REVUE DE STATISTIQUE APPLIQUÉE

L. BEL
L. BELLANGER
V. BONNEAU
G. CIUPERCA
D. DACUNHA-CASTELLE
C. DENIAU
B. GHATTAS
M. MISITI
Y. MISITI
G. OPPENHEIM
J. M. POGGI
R. TOMASSONE

Éléments de comparaison de prévisions statistiques des pics d'ozone

Revue de statistique appliquée, tome 47, n° 3 (1999), p. 7-25

<http://www.numdam.org/item?id=RSA_1999__47_3_7_0>

© Société française de statistique, 1999, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (http://www.sfds.asso.fr/publicat/rsa.htm) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

Rev. Statistique Appliquée, 1999, **XLVII** (3), 7-25

ÉLÉMENTS DE COMPARAISON DE PRÉVISIONS STATISTIQUES DES PICS D'OZONE

L. Bel[1], L. Bellanger[1], V. Bonneau[2], G. Ciuperca[1], D. Dacunha-Castelle[1], C. Deniau[3], B. Ghattas[3], M. Misiti[1], Y. Misiti[1], G. Oppenheim[1] J.M. Poggi[1], R. Tomassone[4]

[1] *Modélisation Stochastique et Statistique, Université Paris Sud, bât. 425, Mathématique, 91405 Orsay*

[2] *Airparif*

[3] *GREQAM, Université de la Méditerranée*

[4] *Mathématique et Informatique, INA Paris*

RÉSUMÉ

Cet article porte sur la prévision des pics de pollution par l'ozone sur Paris et sa région. On expose brièvement les quatre modèles de prévision issus de méthodes variées : non linéaire non paramétrique générale et une spécialisation additive, la méthode classique dans le domaine, de classification-discrimination-régression et enfin la méthode CART de régression par arbre. On insiste sur l'intérêt méthodologique de confronter et de combiner plusieurs méthodes complémentaires. On présente la manière de comparer ces différents modèles de prévision afin de fournir à Airparif (organisme chargé de la surveillance et de la prévision de la qualité de l'air en Ile de France) un modèle opérationnel de prévision des concentrations de l'ozone pour l'été 1997. Ces prévisions ont été fournies en temps réel et les résultats obtenus durant cette période sont commentés.

Mots-clés : *Pollution de l'air, statistique appliquée, prévision, méthodes comparatives.*

ABSTRACT

This paper deals with the forecasting of ozone episodes in the Paris area. We present briefly the four forecasting models obtained from different methods: a general nonparametric method and an additive specialization, the classical method used in this area clustering-discrimination-regression and lastly CART a regression tree method. We focus on methodological interest of comparing and combining several complementary methods. We present the way to compare these different foracasting models in order to deliver to Airparif (the air pollution agency for Paris area) a real-time forecasting model for ozone episodes during summer 1997. Finally the results obtained during this period are commented.

Keywords : *Air pollution, applied statistics, forecasting, comparative methods.*

1. Introduction

Ce travail portant sur la prévision des concentrations maximales d'ozone sur Paris et sa région, a été effectué dans le cadre d'une collaboration entre l'Université d'Orsay et Airparif, organisme chargé de la surveillance et de la prévision de la qualité de l'air en Ile de France. La prévision du maximum d'ozone de la journée, qui a lieu en général l'après-midi, est effectuée à 6h T.U. (c'est-à-dire 8h en été). Il s'agit de prévoir les pics de pollution sur un ensemble de sites répartis sur Paris et sa banlieue, constituant le réseau d'alerte. Pour l'ozone celui-ci comprend huit stations situées à Aubervilliers, Champs-sur-Marne, Créteil, Montgeron, Neuilly, Paris 13, Tour Eiffel et Tour Saint-Jacques. Les mesures effectuées sur ce réseau permettent de déclencher les procédures d'information des autorités administratives et du public qui sont désormais bien connues de la population. Plus précisément, pour chacune des stations on détermine le niveau : niveau 0 si le maximum de pollution est inférieur à 130 μg/m^3, niveau 1 s'il est compris entre 130 et 180 μg/m^3, niveau 2 s'il est compris entre 180 et 360 μg/m^{3}et niveau 3 s'il est supérieur à 360 μg/m^3. Puis on agrège ces niveaux en disant qu'il y a alerte (nous adoptons ici une terminologie différente de celle d'Airparif, qui distingue alerte et alarme) dès qu'au moins deux stations de mesure dépassent 130 μg/m^3. Plus précisément, il y a alerte de niveau i si deux stations au moins sont de niveau supérieur ou égal à i. Les niveaux 0 sont de loin les plus nombreux, les autres étant rares et la plupart de niveau 1.

Bien sûr, Airparif souhaite disposer dans le futur, d'un outil de prévision permettant d'adopter des mesures préventives afin d'éviter les épisodes de forte pollution (niveau 2 ou plus). L'objectif final est de prévoir la veille à midi, les concentrations maximales et l'alerte du lendemain. Le premier pas dans cette direction consiste à prévoir le matin l'alerte de l'après-midi.

Prévoir est par nature un art difficile quel que soit le domaine d'application et la prévision des pics de pollution ne déroge pas à cette règle. En effet, la pollution est la résultante de nombreux phénomènes complexes avec des descriptions physiques riches, à différentes échelles en temps et en espace offrant des visions différentes : météorologiques, physico-chimiques, topographiques (cf. Toupance 88, Jin, Demerjian 93, Vautard *et al.* 97, Walcek, Yuan 95 et [13]). Il existe de nombreux travaux dans cette direction consistant à modéliser la dynamique de fond du système engendrant la pollution, on peut citer par exemple Poulos, Pielke 94. Ces modèles sont très lourds à mettre en œuvre mais demeurent indispensables par exemple pour permettre de prévoir et d'analyser les situations de pollution importée d'une source d'émission lointaine par le déplacement des masses d'air. On peut d'ailleurs s'inspirer de la simplification de l'un de ces modèles pour donner une idée de la pollution de l'air. Les voitures et les industries émettent des précurseurs de l'ozone dans une boîte dont la base est la région parisienne et la hauteur est celle de la couche de mélange. Plus cette hauteur est grande, plus les émissions se diluent facilement et meilleure est la qualité de l'air. Cette hauteur varie suivant l'heure de la journée, la période de l'année et les conditions météorologiques. Les variations peuvent être énormes vis à vis des fluctuations des émissions. L'air contenu dans cette boîte est agité horizontalement par le vent et verticalement par des turbulences.

