

L'ozone dans les Bouches-du-Rhône et une méthodologie de prévision

The ozone in the French Department of Bouches-du-Rhône and a forecasting methodology

Badih GHATTAS*, Laetitia MARY**, Patricia RENZI**, Dominique ROBIN**

Résumé

La prévision statistique du maximum journalier d'ozone sur les stations permanentes de mesure de l'aire métropolitaine marseillaise a été initiée en 1997 par l'équipe de statistiques du GREQAM* et AIRMARAIX** dans l'objectif d'anticiper l'information du public selon la procédure réglementaire du « décret ozone » de 1996.

Deux modèles statistiques basés sur la méthode CART (Classification And Regression Trees, arbres de régression et de classification) sont proposés pour donner une prévision par station et par groupe de stations. Les modèles correspondants s'appuient sur des prédicteurs environnementaux tels que les facteurs météorologiques et les paramètres ayant trait aux niveaux de pollution. La méthodologie insiste sur le choix et la préparation de la base de données (historique de cinq ans), mais aussi sur le développement inédit d'outils théoriques connexes à la méthode, permettant la stabilisation et l'optimisation des résultats.

Ces modèles, implantés à AIRMARAIX depuis mai 1999, ont été testés et validés sur l'été 1998, et mis en opérationnel durant l'été 1999.

Cet article débute par une description des spécificités météorologiques et des émissions particulières du département des Bouches-du-Rhône. La suite présente succinctement la méthodologie puis les résultats des modèles de prévision du maximum d'ozone quotidien par station de mesure et par groupe de stations définissant les zones homogènes d'information du département.

Abstract

In this paper we present the *statistical* approach used in order to forecast the daily maximum ozone concentration at the Bouches-du-Rhône department in France. We first describe the meteorological and emissions specificity of this region. Then we suggest a statistical approach to deal with this problem using three steps.

Two main statistical models are suggested based on CART (Classification And Regression Trees) methodology, giving a prediction over each monitoring site. Finally, two common strategies based on predictions results are tested. Such strategies will help the departmental authorities to make a decision, whether to launch ozone alert over the Bouches-du-Rhône department or not.

La pollution photochimique dans les Bouches-du-Rhône

La pollution photochimique

La pollution photochimique est une pollution secondaire liée à l'oxydation de polluants primaires ou précurseurs [oxydes d'azote (NO_x) et composés organiques volatils (COV)] sous l'effet du soleil.

Les sources de précurseurs dans l'atmosphère sont d'origine anthropique et naturelle. Les polluants primaires sont émis par des sources fixes (industries, stocks de solvants, d'hydrocarbures, de peintures...), par des sources mobiles (transports) mais aussi par les végétaux qui participent aux émissions de COV (terpènes...).

Les principaux composants de la pollution photochimique sont l'ozone, qui est mesuré en qualité d'indicateur, les aldéhydes, l'acide nitrique et des composés nitrés comme les PAN (peroxy acétyle nitrates) [1].

L'ozone est un gaz oxydant puissant dont les effets néfastes sur l'organisme humain et sur la végétation sont reconnus à des teneurs régulièrement atteintes dans les Bouches-du-Rhône durant l'été.

* GREQAM, Groupement de recherche en économétrie quantitative d'Aix-Marseille, Université de la Méditerranée.

** AIRMARAIX, Association pour la surveillance de la qualité de l'air de l'est des Bouches-du-Rhône, du Var et du Vaucluse.

Réglementation

D'un point de vue réglementaire, la directive européenne sur l'ozone du 21 septembre 1992, transcrite en droit français en avril 1996, fixe trois seuils vis-à-vis de la santé : 110 $\mu\text{g}/\text{m}^3$ sur huit heures (seuil de protection de la santé), 180 $\mu\text{g}/\text{m}^3$ sur une heure (seuil d'information de la population) et 360 $\mu\text{g}/\text{m}^3$ sur une heure (seuil d'alerte de la population). Elle énonce également des seuils pour la protection de la végétation : 65 $\mu\text{g}/\text{m}^3$ sur vingt-quatre heures et 200 $\mu\text{g}/\text{m}^3$ sur une heure.

Les seuils d'information et d'alerte sont repris dans un arrêté préfectoral des Bouches-du-Rhône qui précise les procédures opérationnelles d'information des autorités administratives et du public.

Trois zones d'information ont été définies sur les Bouches-du-Rhône :

- la zone **ouest** : ouest de l'étang de Berre ;
- la zone **nord-est** : nord-est de l'étang de Berre, dont Aix-en-Provence, et jusqu'à Cadarache ;
- la zone **sud-est** : sud-est du département, dont Marseille, Aubagne, et qui s'étend jusqu'à La Ciotat.

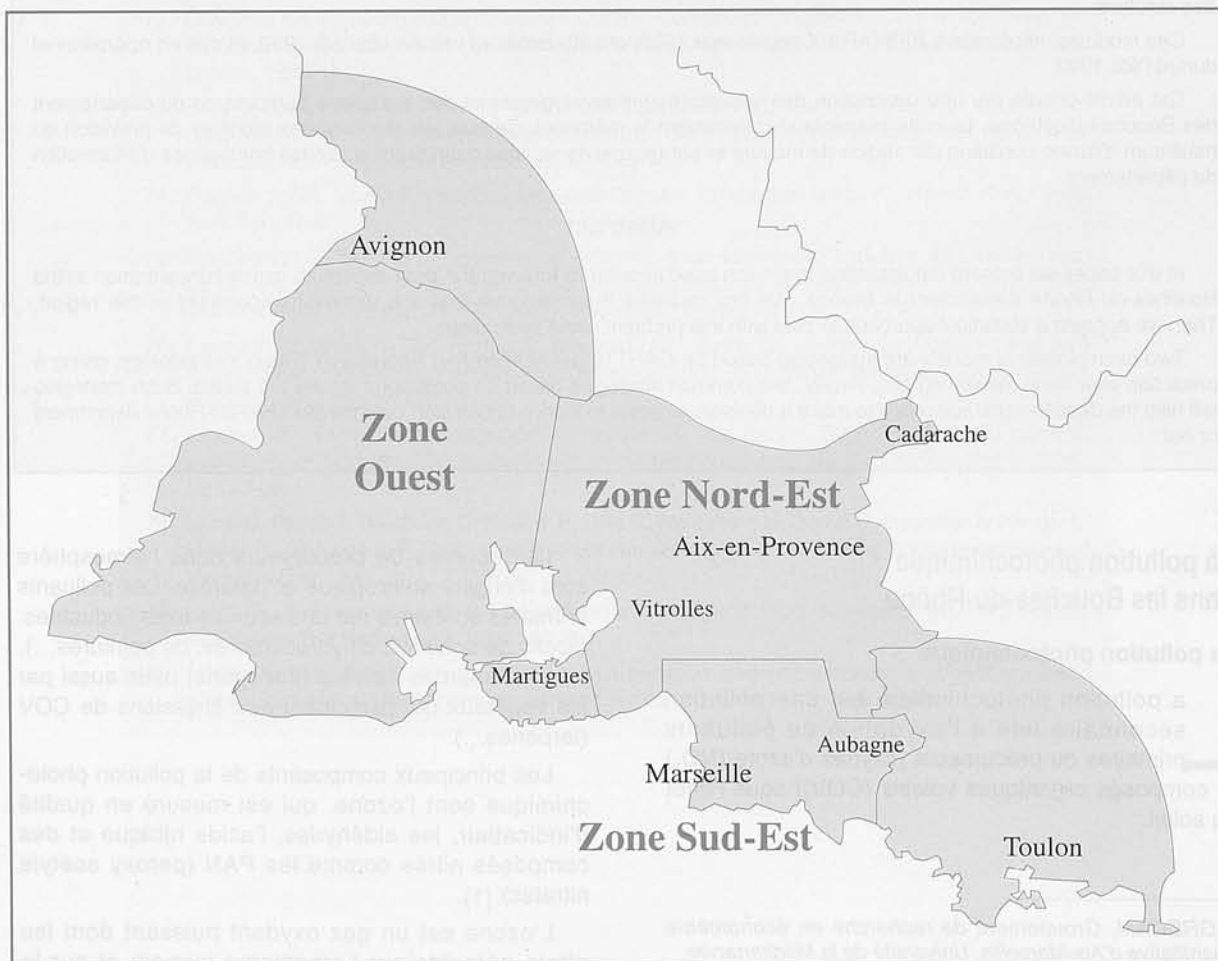
La procédure d'information et de recommandation, est mise en œuvre lorsque deux stations d'une même zone atteignent ou dépassent le seuil 180 $\mu\text{g}/\text{m}^3$ sur une heure ou si chacune des trois zones enregistre au moins un dépassement de ce seuil.

