Comment boucher des trous dans des séries temporelles ?

David NERINI*

***Institut Méditerranéen d'Océanologie**
UMR 7294 MIO CNRS IRD
Campus de Luminy, Case 901
13288 MARSEILLE Cedex 09

Formation ECCOREV, Plateau de l'Arbois, Mai 2012

Analyse de Données Fonctionnelles

- En écologie, beaucoup de données se présentent sous forme de courbes

414 D. Nerini et al. / Journal of Multivariate Analysis 101 (2010) 409–418

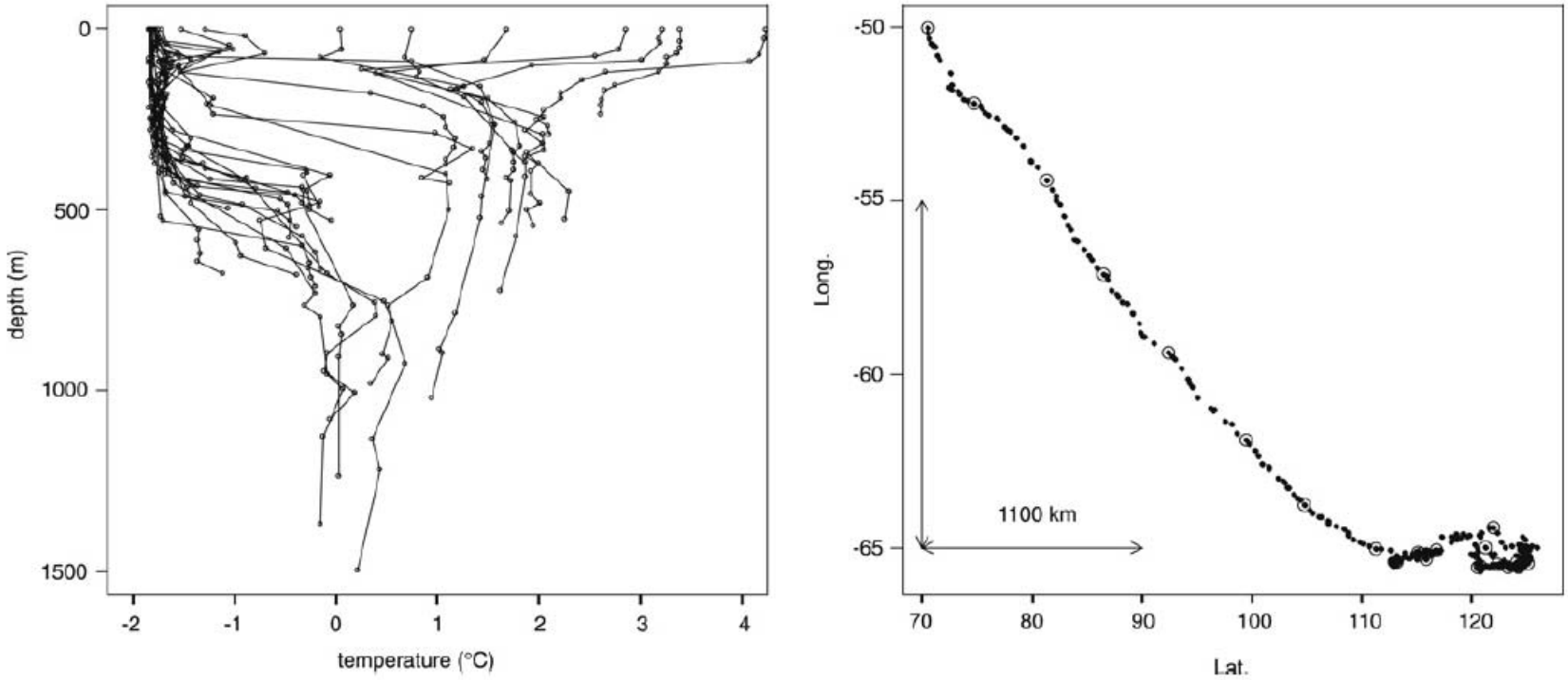


Fig. 1. Example of temperature profiles sampled by elephant seal and spatial location of sampled curves along trajectory (white circles). The cruise duration is 1 month. During the travel, the animal crosses and samples heterogeneous water structures.

Analyse de Données Fonctionnelles

D. Nerini, B. Ghattas / Computational Statistics & Data Analysis 51 (2007) 4984–4993 4985

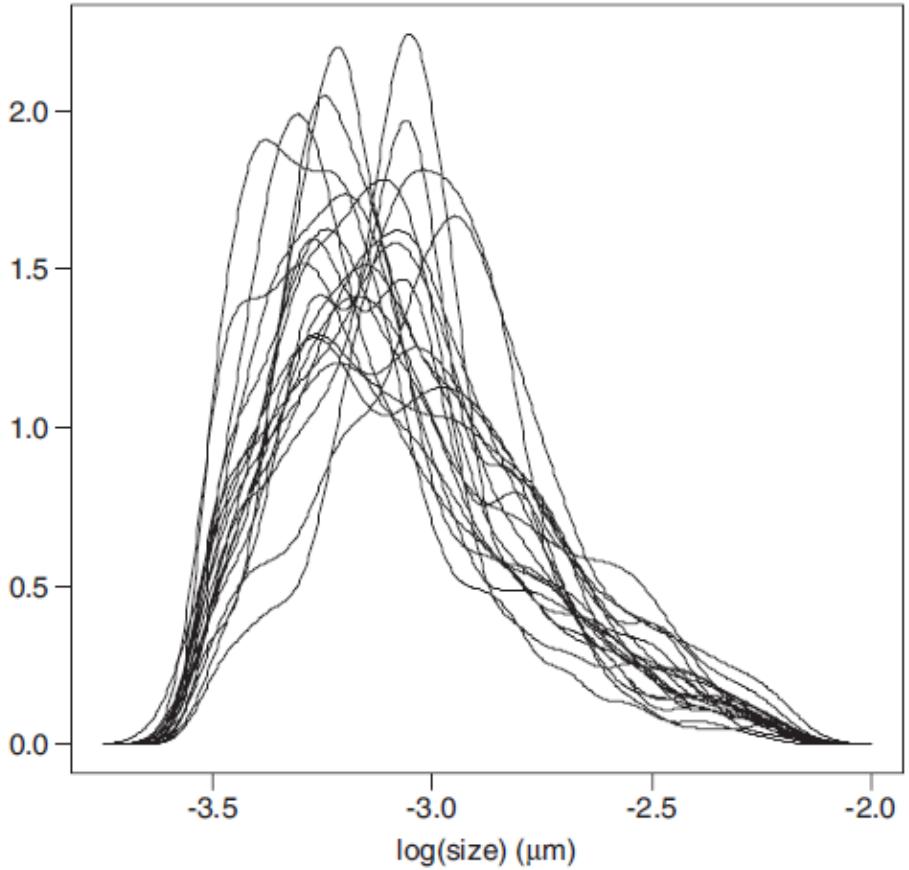
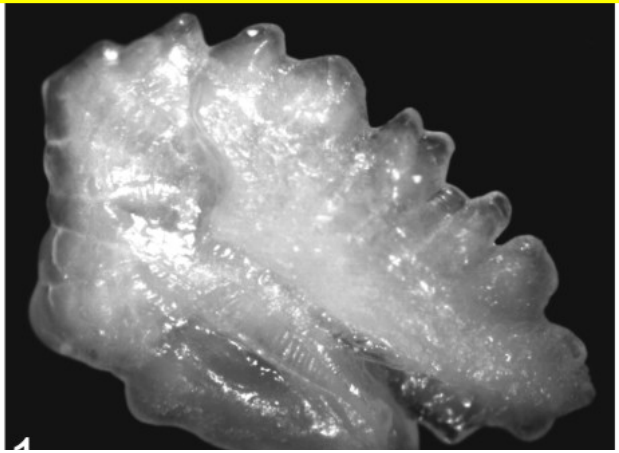
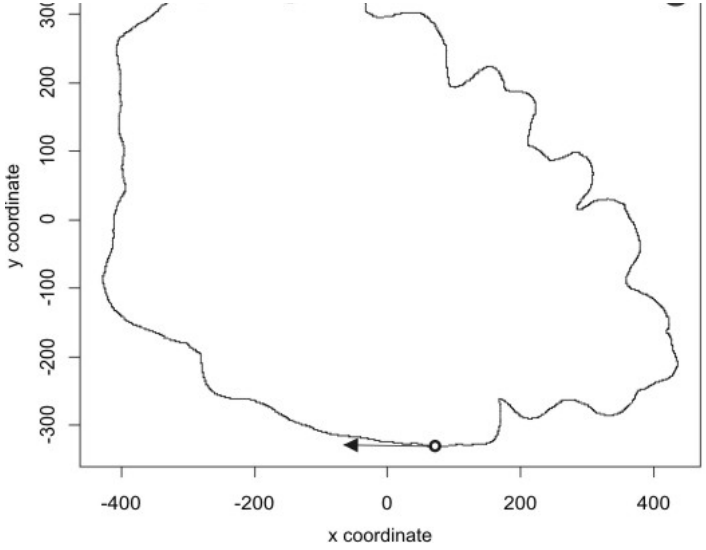


