Modélisation des brises de mer

J. Josse [1] & Eric Matzner-Løber [2]

[1]Lab. de mathématiques appliquées, Agrocampus Rennes,

[2]IRMAR, Université Rennes 2.

Brest

Brest, 14 février 2008

Présentation du sujet

Nous devons prévoir aujourd'hui à 16 TUC le pic de d'O_3 de demain. L'O_3 est un polluant secondaire :

$$NO_2 + O_2 \rightarrow NO + O_3$$

$$NO + COV \rightarrow NO_2 + ?$$

modèle physico-chimique .... modèle statistique.
Mission d'Air Breizh : mesure, analyse, prévision.

Modèles

- Modélisation **globale** ou **Macro-échelle** : **Prev'Air**
  - modèle déterministe
  - récemment correction statistique
- Modélisation **locale** : modèle statistique.

Les données initiales

- 1^{er} janvier 1999 au 31 décembre 2007.
- Données concentration de polluant.
  - Ozone (en $\mu g/m^3$) toutes les heures
    Selection du maximum d'ozone $\Rightarrow$ maxO3.
- Phénomène de persistance: maxO3v.
- Données météo (horaire).
  - La température (Celsius)
  - La nébulosité (0-8)
  - Le vent : vitesse et direction
  - Création de vents projetés (V_x = vitesse $\times$ sin(direction)).
  - Les précipitations (en mm)
  - Phénomène de lessivage.
  - La radiation et la pression (ne peuvent être prévues)

