The area under the ROC curve as a variable selection criterion for multiclass classification problems

M. de Figueiredo[1,2] C. B. Y. Cordella[3] D. Jouan-Rimbaud Bouveresse[1,3] X. Archer[2] J.-M. Bégué[2] D. N. Rutledge[1,*]

[1] UMR Ingénierie Procédés Aliments, AgroParisTech, Inra, Université Paris-Saclay, 91300 Massy, France

[2] Laboratoire Central de la Préfecture de Police, 39bis rue de Dantzig, 75015 Paris, France

[3] UMR Physiologie de la Nutrition et du Comportement Alimentaire, AgroParisTech, Inra, Université Paris-Saclay, 75005 Paris, France

[*]Correspondence : rutledge@agroparistech.fr

UMR G1145 GENIAL

Ingénierie Procédés Aliments
(Food & Process Engineering)

AgroParisTech
INSTITUT DES SCIENCES ET INDUSTRIES DU VIVANT ET DE L'ENVIRONNEMENT
PARIS INSTITUTE OF TECHNOLOGY FOR LIFE, FOOD AND ENVIRONMENTAL SCIENCES

INRA SCIENCE & IMPACT

université PARIS-SACLAY

LCPP

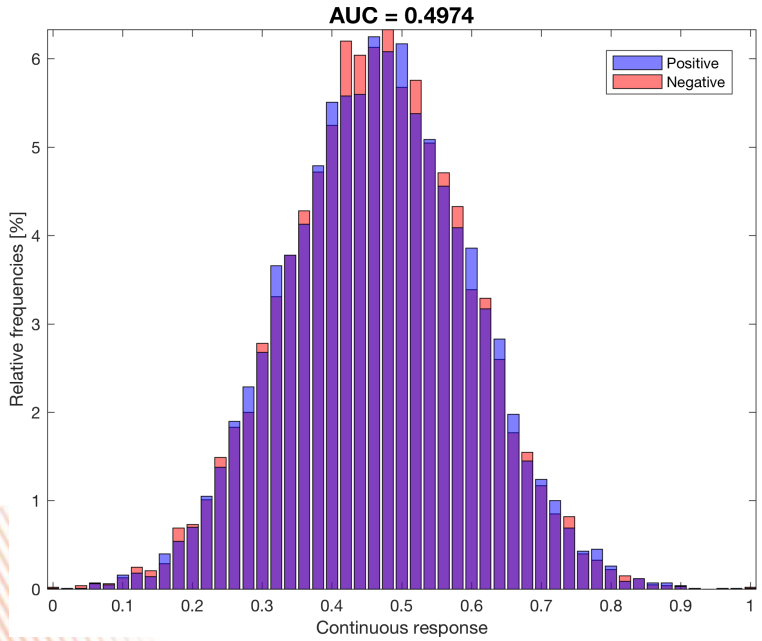
PREFECTURE DE POLICE

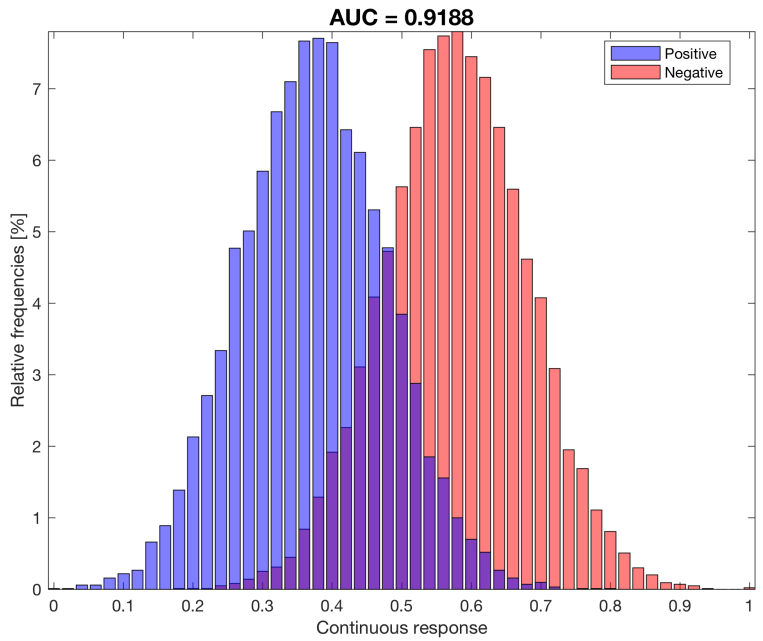
A bit of context

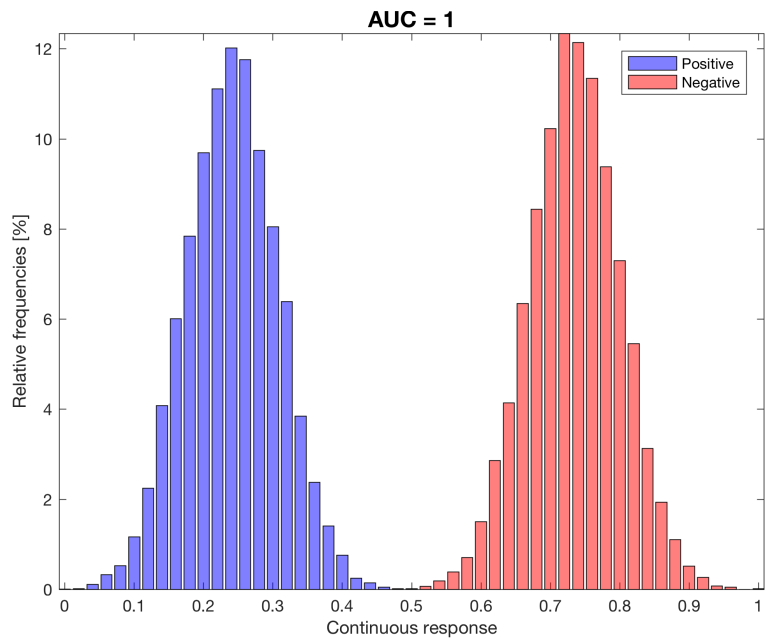
- **ROC analysis is mostly used for two-class problems**
  - Receiver Operating Characteristic (ROC) analysis uses the area under the ROC curve (AUC) as criterion to evaluate the discrimination between two classes
  - The AUC value is a direct measure of a classifier performances representing how well two classes are separated
  - Maximizing the AUC value is equivalent to maximize the separation between two classes
  - Classifier performances can be improved by selecting an appropriate subset of variables that combined together provide an optimal AUC
- **The objective**
  - Nowadays many fields are confronted with multiclass classification problems
  - Demonstrate how two-class ROC analysis and its associated AUC can be used to perform variable selection within a multiclass problem framework

ROC analysis: basic principles

- **Lets consider two vectors of continuous responses whose values range from 0 to 1**
  - Each vector represents a group (positive or negative) that is to be discriminated from the other one
  - The better the separation between the two distributions, the higher the AUC value is

