Quand les données sont des courbes

Philippe BESSE

Laboratoire de Statistique et Probabilités

UMR CNRS 5583

Université Paul Sabatier Toulouse III

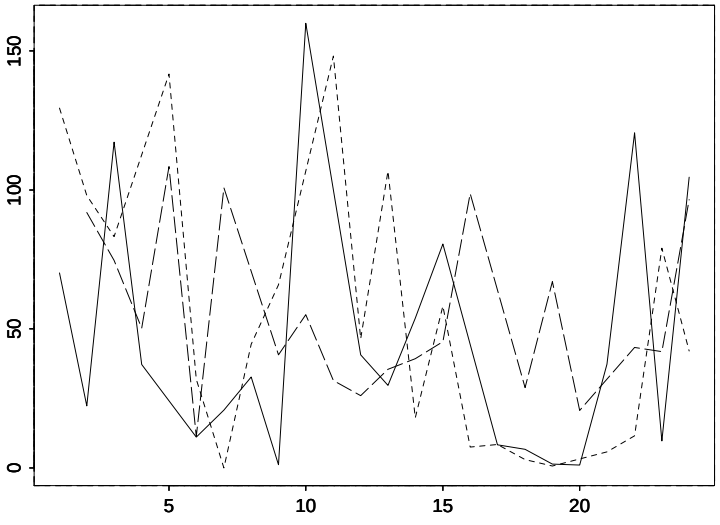
besse@math.ups-tlse.fr

www.lsp.ups-tlse.fr/Besse

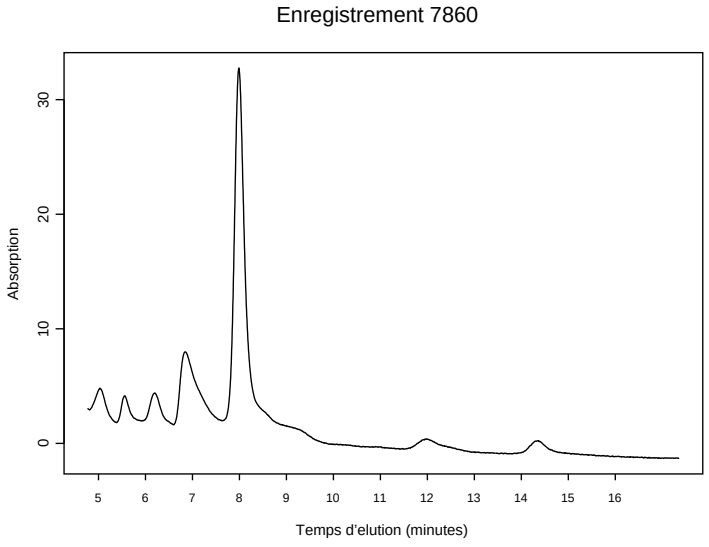
1 Introduction

- Problèmes naturellement de plus en plus fréquents.
- Historique : météorologie, chimiométrie, traitement du signal...
- Bibliographie : Deville (1974), Dauxois et pousse (1976), Besse et Ramsay (1986), Ramsay et Silverman (1997) ...
- Spécificité 1. nature fonctionnelle ou non des données,
  - base de représentation (Fourier, splines, ondelettes, fonctions propres),
  - lissage des courbes ou débruitage du signal,
  - complexité des outils mathématiques

- Spécificité 2. très forte dimensionnalité.
- Exploration (ACP) puis modélisation.
- Intérêt de la démarche sur des exemples.



Trois exemples de courbes décrivant la pluviométrie mensuelle durant 2 ans.



Chromatogramme d'un jus d'orange.

2 Exploration

2.1 ACP de courbes bruitées

2.1.1 Données

- n trajectoires z_i,observées en p d'instants $t_1, \ldots, t_p$
- n répétitions indépendantes d'un modèle de régression non-paramétrique

$$x_j = z(t_j) + \varepsilon_j \; ; \;\; E(\varepsilon_j) = 0, \;\; E(\varepsilon_j \varepsilon_k) = \sigma^2 \delta_k, \;\; j, k = 1, ..., p$$
$$a \leq t_1 < t_2 < ... < t_p \leq b$$