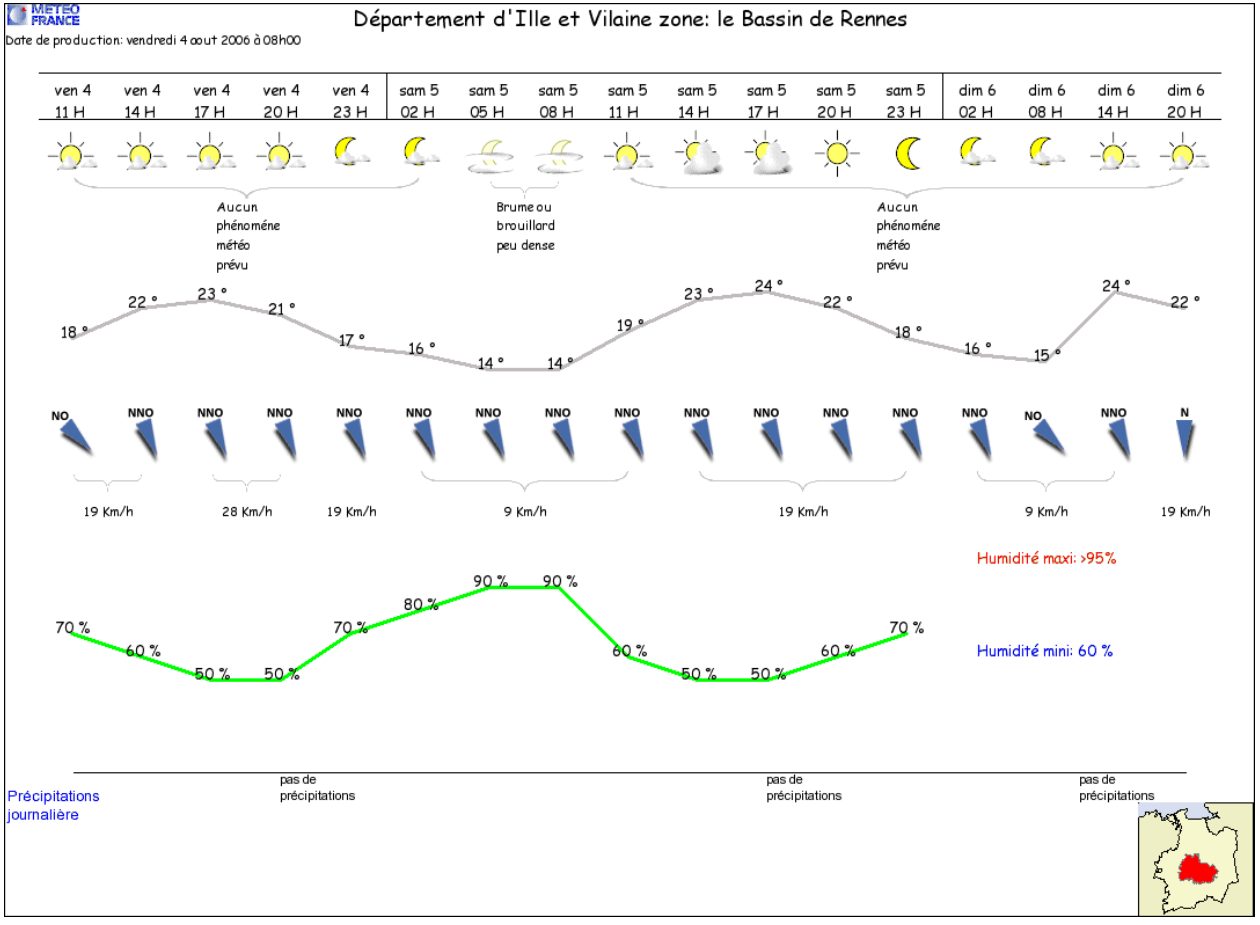
Les données initiales

- 1^{er} janvier 1999 au 31 décembre 2007.
- Données concentration de polluant.
  - Ozone (en $\mu g/m^3$) toutes les heures
    Selection du maximum d'ozone $\Rightarrow$ maxO3.
- Phénomène de persistance: maxO3v.
- Données météo (horaire).
  - La température (Celsius)
  - La nébulosité (0-8)
  - Le vent : vitesse et direction
  - Création de vents projetés (V_x = vitesse $\times$ sin(direction)).
  - Les précipitations (en mm)
  - Phénomène de lessivage.
  - La radiation et la pression (ne peuvent être prévues)

Les données initiales

- 1^{er} janvier 1999 au 31 décembre 2007.
- Données concentration de polluant.
  - Ozone (en $\mu g/m^3$) toutes les heures
    Selection du maximum d'ozone $\Rightarrow$ maxO3.
- Phénomène de persistance: maxO3v.
- Données météo (horaire).
  - La température (Celsius)
  - La nébulosité (0-8)
  - Le vent : vitesse et direction
  - Création de vents projetés (V_x = vitesse $\times$ sin(direction)).
  - Les précipitations (en mm)
  - Phénomène de lessivage.
  - La radiation et la pression (ne peuvent être prévues)

Atmogramme



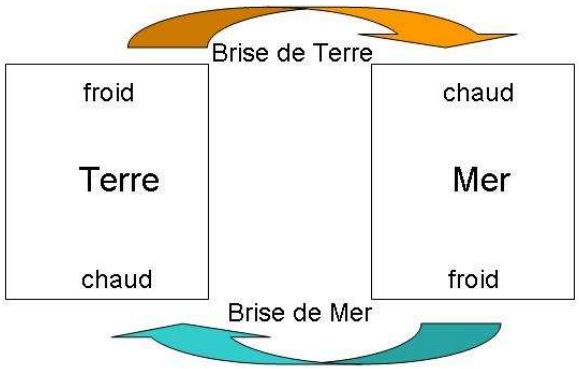
→ Selection des variables à des heures ponctuelles.

Les données de travail

- **maxO3** : Concentration maximale du jour.
- maxO3v : Concentration maximale de la veille.
- T6, T9, T12, T15, T18 : Température à des heures ponctuelles 6 heures, 9 heures, 12 heures, 15 heures et 18 heures.
- Ne6, Ne9, Ne12, Ne15, Ne18 : Nébulosité à des heures ponctuelles.
- Vx6, Vx9, Vx12, Vx15, Vx18 : Projection du vent sur l'axe E-O à des heures ponctuelles.
- Vy6, Vy9, Vy12, Vy15, Vy18 : Projection du vent sur l'axe N-S à des heures ponctuelles.
- Précipitations **Création Base pluie et Base sans pluie**