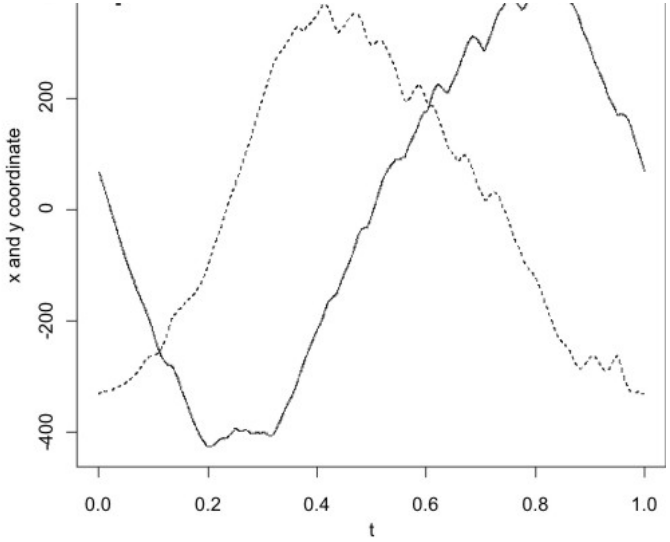
Fig. 1. Zooplankton size spectra. Each curve is a density which arises from a digital imaging system designed for estimating the size distribution of marine zooplankton net samples (Grosjean et al., 2004).

Analyse de Données Fonctionnelles



L'objectif de cette formation est de donner un cadre introductif à l'**analyse de données fonctionnelles**





Démarche générale

Variable fonctionnelle Y échantillonnée

$$\{y(t_j), j = 1, \dots, p\}$$

Il peut y avoir plusieurs observations

Modèle mathématique

- $\hat{y}(t) = f(t, \boldsymbol{\theta})$
- $\hat{y}(t) = m(t)$
- $\begin{cases} \frac{d\hat{y}}{dt}(t) = g(\hat{y}(t), t, \boldsymbol{\theta}) \\ \hat{y}(t = 0) = \hat{y}_0 \end{cases}$

$$y(t) = \hat{y}(t) + \varepsilon(t)$$

Idée fondamentale :

- Il existe une **erreur** entre modèle et données

Fonction de coût

$$SSE_t(\boldsymbol{\theta}) = \sum_{j=1}^{p} K\left(\frac{t - t_j}{h}\right)\left(y(t_j) - \hat{y}(t_j, \boldsymbol{\theta})\right)^2$$

- Estimer $\hat{y}(t)$ en utilisant des données revient souvent à **estimer le paramètre $\boldsymbol{\theta}$**

