

PRÉVISION DES PICS D'OZONE À LORIENT

Julie Josse & Eric Matzner-Løber

Lab. Mathématique Appliquée, Agrocampus Rennes, CS 84215, 35042 RENNES

Lab. de Statistique, IRMAR, UMR 6625, Univ. Rennes 2, CS 24307, 35043 RENNES

Résumé La pollution à l'ozone est un problème de santé publique. Le principe de précaution nécessite d'anticiper les épisodes de pollution pour mettre en œuvre des mesures d'urgence en cas de dépassement du seuil d'alerte. Ceci implique une prévision à court terme (24 h) voire à plus longue échéance. De nombreuses méthodes statistiques sont utilisées pour effectuer au jour j la prévision de la concentration maximale d'ozone pour le lendemain (jour $j + 1$). La ville de Lorient, ville côtière, est sujette aux brises de mer et de terre. Ces brises, en raison de la dilution, de la concentration et de la circulation des polluants qu'elles entraînent d'un jour à l'autre, peuvent considérablement complexifier les épisodes de pollution. Il faut donc en tenir compte dans les prévisions. Nous présentons dans ce travail la manière dont nous avons modélisé les brises et les résultats des prévisions obtenues par régression spline et forêts aléatoires.

Abstract In this work, we focus our attention on the prediction of ozone concentration in the city of Lorient, France. Due to its coastal situation, the Lorient area is regularly subject to sea and land breezes in summertime. To face this problem, we show how we use exogenous variables in order to build 24 hours ahead prediction models. Then we compare results obtained by spline regression and random forest.

Mots clés Prévision, pollution atmosphérique, régression, arbres.

1 Introduction

La pollution à l'ozone est depuis plusieurs années, un thème d'étude récurrent dans de nombreuses disciplines dont la statistique. L'ozone est un polluant secondaire qui se forme naturellement à partir de l'oxygène de l'air. Au delà d'une certaine concentration dans l'atmosphère l'ozone peut se montrer dangereux pour la santé. Il est donc important de prévenir les personnes sensibles en cas d'épisodes de pollutions. De nombreuses méthodes statistiques sont utilisées pour effectuer au jour j la prévision de la concentration maximale d'ozone pour le lendemain (jour $j + 1$) (le lecteur intéressé peut se reporter à Bellanger *et al.* (1999) par exemple). Le choix d'une méthode est souvent très délicat et la prévision des pics de pollution n'échappe pas à ce principe.

Prev'Air est le système de prévision opérationnel pour l'ensemble du territoire depuis juillet 2003. Ce système diffuse quotidiennement des prévisions et cartographies de la qualité de l'air en France et en Europe. Compte tenu de sa portée nationale, il n'est pas exclu que ponctuellement, les prévisions effectuées nationalement diffèrent de celles réalisées localement. L'objectif de ce travail est donc de fournir des modèles de prévision des pics

d’ozone à une échelle plus fine, cela nous permettra d’introduire des caractéristiques locales dans les modèles comme la brise de mer à Lorient.

Nous allons insister dans un premier temps sur la difficulté de construire un jeu de variables explicatives. Nous allons ensuite présenter les méthodes de prévision utilisant la régression spline et les forêts aléatoires.

2 Les données

Nous nous intéressons à une station du réseau Air Breizh : la station urbaine de Lorient. Nous disposons d’une base de données historiques couvrant la période comprise entre le 1^{er} janvier 1999 et le 31 décembre 2005. Cette base contient :

- des données de concentration horaire d’ozone ;
- des données météorologiques horaires mesurées par Météo France : température, nébulosité, radiation, humidité, vent (vitesse et direction), précipitation.

