Random Forests for the prediction of water content by Near-Infrared Hyperspectral Imaging in biscuits.

Philippe Courcoux [1], Eloïse Lancelot [2],
Sylvie Chevallier [2], Alain Le Bail [2], Benoît Jaillais [1]

1. *StatSC / Statistics, Sensometrics & Chemometrics*
2. *GEPEA UMR 6144 CNRS*

ONIRIS Nantes (France)

Context and objectives

Breakage, cracking and friability of biscuits are common defaults during industrial process.
There are known to be caused by heterogeneity of water content in biscuits.

Cracks

Context and objectives

The main objectives of this study are

- the evaluation of the potential of Near-Infrared Hyperspectral Imaging Spectroscopy (NIR-HSI) for the local quantification of water content in biscuits,

- the comparison of regression techniques for the prediction of moisture in cereal products (PLS and random forests).

Commercial brand of biscuits produced in the western France.

Material and methods

Ten biscuits were conditioned in 10 desiccators containing different saturated solutions of salt.
In addition, one biscuit was kept in plastic bag in the lab (hygrometry between 0,42-0,5)

N°	1	2	3	4	5	6	7	8	9	10
Salt	LiCl	CH3COOK	MgCl2	K2CO3	Mg(NO3)2	NaBr	CuCl2	NaCl	KCl	BaCl2
Aw (20°C)	0.114	0.226	0.313	0.44	0.545	0.587	0.684	0.754	0.851	0.907

All biscuits were weighted before and after conditioning.
Dry matter was evaluated after a conditionning time of one week.

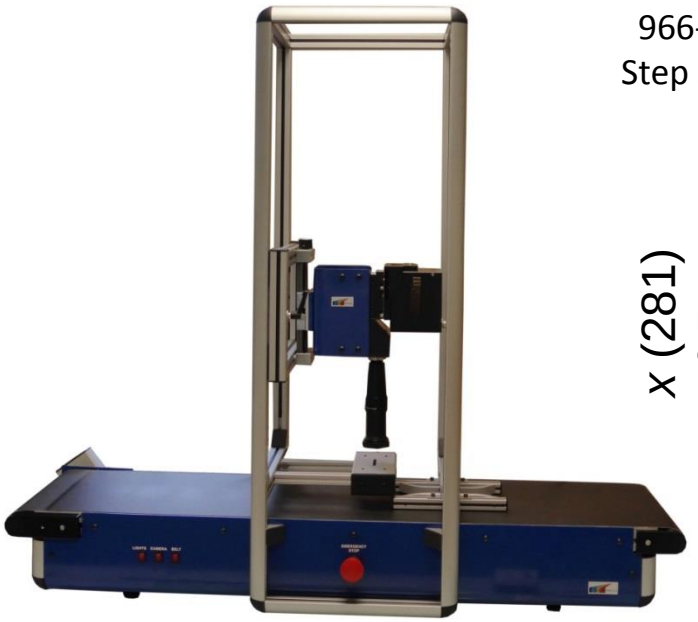
N°	1	2	3	4	5	6	7	8	9	10
Moist.	2.54	3.87	4.84	6.88	7.73	8.42	10.55	12.88	16.56	20.10

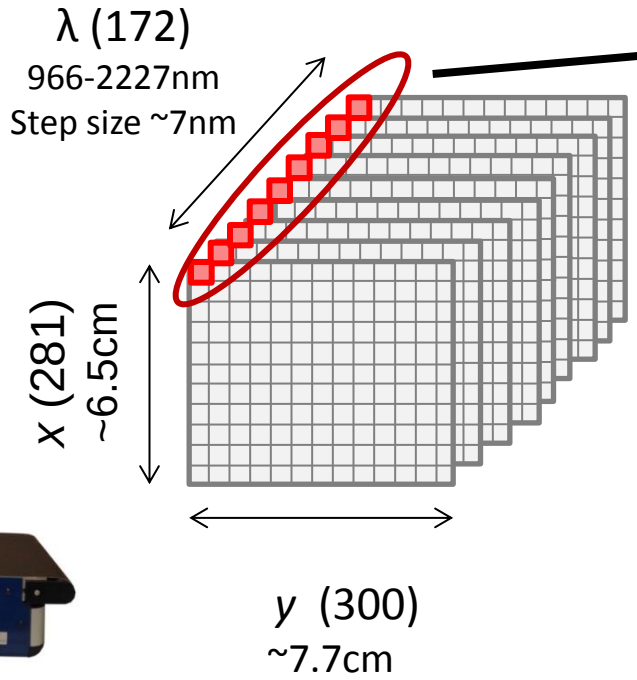
Then, all biscuits were imaged using the NIR-HSI system.

Material and methods

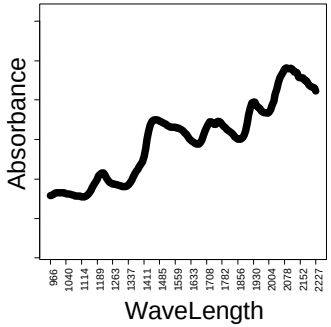
NIR hyperspectral imaging system:

HyperPro (BurgerMetrics)