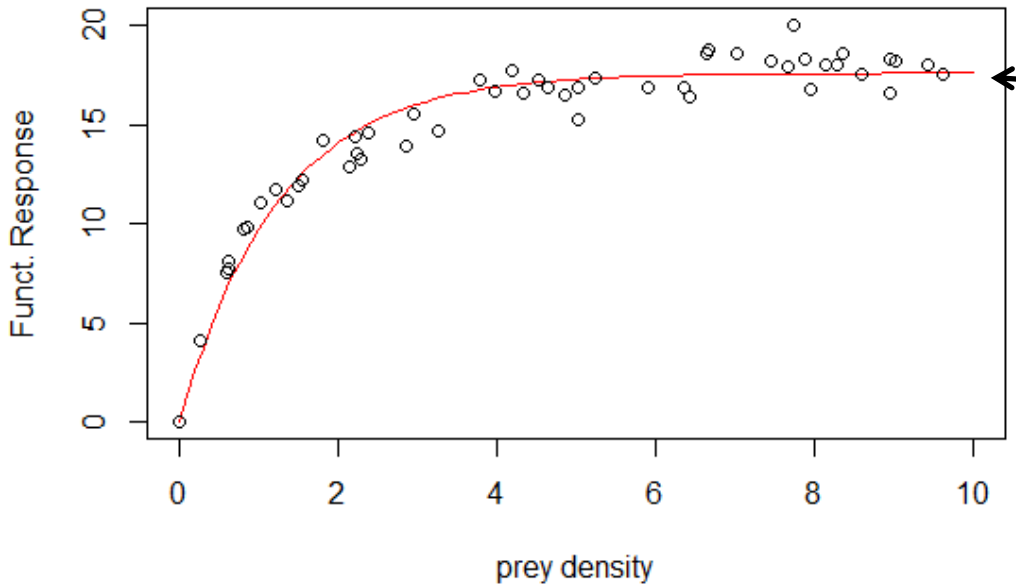
Démarche générale

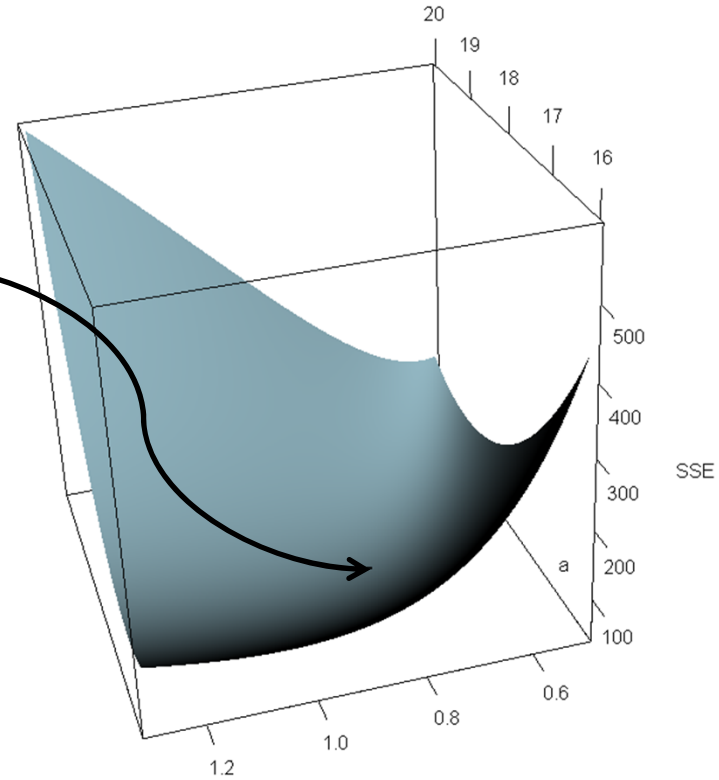
$$\left[\frac{dSSE}{d\boldsymbol{\theta}}(\boldsymbol{\theta}) = 0\right]_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$$

- L'estimation $\widehat{\boldsymbol{\theta}}$ du paramètre $\boldsymbol{\theta}$ est réalisée par minimisation de cette fonction de coût (**méthode des moindres carrés**)

- En fonction de la forme du modèle $\hat{y}(t)$ choisi, la solution n'a pas nécessairement d'**expression algébrique**.

- Dans ce cas, il faudra recourir à des **méthodes numériques de minimisation**

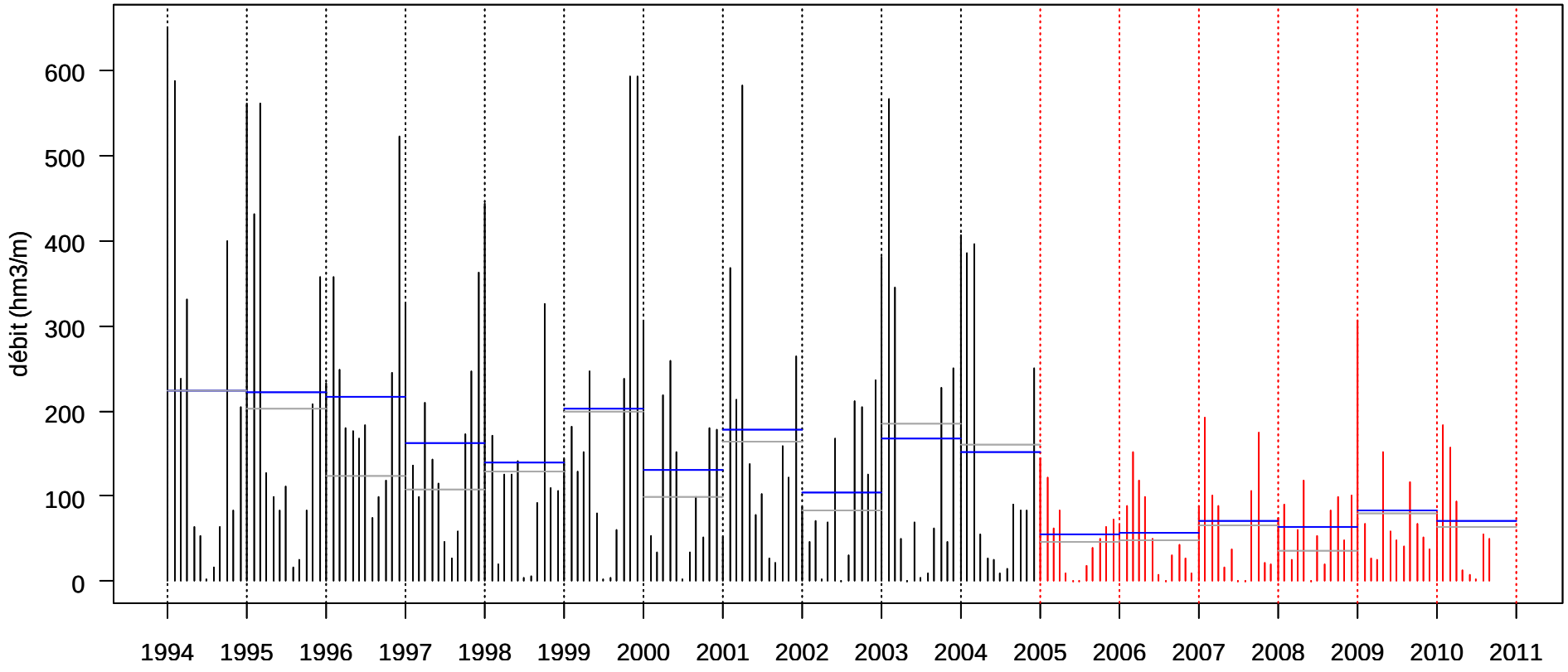




Finalement…

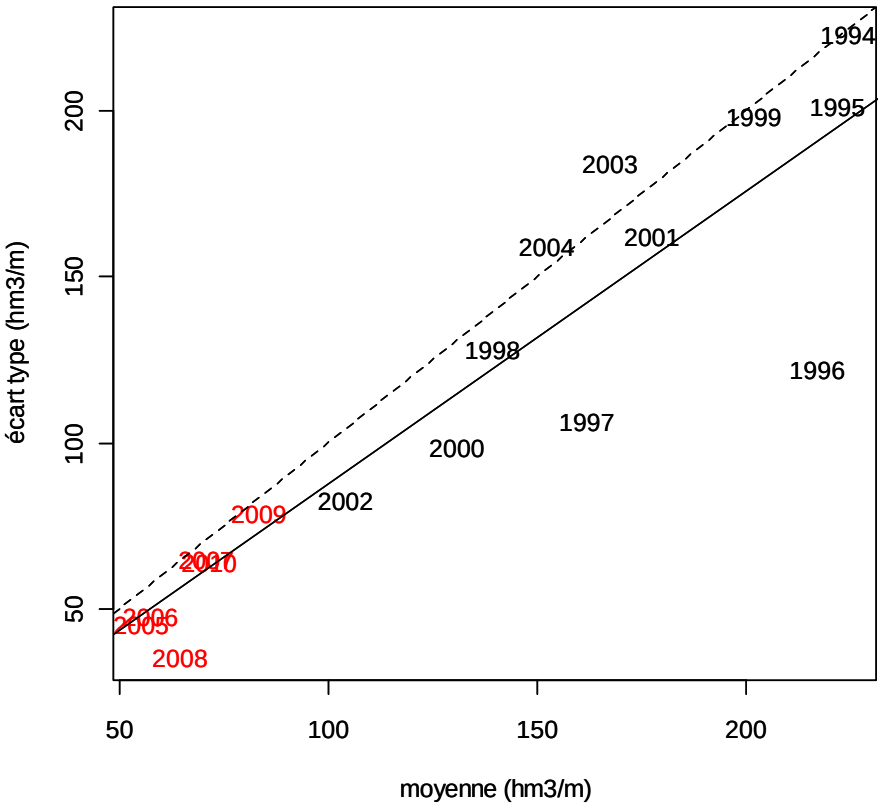
- On appelle donnée fonctionnelle toute donnée multivariée qui fait état d'une relation d'ordre sur sa dimension.
- On appelle variable fonctionnelle toute variable $\mathrm{y}(t)$ possédant un argument t. Les séries temporelles sont des données fonctionnelles.
- L'objectif est de reconstituer la fonction $\mathrm{y}(t)$ en utilisant des données fonctionnelles échantillonnées. On pourra après envisager d'autres méthodes d'analyse des données…