Par opposition à ce type de modélisation, l'approche statistique, quant à elle, est plus flexible et les procédures de prévision plus légères, mais elle est restreinte

à l'analyse et la prévision de la pollution à court terme sur une petite échelle en espace et nécessite un historique suffisamment long. Parmi les nombreux travaux antérieurs utilisant cette approche on peut citer Seinfeld 88, Eder *et al.* 94, Ryan 95, Milionis, Davies 94, Sheifinger *et al.* 96, Wipij, Sally Liu 94. Un seul travail avancé sur un thème très voisin est effectué en France, par D. Brion (Météo-France) et E. Gilibert (Airparif) dans [3]. Après un très important (en qualité et en volume) travail de validation des données, les auteurs prévoient les événements : le niveau 1 sera atteint ou le niveau 1 ne sera pas atteint. La technique est de type prévision parfaite (cf. plus bas) d'un jour pour le lendemain.

Une approche unifiée intriquant les diverses descriptions des mêmes phénomènes aux diverses échelles de temps et d'espace constituerait une modélisation plus fine et efficace.

Les données utilisées sont d'une part des données de pollution : concentrations d'ozone, de monoxyde et dioxyde d'azote, qui proviennent des huit stations du réseau d'alerte d'Airparif et d'autre part de données météorologiques : température au sol, températures à 40 m et à 100 m, les intensités et directions du vent à 58 m et à 110 m qui proviennent du seul mât de Saclay. La période de l'étude comprend les jours d'été de 1992 à 1996, pour lesquels nous disposons de données considérées comme fiables. Notons que dans le travail exposé nous ne disposons pas en temps réel, des mesures ou des prévisions de Météo France.

Quels sont les enjeux de ce travail? Ils sont doublement méthodologiques : méthodologie d'un problème en vraie grandeur et méthodologie statistique.

Sans être extraordinairement volumineuses, les données ont posé des problèmes variés. Qualité et représentativité sont deux parmi les nombreux aspects qui incluent le traitement de données manquantes (environ 15 %), l'évaluation des incertitudes des mesures horaires.

Les études, devant déboucher sur un produit opérationnel implanté sur un calculateur travaillant tous les jours, peuvent recourir à toutes les techniques statistiques de réglage des paramètres, mais doivent être, in fine, suffisamment simples pour être programmées de manière à être maintenues. Les destinataires finaux du travail veulent des procédures, sinon optimales, du moins aussi bonnes que possible. C'est une des raisons de l'introduction de plusieurs méthodes et de la comparaison systématique, certainement lourde mais nécessaire.

Le point de vue statistique est essentiel. D'autres techniques de prévision faisant appel à des modèles déterministes physico-chimiques partent des mêmes approches que les prévisions météorologiques. Ces approches se développent bien, mais seront opérationnelles à l'échelle de cinq à sept ans. Pour le moment les méthodes statistiques sont les plus fiables pour la prévision à court terme de l'ozone (cf. Seinfeld 88 et Sheifinger *et al.* 96). Parmi les questions que peut se poser un statisticien on trouve les suivantes. Sur quels critères se base la comparaison des méthodes? Comment construire un critère synthétique, permettant d'ordonner totalement des procédures? Les comparaisons sont souvent faites sur des résultats en moyenne sur un grand nombre de jours, fournissant une information très globale. Cette étape est nécessaire et classique. Nous avons été conduit à compléter ces travaux de masse par des analyses cliniques de jours associés à de grosses erreurs de prévision. Elles servent à comprendre les causes de dysfonctionnement des prévisions.

Le plan de l'article est le suivant. Après cette introduction, on expose les quatre modèles qui ont été construits en insistant sur l'intérêt de confronter et combiner plusieurs méthodes complémentaires. Dans le paragraphe 3, on présente la technique de comparaison de ces différents modèles de prévision. Enfin un bilan de la campagne de l'été 1997 et une analyse clinique de quelques jours complètent cet article.

2. Les modèles

Quatre modèles ont été étudiés. Ils sont estimés en utilisant des méthodes différentes; modèles et méthodes d'estimation correspondantes seront ainsi baptisés du même nom : classification-discrimination-régression (abrégée parfois dans la suite C.D.R.), régression par arbre (appelée dans la suite CART), non linéaire général et non linéaire additif. Nous les présentons brièvement avant de confronter ces différentes attaques du problème et d'esquisser quelques remarques méthodologiques.

Dans un premier temps, il est possible de préciser l'ensemble de ces modèles d'une manière unifiée, sans entrer dans les détails. Le modèle général est un modèle de régression non linéaire de la forme :

$$Z_j = f(X_j^{(1)}, ..., X_j^{(l)}) + \epsilon_j \tag{1}$$

où j désigne l'indice du jour, les variables $X_j^{(i)}$ sont des variables endogènes ou exogènes et Z_j le maximum de l'ozone. Cette forme générale est celle du modèle appelé non linéaire général. L'estimation consiste alors à estimer f. Les trois autres modèles peuvent être vus comme des spécialisations.

Ainsi le modèle non linéaire additif est une adaptation de (1) au cas où f est additive et séparable, c'est-à-dire que les effets des variables sont supposés additifs et qu'il n'y a pas d'interaction :

$$Z_j = \sum_{i=1}^{l} f^{(i)}(X_j^{(i)}) + \epsilon_j \tag{2}$$

L'estimation consiste à estimer les $f^{(i)}$.

Les modèles CART et classification-discrimination-régression sont de formes très voisines.

