Calibration Report

Dry Matter Quantification in Fresh Grated Cassava Clones using NIRS

High-Throughput Phenotyping Protocols (HTPP), WP3

Kampala, Uganda, 29/11/2022

Ephraim NUWAMANYA, National Crops Resources Research Institute (NaCRRI), Namulonge, Uganda

Babirye Fatumah NAMAKULA, NaCRRI, Namulonge, Uganda

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

Michael KANAABI, NaCRRI, Namulonge, Uganda

Enoch WEMBABAZI, NaCRRI, Namulonge, Uganda

Ivan LYATUMI, NaCRRI, Namulonge, Uganda

Robert KAWUKI, NaCRRI, Namulonge, Uganda

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

To be cited as:

Ephraim NUWAMANYA, Babirye Fatumah NAMAKULA, Fabrice DAVRIEUX, Michael KANAABI, Enoch WEMBABAZI, Ivan LYATUMI, Robert KAWUKI, (2023). *Dry Matter Quantification in Fresh Grated Cassava Clones using NIRS. High-Throughput Phenotyping Protocols (HTPP), WP3.* Kampala, Uganda: RTBfoods Calibration Report, 17 p. https://doi.org/10.18167/agritrop/00731

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)	28/09/2022
Babirye Fatumah NAMAKULA (NaCRRI)	29/11/2022
Fabrice DAVRIEUX (CIRAD)	12/12/2022
Final validation by:	
Fabrice DAVRIEUX (CIRAD)	13/12/2022

CONTENTS

Table of Contents

Table of Figures

List of Tables

ABSTRACT

Context: This scientific report concerns data analysis of two matrices of measured data on fresh grated cassava 1) physico-chemical data and 2) spectral data. The data were collected on fresh cassava in NaCRRI, Uganda.

Place : Uganda, Réunion

Date : 29/11/2022

Authors: Fabrice DAVRIEUX, Fatumah Namakula BABIRYE and Ephraim NUWAMANYA

Content:

The analyses concern 291 cassava genotypes harvested. in 2019 in Namulonge, Uganda. Genotypes came from NaCRRI Cassava breeding program. For this study 291 cassava roots were analysed; 3 uniformly sized non-necrotic roots per clone were sampled. Among the 281 genotypes, 282 were analysed one time and 9 genotypes were analysed 2 times, the total number of spectra is 300.

Dry matter (DM) quantification was achieved in NaCRRI physico-chemical laboratory in Namulonge. Near infrared spectra were scanned in NIR laboratory of NaCRRI in Namulonge. The protocol measurement follows the SOP protocol described for fresh grated cassava: https://doi.org/10.18167/agritrop/00669.

DM content, expressed as % of total weight, vary between 10,74% and 45,00% with an average value of 29,42%. On the basis of the 300 samples 2 data sets were constituted using the Kennard-Stone algorithm: one training set (n = 210) and one test set (n =90).

DM content was calibrated using Modified Partial Least Squares Regression, for the spectral range: 400 nm - 2500nm, that is to say Visible and NIR regions. Calibrations were done on non-pretreated spectra and pretreated spectra using different pre-treatments. The best model was obtained with no pre-treatment applied to the spectra, the R^2 was 0,973 with an SECV equal to 0,894%.

This model was applied to predict samples from the test set, the standard error of prediction (SEP) is equal to SEP = 0,815% and the R^2 for prediction is 0,953. The ratio performance to deviation RPD is equal to 4,63.

The developed NIRs model for quantification of DM content of cassava root presents an accuracy good enough to enable cassava selection based on DM content. The error of the model is SEP = 0,815% which means that a predicted value could be defined with a confidence interval of +/- 1,63 % associated with 95 % of confidence. Furthermore, the database refers to 291 different genotypes representative of the variability of the DM content, thus the model is robust enough for DM quantification in fresh cassava. The procedure can be applied in cassava breeding as routine procedure.