Premier exemple : Hydrologie



- **Modification** de la politique des rejets à partir de 2005
- **Objectif** : évaluation des variations physico-chimiques du milieu (T°,S, %O2)
- **Aucune hypothèse** sur la forme du modèle global
- Travail dans un **cadre linéaire**

La problématique de l'Etang de Berre

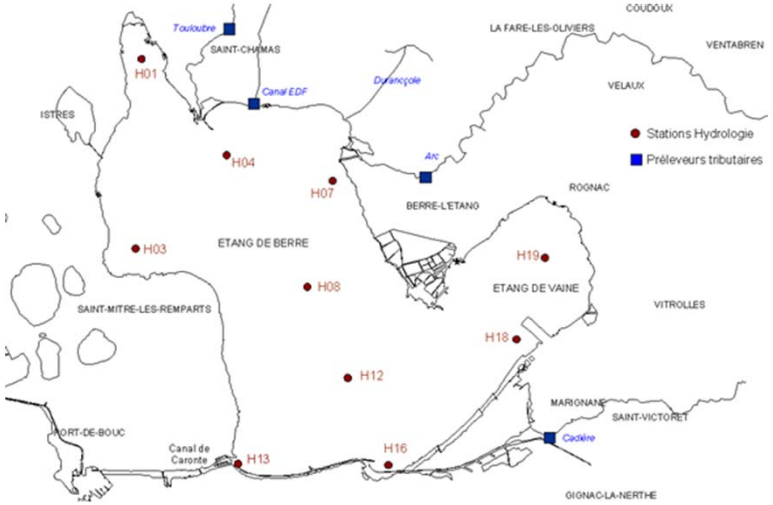


- La variabilité des rejets diminue de manière linéaire avec la moyenne

- Pour chaque **indicateur biologique**, l'expérimentation de nouvelles modalités de rejet s'est-elle accompagnée d'une **évolution positive et significative** temporelle ou spatiale?

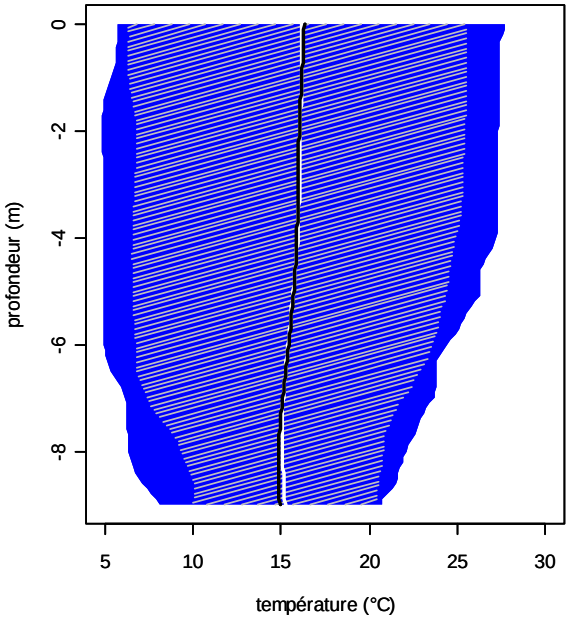
- Les évolutions observées peuvent-elles être mises en relation avec les facteurs physico-chimiques du milieu?

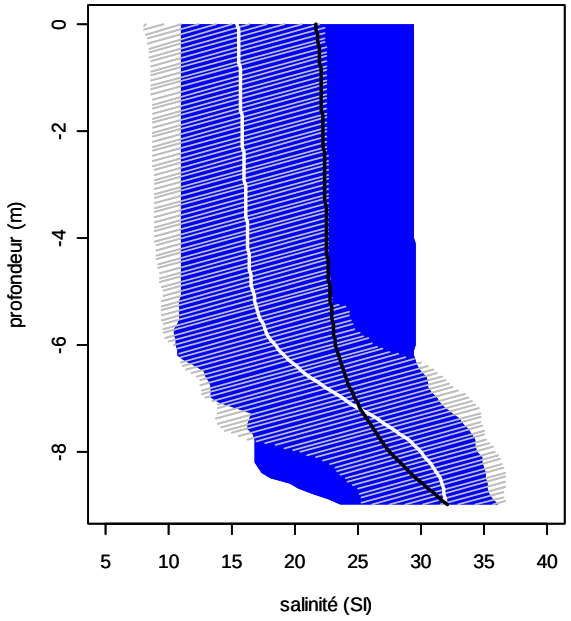
La base de données du GIPREB

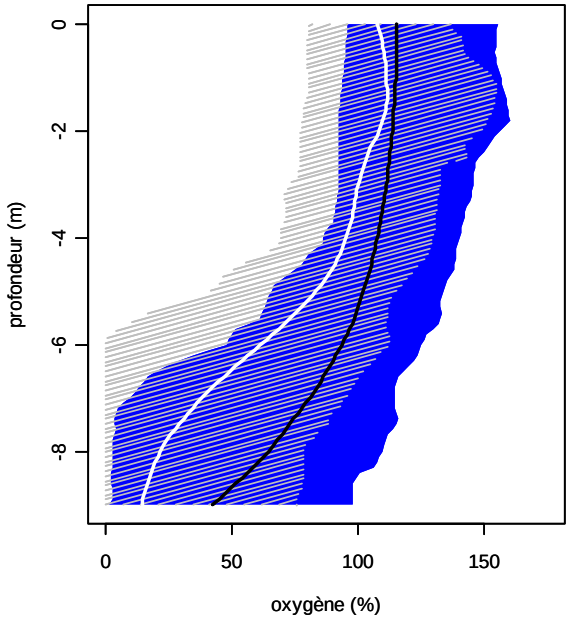


- 1994-2011
  Mesures **bimensuelles** de profils de température (°C), salinité (PSU) et oxygène dissous (% de saturation) sur 10 stations

