Apprentissage, Classification & Prévision

David Nerini

MIO

premier semestre

Motivation

- La description analytique de tous les processus existants et qui interagissent dans le milieu naturel est une tâche impossible. La construction de modèles mathématiques cohérents, à but explicatif, doit être effectuée sous des hypothèses souvent éloignées de la réalité. Cependant, leur élaboration est indispensable puisqu'ils doivent permettre de mieux *comprendre* et d'*expliquer* le fonctionnement partiel des écosystèmes.

Motivation

- La description analytique de tous les processus existants et qui interagissent dans le milieu naturel est une tâche impossible. La construction de modèles mathématiques cohérents, à but explicatif, doit être effectuée sous des hypothèses souvent éloignées de la réalité. Cependant, leur élaboration est indispensable puisqu'ils doivent permettre de mieux *comprendre* et d'*expliquer* le fonctionnement partiel des écosystèmes.
- Une des grandes tâches en statistiques consiste à *décrire* le fonctionnement d'un système à partir de l'observation de variables échantillonnées sur un grand nombre d'individus. L'objectif n'est pas d'étudier les processus qui interagissent dans le système mais de *prévoir* ou d'*anticiper* des modifications de *variables cibles* liées de manière complexe avec des *variables exogènes*.

Motivation

- La description analytique de tous les processus existants et qui interagissent dans le milieu naturel est une tâche impossible. La construction de modèles mathématiques cohérents, à but explicatif, doit être effectuée sous des hypothèses souvent éloignées de la réalité. Cependant, leur élaboration est indispensable puisqu'ils doivent permettre de mieux *comprendre* et d'*expliquer* le fonctionnement partiel des écosystèmes.
- Une des grandes tâches en statistiques consiste à *décrire* le fonctionnement d'un système à partir de l'observation de variables échantillonnées sur un grand nombre d'individus. L'objectif n'est pas d'étudier les processus qui interagissent dans le système mais de *prévoir* ou d'*anticiper* des modifications de *variables cibles* liées de manière complexe avec des *variables exogènes*.
-

$$expliquer \neq prédire$$

Position du problème

- On dispose d'un échantillon $E = \{(y_i, \mathbf{x}_i), i = 1, ..., n\}$ d'une *variable à prédire* Y et de *variables de prévisions* ou *explicatives* $\mathbf{X} = (X_1, ..., X_p)'$. Cet échantillon doit être représentatif de la population dans laquelle il a été prélevé. Cela veut dire que si sa taille n tend vers l'infini, les caractéristiques des variables échantillonnées se rapprochent de celles de leur distribution théorique.

Position du problème

- On dispose d'un échantillon $E = \{(y_i, \mathbf{x}_i), i = 1, ..., n\}$ d'une *variable à prédire* Y et de *variables de prévisions* ou *explicatives* $\mathbf{X} = (X_1, ..., X_p)'$. Cet échantillon doit être représentatif de la population dans laquelle il a été prélevé. Cela veut dire que si sa taille n tend vers l'infini, les caractéristiques des variables échantillonnées se rapprochent de celles de leur distribution théorique.
- On cherche à construire un *modèle statistique* de la forme

$$y = f_\theta(\mathbf{x}) + \varepsilon$$

où f_θ est une fonction qui décrit les relations liant la variable Y aux variables $X_1, ..., X_p$. Sa forme peut appartenir à une famille paramétrique ou ne pas être connue. La détermination de f_θ, directement ou par l'intermédiaire des paramètres θ, est effectuée, la plupart du temps, en minimisant la norme des erreurs ε.

Position du problème

- On dispose d'un échantillon $E = \{(y_i, \mathbf{x}_i), i = 1, ..., n\}$ d'une *variable à prédire* Y et de *variables de prévisions* ou *explicatives* $\mathbf{X} = (X_1, ..., X_p)'$. Cet échantillon doit être représentatif de la population dans laquelle il a été prélevé. Cela veut dire que si sa taille n tend vers l'infini, les caractéristiques des variables échantillonnées se rapprochent de celles de leur distribution théorique.
- On cherche à construire un *modèle statistique* de la forme

$$y = f_\theta(\mathbf{x}) + \varepsilon$$

où f_θ est une fonction qui décrit les relations liant la variable Y aux variables $X_1, ..., X_p$. Sa forme peut appartenir à une famille paramétrique ou ne pas être connue. La détermination de f_θ, directement ou par l'intermédiaire des paramètres θ, est effectuée, la plupart du temps, en minimisant la norme des erreurs ε.

- Une fois le modèle construit, il est possible d'estimer sa fiabilité en utilisant des données qui n'ont pas servi à sa construire dans le but de faire de la *prévision.*

Position du problème

Definition

On appelle Apprentissage Machine (*Machine Learning*) l'ensemble des méthodes statistiques relatives à l'étude des propriétés de E, de l'estimation de f_θ et de son utilisation pour effectuer des prévisions.

- Les techniques mathématiques utilisées pour la construction des modèles et la prévision utilisent énormément de données.

Position du problème

Definition

On appelle Apprentissage Machine (*Machine Learning*) l'ensemble des méthodes statistiques relatives à l'étude des propriétés de E, de l'estimation de f_θ et de son utilisation pour effectuer des prévisions.

- Les techniques mathématiques utilisées pour la construction des modèles et la prévision utilisent énormément de données.
- L'étude de leurs propriétés et leur utilisation dans un cadre opérationnel sont souvent basées sur des algorithmes informatiques particuliers (*Machine Learning*)

Position du problème

Definition

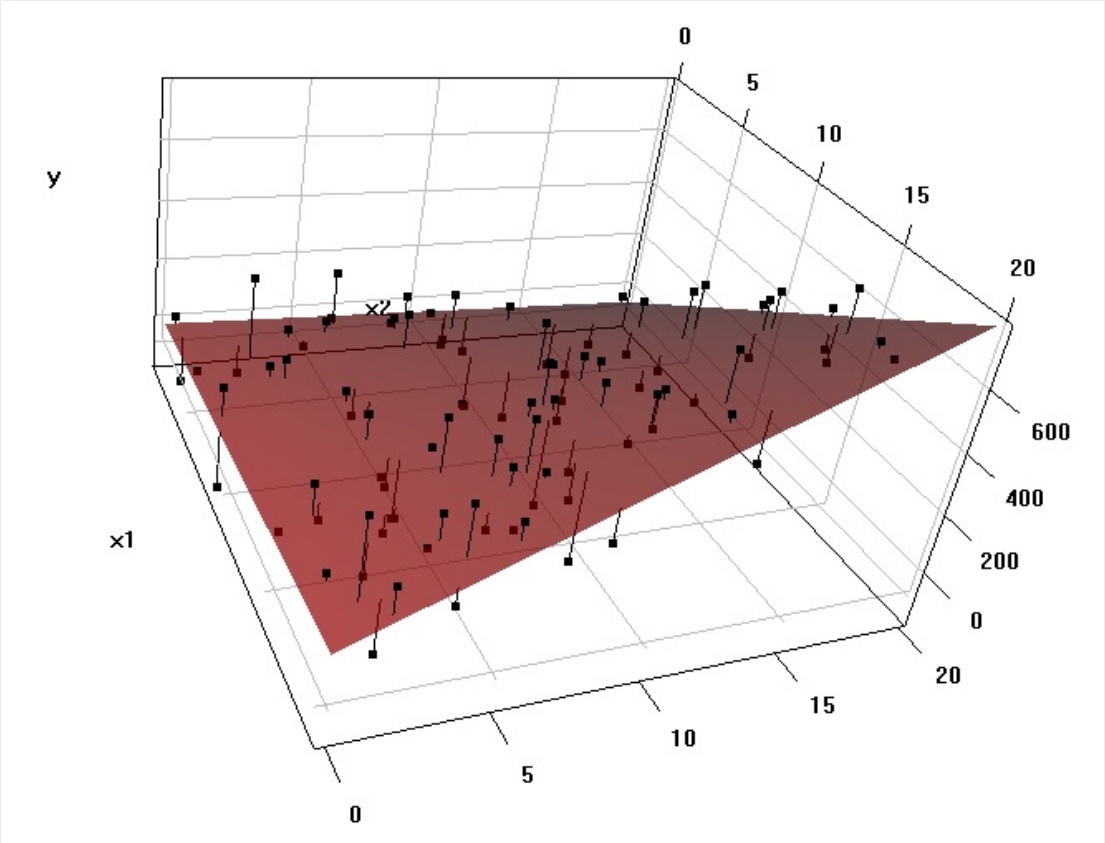
On appelle Apprentissage Machine (*Machine Learning*) l'ensemble des méthodes statistiques relatives à l'étude des propriétés de E, de l'estimation de f_θ et de son utilisation pour effectuer des prévisions.

- Les techniques mathématiques utilisées pour la construction des modèles et la prévision utilisent énormément de données.
- L'étude de leurs propriétés et leur utilisation dans un cadre opérationnel sont souvent basées sur des algorithmes informatiques particuliers (*Machine Learning*)
- Le choix des méthodes repose essentiellement sur le type de données avec lesquelles l'écologue travaille.

Exemples : Y et X réels

Example (Régression linéaire multiple)

$$y = k + ax_1 + bx_2 + cx_1x_2 + \varepsilon$$

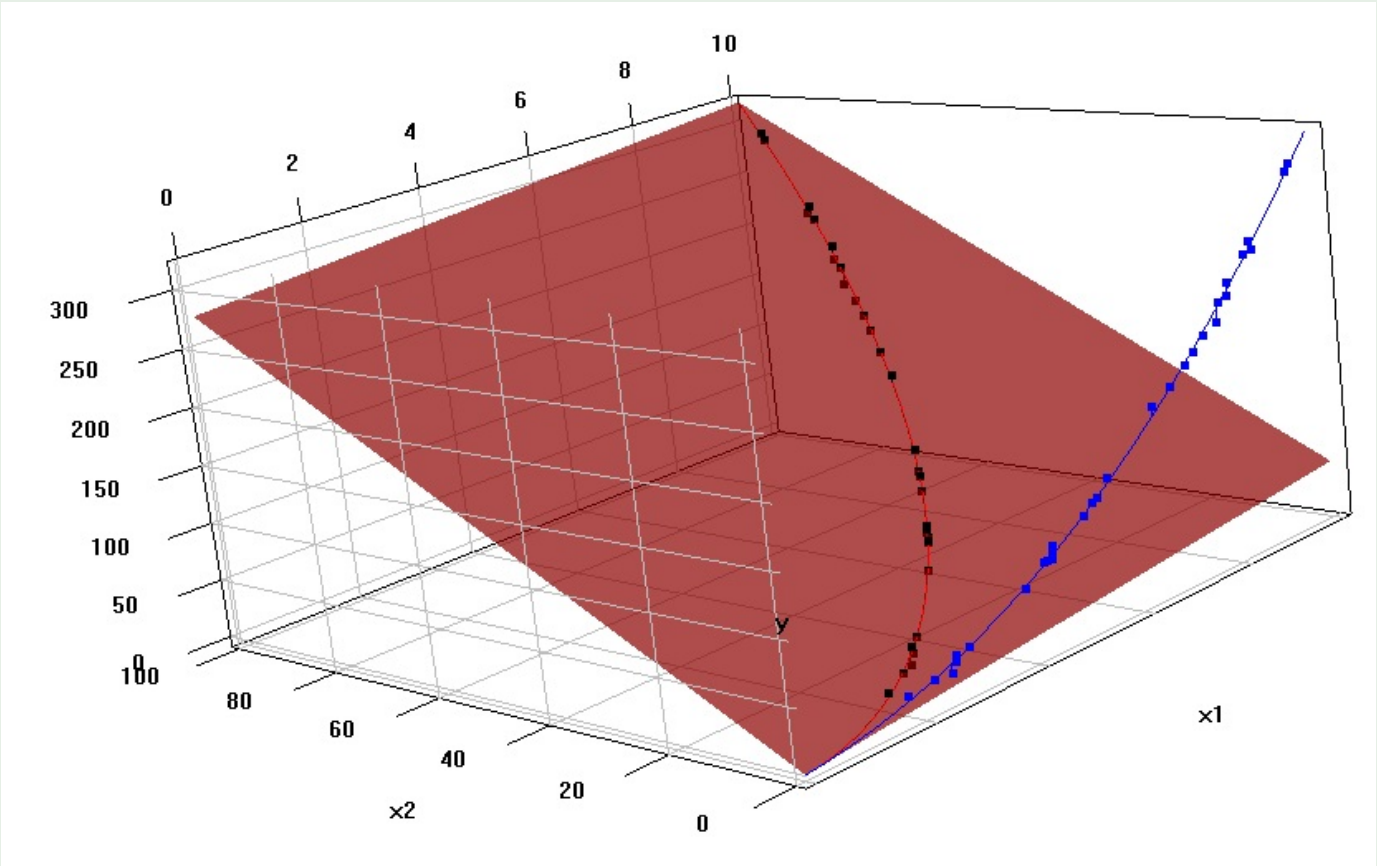


- On cherche à *estimer* un modèle à partir de données, sous-entendu qu'un *modèle théorique* est supposé existant.