Key Words: Dry Matter, Quantification, Fresh Cassava, NIRS

1 MATERIAL AND METHODS

1.1 Material

Cassava root samples used for analysis were planted and harvested at Namulonge, National Crops Resources Research Institute, 300 samples were analyzed corresponding to 291 genotypes. Three uniformly sized non-necrotic roots per clone were sampled. Dry matter content quantification was achieved in NaCRRI physico-chemical laboratory in Namulonge using the Oven drying method. Near infrared spectra were scanned in Nutrition and analytical laboratory of NaCRRI in Namulonge.

1.2 Dry matter quantification

The DM content of cassava starch was determined by oven drying.

Cassava roots in duplicate were pealed with a knife and whole root grated using a grater. One hundred grams of fresh cassava grated roots was measured using a weighing balance and this was then oven dried at 105°C for 24 hours. After drying, the sample was weighed again and DMC was determined according to the formula below

$$Dry\ matter\ content = \frac{Wt - Wd}{Wt}\ X100\%$$

Where Wt- Total fresh weight, Wd – Dry weight.

1.3 Near Infrared measurements

The NIR spectra acquisition was done on fresh grated cassava roots using a DS2500 (FOSS) spectrometer. Three roots per clone were sampled and scanned, one spectrum per root was performed. The protocol measurement follows the SOP protocol described for fresh grated cassava: https://doi.org/10.18167/agritrop/00669. All treatments and data analysis for exploration and calibration were done using R, (R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/)

And the package rchemo (Lesnoff M. rchemo: Dimension reduction, Regression and Discrimination for chemometrics. - https://github.com/mlesnoff/rchemo: R package version 0.0-17). (R script in annex I)

2 RESULTS

2.1 Dry matter content

Table 1: Descriptive statistics for dry matter values

Statistic	N	Min	Max	Mean	SD
Dry matter	300	10.74	45.00	29.58	5.08

DM content vary between 10,7% and 45% (Table 1) which correspond to a wide range with a relative high dispersion (SD = 5,8 %, fig. 1). Thus, the dataset was diverse enough to be representative of DM content within cassava genotypes in selection. And the range and distribution of the DM values are suitable for modulization using NIR fingerprints.

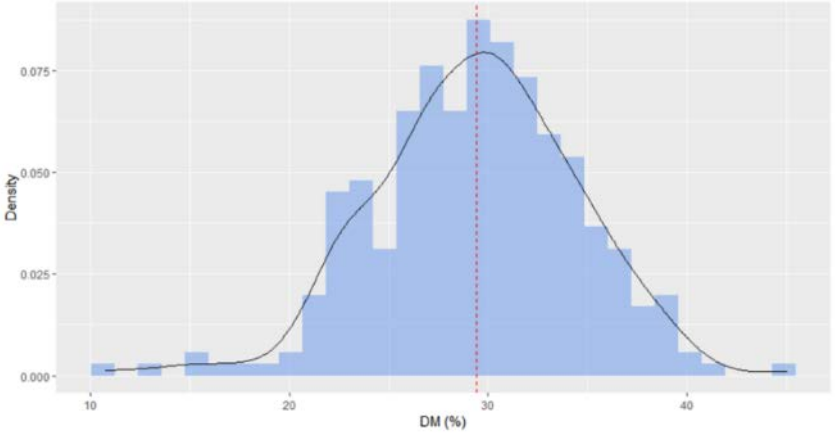

Figure 1: Histogram of dry matter content (%)

2.2 Statistics of Calibration and Validation sets

On the basis of the 300 samples 2 data sets were constituted: one training set (n = 210) and one test set (n = 90) for this 70% of the samples were picked using the Kennard-stone algorithm for the training set while the remaining 30% were kept as test set.

The test and training sets are within the range of variability of the whole data (table 2). The dispersion observed for training set covers the test set variability, which insure a correct evaluation of the model developed on the training set (fig. 2).