Spectrum of a given point xy



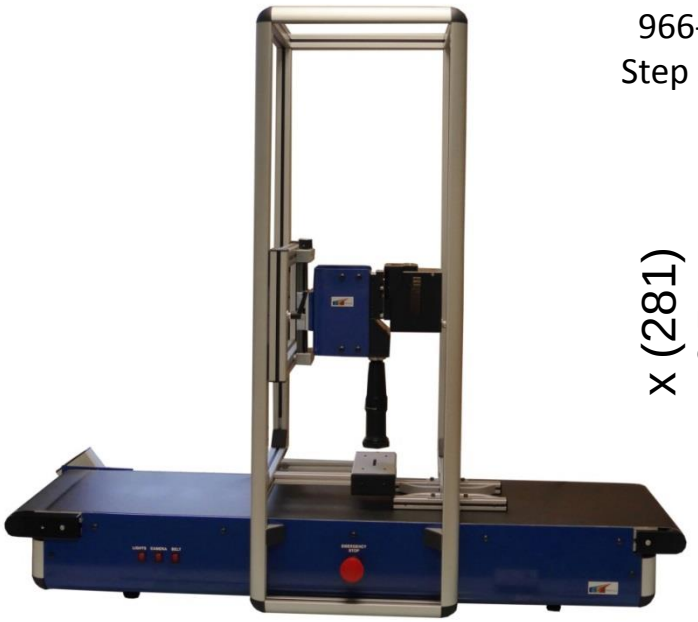
“ Pushbroom” hyperspectral imaging system :

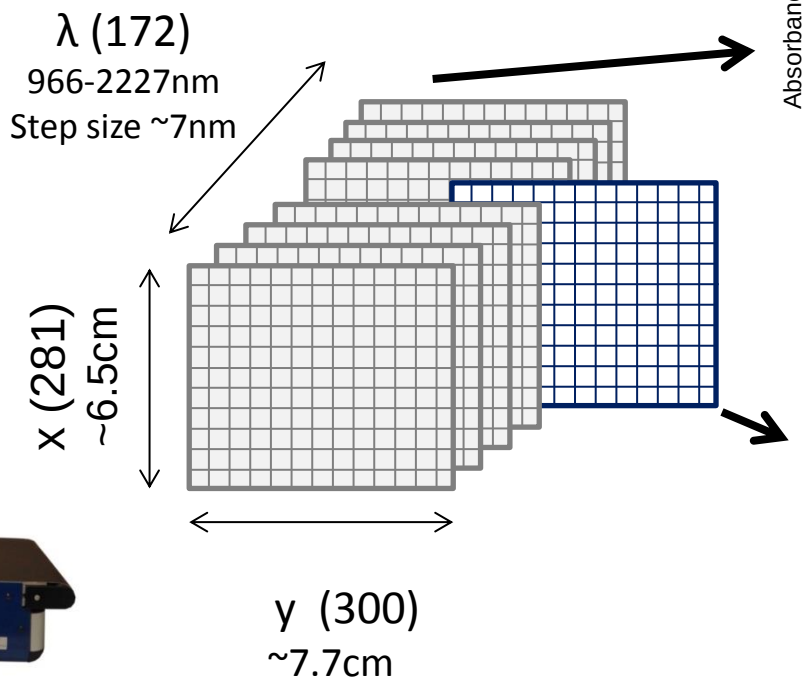
Each pixel of this image cube is a spectrum. The cube may be considered as an image of spectra.

Material and methods

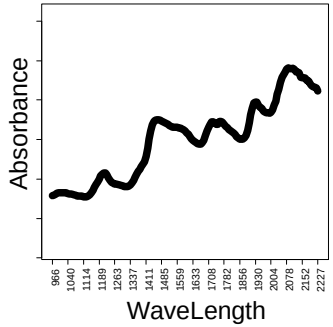
NIR hyperspectral imaging system:

HyperPro (BurgerMetrics)





Spectrum of a given point xy



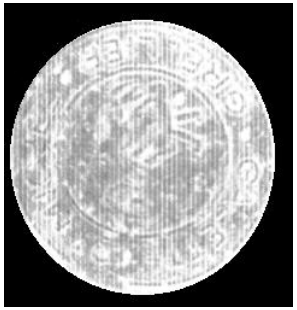


Image obtained at the selected wavelength

1min/image

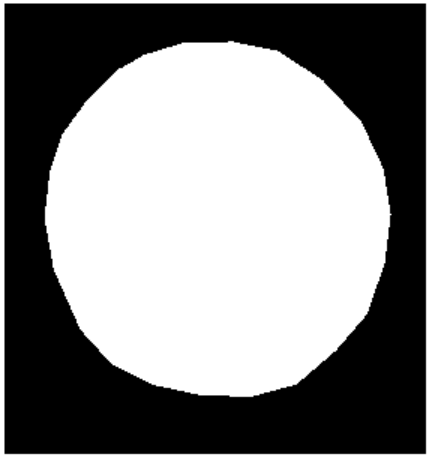
" Pushbroom" hyperspectral imaging system :

Each slice of the image cube (at a given wavelength) is a grey-level image. So, this cube may also be considered as a spectrum of images.