Exemples : Y et X réels

Example (Régression polynômiale)

$$y = k + ax + bx^2 + \varepsilon$$

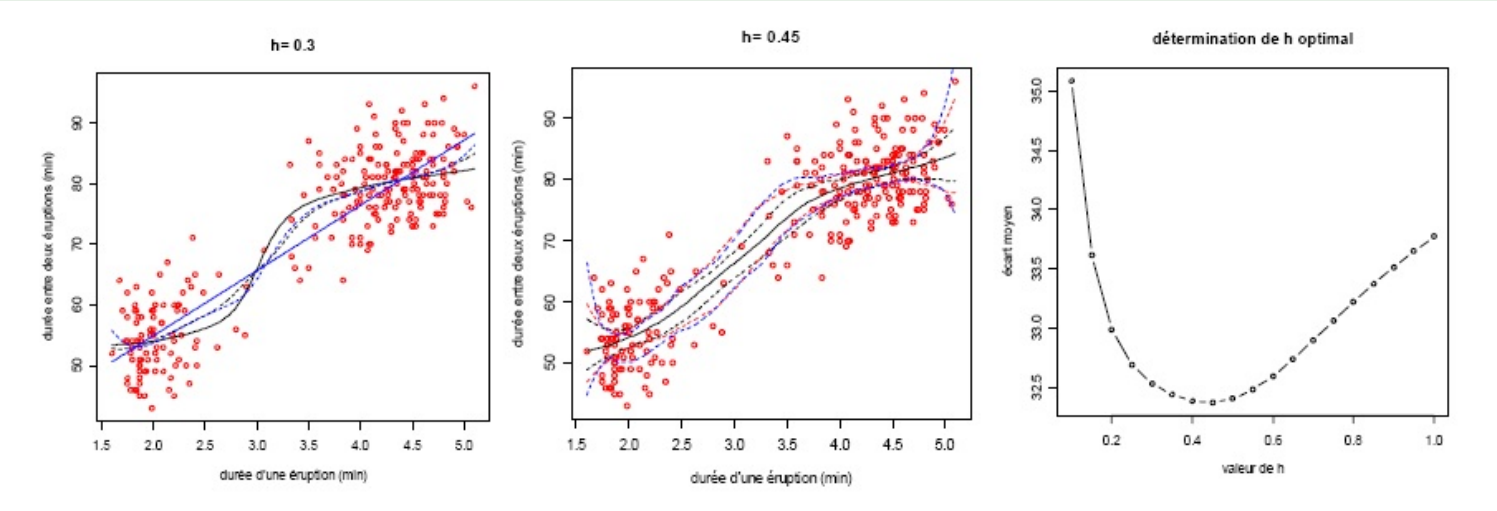


- La forme du modèle dépend des caractéristiques des variables étudiées même si la problématique est identique au cas précédent.

Exemples : Y et X réels

Example (Régression non-paramétrique)

$$y = f(x) + \varepsilon$$

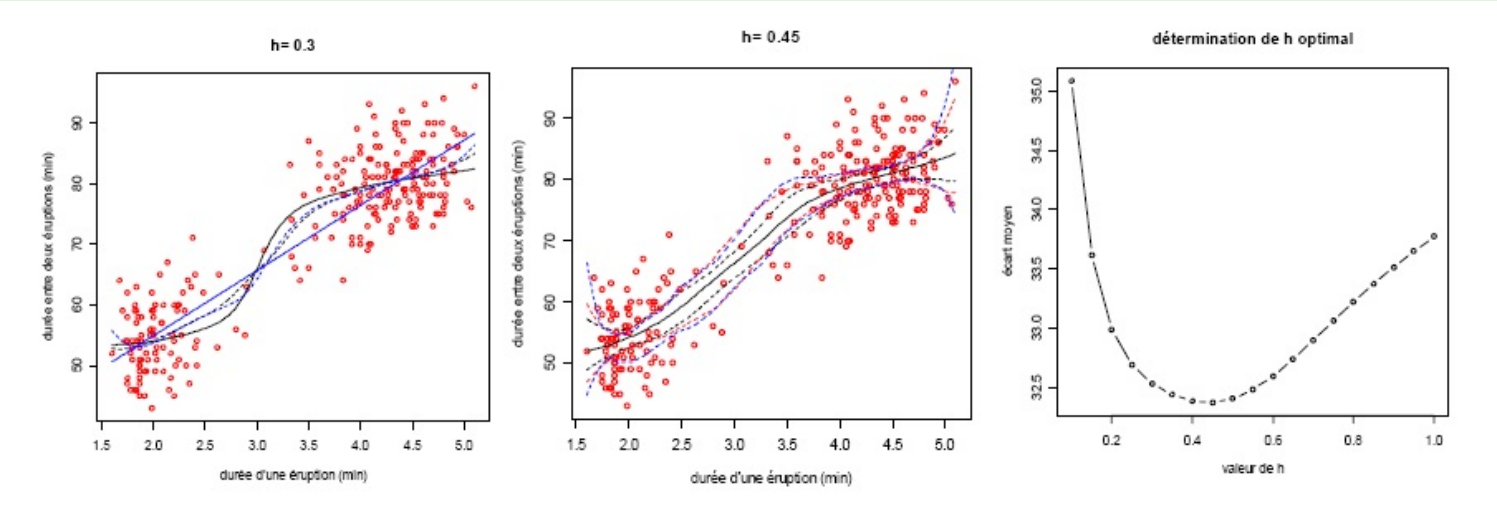


- L'estimation de la fonction est directement réalisée à l'aide des données sans passer par une *estimation de paramètres*

Exemples : Y et X réels

Example (Régression non-paramétrique)

$$y = f(x) + \varepsilon$$



- L'estimation de la fonction est directement réalisée à l'aide des données sans passer par une *estimation de paramètres*
- La structure du modèle est complètement dépendante de la qualité de l'échantillon

Estimation des MC

- Dans les trois cas précédents, l'*estimation* de f_θ est effectuée par *moindres carrés* à l'aide de l'*échantillon d'apprentissage* E. On cherche en fait à minimiser la norme des erreurs ε

$$\|\varepsilon\|^2 = RSS\left(f_\theta\right) = \sum_{i=1}^{n} K\left(x_i\right)\left(y_i - f_\theta\left(x_i\right)\right)^2$$

Estimation des MC

- Dans les trois cas précédents, l'*estimation* de f_θ est effectuée par *moindres carrés* à l'aide de l'*échantillon d'apprentissage* E. On cherche en fait à minimiser la norme des erreurs ε

$$\|\varepsilon\|^2 = RSS\left(f_\theta\right) = \sum_{i=1}^{n} K\left(x_i\right)\left(y_i - f_\theta\left(x_i\right)\right)^2$$

- La forme de la fonction poids $K\left(x_i\right)$ dépend de la méthode employée

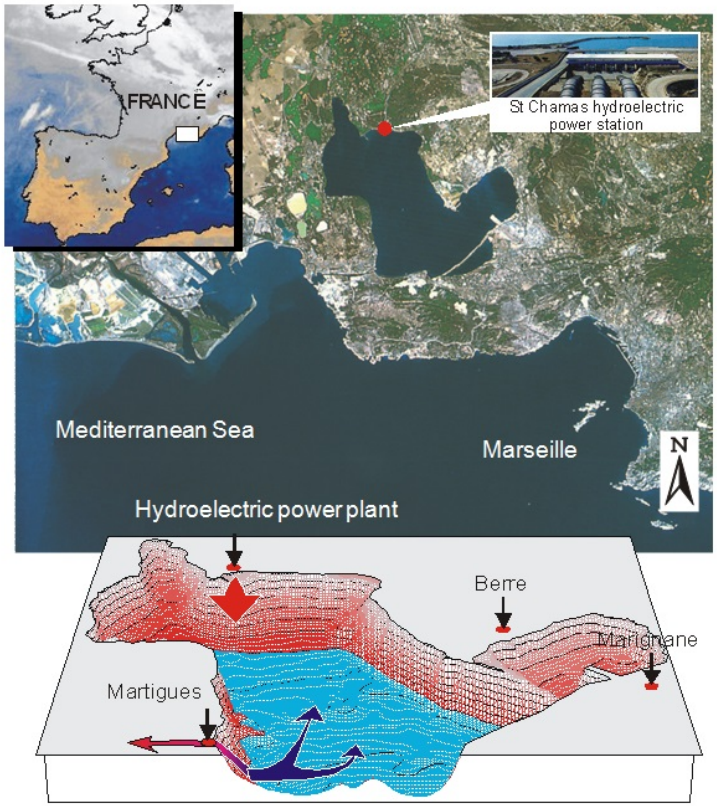
Estimation des MC

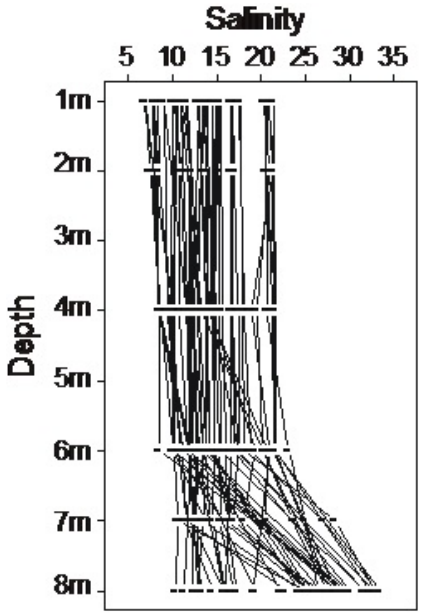
- Dans les trois cas précédents, l'*estimation* de f_θ est effectuée par *moindres carrés* à l'aide de l'*échantillon d'apprentissage* E. On cherche en fait à minimiser la norme des erreurs ε

$$\|\varepsilon\|^2 = RSS\left(f_\theta\right) = \sum_{i=1}^{n} K\left(x_i\right)\left(y_i - f_\theta\left(x_i\right)\right)^2$$

- La forme de la fonction poids $K\left(x_i\right)$ dépend de la méthode employée
- La méthode des MC est à la base de la construction de la plupart des modèles

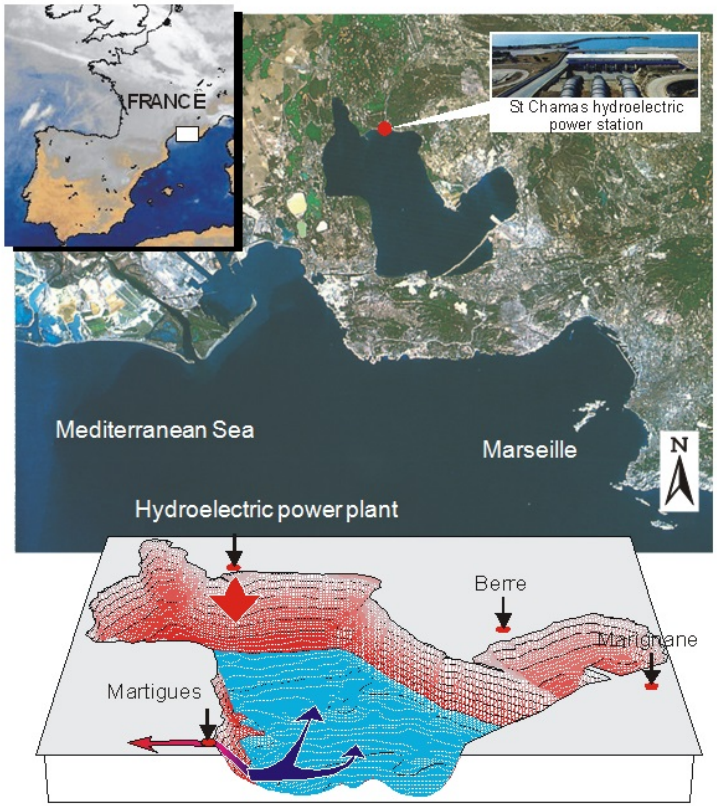
Exemples : Y catégorielle et X quelconques

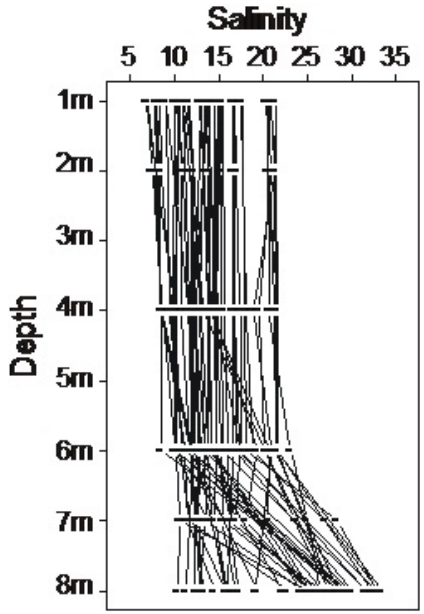




- On construit une *variable qualitative* $Y \in \{e_1, ..., e_m\}$ qui décrit les états hydrologiques de l'étang

Exemples : Y catégorielle et X quelconques

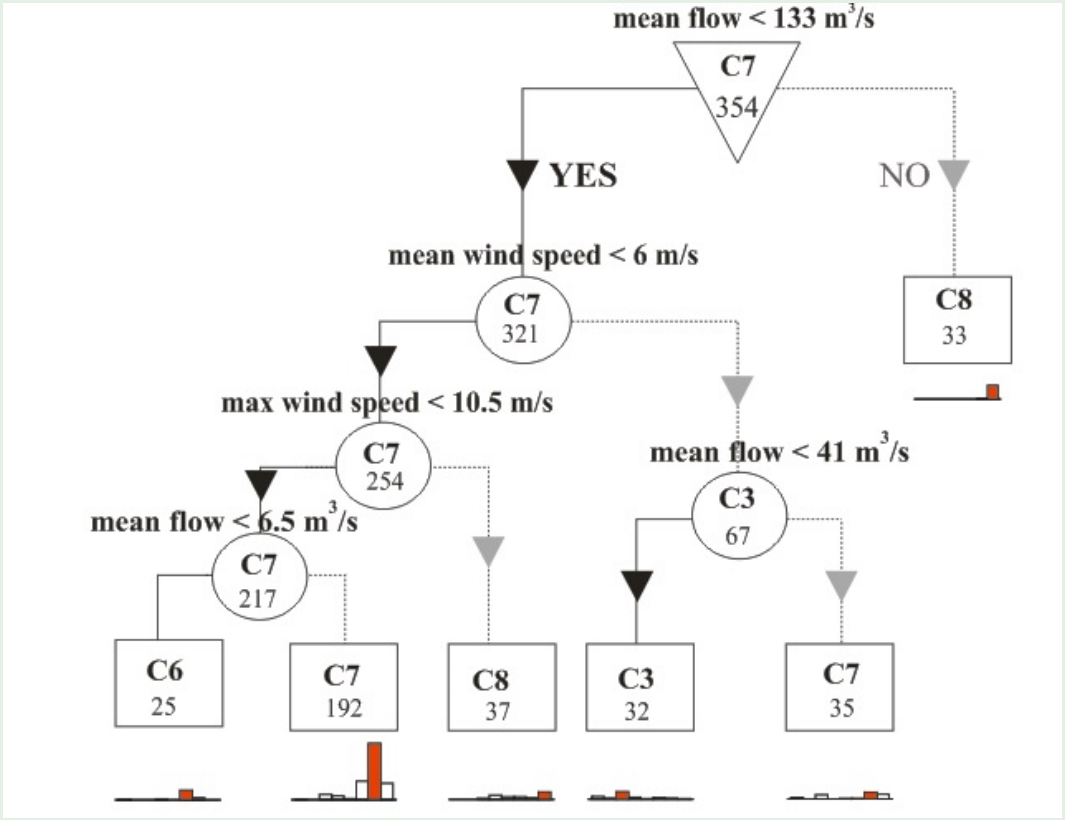




- On construit une *variable qualitative* $Y \in \{e_1, ..., e_m\}$ qui décrit les états hydrologiques de l'étang
- Prévision des conditions hydrologiques qualitatives en fonction d'un forçage externe (vent et débits)