Table 2: Descriptive statistics of dry matter content per set of samples

Dry matter	N	Min	Max	Mean	SD
Training Set	210	10,74	45,00	28,93	5,48
Test Set	90	22,56	38,44	30,54	3,80

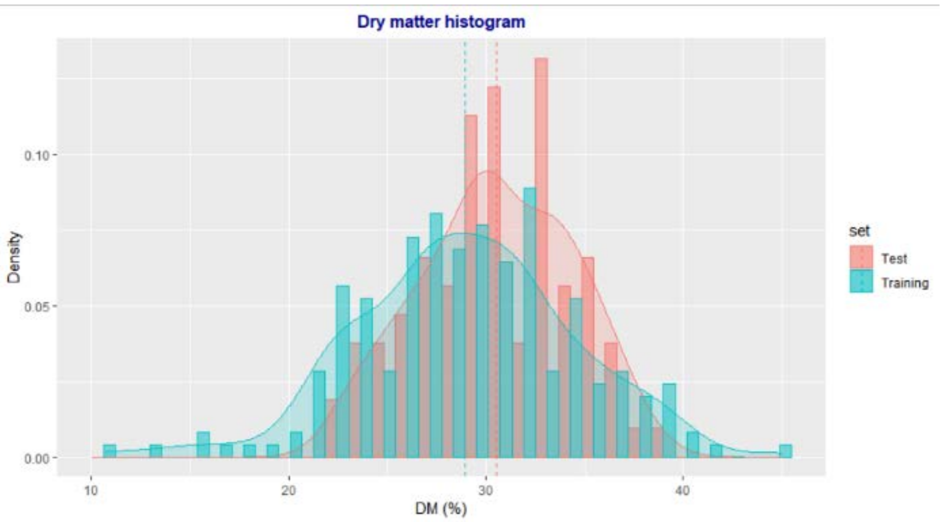

Figure 2: Histogram for Dry matter per set

2.3 Near Infrared Spectroscopy

The representation of the 300 spectra (fig.3) highlights the variability within the database in terms of response (absorbances), especially for water absorption bands (1500 and 1900 nm). All spectra presented similar patterns, no atypical spectra were detected. Genotype V54R3D presented a spectrum with higher absorption especially for water bands (1500 nm and 1900 nm), this genotype corresponds to the minimum observed for DM content (10,74%)

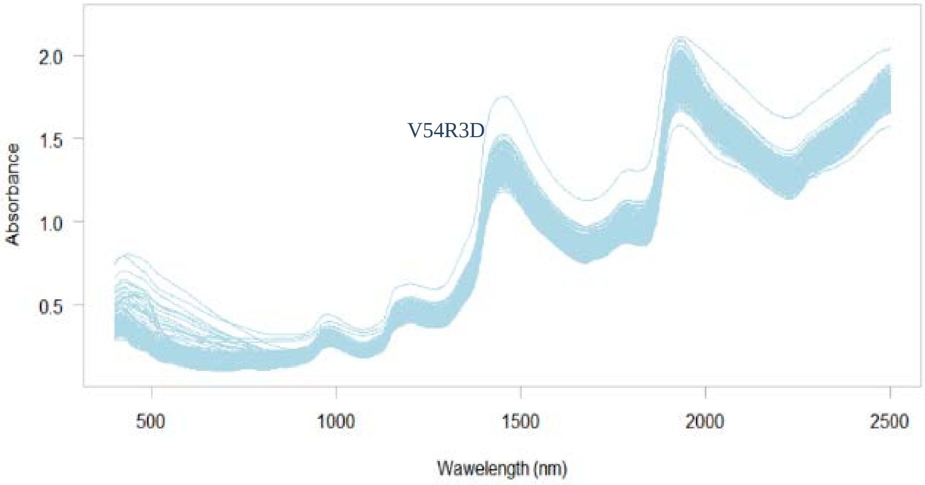

Figure 3: Absorbance spectra for the 300 samples of fresh cassava

2.4 Principal Components Analysis

A PCA calculated on the spectra of the samples (300) led to 89.80% of variance explained by the 2 first PCs, (Figure 4). The Mahalanobis distances (GH) from the average spectrum were calculated on the PCs scores for all the spectra, 6 spectra presented GH values higher than the limit (GH=3). The maximum GH distance is observed for sample VS54R3D (GH = 6.1), which corresponds to the genotype with lower DM content (10,74%). Considering these distances, all the samples are kept in the database.