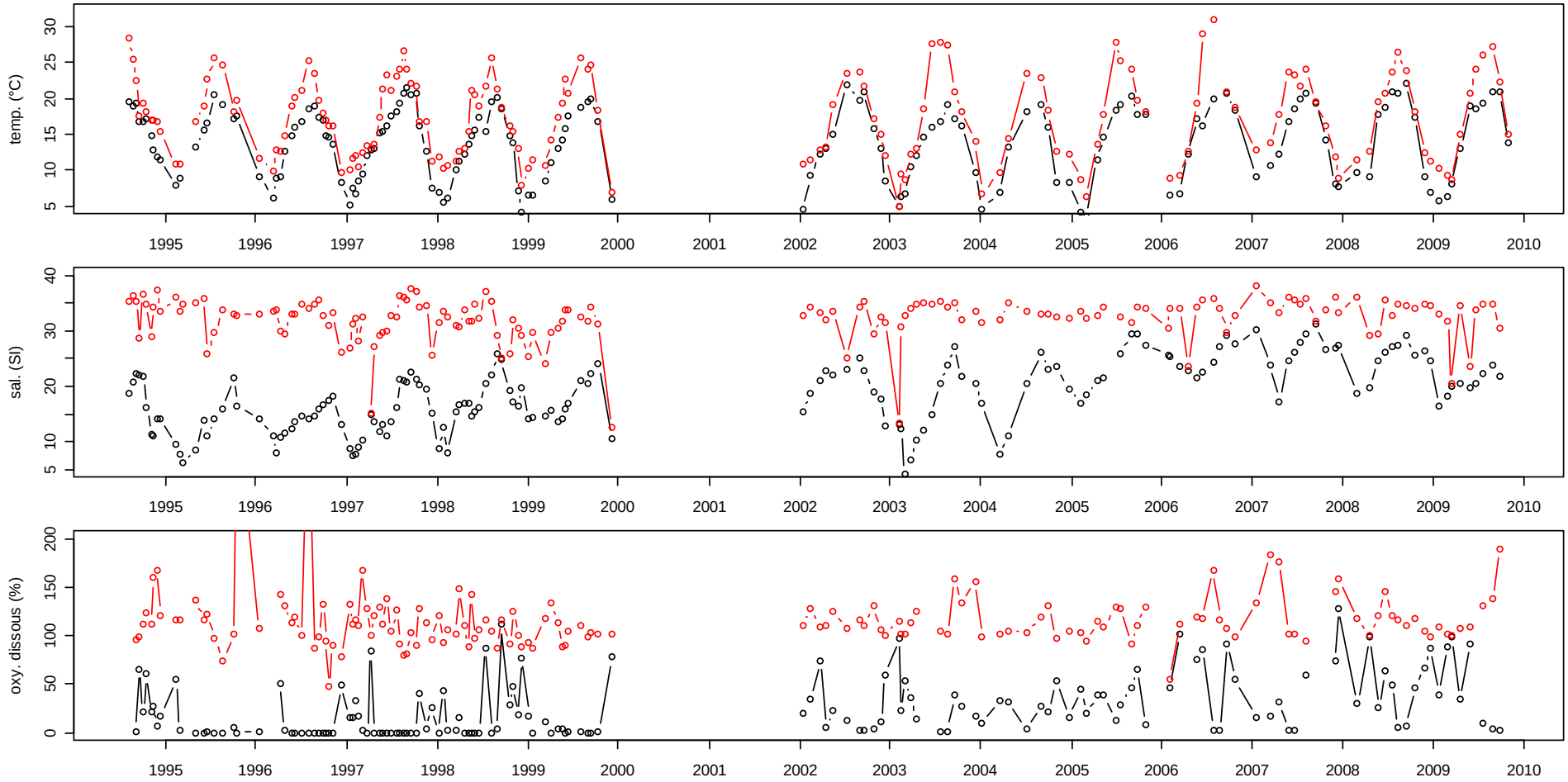
- 1994-2011
  Suivi **semestriel** de la macrofaune benthique sur ces mêmes points





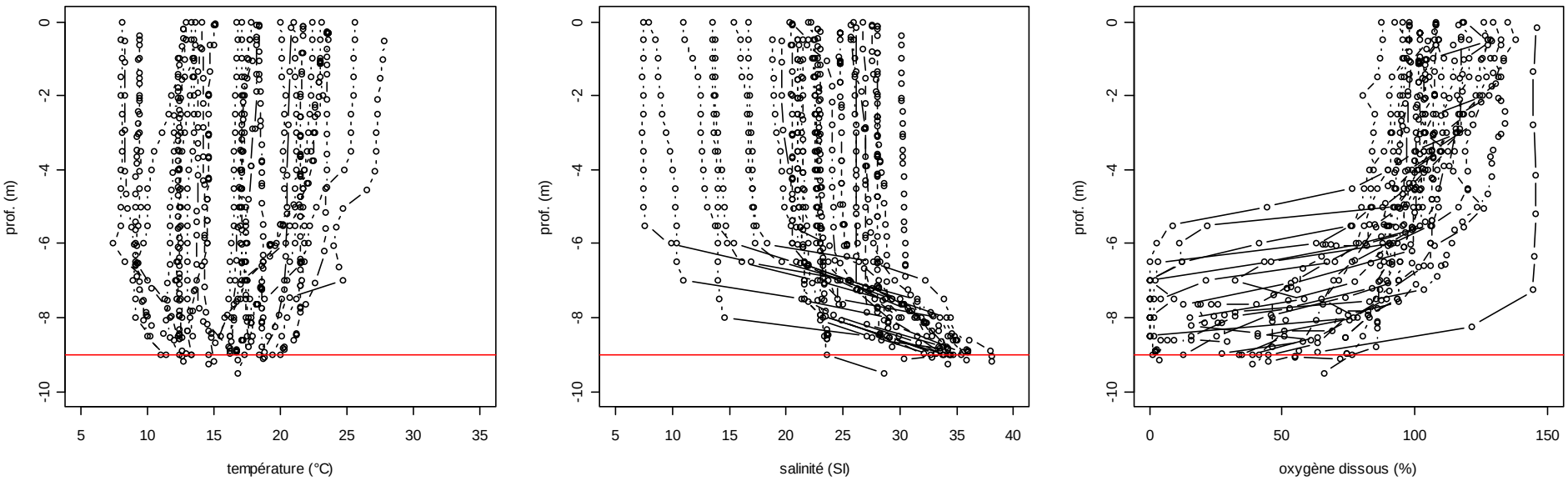


Indicateurs physiques : température, salinité et oxygène



- L'évolution globale des variables physiques montre une forte variabilité fond / surface
- Existe-t-il une différence de structure avant et après 2005?

Des données sous forme de courbes



- On dispose d'un ensemble de 170 profils de 3 variables physiques
- Ils sont échantillonnés sur la station centrale (H12) en conditions météorologiques similaires

- L'analyse de telles données nécessite d'utiliser des méthodes permettant d'intégrer simultanément :
  - la **structure fonctionnelle** de ces données
  - l'**aspect multivarié** (une observation = 3 courbes)