Exemples : Y catégorielle et X quelconques

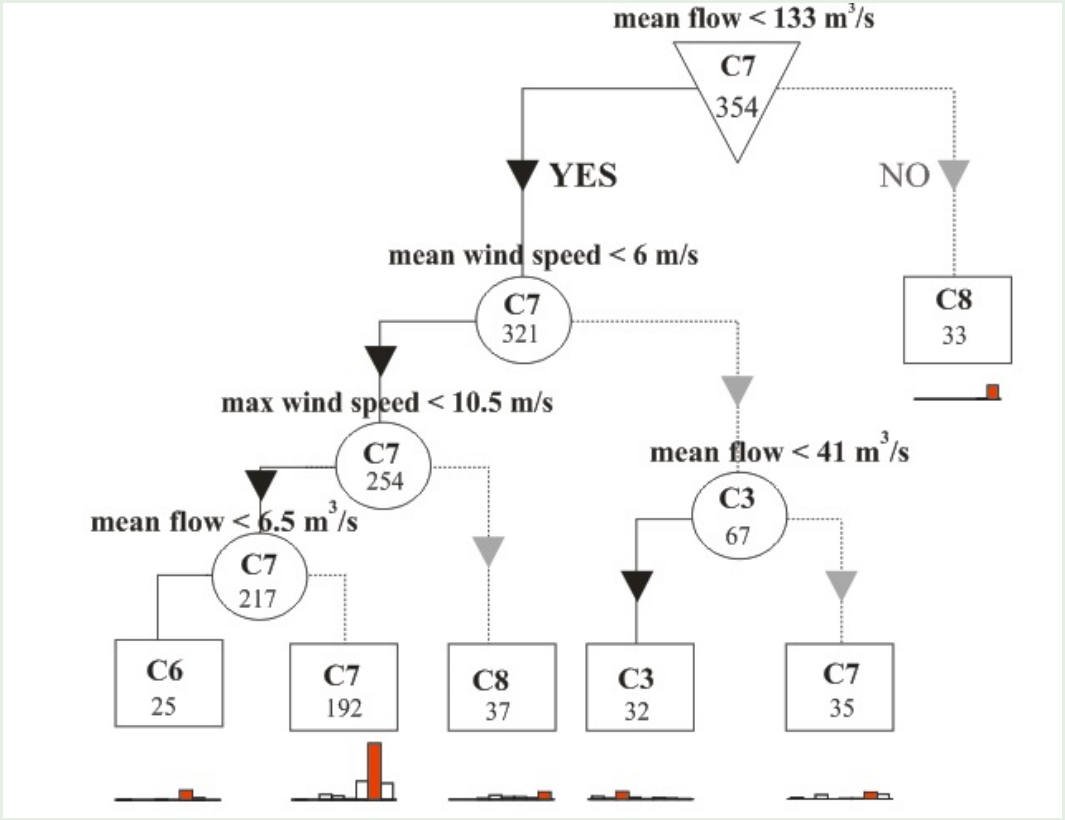
Example (Arbre de classification)



- $y = f_L(\mathbf{x}) = \sum_{j=1}^{q} c_j I(\mathbf{X} \in r_j)$

Exemples : Y catégorielle et X quelconques

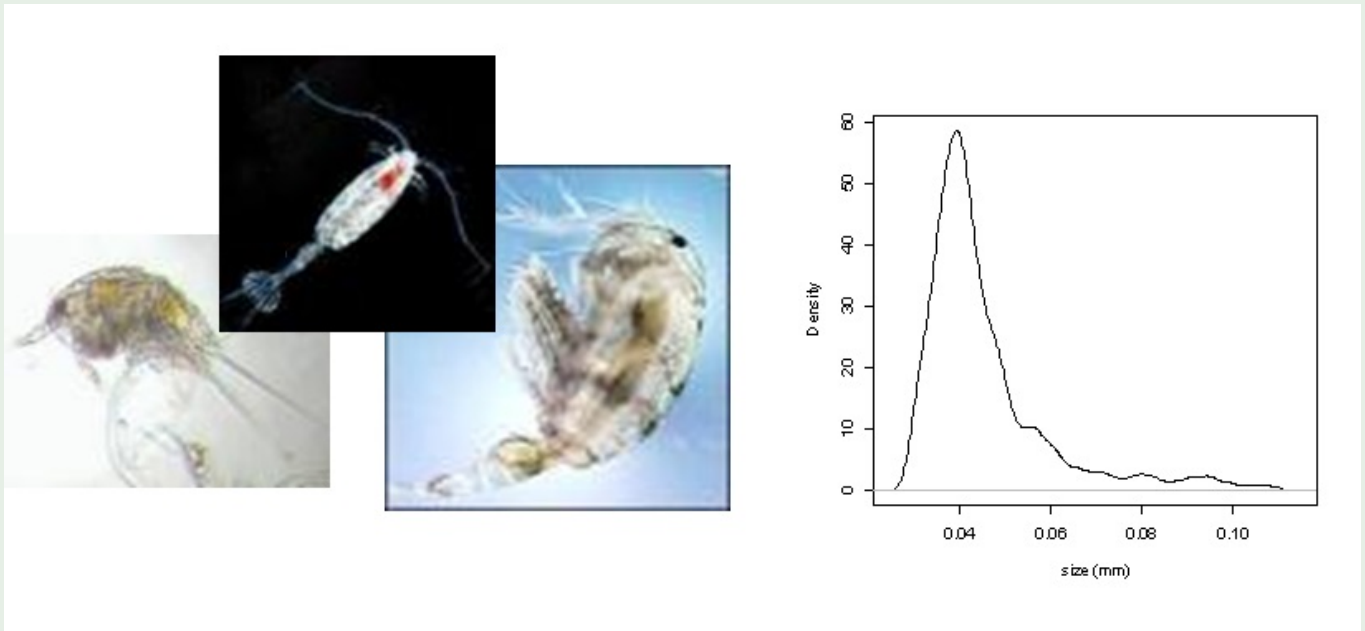
Example (Arbre de classification)



- $y = f_L(\mathbf{x}) = \sum_{j=1}^{q} c_j I(\mathbf{X} \in r_j)$
- La classe c_j élue est celle qui a la probabilité maximale dans un noeud terminal donné

Exemples : Y fonctionnelle et X quelconques

Example (Spectres de taille)



- Prévision du spectre de taille du zooplancton en fonction de variables environnementales

Exemples : Y fonctionnelle et X quelconques

Example (Arbre de régression)

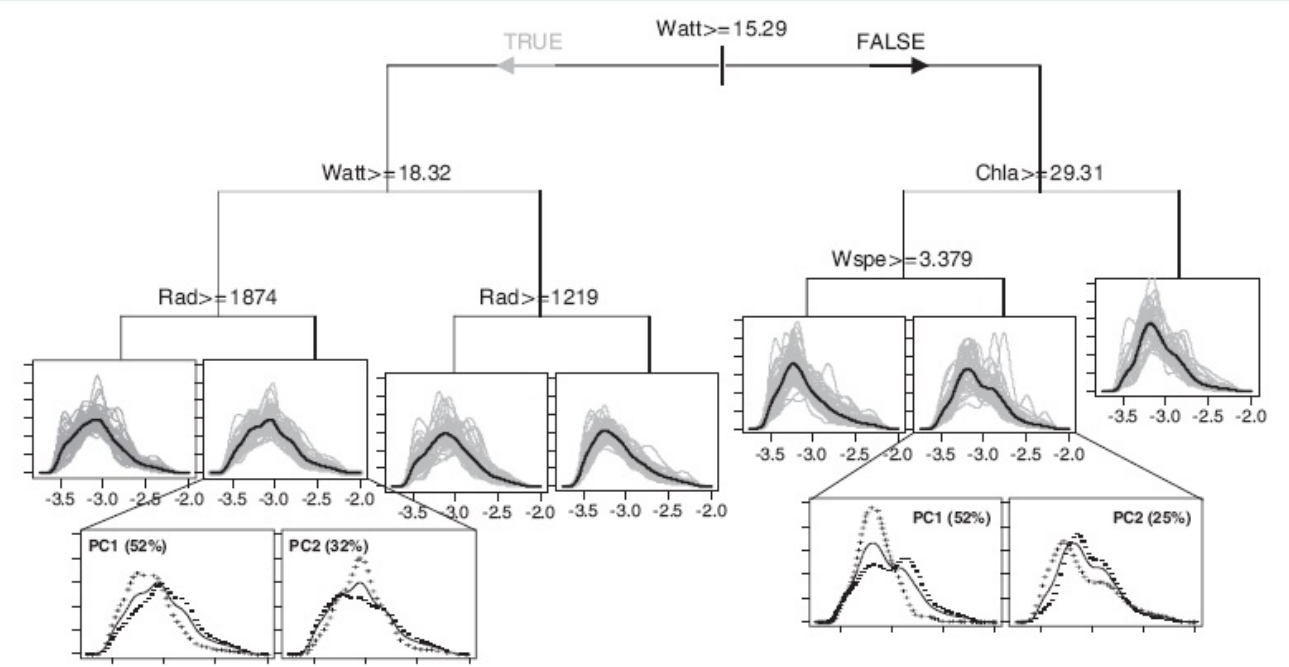


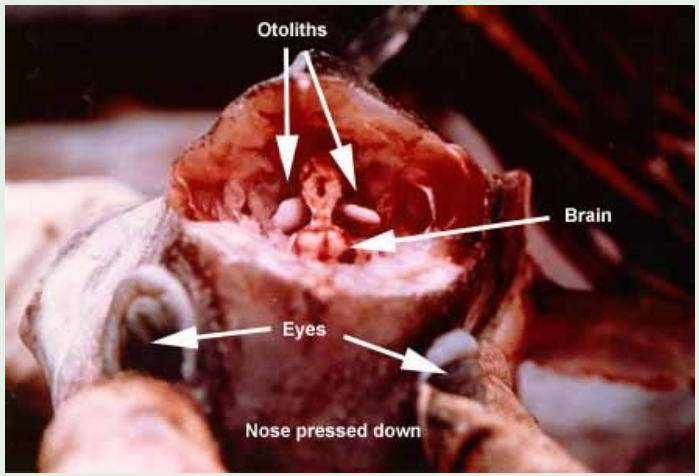
Fig. 6. A Kullback–Leibler regression tree constructed over 395 observations of size density. Only four predictive variables appeared as decision rules. A terminal node contains the observations (grey lines) and the predicted density curve (black line). An example of local PCA is displayed for two terminal nodes.

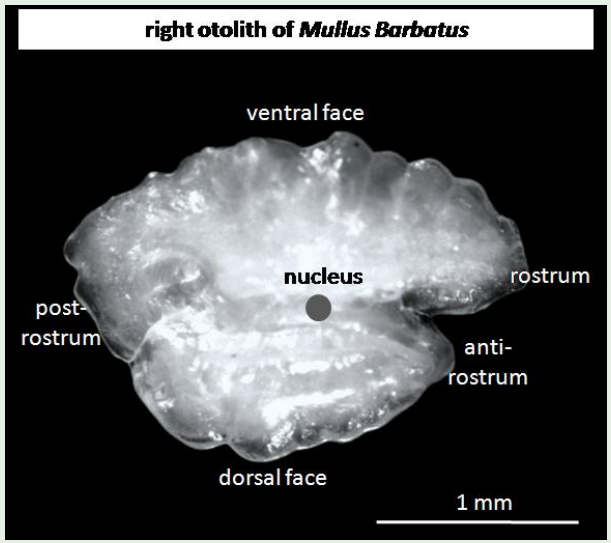
- le modèle est de la forme

$$y = f_L(\mathbf{x}) = E(Y/\mathbf{X} = \mathbf{x}) = \sum_{j=1}^{q} f_j I(\mathbf{X} \in r_j)$$

Exemples : Y catégorielle et X fonctionnelle bivariée

Example (location and shape)

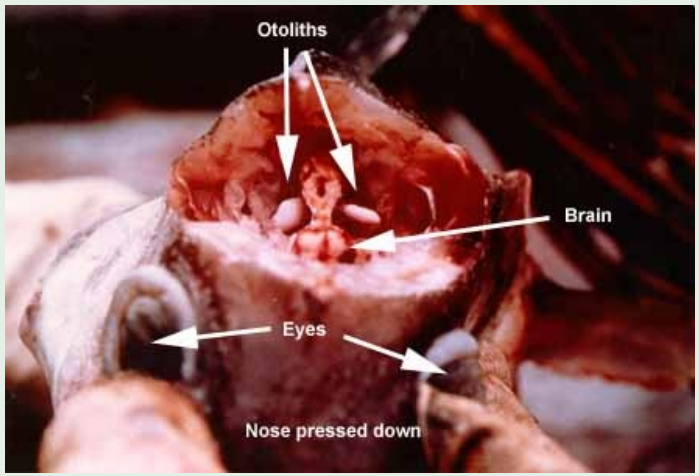


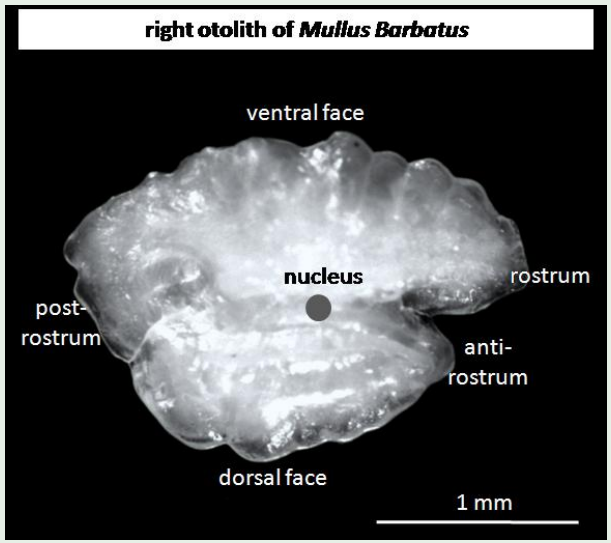


- a small piece of calcium carbonate located in the head of *Teleostean* fishes

Exemples : Y catégorielle et X fonctionnelle bivariée

Example (location and shape)