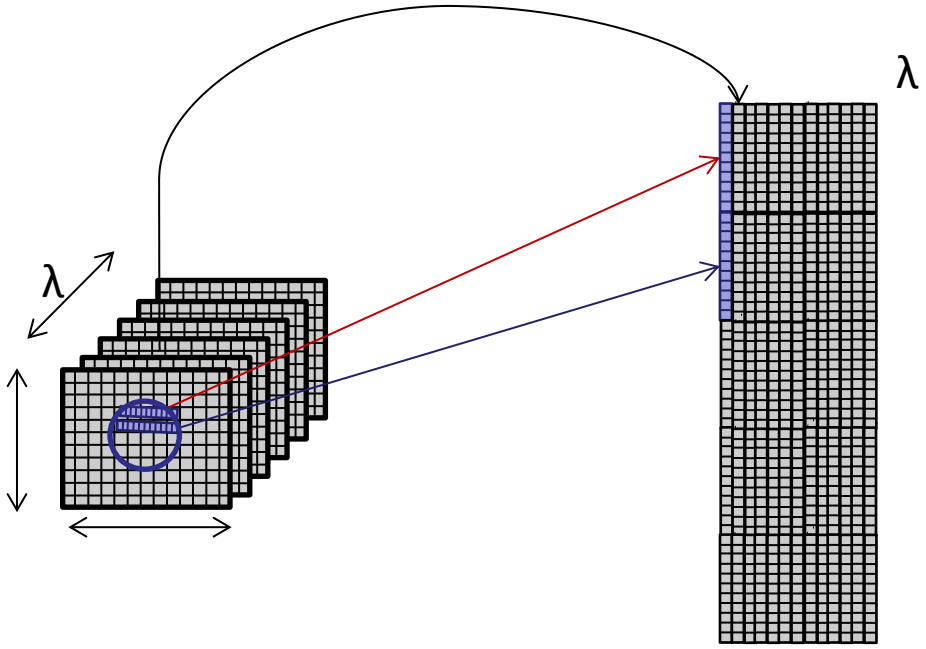
Data pre-treatment

1. ROI selection

An automatic segmentation is performed on the averaged image for each data cube, resulting on a binary image (pixel values equal to 1 for the ROI and to 0 for the background).



2. Cube unfolding



3. Smoothing+SNV

Data processing

Random sampling:
1000 pixels in each image

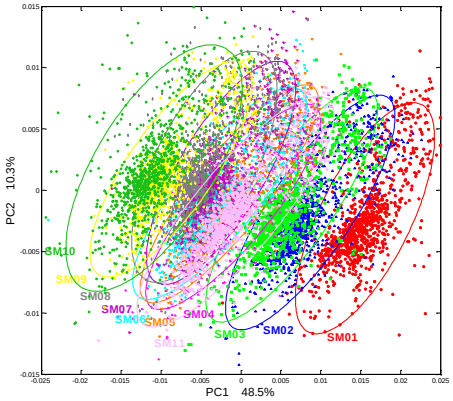
K unfolded cubes

172 × 1000

Concatenation

*K*000 × 172

PCA analysis

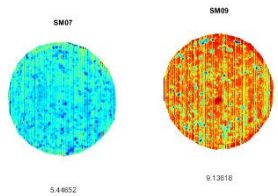


X: *K*000 × 172

y: Measured % water, *K*000

Regression

Construction of predicted images



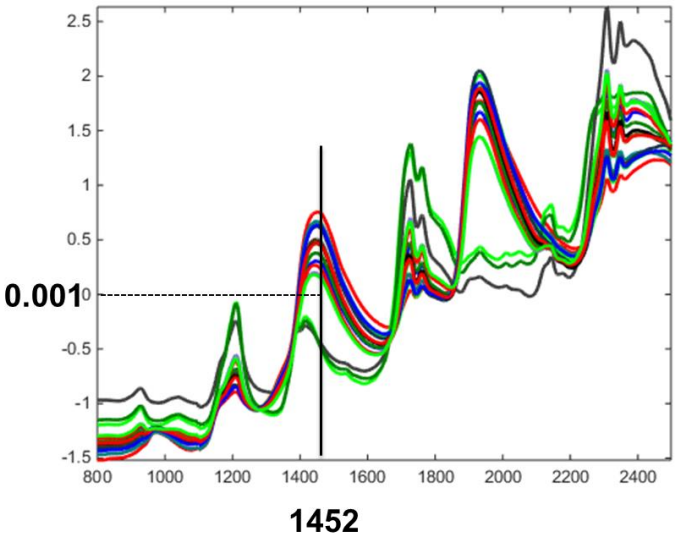
Regression tree

Classification and Regression Tree
CART

Recursive partitioning technique

⟹ Binary tree :

- the root contains all the samples,
- each node is determined by a variable and a cut-off value,
- the leaves form a partition of the samples.



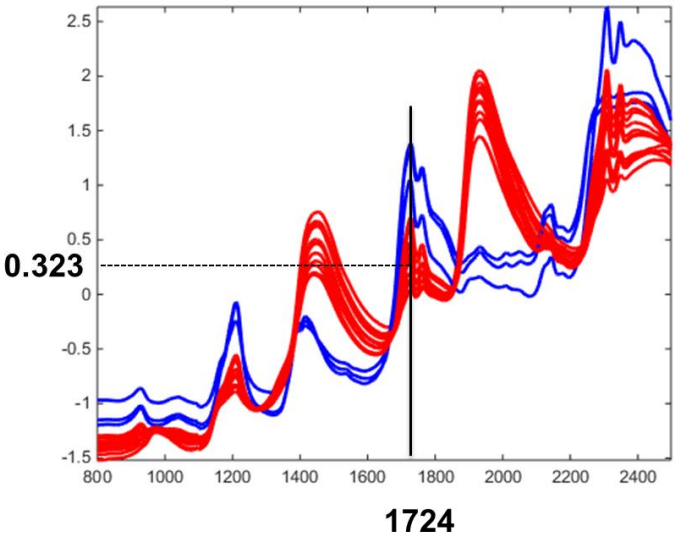
Regression tree

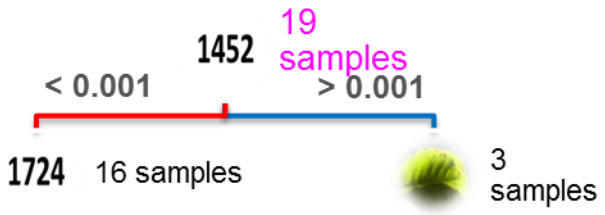
Classification and Regression Tree
CART

Recursive partitioning technique

⟹ Binary tree :

- the root contains all the samples,
- each node is determined by a variable and a cut-off value,
- the leaves form a partition of the samples.