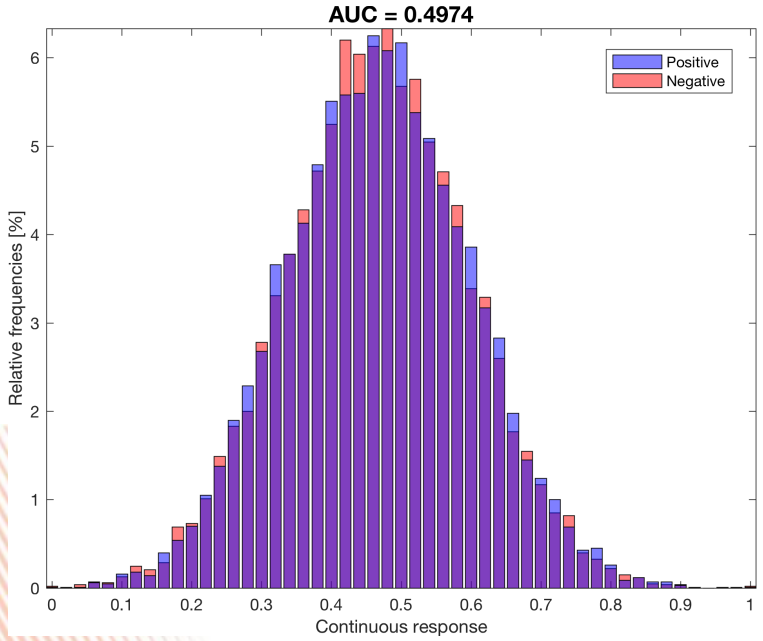


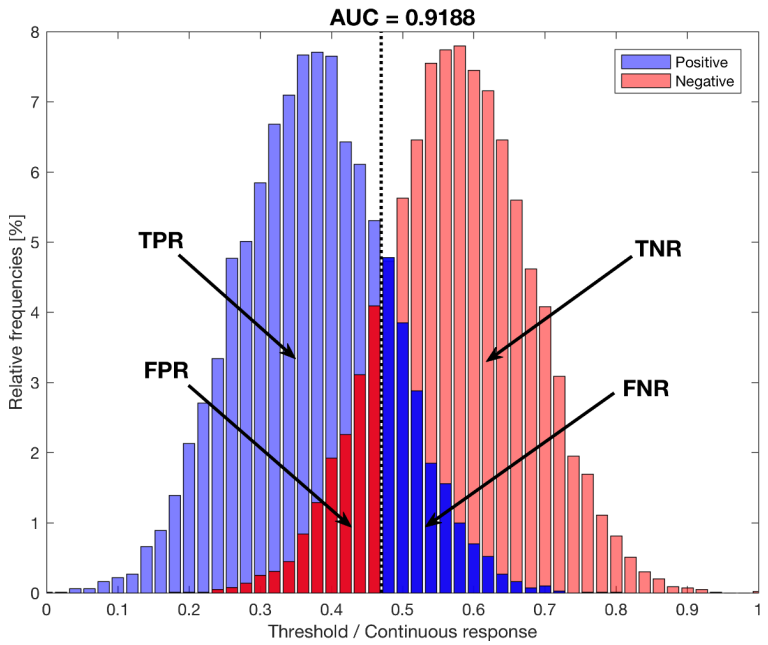


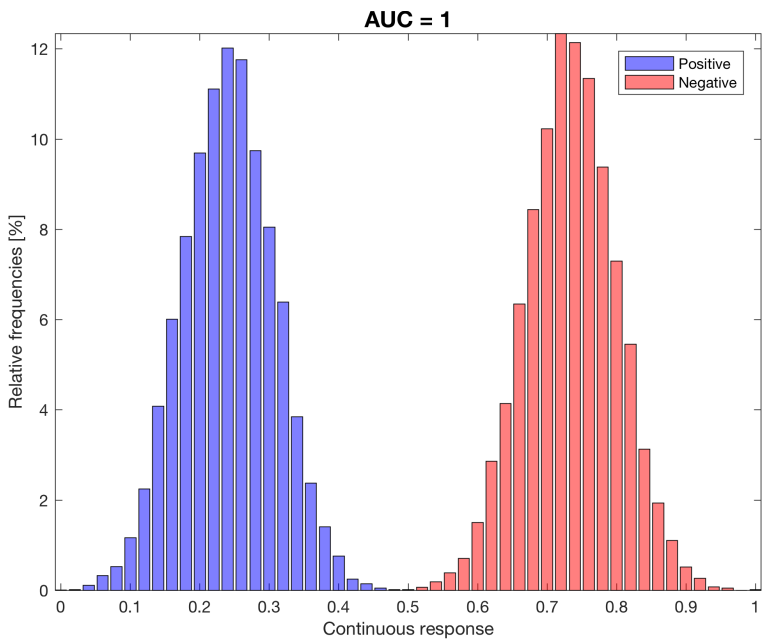


ROC analysis: basic principles

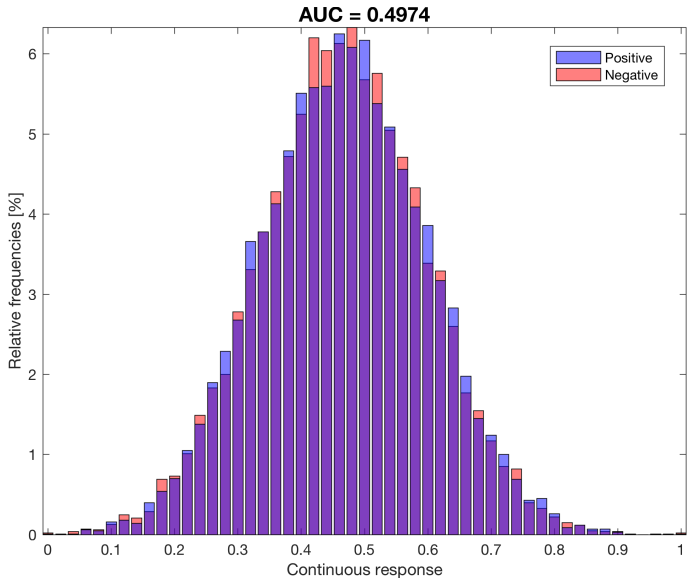
- **Lets consider two vectors of continuous responses whose values range from 0 to 1**
  - Example of decision threshold defining classification rates
  - How is the AUC calculated? Force variation of the threshold between 0 and 1
  - At each step, the TPR and FPR are recorded
  - The ROC curve is the representation of the TPR recorded as a function of the FPR