2.1.2 Modèle et estimation

- Estimation simultanée de n régressions non paramétriques
- contraintes de régularité et de dimension :

$$\mathbf{x}_i = \mathbf{z}_i + \boldsymbol{\varepsilon}_i; \quad i = 1, \ldots, n \qquad \text{avec} \begin{cases} \mathbb{E}(\boldsymbol{\varepsilon}_i) = 0 \text{ et } \mathbb{E}(\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i) = \sigma^2 \mathbf{I}, \\ \sigma \text{ inconnue}, \ (\sigma > 0) \\ \mathbf{x}_i \text{ indépendant de } \boldsymbol{\varepsilon}_{i'}, \ i' = 1, \ldots, n, \\ \mathbf{x}_i \in A_q \text{ p.s. et } \|\mathbf{x}_i\|_m^2 \leq c \text{ p.s.}. \end{cases}$$

$$\min_{\mathbf{z}_i, A_q} \left\{ \sum_{i=1}^{n} w_i \left(\|\mathbf{z}_i - \mathbf{x}_i\|_{\mathbf{I}}^2 + \ell \|\mathbf{z}_i\|_{\mathbf{M}}^2 \right) ; \mathbf{z} \in A_q , \ \dim A_q = q \right\}$$

PROPOSITION **1**. — *La solution du problème est donnée par :*

$$\widehat{\mathbf{z}}_i = \mathbf{A}_\ell^{1/2} \widehat{\mathbf{P}}_q \mathbf{A}_\ell^{1/2} \mathbf{x}_i + \mathbf{A}_\ell \overline{\mathbf{x}} \ , \ i = 1, \ldots, n.$$

La matrice $\widehat{\mathbf{P}}_q = \mathbf{V}_q \mathbf{V}_q'$ *est la projection orthogonale sur le sous-espace* $\widehat{E}_q$ *engendré par les* q *vecteurs propres de la matrice*

$$\mathbf{A}_\ell^{1/2} \mathbf{S} \mathbf{A}_\ell^{1/2}.$$

associés aux q *plus grandes valeurs propres.*

- Interpolation spline des valeurs de $\widehat{z}_i$.

Solution équivalente :

- D.V.S. de $(\mathbf{X}\mathbf{A}_\ell^{1/2}, \mathbf{I}, \mathbf{D})$
- donc diagonalilsation de $\mathbf{A}_\ell^{1/2}\mathbf{S}\mathbf{A}_\ell^{1/2}$.
- Trajectoires projetées sur Vect.

$$\widetilde{\mathbf{v}}_i \;=\; \mathbf{A}_\ell^{1/2}\mathbf{v}_j, \quad j = 1, \ldots, q.$$

- Trajectoires estimées $\hat{\mathbf{z}}_i$ se décomposent de manière équivalente sur la base $\mathbf{A}_\ell^{-1}$-orthonormée des $\{\widetilde{\mathbf{v}}_j\}$ par projection des données transformées $\mathbf{A}_\ell\mathbf{x}_i$:

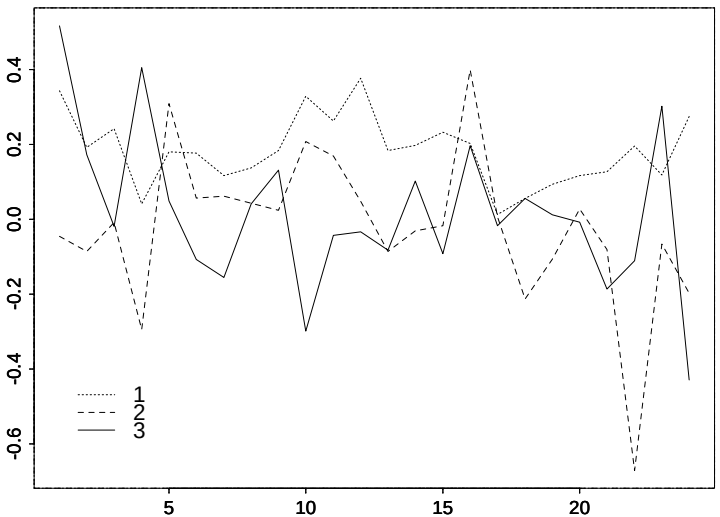
$$\hat{\mathbf{z}}_i \;=\; \mathbf{A}_\ell\bar{\mathbf{x}} + \sum_{j=1}^{q} \langle \widetilde{\mathbf{v}}_j, \mathbf{A}_\ell\mathbf{x}_i \rangle_{\mathbf{A}_\ell^{-1}} \widetilde{\mathbf{v}}_j.$$

- Dimension et paramètre de lissage : optimiser un critère de stabilité.