Pour prévoir la concentration d’ozone du jour $j + 1$ (variable à prédire, notée $\max O_3$), nous utilisons la concentration maximale d’ozone observée le jour j (en effet, la concentration maximale relevée le jour reste assez voisine de la valeur relevée la veille, ce phénomène de persistance est primordial pour la prévision) et des prévisions météorologique fournies par Météo France le jour j pour le jour $j + 1$. Il faut donc, dans la construction des modèles de prévision, tenir compte de variables que peut prévoir Météo France. Dans notre cas, nous obtenons des prévisions toutes les 3 heures de la température, de la nébulosité, du vent et de la précipitation sous la forme d’indications comme “petite pluie le matin”, ou bien “orage prévu”... Nous ne pouvons donc pas utiliser l’information quantitative de la précipitation mais il est quand même pertinent d’envisager de réaliser deux modèles: un modèle “pluie” et un modèle “sans pluie”. Nous avons déterminé une variable de précipitation cumulée associée à un seuil pertinent pour construire une base “pluie” et une base “sans pluie”. Nous ne présentons que les modèles obtenus avec la base “sans pluie” car la concentration d’ozone diminue lors de phénomènes pluvieux. La variable “vent” est composée de la vitesse en mètre par seconde et d’une direction en degré (0 correspond au Nord). Nous travaillons alors avec le vent projeté sur l’axe Est-Ouest : $Vx = vitesse * \sin(direction)$ et sur l’axe Nord-Sud.

L’approche statistique consiste à explorer l’ensemble des données mesurées et à construire un jeu de variables explicatives permettant de prédire au mieux les valeurs de la variable à expliquer. Les variables explicatives présentées précédemment ne sont pas suffisantes pour prévoir les concentrations maximales d’ozone. Nous avons donc créé plusieurs variables statistiques d’intérêt pour essayer de prendre en compte l’influence du vent à Lorient et notamment la brise de mer. Les circulations de brises de mer et de terre sont dues au gradient de pression généré par la différence de températures entre la masse d’air surmontant le continent et celle de la mer. Cette particularité météorologique des régions

côtières complexifie considérablement la prévision des épisodes de pollution. Le graphique 1 matérialise l'évolution du vent sur une journée de brise de mer à la station de Lorient.

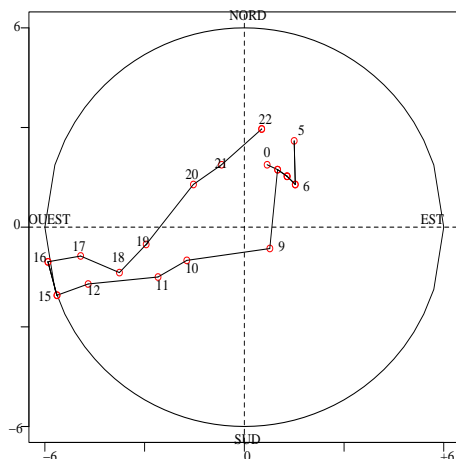


Figure 1: Journée de brise de mer du 22 avril 2002

Nous retrouvons les principales caractéristiques des brises : un vent faible le matin (inférieur à 3 m/s entre 0 et 7 heures) pour que la brise puisse s'installer, un changement de direction du vent d'au moins 90 degrés de la terre vers la mer accompagné d'une augmentation de la vitesse du vent de moins de 6 m/s.

Nous avons donc créé de nombreuses variables explicatives pour prendre en compte ce phénomène. Les variables explicatives sont maintenant au nombre de 113.

3 Modélisation

Nous considérons le modèle suivant :

$$Y_i = r(X_{i,1}, \dots, X_{i,p}) + \varepsilon_i.$$

où Y_i représente le maximum de la concentration du lendemain $j + 1$, les X_1, \dots, X_p sont les p variables explicatives (par exemple l'ozone du jour j , la température du lendemain à 12h ... il y en a 113) et ε est une variable aléatoire. Dans le cas multidimensionnel, la régression non paramétrique présente plusieurs problèmes et en particulier le problème dit du fléau de la dimensionnalité ("curse of dimensionality"). Les modèles additifs et les méthodes de régression par arbre proposent des solutions pour réduire la dimension.