Commençons par le modèle CDR qui s'exprime ainsi :

$$Z_j = \sum_{i=1}^{nc} 1_{C_i}(X_j^{(1)}, ..., X_j^{(l)}) \sum_{k=1}^{l} \alpha_{k,i} X_j^{(k)} + \epsilon_j \tag{3}$$

où nc est le nombre inconnu de classes. Estimer le modèle consiste à construire les C_i qui sont définis par les étapes de classification et de discrimination et à estimer les $\alpha_{k,i}$ par une régression à l'intérieur de chacune des classes.

Le modèle CART s'écrit :

$$Z_j = \sum_{i=1}^{nf} 1_{C_i}(X_j^{(1)}, ..., X_j^{(l)})\alpha_i + \epsilon_j \qquad (4)$$

où nf est le nombre inconnu de feuilles de l'arbre sous-jacent à la méthode. Estimer le modèle consiste à construire les classes C_i qui sont des pavés explicites de l'espace des variables explicatives et à estimer les α_i par les moyennes, dans les classes, de la variable expliquée.

2.1. La méthode classification-discrimination-régression

Cette méthode utilise les techniques classiques, à savoir la classification, la discrimination et la régression linéaire (cf. Tomassone *et al.* 93). Ce type de procédure est largement utilisé dans le domaine de la prévision de la pollution, on peut citer par exemple Ryan 95 et Eder *et al.* 94. L'originalité, importante à notre avis et qui a séduit les équipes de surveillance de la qualité de l'air, réside ici dans l'utilisation des profils d'ozone pour construire la classification des jours, au lieu d'utiliser des types de temps déduits des variables météorologiques.

Pour chacune des huit stations, on procède de la façon suivante : classification en fonction des profils d'ozone; discrimination en fonction des composantes principales calculées à partir de l'ensemble initial d'une centaine de variables disponibles à 6 h T.U.; classement des jours suivant la probabilité d'appartenance à chaque classe (déterminée dans l'étape précédente); puis régression linéaire dans chaque classe : ACP par classe sur les mêmes variables qui sont utilisées dans la discrimination et choix des meilleurs régresseurs parmi les composantes principales. La prévision est obtenue à partir des résultats de la discrimination et de la régression.

2.2. La méthode CART

Cette méthode non paramétrique construit une partition en pavés de base rectangulaires de l'espace des variables explicatives, et attribue une valeur de la variable expliquée à chaque ensemble de la partition. Elle est présentée dans Breiman *et al.* 84 et Venables, Ripley 94. La méthode CART construit une partition visualisée à l'aide d'un arbre binaire, optimale pour un critère de somme des carrés des erreurs intra-classe. L'erreur commise sur la prévision un jour donné est mesurée par l'écart de la pollution réalisée du jour à la moyenne des pollutions des jours de la classe. L'algorithme de construction procède en plusieurs étapes : construction d'un premier arbre, souvent trop grand, suivie d'une étape d'élagage qui choisit le meilleur arbre ayant un nombre de feuilles donné. Cette méthode est simple et facile à comprendre par son principe de découpage en arbre de régression. Elle permet ainsi l'analyse des causes de mauvaises prévisions et une approche clinique des jours atypiques.

2.3. La méthode non linéaire générale

C'est une méthode de prévision non paramétrique à noyau (cf. Härdle 90) basée sur l'estimation de f dans (1) fonction de variables endogènes et exogènes.

L'estimateur peut s'écrire sous la forme :

$$\hat{Z}_j = \sum_{k=1}^{j-1} \omega_k \cdot Z_k \tag{5}$$

où $\sum_{k=1}^{j-1} \omega_k = 1.$

En d'autres termes $\hat{Z}_j$ s'obtient comme le barycentre des $(Z_k)_k$ affectés des poids ω_k. Le coefficient ω_k s'interprète (cf. Poggi 94) comme un indice de similarité du jour du passé repéré par $(X_k^{(1)}, ..., X_k^{(l)})$ au jour témoin caractérisé par $(X_j^{(1)}, ..., X_j^{(l)})$ représentant le présent. La similarité est mesurée par l'image d'une distance entre les deux vecteurs précédents par un noyau de probabilité, de sorte que plus la distance entre un jour du passé et le jour présent est grande (resp. petite) et plus la valeur de ω_k sera petite (resp. grande). La prévision $\hat{Z}_j$ s'obtient donc comme le barycentre des concentrations d'ozone Z_k de tous les jours du passé affectés des coefficients de similarité ω_k de ces blocs au bloc témoin. Cette remarque implique donc que le calcul de la prévision passe par une étape de sélection des jours les plus ressemblants à la situation présente, permettant ainsi l'analyse des causes de mauvaises prévisions et une approche clinique des jours atypiques.

2.4. La méthode non linéaire additive

Cette méthode est présentée dans Hastie, Tibshirani 90 et Härdle 90. Elle propose un compromis entre le modèle non linéaire général présenté ci-dessus et le modèle additif le plus simple qui est le modèle linéaire. Les effets des variables explicatives sont non linéaires mais sont supposés additifs. Cette méthode présente l'avantage d'être flexible et de conduire à un modèle explicite, facile à interpréter.

La méthode d'estimation non paramétrique des fonctions $f^{(i)}$ dans (2) adoptée est basée sur l'estimateur dit par intégration marginale (cf. Linton, Härdle 96).

2.5. Les stratégies de prévision

La bibliographie de la prévision statistique dans les sciences de l'atmosphère, telle qu'elle apparaît dans le livre de Wilks 95, présente plusieurs techniques. La distinction majeure se fait entre les techniques tirant bénéfice des «sorties de modèles» et les autres. On désigne sous le vocable «sortie de modèle» la prévision d'une grandeur faite par une technique déterministe, souvent construite par discrétisation des équations de mécanique des fluides. Les techniques s'appellent Prévision Parfaite (Perfect Prognosis, on peut citer par exemple Karl 79), MOS (Model Output Statitics, on peut citer par exemple Clark, Karl 82, Brion, Gilibert 97) ou méthodes classiques. Dans la mise en œuvre des procédures, on doit distinguer l'étape de construction, appelée dans d'autres contextes étape d'identification, de l'étape de prévision.