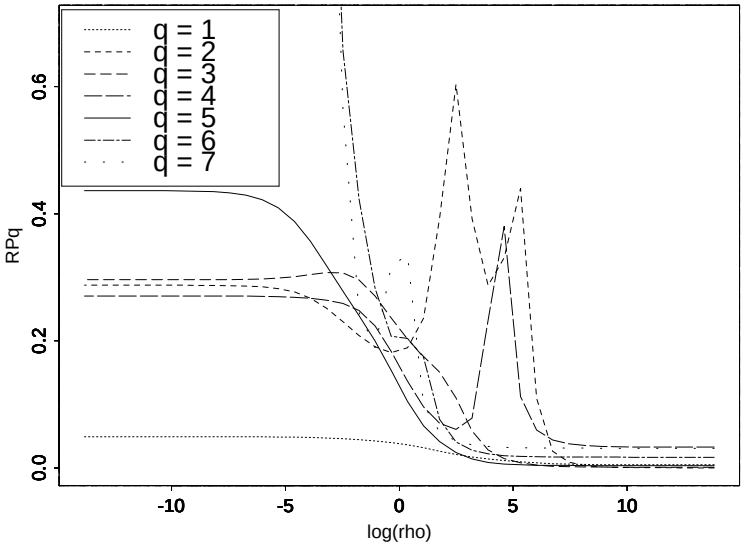
2.2 Exemples : ACP de séries climatiques

2.2.1 ACP des précipitations

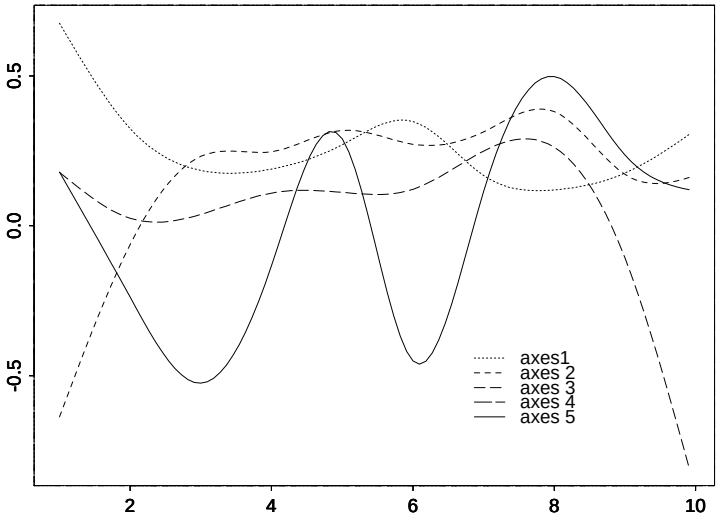
- Données : 130 courbes de précipitations mensuelles sur deux ans.



Trois premières fonctions propres de l'ACP classique.



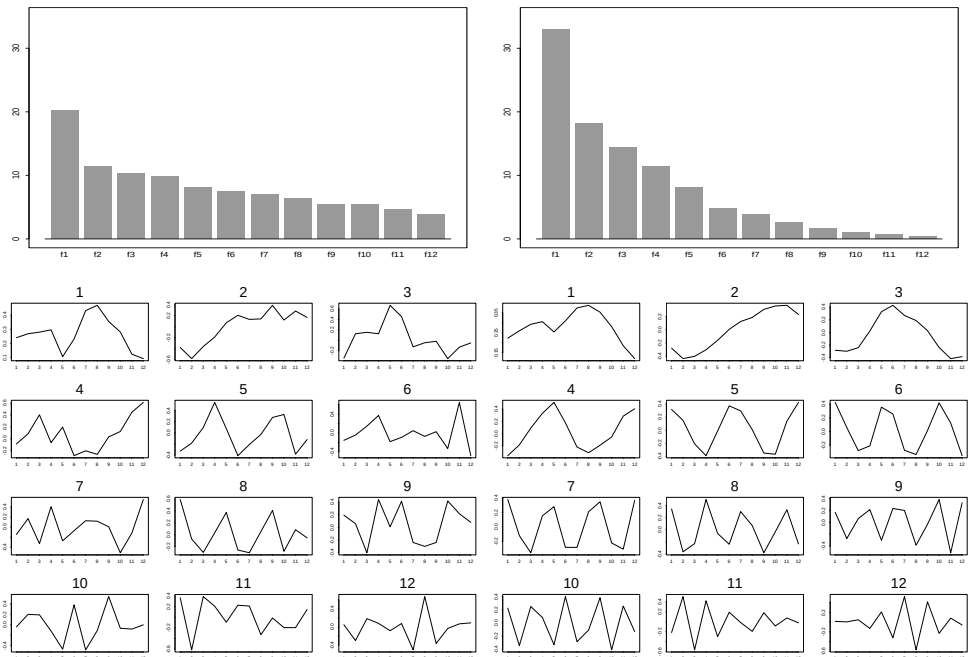
Stabilité du sous-espace de projection.