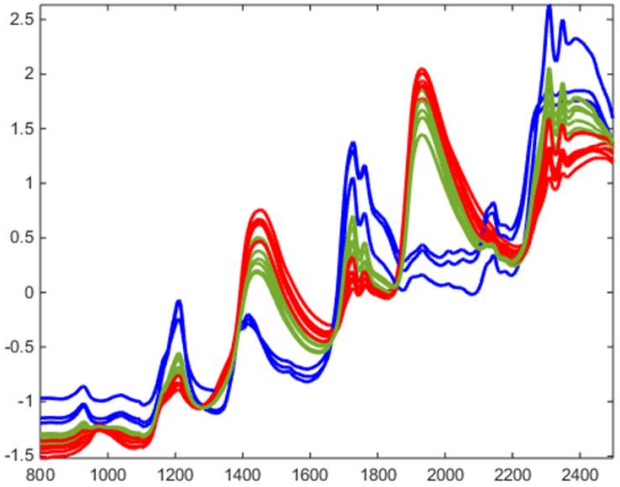
Regression tree

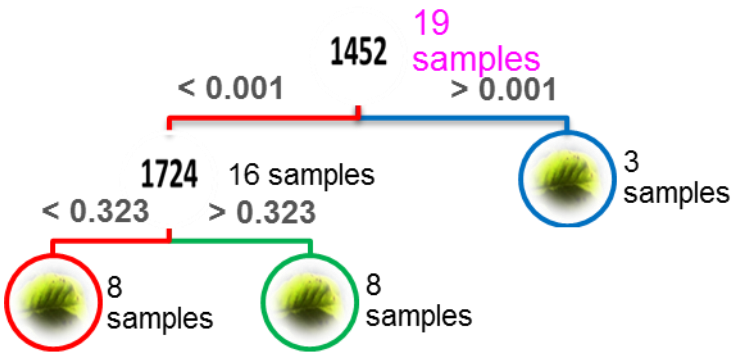
Classification and Regression Tree
CART

Recursive partitioning technique

⟹ Binary tree :

- the root contains all the samples,
- each node is determined by a variable and a cut-off value,
- the leaves form a partition of the samples.





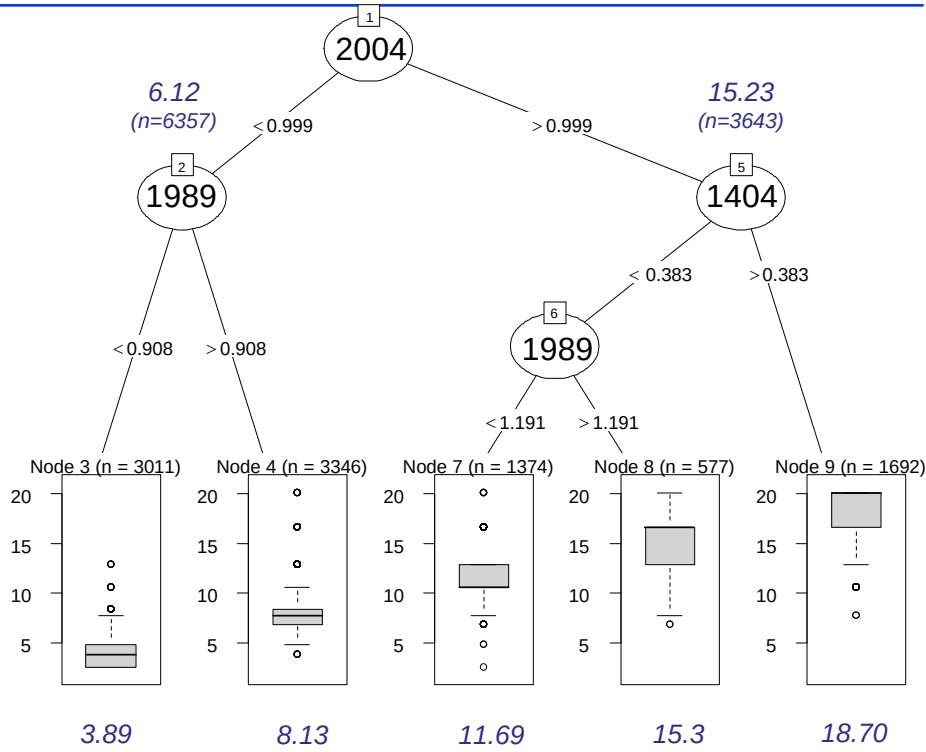
Regression tree

Classification and Regression Tree
CART

Recursive partitioning technique

⟹ Binary tree :

- the root contains all the samples,
- each node is determined by a variable and a cut-off value,
- the leaves form a partition of the samples.



At each node, a partition of samples is realized by choosing a variable X_j and a cut-off value s.

$$\boldsymbol{R_1(j,s) = \{X / X_j \le s\}} \textbf{ and } \boldsymbol{R_2(j,s) = \{X / X_j > s\}}$$

The couple (variable, cut-off) is chosen to minimize the within sum of squares (measure of the impurity of the two regions)

$$SSW = \sum_{i \in R_1} (y_i - \bar{y}_1)^2 + \sum_{i \in R_2} (y_i - \bar{y}_2)^2$$

Breiman L., Friedman J., Olshen R., Stone C. (1984). Classification And Regression Trees. Chapman & Hall

Regression trees

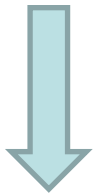
Advantages

- Provide decision rules,
- Handle non linear links,
- No need of distributional assumptions
- …

However…

In case of highly correlated variables :