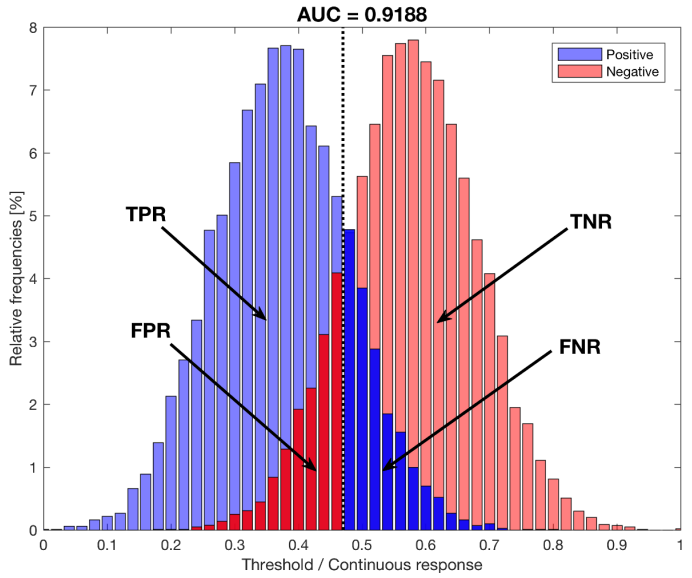


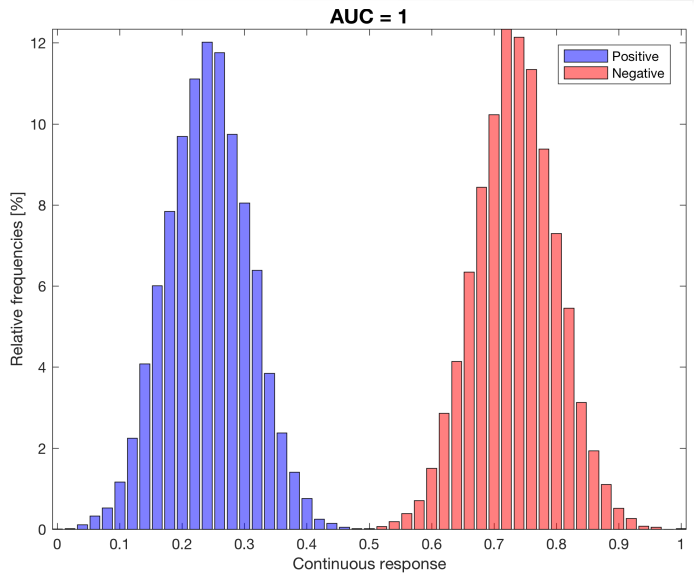


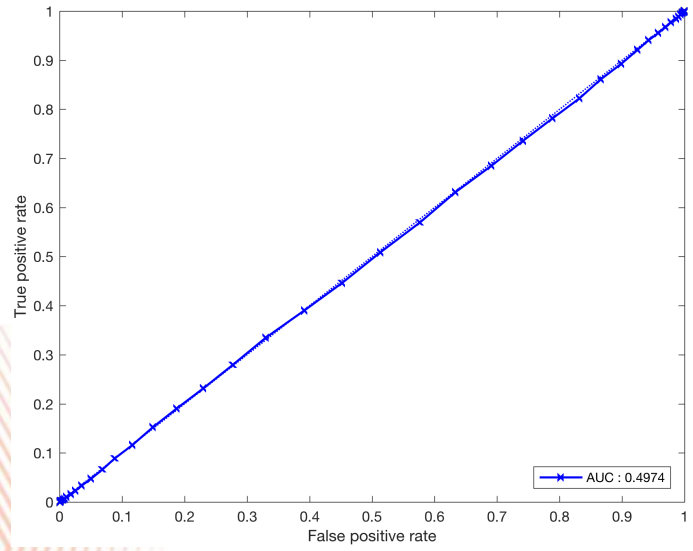


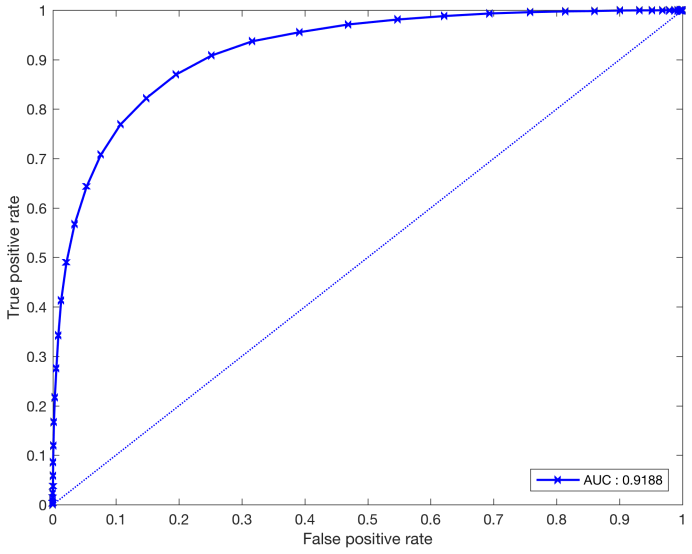
ROC analysis: basic principles

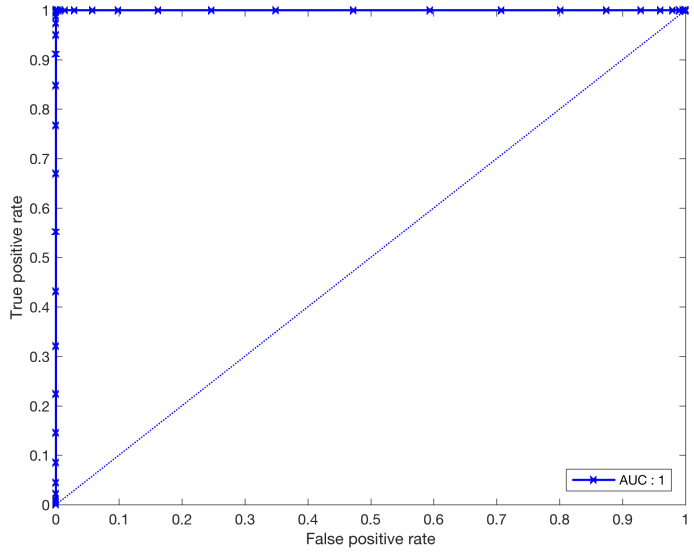










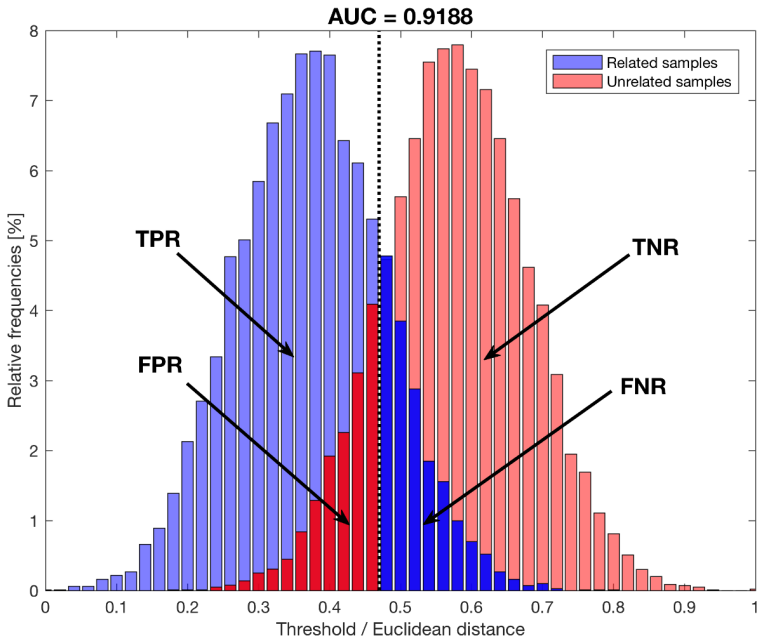


Examples of ROC curves for a varying threshold with a step of 0.2

Similarity measurements

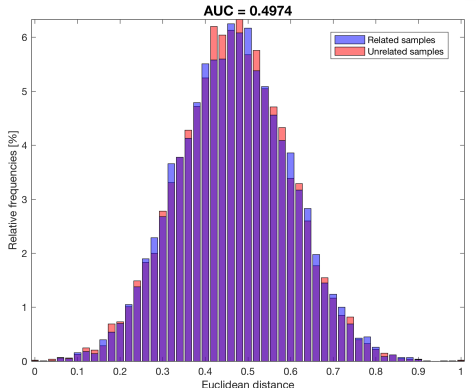
- **Lets consider now that the continuous responses are Euclidean distances**
  - Each vector element corresponds to the distance between two samples for a given set of variables
  - Either pairs of samples belong to the same class (related samples) and should be close to each other or the pairs of samples belong to different classes (unrelated samples) and should be far from each other