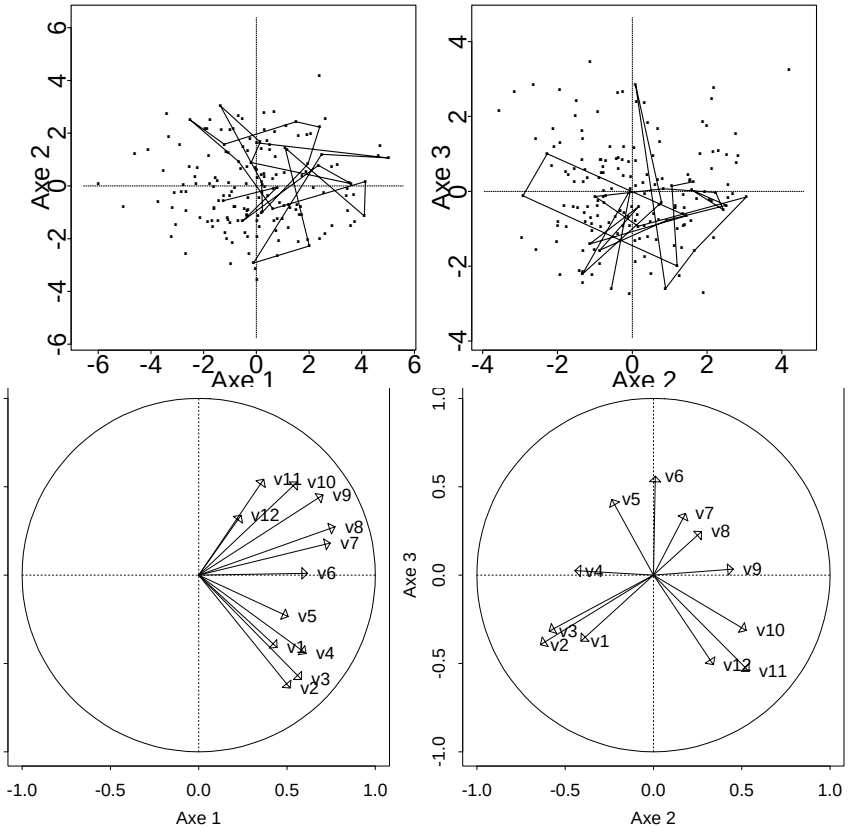


Cinq premières composantes principales.

2.2.2 ACP de températures

- Données : températures centrales en Angleterre depuis 1659.
- Moyennes mensuelles de plusieurs villes.




Axe 2
Axe 1
Axe 3
Axe 2
v11
v10
v9
v12
v8
v7
v6
v5
v1
v4
v3
v2
Axe 1
Axe 3
Axe 2

3 Modélisation de données fonctionnelles

3.1 Introduction

- Nature fonctionnelle des données comme pour exploration.
- Dimensionnalité : colinéarité, conditionnement, identifiabilité.
  - Régression sur composantes principales (El Niño).
  - Régularisation comme en *ridge*.
  - Régression PLS.
  - Sélection de fonctions de base (chromatogrammes).

3.2 Modèle linéaire et extensions

Ramsay et Dalzell, 1991 ; Hastie et Mallows, 1993).
soit (X, Y) un couple de variables aléatoires à valeurs dans $H \times \mathbb{R}$

$$I\!E[Y\,|X = x] \;= \int_0^1 \psi(t)x(t)\;dt, \qquad x \in H,$$

$\psi \in H$ coefficient de régression fonctionnel à estimer.

3.2.1 Modèle linéaire fonctionnel

- Γ l'opérateur de covariance de X
- Δ l'opérateur de covariance croisée de (X, Y).

$$\Delta x \;=\; \int_0^1 \mathbb{E}\left[X(t)Y\right] x(t)dt,\; x \in H.$$

On vérifie :

$$\Delta \;=\; \Psi\Gamma,$$

où $\Psi(x) =< \psi, x >$.
Problème : inverser Γ
Solutions :

- Dans l'espace engendré par les q premiers vecteurs propres (ACP).
- Régularisation comme en régression *ridge*.