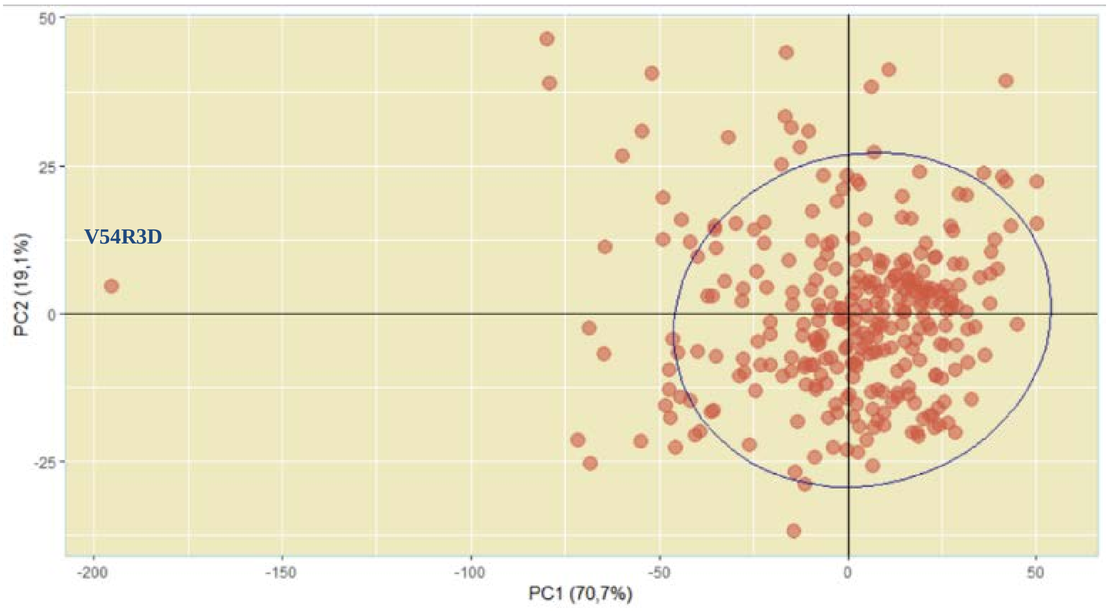

Figure 4: Scatter plot of samples scores for the two first PCs

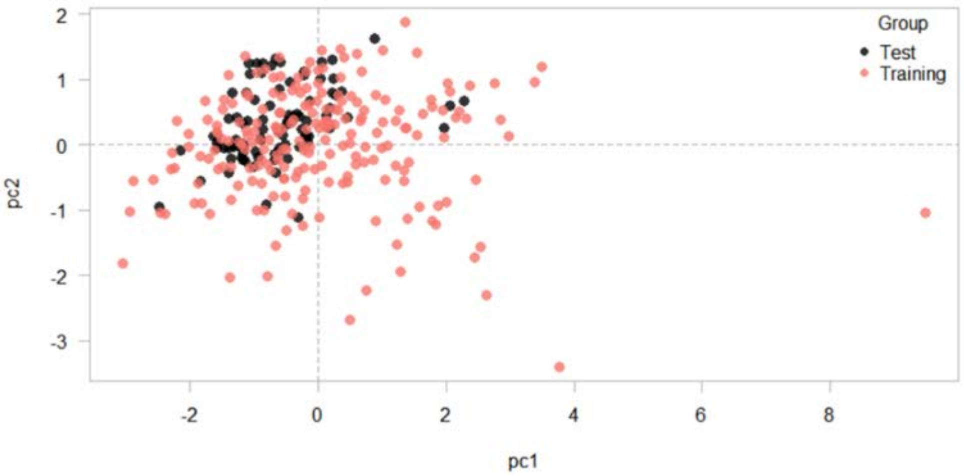

Figure 5: Scatter plot of test samples scores by projection onto PCA space of train samples

The projection of the test samples onto PCs space of the train samples (fig.5) confirms that the training set spectral variability is well representative of the total variability of the population.

