A.N.F. Plasma induit par laser pour l'analyse de la matière. 15-17 novembre 2021, Orléans

Approches statistiques multivariées

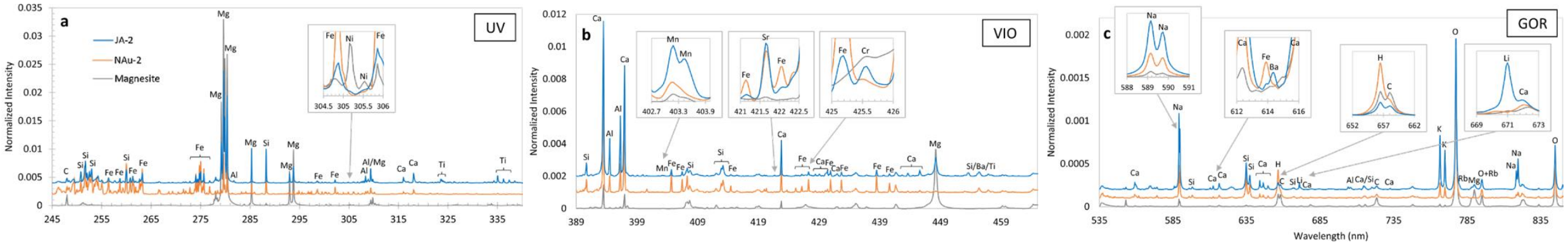
Bruno Bousquet, Université de Bordeaux

Sommaire

Pourquoi faire appel à des approches multivariées ?

⇒ ***Lorsque la question posée nécessite de prendre en compte plusieurs variables !!!***

Exemple 1: âge, poids, activité physique, régime alimentaire…=> maladie cardiaque
Exemple 2: composition chimique, propriétés physiques => matériau recyclé ou non



Un spectre LIBS contient des centaines de variables !

Lorsque les échantillons ne partagent pas les mêmes caractéristiques physiques et/ou chimiques

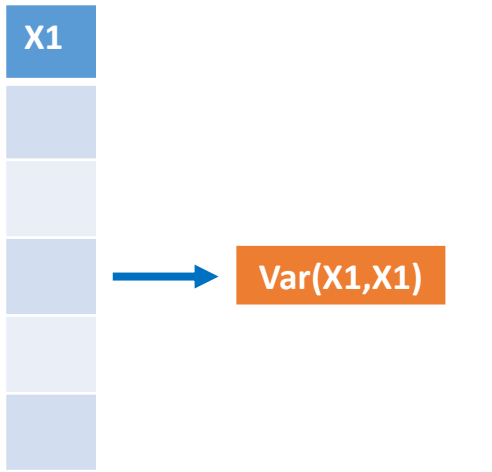
- **Modèle de régression univarié inutilisable**
- **L'approche multivariée permet de prendre en compte les effets de matrice**

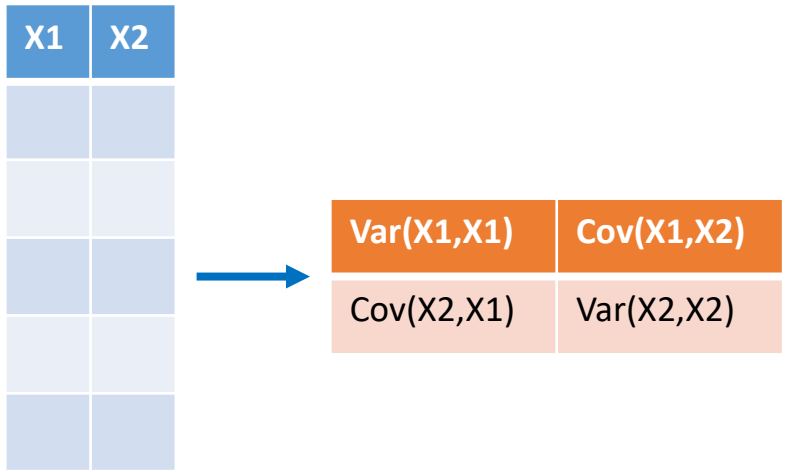
L'observation des données

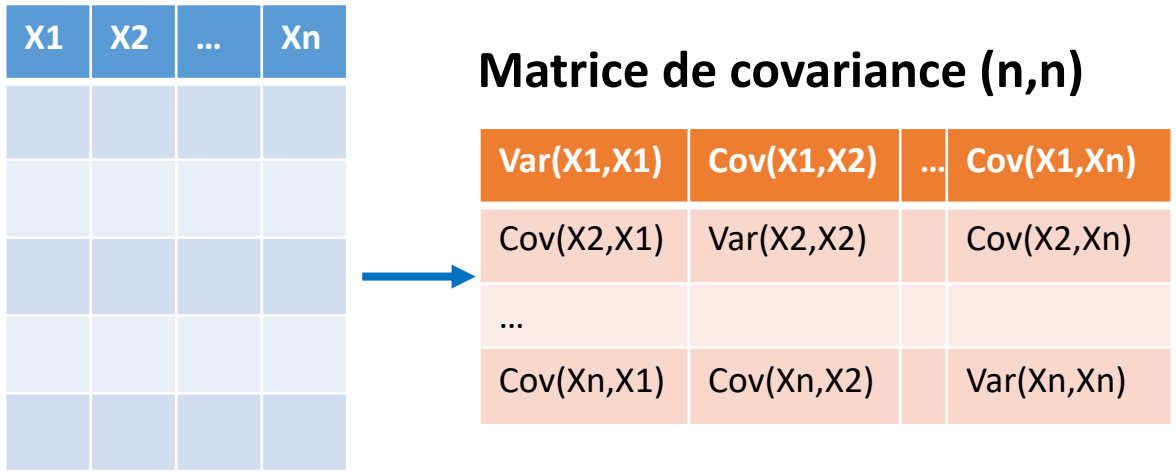
Covariance

Comment observer les différences entre 1000 spectres LIBS contenant chacun plus de 1000 variables ?