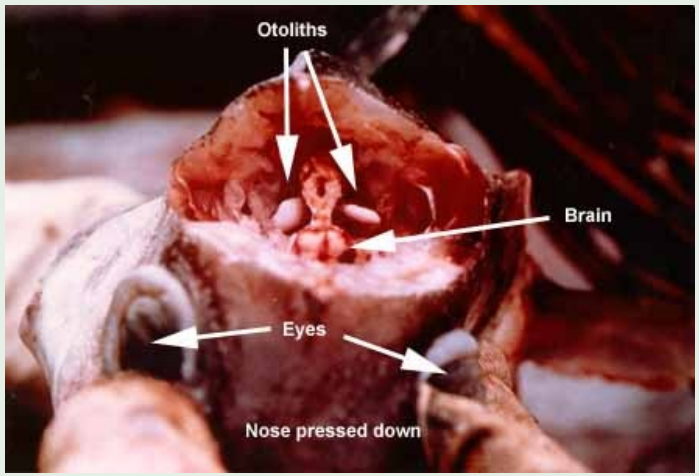


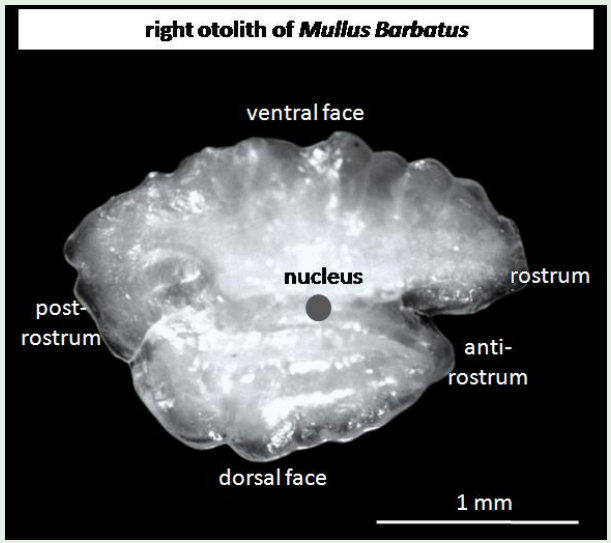


- a small piece of calcium carbonate located in the head of *Teleostean* fishes
  - acceleration perception and sense of balance

Exemples : Y catégorielle et X fonctionnelle bivariée

Example (location and shape)

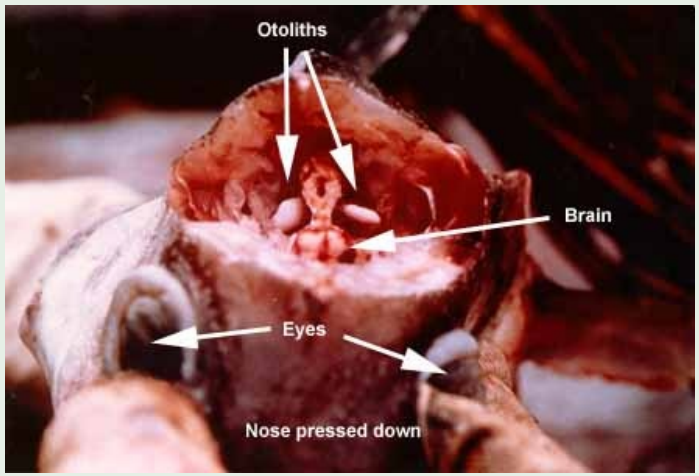


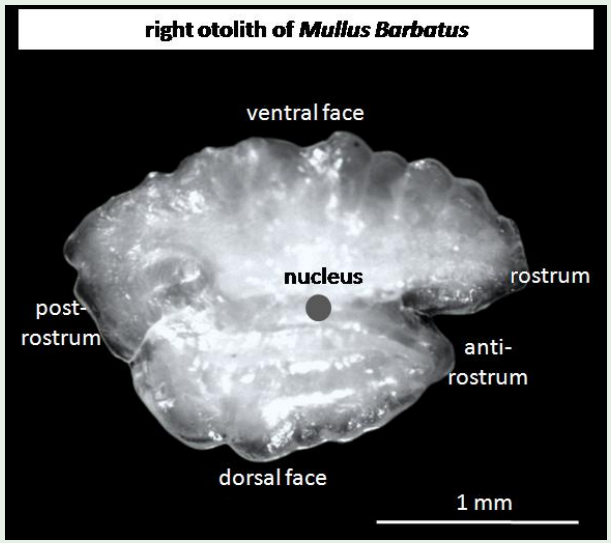


- a small piece of calcium carbonate located in the head of *Teleostean* fishes
  - acceleration perception and sense of balance
  - tridimensional perception and aid in hearing functions

Exemples : Y catégorielle et X fonctionnelle bivariée

Example (location and shape)

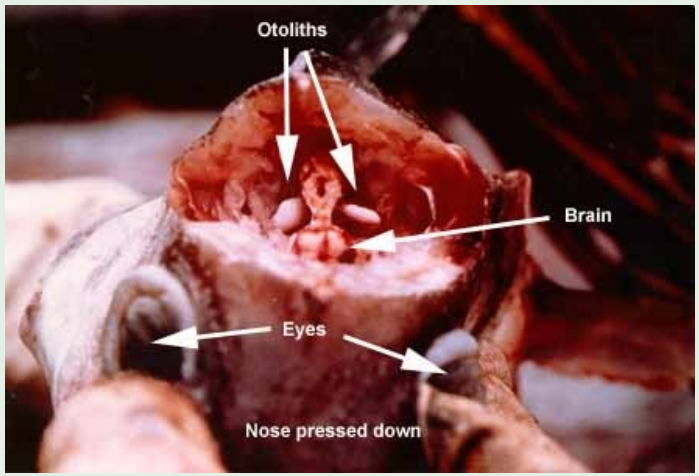


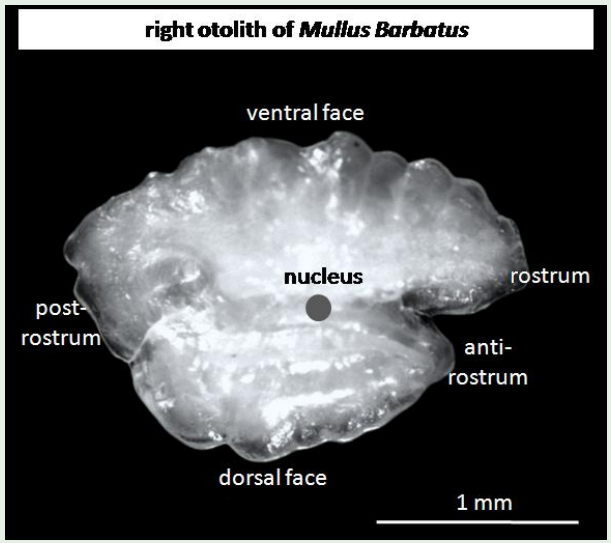


- a small piece of calcium carbonate located in the head of *Teleostean* fishes
  - acceleration perception and sense of balance
  - tridimensional perception and aid in hearing functions
- 3 pairs

Exemples : Y catégorielle et X fonctionnelle bivariée

Example (location and shape)

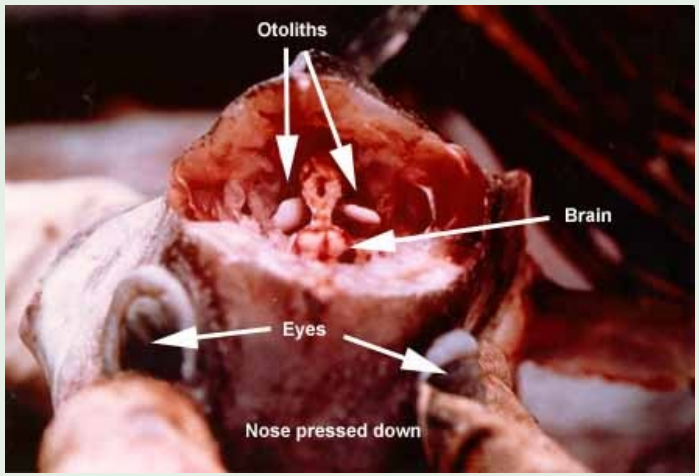


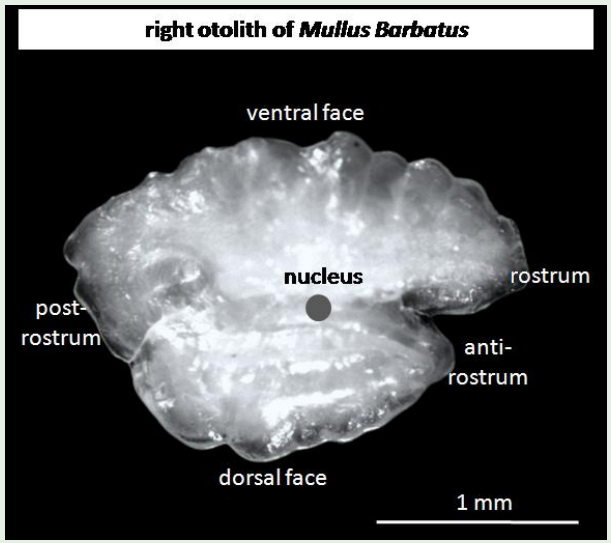


- a small piece of calcium carbonate located in the head of *Teleostean* fishes
  - acceleration perception and sense of balance
  - tridimensional perception and aid in hearing functions
- 3 pairs
- *sagitta* pair

Exemples : Y catégorielle et X fonctionnelle bivariée

Example (location and shape)

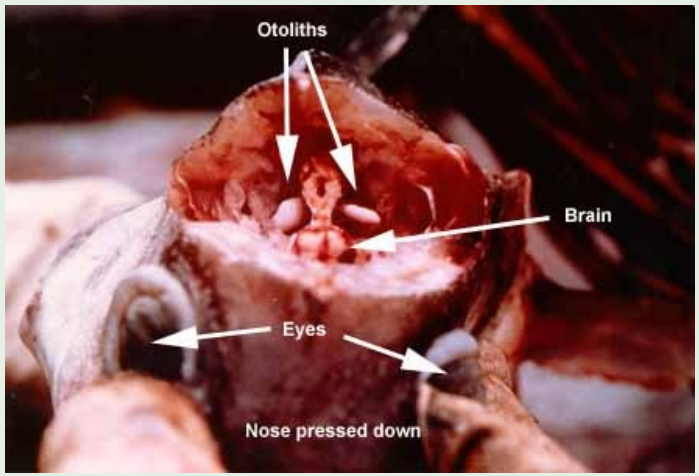


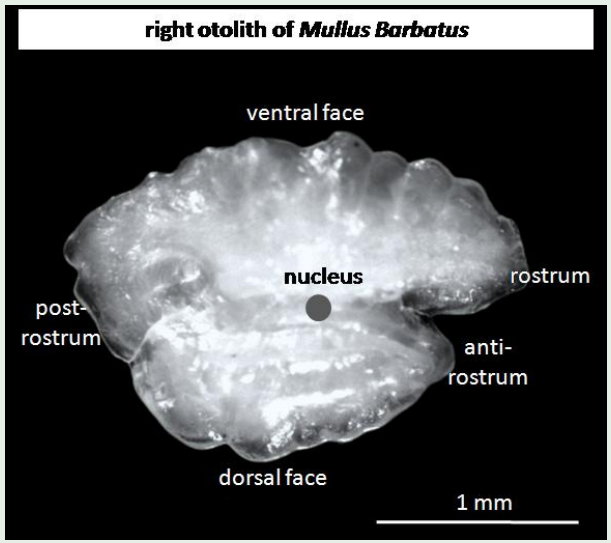


- a small piece of calcium carbonate located in the head of *Teleostean* fishes
  - acceleration perception and sense of balance
  - tridimensional perception and aid in hearing functions
- 3 pairs
- *sagitta* pair
  - the biggest one

Exemples : Y catégorielle et X fonctionnelle bivariée

Example (location and shape)

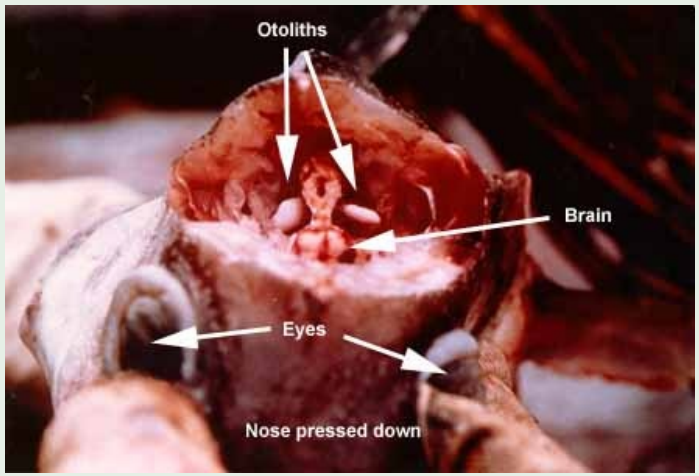


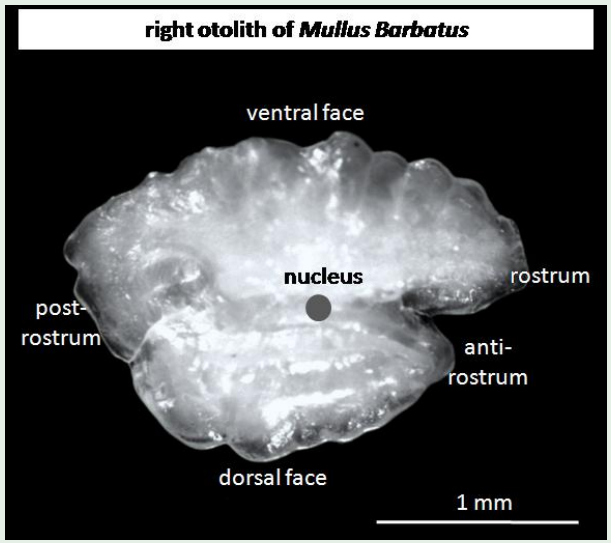


- a small piece of calcium carbonate located in the head of *Teleostean* fishes
  - acceleration perception and sense of balance
  - tridimensional perception and aid in hearing functions
- 3 pairs
- *sagitta* pair
  - the biggest one
  - grows from the date of hatch to the time of death

Exemples : Y catégorielle et X fonctionnelle bivariée

Example (location and shape)





- a small piece of calcium carbonate located in the head of *Teleostean* fishes
  - acceleration perception and sense of balance
  - tridimensional perception and aid in hearing functions
- 3 pairs
- *sagitta* pair
  - the biggest one
  - grows from the date of hatch to the time of death
  - growth starts from *nucleus*

What bring otolith studies ?