Soit un échantillon $(\mathbf{x}_i, y_i),\; i = 1, \ldots, n$ où $\mathbf{x}_i = (x_i(t_j),\; j = 1, \ldots, p) \in \mathbb{R}^p$ obtenu par discrétisation de la fonction ou courbe x_i.

3.2.2 Régression sur composantes principales

$$\widehat{\boldsymbol{\psi}}(t) \;=\; \sum_{j=1}^{q} \beta_j \widehat{v}_j(t).$$

- Les composantes principales $\langle \widehat{z}_i, \widehat{v}_j \rangle_{\mathbf{L^2}} = \mathbf{z}_i' \mathbf{N} \mathbf{v}_j$ sont les variables explicatives

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^q} \sum_{j=1}^{n} w_i \left(Y_i - \sum_{j=1}^{q} \beta_j \mathbf{z}_i' \mathbf{N} \mathbf{v}_j \right)^2 .$$

3.2.3 Approche par pénalisation

Hastie et Mallows (1993), Marx et Eilers (1999)

$$\min_{\boldsymbol{\psi}} \sum_{i=1}^{n} w_i \left(Y_i - \boldsymbol{\psi}'\mathbf{N}\mathbf{x}_i\right)^2 + \ell \left\|\psi\right\|_2^2,$$

$$\min_{\boldsymbol{\psi}} \boldsymbol{\psi}'\mathbf{N}\mathbf{S}\mathbf{N}\boldsymbol{\psi} - 2\,\boldsymbol{\psi}'\mathbf{N}\mathbf{Z}'\mathbf{D}\mathbf{Y} + \ell\,\boldsymbol{\psi}'\mathbf{M}\boldsymbol{\psi}.$$

La solution s'écrit alors

$$\widehat{\boldsymbol{\psi}} \;=\; \left(\mathbf{N}\mathbf{S}\mathbf{N} + \ell\mathbf{M}\right)^{-1}\mathbf{N}\mathbf{Z}'\mathbf{D}\mathbf{Y}.$$

- Choix du paramètre de lissage.

3.2.4 Modèle linéaire général et autres extensions

Marx et Eilers (1999)

$$\mathbb{E}[Y|X = x] = g^{-1}\left(< \psi, x >\right), \qquad x \in H,$$

où g est la fonction lien de type logit, log, ...

- Vraisemblance pénalisée :

$$\max_{\boldsymbol{\psi}} \sum_{i=1}^{n} \log \mathcal{L}(y_i, \mathbf{x}_i, \boldsymbol{\psi}) - \ell \left\|\psi\right\|_2^2$$

3.3 Prévision fonctionnelle

3.3.1 Problème

Bosq (2000), Besse et Cardot (1996) et Besse et coll. (2000).

- $(X_t)_{t\in\mathbb{R}}$ découpé en intervalles δ,
- $(Z_n)_{n\in\mathbb{Z}}$, à valeurs dans $H = L^2[0,1]$

3.3.2 Modèle ARH(1)

- $(Z_i)_{i\in\mathbb{Z}}$ processus auto-régressif hilbertien du premier ordre d'espérance $a \in H$ et d'opérateur d'auto-corrélation ρ :
- $\forall i \in \mathbb{Z}, \quad Z_i - a = \rho\,(Z_{i-1} - a) + \epsilon_i$.
- $\sum_{n\geq 1} \|\rho^n\| < +\infty$.
- $\{\epsilon_i\}$ est centré i.i.d. dans H de variance finie,
- $\mathbb{E}\,\|\epsilon_i\|_H^2 = \sigma^2 < +\infty$.
- $\mathbb{E}(Z_{i+1}|Z_i, Z_{i-1}, \ldots) - a = \rho(Z_i - a)\ , \quad i \in \mathbb{Z}$.