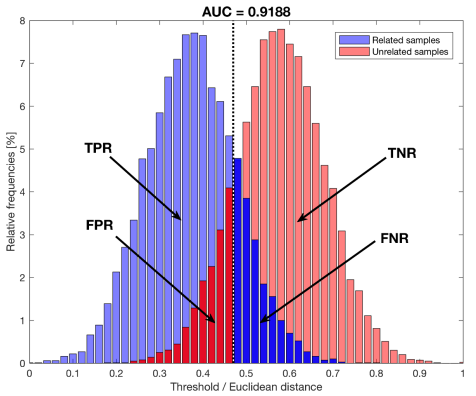
- **A similarity measurement quantifies how much alike two samples are**
  - Elements in the vector of (un)related samples correspond to the Euclidean distance between pairs of samples (not) belonging to the same class

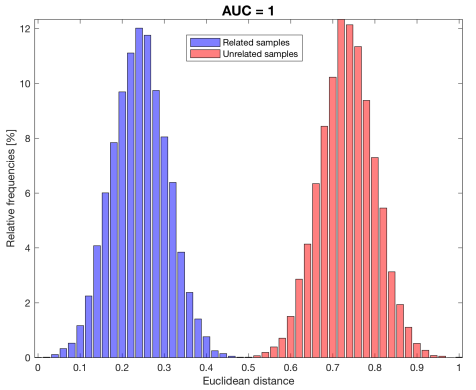


The variable selection method

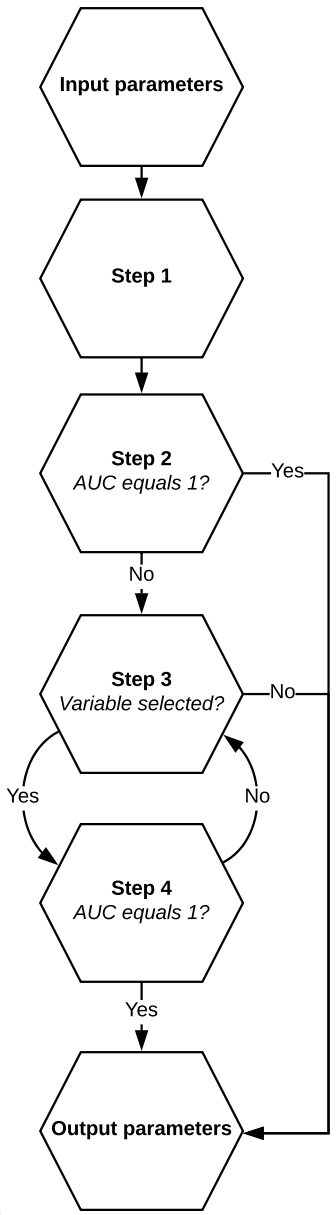
- **The goals of the variable selection method based on the AUC criterion**
  - Select a subset of variables maximizing the AUC value
  - Maximizing the AUC value is equivalent to maximize the separation between the distributions of related and unrelated samples
  - Maximizing the separation between the two distributions is equivalent to bringing closer together in the multivariate space samples belonging to the same class and, at the same time, separating groups of samples belonging to different classes
  - This point is true whether we have a two-class problem at hand or a multiclass one
  - As long as the separation between the two distributions is maximized, no matter the number of classes, it is always possible to reduce the problem to a two-class one
  - If the AUC represents how well two classes are separated, it can also represent a global value of how well several classes are separated







The variable selection method



X_in	Groups	Increment
There are *n* samples in the rows and *p* variables in the columns	Binary matrix with samples on the rows and actual groups on the columns	Minimal increase of the AUC for a variable to be selected

The algorithm calculates for each variable in **X_in** an individual AUC value and sorts decreasingly the columns in **X_in** according to the AUCs estimated

Note : The sorted data matrix is used all along in the following steps

The algorithm stores in the ouput data matrix **X_out** the variable with the highest AUC value, meaning the first column of the matrix sorted in **step 1** and removes it from **X_in**

The algorithm iteratively runs through the columns in **X_in** looking for a variable which combined to the already stored variable(s) in **X_out** increases the global AUC value of at least a specific 'Increment' value

Note : The algorithm stops only if the global AUC value equals 1 or if there is no variable selected in step 3

The algorithm iteratively runs through the variables stored in **X_out** looking for increasing the AUC value by eleminating one or several variables

Note : Backward step looking for redundancies in selected variables

X_out	VarIndexOut
Combination of variables maximizing the AUC value	Index of the selected variable in the input parameter **X_in**

- **AUC estimation**
  - Everytime an AUC value is estimated according to Hand & Till (2001), vectors of intra- and inter-similarities are calculated beforehand for a given subset of variables
  - This estimation is not based on the effective representation of the ROC curve

$$\widehat{AUC} = \frac{S_0 - n_1(n_1+1)/2}{n_0 n_1}$$

Hand, D. J., & Till, R. J. (2001). A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. Machine Learning, 45(2), 171-186.

A practical example : the dataset, samples preparation and analysis

Dataset

190 gasoline samples representing **4 qualities** (SP95, SP98, SP95-E10 and SP-E85) and bought over **4 seasons** (fall, winter, spring and summer) in 20 different gas stations

QC sample

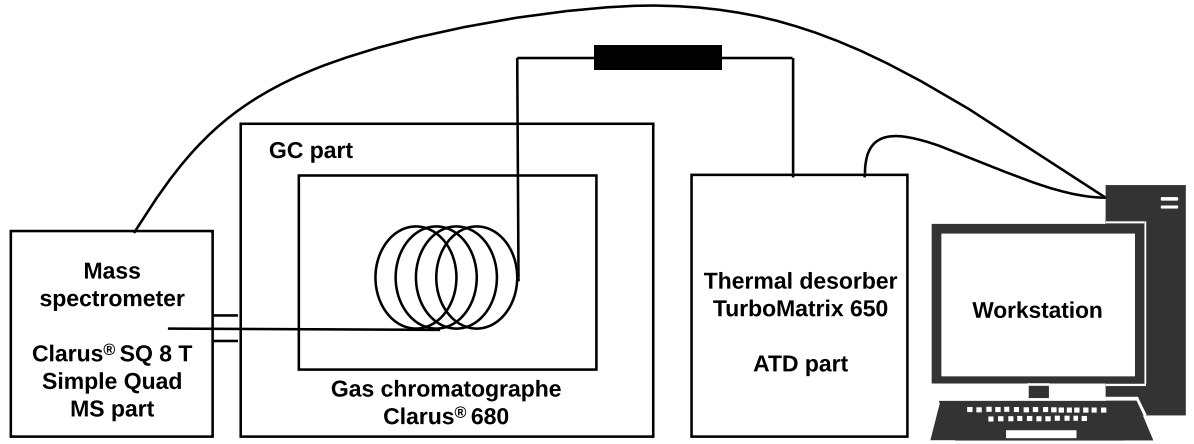
Mix of 98 gasoline samples prepared and analyzed in parallel with regular samples

Samples preparation

Passive headspace thermal extraction onto Tenax TA® tubes

Samples prepared in triplicates

Samples analysis by ATD-GC-MS



Data samples acquired

570 gasoline samples
86 QC samples

A practical example : data preparation

Features detection and integration with XCMS (Smith et al., 2006)

723 features detected and integrated

Normalization increasing repeatability: corrects for differences in the passive headspace extraction process

$$Ratio_{A,B} = \frac{A}{A+B}$$

(Daling et al., 2002)

522,006 ratios

Calculates ratios from all possible combinations between pairs of features with repetition

Ratios of QCs with RSD ≤ 5 %

110,279 ratios

(1) Selects repeatable ratios over time;
(2) Increases speed in next step;

Variable selection

***x* ratios**

Selects ratios maximizing the AUC

The number of ratios selected depends on the property investigated

Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R., & Siuzdak, G. (2006). XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. Analytical Chemistry, 78(3), 779-787.
Daling, P. S., Faksness, L.-G., Hansen, A. B., & Stout, S. A. (2002). Improved and Standardized Methodology for Oil Spill Fingerprinting. Environmental Forensics, 3(3-4), 263-278.