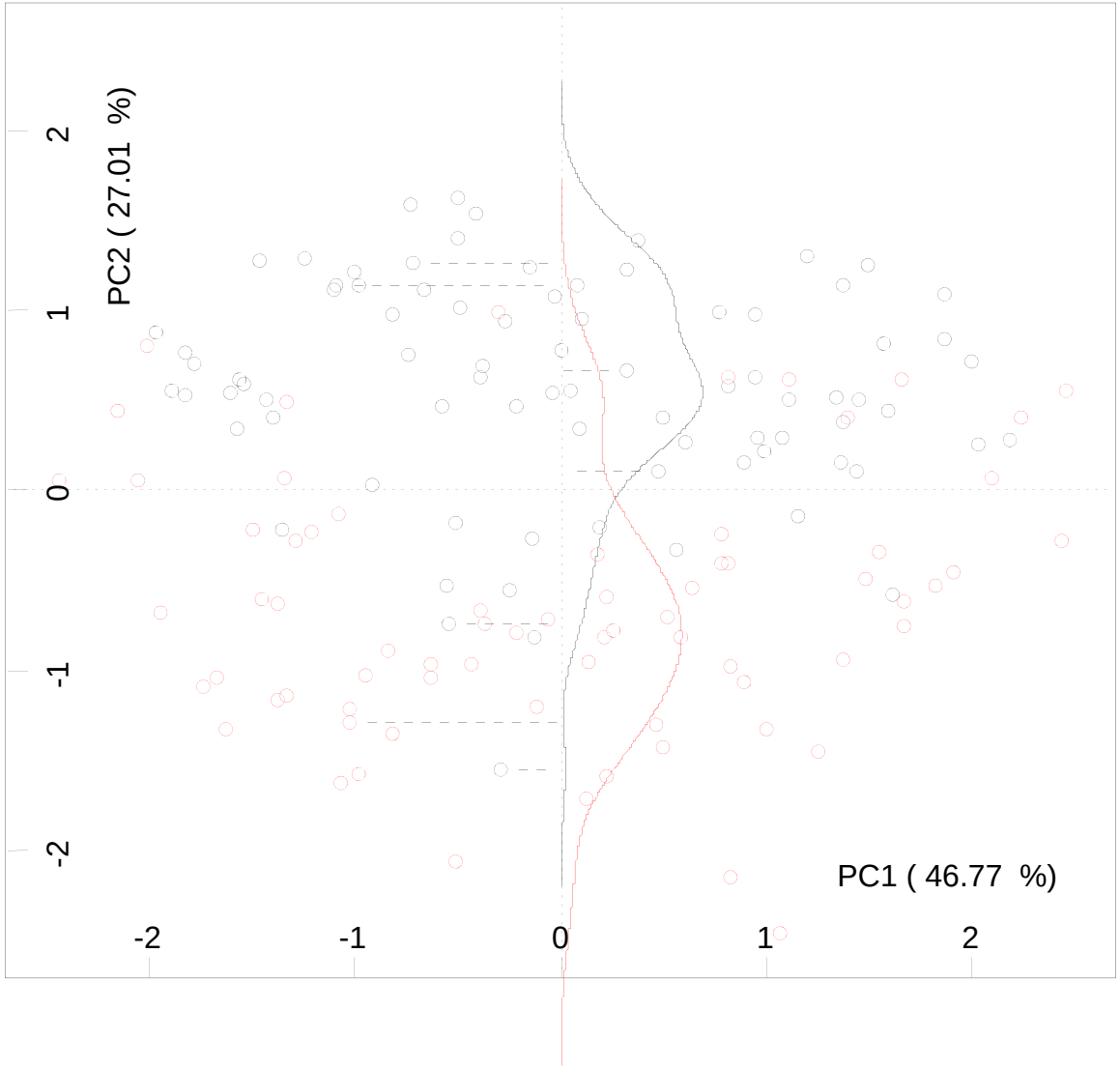
→ **Analyse en Composantes Principales Fonctionnelles**

ACP de courbes : représentation des individus

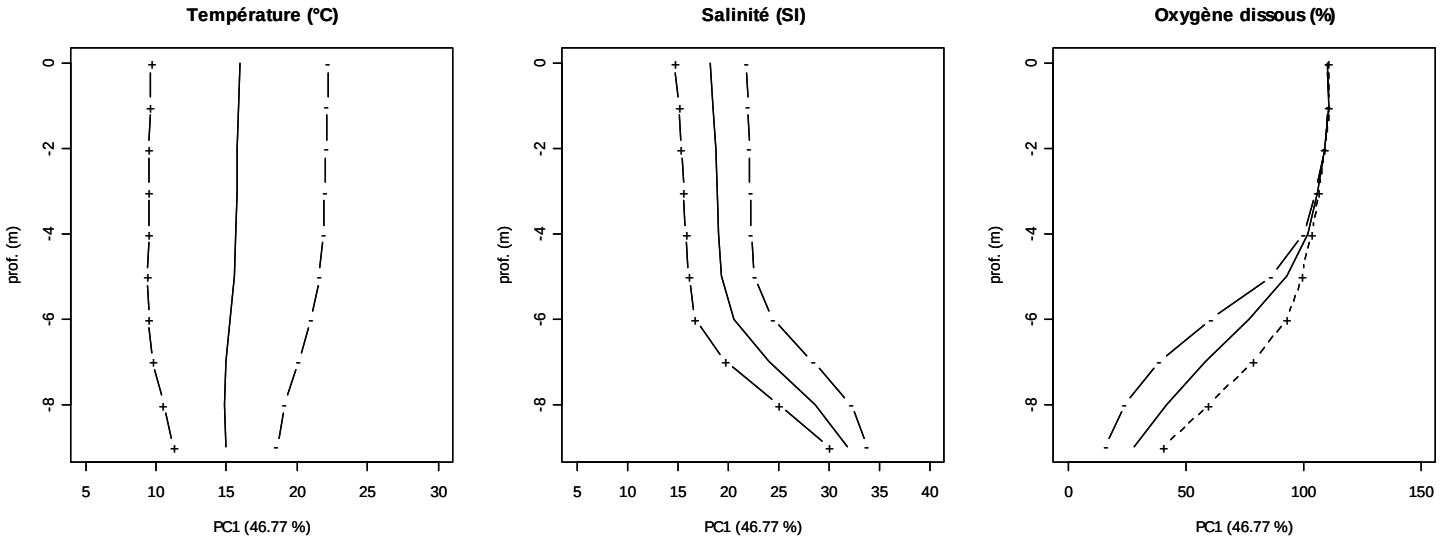
- On cherche à exprimer chaque **profil multivarié** sous la forme d'une combinaison linéaire de **facteurs explicatifs** indépendants deux à deux

- L'analyse est réalisée en donnant le **même poids** aux différents blocs de courbes (T°, S, O2)

- On cherche également à **réduire la dimension** de représentation de ces observations pour mieux comprendre les facteurs explicatifs