$\Gamma = \mathbb{E}((Y_i - a) \otimes (Y_i - a))$ et $\Delta = \mathbb{E}((Y_i - a) \otimes (Y_{i+1} - a))$ vérifient :

$$\Delta = \rho\Gamma.$$

- Inverser Γ pour estimer ρ.
- Régression sur composantes principales.
- $\widehat{\rho}$ approché dans la base engendrée par les premières fonctions propres de Γ :

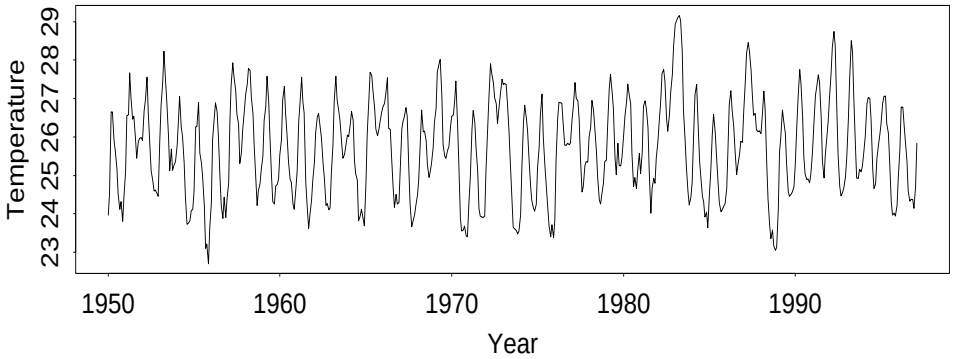
$$\widehat{\rho}(s,t) = \sum_{j=1}^{q}\sum_{j'=1}^{q} \widehat{\beta}_{jj'}\widehat{v}_j(t)\widehat{v}_{j'}(s).$$

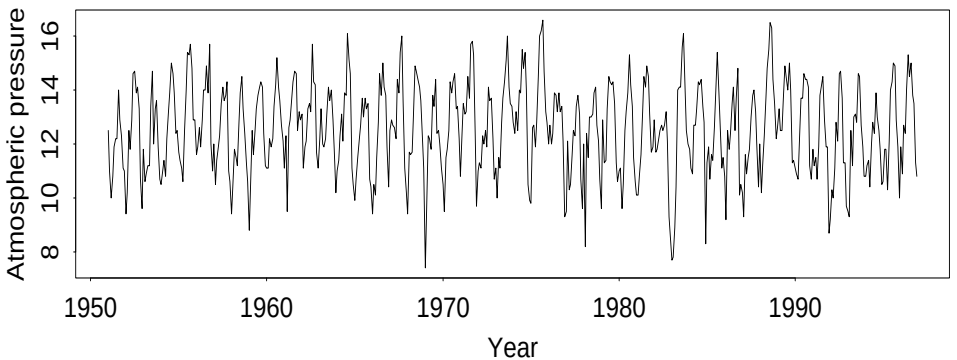
$$\min_{\hat{z}_i \in H_q} \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{p}\sum_{j=1}^{p}\left(z_i(t_j)-\widehat{z}_i(t_j)\right)^2 + \ell\left\|D^2\widehat{z}_i\right\|_{L^2}^2\right)$$

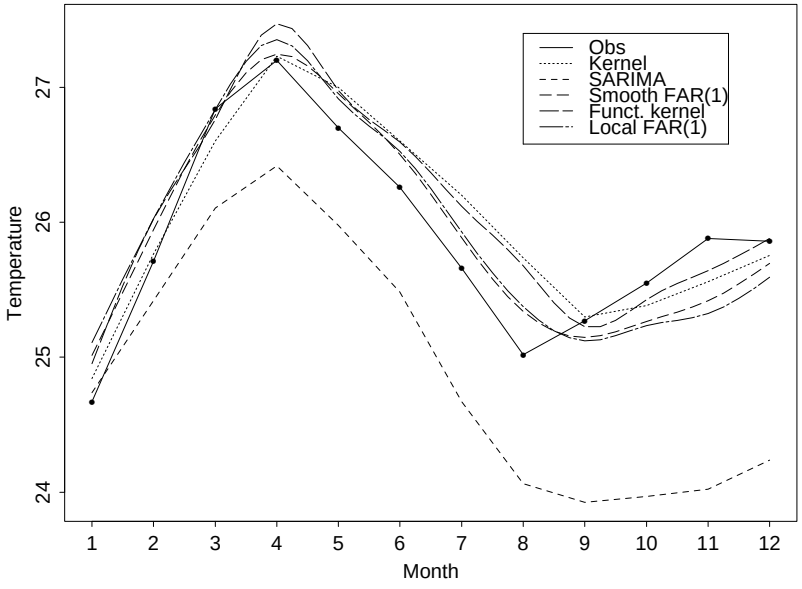
$$\widehat{\mathbf{\Gamma}}_{q,\ell} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\mathbf{x}}_i\widehat{\mathbf{x}}_i'\mathbf{N}, \quad \widehat{\mathbf{\Delta}}_{q,\ell} = \frac{1}{n-1}\sum_{i=1}^{n-1}\widehat{\mathbf{x}}_{i+1}\widehat{\mathbf{x}}_i'\mathbf{N}.$$