A practical example : multivariate analysis

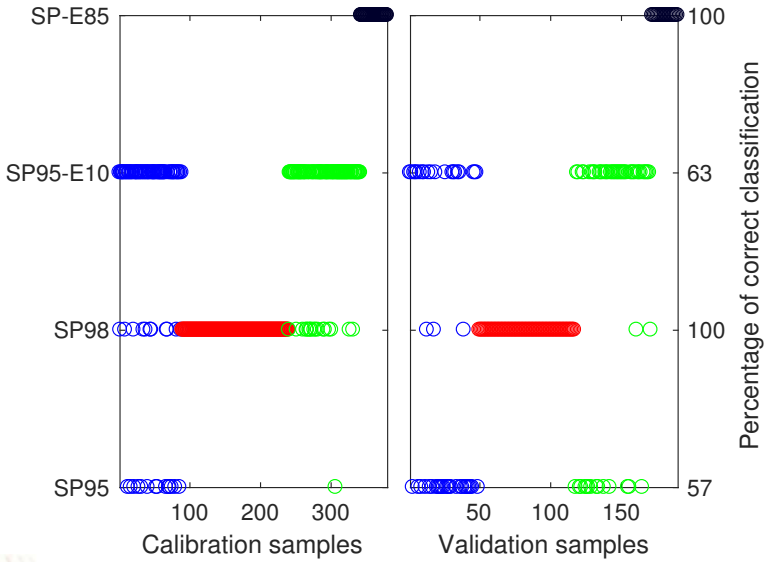
- **Cross-validation PLS-DA**
  - Using the SAISIR toolbox (Cordella, C. B. Y. et Bertrand, D., 2014)
  - 100 holdout/testset cross-validation PLS-DA using random subsets of 2/3 of the samples for the calibration and the remaining 1/3 as a validation set
  - Cross-validation PLS-DA was performed on the gasoline dataset with and without variable selection to evaluate the predictability of the models
  - The accuracy (percentage of correct classification) was used to evaluate the predictability of the models

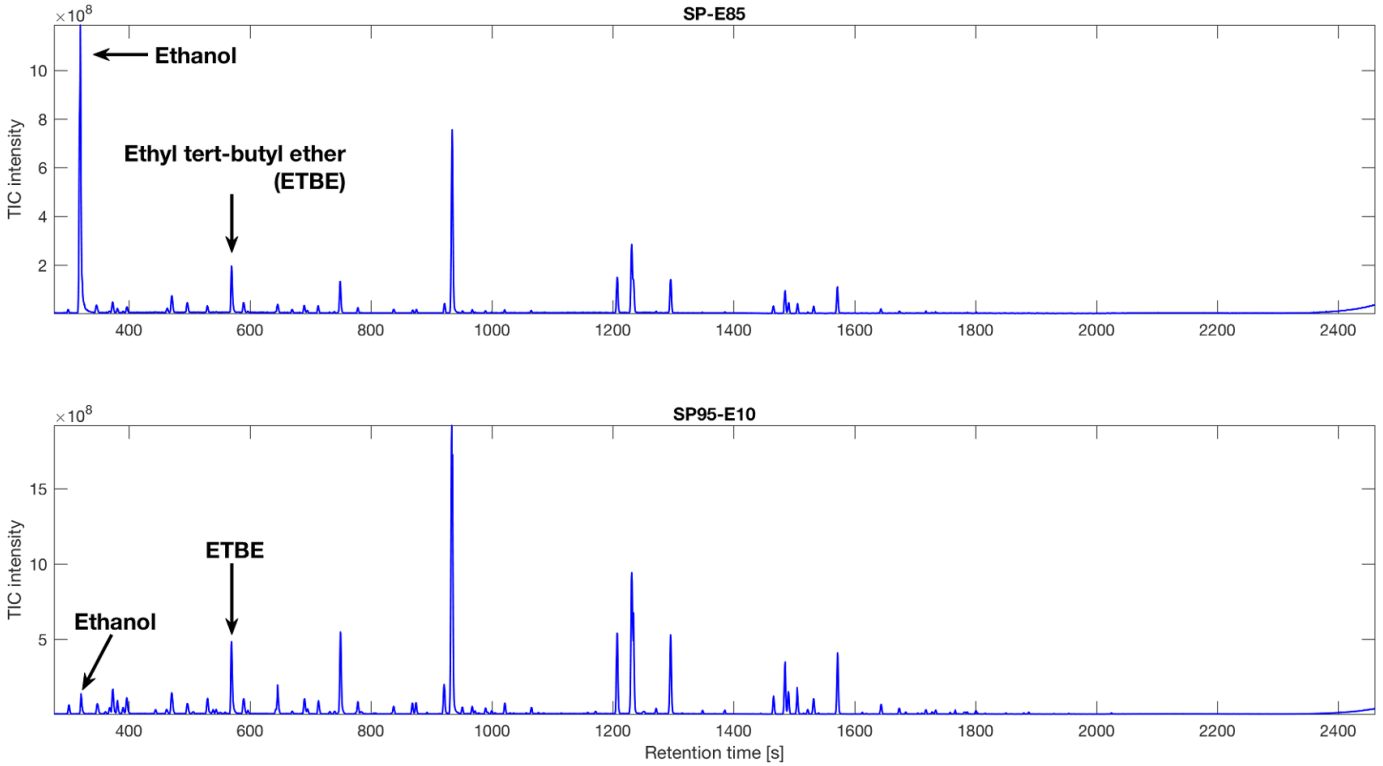
Cordella, C.B.Y. & Bertrand, D. SAISIR: A new general chemometric toolbox. TrAC Trends in Analytical Chemistry 54, 75-82, 2014.