⇒ Il faut transposer à N-dimensions les analyses statistiques que l'on sait faire à 1 dimension :
calcul de moyenne et d'écart-type (ou variance =c arré de l'écart-type)







1 variable X1

2 variables X1 et X2

n variables X1 à Xn

Analyse en composantes principales -ACP

La matrice des covariances contient des milliers voire des millions de valeurs numériques

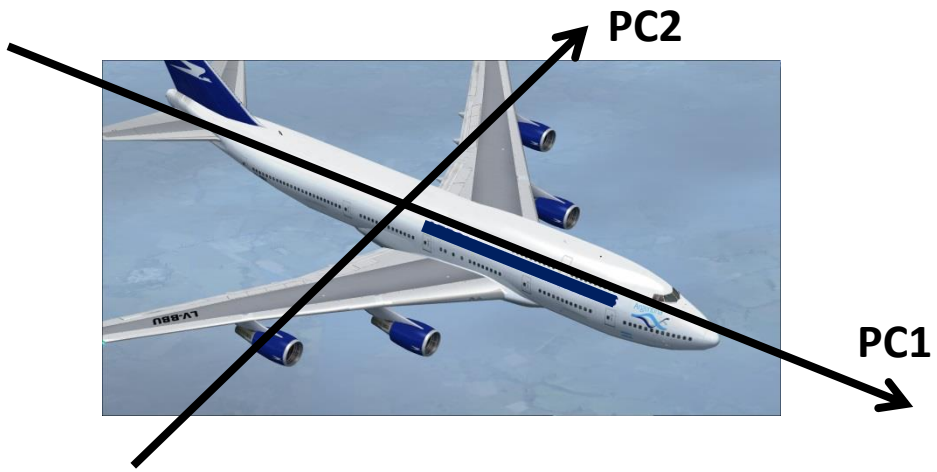
⇒ Résultats les plus marquants dans un espace de plus petite dimension
⇒ C'est ce que fait **l'analyse en composantes principales** (ACP ou PCA)

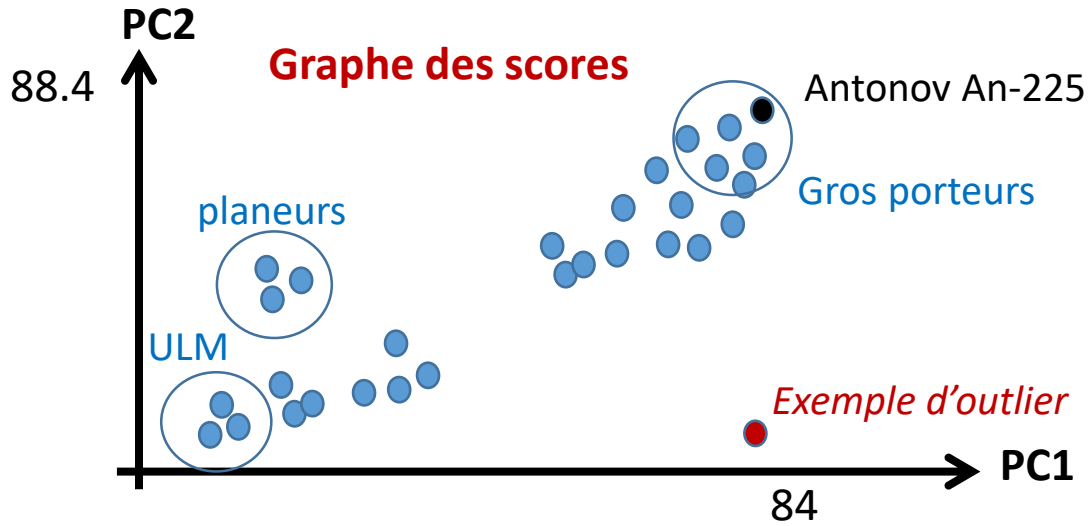
Exemple: comment différencier un millier d'avions entre eux ?

Les variables sont : longueur, hauteur, envergure et forme des ailes, la dimension de la queue, le nombre de réacteurs, etc…

Pour simplifier, considérons que le paramètre qui varie le plus d'un avion à l'autre soit la longueur et le suivant l'envergure.

Dans ce cas l'axe mesurant la longueur devient celui de la première composante principale PC1 et celui mesurant l'envergure, PC2.



