better and faster classification.

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/