La procédure d'alerte est déclenchée lorsque au moins deux stations du département atteignent 360 $\mu\text{g}/\text{m}^3$ sur une heure.

Historique des mesures dans les Bouches-du-Rhône

Dans les Bouches-du-Rhône, la surveillance de l'ozone est effectuée depuis une quinzaine d'années. Le parc d'analyseurs d'ozone a fortement crû suite à la directive de 1992 et la Loi sur l'air de 1996. En 1998, 24 analyseurs permanents étaient installés dans le département.

Ce réseau de capteurs est géré par deux associations, AIRMARAIX sur l'est des Bouches-du-Rhône (Marseille, Aix-en-Provence, Aubagne) et AIRFOBEP sur le pourtour de l'étang de Berre (voir Carte de localisation ci-dessous).



Carte des différentes zones d'information des Bouches-du-Rhône.

Map of the various information areas of Bouches-du-Rhône.

Dans les Bouches-du-Rhône, le seuil de $180 \mu\text{g}/\text{m}^3$ sur une heure est atteint entre 1 jour sur 5 et 1 jour sur 3 suivant les années (maximums relevés en 1997 et 1998 avec respectivement 51 et 46 jours). Ce constat est à relier notamment à l'augmentation du parc d'analyseurs qui permet une meilleure couverture du département (surtout depuis 1997) (Tableau 1).

Les épisodes se déroulent de mai à septembre, parfois jusqu'au début du mois d'octobre, avec une prédisposition pour les mois les plus chauds (juillet et août : une dizaine de jours par mois en moyenne de dépassements de ce seuil de 1994 à 1998). Le mois de juillet 1994 a enregistré le maximum de journées relevant au moins un dépassement du seuil $180 \mu\text{g}/\text{m}^3$ sur une heure avec 20 jours, soit 2 jours sur 3.

Le contexte géographique et climatologique

Pour expliquer la fréquence élevée des dépassements du seuil $180 \mu\text{g}/\text{m}^3$, deux facteurs peuvent être mis en avant : **le climat méditerranéen** chaud et très ensoleillé et la présence de deux pôles de précurseurs importants avec **l'agglomération marseillaise** (1,2 millions d'habitants) et **le pôle industriel de l'étang de Berre** qui représente le quart des émissions nationales en matière de COV.

Les conditions météorologiques favorables au développement des processus photochimiques sont les **situations anticycloniques** accompagnées de **régimes de brises**. Ces conditions météorologiques engendrent des déplacements de masses d'air très caractéristiques et correspondent à une structure de l'atmosphère verticale bien particulière (inversion thermique nocturne).

Ces régimes se matérialisent par une alternance de brises deux fois par jour conditionnée par le

contraste thermique terre-mer. La brise d'Est durant la nuit transporte en mer les polluants émis sur la bande côtière. Avec l'ensoleillement, la brise d'Ouest s'établit et les polluants accumulés pendant la nuit en mer refluent vers le littoral et s'ajoutent alors à ceux émis sur la bande côtière. La pointe d'ozone est alors constatée sur la bande côtière, puis dans l'intérieur des terres jusqu'à près de 200 kilomètres des sites d'émission.

Le processus de formation de l'ozone est favorisé en périphérie de la partie agglomérée de Marseille où le rapport COV/NO_x est plus propice.

Les phénomènes côtiers de brises de mer-brises de terre sont relayés dans l'intérieur des terres par les brises de pentes qui suivent le même schéma la nuit et le jour que les précédentes. Ces brises s'expliquent également par des contrastes thermiques.

Les conditions favorables à l'établissement des processus photochimiques peuvent se résumer en quatre points : un ensoleillement fort, une température élevée ($>25^\circ\text{C}$), des conditions météorologiques peu dispersives (régime de brises) et bien sûr la présence de précurseurs (COV et NO_x).

Une journée polluée type

La journée du 15 août 1997 illustre un épisode typique de pollution photochimique (Figure 1, p. 416) **en régime de brises pures** (Figure 2, p. 416) sur le département des Bouches-du-Rhône.

Le phénomène se développe tout d'abord sur **la bande côtière**, puis transite sur **le Nord-Est** et les départements limitrophes. Les tons orangé-rouge traduisent le dépassement du seuil d'information de la population ($180 \mu\text{g}/\text{m}^3/\text{h}$), qui dans cet épisode concerne **l'ensemble du Nord-Est** du département

Tableau 1.
Évolution du nombre de jours de dépassements du seuil $180 \mu\text{g}/\text{m}^3$ sur une heure dans les Bouches-du-Rhône entre le 1^{er} mai et le 30 septembre de 1994 à 1998.

Evolution of the number of overtaking days of the $180 \mu\text{g}/\text{m}^3$ threshold per hour in Bouches-du-Rhône between May 1st and September 30th from 1994 to 1998.

	1994	1995	1996	1997	1998
Mai	1	3	3	6	9
Juin	10	4	12	4	11
Juillet	20	13	11	11	8
Août	10	5	4	13	16
Septembre	1	1	2	17	2
Total	42	26	32	51	46
% de jours sur 5 mois	27 %	17 %	21 %	33 %	30 %
Nombre de sites	4	8	16	23	24
Total sur les 4 sites existant depuis 1994	42	23	27	33	27

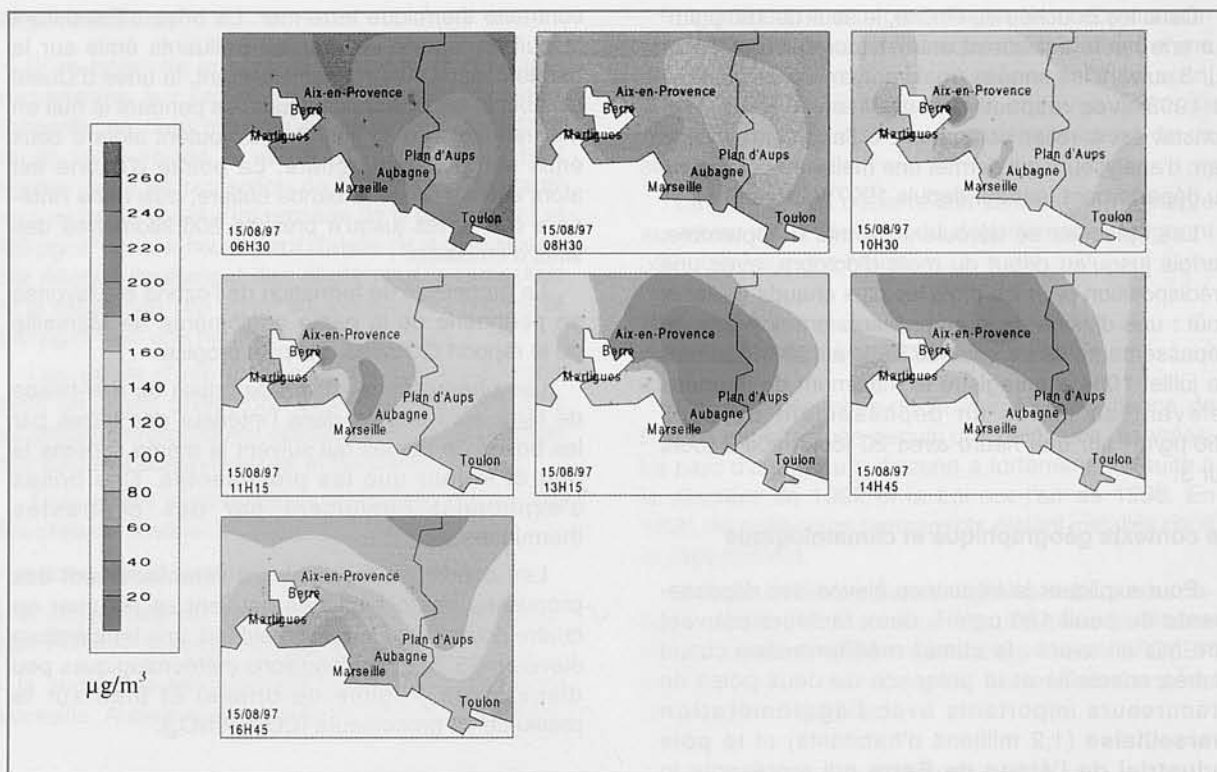


Figure 1.