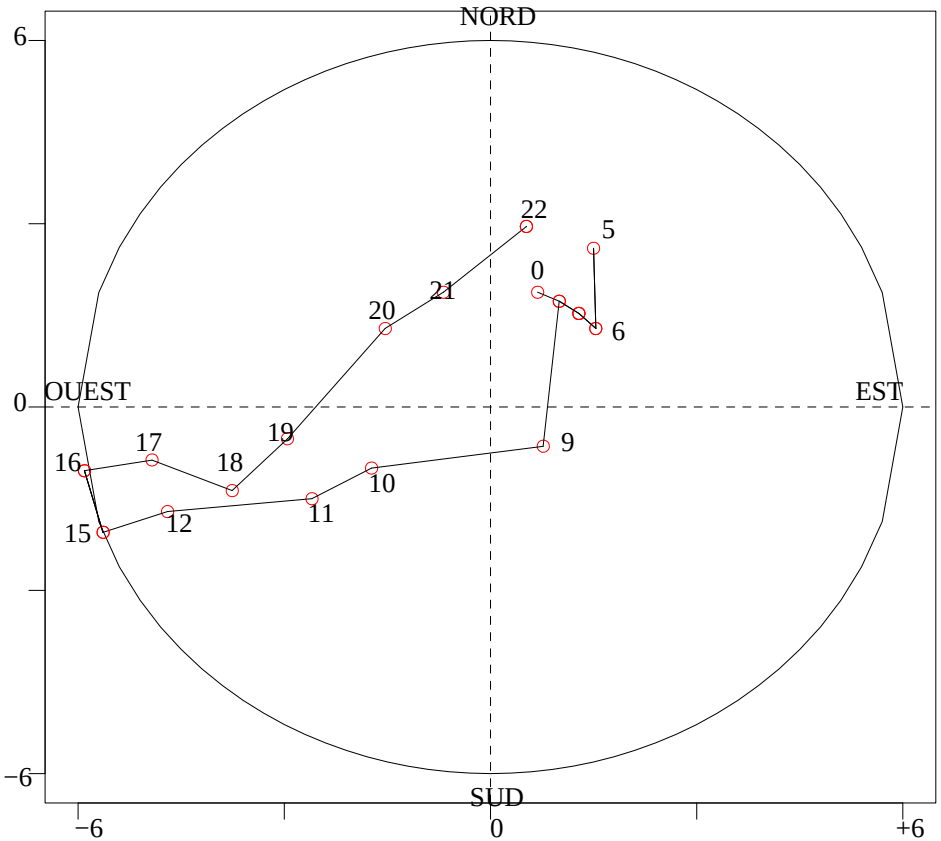
Le problème des villes côtières: la brise de mer

- Gradient de pression généré par la différence de température entre la masse d'air surmontant le continent et celle de la mer.



- Dilution, concentration, circulation des polluants $\longrightarrow$ complexifier les épisodes de pollution.
- Caractéristique de la brise de mer :
  - Changement diurne de la direction du vent
  - Renforcement de la vitesse du vent
  - Diminution de la température, augmentation de l'humidité...

La brise de mer



Création de variable

- Vitesses cumulées...
- Gradient de vitesses, d'angle...
- Calcul des différences de vents...

$\Rightarrow$ Création de 113 variables explicatives.

Deux grandes familles d'approches

$$Y = r(x) + \varepsilon$$

- Deux grandes familles d'approches
  - Approche paramétrique
    -Hypothèse restrictive sur la fonction de régression.
    -Nombre fini de paramètres à estimer.
  - Approche non paramétrique
    -Aucune hypothèse a priori sur l'allure de la fonction de régression.

- Principales méthodes de régression non paramétriques : noyaux, splines...

Modèles additifs

- $Y = r(X_1, ..., X_p) + \varepsilon$
- On va simplifier la forme de la fonction de régression r en supposant que cette fonction est additive:

$$Y = \alpha_0 + \sum_{j=1}^{p} r_j(X_j) + \varepsilon$$

- Genéralise les modèles de régression linéaire.
- Permet une interprétation de l'effet marginal de chacune des variables sur la fonction de régression.
- **Méthode d'estimation :** Intégration marginale, Backfitting.

CART (Breiman): régression par arbre

- La fonction de régression est la suivante:

$$r(x) = \sum_{j=1}^{p} c_j \mathbb{1}_{\{x \in F_j\}}$$

- Avantages:
  - simple, facile à comprendre et à interpréter
  - œuvre simultanément sur le plan descriptif et décisionnel
- Inconvénient: manque de stabilité $\rightarrow$ **Agrégation par bootstrap**.

Forêts aléatoires

- Amélioration du Bagging
- Collection d'arbres, h_k $k = 1, .., K$
  - Construction d'un échantillon bootstrapé à partir de $(X_1, Y_1), ..., (X_n, Y_n)$
  - Tirage aléatoire d'un sous-ensemble de prédicteurs q
  - Construction d'un arbre avec les q variables (CART)
- Prédiction finale:

$$h(X) = \frac{1}{K} \sum_{k=1}^{K} h_k(X).$$

L_2-Boosting et réduction de biais

(X_i, Y_i) n paires d'observations

$$\begin{aligned} Y_i &= m(X_i) + \varepsilon_i \\ Y &= m + \varepsilon. \end{aligned}$$

Solution possible : les lisseurs

$$\hat{Y} = \hat{m} = S_\lambda Y,$$

où S_λ est la matrice de lissage.