- Fish stocks, populations and species can be studied through the size, composition and shape of their otoliths
- Chemical composition and shape provide informations about fish environment (temperature, migration pathways) and dietary pattern
- Lateral section shows daily growth rings that record age and growth patterns

Example (Otolith and tree rings)

Shape analysis

- We aim to compare shape of a collection of *Mullus barbatus* otoliths

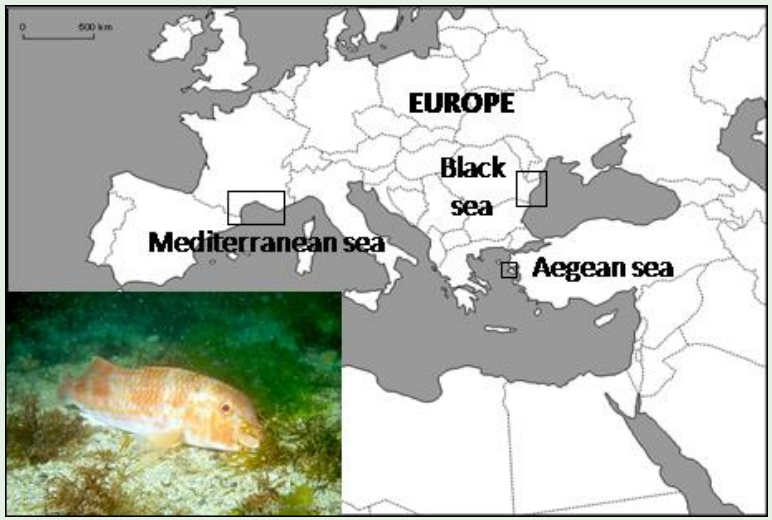
Shape analysis

- We aim to compare shape of a collection of *Mullus barbatus* otoliths
  - **objective** : discriminate individuals belonging to same species but sampled at different places

Shape analysis

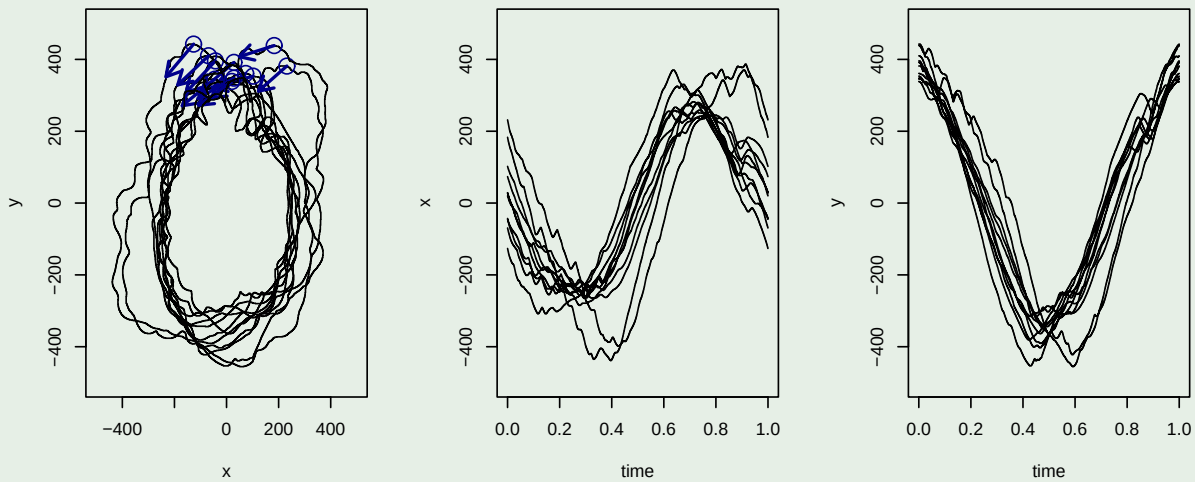
- We aim to compare shape of a collection of *Mullus barbatus* otoliths
  - **objective** : discriminate individuals belonging to same species but sampled at different places

Example (Sampling location & *Mullus barbatus*)



Exemples : Y catégorielle et X fonctionnelle bivariée

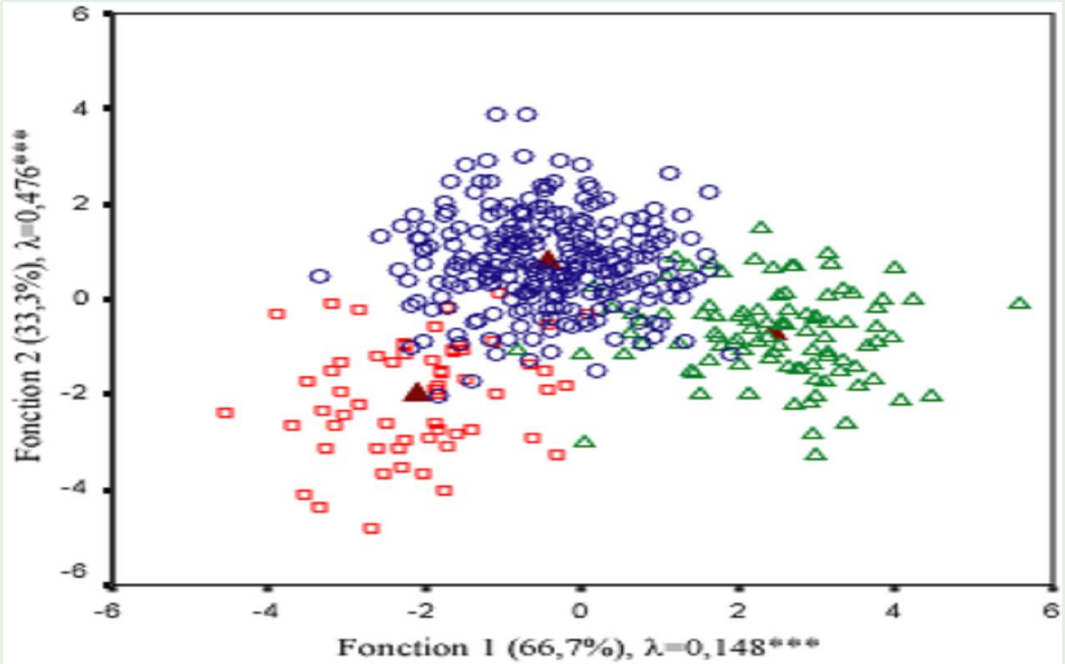
Example (Sampling location & *Mullus barbatus*)



- Problème de recalage (analyse Procuste), objets bidimensionnels
- Peut-on utiliser leurs formes pour connaître leur provenance ?

Exemples : Y catégorielle et X fonctionnelle bivariée

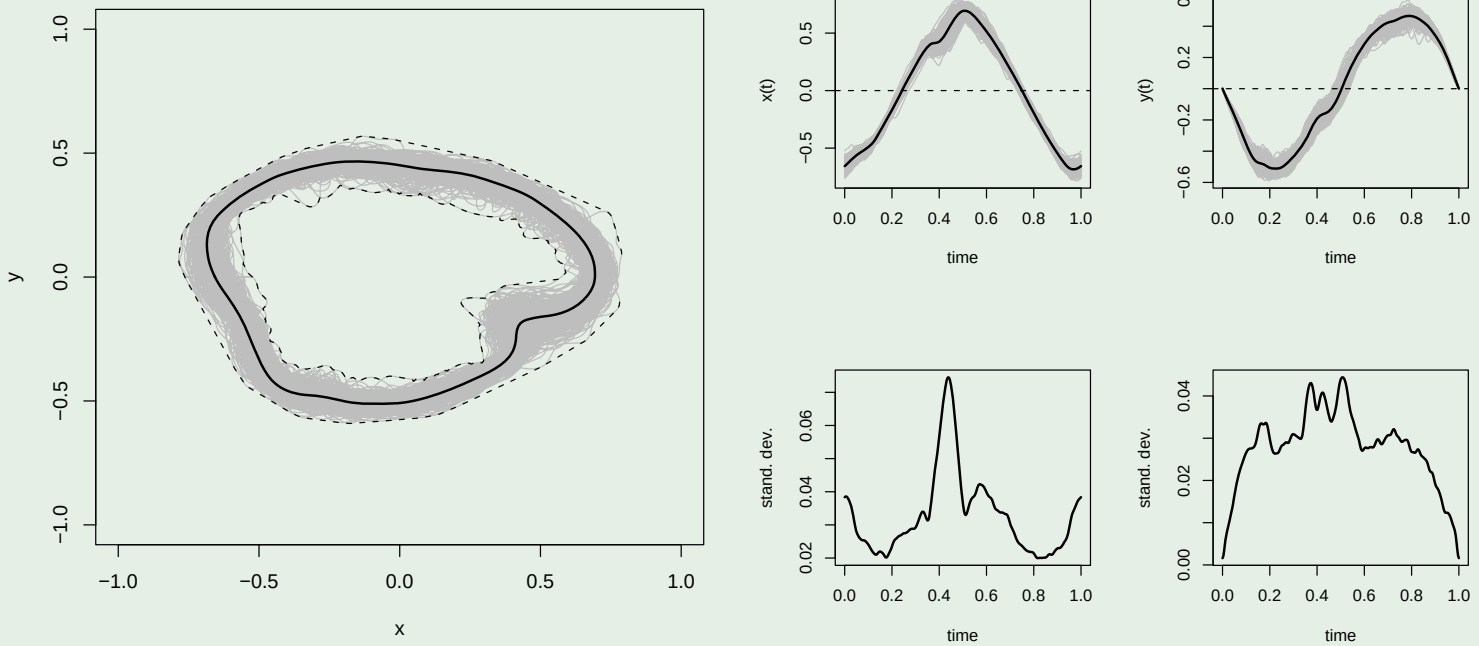
Example (Analyse Discriminante)



- Séparer au mieux les k classes a priori sur l'ensemble des individus échantillonnés à partir des p prédicteurs

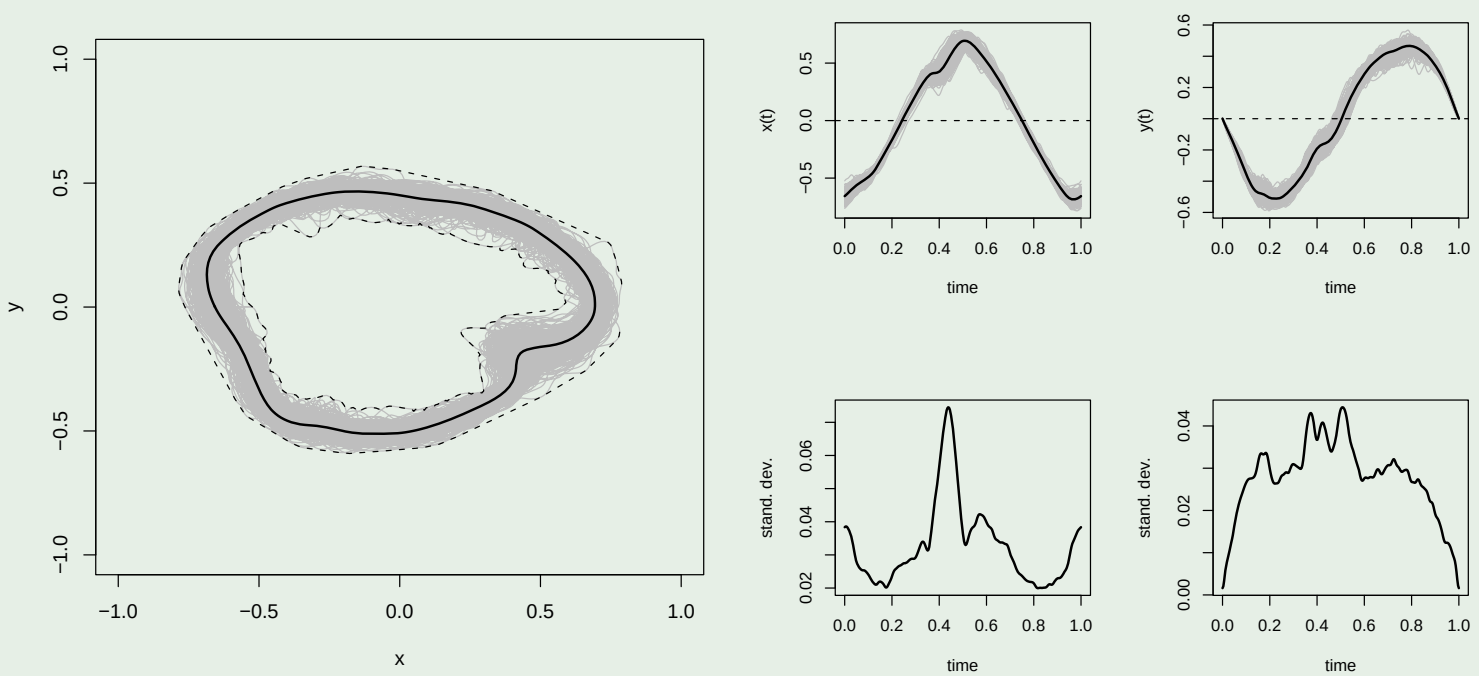
Exemples : Y catégorielle et X fonctionnelle bivariée