A practical example : results & discussion for gasoline quality

Note that columns marked by (*) concern results with a prior variable selection process

Property	Increment	Ratios selected	AUC	LVs*	Accuracy*	LVs	Accuracy
Quality	0.001	4	0.9877	2	79.45±3.56 %	3	77.20±3.82 %

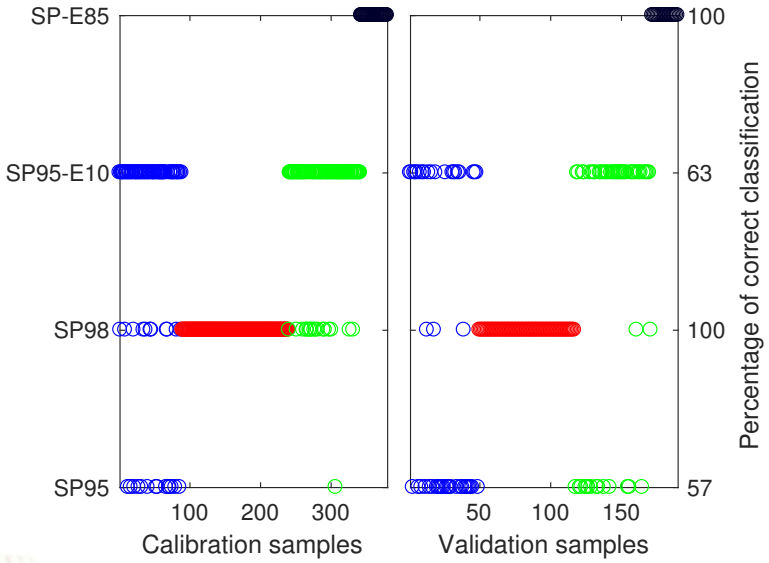


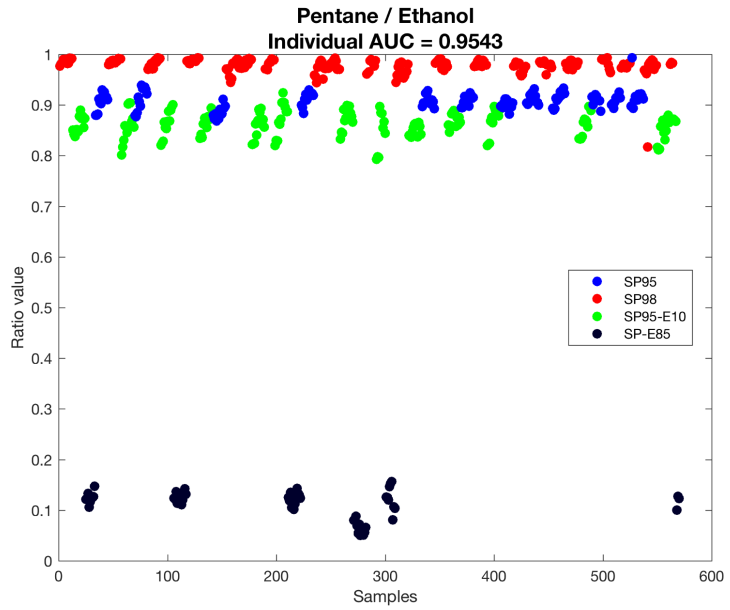


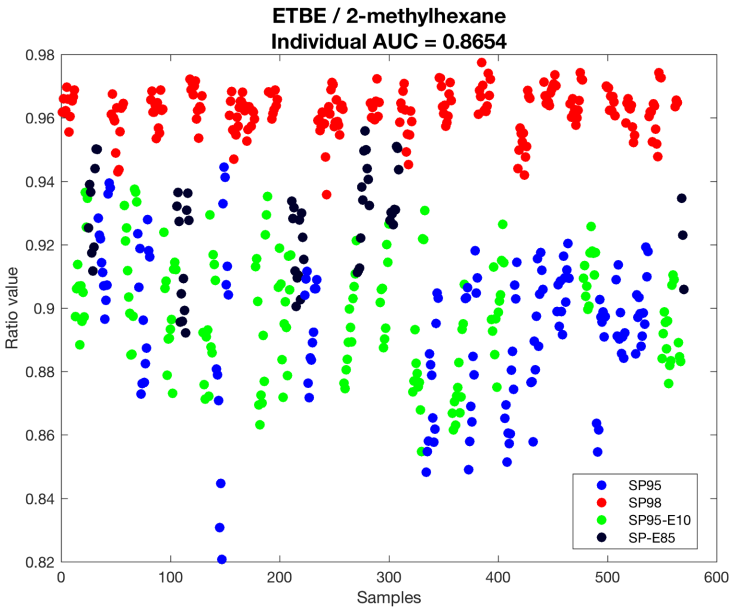
A practical example : results & discussion for gasoline quality

Note that columns marked by (*) concern results with a prior variable selection process

Property	Increment	Ratios selected	AUC	LVs*	Accuracy*	LVs	Accuracy
Quality	0.001	4	0.9877	2	79.45±3.56 %	3	77.20±3.82 %



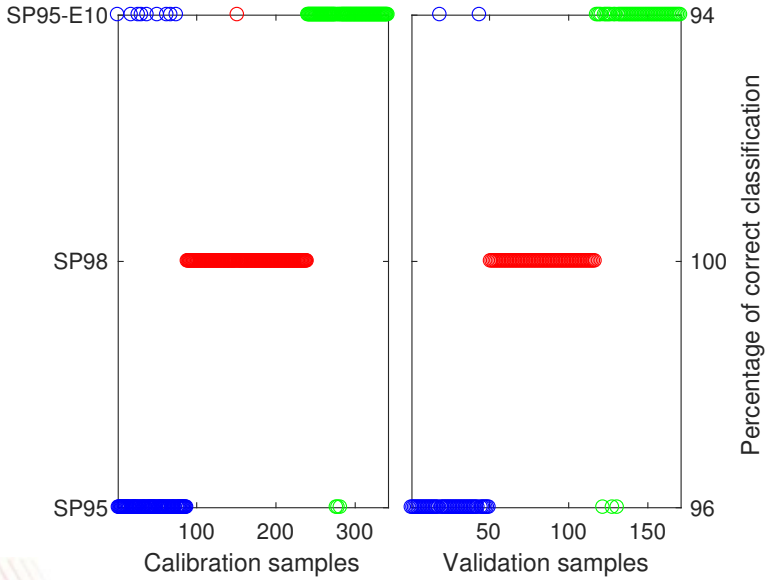


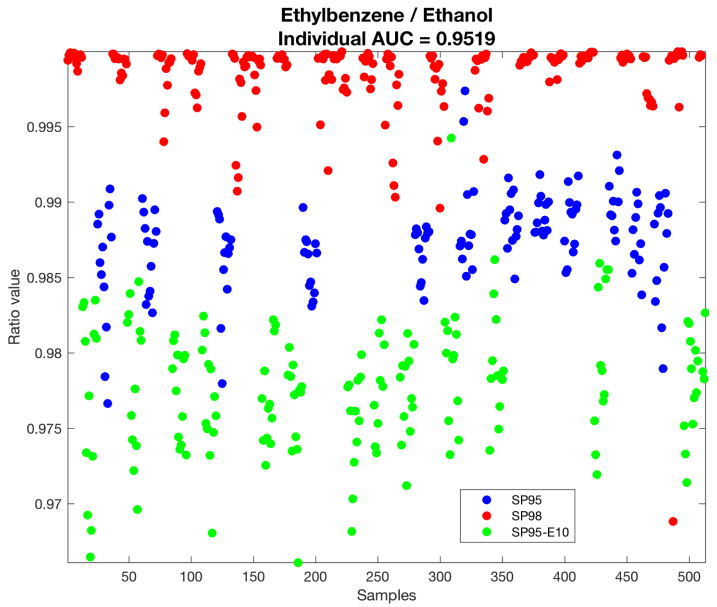


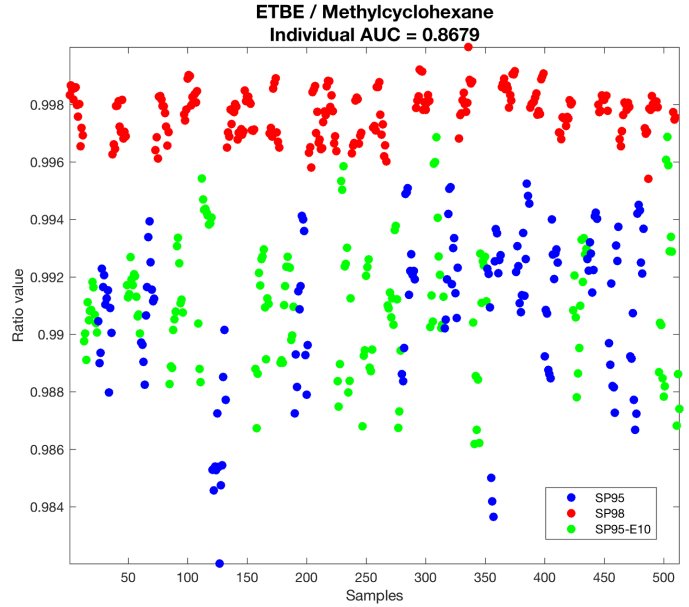
A practical example : results & discussion for gasoline quality