- Soit $\widehat{\mathbf{\Gamma}}_{q,\ell}^{-1}$, l'inverse généralisé de $\widehat{\mathbf{\Gamma}}_{q,\ell}$: $\widehat{\boldsymbol{\rho}}_{q,\ell} = \widehat{\mathbf{\Gamma}}_{q,\ell}^{-1}\,\widehat{\mathbf{\Delta}}_{q,\ell}$.
- Prévision de $\mathbf{z}_{n+1}$: $\widehat{\mathbf{z}}_{n+1} = \widehat{\boldsymbol{\rho}}_{q,\ell}\widehat{\mathbf{x}}_n + \mathbf{A}_\ell\bar{\mathbf{z}}$.

3.3.3 Prévision d'El Nino







Comparaison de l'année 1986 de El Niño avec différentes prévisions.

Comparaisons des erreurs quadratiques et erreurs absolues relatives de prévision sur la période 1987-1996.

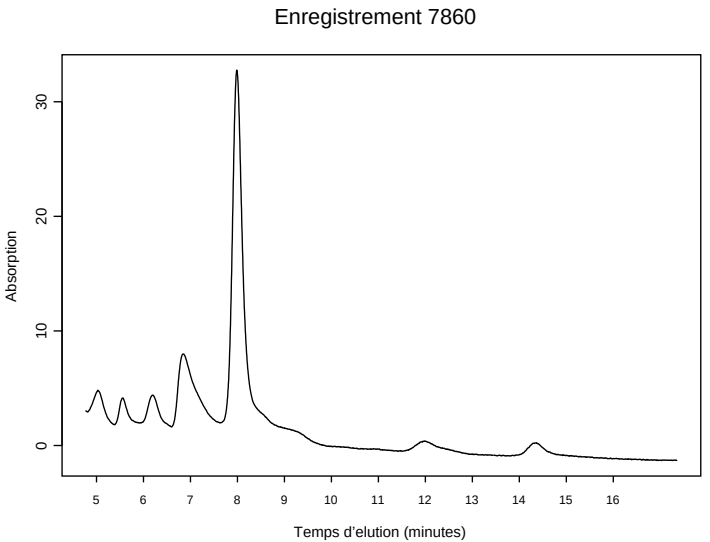
	El Niño index		S. Osc. index	
Prédicteur	EQMP	EAMR	EQMP	EAMR
Climat.	0.73	2.5 %	0.91	6.3 %
SARIMA	1.45	3.7 %	0.95	6.2 %
Noyau	0.60	2.3%	0.87	6.1%
Noyau fonct.	0.58	2.2%	0.82	6.0%
ARH(1) lisse	0.55	2.3%	**0.78**	**5.8%**
ARH(1) local	**0.53**	**2.2%**	0.82	5.8%

3.4 Modélisation par arbre

3.4.1 Objectif

Besse et Farkas (2003)

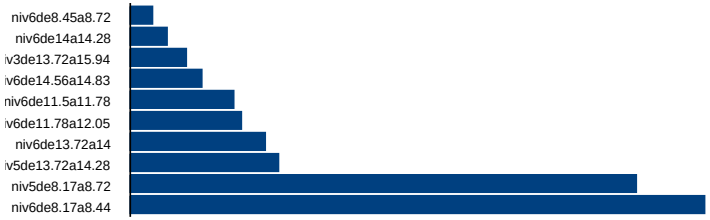
- Discriminer des jus d'orange présence/absence de pasteurisation.



- Quelle base : Fourier, fonctions propres, splines ou ondelettes ?
- Quelles méthode de discrimination : analyse discriminante, régression logistique, k plus proches voisins, réseaux de neurones, arbres binaires (CART)...
- Dimensionnalité : 6912 mesures pour 64 courbes
- Problème de localisation du pic : ondelettes.
- Comparaison des méthodes de Sélection :
- Procédure de comparaison (B fois)

1. extraction aléatoire d'un échantillon test,
2. estimation du modèle sur l'autre partie (apprentissage),
3. optimisation du modèle par validation croisée (sélection de variable ou élagage),
4. estimation de l'erreur sur l'échantillon test.