Exemple d'épisode photochimique sur les Bouches-du-Rhône. Évolution des niveaux d'ozone le 15 août 1997.
Cartes réalisées par interpolation mathématique des données horaires d'ozone sur les stations de mesure des Bouches-du-Rhône.

Code de calcul : « Krigeage linéaire » – logiciel : Surfer – niveaux d'ozone en $\mu\text{g}/\text{m}^3$ – temps en heures TU.

Example of photochemical episode in Bouches-du-Rhône. Evolution of ozone levels on August 15, 1997.
Maps realized by mathematical interpolation of ozone hour datas measured in the Bouches-du-Rhône sites.
Code of calculation : « linear kriging » – Software : Surfer – Concentration units : $\mu\text{g}/\text{m}^3$ – Calculus in hours UT.

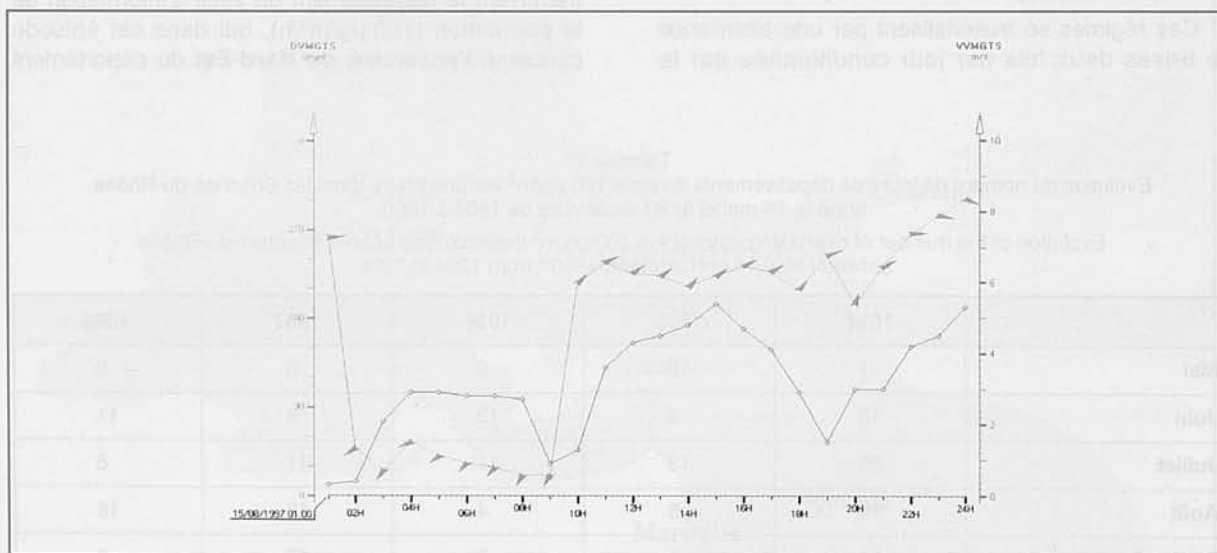


Figure 2.

Vitesse (VVMGTS) et direction du vent (DVMGTS) à la station Martigues-Gatasse durant la journée du 15 août 1997.
L'alternance et la direction des brises sont typiques : brise de terre (direction nord-est) ; brise de mer (direction sud-ouest).
L'évolution de la vitesse du vent est aussi très classique, vent faible pendant la nuit avec un minimum au moment de l'alternance des brises, puis un renforcement au cours de la journée.

Wind speed (VVMGTS) and direction (DVMGTS) in the Martigues-Gatasse site during the day of 15-08-1997.
Breezes alternation and direction are typical : earth breeze (Northeast direction) ; sea breeze (Southwest direction).
The evolution of the wind speed is also very classic, weak wind during the night with a minimum at the time of breezes alternation, then an intensification during the day.

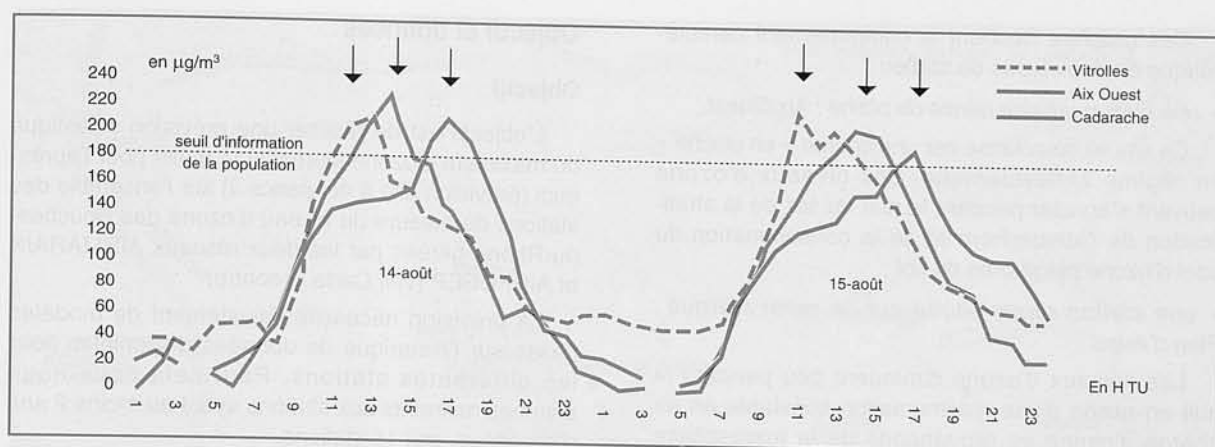


Figure 3.

Évolution des niveaux d'ozone sur trois stations des Bouches-du-Rhône selon un axe sud-ouest/nord-est, les 14 et 15 août 1997.

Ces deux journées se caractérisent par des régimes de brises. La chronologie de passage de la masse d'air du sud-ouest au nord-est est nettement visible : Vitrolles, Aix et Cadarache.

Evolution of ozone levels on three sites of Bouches-du-Rhône according to an axis Southwest/Northeast on August 14 and 16, 1997. These two days are characterized by regimes of breezes. The chronology of the air mass evolution from the Southwest to the Northeast is sharply visible : Vitrolles, Aix and Cadarache.

(Figure 1 ci-contre). Cette situation illustre bien le **caractère régional** de la pollution photochimique. Elle montre également que l'heure du maximum d'ozone est fonction de l'éloignement aux sites émetteurs. La figure 3 confirme ce point.

Des niveaux faibles la nuit et en début de matinée traduisent l'absence de photochimie et les réactions plus ou moins importantes de consommation de l'ozone (dépôt sec au sol et réactions avec les réducteurs tel que le monoxyde d'azote), puis une augmentation très rapide des niveaux sur la bande côtière. La masse d'air photochimiquement active transite ensuite

dans l'intérieur des terres, véhiculée par la brise de mer.

À noter sur la vignette de 6 h 30 TU (Figure 1), le comportement de l'ozone dans la zone de Plan d'Aups à l'est, qui se caractérise par la présence de niveaux plus élevés que sur le reste du département. Ce phénomène s'explique par l'influence du relief sur les conditions météorologiques locales. Ce dernier entraîne un brassage constant, ce qui empêche la stratification verticale de l'atmosphère. Le site est constamment réalimenté avec l'air de la troposphère libre (Figure 4).