Note that columns marked by (*) concern results with a prior variable selection process

Property	Increment	Ratios selected	AUC	LVs*	Accuracy*	LVs	Accuracy
Quality	0.001	4	0.9877	2	79.45±3.56 %	3	77.20±3.82 %
Quality (no SP-E85)	0.001	9	0.9862	2	96.87±1.14 %	5	79.15±2.67 %



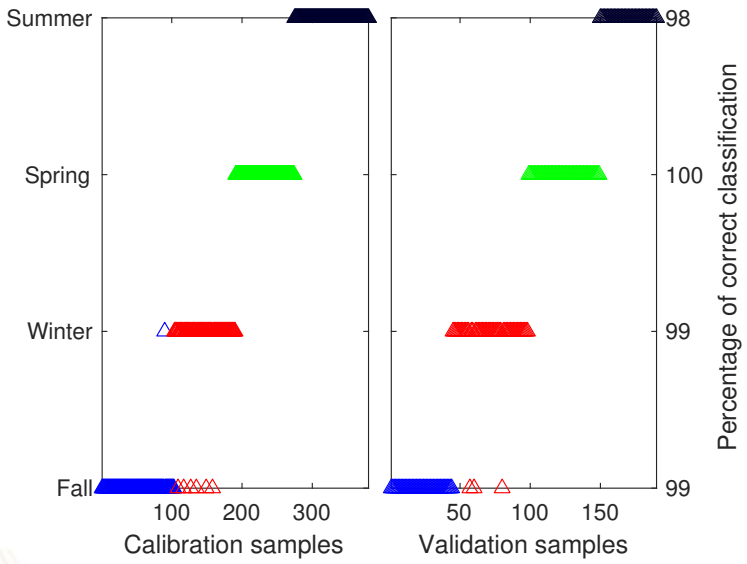


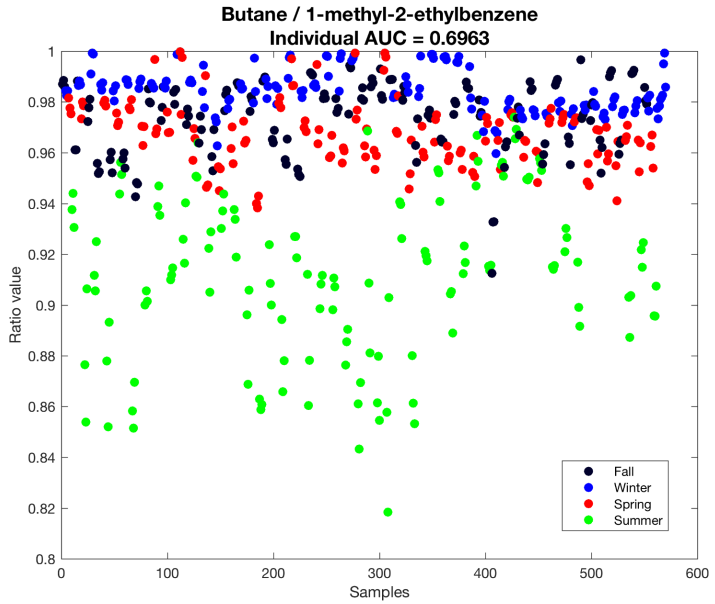


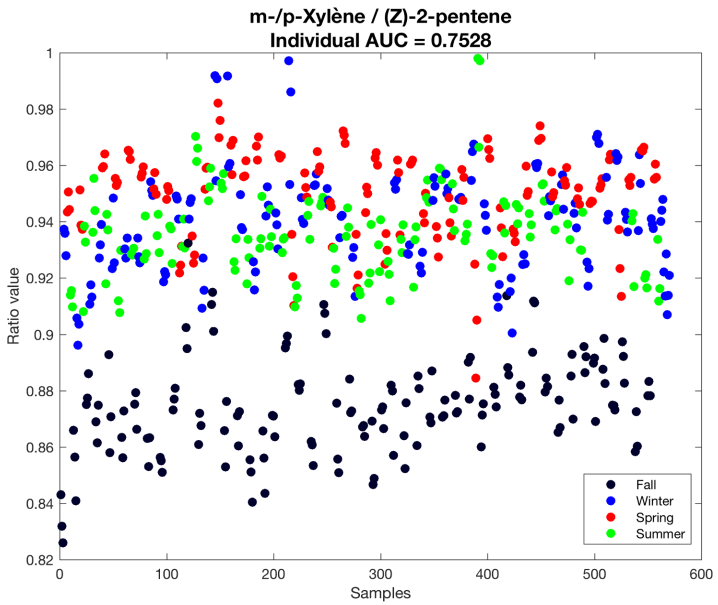
A practical example : results & discussion for gasoline quality

Note that columns marked by (*) concern results with a prior variable selection process

Property	Increment	Ratios selected	AUC	LVs*	Accuracy*	LVs	Accuracy
Quality	0.001	4	0.9877	2	79.45±3.56 %	3	77.20±3.82 %
Quality (no SP-E85)	0.001	9	0.9862	2	96.87±1.14 %	5	79.15±2.67 %
Season	0.001	21	0.9789	4	98.88±0.68 %	9	97.67±1.14 %







A conclusion

- **An adequate variable selection procedure should in general :**
  - Substantially reduce the dimensionality of the data
  - Improve the predictability of the models built
  - Facilitate the interpretation of the chemical systems

- **The variable selection presented here :**
  - Performs data dimensionality reduction by selecting variables in their orginal space and not by building linear combinations of the original variables (reduction close to 100% in the given example)
  - Proposes a way to use the traditional two-class ROC analysis within a multiclass context
  - Can implement different distance measures other than Euclidean distance
  - Can be sped up by avoiding the calculation of similarities between all possible pairs of samples
  - Because of the evermore growing amounts of data produced by modern analytical procedures, being able to reduce the dimensionality of the data while maximizing the predictive power of the models and their interpretability if of utmost importance

UMR G1145 GENIAL

Ingénierie Procédés Aliments
(Food & Process Engineering)

AgroParisTech
INSTITUT DES SCIENCES ET INDUSTRIES DU VIVANT ET DE L'ENVIRONNEMENT
PARIS INSTITUTE OF TECHNOLOGY FOR LIFE, FOOD AND ENVIRONMENTAL SCIENCES