Les prévisions sont faites à 06 h T.U. le jour j et l'on prévoit l'après-midi du jour j.

- Méthode 1 : Prévision parfaite (notée PP dans la suite).
  - L'étape de construction utilise les observations de variables météorologiques du jour j (donc non disponibles) ainsi que d'autres observations qui sont quant à elles, disponibles avant 06h T.U.;
  - L'étape de prévision utilise la prévision faite par Météo-France de ces variables météorologiques non observées.
- Méthode 2 notée OBS dans la suite. Cette méthode utilise seulement des observations disponibles à 06h T.U.;
- Méthode 3 notée MOS dans la suite. Elle construit le modèle et prévoit en utilisant des prévisions des variables météorologiques de Météo-France.

Les techniques employées dans les procédures non linéaires, sont du type PP alors que celles utilisés par CART et la CDR sont de type OBS.

Plus précisément les variables utilisées dans les procédures non linéaires, pour la prévision du jour j sont la température maximale du jour j, le vent moyen à 58 m dans l'après-midi du jour j et la concentration maximale d'ozone du jour $j-1$ (la veille).

Les prévisions de la température maximale et du vent moyen à 58 m dans l'après-midi du jour j peuvent provenir de deux sources, soit de Météo-France (l'historique disponible n'est que de deux ans et ce sont des prévisions données la veille au soir), mais les données ne sont pas disponibles en temps réel, soit par une technique de prévision non linéaire similaire à celle utilisée pour prévoir l'ozone.

Les variables utilisées par CART et la CDR sont les variables de pollution et de climat effectivement mesurées et listées dans l'introduction, à différentes heures, ainsi que des variables calculées à partir de ces dernières, comme les gradients de températures aux diverses altitudes.

Habituellement les prédicteurs de MOS, en météo, sont des prévisions (appelées sorties de modèle) calculées par des méthodes déterministes. Nous n'avons pu pour le moment, expérimenter l'approche MOS en raison de l'indisponibilité des prévisions.

Les avantages et les inconvénients des approches sont difficiles à évaluer sereinement. Quelques points sont souvent cités, par exemple dans Wilks 95. Rappelons que ces méthodes incluent un prédicteur de température et de vent que nous appellerons ici prédicteur intermédiaire.

- En faveur de MOS. La phase de construction et la phase de prévision utilisent des informations homogènes entre elles. Un biais systématique du prédicteur intermédiaire ne perturbe pas la prévision.
- En défaveur de MOS. Dans sa phase de construction, MOS a besoin de l'historique des prévisions, parfois difficile à obtenir. En France, par exemple, les techniques de prévision des températures changent régulièrement. Des adaptations de ces techniques se font tous les ans. Les prévisions, s'améliorent en général, bien sûr, mais les modifications introduisent souvent des discontinuités dans les prévisions.
- En faveur de PP.

  - Dans sa phase de construction, PP a besoin de l'historique des observations, assez facile à obtenir. Dans sa phase de prévision de la pollution, PP a seulement besoin des prévisions de température et de la vitesse du vent pour ce même jour;
  - La technique utilisée pour assurer la prévision climatique, pourvu qu'elle soit de bonne qualité est indifférente. Elle peut même changer avec le temps. C'est un second avantage.
- En défaveur de PP. Pour que PP fonctionne, le prédicteur intermédiaire doit être de bonne qualité. S'il introduit un biais rien ne peut le rectifier, comme le souligne Wilks.

2.6. Pourquoi utiliser plusieurs modèles?

La principale raison est l'absence d'éléments de référence. De ce point de vue, une redondance de modèles issus de méthodes variées, utilisant des stratégies complémentaires conduit à une sorte de robustesse.

On peut distinguer deux familles en examinant les modèles actuellement implémentés à Airparif :

- les modèles non linéaires, généraux et additifs, qui ont été construits en utilisant la technique dite de «prévision parfaite». Les variables utilisées pour caler le modèle sont issus d'une procédure de sélection de variables et confirmées par CART qui peut donner de bons indices pour l'identification de tels modèles non linéaires. En outre, des relations statistiques désormais bien connues (cf. Sheifinger *et al.* 96) existent entre l'ozone et quelques variables météorologiques dont la température, la radiation UV, la radiation globale, la vitesse du vent. La température est un des prédicteurs les plus corrélés et le plus facilement prévisible. Le terme de persistance (l'ozone de la veille) est une sorte de variable auxiliaire pour la prise en compte d'autres variables non disponibles ou non identifiées;
- les modèles CART et CDR se placent à 06 h T.U. et ne considèrent que les variables dont les mesures sont disponibles. C'est une des raisons pour lesquelles le nombre de variables utilisées par ces deux autres méthodes est beaucoup plus important. En effet, l'introduction de peu de variables climatiques réalisées cachent en un certain sens de très nombreuses variables disponibles à 06 h T.U. La température réalisée est dans le cas de l'ozone, une variable très fortement liée à un ensemble important de variables explicatives. Evidemment l'intérêt de la prévision parfaite, suspecte si la prévision des variables climatiques est basée sur les mesures disponibles à 06 h T.U., est manifeste dans le cas pour lequel les variables météorologiques sont à la fois faciles à mesurer et à prévoir et que le travail de prévision est effectué par Météo-France sur la base de mesures auxquelles nous n'avons pas accès et de modèles à grande échelle en espace.