- caution must be taken for interpreting the tree
- lack of robustness of the obtained tree



Random Forests

Random Forests

Building of high number of regression trees with a double randomisation process:

- Resampling of observations (bootstrap / bagging)
- Random selection of variables at each node of each tree

⟹ **Prediction of moisture for unknown samples**

Prediction by a random forest
=
mean of the predictions
given by all the trees of this forest

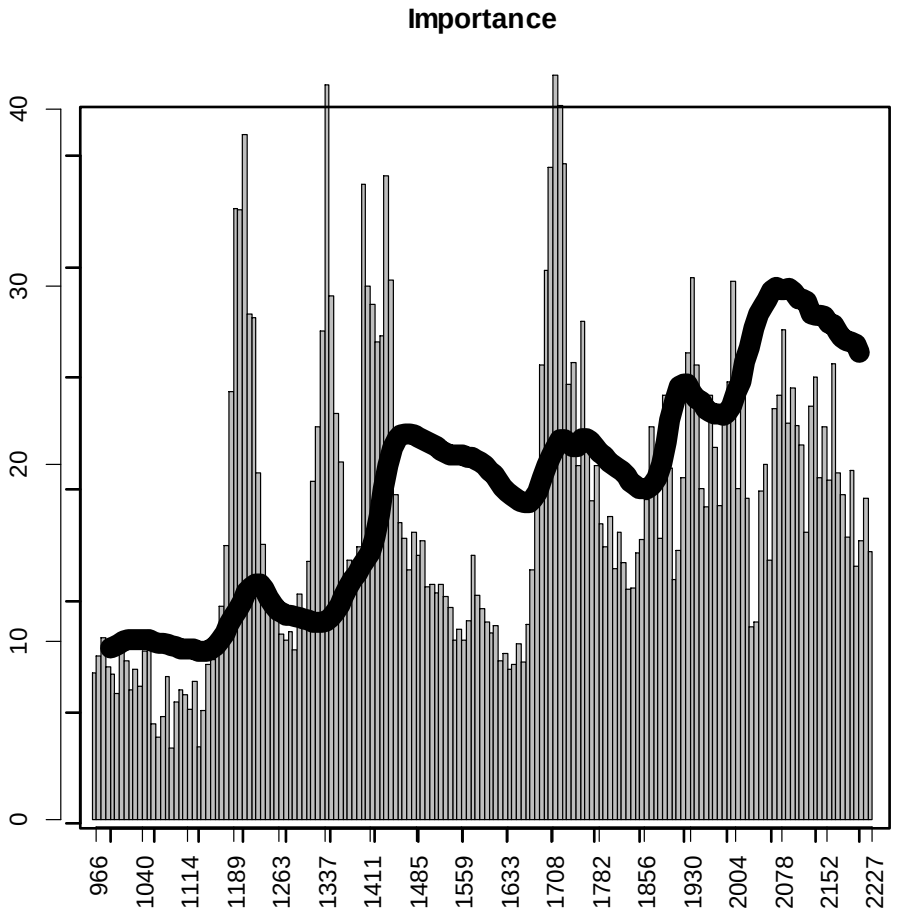
⟹ **Computation of the importance of each predictor**
VI : Variable Importance

Based on the mean increase of the error of prediction for Out-of-Bag observations (OOB) after random permutation of data.

Breiman, L. (2001). Random Forests. Machine Learning, 45, 5–32

Random Forests

Importance of each wavelength for the prediction of moisture in biscuits



Selection of variables

Using the importance provided by random forests for selecting the variables:

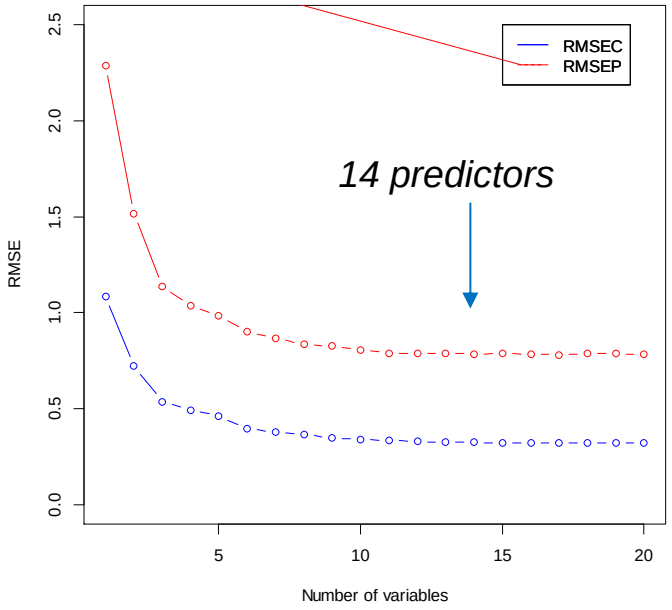
1. Order the variables and select a given number of predictors

2. Introduce each variable in the model and choose the one minimizing the RMSE in prediction

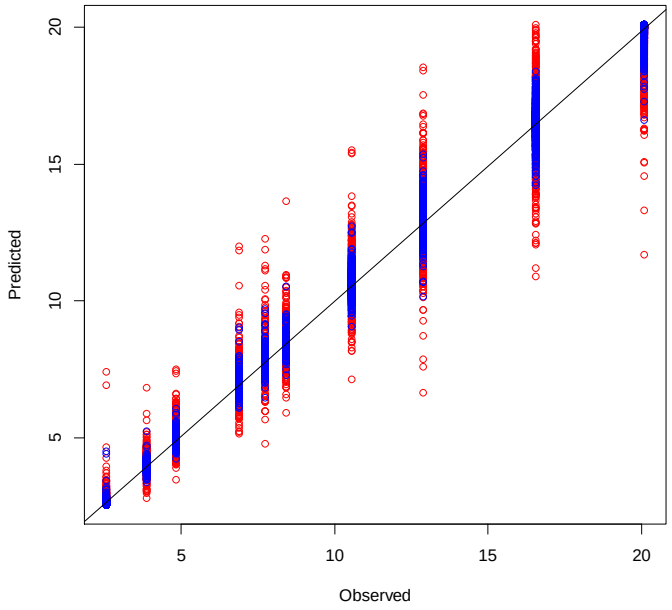
3. Repeat the step 2 until no improvement in the accuracy of the model is observed

Genuer R. et al. (2010) Variable selection using random forests. Pattern Regognition Letters, 31(14) 2225-2236.