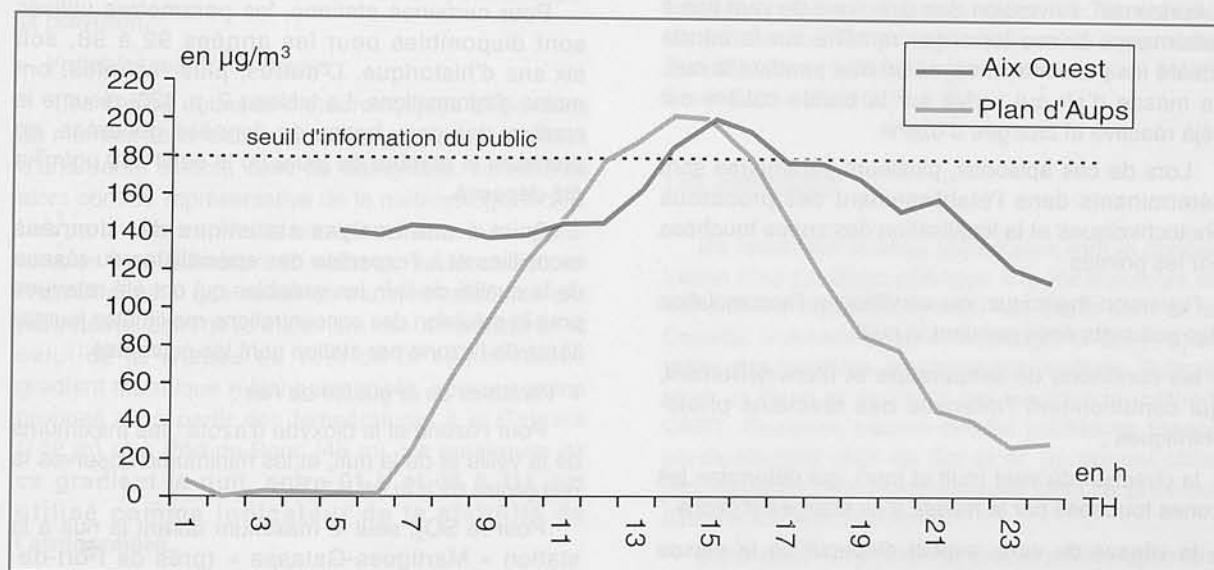


Figure 4.

Évolution des niveaux d'ozone à Aix-en-Provence et Plan d'Aups.

Evolution of ozone levels in Aix-en-Provence and Plan d'Aups.

Ces graphes illustrent le comportement caractéristique de deux types de station :

- *une station urbaine dense de plaine* : Aix Ouest.

Ce site se caractérise par une courbe « en cloche » en régime anticyclonique. Les niveaux d'ozone peuvent s'annuler pendant la nuit du fait de la stratification de l'atmosphère et de la consommation du pool d'ozone piégé près du sol ;

- *une station rurale située sur un relief marqué* : Plan d'Aups.

Les niveaux d'ozone diminuent peu pendant la nuit en raison d'une réalimentation constante en air chargé d'ozone en provenance de la troposphère libre [1]. La pointe traduit l'arrivée d'une masse d'air polluée.

Évolution de l'atmosphère au cours de la journée et paramètres déterminants dans le phénomène

L'apparition du soleil provoque une augmentation généralisée et hétérogène des niveaux d'ozone. Ce constat est lié à deux phénomènes principaux de l'atmosphère :

- *vertical* : le réchauffement du sol, lié au rayonnement, entraîne des mouvements de convection verticaux qui détruisent la stratification de l'atmosphère nocturne réalimentant le sol en air de la couche limite fossile dont le niveau en ozone à cette saison est de l'ordre de $100 \mu\text{g}/\text{m}^3$ dans la région. Ce brassage a également pour effet de ramener au sol les polluants stockés et accumulés pendant la nuit dans la couche intermédiaire. C'est surtout ce deuxième point qui explique le caractère « explosif » de la pointe dans le secteur de Berre/Vitrolles, puis de Marseille ;

- *horizontal* : l'inversion des directions de vent liée à l'alternance brises terre/mer ramène sur la bande côtière les polluants envoyés en mer pendant la nuit. La masse d'air qui arrive sur la bande côtière est déjà réactive et chargée d'ozone.

Lors de ces épisodes, plusieurs paramètres sont déterminants dans l'établissement des processus photochimiques et la localisation des zones touchées par les pointes :

- l'inversion thermique, qui conditionne l'accumulation des polluants émis pendant la nuit ;
- les conditions de température et d'ensoleillement, qui conditionnent l'intensité des réactions photochimiques ;
- la direction du vent (nuit et jour), qui détermine les zones touchées par la masse d'air chargée d'ozone ;
- la vitesse du vent, aspect dispersif de la masse d'air.

Ces paramètres sont exploités dans la prévision des pics de pollution par l'ozone.

Objectif et données

Objectif

L'objectif est de réaliser une prévision statistique du maximum d'ozone journalier le matin pour l'après-midi (prévision dite à échéance J) sur l'ensemble des stations de mesure du niveau d'ozone des Bouches-du-Rhône gérées par les deux réseaux AIRMARAIX et AIRFOBEP (voir Carte ci-contre).

La prévision nécessite l'ajustement de modèles basés sur l'historique de données disponibles pour les différentes stations. Pour cela nous sommes restreints aux stations ayant au moins 2 ans d'historique, soit 16 stations.

Pour chaque station nous avons recueilli un ensemble d'informations à partir desquelles nous avons construit les bases de données.

Les données

Une distinction a été faite entre la phase de construction des modèles et celle de leur évaluation. La première phase a consisté en la génération d'une base de données, dite d'apprentissage (ou de construction) utilisant les observations des périodes estivales de 1992 à 1997. Lors de la deuxième phase, la même base de données a été générée mais en utilisant les données de l'été 1998, période qui a donc servi de test pour les modèles.

Pour constituer les bases de données par station, des paramètres supposés être déterminants dans la formation de l'ozone ont été utilisés, notamment les paramètres météorologiques observés et prévus à Marignane (par Météo-France) et à d'autres stations de mesures des réseaux de la qualité de l'air. Des données concernant certains polluants (dioxyde de soufre (SO_2), dioxyde d'azote (NO_2) et ozone (O_3)) ont aussi été utilisées pour les 16 stations citées.

Pour certaines stations, les paramètres utilisés sont disponibles pour les années 92 à 98, soit six ans d'historique. D'autres, plus récentes, ont moins d'informations. Le tableau 2, p. 420, résume le contenu des deux bases de données générées, en précisant le nombre de jours où le seuil $180 \mu\text{g}/\text{m}^3$ a été dépassé.

Suite à une analyse statistique des données recueillies et à l'expertise des spécialistes du réseau de la qualité de l'air, les variables qui ont été retenues pour la prévision des concentrations maximales journalières de l'ozone par station sont les suivantes :

- *Variables de la qualité de l'air* :

Pour l'ozone et le dioxyde d'azote : les maximums de la veille et de la nuit, et les minimums observés la nuit (entre 01 h et 06 h TU).

Pour le SO_2 , seul le maximum durant la nuit à la station « Martigues-Gatasse » (près de Port-de-Bouc) a été retenu. En effet, il apparaît comme un **indicateur de la stabilité de l'atmosphère** car ce site est situé en altitude, et traduit ainsi, en régime de