En outre d'autres aspects distinguent les deux groupes de méthodes :

- tout d'abord dans chacun des groupes :
  - les avantages de principe de CART sont clairs : la méthode est à la fois plus automatique et le modèle représenté par un arbre dont les nœuds sont étiquetés par des conditions portant sur les variables initiales, est plus facile à interpréter que la procédure classification-discrimination-régression qui utilise

des composantes principales. En revanche, la stratégie de conception de CART est plus figée. Les deux procédures se complètent et CART peut aussi être utilisé dans un contexte d'analyse des données pour la compréhension des phénomènes;

- le non linéaire additif est une spécialisation du non linéaire général. Il se démarque de celui-ci par deux inconvénients : l'absence de termes d'interaction entre variables dans le modèle et l'absence d'interprétation en termes de ressemblance des jours du passé au jour présent. Il présente deux avantages : on souffre moins du «fléau» de la dimension (surtout dans les zones de l'espace d'état de faible densité) puisque les vitesses de convergence des estimateurs sous des hypothèses habituelles dans un contexte de régression, sont les vitesses non-paramétriques unidimensionnelles (cf. Linton, Härdle 96) et le modèle est plus explicite. Après estimation, qui peut être vue comme une estimation préliminaire qui capture la forme des fonctions, on peut post-traiter les estimateurs pour instiller de l'information paramétrique.
- la modélisation additive a permis de préciser ou confirmer les aspects statistiques suivants concernant le modèle non linéaire général : l'effet de la température est le plus important, l'effet de la variable endogène (terme de persistance, influence de la pollution d'un jour sur celle du lendemain) est le moins important, les effets du climat varient peu géographiquement (deux groupes de stations se dégagent, ils distinguent «ville» et «campagne») et les effets d'interaction ne sont pas importants puisque les performances des deux modèles non linéaires sont très comparables.

- les méthodes non linéaires ont tendance à délivrer des prévisions dont la variabilité spatiale est en général moins importante que celles des mesures. Contrairement à la méthode de classification-discrimination-régression (et dans une moindre mesure CART) dont la tendance est plutôt de bien suivre cette variabilité spatiale. Ceci vient du fait que les modèles sont ajustés finement station par station dans le dernier cas et sans utiliser les informations sur le climat réalisé, alors que dans le premier l'effet du climat réalisé est très fort et gomme plus la variabilité géographique.

3. Comparaison des résultats des divers modèles

3.1. Approche globale et décisionnelle

3.1.1. Schéma des plans d'essais

Les paramètres de chaque méthode sont réglés indépendamment sur un ensemble de jours appelé ensemble d'apprentissage, par des optimisations qui ne sont pas présentées ici. Les réglages sont parfois longs à conduire et dépendent du critère d'optimisation.

Une fois les paramètres ajustés, on fait pour chaque méthode, une prévision sur un ensemble de jours identique pour toutes les méthodes, appelé échantillon test.

Les comparaisons des méthodes sont faites par un outil de dépouillement multi-critères commun.

– Ensemble d'apprentissage

L'ensemble d'apprentissage, pour toutes les méthodes, correspond à la période du 01.01.92 au 31.12.96. Les ensembles d'apprentissage et de test sont disjoints. Ce choix classique nous place dans une situation réaliste : en effet les jours, où le maximum de la concentration d'ozone est prévu, n'interviennent pas dans le réglage des modèles de prévision.

– Echantillon test

Sept échantillons tests sont définis : trois sont des étés entiers, trois sont des quinzaines et le dernier est un échantillon stratifié construit par un tirage au sort :

- étés entiers :
  - été 1994 (pollué : *6* alertes 2, *15* alertes 1);
  - été 1995 (pollué : *4* alertes 2, *22* alertes 1);
  - été 1996 (peu pollué : *0* alerte 2, *17* alertes 1).
- quinzaines
  - 01-07-94 au 15-07-94 (polluée : *3* alertes 2, *3* alertes 1);
  - 01-08-95 au 15-08-95 (polluée : *1* alerte 2, *8* alertes 1);
  - 01-08-96 au 15-08-96 (peu polluée : *0* alerte 2, *1* alerte 1).
- échantillon stratifié de 33 jours (*2* alertes 2, *11* alertes 1).

– Critères de comparaison

En dehors des critères classiques portant sur la validation et l'évaluation des résultats de prévision (comme par exemple la moyenne et les quantiles des erreurs absolues), des critères adaptés au problème décisionnel qui nous intéresse sont déduits du tableau d'alertes suivant :

		Prévu		
		Niveau 0	**Niveau 1**	**Niveau 2**
Réalisé	**Niveau 0**	Bonne alerte	Fausses alertes	
	Niveau 1		Bonne alerte	
	Niveau 2	Alertes non faites		Bonne alerte

On trouve en ligne les niveaux (si l'on raisonne station par station) ou les alertes réalisés et en colonne les niveaux (ou alertes) prévus. On répartit ainsi les jours considérés dans ce tableau. Le tableau idéal est diagonal. Le bloc au dessus de la diagonale représente les fausses alertes tandis que le bloc sous la diagonale concerne les alertes non faites. L'écrasante majorité des jours sont dans la case 0-0.

Partant de la répartition des jours dans ce tableau, on peut considérer des indicateurs scalaires de deux types. Le premier fournit un coût global et il est calculé comme la somme des produits terme à terme du tableau d'alertes et d'un tableau de coefficients reflétant le coût associé à chacune des configurations. Le second consiste à comptabiliser les occurences de certains événements pertinents.

Deux types de critères ont donc été retenus :

- un type de critère sur les tableaux d'alertes. On se donne trois matrices de coûts symétriques.

	Coût uniforme			**Coût Santé Publique**			**Coût préfecture**		
	Niv 0	**Niv 1**	**Niv 2**	**Niv 0**	**Niv 1**	**Niv 2**	**Niv 0**	**Niv 1**	**Niv 2**
Niv 0	0	1	1	0	1	3	0	1	3
Niv 1	1	0	1	1	−2	2	1	0	2
Niv 2	1	1	0	3	2	−3	3	2	0

Le critère de coût uniforme pénalise, à poids égal, les fausses alertes ou les alertes non faites et n'avantage pas les bonnes alertes. Le critère de coût «santé» pénalise davantage les fausses alertes et les alertes non faites au niveau 2 par rapport au niveau 0 et avantage les alertes bien prévues de niveaux 1 et 2. Le critère de coût «préfecture» pénalise, comme pour le coût «santé», les fausses alertes et les alertes non faites mais sans privilégier les bonnes alertes.