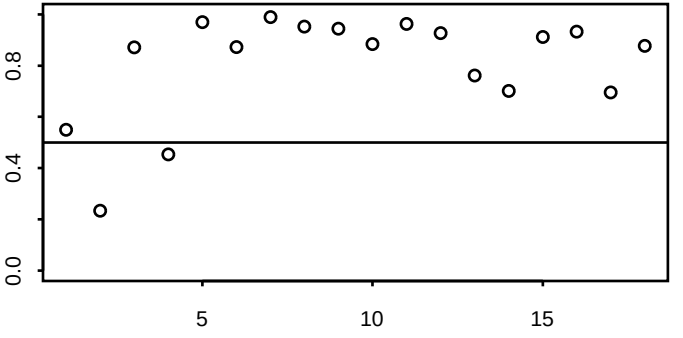
- Échec de la régression logistique.
- Grande instabilité d'un arbre binaire seul.
- Forêt aléatoire avec graphe d'importance :



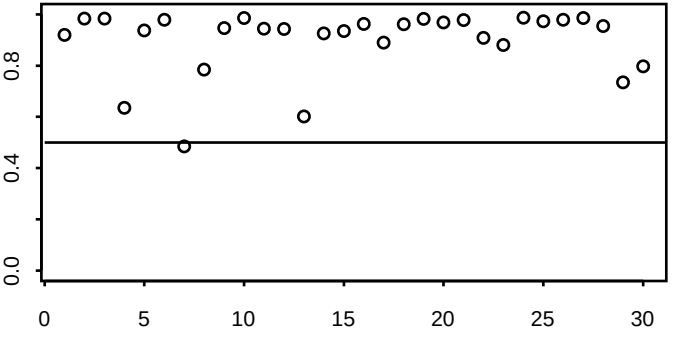
Importance des variables.

jeuVC n° 2 , mtry= 19 , ntrees= 1000

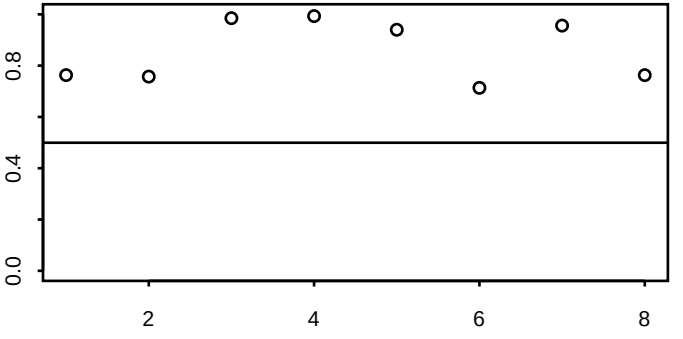
u.s. Train NonPast.: 11.1/100 mc



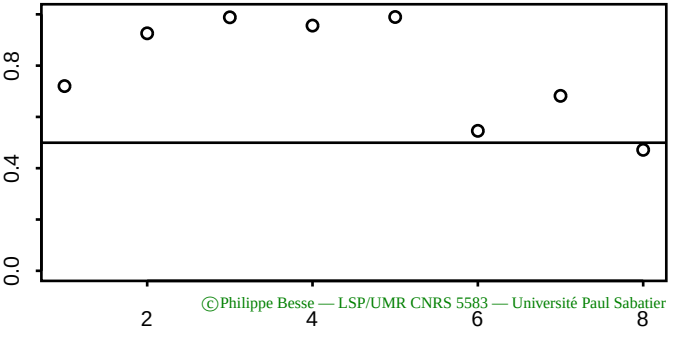
u.s. Train Pasteurises: 3.3/100 mc



u.s. Test NonPast.: 0/100 mc



u.s. Test Pasteurises: 12.5/100 mc



3.5 Conclusion

- Coût/intérêt de la démarche.

1. Plongementdans un espace fonctionnel (splines, Fourier, ondelettes) adaptée au problème posé : hypothèse ou non de régularité par opposition à la présence jugée significativite de singularités.
2. Recaler les courbes par une transformation non linéaire de l'échelle des temps (*curve registration*).
3. Dimensionnalité et choix de la technique de modélisation ; régularisation (lissage), ACP ou sélection.
4. Optimisation conjointe des paramètres.

Table des matières