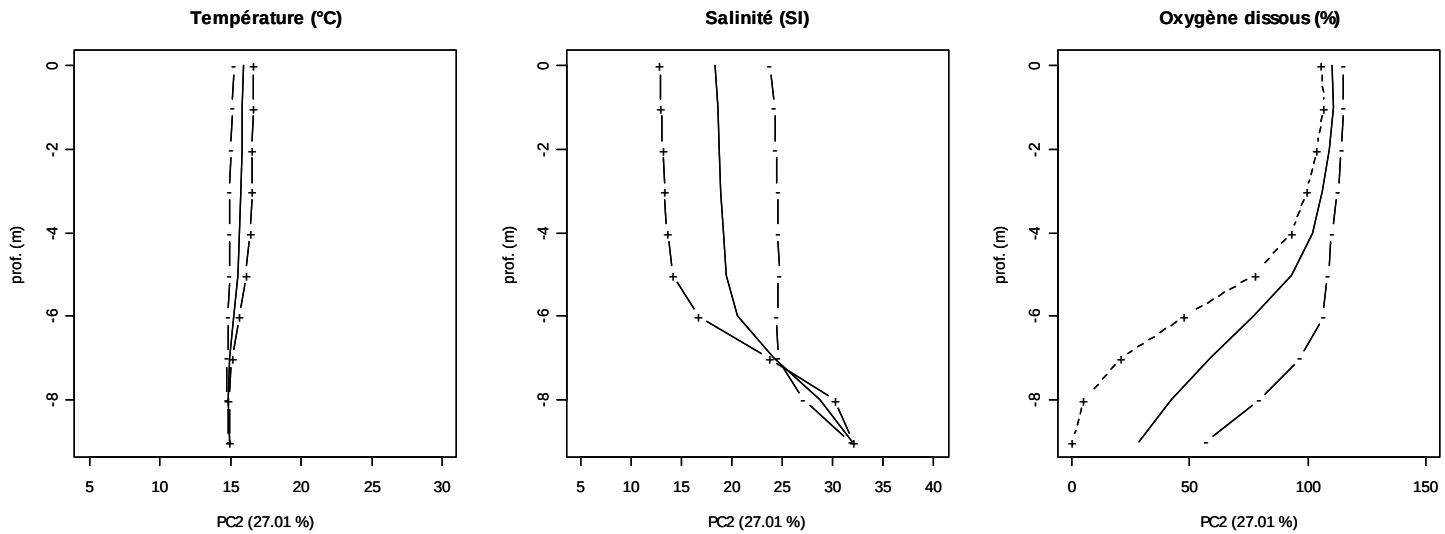


ACP de courbes : analyse des facteurs de variabilité



- Facteur saisonnier (47% de la variabilité)

- Changement des régimes de turbinage (27% de la variabilité)



Ajustement des courbes

Choix de fonctions de base

$$\hat{y}(t) = \sum_{k=1}^{K} \alpha_k \phi_k(t) = {}^t\boldsymbol{a}\boldsymbol{\phi}(t)$$

- Estimer $\hat{y}(t)$ en utilisant des données revient à estimer $\boldsymbol{a}$ avec $\hat{\boldsymbol{a}} = ({}^t\boldsymbol{\phi}\boldsymbol{\phi})^{-1}\,{}^t\boldsymbol{\phi}\boldsymbol{y}$

- Le problème d'estimation se ramène à un problème de **régression linéaire!**

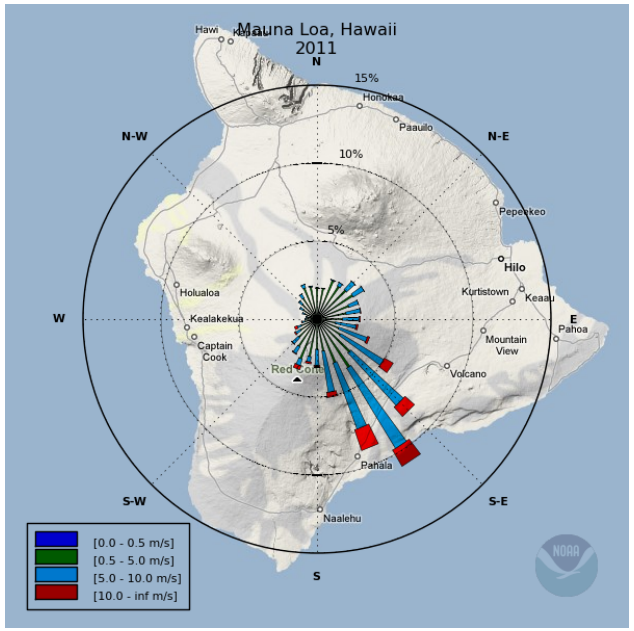
Fonction de coût

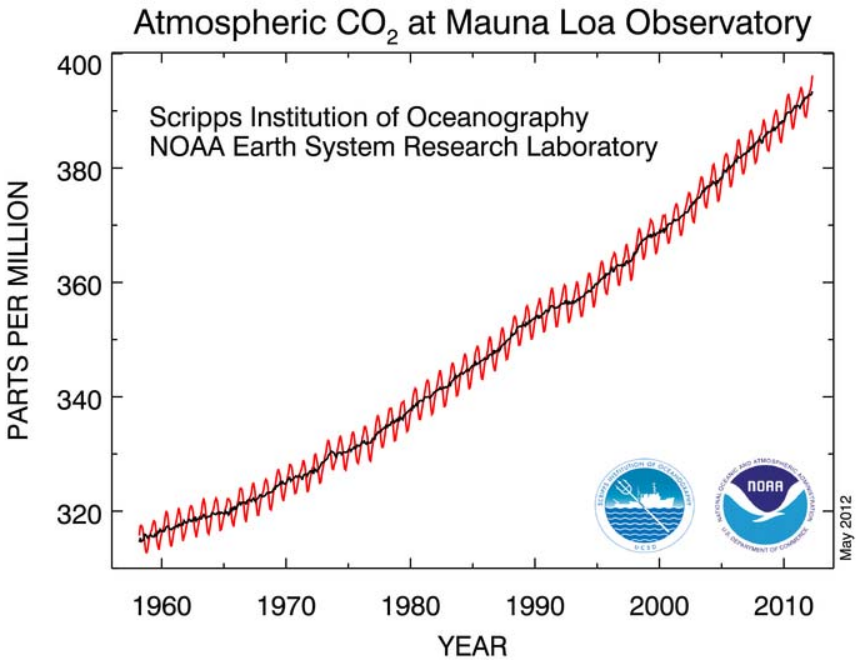
$$\mathrm{SSE}(\boldsymbol{a}) = \sum_{j=1}^{p} \left(y(t_j) - {}^t\boldsymbol{a}\boldsymbol{\phi}(t_j)\right)^2$$

- Il existe un grand choix possible de **fonctions de base** (série de Fourier, polynômes, …) dont la forme va dépendre du problème rencontré.

- Dans notre cas, on utilisera des splines cubiques d'approximation

Deuxième exemple : Mauna Loa, CO2





- **Série longue** de concentration (*ppm*) en C02 atmosphérique
- **Objectif** : séparer la tendance globale des variations saisonnières
- **Aucune hypothèse** sur la forme du modèle global

Deuxième exemple : Régression Non-paramétrique

Modèle mathématique

$$y(t) = m(t) + \varepsilon(t)$$

- Il arrive parfois que l'on ne souhaite pas donner de forme analytique particulière à $m(t)$

- On souhaitera simplement que $m(t)$ s'adapte le mieux possible aux données. Pour cela, on utilise un système de poids particulier (**fonction à noyau**)

Fonction de coût

$$SSE_t(\boldsymbol{a}) = \sum_{j=1}^{p} K\left(\frac{t - t_j}{h}\right)\left(y(t_j) - m_{\boldsymbol{a}}(t)\right)^2$$

- Le problème d'estimation se ramène à un problème de **régression linéaire** mais **conditionnellement à** t. La forme de $m(t)$ est contrôlée par h.