Example



Exemples : Y catégorielle et X fonctionnelle bivariée

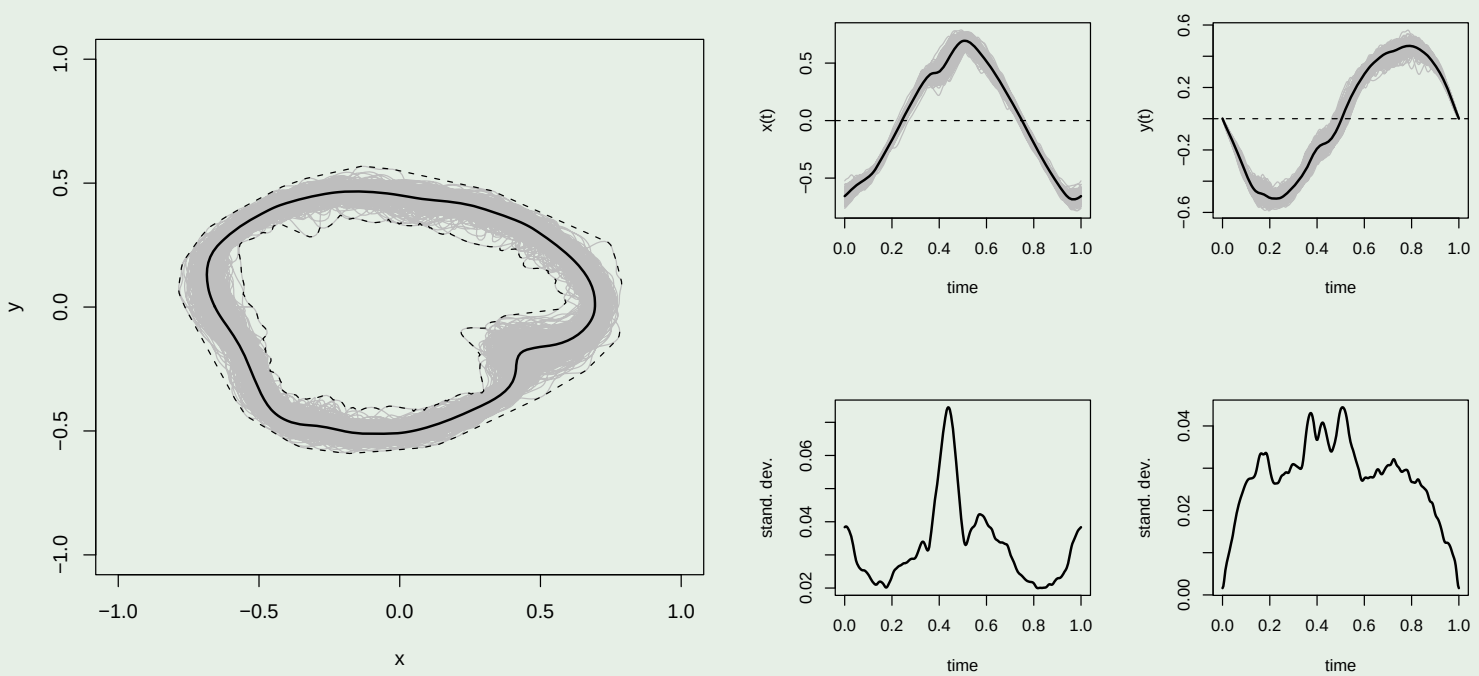
Example



- The average otolith is defined as $\{(\overline{x}(t), \overline{y}(t)), t \in [0;1]\}$

Exemples : Y catégorielle et X fonctionnelle bivariée

Example



- The average otolith is defined as $\{(\overline{x}(t), \overline{y}(t)), t \in [0;1]\}$
- The variance can be computed and allows to identify variability zones

Apprentissage supervisé et non-supervisé

Definition

L'ensemble des méthodes qui cherchent à reproduire une variable Y étant données des observations simultanées de $\mathbf{X}$ font partie des méthodes d'*apprentissage supervisé*. Il s'agit de construire un modèle

$$y = f_\theta(\mathbf{x}) + \varepsilon$$

à partir d'un échantillon d'apprentissage E.

- Et s'il n'y a pas d'information a priori ?

Apprentissage supervisé et non-supervisé

Definition

L'ensemble des méthodes qui cherchent à reproduire une variable Y étant données des observations simultanées de $\mathbf{X}$ font partie des méthodes d'*apprentissage supervisé*. Il s'agit de construire un modèle

$$y = f_\theta(\mathbf{x}) + \varepsilon$$

à partir d'un échantillon d'apprentissage E.

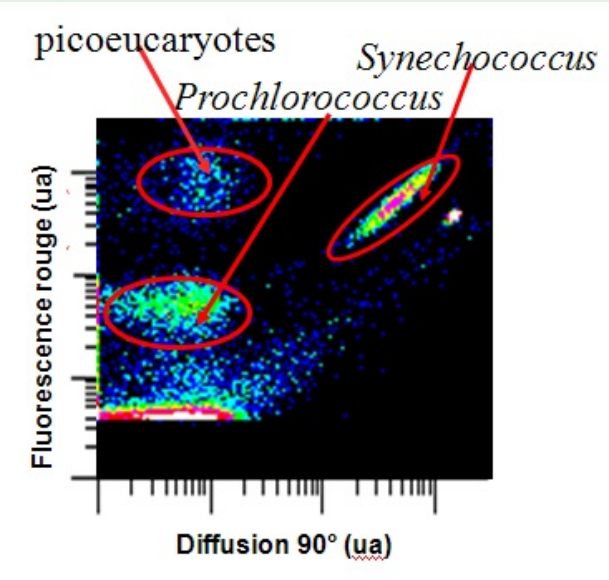
- Et s'il n'y a pas d'information a priori ?

Definition

Lorsqu'on cherche à regrouper des objets qui se ressemblent sur la base d'un échantillon des seules variables $\mathbf{X}$, on parlera d'*apprentissage non supervisé*. On ne connaît pas a priori l'appartenance d'un individus à une classe qui serait donnée par Y : il faut déterminer un nombre de classes optimal

Exemple : X réels

Example (cytométrie en flux)



- On utilisera dans ce cas des *méthodes de classification*

Exemple : X réels

Example (Hydrologie)

D. Nerini et al. / C. R. Acad. Sci. Paris, Sciences de la Terre et des planètes / Earth and Planetary Sciences 332 (2001) 457–464

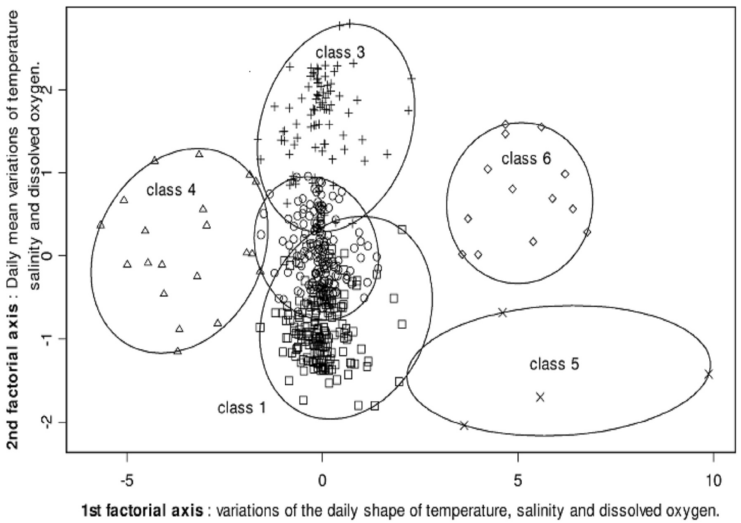


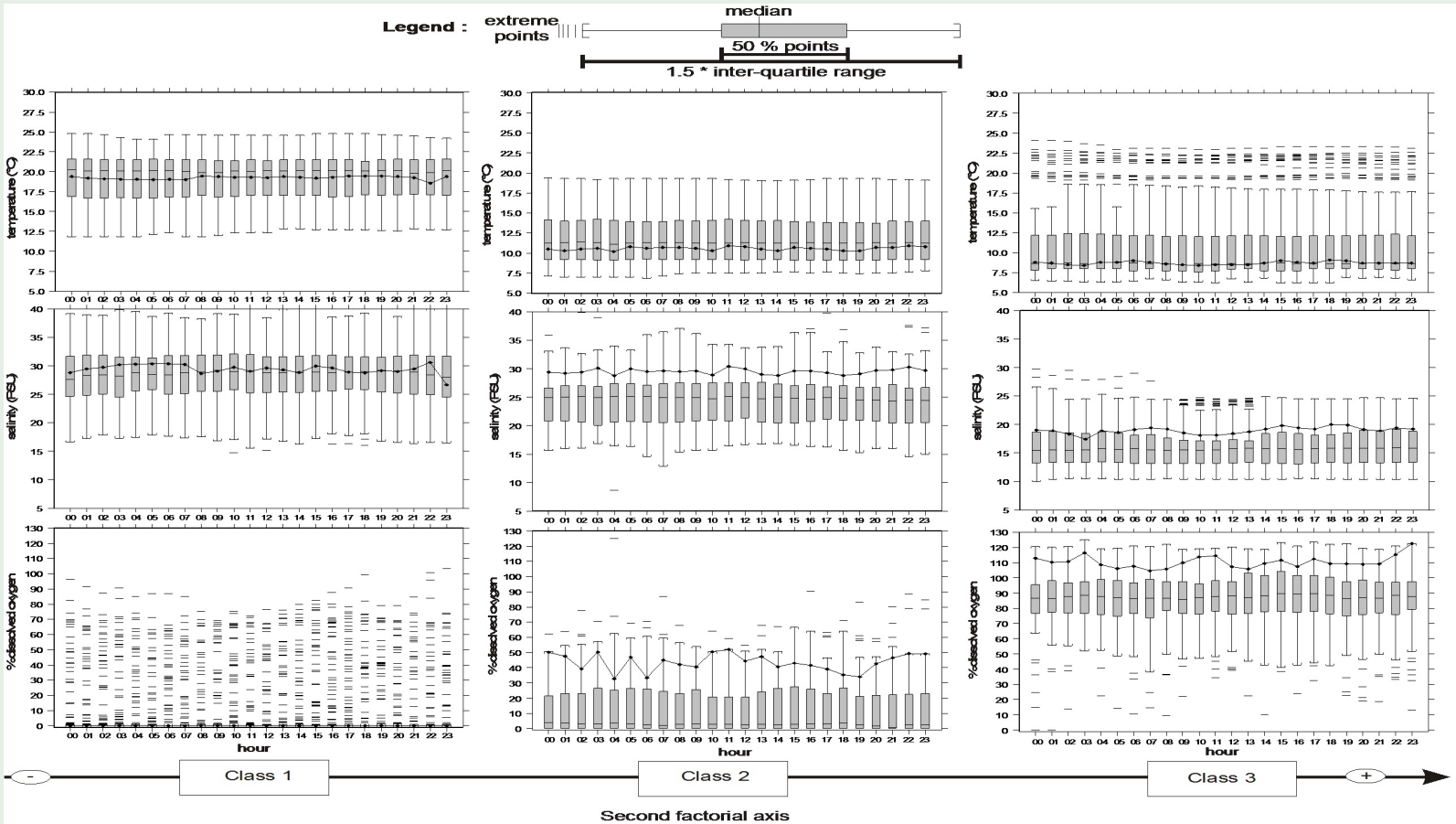
Figure 2. Plan factoriel 1/2 de l'AFM et ellipses de volume minimum [12] associées à chacune des classes issues de la CAH sur les huit premiers facteurs de l'AFM.

Figure 2. Bivariate plot (principal components 1/2) and minimum volume ellipsoids [12] associated to the hierarchical clustering of the data using eight principal components of the MFA.