3.1 Les modèles additifs

Le modèle additif propose une réponse au problème de dimensionnalité. Nous supposons alors que la fonction de régression possède la forme suivante :

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

où $E[\varepsilon] = 0$ et $Var(\varepsilon) = \sigma^2$; α_0 est une constante, r_j sont des fonctions unidimensionnelles telles que avec $E_{X_j}[r_j] = 0$, condition d'identifiabilité. Les fonctions r_j sont arbitraires et supposées lisses. Elles sont estimées à partir des données.

Ce modèle est plus restrictif qu'un modèle de régression non paramétrique général mais moins restrictif qu'un modèle de régression linéaire. La régression par modèle additif non paramétrique présentée dans Hastie et Tibshirani (1990) estime la dépendance entre une variable réponse Y et plusieurs variables explicatives $X = (X_1, \dots, X_p)$ d'une façon flexible et interprétable, en restreignant les dépendances à une somme de fonctions monovariées. Cette structure simple permet également de représenter l'effet de chaque variable, ce qui facilite l'interprétation des solutions. Nous pouvons utiliser tous les lisseurs comme fonction de régression partielle r_j dans les modèles additifs. Nous avons choisi de nous intéresser aux splines de régression. Notre modèle devient donc:

$$y_i = \alpha_0 + s(x_{i1}) + s(x_{i2}) + \dots + s(x_{ip}) + \varepsilon_i.$$

Les splines sont des polynômes par morceaux d'ordre m (de degré $m-1$) qui se raccordent ainsi que certaines de leurs dérivées en k points appelés nœuds intérieurs. L'espace des fonctions splines est de dimension $r = m + k$. Cet espace est entièrement caractérisé par le degré, le nombre des nœuds et leur position. Nous utilisons comme base, les B-splines. Pour chaque variables explicatives, on utilise une base B-spline B_j , avec une séquence de nœuds associée. Ainsi, nous avons

$$Y = B\beta + \varepsilon \quad \text{où} \quad B = [B_1, \dots, B_p].$$

Lorsque les base B_j sont définies, les estimateurs des paramètres peuvent être calculés à l'aide des moindres carré. Le modèle additif devient donc un modèle linéaire. Cette propriété est donc très commode. L'inconvénient majeur de cette méthode est la difficulté de choisir le nombre et la position des nœuds pour chaque variable explicative. Pour chaque variable, nous positionnons 6 nœuds intérieurs aux quantiles 0.25, 0.5, 0.6, 0.7, 0.8 et 0.9 et choisissons un degré 2. Le nombre équivalent de paramètres pour chaque variable est donc de 9 ($6+2+1$). Nous avons aussi examiné les résidus partiels pour nous aider quant au choix de l'emplacement des nœuds. Nous avons choisi les variables explicatives de façon ascendante en ne conservant que les variables les plus pertinentes pour effectuer la prévision (utilisation d'un sous-échantillon).

3.2 Forêts aléatoires

Les techniques de régression par arbre sont très utiles en régression lorsqu'on est en présence d'un nombre important de variables explicatives et qu'on s'attend à une relation très complexe entre la variable à expliquer (la réponse) et les variables explicatives. La méthode CART est présentée dans Breiman *et al.* (1984). Elle construit une partition à l'aide d'un arbre binaire, optimale pour un critère de somme des carrés des erreurs intra-classes. Les arbres divisent l'espace des variables explicatives en un ensemble d'hypercubes. Un modèle simple est un modèle où une constante est ajustée sur chaque hypercube. Soit la fonction de régression suivante:

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

Les c_j sont des constantes et les F_j constituent une partition de \mathbb{R}^p . La surface de régression est constante par morceau. Il s'agit donc d'estimer les c_j mais aussi de déterminer le nombre k et les classes F_j de la partition. Quand les F_j sont connus, le meilleur estimateur au sens des moindres carrés des c_j est donné par la moyenne des Y dans la classe F_j . Trouver la meilleure partition au sens de la minimisation de la somme des carrés n'est pas possible. On utilise alors un algorithme. La méthode proposée pour construire la partition F_j est la méthode séquentielle de régression par arbre.