2.5 Calibration

DM content was calibrated using Partial Least Squares Regression, for the spectral range: 400 nm - 2500nm, that is to say Visible and NIR regions. Calibrations were done on non-pretreated spectra and pretreated spectra using different pre-treatments. The best model was selected based on highest R²c, R²cv, lowest SEC and SECV, minimum PLS factors and highest ratio SD/SEP (Table 3). The best model was obtained with no pre-treatment applied to the spectra, the R² was 0,937 with an SECV equal to 0,894%

Table 3 : statistic parameters for the DM calibration and validation

Constituent	N	Mean	SD	SEC	SECV	R²cv	Np	SEP	R²p	RPDp
DM	210	28.93	5.48	0.764	0.894	0.973	90	0.815	0.953	4.63

This model was applied to predict samples from the test set, the standard error of prediction (SEP) is equal to SEP = 0,815% and the R² for prediction is 0,953. The ratio performance to deviation RPD is equal to 4.63. These performances indicated that this model is efficient to predict DM content in fresh cassava roots. The scatter plot of predicted values versus laboratory values illustrates the quality of the fitting (figure.6)

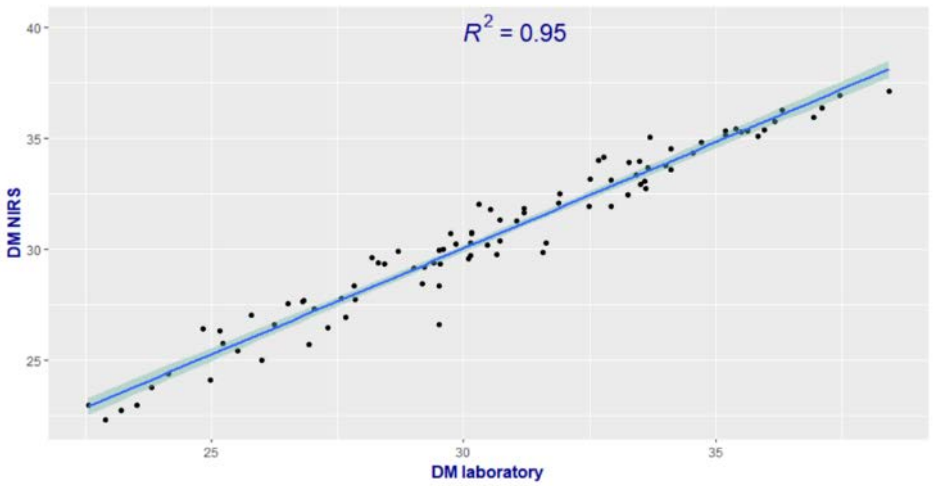

Figure 6: Scatter plot of Predicted values of DM vs laboratory values for the test set

3 CONCLUSION

The developed NIRs model for quantification of DM content of cassava root presents an accuracy sufficient to enable cassava selection based on Dry matter content expressed as % of total weight. The error of the model is SEP = 0.815 % which means that a predicted value could be defined with a confidence interval of +/- 1.63 % associated with 95 % of confidence. Furthermore, the database refers to 291 different genotypes representative of the variability of the dry matter content, thus the model is robust enough for Dry matter content quantification in fresh cassava. The procedure can be applied, as routine analysis, in cassava breeding programs as well scaled to industry.