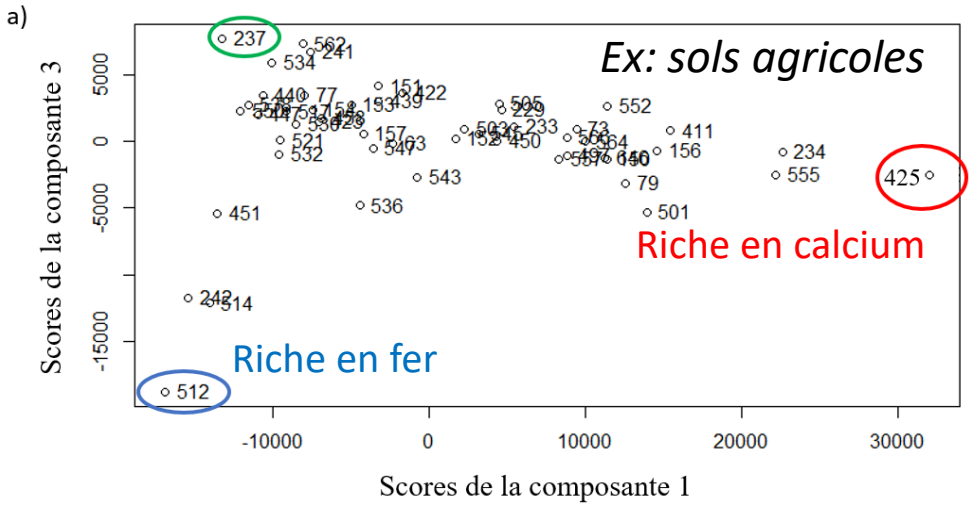


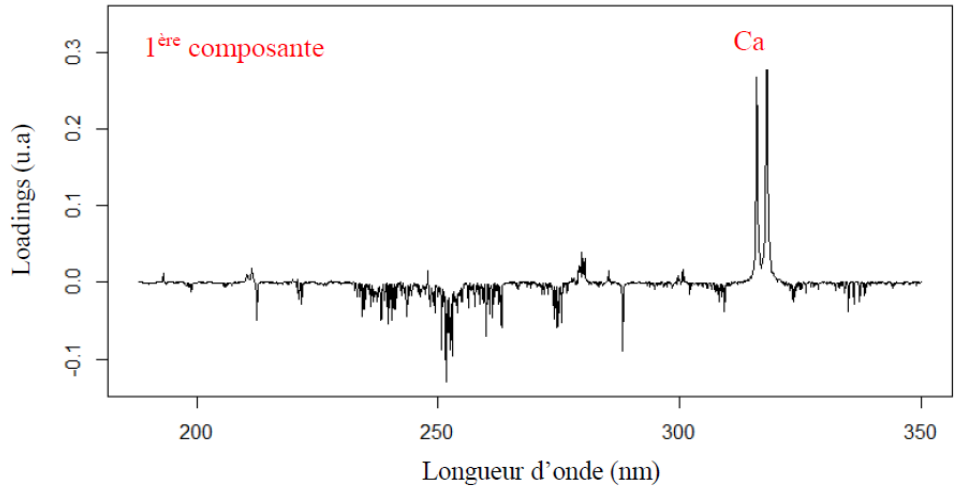
ACP appliquée aux données LIBS

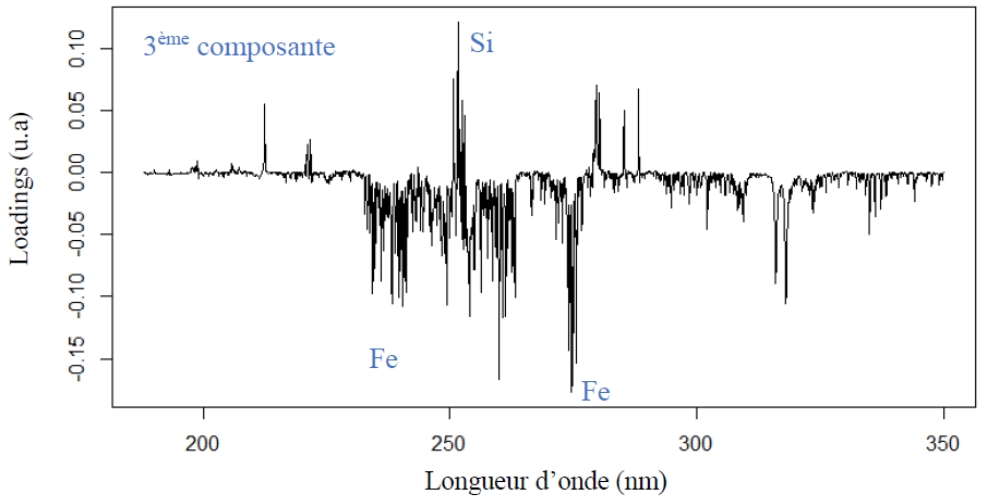
⇒ Les composantes principales PC sont des combinaisons linéaires des variables (longueurs d'onde)

⇒ Chaque point dans les graphes des scores représente un spectre LIBS

Matrice X	λ1	λ2	λ3	…	λn
Spectre 1	I11	I12	I13		I1n
Spectre 2	I21	I22	I23		I2n
Spectre 3	I31	I32	I33		I3n
…	*Données centrées (on soustrait la valeur moyenne)*				
Spectre p	Ip1	Ip2	Ip3		Ipn







Bilan

⇒ ***Une observation des spectres LIBS est fortement recommandée en préalable à de leur analyse***

- **Détection des outliers :**
  - Dû à l'échantillon lui-même (composition chimique; propriétés physiques très différente de la moyenne des échantillons observés)
  - Dû à une anomalie des conditions expérimentales
  - *On exclut les données anormales !*

- **Détection de groupes (classes) de points :**
  - Meilleure connaissance du jeu de données
  - A l'œil nu ou par calcul de distances entre points
  - Interprétations: composition chimique, propriétés physiques, besoins en normalisation

- **Très grande réduction de dimensionnalité ; compression**

- **Mise en évidence de corrélation/anti-corrélation entre variables**

L'analyse quantitative

L'analyse quantitative - principe

Construire un modèle permettant de relier les données LIBS aux valeurs de concentration de l'élément d'intérêt.

Variable d'entrée X1
Signal extrait du spectre LIBS pour la raie d'émission choisie

Modèle univarié
Régression linéaire par moindres carrés

Donnée de sortie Y
Valeur prédite de concentration pour l'élément d'intérêt

- Pour un élément donné, les performances dépendent de:
  la raie d'émission, la méthode d'extraction des données et la normalisation.
- Pour construire le modèle (apprentissage), il faut un lot de spectres pour lesquels la concentration est connue : ***lot de calibration***
- Pour valider le modèle, il faut un lot de spectres pour lesquels la concentration est connue mais qui n'ont pas servi à construire le modèle; ***lot de validation***
- On peut estimer la ***précision du modèle grâce aux limites de prédiction***

L'analyse quantitative – bases de l'approche multivariée

Variables d'entrée X
extraites du spectre LIBS

→

Modèle
Régression multi-linéaire par moindres carrés partiels - PLS

→

Donnée de sortie Y
Valeur prédite de concentration pour l'élément d'intérêt

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Nk} \end{bmatrix}$$

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

- Spectres complet ou variables sélectionnées
- Avec/sans normalisation
- Matrice de données X

- A partir des spectres du lot de calibration, on calcule les composantes qui permettent de décrire le maximum de variance **de manière corrélée** aux valeurs de concentration.

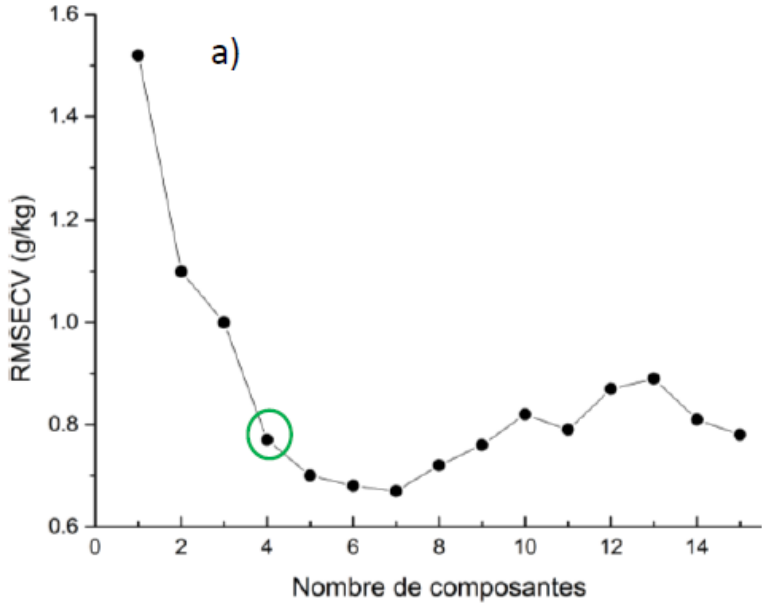
$$\begin{cases} X = t.P^T + E \\ y = t.q + f \end{cases}$$

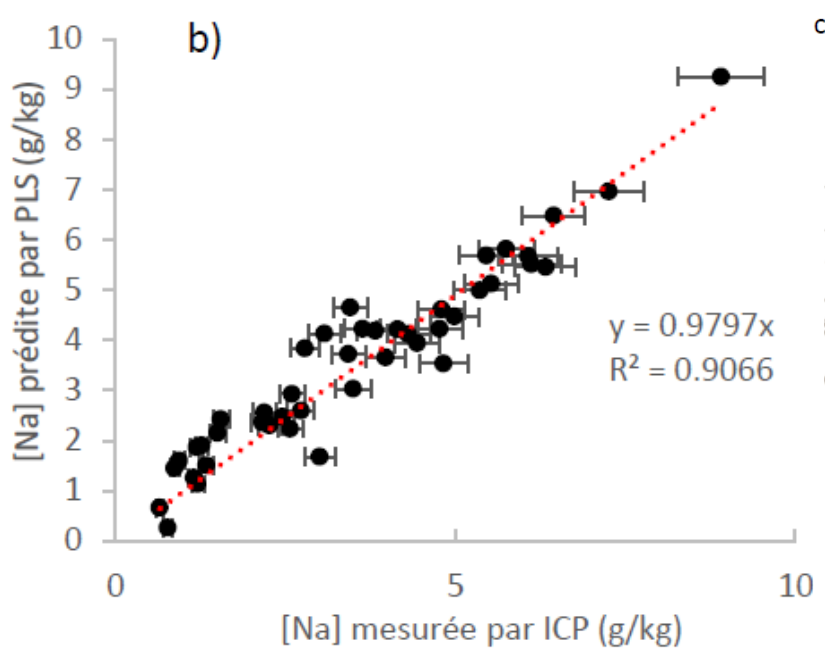
t: scores
P, q: loadings
E, f: erreurs que l'on cherche à minimiser