L_2-Boosting et réduction de biais

(X_i, Y_i) n paires d'observations

$$Y_i = m(X_i) + \varepsilon_i$$

$$Y = m + \varepsilon.$$

Solution possible : les lisseurs

$$\hat{Y} = \hat{m} = S_\lambda Y,$$

où S_λ est la matrice de lissage.

Lisseurs classiques

- moyenne mobile $S_{ij} = 1/\text{nbr de } X \text{ dans le voisinage}$
- lisseur par cellule $S_{ij} = 1/\text{nbr de } X \text{ dans la cellule}$
- lisseur à noyau $S_{ij} = K_h(X_i - X_j)/\sum_l K_h(X_i - X_l)$
- K-pp voisins $S_{ij} = 1/K$ si $X_j \in Kpp(X_i)$
- spline de régression $S = B(B'B)^{-1}B'$
- spline de lissage $S = N(N'N + \lambda\Omega_N)^{-1}N'$

Solution

Considérons le modèle

$$Y \quad = \quad m + \varepsilon \quad \mathbb{E}\varepsilon = 0 \quad \mathrm{Var}(\varepsilon) = \sigma^2 I$$

et une matrice de lissage S_λ.

L'estimateur proposé est

$$\hat{m}_1 \quad = \quad S_\lambda Y.$$

Biais et variance

$$\hat{m}_1 \quad = \quad S_\lambda Y.$$

Le biais

$$B(\widehat{m}_1) \quad = \quad \mathbb{E}[\hat{m}_1|X] - m = (S_\lambda - I)m.$$

La variance

$$V(\widehat{m}_1|X) \quad = \quad S_\lambda S_\lambda' \sigma^2.$$

Estimation du biais

$$\hat{m}_1 = S_\lambda Y.$$

Le biais

$$B(\widehat{m}_1) = \mathbb{E}[\hat{m}_1|X] - m = (S_\lambda - I)m.$$

Un estimateur possible est

$$\begin{aligned} \hat{b}_1 &= (S_\lambda - I)\hat{m}_1 = (S_\lambda - I)S_\lambda Y \\ &= -S_\lambda(I - S_\lambda)Y = -S_\lambda R_1. \end{aligned}$$

Estimateur corrigé : twiced

$$\begin{aligned}\widehat{m}_2 &= \widehat{m}_1 - \widehat{b}_1 \\ &= (S + S(I-S))Y \\ &= (I - (I-S)^2)Y,\end{aligned}$$

et on peut recommencer

$$\begin{aligned}R_{k-1} &= (I-S)^{k-1}Y \\ \widehat{b}_k &= -SR_{k-1} = -(I-S)^{k-1}SY \\ \widehat{m}_k &= \widehat{m}_{k-1} - \widehat{b}_k = \widehat{m}_{k-1} + SR_{k-1}.\end{aligned}$$

Estimateur itéré

$$\begin{aligned}\widehat{m}_k &= S[I + (I-S) + (I-S)^2 + \cdots + (I-S)^{k-1}]Y \\ &= [I - (I-S)^k]Y \\ &= S_k Y.\end{aligned}$$

$$\widehat{m}_k(x) = S(x)^t \widehat{\beta}_k.$$

Estimateur itéré

$$\begin{aligned}\widehat{m}_k &= S[I + (I - S) + (I - S)^2 + \cdots + (I - S)^{k-1}]Y \\ &= [I - (I - S)^k]Y \\ &= S_k Y.\end{aligned}$$

$$\widehat{m}_k(x) = S(x)^t \widehat{\beta}_k.$$

Comportement

Supposons que toutes les valeurs singulières λ_j de $I - S$ satisfassent $-1 < \lambda_j(I - S) < 1$ alors

$$\|\hat{b}_k\| < \|\hat{b}_{k-1}\| \quad \text{et} \quad \lim_{k\to\infty} \hat{b}_k = 0$$
$$\|R_k\| < \|R_{k-1}\| \quad \text{et} \quad \lim_{k\to\infty} R_k = 0$$
$$\lim_{k\to\infty} \widehat{m}_k = Y \quad \text{et} \quad \lim_{k\to\infty} \mathbb{E}[\|\widehat{m}_k - m\|^2] = n\sigma^2.$$