4 APPENDIX: R SCRIPT

```
# Load libraries_____________________________________________________
library(tidyverse)
library(rchemo)
library(psych)
library(ggpubr)
library(caret)
#_______________________________________________________________________________
# Load data files and create working files_________________________________
nacrri <- read.csv2 ("data/nacrridm.csv", header=T, dec=".",
stringsAsFactors = T, na.strings = c ("NA",""))
DM <- as.data.frame(nacrri$DM) # dry matter (Y)
colnames(DM) <- "DM"
XDM <- select(nacrri, starts_with("X")) # Spectra (X)
colnames (XDM) <- seq(400,2498,2) #

## descriptive statistics for DM_____________________________________________
summary(DM)
dmstat <- describe(DM$DM)
rownames(dmstat) <- "DM"
dmstat <- dmstat[,-1]
dmstat <- round(dmstat,3)
##

## Box plot for DM________________________________________________________________
group <- data.frame(G = rep("DM",300), DM=DM$DM)
ggplot(group, aes( y = DM, x = G ))+
geom_boxplot(alpha = 0.5, outlier.alpha = 0, width = 0.2, color = "dark blue",
fill = "lightgoldenrod")+
geom_jitter( width = 0.1, color = "cornflowerblue")+ stat_summary(fun = mean,
geom = "point",
shape = 18, size = 3.5, color = "red")+ ggtitle("BoxPlot DM")+
xlab("") +
theme (
legend.position="none",plot.title = element_text (size=12, color = "dark blue",
face = "bold", hjust = 0.5 )
)
```

```
## histogram for DM with density curve________________________________________
ggplot ( DM, aes(x = DM)) +
geom_histogram (aes (y=..density..), position="identity", alpha=0.5,
fill = " cornflowerblue" )+
geom_density (alpha =0.3) +
labs ( title = "Dry matter", x= "DM (%)", y = "Density") +
theme (
plot.title = element_text (size=12, color = "dark blue", face = "bold", hjust =
0.5)
)
```

```
## Spectra plot_______________________________________________________________
plotsp(XDM, col = "lightblue", xlim = c(400, 2500),xlab = "Wawelength (nm)", ylab
= "Absorbance")
title("NaCRRI fresh ground Cassava spectra")
```

```
## Parameters for cross validation_________________________________________
ntot <- nrow(XDM)
segm <- segmkf(n = ntot, K = 4, nrep = 50) # 4 blocks of 75 lines (300/4 =
75) repeated 50 times
nlv <- 1:50 # latent variables (LV) from 1 to 50
res <- gridcvlv(XDM, DM, segm,
score = msep, fun = plskern,
nlv = nlv, verb = TRUE)
#____________________________________________________________________________
```

```
## Error evolution as a function of #LV_______________________________________
evol <- res$val
ggplot(evol, aes(x = nlv, y = y1)) +
geom_point(size=3, color="darkblue")+
geom_line(color="red",linetype="dashed")+
labs (title = "RMSEP as a function of number of LV", x= "Latent
variables", y="RMSEP")+
theme(
plot.title = element_text(color="red", size=14, face="bold.italic", hjust = 0.5),
axis.title.x = element_text(color="blue", size=14, face="bold"), axis.title.y =
element_text(color="#993333", size=14, face="bold")
)
```

```
u <- evol[evol$y1 == min(evol$y1), ][1, , drop = FALSE] ## extraction of the #LV
## corresponding to minimum error
```

## Developing model PLSR and performances

```
fm <- plskern(XDM, DM, nlv = u$nlv)
pred <- predict(fm, XDM)$pred
performances <- data.frame (MSEP = msep(pred,DM), RMSEP= rmsep(pred,DM), SEP=
sep(pred, DM), Bias = bias(pred, DM), R2 = r2(pred, DM), RPD = rpd(pred, DM),
RPDR = rpdr(pred, DM))
rownames(performances) <- "Dry Matter"
colnames(performances)[5] <- "R²"
performances <- round(performances,3)
residuals <- residreg (pred, DM)
cor2(pred, DM)
DM_DMpred <- data.frame(DM = DM$DM, DMpred = pred)
dim(DM_DMpred)
colnames (DM_DMpred)[2] <- "DMpred"
```

## Scatter plot DM vs predicted DM

```
ggplot(DM_DMpred, aes( x = DM, y= DMpred)) +
geom_point()+
geom_smooth(method = lm, fill="#69b3a2", se=TRUE)+
stat_regline_equation(label.y = 35, aes(label = ..eq.label..)) + ## Add
regression equation to the plot
# add R² to the plot with the position Y = (40%) and x = (30%) of DM
stat_regline_equation(label.y = 40, label.x = 30, aes(label =
..rr.label..), size=6, color="darkblue")+
labs(title = "DRy matter", x = "DM laboratory", y = "DM NIRS")+
theme (plot.title = element_text(face = "bold", colour = "darkblue" , size =14,
hjust=0.5))+
theme(axis.title.x = element_text(face = "bold", colour = "darkblue", size =
12))+
theme(axis.title.y = element_text(face = "bold", colour = "darkblue", size =
12))
```