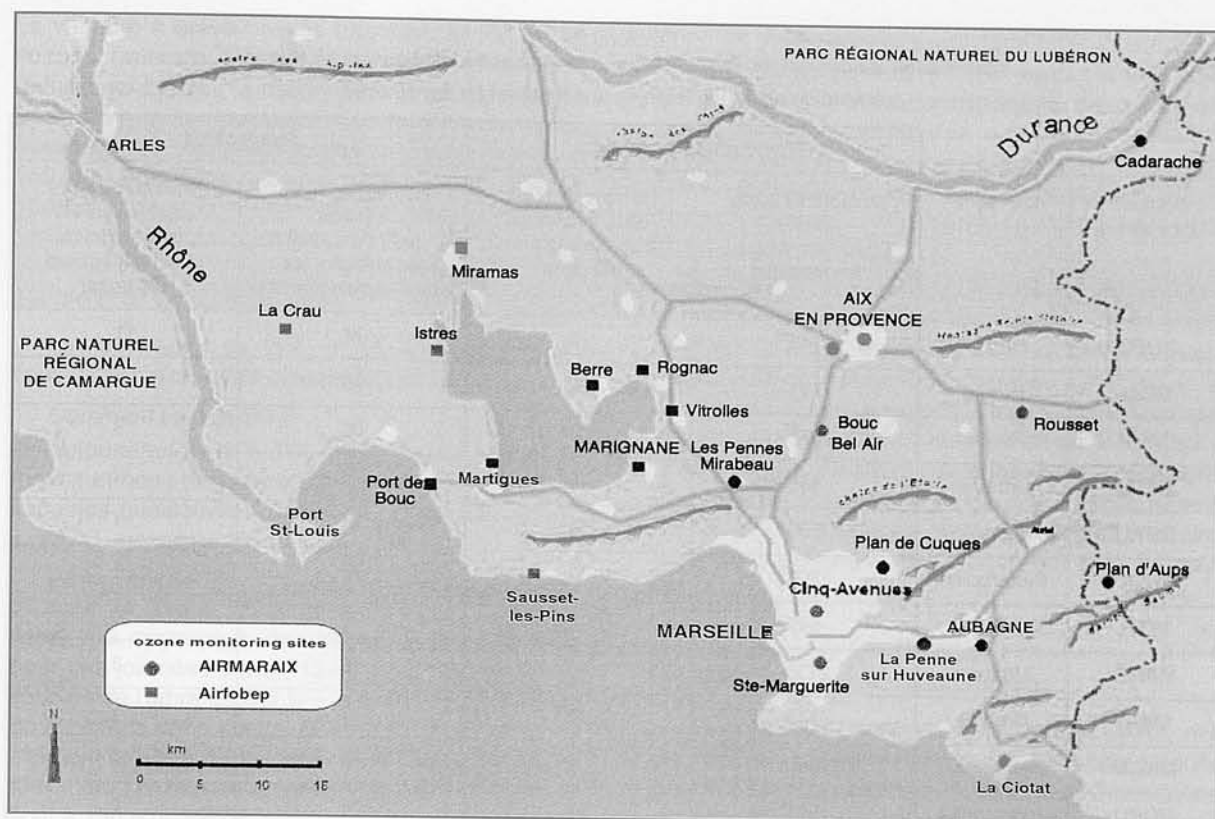


Figure 5.

Stations des Bouches-du-Rhône utilisées dans cette étude :
 Plan d'AUPS – « AUPS », Bouc Bel Air – « BBAI », Cadarache – « CADA », Cinq Avenues – « CINQ »,
 La Crau – « CRAU », Plan de Cuques – « CUQU », Istres – « ISTR », Sainte-Marguerite – « MARG »,
 Les Pennes-Mirabeau – « MIRA », Martigues – « MNDM », Port-de-Bouc – « PDBC », La Penne-sur-Huveaune –
 « PENH », Rognac – « RBRT », Rousset – « ROUS », Sausset-les-Pins – « SSLP », Vitrolles – « VTRL ».

Bouches-du-Rhône sites used in this study.

stabilité, l'enrichissement en polluants de la couche intermédiaire.

Ces paramètres sont fournis quotidiennement par les réseaux AIRMARAIX et AIRFOBEP pour effectuer la prévision.

• Variables météorologiques :

Les seules données météorologiques disponibles au moment de la réalisation de cette étude proviennent d'une seule station, celle de Marignane, considérée alors comme représentative de la météorologie locale. Les paramètres suivants ont été utilisés : la température, le vent (direction et vitesse), la nébulosité, et l'humidité, ainsi que certaines transformations de ces paramètres comme le maximum des températures et celui de la vitesse de vent de la veille. Aucun gradient thermique n'étant disponible, nous en avons proposé un à partir des températures à la Gatasse (142 m) et à Port-de-Bouc (45 m). **Le maximum de ce gradient la nuit, entre 01 h et 06 h TU, est utilisé comme indicateur de la stabilité de l'atmosphère.**

Pour la construction des modèles, ce sont les observations de ces paramètres entre 1 h et 18 h TU qui ont été utilisées.

La prévision

Il s'agit de prévision parfaite : l'étape de construction des modèles utilise les observations des variables météorologiques jusqu'à 18 h TU, et l'étape de prévision utilise les observations de ces mêmes variables jusqu'à 06 h TU et les prévisions de Météo-France entre 09 h et 18 h TU.

Les méthodes

De nombreux travaux statistiques liant la concentration d'ozone troposphérique à la météorologie ont été publiés essentiellement aux États-Unis et au Canada. Robeson et al. et Shefner et al. [2, 3] ont utilisé des modèles de régression multiple, Burrows et al. et Ryan et al. [4, 5] ont employé la méthode CART. Quelques travaux ont été publiés en France, principalement ceux de Bel et al. [6] qui ont utilisé des modèles non paramétriques pour la prévision d'ozone sur l'Ile-de-France.

Dans cette revue-même, une première approche de la prévision de l'ozone dans les Bouches-du-Rhône, où l'expertise du réseau est essentielle, a été publiée récemment [7].

Tableau 2.
Informations concernant les fichiers d'apprentissage et de test pour les 16 stations.
Information about the learning and test files for the 16 sites.

Stations de mesure	Catégorie du site	Fichier d'apprentissage		Fichier test	
		Nombre de jours utilisés pour construire les modèles	Nombre de jours où le maximum d'ozone dépasse $180 \mu\text{g}/\text{m}^3$	Nombre de jours valables pour l'été 1998 (en enlevant les journées manquantes)	Nombre de jours valables (1998) où le maximum observé dépasse $180 \mu\text{g}/\text{m}^3$
AUPS	Rurale	748	76	74	11
BBAI	Rurale	721	36	75	2
CADA	Rurale	577	33	60	3
CINQ	Urbaine	445	16	75	2
CRAU	Rurale	559	14	74	1
CUQU	Périurbaine	396	13	73	4
ISTR	Urbaine	349	2	74	0
MARG	Urbaine	338	10	75	3
MIRA	Urbaine	349	12	73	1
MNDM	Urbaine	340	5	75	3
PDBC	Industrielle	801	54	75	2
PENH	Périurbaine	403	17	74	5
RBRT	Industrielle	647	39	74	8
ROUS	Rurale	747	44	72	2
SSLP	Urbaine	796	33	73	6
VRTX	Industrielle	799	98	74	4

La relation entre le maximum d'ozone et les paramètres météorologiques ne semble avoir aucune propriété linéaire, ce qui explique le choix de la modéliser à l'aide de techniques non linéaires et non paramétriques.

Deux méthodes ont été utilisées. L'une est la méthode de base : c'est la méthode de régression et de classification par arbre (CART [8]), et l'autre, plus récente, agrège des prévisions obtenues par différents arbres de régression.

Les méthodes CART

Les méthodes CART sont maintenant populaires en raison de leur simplicité et de l'aspect visuel du modèle construit. Leur intérêt est multiple. Il s'agit de techniques de classification, de discrimination et de prévision, simples et performantes, mais aussi d'un outil de statistique descriptive très intéressant. Elles permettent d'utiliser des variables explicatives aussi bien quantitatives que qualitatives et ne limitent pas

leur nombre. En pratique, rajouter des variables explicatives à un tel modèle est sans inconvénient.

La visualisation des résultats à l'aide d'un arbre binaire permet la mise en évidence des variables explicatives essentielles pour la prévision du maximum quotidien de l'ozone pour la station considérée.

Introduction

Les méthodes de classification et de régression par arbre ont été popularisées par Breiman et al. [8] ; en langue française les travaux de Ghattas [9, 10] décrivent ces méthodes et les illustrent à l'aide d'exemples empruntés à la prévision du maximum quotidien d'ozone.

La suite présente les idées de la régression par arbre* en faisant référence à un exemple simplifié.

* Pour les arbres de classification, la variable à prévoir est qualitative (le maximum de l'ozone dans ce cas est découpé en quatre niveaux selon les seuils réglementaires).

La variable à prévoir est le maximum quotidien de l'ozone (maxo3), les variables explicatives sont réduites ici à trois : le maximum d'ozone de la veille (maxo3v), la température maximale de la veille (maxtcv), et le maximum du gradient vertical observé entre 00 h et 06 h (maxgrd). Un exemple d'arbre construit pour la station de Vitrolles avec ces variables est donné dans la figure 6, p. 422.

L'arbre et son utilisation

La visualisation de l'outil de prévision conduit à la construction d'un *arbre binaire* (Figure 6).

Comment ça marche ?

Aujourd'hui à 06 h TU, nous devons prévoir le maxo3 et nous disposons des observations des trois variables qualitatives : maxo3v = 140 $\mu\text{g}/\text{m}^3$, maxtcv = 25 °C, maxgrd = -1.