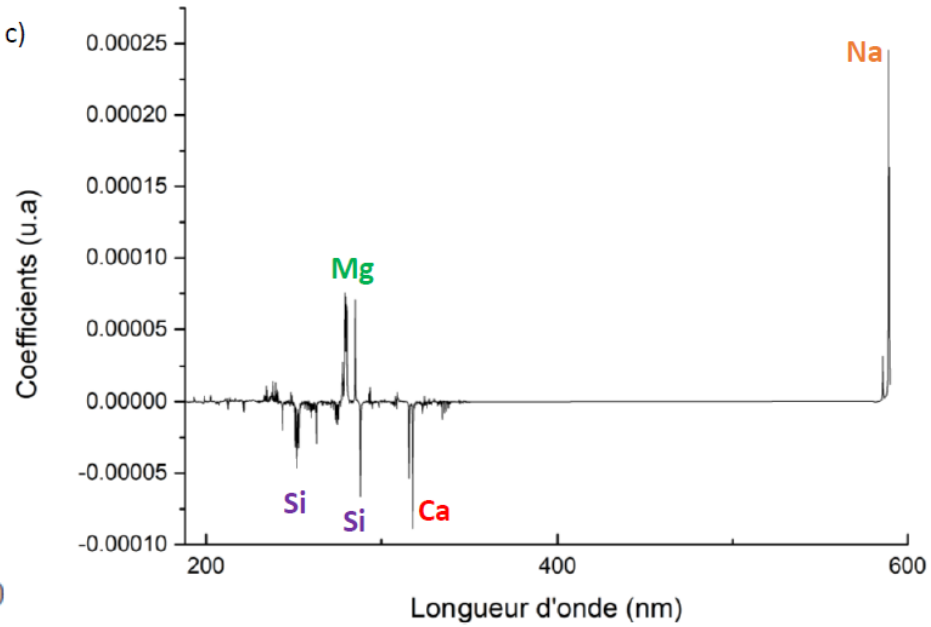
Le nombre de composantes optimum est obtenu par ***validation croisée***

L'analyse quantitative – calibration

Exemple : quantification par PLS-LIBS du sodium (Na) dans des sols agricoles







Root-mean square errors - RMSE

$$RMSE = \sqrt{\frac{\sum_{p=1}^{N}\left(c_p - \widehat{c_p}\right)^2}{N}}$$

c_p de référence

$\widehat{c_p}$ prédites

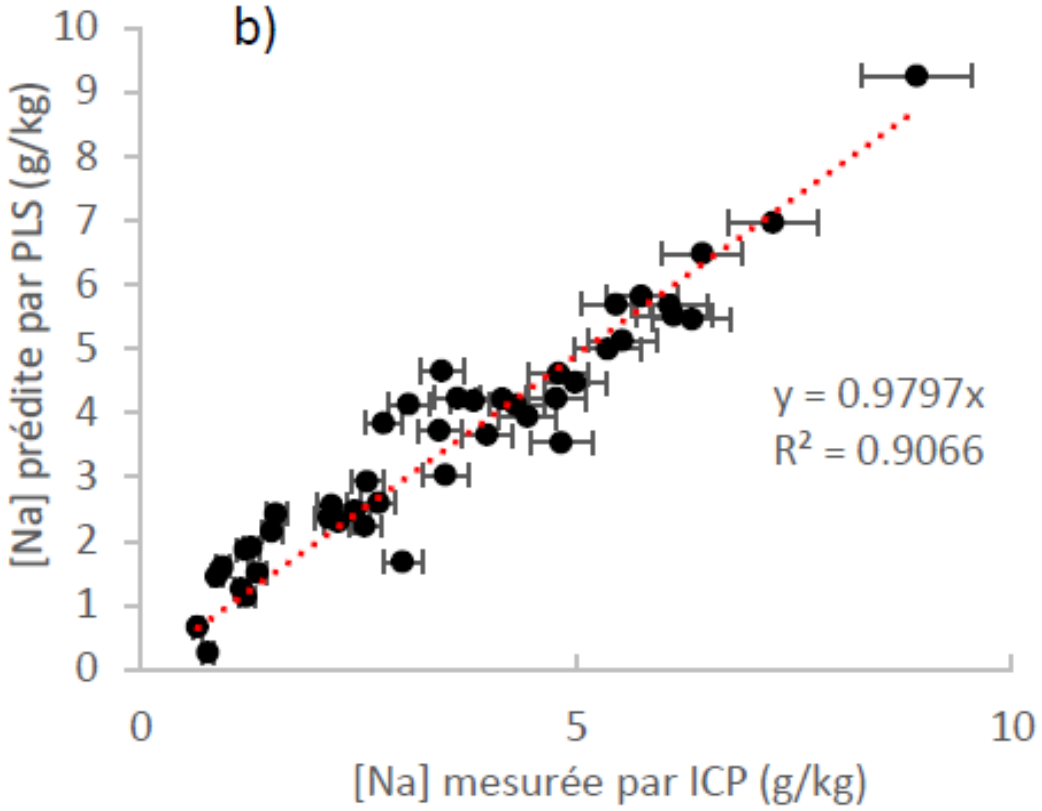
N # spectres

La quantification par PLS s'appuie sur les raies d'émission de :

- Na
- Mg (corrélation)
- Si et Ca (anti-corrélation)

L'analyse quantitative – validation

Exemple : quantification par PLS-LIBS du sodium (Na) dans des sols agricoles



La validation est une étape nécessaire pour considérer que le modèle PLS est adapté pour prédire les valeurs de concentration d'échantillons inconnus.