Les intérêts de cette méthode sont multiples : cette méthode est simple, facile à comprendre, facile à utiliser et à interpréter. De plus, elle ne limite pas le nombre de variables explicatives. Elle œuvre simultanément sur le plan descriptif (étiquetage des nœuds) et sur le plan décisionnel (construction d'une règle de décision). Elle permet donc de produire des règles de décision et d'aider à la compréhension du phénomène. L'inconvénient majeur de cette méthode, est son manque de stabilité. Un petit changement sur l'échantillon d'apprentissage peut avoir des conséquences importantes. Les forêts aléatoires (random forests) sont dues à Breiman (2001) qui propose une amélioration du bagging spécifique aux modèles définis par des arbres binaires (CART). Les forêts aléatoires sont une collection d'arbres, h_k , $k = 1, \dots, K$, où chaque arbre est construit à partir d'un échantillon bootstrapé (tirage avec remise dans l'échantillon de départ). Cependant, au lieu de chercher à chaque étape la variable et le nœud qui minimisent la somme des carrés des écarts à la moyenne, un sous ensemble de variables choisi aléatoirement est utilisé. L'algorithme de construction de chaque arbre est le suivant :

- Construction d'un échantillon bootstrapé à partir de $(X_1, Y_1), \dots, (X_n, Y_n)$;
- Tirage aléatoire d'un sous ensemble de q prédicteurs ;
- Construction d'un arbre avec les q variables (construit à l'aide de l'algorithme CART).

Finalement la prédiction finale est la moyenne des prédictions.

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

Breiman (2001) a montré que les forêts aléatoire bénéficient d'excellentes qualités prédictives en présence d'un grand nombre de prédicteurs.

4 Evaluation des résultats

L'élaboration d'un modèle statistique de prévision de la concentration d'ozone se fait donc à partir d'un échantillon de données dit échantillon d'apprentissage. Nous travaillerons sur des données d'apprentissage concernant la période d'été 1999 à 2005. Nous testons ensuite notre modèle sur un échantillon test couvrant la période allant du 1er avril 2006 au 30 septembre 2006 soit une période de 152 jours sans pluie. Le critère de qualité des modèles de prévision que nous utilisons est (MAPE Mean absolute percentage error) :

$$MAPE = 100 * \frac{1}{n} \sum_{i=1}^n \frac{|\hat{Y}_i - Y_i|}{Y_i}$$

Afin de comparer les différents modèles présentés, nous avons calculé les MAPE sur l'ensemble des jours sans pluie. Les résultats sont résumés dans le tableau suivant:

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

Table 1: Prévion : Avril 2006 - Septembre 2006, 152 jours sans pluie

Les modèles proposés améliorent considérablement le modèle simpliste basé sur la persistance. Les forets aléatoires donnent de très bon résultats de prévision, cependant l'implémentation pour une utilisation quotidienne est plus lourde que l'implémentation de la régression spline. En effet dans le premier cas, nous utilisons 113 variables explicatives alors que dans le second cas, nous en utilisons 5.

Bibliographie

- [1] Bellanger, L. et Bel, L. et Tomassone, T. (1999) Eléments de comparaisons de prévisions statistiques des pics d'ozone, *Revue de Statistique Appliquée*, 47, 7-25.
- [2] Hastie, T. J. et Tibshirani, R. J. (1990) *Generalized additive models*, Chapman and Hall
- [3] Breiman, L. (2001) Random Forests, *Machine Learning*.
- [4] Breiman, L. et Friedman, J. et Olsen, R. et Stone, C. (1998), *Classification and Regression Trees*, CRC PRESS.