- un type de critère sur des composites synthétiques : ces critères, comparant les valeurs de la concentration d'ozone prévue et réalisée, sont calculés à partir d'un tableau de contingence binaire de la forme suivante :

	Prévu 0	**Prévu 1 ou 2**
Réalisé 0	t_{00}	t_{01}
Réalisé 1 ou 2	t_{10}	t_{11}

- un critère représente le taux de non détection, c'est-à-dire le nombre d'alertes non faites sur le nombre d'alertes :

$$t_1 = \frac{t_{10}}{t_{11} + t_{10}} \tag{6}$$

- un critère représente le taux de fausses alertes, c'est-à-dire le nombre de fausses alertes sur le nombre d'alertes :

$$t_2 = \frac{t_{01}}{t_{11} + t_{01}} \tag{7}$$

- mentionnons un autre critère appelé le threat score largement utilisé en météorologie et en pollution, défini par le nombres d'alertes faites sur le nombre d'alertes prévues ou réalisées :

$$t_3 = \frac{t_{11}}{t_{11} + t_{10} + t_{01}} \tag{8}$$

Plus les valeurs des deux premiers critères sont petites et meilleure est la prévision. C'est l'inverse pour le dernier critère. Aucun de ces critères ne tient compte de la case 0-0, ce qui est naturel puisque son introduction gonflerait artificiellement les performances en masquant l'évaluation de la capacité à minimiser les fausses alertes et les alertes non faites.

3.1.2. Ensemble des modèles

Chacune des méthodes a été utilisée pour prévoir le maximum d'ozone sur les sept échantillons tests en calant le modèle sur l'ensemble d'apprentissage exempté de la période test correspondante. Pour chacune de ces périodes, nous avons déterminé les jours de prévision communs à toutes les méthodes, ceci est nécessaire à cause de la présence de données manquantes (non interpolées) qui a des impacts différents sur les méthodes. Puis nous avons calculé pour chacun les critères de comparaison et octroyé un point par meilleur critère obtenu (le nombre de points maximum pouvant être atteint par une méthode étant de 63). Les résultats sont détaillés dans le rapport Bel *et al.* 97 et peuvent être résumés comme suit.

Les méthodes sont assez proches pour les critères de coûts d'alertes, exceptée CART assez nettement distancée. En revanche, pour les critères de mesures d'erreurs, la méthode classification-discrimination-régression domine plus nettement, ainsi que pour la comparaison des taux t_1 et t_2, basés eux aussi sur les tableaux d'alertes. Cependant la méthode additive occupe une seconde place générale.

Observons sur des périodes polluées, les tableaux des alertes prévues/réalisées :

- été 1994 (*6* alertes 2, *15* alertes 1); nombre de jours communs à toutes les méthodes : 43.

	Non linéaire général			**Non linéaire additif**			**CDR**		
	Niv 0	**Niv 1**	**Niv 2**	**Niv 0**	**Niv 1**	**Niv 2**	**Niv 0**	**Niv 1**	**Niv 2**
Niv 0	30	1	0	28	3	0	24	7	0
Niv 1	6	2	0	3	4	1	2	6	0
Niv 2	3	1	0	2	1	1	0	4	0

	CART		
	Niv 0	**Niv 1**	**Niv 2**
Niv 0	30	1	0
Niv 1	7	1	0
Niv 2	3	1	0

Sur cette première période, seule la méthode additive prévoit une alerte de niveau 2 mais on observe deux sous-estimations de niveau 2 en niveau 0. Il en est de même pour les méthodes non linéaire générale et CART qui, elles, en sous-estiment 3 en niveau 0.

La méthode CDR, si elle prévoit un plus grand nombre de niveau 1, ne prévoit aucun niveau 2. Ces niveaux 2 réalisés sont cependant tous prévus en niveau 1.

- échantillon stratifié de 33 jours (*2* alertes 2, *11* alertes 1); nombre de jours communs à toutes les méthodes : 24.

	Non linéaire général			**Non linéaire additif**			**CDR**		
	Niv 0	**Niv 1**	**Niv 2**	**Niv 0**	**Niv 1**	**Niv 2**	**Niv 0**	**Niv 1**	**Niv 2**
Niv 0	15	1	0	14	2	0	12	4	0
Niv 1	6	1	0	6	1	0	4	3	0
Niv 2	1	0	0	1	0	0	0	1	0

	CART		
	Niv 0	**Niv 1**	**Niv 2**
Niv 0	14	2	0
Niv 1	7	0	0
Niv 2	1	0	0

Sur la seconde période, aucune méthode ne prévoit le niveau 2; seule la méthode CDR le sous-estime en niveau 1, les autres en niveau 0. De même, malgré quatre fausses alertes de niveau 0 en niveau 1, c'est la méthode CDR qui prévoit le plus grand nombre de niveau 1.

3.1.3. Comparaison avec la méthode de persistance

Adopter la persistance comme méthode de prévision consiste à prévoir le maximum d'ozone d'un jour j par le maximum d'ozone réalisé au jour $j-1$. La comparaison avec une telle méthode peut surprendre, mais elle est néanmoins classique dans ce domaine difficile. Elle donne des résultats dont la qualité est médiocre mais pas catastrophique à cause de la relative stabilité des conditions météorologiques en région parisienne.

Regardons, sur les deux mêmes périodes et sur les jours en commun, les tableaux des alertes prévues/réalisées :

	Eté 1994			Echantillon stratifié		
	Persistance			Persistance		
	Niv 0	Niv 1	Niv 2	Niv 0	Niv 1	Niv 2
Niv 0	24	4	3	10	5	1
Niv 1	4	4	0	5	2	0
Niv 2	1	2	1	1	0	0

Les défauts de cette prévision sont de faire des fausses alertes de niveaux 2 sur des niveaux 0 de pollution et de sous-estimer des niveaux 2 en niveaux 0. Fort logiquement, toutes les méthodes exposées précédemment sont nettement meilleures que la prévision par persistance.

3.1.4. Les modèles opérationnels

Pour conclure sur cette comparaison sur laquelle sont basés les modèles opérationnels, on peut noter que le threat score est environ de 35 à 40 %. Ces performances, apparemment modestes, sont comparables aux résultats des prévisions qui se pratiquent ailleurs dans le monde pour ce même polluant. L'ordre de grandeur des moyennes des erreurs absolues est de 20 à 27 $\mu g/m^3$; aucune des méthodes ne prévoit la plupart des alertes de niveau 2. Globalement toutes ces méthodes sont équivalentes au vu de ce protocole de comparaison.