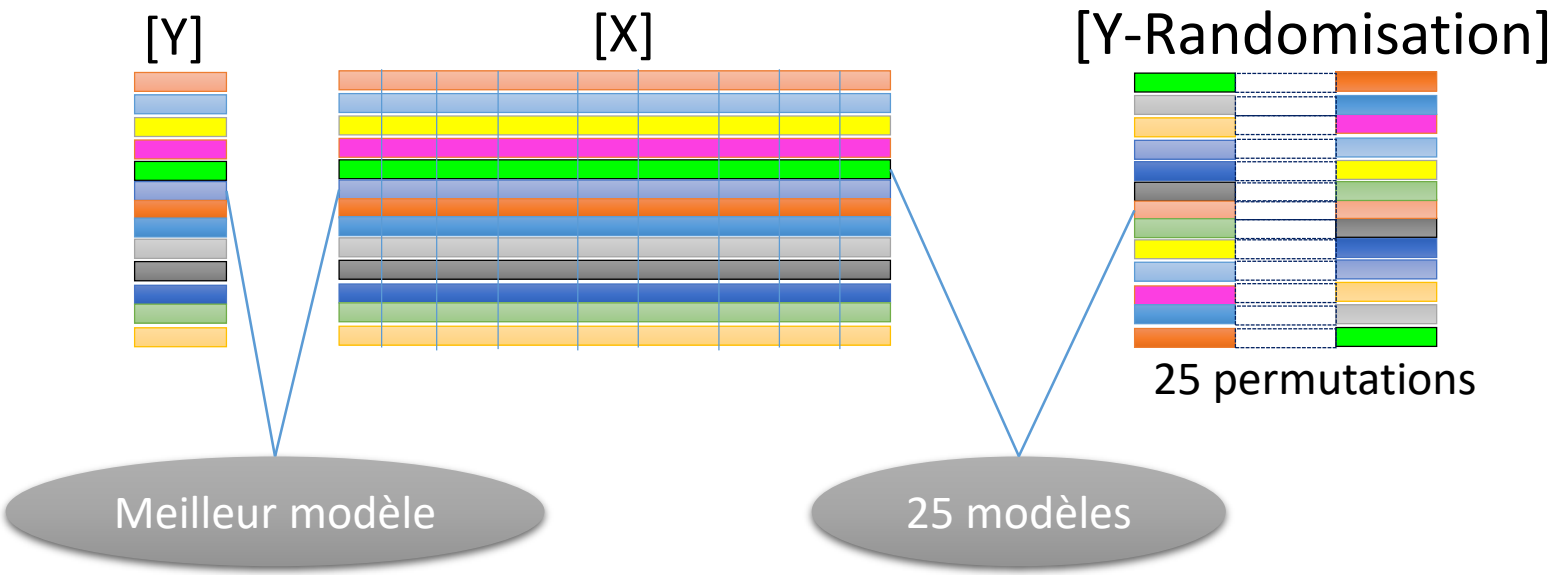
RMSECV = 0,76 g/kg

Pour 11 échantillons du lot de validation

[Na] g/kg	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11
Connue	7.57	1.23	7.90	3.19	5.67	6.27	6.30	5.60	5.91	6.81	3.78
Prédite	7.09	1.85	9.23	3.31	5.94	7.07	5.45	6.29	6.87	6.93	5.04
Incertitude	0.81	0.37	0.69	0.38	0.45	0.40	0.52	0.61	0.54	0.87	0.32

RMSEP = 0,79 g/kg

L'analyse quantitative – validation (suite)



Le meilleur modèle est considéré **robuste** lorsque :