Dans l'ellipse du haut de l'arbre (la racine) on dispose de toutes les observations ($n = 822$ journées) et la moyenne du maximum quotidien de l'ozone pour ces journées est de 124,4 $\mu\text{g}/\text{m}^3$. La première coupure de l'arbre est sur le « maxo3v ». Le maximum de l'ozone la veille étant supérieur à 130,5 $\mu\text{g}/\text{m}^3$, on s'achemine dans la branche droite de l'arbre (si le maximum de l'ozone la veille avait été inférieur ou égal à 130,5 $\mu\text{g}/\text{m}^3$, on serait allé dans la branche de gauche). Le maximum du gradient de la nuit valant -1, on poursuit vers la droite (maxgrd > -1,3875), puis vers la droite (maxtcv = 25 °C > 22,9 °C), puis vers la gauche (maxo3v < 176). La feuille (entourée d'un rectangle) dans laquelle on se retrouve, indique la valeur 167,5 qui est la valeur maximale prévue de l'ozone pour le jour courant. Le nombre de journées qui se retrouvent dans la même situation vis-à-vis des variables considérées (et donc dans la même feuille) est $n = 53$.

C'est suivant l'application de ces règles que seront calculées les prévisions chaque jour, à l'aide du modèle visualisé par l'arbre.

Idée de la construction de l'arbre

La construction s'effectue à l'aide d'un échantillon dit d'apprentissage portant sur les observations quotidiennes des quatre variables par exemple, au cours des cinq dernières périodes estivales (de mai à septembre chaque année).

Dans la racine, se situent toutes les observations de l'échantillon d'apprentissage. La première division de l'échantillon en deux classes (maxo3v $\leq$ 130,5 et maxo3v > 130,5) est celle qui, de toutes les divisions possibles du même type obtenues à l'aide de chacune des trois variables explicatives, minimise le critère de la somme des variances de chacune des deux classes réalisées. Donc la division de l'échantillon en deux classes dépend d'une variable (ici maxo3v) et d'un seuil sur cette variable (ici 130,5).

Chacune de ces classes constitue un nœud de l'arbre (représenté par une ellipse dans la figure 6, p. 422) ; sur chacun de ces nœuds la même procé-

dure de division binaire est appliquée et conduit à d'autres divisions. On a ensuite recours à une règle d'arrêt de la procédure (détaillée dans [9]) qui conduit à la réalisation d'un arbre terminal dont le nombre de nœuds extrêmes (appelés *feuilles* de l'arbre et symbolisés par des rectangles sur la même figure) est réduit. Leur rôle dans le calcul de la prévision a été vu précédemment.

Suite à de nombreux essais de différentes variantes des méthodes CART, nous avons choisi d'utiliser les arbres de régression et ce en construisant les modèles sur la base de données « réduite » : la moitié des observations pour lesquelles le maximum observé d'ozone était en dessous du seuil 130 $\mu\text{g}/\text{m}^3$ a été retirée. En effet ces observations sont très nombreuses et la réduction de leur nombre permet d'obtenir une meilleure performance des modèles pour les journées plus polluées qui sont en petit nombre.

Combinaison de prédicteurs, agrégation d'arbres par bootstrap

L'inconvénient majeur des méthodes CART, mis en évidence par Breiman [11] est l'*instabilité* traduite par le fait qu'une légère perturbation de l'échantillon d'apprentissage peut conduire à un arbre (donc à une prévision) très différent.

La deuxième méthode proposée ici pour la prévision de l'ozone consiste à agréger des arbres (ou prédicteurs) construits sur des échantillons *bootstrap* de l'échantillon de construction.

Comment construire de tels échantillons ?

On tire au hasard, avec remise N observations dans l'échantillon de construction (lui-même de taille N).

Un modèle est alors construit sur chaque échantillon bootstrap (arbre de régression ou de classification). Le prédicteur agrégé est obtenu comme la moyenne arithmétique des prédicteurs obtenus sur les différents échantillons bootstrap.

Breiman a étudié les propriétés de ces algorithmes dans un contexte statistique. Il montre que l'agrégation de prédicteurs améliore la qualité de la prévision pour les arbres de régression, et donne un modèle plus stable, c'est-à-dire moins sensible aux perturbations des données.

Dans cette étude, 50 arbres de régression ont été agrégés par la méthode bootstrap. L'inconvénient est que le modèle agrégé ne peut plus être visualisé par un arbre.

Prévision par station et prévision par zone d'information **Stratégies de prévision**

L'information diffusée au public concernant la prévision de l'ozone résulte d'une combinaison des résultats des prévisions fournies par les modèles. Cette combinaison consiste à appliquer des « straté-

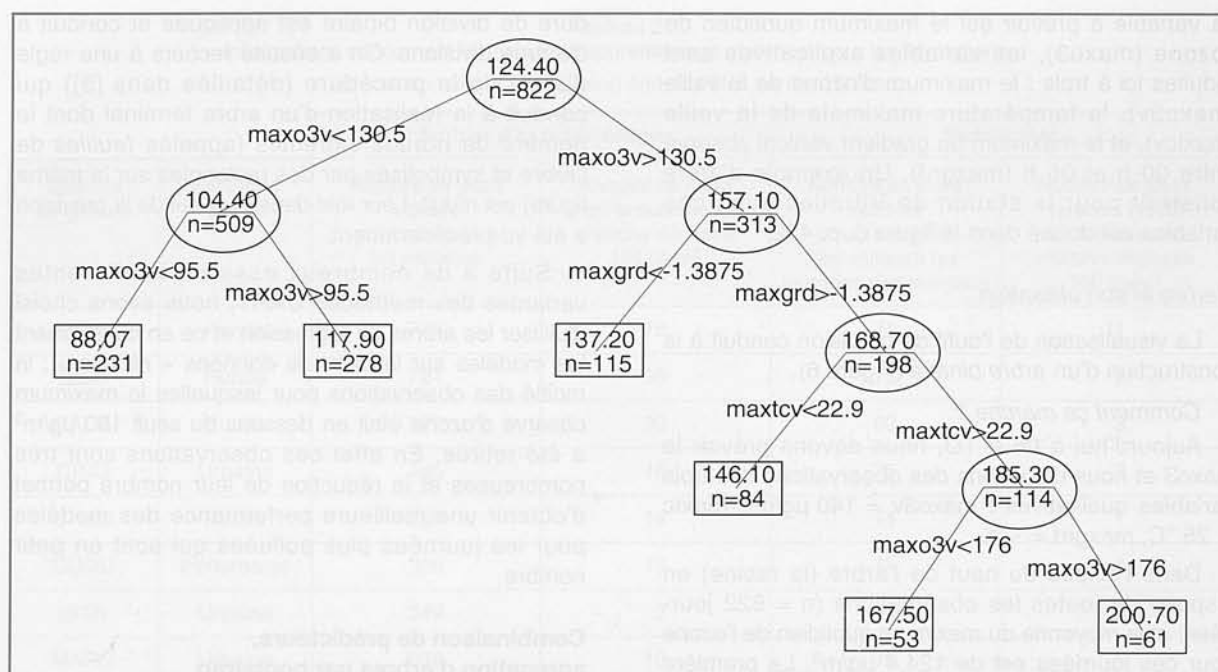


Figure 6.

Arbre de régression à 6 feuilles. Variable expliquée « maxo3 », variables explicatives : « maxo3v », « maxtcv » et « maxgrd ». Pour chaque nœud, la moyenne de la variable expliquée pour les observations dans ce nœud est indiquée ainsi que le nombre d'observations.

Regression tree with 6 sheets. Explained variable "maxo3", explicative variables "maxo3v", "maxtcv" and "maxgrd". For every knot, the average of explained variable for observations in this knot is indicated, as well as the number of observations.

gies », élaborées par le réseau et basées sur des zones géographiques bien déterminées.

Pour les Bouches-du-Rhône, trois zones d'information du public sont utilisées : zone ouest des Bouches-du-Rhône (zone 1), zone nord-est (zone 2) et zone sud-est (zone 3).

Zone 1 : « CRAU », « ISTR », « MNDM », « PDBC », et « SSLP ».

Zone 2 : « AUPS », « BBAI », « CADA », « RBRT », « ROUS », et « VTRL ».

Zone 3 : « CINQ », « CUQU », « MARG », « MIRA », et « PENH ».

