



Prévision des pics d'ozone à Lorient

J. Josse ¹ & Eric Matzner-Løber ²

¹Lab. de mathématiques appliquées, Agrocampus Rennes,

²Lab. de Statistiques, Université Rennes 2,

Air Breizh.

SFDS

Angers, 11 juin 2007



Plan

- **Présentation.**
- **Problème, objectif.**
- **Les données.**
- **Méthodologie.**
- **Résultats.**
- **Conclusion, perspectives.**



Présentation du sujet

- **Ozone phénomène complexe** : $NO_2 + O_2 \leftrightarrow NO + O_3$.
- **Pollution à l'ozone est un problème aux enjeux importants** : santé publique, végétation...
- **Mission d'Air Breizh** : mesure, analyse, prévision \Rightarrow Envoie tous les jours avant 17 heures l'indice ATMO du lendemain aux autorités.
- **Prev'Air** : modèle déterministe.
 - Simulation de la qualité de l'air.
 - Macro-échelle.

\Rightarrow **Idée** : modèle statistique pour Lorient : local.



Problématique

- Nous voulons prévoir aujourd'hui avant 17 heures le pic d'ozone de demain.
- Prévision de la concentration maximale d'ozone à partir de données météorologiques prévues par Météo France et de données de concentration de polluant.



Les données initiales

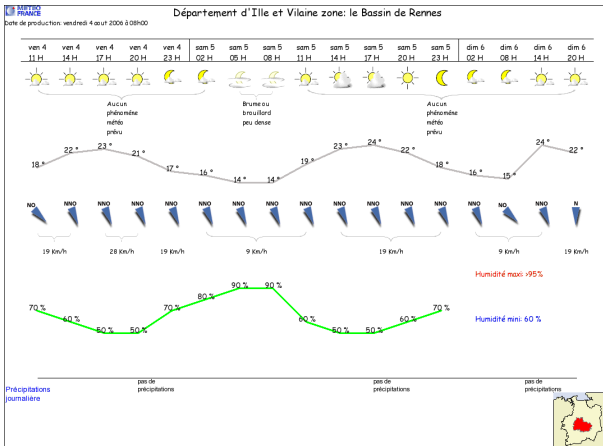
- 1^{er} janvier 1999 au 31 décembre 2006.
- Données concentration de polluant.
 - Ozone (en $\mu g/m^3$) toutes les heures
- Données météo (horaire).
 - La température (Celsius)
 - La nébulosité (0-8)
 - Le vent : vitesse et direction
 - Les précipitations (en mm)
 - La radiation et la pression (ne peuvent être prévues)
- Selection du maximum d'ozone \Rightarrow variable à expliquer : maxO3.
- Création de vents projetés ($V_x = \text{vitesse} \times \sin(\text{direction})$).
- Selection des variables à des heures ponctuelles.



Les données initiales

- 1^{er} janvier 1999 au 31 décembre 2006.
- Données concentration de polluant.
 - Ozone (en $\mu g/m^3$) toutes les heures
- Données météo (horaire).
 - La température (Celsius)
 - La nébulosité (0-8)
 - Le vent : vitesse et direction
 - Les précipitations (en mm)
 - La radiation et la pression (ne peuvent être prévues)
- Selection du maximum d'ozone \Rightarrow variable à expliquer : maxO3.
- Création de vents projetés ($V_x = \text{vitesse} \times \sin(\text{direction})$).
- Selection des variables à des heures ponctuelles.

Atmogramme





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.



Les précipitations

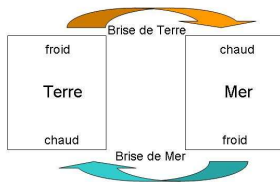
- Phénomène de lessivage.
- Heure d'apparition des pics: 11 heures-18 heures.
- Précipitations cumulées le soir, le matin?

⇒ **Création Base pluie et Base sans pluie.**



Le problème des villes côtières: la brise de mer

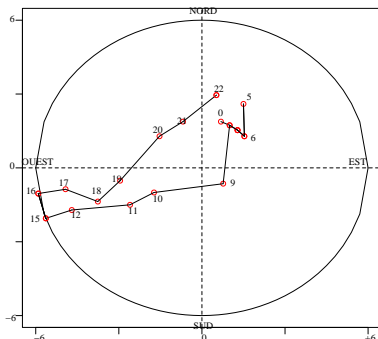
- Gradient de pression généré par la différence de température entre la masse d'air surmontant le continent et celle de la mer.



- Dilution, concentration, circulation des polluants → 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



⇒ Prise en compte de ce phénomène: création de nombreuses variables explicatives (113).

⇒ Prévoir les concentrations maximales d'ozone: **phénomène complexe: création de 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.
 - Modèle non indexable par un nombre fini de paramètres.
- Principales méthodes de régression non paramétriques :
noyaux, splines...



Modèles additifs

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

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

- Généralise 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, MCO.



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 → **Agrégation par bootstrap.**



Forêts aléatoires

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



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

$$MaxO3 = s_1(T15) + s_2(maxO3v) + s_3(gradvent) + s_4(Ne6) + s_5(gradtemp) + \varepsilon$$

- Forêts.

Résultats pour l'été 2006

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

Table: Prévision: Avril 2006 - Septembre 2006.

- Comparaison des 2 méthodes: régression spline et forêts
- Amélioration du modèle naif de persistance: régression linéaire de maxO3 sur maxO3v



Bilan

- Elaboration d'une méthodologie pour la prévision des pics d'ozone à Lorient.
- Sous-estime les très fortes concentrations (événements rares)
- Phase opérationnelle: prévision en temps réel (comparaison avec les prévisions de Prev'Air) + interface graphique.
- Combiner plusieurs méthodes
- Nouvelles variables
- Ponderer les fortes valeurs
- Corriger la prévision fournie par Prev'Air



Bilan

- Elaboration d'une méthodologie pour la prévision des pics d'ozone à Lorient.
- Sous-estime les très fortes concentrations (événements rares)
- Phase opérationnelle: prévision en temps réel (comparaison avec les prévisions de Prev'Air) + interface graphique.
- Combiner plusieurs méthodes
- Nouvelles variables
- Ponderer les fortes valeurs
- Corriger la prévision fournie par Prev'Air