- R^2 >> R^2 Random
- RMSE << RMSE Random

Mise en garde - Corrélation n'est pas causalité

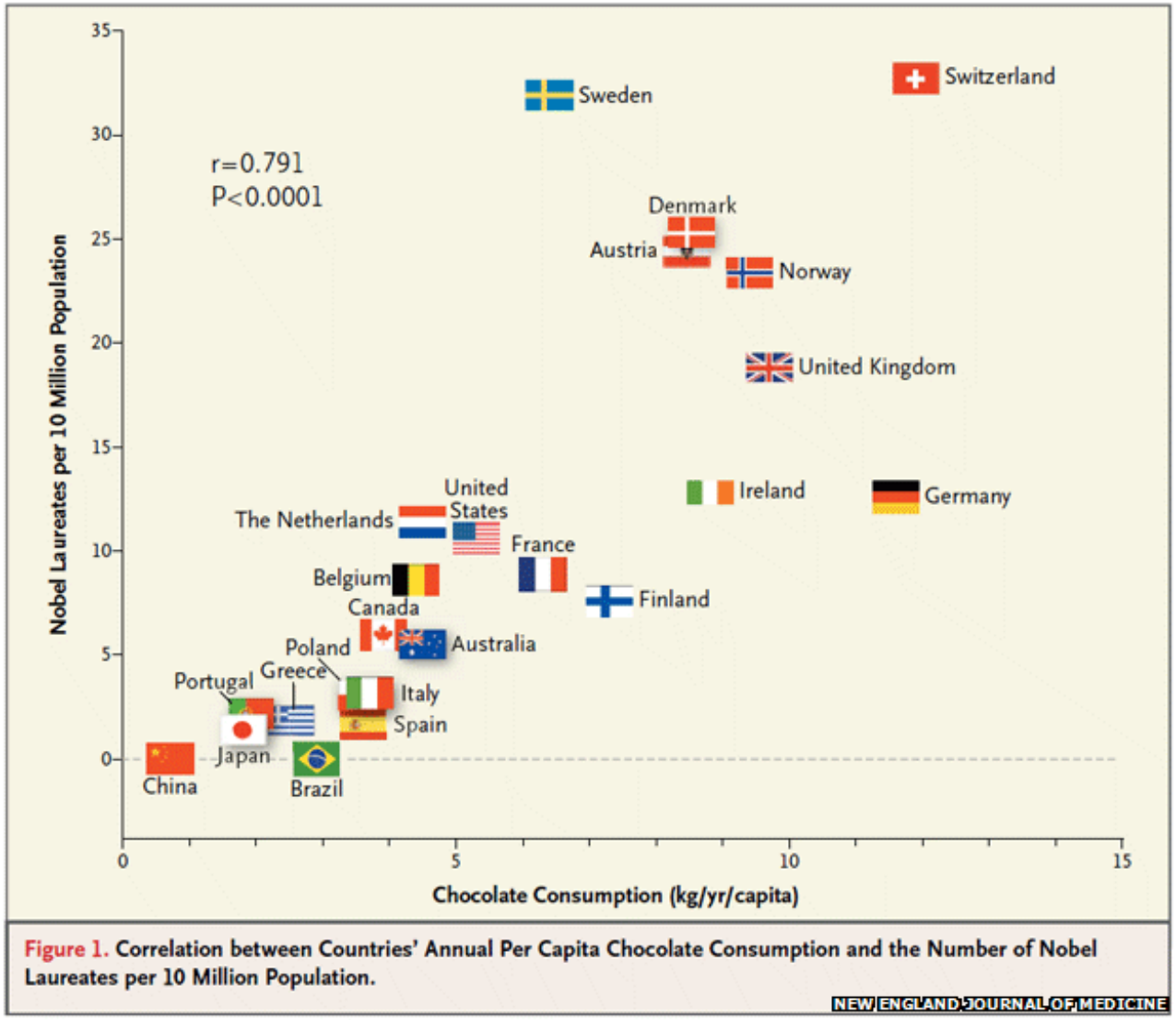


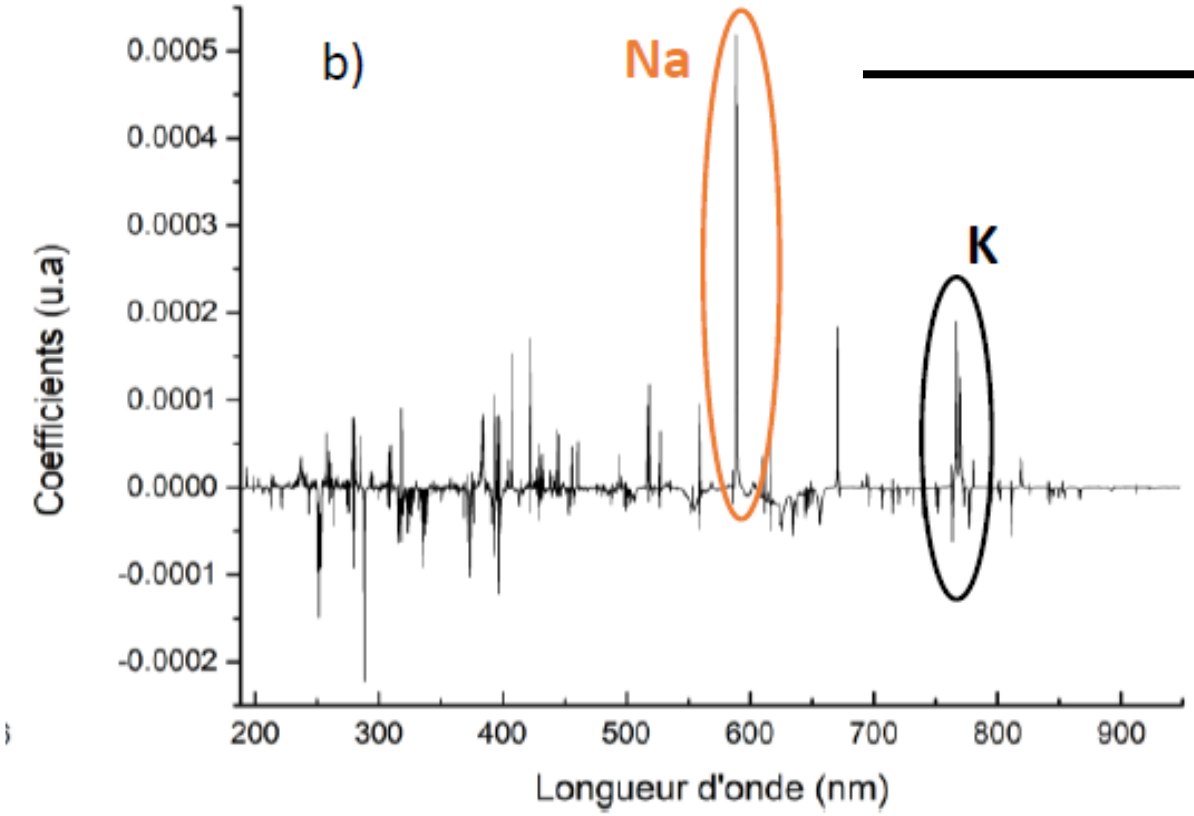
Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

NEW ENGLAND JOURNAL OF MEDICINE

D'après ce graphe, on a plus de chance d'obtenir le prix Nobel si on habite dans un pays qui a la plus grande consommation de chocolat !!!

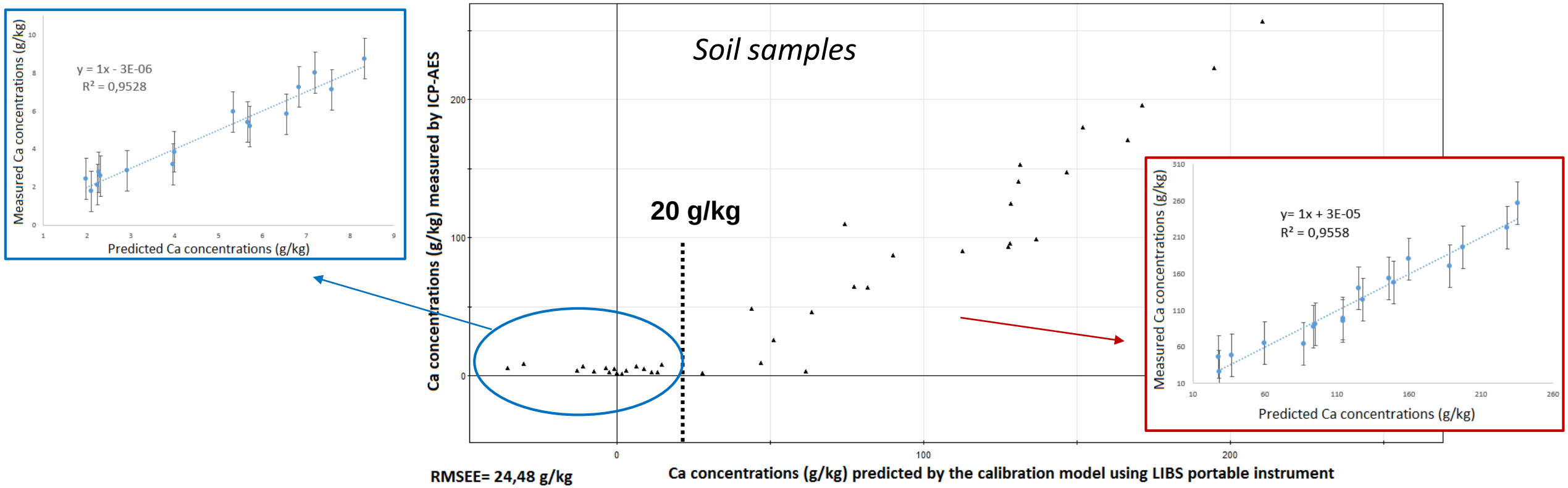
Mise en garde - Corrélation n'est pas causalité (suite)

Quantification par PLS-LIBS du potassium (K) dans des sols agricoles



Résultat inattendu, dû à une **corrélation entre K et Na**.

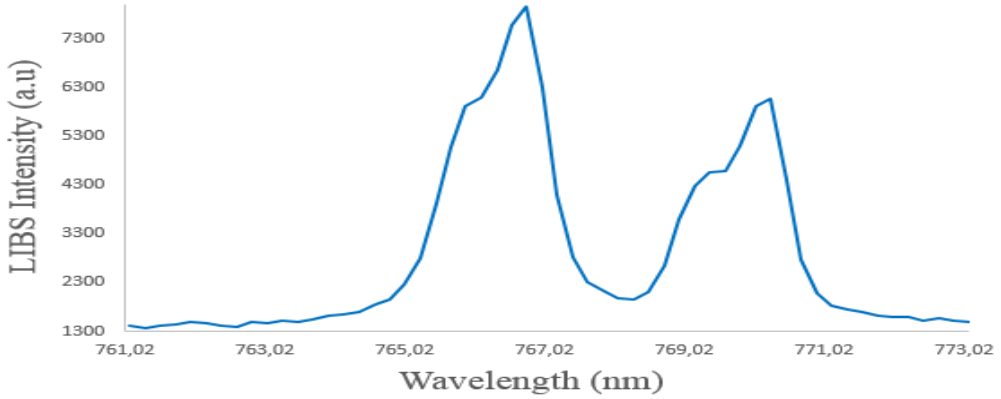
Mise en garde – large gamme de concentrations