Troisième exemple : réponses fonctionnelles

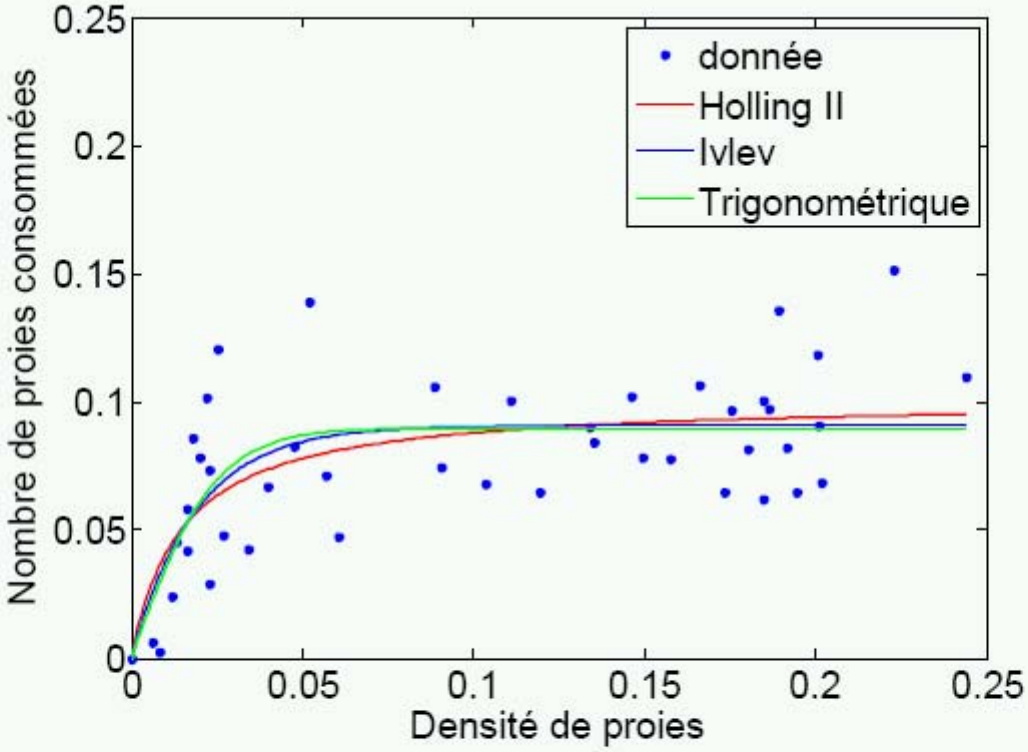


FIGURE 3.3: Données de réponse fonctionnelle des protozoaires (Canale et al., 1973), et courbes des trois réponses fonctionnelles estimées par la méthode du simplexe de Nelder-Mead. Les fonctions ont une forme très similaire, même si leur formulation mathématique est différente.

Troisième exemple : réponses fonctionnelles

Holling type 2

$$f_H(x) = \frac{ax}{1 + bx}$$

Ivlev

$$f_I(x) = c\big(1 - exp(-dx)\big)$$

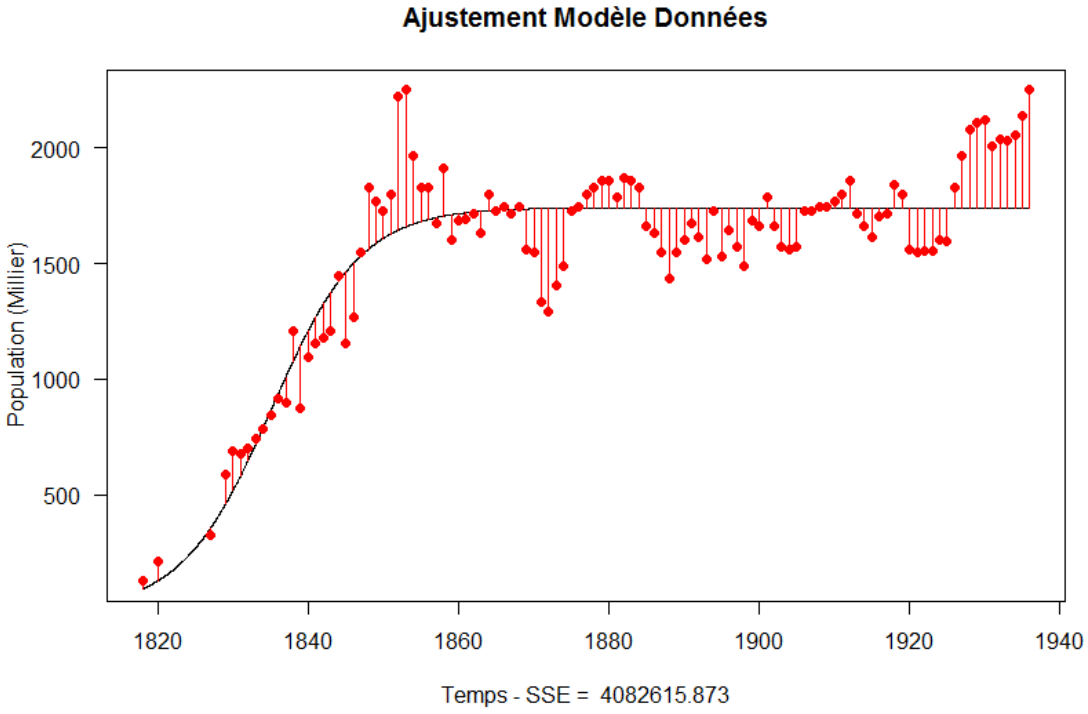
Fonction de coût

$$\mathrm{SSE}(\boldsymbol{\theta}) = \sum_{j=1}^{p} \Big(y(x_j) - f_{\boldsymbol{\theta}}(x_j)\Big)^2$$

- Dans ces deux cas, la solution qui minimise la fonction de coût n'a pas d'expression analytique.
- Il va falloir dans ce cas, utiliser des **méthodes numériques d'optimisation**
- Elles appartiennent à deux grandes familles : les **méthodes locales** et les **méthodes globales**

Quatrième exemple : Dynamique de population

Pour la qualité de leur laines, des moutons ont été introduit au début du XIXème siècle en Tasmanie. Cette introduction a largement contribué à la disparition de l'émeu de Tasmanie et du grand tigre, entre autres espèces animales.



La loi Logistique

Modèle mathématique

$$\begin{cases} \dfrac{dx}{dt} = rx\left(1 - \dfrac{x}{K}\right) \\ x(t=0) = x_0 \end{cases}$$

Solution Analytique

$$x(t) = \frac{K}{1 + \left(\dfrac{K}{x_0} - 1\right) exp(-rt)}$$

Moyennant une petite magouille…

$$x(t) = \frac{K}{1 + exp\big(-r(t - t_0)\big)}$$

La Logistique est une sigmoïde translatée!

- On peut utiliser les deux formes de la solution pour trouver les paramètres. L'utilisation de l'une ou de l'autre forme modifiera la fonction de coût.

Et si l'on ne connaît pas la solution de l'EDO?

Modèle mathématique

$$\begin{cases} \frac{dx}{dt}(t) = rx(t)\left(1 - \frac{x(t)}{K}\right) \\ x(t=0) = x_0 \end{cases}$$

- On utilise un **schéma numérique** pour trouver une estimation discrète de la solution.

- Le **schéma d'Euler** est le plus robuste. Il consiste à approcher la dérivée par son **taux de variation.**

$$\frac{x(t+\Delta t) - x(t)}{\Delta t} = \mathrm{rx}(t)\left(1 - \frac{x(t)}{K}\right), \Delta t \ll 1$$

$$x(n\Delta t) = x_0 + \Delta t \sum_{k=0}^{n-1} \mathrm{rx}(k\Delta t)\left(1 - \frac{x(k\Delta t)}{K}\right), n = 1,2,3, \dots, N$$

- Si l'on se déplace avec un pas de temps Δt, une formule de récurrence permet d'approcher la solution du modèle. Plus il y aura de **non-linéarités** dans la dynamique, plus Δt devra être choisi petit.