Results



RMSE as a function of the number of variables introduded in the model.

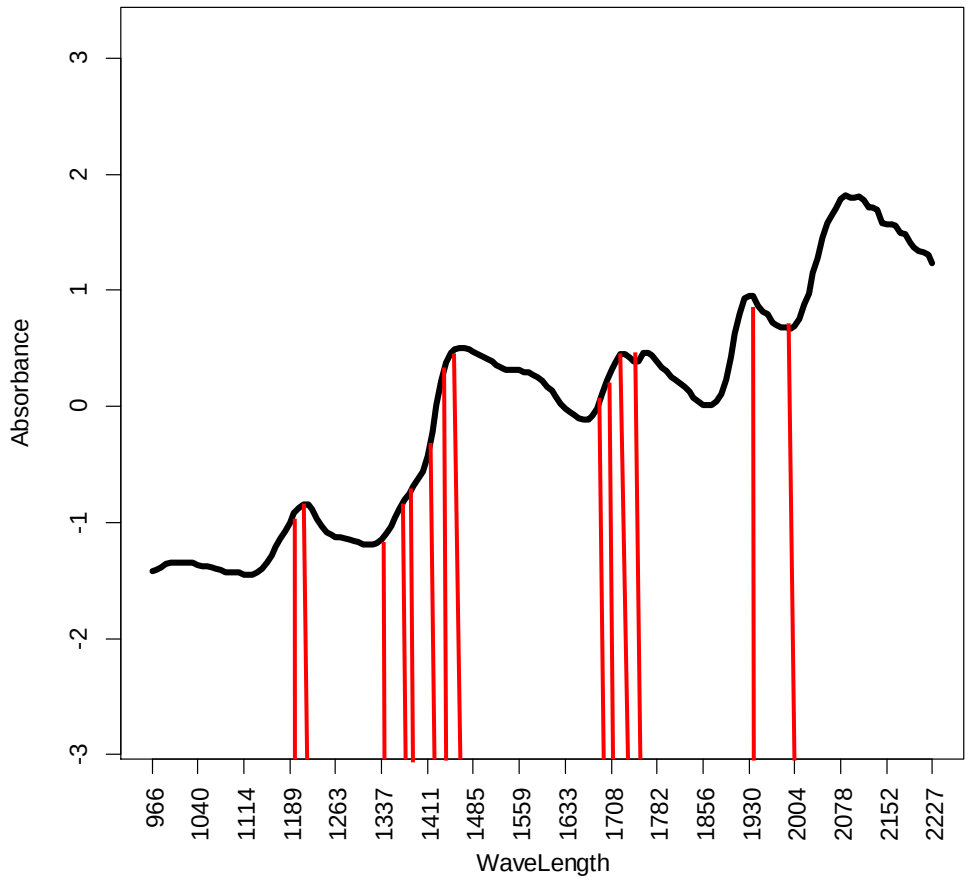


Predicted *vs* observed values of moisture.

	RMSEC	RMSEP	R^2 C	R^2 P
Random Forest	0.3256	0.7853	0.9963	0.97870

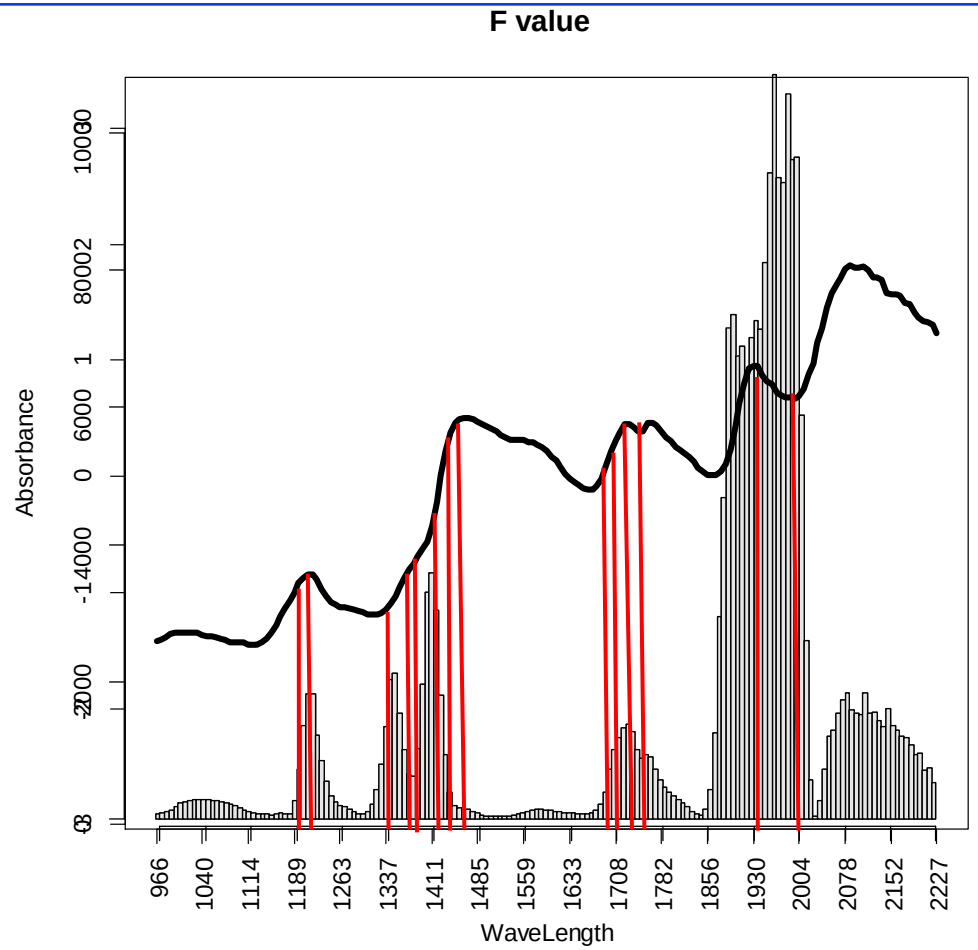
Very good quality in prediction for the selected model.