Ces zones ont été obtenues suite à une expertise des réseaux et à une étude statistique préalable sur la classification des stations de mesure de l'ozone du département [12].

Cette typologie peut constituer un facteur décisionnel pour effectuer les prévisions station par station ou sur des zones déterminées a priori par ces résultats.

En se basant sur ces trois zones, voici les stratégies utilisées dans les Bouches-du-Rhône pour l'information, du public :

La première, dite stratégie *locale*, correspond au niveau d'information par zone, c'est-à-dire qu'un niveau d'information est prévu dans une zone lorsque, pour au moins deux capteurs d'une même zone, les prévisions dépassent le seuil $180 \mu\text{g}/\text{m}^3$ sur une heure.

La deuxième, dite *globale*, correspond au niveau départemental, qui est déclenché lorsque au moins un capteur de chacune des trois zones dépasse le seuil $180 \mu\text{g}/\text{m}^3$ sur une heure.

Les résultats obtenus en appliquant ces stratégies sont fournis dans le paragraphe suivant.

Résultats de la prévision sur l'été 1998

Sur l'ensemble de la période sur laquelle la prévision est testée (l'été 1998, voir Tableau 2, p. 420), certaines journées ont été retirées à cause de données manquantes ou de prévisions de paramètres météorologiques non fournies. *Notons que les jours dépassant $180 \mu\text{g}/\text{m}^3$ sont rares et donc difficiles à prévoir.*

Résultats des prévisions par station

Pour les deux modèles sélectionnés, les résultats de la prévision pour l'été 1998 sont montrés dans le tableau 3, ci-contre.

Le seuil $180 \mu\text{g}/\text{m}^3$ est considéré pour séparer les concentrations d'ozone en deux niveaux, le niveau 0 et le niveau 1 correspondant aux niveaux de concentrations, inférieure et supérieure à ce seuil.

L'erreur de classification donne la proportion des jours où le niveau est mal prévu. Les fausses alertes (FA) correspondent au nombre de jours où un niveau 1 est prévu mais un niveau 0 est observé, les alertes

Tableau 3.

Résultats de prévision pour les deux méthodes. L'échantillon test contient les jours de l'été 1998 d'avril à octobre.

TFC : taux de fausse classification, TS : Threat Score, FA : fausses alertes, AM : alertes manquées.

L'étoile* pour le Threat Score signifie qu'aucune observation de plus de 180 µg/m³ n'a été constatée, et aucune n'a été prévue.

Forecast results for the two methods. The test sample contains the days of the 1998 summer, from April to October.

TFC = Rate of False Classification, TS = Threat Score, FA = False Alert, AM = Missed Alerts.

The star* for the Threat Score means that no observation more of 180 µg/m³ was noted, and none was foreseen.

Stations	Agrégation bootstrap des arbres de régression avec la base de données complète				Arbres de régression avec pénalités avec la base de données réduite**			
	TFC	TS	FA	AM	TFC	TS	FA	AM
AUPS	32,6	16,7	4	11	46,7	17,4	9	10
BBAI	15,2	50,0	0	1	40,2	14,3	5	1
CADA	36,1	14,3	1	5	38,6	11,1	3	5
CINQ	16,3	0,0	0	3	35,9	11,1	6	2
CRAU	22,8	0,0	0	4	39,1	0,0	0	4
CUQU	28,3	0,0	2	5	32,6	0,0	2	5
ISTR	12,0	*	0	0	27,2	*	22	3
MARG	21,7	0,0	0	4	50,0	0,0	0	4
MIRA	20,9	0,0	0	1	36,3	0,0	0	1
MNDM	21,7	0,0	0	5	39,1	0,0	0	5
PDBC	27,2	0,0	1	4	34,8	14,3	3	3
PENH	25,0	0,0	1	6	41,3	11,8	11	4
RBRT	34,8	0,0	1	13	43,5	50,0	3	5
ROUS	25,6	20,0	1	3	34,4	7,1	10	3
SSLP	31,1	0,0	0	7	46,7	0,0	1	7
VTRL	27,2	35,7	6	3	45,7	25,0	8	4

** L'algorithme de régression par arbres peut prendre en compte un tel paramètre. Il est utilisé pour donner du poids aux erreurs de classification. Habituellement, des pénalités plus fortes sont données quand les différences entre les niveaux observés et prévus de la variable expliquée sont grandes.

manquées (AM) sont les jours où l'on prévoit un niveau 0 mais un niveau 1 est observé. Enfin, le Threat Score est calculé de la manière suivante :

$$\frac{100 * N_1}{FA + AM + N_1}$$

où N_1 est le nombre de bonnes prévisions du niveau 1. Ce pourcentage caractérise la qualité de la prévision des niveaux 1.

Dans le tableau 3, on constate que le modèle de régression par arbres donne de moins bons résultats que le modèle bootstrap. Le taux de fausse classification ainsi que le Threat Score et le nombre de fausses alertes l'illustrent. Les valeurs du Threat Score sont très sensibles aux prévisions du niveau 1 car ces observations sont rares (voir Tableau 2, p. 420, dernière colonne).

La prévision des maximums d'ozone par station le matin pour l'après-midi est une prévision très fine et

très complexe, ce qui explique le risque d'erreur important observé dans ces résultats.

Par la suite, on verra que les prévisions obtenues à l'aide des stratégies sont bien meilleures.

Utilisation de stratégies

Les résultats de l'application des deux stratégies proposées sont présentés sous la forme de tableaux de croisement à deux dimensions. Les éléments de la diagonale représentent le nombre de bonnes prévisions en ozone pour les deux niveaux (celui inférieur à 180 µg/m³ en haut à gauche et celui supérieur en bas à droite). L'élément sous la diagonale donne le nombre d'alertes manquées, et celui au-dessus, le nombre de fausses alertes.

La somme des colonnes donne le nombre total de prévisions et la somme des lignes, le nombre total d'observations dans chaque niveau.

Tableau 4.
Résultats de la première stratégie.
Results of the first strategy.

Méthode	Zone 1	Zone 2	Zone 3
Agrégation d'arbres de régression	65	49	65
	2	1	1
	TFC = 5/71	TFC = 7/61	TFC = 4/72
Arbres de régression	68	55	66
	3	5	2
	TFC = 3/71	TFC = 10/66	TFC = 4/72

La première stratégie est appliquée par zone ; les résultats sont donc donnés sur les trois zones définies précédemment (voir p. 421-422, Préviation par station et préviation par zone d'information – Stratégies de préviation). Le tableau 4 présente ces résultats pour les deux modèles.

En ce qui concerne la seconde stratégie (globale au département), les résultats pour les deux modèles sont montrés dans le tableau 5.

Tableau 5.
Résultats de la seconde stratégie.
Results of the second strategy.

Agrégation par arbres de régression	Arbres de régression
68	73
4	1
TFC = 10/80	TFC = 5/80

Pour ces stratégies, la qualité des prévisions dépend des zones. En effet, certaines connaissent des phénomènes météorologiques plus complexes que d'autres, et donc plus difficiles à comprendre et à prévoir. C'est le cas de la zone ouest de l'étang de Berre qui, de ce fait, enregistre un pourcentage d'alertes manquées plus important que les autres.

Sinon, les résultats sont assez satisfaisants dans l'ensemble, particulièrement pour l'agrégation par « bootstrap » avec la première stratégie et pour la régression avec la deuxième (en effet, à alertes manquées égales, il y a moins de fausses alertes).

Utilisation de la météorologie observée

La nécessité d'améliorer le modèle a permis de comprendre certaines lacunes des données et donc des modèles de préviation actuels.

Une seule station de données météorologiques a servi pour la préviation : celle de Marignane qui a été considérée comme représentative des 16 stations sur lesquelles les prévisions sont faites. Cette hypo-

thèse n'est peut être pas vraie en raison des variations importantes de la « météo » locale. Des données et des prévisions sur plusieurs autres sites météorologiques n'étaient pas encore disponibles au moment où cette étude a été réalisée. Leur intégration est prévue pour l'été 2000.

Nous avons essayé de quantifier l'importance de la préviation des paramètres météorologiques pour la préviation de l'ozone. Dans les données de l'été 1998 (période de test), les prévisions « météo » ont été remplacées par les observations. L'amélioration de la préviation de l'ozone résultant de cette opération est montrée dans les tableaux 6 et 7 ci-contre.