- On utilisera dans ce cas des *méthodes de classification*

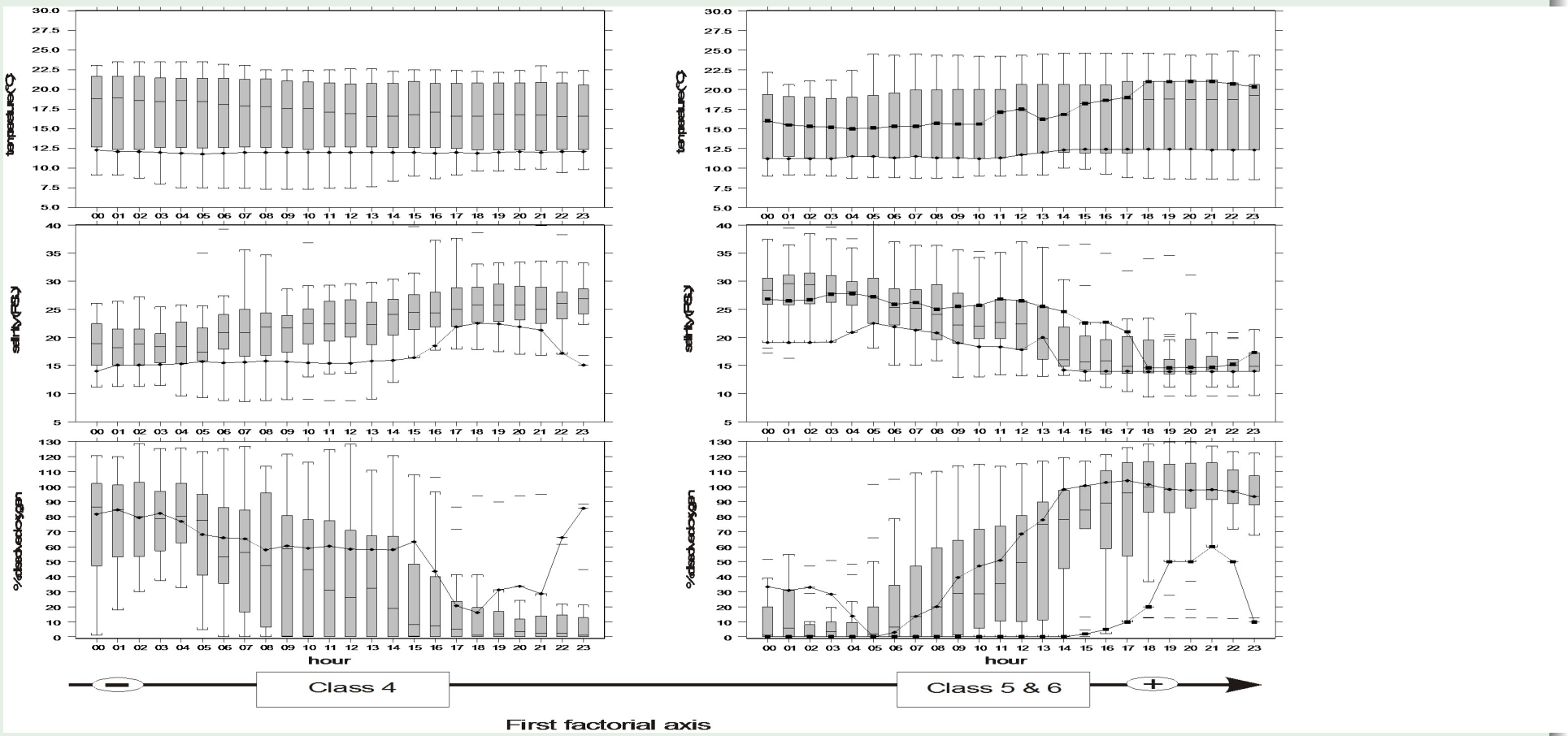
Exemple : X réels

Example (Hydrologie)



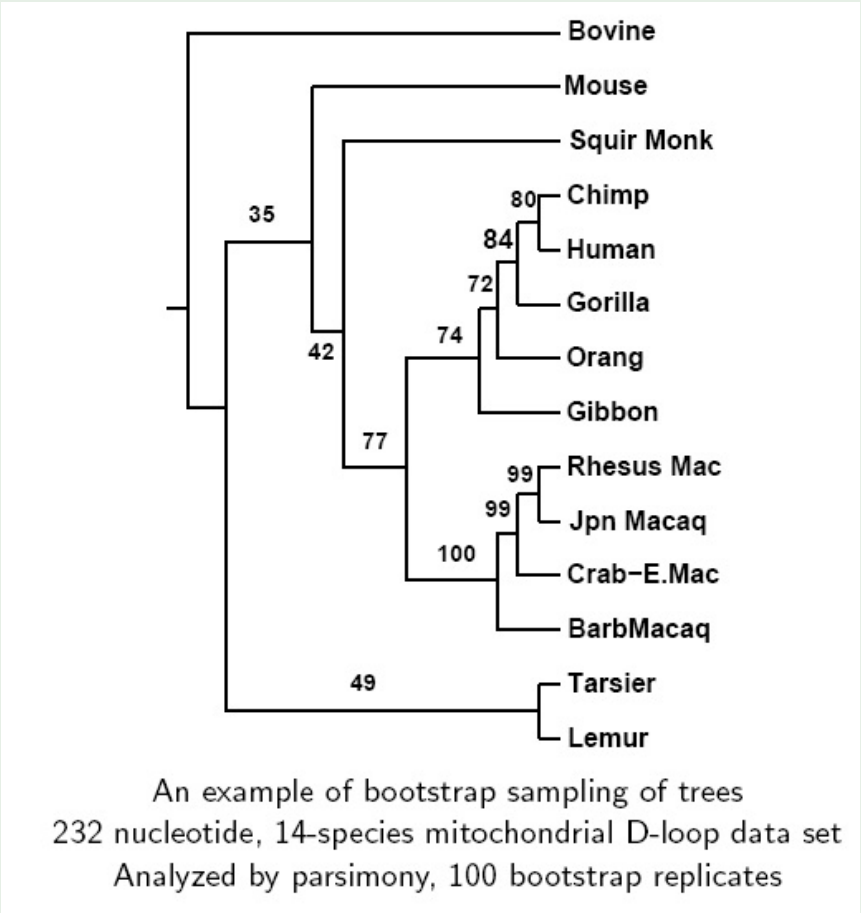
Exemple : X réels

Example (Hydrologie)



Exemple : X qualitatifs

Example (Phylognie)



An example of bootstrap sampling of trees
232 nucleotide, 14-species mitochondrial D-loop data set
Analyzed by parsimony, 100 bootstrap replicates

- Un arbre de classification ascendante est construit à partir d'un échantillon de "mots"

Récapitulatif

- L'apprentissage statistique suppose l'existence d'un *échantillon de construction* E. Il est constitué de p *variables explicatives* ou *prédictives* notées sous la forme d'un vecteur $\mathbf{X}$.

Récapitulatif

- L'apprentissage statistique suppose l'existence d'un *échantillon de construction* E. Il est constitué de p *variables explicatives* ou *prédictives* notées sous la forme d'un vecteur $\mathbf{X}$.

1. $\mathbf{X}_{\mathbb{R}}$ toutes quantitatives,

Récapitulatif

- L'apprentissage statistique suppose l'existence d'un *échantillon de construction* E. Il est constitué de p *variables explicatives* ou *prédictives* notées sous la forme d'un vecteur $\mathbf{X}$.

1. $\mathbf{X}_{\mathbb{R}}$ toutes quantitatives,
2. $\mathbf{X}_{\Theta}$ toutes qualitatives,

Récapitulatif

- L'apprentissage statistique suppose l'existence d'un *échantillon de construction* E. Il est constitué de p *variables explicatives* ou *prédictives* notées sous la forme d'un vecteur $\mathbf{X}$.

1. $\mathbf{X}_{\mathbb{R}}$ toutes quantitatives,
2. $\mathbf{X}_{\Theta}$ toutes qualitatives,
3. $\mathbf{X}_{\mathbb{R}\cup\Theta}$ un mélange de qualitatives et quantitatives.

Récapitulatif

- L'apprentissage statistique suppose l'existence d'un *échantillon de construction* E. Il est constitué de p *variables explicatives* ou *prédictives* notées sous la forme d'un vecteur $\mathbf{X}$.

1. $\mathbf{X}_{\mathbb{R}}$ toutes quantitatives,
2. $\mathbf{X}_{\Theta}$ toutes qualitatives,
3. $\mathbf{X}_{\mathbb{R}\cup\Theta}$ un mélange de qualitatives et quantitatives.

- La variable à expliquer (à prédire, cible) peut être :

Récapitulatif

- L'apprentissage statistique suppose l'existence d'un *échantillon de construction* E. Il est constitué de p *variables explicatives* ou *prédictives* notées sous la forme d'un vecteur $\mathbf{X}$.

1. $\mathbf{X}_{\mathbb{R}}$ toutes quantitatives,
2. $\mathbf{X}_{\Theta}$ toutes qualitatives,
3. $\mathbf{X}_{\mathbb{R}\cup\Theta}$ un mélange de qualitatives et quantitatives.

- La variable à expliquer (à prédire, cible) peut être :

1. $Y \in \mathbb{R}$ quantitative,

Récapitulatif

- L'apprentissage statistique suppose l'existence d'un *échantillon de construction* E. Il est constitué de p *variables explicatives* ou *prédictives* notées sous la forme d'un vecteur $\mathbf{X}$.

1. $\mathbf{X}_{\mathbb{R}}$ toutes quantitatives,
2. $\mathbf{X}_{\Theta}$ toutes qualitatives,
3. $\mathbf{X}_{\mathbb{R}\cup\Theta}$ un mélange de qualitatives et quantitatives.

- La variable à expliquer (à prédire, cible) peut être :

1. $Y \in \mathbb{R}$ quantitative,
2. $Z \in \{0, 1\}$ qualitative à 2 modalités,

Récapitulatif

- L'apprentissage statistique suppose l'existence d'un *échantillon de construction* E. Il est constitué de p *variables explicatives* ou *prédictives* notées sous la forme d'un vecteur $\mathbf{X}$.

1. $\mathbf{X}_{\mathbb{R}}$ toutes quantitatives,
2. $\mathbf{X}_{\Theta}$ toutes qualitatives,
3. $\mathbf{X}_{\mathbb{R}\cup\Theta}$ un mélange de qualitatives et quantitatives.

- La variable à expliquer (à prédire, cible) peut être :

1. $Y \in \mathbb{R}$ quantitative,
2. $Z \in \{0, 1\}$ qualitative à 2 modalités,
3. $T \in \Theta$ qualitative.

Récapitulatif

- L'apprentissage statistique suppose l'existence d'un *échantillon de construction* E. Il est constitué de p *variables explicatives* ou *prédictives* notées sous la forme d'un vecteur $\mathbf{X}$.

1. $\mathbf{X}_{\mathbb{R}}$ toutes quantitatives,
2. $\mathbf{X}_{\Theta}$ toutes qualitatives,
3. $\mathbf{X}_{\mathbb{R}\cup\Theta}$ un mélange de qualitatives et quantitatives.

- La variable à expliquer (à prédire, cible) peut être :

1. $Y \in \mathbb{R}$ quantitative,
2. $Z \in \{0, 1\}$ qualitative à 2 modalités,
3. $T \in \Theta$ qualitative.
4. inexistante

Récapitulatif

- Trois objectifs principaux sont poursuivis dans les applications classiques d'apprentissage :

Récapitulatif

- Trois objectifs principaux sont poursuivis dans les applications classiques d'apprentissage :

1. **Exploration multidimensionnelle ou réduction de dimension** : production de graphes, d'un sous-ensemble de variables représentatives des données initiales ou d'un ensemble de composantes préalable à l'utilisation d'une autre technique (ex. Berre)

Récapitulatif

- Trois objectifs principaux sont poursuivis dans les applications classiques d'apprentissage :

1. **Exploration multidimensionnelle ou réduction de dimension** : production de graphes, d'un sous-ensemble de variables représentatives des données initiales ou d'un ensemble de composantes préalable à l'utilisation d'une autre technique (ex. Berre)
2. **Classification (clustering) ou segmentation** : production d'une variable qualitative à partir de données quantitatives ou qualitatives.

Récapitulatif

- Trois objectifs principaux sont poursuivis dans les applications classiques d'apprentissage :

1. **Exploration multidimensionnelle ou réduction de dimension** : production de graphes, d'un sous-ensemble de variables représentatives des données initiales ou d'un ensemble de composantes préalable à l'utilisation d'une autre technique (ex. Berre)
2. **Classification (clustering) ou segmentation** : production d'une variable qualitative à partir de données quantitatives ou qualitatives.
3. **Modélisation** (Y ou Z) ou **Discrimination** (Z ou T) : production d'un *modèle de prévision* de Y (resp. Z, T)

Récapitulatif

- Trois objectifs principaux sont poursuivis dans les applications classiques d'apprentissage :

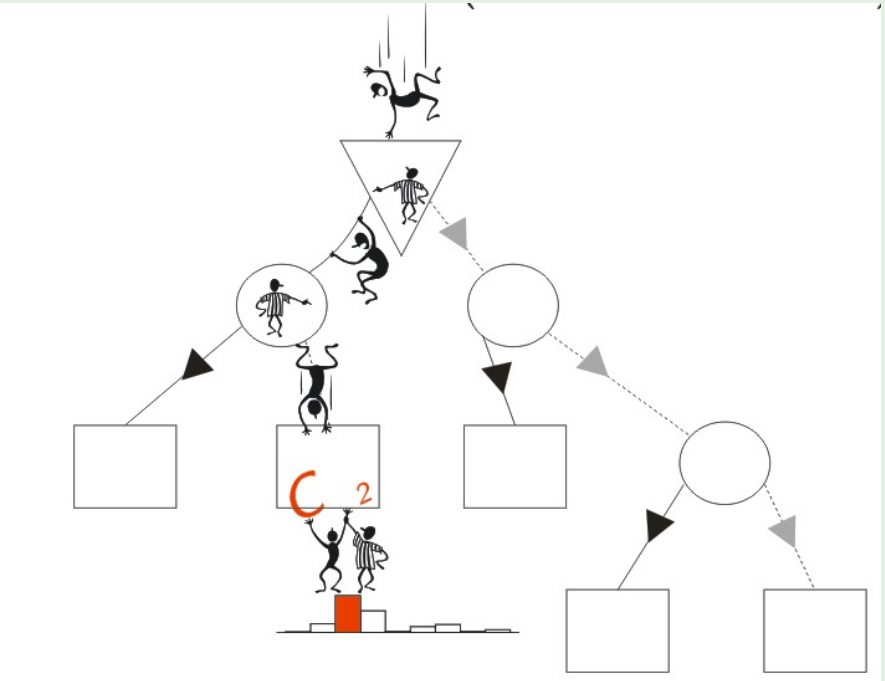
1. **Exploration multidimensionnelle ou réduction de dimension** : production de graphes, d'un sous-ensemble de variables représentatives des données initiales ou d'un ensemble de composantes préalable à l'utilisation d'une autre technique (ex. Berre)
2. **Classification (clustering) ou segmentation** : production d'une variable qualitative à partir de données quantitatives ou qualitatives.
3. **Modélisation** (Y ou Z) ou **Discrimination** (Z ou T) : production d'un *modèle de prévision* de Y (resp. Z, T)

- Une fois ces étapes déterminées, on peut passer à la *sélection* et à l'*utilisation* des modèles construits dans un but prévisionnel.