Variables selected



Position of the 14 variables selected in the final model.

Variables selected

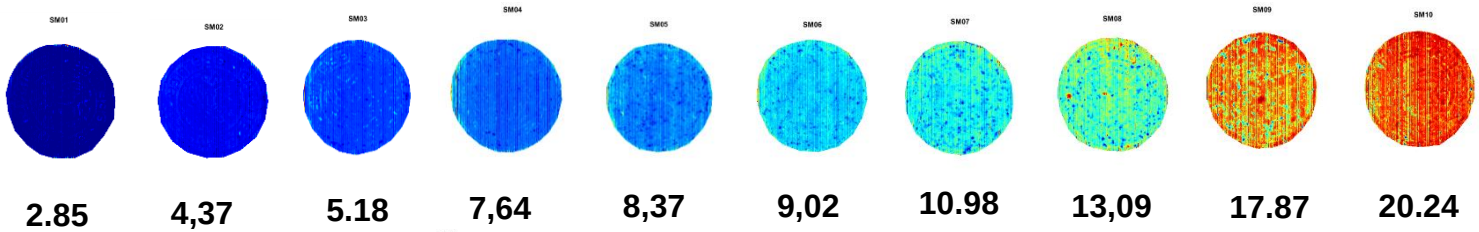


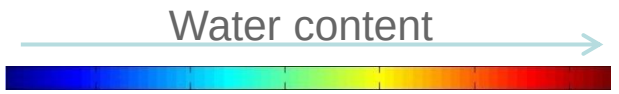
F-value of the variables for the discrimination of water content values

Position of the 14 variables selected in the final model.

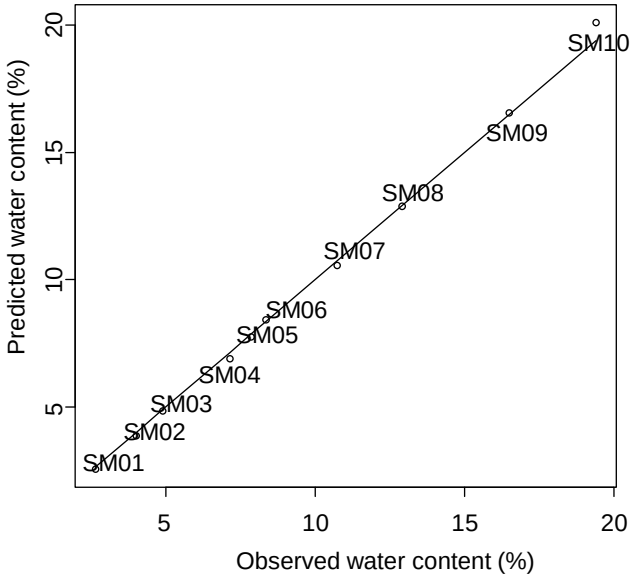
Predicted images

False color images
obtained after applying the prediction model to the biscuits
(and mean predicted value of moisture)