3.2. Les résultats sur l'été 1997

Examinons enfin les résultats sur l'été 1997. Cette période a été assez polluée : 16 alertes de niveau 1 et 4 de niveau 2 ont été relevées sur 138 jours étudiés. Ce dernier été a constitué une période de validation en vraie grandeur des méthodes mises au point et des procédures opérationnelles associées. Les prévisions que nous avons élaborées ont été utilisées pour informer la population par le biais du service minitel 3614 Airparif.

– Disponibilité de la prévision

La période durant laquelle les modèles ont été opérationnels, comporte en tout 138 jours. Sur ces 138 jours, la méthode additive en a prévu 136 et la méthode non linéaire 131. Les méthodes classification-discrimination-régression et CART, utilisant beaucoup plus de variables, en ont prévu respectivement 124 et 122; les jours de données manquantes étant malheureusement des jours où des niveaux 2 ont été atteints.

Sur l'ensemble des jours de prévision de chacune des méthodes, la moyenne des erreurs absolues est bonne (environ de 20 $\mu g/m^3$) mais cependant un peu plus élevée pour la méthode non linéaire générale.

– Tableaux d'alertes et Threat score

En regardant les tableaux d'alertes fournis par chaque méthode, on se rend immédiatement compte que la méthode additive a fait dans l'ensemble de meilleures prévisions que les autres.

	Non linéaire additif			**CART**			**CDR**		
	Niv 0	**Niv 1**	**Niv 2**	**Niv 0**	**Niv 1**	**Niv 2**	**Niv 0**	**Niv 1**	**Niv 2**
Niv 0	115	2	0	103	3	0	102	7	0
Niv 1	5	9	1	10	5	0	7	7	0
Niv 2	0	2	2	1	0	0	0	1	0

	Non linéaire général		
	Niv 0	**Niv 1**	**Niv 2**
Niv 0	111	1	0
Niv 1	11	4	0
Niv 2	2	2	0

En effet, sur les quatre niveaux 2 réalisés, elle a fait deux bonnes alertes et deux sous-estimations en niveaux 1. Ces deux sous-estimations sont valables également pour la méthode non linéaire générale mais elle commet une erreur plus importante en prévoyant les deux autres niveaux 2 en niveaux 0. Les méthodes CART et CDR ont pu prévoir un seul de ces quatre jours et elles l'ont toutes les deux sous-estimé en niveau 1.

En ce qui concerne les niveaux 1, la méthode additive en prévoit correctement neuf; deux de plus que la méthode CDR et quatre à cinq de plus que pour les deux autres méthodes.

De plus la méthode additive fait deux fois moins de sous-estimations (cinq au total) de niveau 0 pour un niveau 1 de réalisé que les méthodes CART et non linéaire générale; la méthode CDR en commettant 7.

En ce qui concerne les fausses alertes de niveau 1 pour des niveaux 0 réalisés, les résultats sont corrects (entre 1 et 3) pour les méthodes additive, CART et non linéaire générale, la méthode CDR en faisant sept.

Tous ces commentaires se résument assez bien avec le Threat score de chacune des méthodes. Le meilleur est pour la méthode additive avec 66.7 % ce qui est environ le double des Threat scores des autres modèles variant entre 26 et 36 %.

– Conclusions

Les méthodes donnant les meilleurs résultats sur cette phase opérationnelle sont la méthode additive et dans une moindre mesure la méthode CDR.

La méthode additive grâce à une meilleure disponibilité de la prévision, a pu faire plus de prévisions du fait du nombre plus restreint de variables et elle a commis moins de fausses alertes et de sous-estimations.

3.3. Approche clinique sur quelques jours

En complément des approches statistiques globales qui ne mesurent qu'une qualité moyenne, nous recherchons des causes de mauvaise prévision par des analyses cliniques jour par jour. Le mois d'août est le mois le plus pollué de l'été 1997 avec 11 niveaux 1 et 4 alertes de niveaux 2. Analysons quelques jours plus précisément.

– 18 août, niveau 1 réalisé :

Seules les méthodes additive et CART l'ont bien prévu; les méthodes CDR et non linéaire générale l'ayant sous-estimé en niveau 0.

Pour la méthode non linéaire générale, cette erreur est due à la sous-estimation de la température 26.6 °C contre 32 °C réalisée; la prévision du vent étant convenable. La sous-estimation de la température est due au gradient de température. Le prédicteur intermédiaire a depuis été modifié pour corriger cet effet.

Pour la méthode CDR, cette sous-estimation provient de la variable gradient de température qui chute progressivement à partir de 23h.

– 12 août, niveau 2 réalisé :

Seules les méthodes additive et non linéaire générale ont pu faire une prévision. Elles sous-estiment en un niveau 1 pour la méthode additive et en un niveau 0 pour la méthode non linéaire générale.

L'erreur provient de la faible prévision de température 29.8 °C contre 32 °C réalisée; la prévision du vent étant bonne. Pour valider cette idée avec la méthode additive, la prévision a été refaite avec les réalisations climatiques (température et vitesse de vent), et la prévision de la pollution reste au niveau 1.

– 17 août, niveau 2 réalisé :

Seule la méthode additive l'a bien prévu; les trois autres méthodes ayant commis des sous-estimations, de niveau 1 pour les méthodes CDR et non linéaire générale et de niveau 0 pour la méthode CART.

Pour la méthode non linéaire générale, l'erreur de prévision est surprenante étant donné la prévision de température 30 °C contre 28.9 °C réalisée et la prévision du vent 1.4 m/s contre 2.6 m/s de réalisé, plutôt favorables à un niveau 2. Ce défaut a été

corrigé en adaptant différemment les fenêtres de l'estimateur à noyau dans les zones de faible densité.