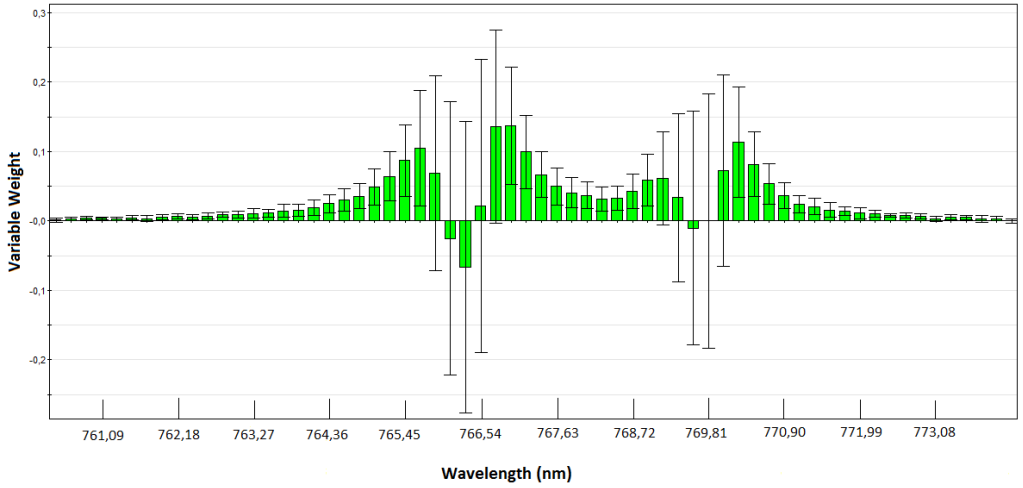
Ne pas hésiter à construire deux modèles (faibles /fortes concentration) si besoin !

Mise en garde – plasma opaque

Exemple : Potassium dans des échantillons de sol agricole



PLS



Le poids des variables dans le modèle PLS montre que l'intensité pour la variable au milieu de chaque raie d'émission n'est pas corrélée à la concentration !

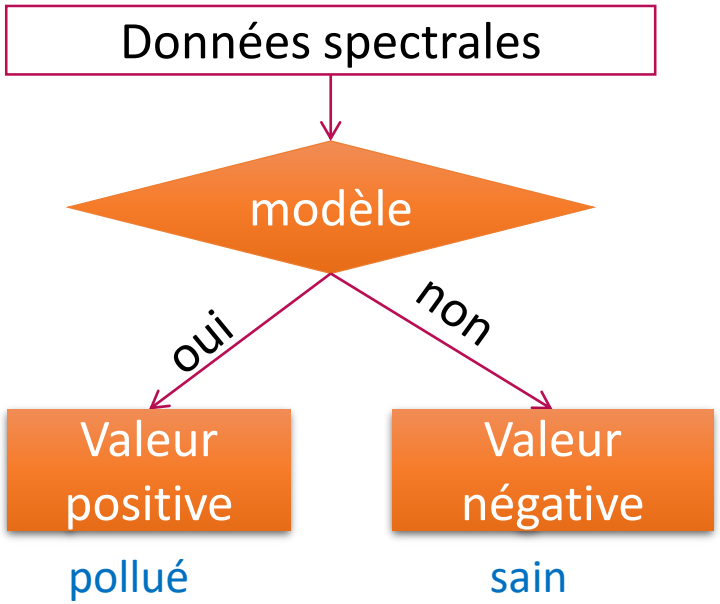
Sélection de variables

Le tri

Le principe de base du tri simple supervisé

Apprentissage à partir d'un lot d'échantillons connus

Exemple : l'échantillon est-il pollué ?



Données de référence

Données LIBS	Positive - pollué	Négative - sain
Positive pollué	**TP** Vrai positif	**FP** Faux positif
Négative sain	**FN** Faux négatif	**TN** Vrai négatif

sensibilité = TP/(TP+FN)	**spécificité** = TN/(FP+TN)

Précision (%) =100*(TP+TN)/(TP+FP+FN+TN)

Meilleur modèle : sensibilité et spécificité les plus élevées.

Le principe de base du tri multiple supervisé

Apprentissage à partir d'un lot d'échantillons connus

L'échantillon appartient-il à la classe C1, C2, C3…

Pour chaque classe Ci :

Données LIBS	Données de référence : Positive	Données de référence : Négative
Positive	**TP** Vrai positif	**FP** Faux positif
Négative	**FN** Faux négatif	**TN** Vrai négatif
	sensibilité = TP/(TP+FN)	**spécificité** = TN/(FP+TN)

Précision du modèle global

$$= \frac{\text{nombre d'échantillons correctement classés (toutes les classes)}}{\text{nombre total d'échantillons}}$$

Meilleur modèle : sensibilité et spécificité les plus élevées pour chaque classe
+ précision globale la plus élevée.

Le tri à partir des spectres LIBS

Variables d'entrée
extraites du spectre LIBS

→

Modèle de classification

→

Donnée de sortie
Valeur numérique c_j associée à la classe Cj

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Nk} \end{bmatrix}$$

Ex: PLS-**DA**
analyse discriminante

$$y = \begin{bmatrix} c_1 \\ \vdots \\ c_M \end{bmatrix}$$

M classes

Mêmes calculs que pour le modèle PLS !

Exemple : provenance d'échantillons d'obsidienne

Table 3 Confusion matrix for the training set

	BDT	LDV	LI	PA	SA	SB1	SB2	SC
BDT	18	0	0	0	0	0	0	0
LDV	0	18	0	0	0	0	0	0
LI	0	0	7	0	0	0	0	0
PA	0	0	0	23	0	0	0	0
SA	0	0	0	0	13	0	0	0
SB1	0	0	0	0	0	16	0	0
SB2	0	0	0	0	1	0	14	0
SC	0	0	0	0	0	0	0	45

Table 4 Confusion matrix for the validation set

	BDT	LDV	LI	PA	SA	SB1	SB2	SC
BDT	7	0	0	0	0	0	0	0
LDV	0	8	0	0	0	0	0	0
LI	0	0	4	0	0	0	0	0
PA	0	0	0	11	0	0	0	0
SA	0	0	0	0	6	0	1	0
SB1	0	0	0	0	0	7	0	1
SB2	0	0	0	0	0	2	5	0
SC	0	0	0	0	0	1	0	20

Bonnes pratiques

Bonnes pratiques

1. Comprendre les données et les questions analytiques

   ‘‘Garbage IN => Garbage OUT’’

2. Choisir le meilleur type de modèle supervisé
3. Préparer les données
4. Optimiser le modèle
5. Valider et tester le modèle