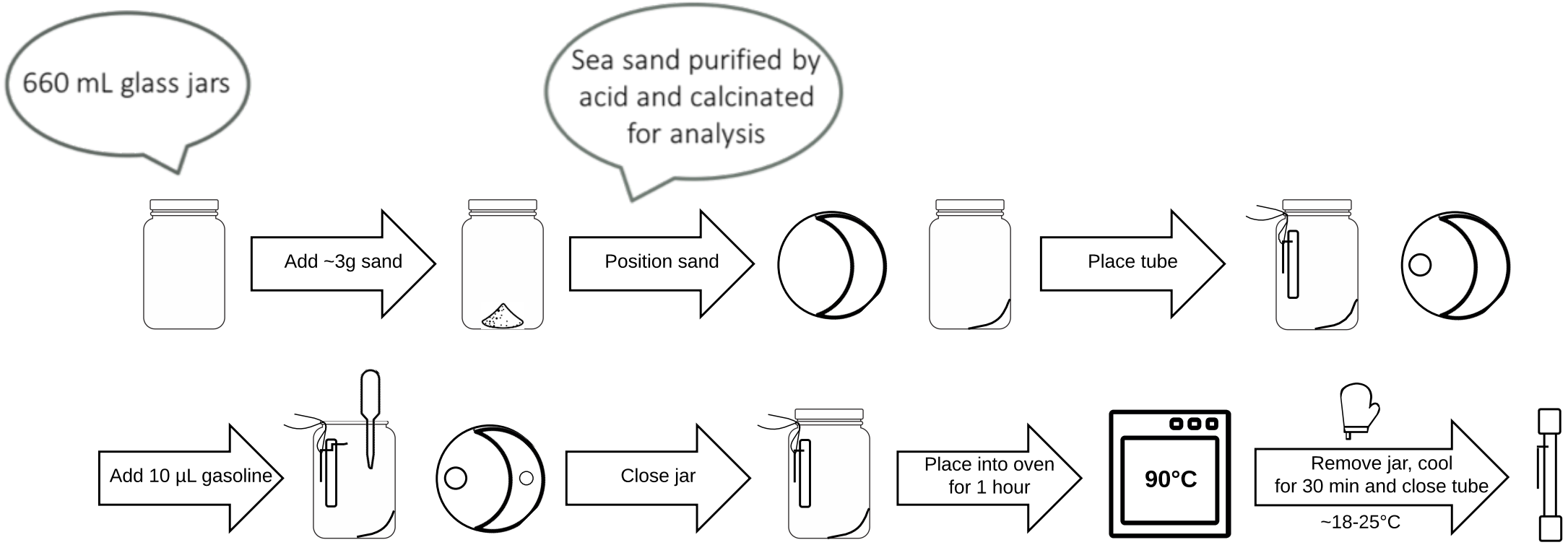
INRA SCIENCE & IMPACT

université PARIS-SACLAY

LCPP

PREFECTURE DE POLICE

Additional information – Preparation of samples



Additional information – Analysis of samples

Mass spectrometer parameters :
Mass range : 30 to 450 m/z;
Detection from 0 to 45.75 min;
Positive ionisation mode EI+;
Centroid aquisition;
Scan time : 0.5 sec;
Interscan delay : 0.1 sec;
Threshold (ion count) : 0

MS part

Mass spectrometer Simple Quad Clarus® SQ 8 T

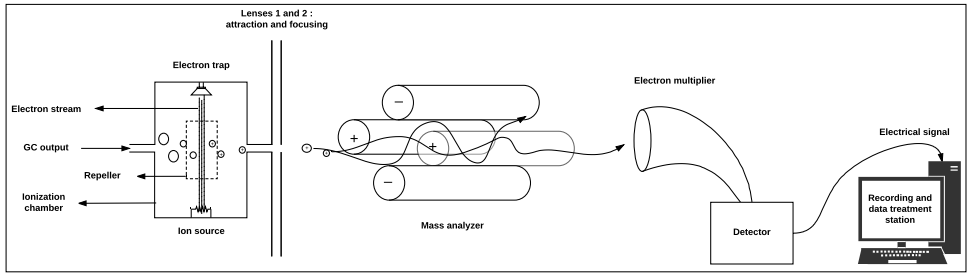
GC part

Gas chromatography Clarus® 680

Thermal desorber TurboMatrix 650

TD part

Recording and data treatment station



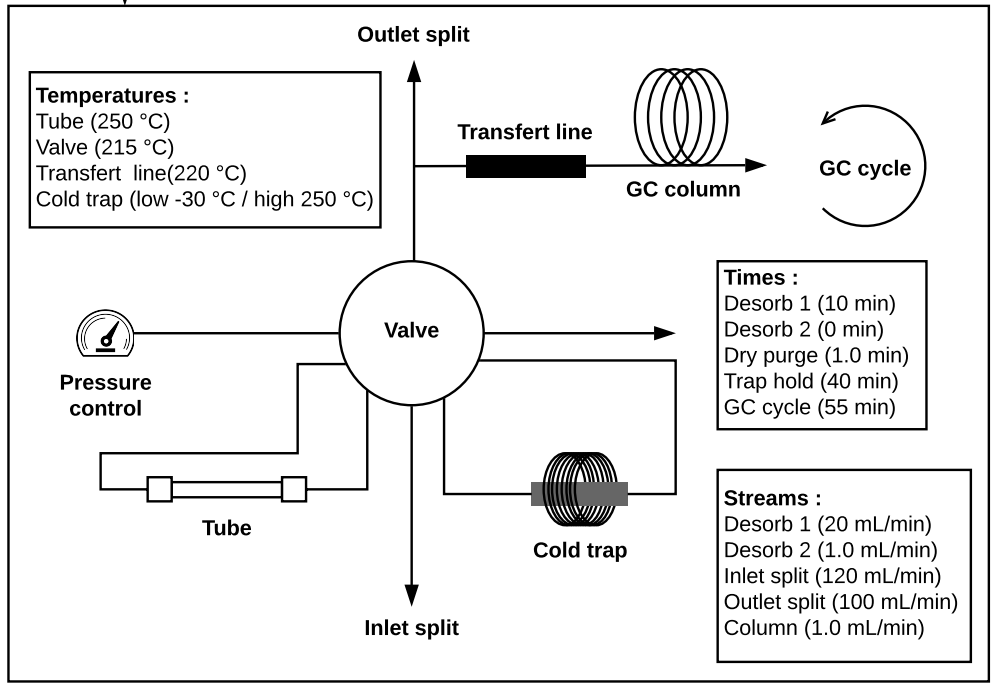
GC parameters

GC Capillary column :
Type : Restek Rxi®-1MS Crossbond® 100% dimethylpolysiloxane;
Length : 60m; Diametre : 250 µm;
Thickness : 0.5 µm;

Oven temperature program :
Initial temperature : 35.0°C for 5.0 min;
Ramp 1 : 4.0°C/min until 150°C;
Hold for 0.0 min;
Ramp 2 : 20.0°C/min until290 °C;
Hold for 5.0 min;
GC run time : 45.75 min;

TD parameters

Temperatures :
Tube (250 °C)
Valve (215 °C)
Transfert line(220 °C)
Cold trap (low -30 °C / high 250 °C)



Times :
Desorb 1 (10 min)
Desorb 2 (0 min)
Dry purge (1.0 min)
Trap hold (40 min)
GC cycle (55 min)

Streams :
Desorb 1 (20 mL/min)
Desorb 2 (1.0 mL/min)
Inlet split (120 mL/min)
Outlet split (100 mL/min)
Column (1.0 mL/min)