– 22 août, niveau 0 réalisé :

Trois des méthodes (l'additif, CART et la CDR) ont fait sur ce jour une surestimation de niveau 1. Toutefois, en regardant les conditions climatiques réalisées (une température de 31 °C et un vent de 4 m/s), on peut dire que la prévision d'un niveau 1 était raisonnable, d'autant plus que quatre stations ont mesuré des concentrations d'ozone supérieures à 120 $\mu g/m^3$. C'est un jour que l'on peut considérer comme atypique puisque c'est un jour de niveau 0, précédé par un jour de niveau 2 et suivi par un jour de niveau 1.

4. Conclusion

Les prévisions obtenues sont de qualité comparable aux résultats des équipes internationales (cf. Fromage 96). Nos prévisions constituent la matière première des informations diffusées sur le minitel 3614 Airparif.

Néanmoins il reste de nombreuses voies d'amélioration à explorer. Sans entrer dans les détails, citons-en rapidement deux.

Nous avons vu, au cours de l'analyse des résultats de chaque méthode, que les niveaux 1 sont assez souvent bien prévus mais que les niveaux 2 ne le sont que rarement. L'amélioration de la prévision des niveaux 2 est donc primordiale; peut-être faudrait-il pondérer davantage les jours observés pollués pour compenser leur faible fréquence dans la base de données.

On pense que l'amélioration majeure proviendra de nouvelles variables météorologiques, telles que la direction du vent, la pression et l'humidité ou la hauteur de la couche de mélange.

Le travail futur consistera à prévoir quatre polluants (ozone, dioxyde d'azote, dioxyde de soufre et les poussières) un jour pour le lendemain.

Références

[1] BEL L., BELLANGER L., BONNEAU V., CIUPERCA G., DACUNHA-CASTELLE D., DENIAU C., GHATTAS B., MISITI M., MISITI Y., OPPENHEIM G., POGGI J-M., TOMASSONE R. (1997). *Prévisions des pointes de pollution dans la région parisienne, O_3 et NO_2 : phase opérationnelle* . Rapport de contrat de recherche Airparif, oct., 193 p.

[2] BREIMAN L., FRIEDMAN J. H., OLSHEN R. A., STONE C. J. (1984). *Classification and Regression Tree*. Wadsworth, Belmont, CA.

[3] BRION D., GILIBERT E. (1997). *Forecasting Atmospheric Pollution Peaks over the Ile-de-France*. Preprint, 36 p.

[4] CLARK T.L., KARL T.R. (1982). *Application of prognostic meteorological variables to forecasts of daily maximum one-hour ozone concentration in the northeastern United States*. Journ. Applied Meteo., vol. 21 (11), 1662–1671.

[5] EDER B.K., DAVIS J.M., BLOOMFIELD P. (1994). *An automated classification scheme designed to better elucidate the dependence of ozone on meteorology*. Journal of Applied Meteorology, vol. 33, 1182–1198.

[6] FROMAGE A. (1996). *Prévision des pointes de pollution atmosphérique : état de l'art dans le monde et perspectives pour la région Ile-de-France*. Mémoire Ecole des Mines de Paris.

[7] HÄRDLE W. (1990). *Applied nonparametric regression*. Cambridge Univ. Press.

[8] HASTIE T., TIBSHIRANI R. (1990). *Generalized additive models*. Chapman & Hall.

[9] JIN S., DEMERJIAN K. (1993). *A photochemical box model for urban air quality study*. Atmos. Environ., 27B, 371–387.

[10] KARL T.R. (1979). *Potential application of model output statistics (MOS) to forecasts surface ozone concentrations*. Journ. Applied Meteo., vol. 18 (3), 254–265.

[11] LINTON O.B., HÄRDLE W. (1996). *Estimation of additive regression models with known links*. Biometrika, vol. 83, 529–540.

[12] MILIONIS A.E., DAVIES T.D. (1994). *Regression and stochastic models for air pollution, I and II*. Atmospheric Environment, 28 (17), 2801–2822.

[13] National Research Council (1991). *Rethinking the ozone problem in urban and regional air pollution*. National Academy Press, Washington, D.C.

[14] POGGI J-M. (1994). *Prévision non paramétrique de la consommation électrique*. Revue de Statistique Appliquée, XLII (4), 83–98.

[15] POULOS G.S., PIELKE R.A. (1994). *A numerical analysis of Los Angeles basin pollution transport to the Grand Canyon under stably stratified, southwest flow conditions*. Atmospheric Environment, 28 (20), 3329–3357.

[16] RYAN W.F. (1995). *Forecasting severe ozone episodes in the baltimore metropolitan area*. Atmospheric Environment, 29 (17), 2387–2398.

[17] SEINFELD J.H. (1988). *Ozone air quality models : a critical review*. JAPCA, vol. 38, 616–645.

[18] SHEIFINGER H., STOHL A., KROMP-KOLB H., SPANGL W. (1996). *A statistical method for predicting daily maximum ozone concentrations*. Gefahrstoffe-Reinhaltung der Luft, vol. 56, 133–137.

[19] TOMASSONE R., DERVIN C., MASSON J-P. (1993). *Biométrie, modélisation des phénomènes biologiques*. Masson, Paris.

[20] TOUPANCE G. (1988). *L'ozone dans la basse troposphère. Théorie et pratique*. Univ. de Créteil.

[21] VAUTARD R., BEEKMANN M., HONORÉ C. (1997). *Inverse modeling of pollution episodes over the Paris area using a photochemical box model*. Air pollution modelling, Bologne, 16-18 sept., proceedings in press.

[22] VENABLES W.N., RIPLEY B.D (1994). *Modern Applied Statistics with S-Plus*. Springer-Verlag.

[23] WILKS D.R. (1995). *Statistical methods in the atmospheric sciences : an introduction*. Academic Press.

[24] WALCEK C.J., YUAN H.H. (1995). *Calculated influence of temperature-related factors on ozone formation rates in the lower troposphere*. J. of Applied Meteorology, vol. 34, 1056–1069.

[25] WIPIJ D., SALLY LIU L-J. (1994). *Prediction models for personal ozone exposure assessment* in «Case Studies in Biometry». Lange N. *et al.* (Ed) Wiley, 41–56.