## Matrix with DM, DMpred and residuals

```
DM_DMpred <- data.frame(DM = DM_DMpred$DM, DMpred = DM_DMpred$DMpred,
residuals = (DM_DMpred$DM-DM_DMpred$DMpred ) )
summary(DM_DMpred)
```

```
# Histogram of residuals
ggplot ( DM_DMpred, aes(x = residuals)) +
geom_histogram (aes (y=..density..), position="identity", alpha=0.5, fill
= " cornflowerblue" )+
geom_density (alpha =0.3, fill = "#FF6666") +
geom_vline (aes(xintercept=mean(residuals)), color="blue",
linetype="dashed", size=0.5)+
labs ( title = "Residuals histogram", x= "Residuals", y = "Density") +
theme (
plot.title = element_text (size=12, color = "dark blue", face = "bold",
hjust = 0.5)
)
```

```
##------------## kennard-stone sampling ##__________________________________________
## Kennard-Stone sampling using euclidean distance
k <- (300*0.70)
sets <- sampks(XDM, k = k, diss = "eucl")
trainIndex <- sets$train
XDMtrain <- XDM[trainIndex,]
DMtrain <- DM [trainIndex,]
XDMtest <- XDM[-trainIndex,]
DMtest <- DM [-trainIndex,]
```

```
## develop model PLSR on train set (n= 210) tuning using repeated k folds CV
ntot <- nrow(XDMtrain)
segm <- segmkf(n = ntot, K = 4, nrep = 50) # 4 blocks of 53 lines repeated
50 times)
nlv <- 1:50
res <- gridcvlv(XDMtrain, DMtrain, segm,
score = msep, fun = plskern, nlv = nlv, verb = TRUE)
u <- evol[evol$y1 == min(evol$y1), ][1, , drop = FALSE]
```

```
## predict test set______________________________________________________________
fm <- plskern(XDMTRAIN, DMTRAIN, nlv = u$nlv)
pred <- predict(fm, XDMtest)$pred

performances <- data.frame (MSEP = msep(pred,DMtest), RMSEP=
rmsep(pred,DMtest), SEP= sep(pred, DMtest), Bias = bias(pred, DMtest), R2 =
r2(pred, DMtest), RPD = rpd(pred, DMtest), RPDR = rpdr(pred, DMtest))

rownames(performances) <- "Dry Matter"
colnames(performances)[5] <- "R²"
performances <- round(performances,3)

residuals <- as.data.frame( residreg (pred, DMtest) )
colnames(residuals)[1] <- "residuals"
DM_DMpred <- data.frame(DM = DMtest, DMpred = pred) dim(DM_DMpred)
colnames (DM_DMpred)[2] <- "DMpred"

## Scatter plot DM vs DMpred
ggplot(DM_DMpred, aes( x = DM, y= DMpred)) +
geom_point()+
geom_smooth(method = lm, fill="#69b3a2", se=TRUE)+
stat_regline_equation(label.y = 40, label.x = 30, aes(label =..rr.label..),
size=6, color="darkblue")+
labs(title = "DRy matter", x = "DM laboratory", y = "DM NIRS")+
theme (plot.title = element_text(face = "bold", colour = "darkblue" , size =14,
hjust=0.5))+
theme(axis.title.x = element_text(face = "bold", colour = "darkblue", size =
12))+
theme(axis.title.y = element_text(face = "bold", colour = "darkblue", size = 12))

## Histogram of residuals
ggplot ( residuals, aes(x = residuals)) +
geom_histogram (aes (y=..density..), position="identity", alpha=0.5, fill = "
cornflowerblue" )+
geom_density (alpha =0.3, fill = "#FF6666") +
geom_vline (aes(xintercept=mean(residuals)), color="blue",
linetype="dashed", size=0.5)+
labs ( title = "Residuals histogram", x= "Residuals", y = "Density") + theme (
plot.title = element_text (size=12, color = "dark blue", face = "bold", hjust =
0.5)
```

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/