Réciproquement si $|\lambda_j(I - S)| > 1$, alors

$$\lim_{k\to\infty} \|\hat{b}_k\| = \lim_{k\to\infty} \|R_k\| = \lim_{k\to\infty} \|\widehat{m}_k\| = \infty.$$

L_2 Boosting avec lisseur S

Etape 1 A partir de $\{(X_i, Y_i), i = 1, \ldots, n\}$, ajuster le régresseur

$$\hat{F}_1(x) \quad = \quad SY,$$

Etape 2 Calculer les résidus $U_i = Y_i - \hat{F}_k(X_i)$.
Ajuster le régresseur aux résidus noté $\hat{f}_{k+1}(.) = SU$.
Mettre à jour

$$\hat{F}_{k+1}(.) \quad = \quad \hat{F}_k(.) + \hat{f}_{k+1}(.).$$

Etape 3 Itérer l'étape 2 pour obtenir

$$\hat{F}_{k+1}(.) \quad = \quad [I - (I - S)^k]Y$$

THEOREMES

Si le design est fixe ou uniform, dès que la valeur de K est plus grande que 1 et plus petite que n, au moins une valeur singulière de $I - S_\lambda$ est plus grande que 1.

Considérons un lisseur à noyau et un échantilon $X_1, \ldots, X_n$ de loi f. Si la transformée de Fourier de $K(\cdot)$ n'est pas une mesure finie positive alors avec une probabilité approchant 1 lorsque n tend vers ∞, le maximum du spectre de $I - S$ est plus grand que 1.

Multivarié

Nous avons p variables explicatives et nous avons donc

$$S_{\lambda_i}(X_i) \quad i = 1, \cdots, p.$$

Nous choisissons le même ddl pour chaque variable.
A chaque itération nous choisissons le

$$S_{\lambda_i}(X_i)$$

qui minise l'erreur résiduelle. On obtient

$$\begin{aligned} \widehat{m}_k &= S_{i_1} Y + S_{i_2}(I - S_{i_1})Y + S_{i_3}(I - S_{i_2})(I - S_{i_1})Y \\ &+ \cdots + S_{i_k}(I - S_{i_{k-1}}) \times \cdots \times (I - S_{i_1})]Y. \end{aligned}$$

Multivarié

Nous avons p variables explicatives et nous avons donc

$$S_{\lambda_i}(X_i) \quad i = 1, \cdots, p.$$

Nous choisissons le même ddl pour chaque variable.
A chaque itération nous choisissons le

$$S_{\lambda_i}(X_i)$$

qui minise l'erreur résiduelle. On obtient

$$\begin{aligned} \widehat{m}_k &= S_{i_1} Y + S_{i_2}(I - S_{i_1})Y + S_{i_3}(I - S_{i_2})(I - S_{i_1})Y \\ &+ \cdots + S_{i_k}(I - S_{i_{k-1}}) \times \cdots \times (I - S_{i_1})]Y. \end{aligned}$$

Multivarié

Nous avons p variables explicatives et nous avons donc

$$S_{\lambda_i}(X_i) \quad i = 1, \cdots, p.$$

Nous choisissons le même ddl pour chaque variable.
A chaque itération nous choisissons le

$$S_{\lambda_i}(X_i)$$

qui minise l'erreur résiduelle. On obtient

$$\begin{aligned} \widehat{m}_k &= S_{i_1} Y + S_{i_2}(I - S_{i_1})Y + S_{i_3}(I - S_{i_2})(I - S_{i_1})Y \\ &+ \cdots + S_{i_k}(I - S_{i_{k-1}}) \times \cdots \times (I - S_{i_1})]Y. \end{aligned}$$

Résultats

- Apprentissage (été 1999-2005)/ Validation (été 2006).
- Choix d'un critère d'erreur: $MAPE = 100 * \frac{|PRED - OBS|}{|OBS|}$.
- Base sans pluie.
- **Modèle additif avec splines de régression:**
  - Selection de variables: pas à pas

$$\begin{aligned} MaxO3 &= s_1(T15) + s_2(maxO3v) + s_3(gradvent) + s_4(Ne6) + \\ &\quad s_5(gradtemp) + \varepsilon \end{aligned}$$

- **Forêts**.
- **Boosting noyau gaussien**.

Résultats pour l'été 2006

Modèle	Persistance	Reg Splines	Forêts	Boosting
MAPE	17.7 %	14.1%	12.3%	??

Table: Prévision: Avril 2006 - Septembre 2006.