ELSEVIER

Spectrochimica Acta Part B: Atomic Spectroscopy

Volume 101, 1 November 2014, Pages 171-182

Review

Good practices in LIBS analysis: Review and advices

J. El Haddad, L. Canioni, B. Bousquet

Comprendre les données et les questions analytiques

- Quantification
- Tri
- Nombre de données de référence
- Etendue des valeurs de concentration
- Représentativité des classes
- Effets de matrice
- Conditions expérimentales
- Données aberrantes (outliers)
- Variables corrélées

Méthodes non-supervisées

PCA
Principal Component Analysis

MCR-ALS
Multivariate Curve Resolution-Alternating Least Squares

ICA
Independent Component Analysis

Kohonen Artificial Neural Network

Choisir le meilleur type de modèle supervisé

- Quantification

linéaires

- PLS
  Partial Least Squares
- PCR
  Principal Component Regression

non-linéaires

- ANN
  Artificial Neural Networks
- SVM (SVM-R, SVM-C)
  Support Vector Machine
- RF
  Random Forest

- Tri

linéaires

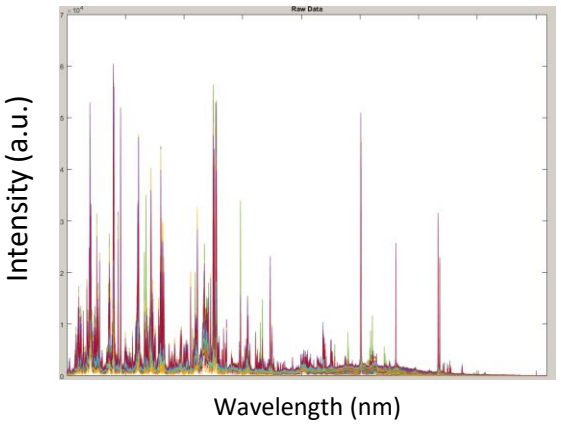
- SIMCA
  Soft Independent Modeling of Class Analogy
- PLS-DA
  Partial Least squares-Discriminant Analysis

non-linéaires

- ANN
  Artificial Neural Networks
- SVM (SVM-R, SVM-C)
  Support Vector Machine
- RF
  Random Forest

Préparer les données

Ligne de base ?
Bruit ?



LIBS intensity spectrum data						
		Wavelength (nm)				
		223.229	223.3727	---	994.3837	994.6131
Spectra	S 01	150.1	158.4	---	22.1	23.3
	S 02	164.8	174.2	---	21.5	21.1
	S 03	173.6	180.8	---	13.3	14.4
	S 04	230.9	221.4	---	25.1	26.9
	S 05	254.8	247.1	---	20.4	22.0
	S 06	235.0	226.9	---	21.4	22.4
	S 07	149.9	157.4	---	19.4	20.5
	S 08	163.1	171.7	---	15.8	16.5
	S 09	174.0	181.0	---	20.0	21.3
	S 10	247.0	244.0	---	25.4	26.1
	Sn	---	---	---	---	---

Scaling

Normalization

- Division par **l'aire totale** du spectre
- **SNV** (standard normal variate) : soustraction de la moyenne puis division par l'écart type des valeurs (ligne)

- **Données centrées** (mean centering) : on soustrait à chaque valeur d'intensité la valeur moyenne de la colonne
- **Données centrées réduites** (unit variance scaling) : on soustrait à chaque valeur d'intensité la valeur moyenne de la colonne puis on divise chaque nouvelle valeur par l'écart-type des valeurs de la colonne ; donne le même poids à chaque variable

Optimiser le modèle

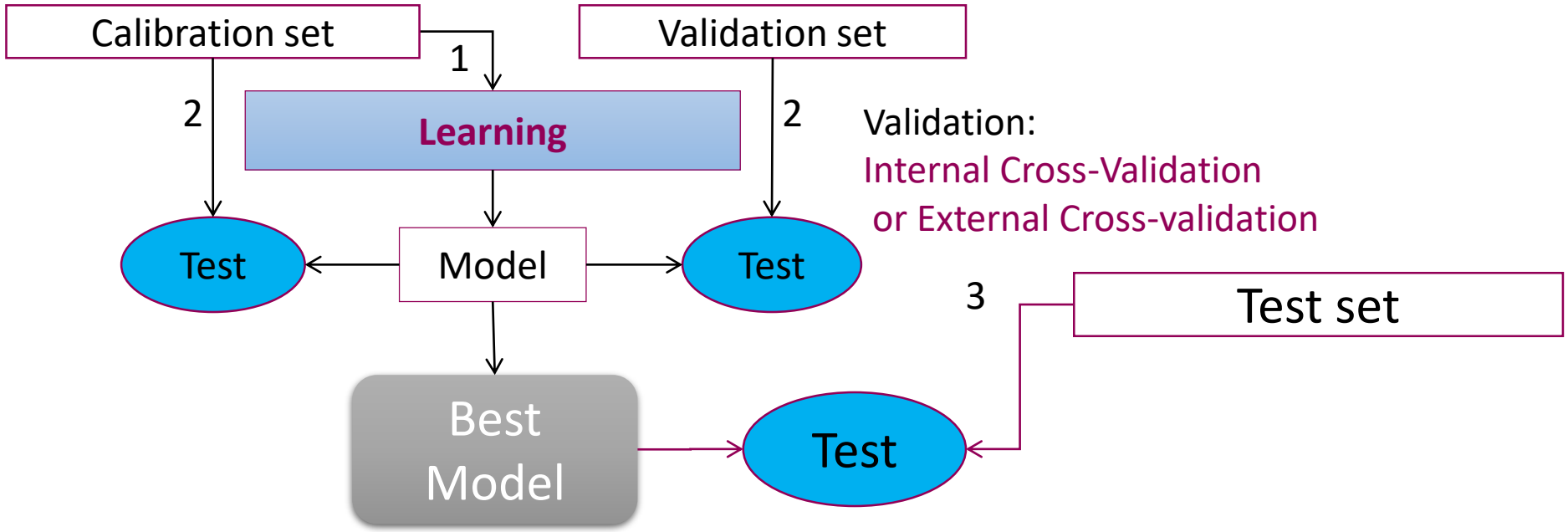
Réduire le nombre de variables d'entrée

- Réduit le bruit, le temps de calcul
- Évite le sur-apprentissage
- Guide l'apprentissage en se basant sur des informations physiques et non sur du bruit
- Nécessaire pour certains modèles (ex: ANN)

Déterminer les meilleurs paramètres du modèle

- Nombre de composantes
- Vitesse d'apprentissage
- Etc.

Valider et tester le modèle



Pour chaque test, on calcule les indicateurs de performance

Si les valeurs des indicateurs sont semblables pour le lot de calibration et de validation
$\Rightarrow$ Le modèle ne souffre pas de sur-apprentissage
Si les valeurs des indicateurs sont semblables pour le lot de test et de validation
=> Le modèle est généralisable à des échantillons inconnus

Questions ?

Fin !

bruno.bousquet@u-bordeaux.fr