Prévision

- L'ensemble des méthodes utilisées peuvent servir à faire des prévisions

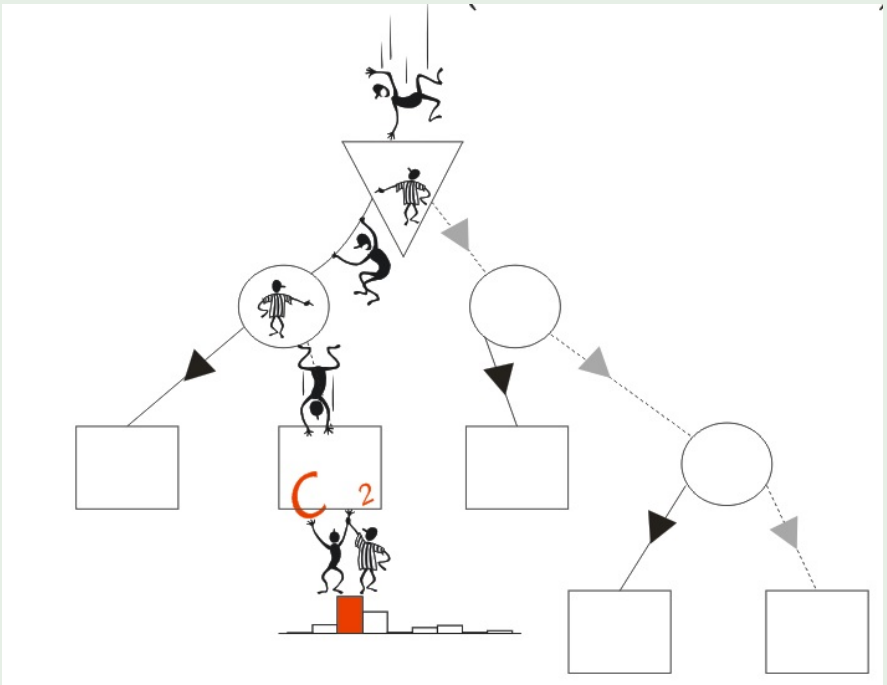
Example (Arbre de classification)



Prévision

- L'ensemble des méthodes utilisées peuvent servir à faire des prévisions

Example (Arbre de classification)



- Comment estimer le bon comportement du modèle ?

Prévision

- Les performances d'un modèle sont basées sur des critères de *qualité de prévision* qui visent à rechercher des *modèles parcimonieux*. Ces modèles fournissent un compromis entre une complexité limitée (faible nombre de paramètres ou flexibilité) et une bonne capacité de généralisation. L'interprétabilité passe au deuxième plan

Prévision

- Les performances d'un modèle sont basées sur des critères de *qualité de prévision* qui visent à rechercher des *modèles parcimonieux*. Ces modèles fournissent un compromis entre une complexité limitée (faible nombre de paramètres ou flexibilité) et une bonne capacité de généralisation. L'interprétabilité passe au deuxième plan

Example

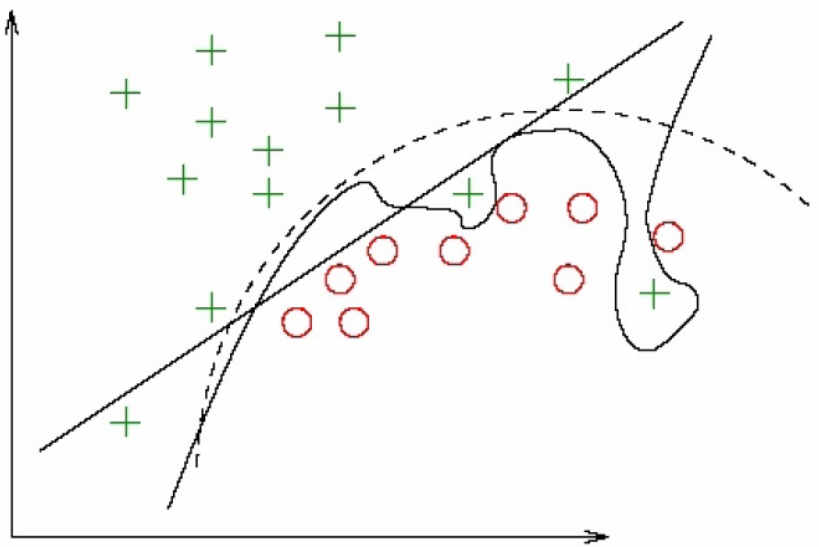
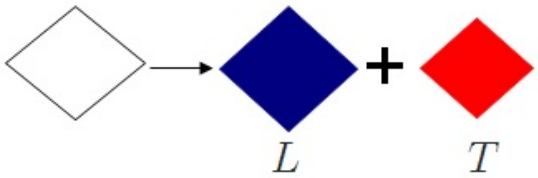
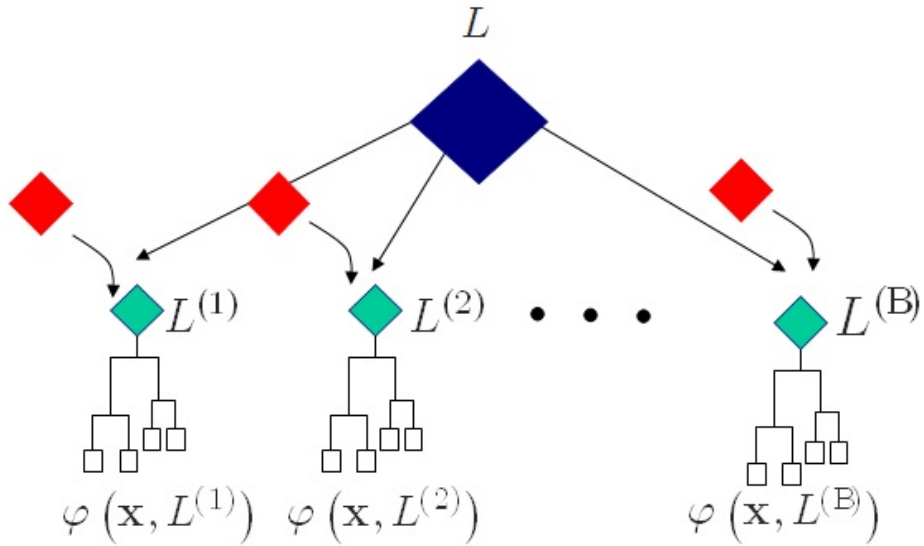


FIG. 1.2 – *Sous-ajustement linéaire et sur-ajustement local (proches voisins) d'un modèle quadratique.*

Modèles agrégés (bagging)

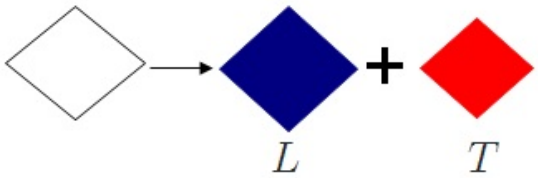
- Il existe un moyen efficace de tester la *capacité de prévision* d'un modèle grâce à la technique du Bootstrap

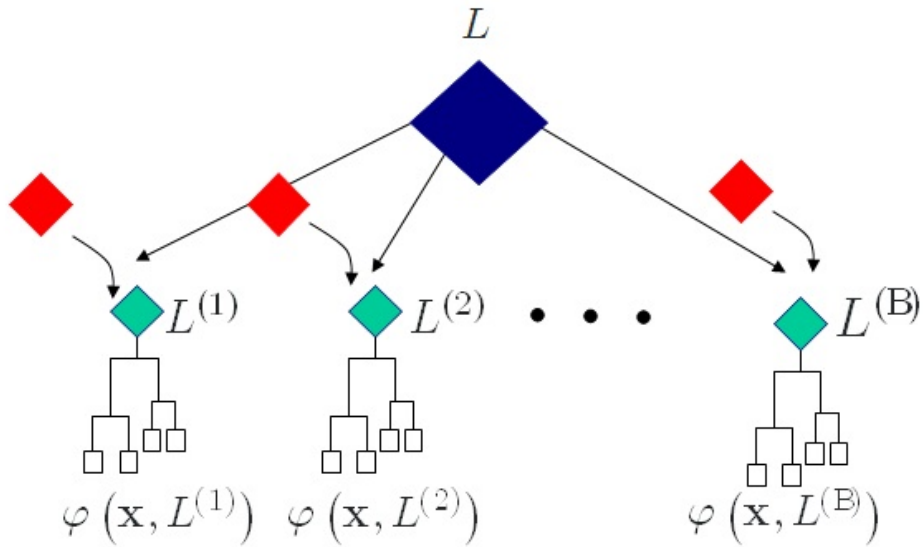




Modèles agrégés (bagging)

- Il existe un moyen efficace de tester la *capacité de prévision* d'un modèle grâce à la technique du Bootstrap

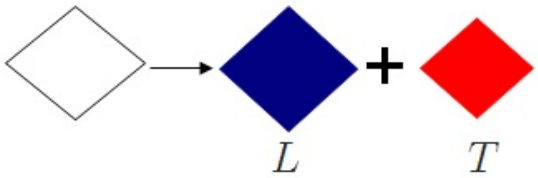


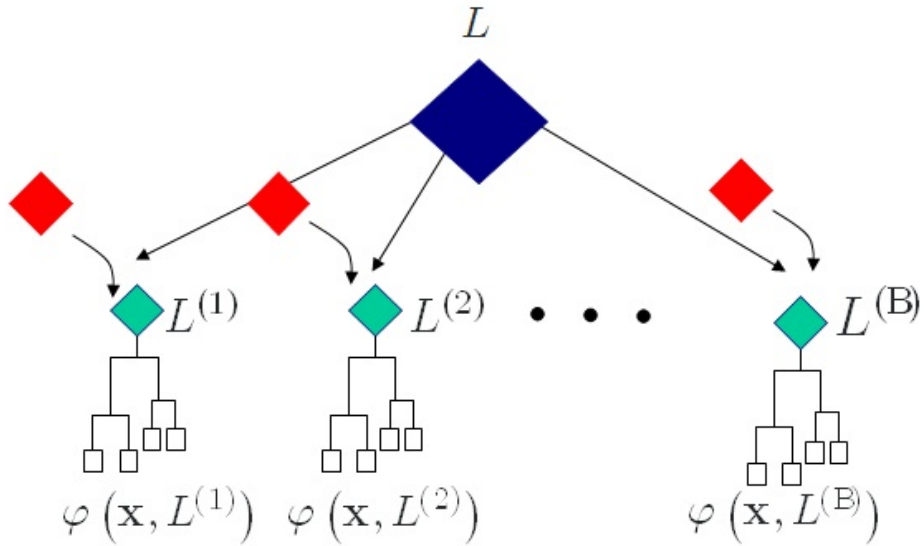


- La prévision est donnée par $\varphi_B(\mathbf{x}) = moy_i\left(\varphi\left(\mathbf{x}, \mathbf{L}^{(i)}\right)\right)$

Modèles agrégés (bagging)

- Il existe un moyen efficace de tester la *capacité de prévision* d'un modèle grâce à la technique du Bootstrap





- La prévision est donnée par $\varphi_B(\mathbf{x}) = moy_i\left(\varphi\left(\mathbf{x}, \mathbf{L}^{(i)}\right)\right)$
- On peut soit utiliser B fois le même modèle soit en utiliser plusieurs : la stratégie du choix du modèle n'est pas figée

Finalement

- Ce que vous devez savoir faire en sortant de ce cours :

Finalement

- Ce que vous devez savoir faire en sortant de ce cours :

1. Comprendre ce qu'est un **tableau de données** et les différentes **variables** qui le composent

Finalement

- Ce que vous devez savoir faire en sortant de ce cours :

1. Comprendre ce qu'est un **tableau de données** et les différentes **variables** qui le composent
2. Donner les **bases géométriques** nécessaires à la **construction des modèles** et à la compréhension de certaines notions en statistiques

Finalement

- Ce que vous devez savoir faire en sortant de ce cours :

1. Comprendre ce qu'est un **tableau de données** et les différentes **variables** qui le composent
2. Donner les **bases géométriques** nécessaires à la **construction des modèles** et à la compréhension de certaines notions en statistiques
3. Aborder quelques méthodes de **classification**, **segmentation** et d'**aide à la décision** (AFD)

Finalement

- Ce que vous devez savoir faire en sortant de ce cours :

1. Comprendre ce qu'est un **tableau de données** et les différentes **variables** qui le composent
2. Donner les **bases géométriques** nécessaires à la **construction des modèles** et à la compréhension de certaines notions en statistiques
3. Aborder quelques méthodes de **classification**, **segmentation** et d'**aide à la décision** (AFD)
4. Donner les bases de la construction de modèles agrégés par **bootstrap**

Finalement

- Ce que vous devez savoir faire en sortant de ce cours :

1. Comprendre ce qu'est un **tableau de données** et les différentes **variables** qui le composent
2. Donner les **bases géométriques** nécessaires à la **construction des modèles** et à la compréhension de certaines notions en statistiques
3. Aborder quelques méthodes de **classification**, **segmentation** et d'**aide à la décision** (AFD)
4. Donner les bases de la construction de modèles agrégés par **bootstrap**
5. Appliquer quelques unes de ces méthodes sur ordinateur

Finalement

- Ce que vous devez savoir faire en sortant de ce cours :

1. Comprendre ce qu'est un **tableau de données** et les différentes **variables** qui le composent
2. Donner les **bases géométriques** nécessaires à la **construction des modèles** et à la compréhension de certaines notions en statistiques
3. Aborder quelques méthodes de **classification**, **segmentation** et d'**aide à la décision** (AFD)
4. Donner les bases de la construction de modèles agrégés par **bootstrap**
5. Appliquer quelques unes de ces méthodes sur ordinateur

- Mon (piêtre) objectif est que vous puissiez ouvrir ensuite un livre qui traite de ce sujet sans effroi !