Malgré l'amélioration globale, ces résultats semblent confirmer l'idée de la mauvaise adéquation de la seule station de Marignane à ce problème concernant la totalité de l'aire métropolitaine marseillaise.

D'autres causes liées directement au modèle sont aussi certainement responsables des imperfections de la préviation : les paramètres ne sont peut-être pas assez représentatifs des phénomènes photochimiques ; peut-être faudrait-il enrichir la base de données de nouveaux éléments. Une analyse clinique d'une préviation quotidienne nous permettra d'en juger.

Phase opérationnelle de l'été 1999

Pour l'été 1999, le nombre de dépassements du niveau d'information du public, traité par la préviation, est de 9 pour la zone nord-est, 6 pour la zone sud-est et 3 pour la zone ouest.

Les résultats des prévisions obtenus par les modèles suite à l'application des stratégies décrites ci-dessus sont donnés dans le tableau 8 ci-contre.

Ces résultats obtenus pendant la phase opérationnelle ont été expertisés par le prévionniste d'AIMARAIX et leur modifications après expertise est résumée dans le tableau 9.

Notons d'abord que les résultats de la phase opérationnelle de l'utilisation des modèles sont globalement meilleurs que ceux obtenus précédemment dans la phase test (été 1998).

Tableau 6.
Résultats de la première stratégie.
Results of the first strategy.

Méthode	Zone 1		Zone 2		Zone 3	
Arbres de régression agrégés	66	2	46	9	66	2
	3	0	1	5	1	3
	TFC = 5/71		TFC = 10/61		TFC = 3/72	
Arbres de régression	68	0	47	8	66	2
	2	1	3	3	1	3
	TFC = 2/71		TFC = 11/61		TFC = 3/72	

Tableau 7.
Résultats de la seconde stratégie.
Results of the second strategy.

Arbres de régression agrégés		Arbres de régression	
67	7	74	0
1	5	4	2
TFC = 8/80		TFC = 4/80	

PRÉVISION BRUTE DES MODÈLES CART ET BAGGING

Tableau 8.
Résultats de la prévision pour l'été 1999.
Results of the forecast for the 1999 summer.

Méthode	Zone 1 Nord-Est		Zone 2 Sud-Est		Zone 3 Ouest	
Arbres de régression	71	10	78	6	85	2
	1	8	3	3	2	1
	TFC = 11/90		TFC = 10/90		TFC = 4/90	

PRÉVISION AVEC EXPERTISE

Tableau 9.
Résultats de la prévision pour l'été 1999 avec expertise humaine.
Results of the forecast for the 1999 summer with the analysis of an expert.

Méthode	Zone 1 Nord-Est		Zone 2 Sud-Est		Zone 3 Ouest	
Arbres de régression	74	7	78	6	84	3
	0	9	1	5	2	1
	TFC = 7/90		TFC = 7/90		TFC = 5/90	

L'expertise humaine a permis de réduire le nombre de fausses alertes et de récupérer de bonnes prévisions, principalement pour les zones nord-est et sud-est.

Une bonne partie des erreurs de prévision a pu être attribuée à des erreurs de prévision de paramètres météorologiques, principalement la direction du vent (rappelons que la direction du vent utilisée est celle observée et prévue à Marignane, qui

semble n'être représentative que pour les stations à proximité de ce site). Les deux premières zones (nord-est et sud-est) sont constituées de stations pour lesquelles on dispose de beaucoup d'observations, alors que la zone ouest est constituée de stations récentes pour lesquelles l'historique des observations est assez court. De plus, le nombre de dépassements observés durant l'été 99 pour ces zones est beaucoup plus important que celui disponible dans le passé.

Conclusions et perspectives

Cet article présente les premiers résultats obtenus dans le cadre de l'utilisation de méthodes statistiques pour la prévision de l'ozone. Des modèles basés sur des données de 1992 à 1997 ont été construits et testés sur la période estivale de 1998. La performance de ces méthodes a été analysée en terme d'application de stratégies d'information du public sur le dépassement du seuil de 180 $\mu\text{g}/\text{m}^3$ d'ozone.

Notons d'abord que les résultats obtenus pour l'été 1998 sont assez satisfaisants pour cette première approche statistique. Cette période n'a pu bénéficier d'expertise humaine et les résultats bruts des modèles ont été utilisés. Seules quelques erreurs de prévision de l'ozone ont pu être attribuées à des erreurs de prévision de paramètres météorologiques utilisés dans nos modèles.

Les méthodes décrites ici ont ensuite été utilisées en phase opérationnelle durant l'été 1999, où la prévision a été réalisée au quotidien et revue par un expert du réseau AIRMARAIX. Le bilan de l'été 1999 a été plus que satisfaisant. L'expertise des résultats fournis par les méthodes semble actuellement indispensable.

Plusieurs points ont été soulevés en vue de l'amélioration des résultats de la prévision de l'ozone :

- l'intégration aux modèles de 9 autres stations de Météo-France dans les Bouches-du-Rhône ;

- l'intégration aussi de données de radiosondage fournies par Météo-France (observations et prévisions), qui sont en cours de traitement.

Enfin, les résultats présentés ici concernent la prévision à échéance J, le matin pour l'après-midi. L'été 2000 servira de phase de test des mêmes méthodes dont la mise en œuvre vient de commencer pour la prévision à échéance J+1, l'après-midi pour le lendemain.

Les données de la qualité de l'air ont été fournies par les réseaux AIRMARAIX et AIRFOBEP, et les données météorologiques par Météo-France. Les programmes utilisés ont été écrits avec le logiciel S+.

Mots clés

Pollution photochimique. Météorologie. Prévision statistique. Maximum d'ozone journalier. Arbres de régression. Stratégies d'information.

Keywords

Photochemical pollution. Meteorology. Statistical forecast. Daily ozone maximum. Regression tree. Strategies of information.

Références

1. Toupance G. L'ozone dans la basse troposphère – Théorie et application. Pollution Atmosphérique, Janvier-Mars 1988 : 32-60.
2. Robeson SM, Steyn DG. Evaluation and comparison of statistical forecast models for daily maximum ozone concentrations. Atmos Environ 1990 ; 246 : 303-12.
3. Shefingier H, Stohl A, Kromp-Kolb H, Spangl W. A statistical method for predicting daily maximum ozone concentration. Gefahrstoffe Reinhaltung der Luft 1996 ; 56 : 133-7.
4. Burrows W, Benjamin M, Beauchamp S, Lord ER, McCollor D, Thomson B. CART Decision-Tree Statistical Analysis and Prediction of Summer Season Maximum Surface Ozone for the Vancouver, Montreal, and Atlantic Regions of Canada. Journal of Applied Meteorology 1994 ; 34 : 1848-62.
5. Ryan WF. Forecasting severe ozone episodes in the Baltimore metropolitan area. Atmos Environ 1995 ; 29 (17) : 2387-99.
6. Bel L, Bellanger L, Bonneau V, Ciuperca G, Dacunha-Castelle D, Deniau C, Ghattas B, Misiti M, Misiti Y, Oppenheim G, Poggi JM, Tomassone R. Éléments de comparaison de prévisions statistiques des pics d'ozone. Revue de Statistique Appliquée 1999 ; XLVII (3) : 7-25.
7. Valfre G, Thieleke R, Léopold A, Mesbah BA. Prévision des pics de pollution par l'ozone dans la région de l'étang de Berre. Pollution Atmosphérique 1999 ; 163 : 58-66.
8. Breiman L, Friedman JH, Olshen R, Stone CJ. Classification And Regression Trees. Wadsworth, Belmont CA, 1984.
9. Ghattas B. Prévisions des pics d'ozone par arbres de régression simples et agrégés par bootstrap. Revue de Statistique Appliquée 1999 ; XLVII (2) : 61-80.
10. Ghattas B. Prévision par arbres de classification. Mathématiques Informatique et Sciences Humaines 1999 ; 146 : 31-49.
11. Breiman L. Bagging Predictors, Machine Learning 1996 ; 24 : 123-40.
12. Deniau C, Ghattas B. Essai de classification des stations de mesure d'ozone du réseau AIRMARAIX-AIRFOBEP. Rapport de contrat de recherche GREQAM-AIRMARAIX 1998.