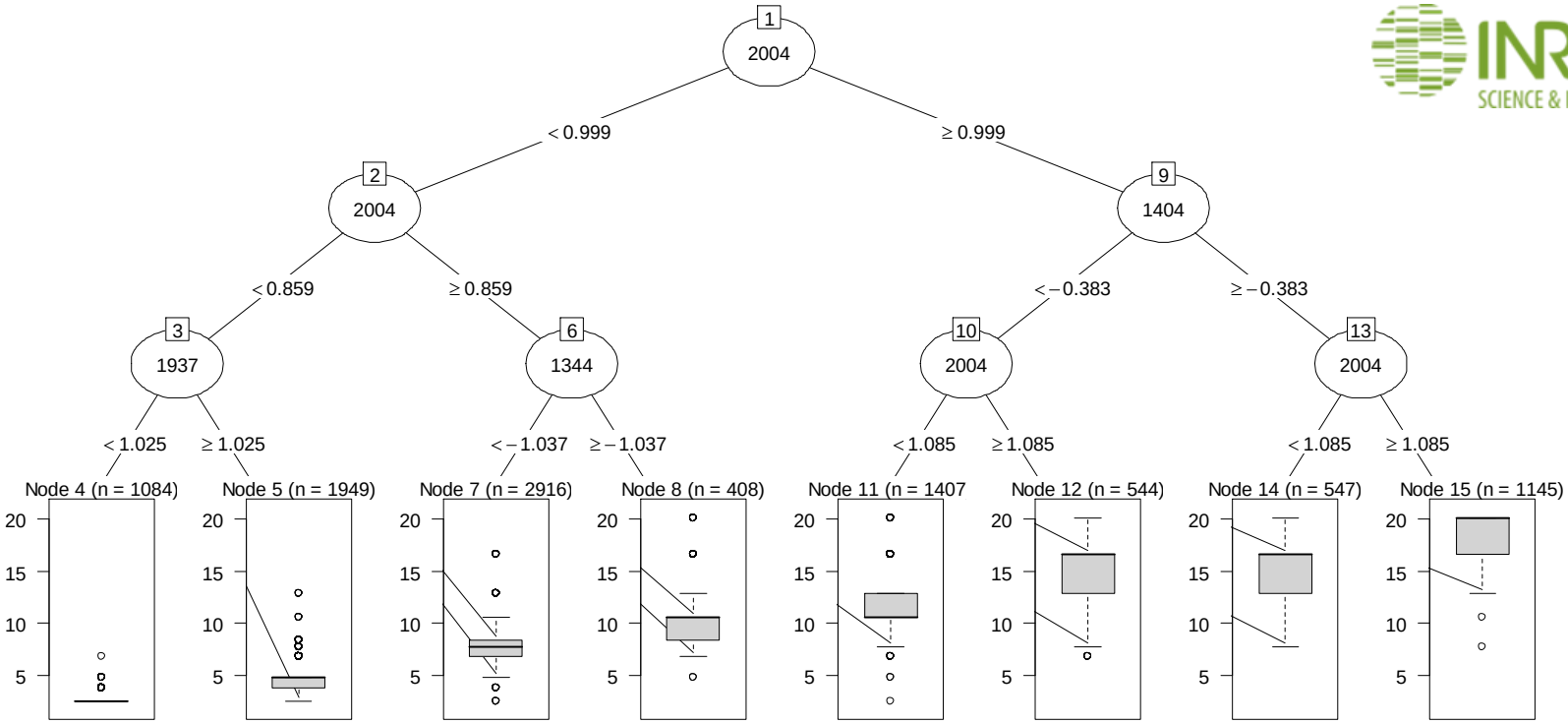




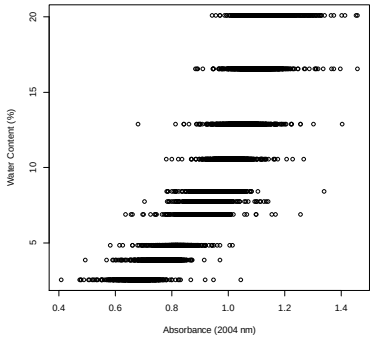
Comparaison of observed
and predicted values
of moisture
R^2=0.998



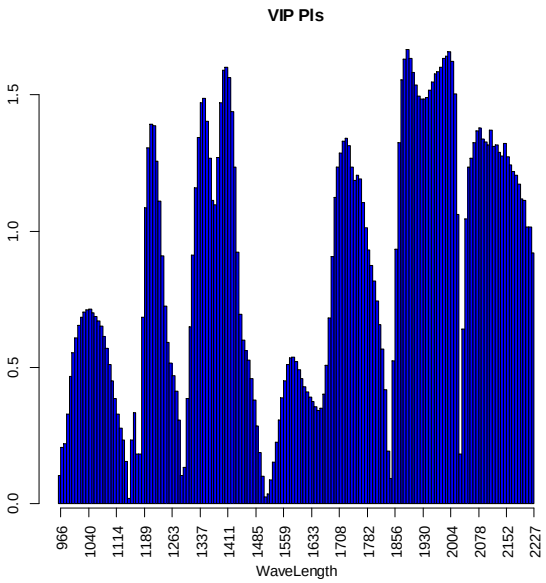
Regression tree with selected variables



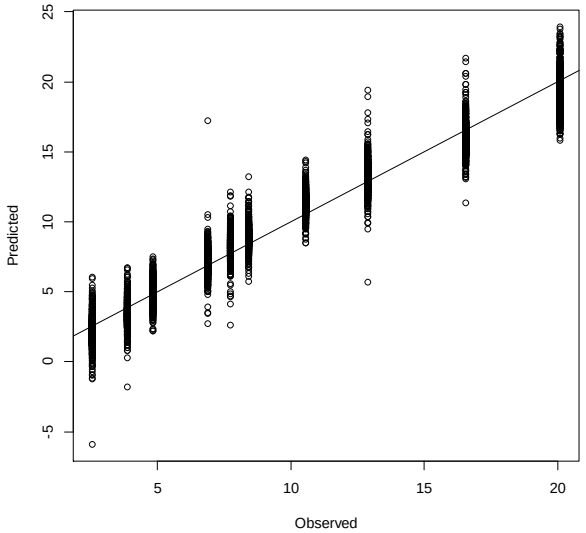
Moisture of biscuits
as a function of the absorbance at 2004 nm



Random forests *vs* PLS



Variable Importance in Projection (PLS)



Predicted *vs* Observed values (PLS)

	RMSE C	RMSE P	R^2 C	R^2 P
PLS	0.8037	0.8552	0.9776	0.9748
Random Forest	0.3256	0.7853	0.9963	0.9787

Results for the prediction of moisture (PLS vs Random forests)

To conclude…

NIR-HSI seems a very promising method to predict moisture inside food products and could probably be implemented on line at different steps of the process.

NIR imaging is able to give a spatial distribution of water content in biscuits and this may be related to the friability of the material.

Random Forests give parsimonious and accurate model for the prediction of water contents.
Several advantages over other regression techniques:

- Handle non linear relationships,
- No need of distributional hypotheses,
- Provide decision rules.