

Développement d'un modèle statistique de prévision à 24 heures d'un dépassement du seuil d'information de la population pour l'ozone

Development of a statistical prediction model for overstepping ozon rates

François ENGEL (*), Christian VIEL (*), Hervé CHANUT (*)

RÉSUMÉ

L'étude met en relation les occurrences de forte pollution par l'ozone à Strasbourg, diagnostiquées à partir des réseaux de mesures de l'ASPA et de l'UMEG, et des données météorologiques issues des stations au sol ou du modèle numérique ARPEGE de Météo-France. L'analyse discriminante permet de distinguer les cas de dépassement des non-dépassements d'une valeur seuil de concentration en ozone à partir de paramètres météorologiques observés ou prévus. Cette méthode représente une amélioration substantielle par rapport à une prévision par persistance.

ABSTRACT

This study investigates the relationships between events of high levels of ozone pollution in Strasbourg detected by the measuring networks of ASPA and UMEG, and meteorological data derived from surface stations or from the ARPEGE numerical model of Meteo-France. With a discriminant analysis on observed or forecasted meteorological parameters, it is possible to discriminate cases where ozone concentrations exceed a given threshold. This technique performs significantly better than a persistence forecast.

1. Introduction

L'objectif est de développer un modèle statistique permettant de prévoir à une échéance de 24 heures le dépassement du seuil d'information de la population pour l'ozone à Strasbourg. Il ne s'agit pas d'une prévision fine. D'abord parce que l'on se limite dans un premier temps à l'information météorologique, sans considération sur les sources de pollution ou sur la chimie de l'ozone. Ensuite parce qu'il suffit qu'un des capteurs de l'agglomération dépasse le seuil de $180 \mu\text{g}/\text{m}^3$ en valeur horaire pour que la procédure d'information de la population soit lancée.

Le modèle statistique doit intégrer l'expérience des situations passées et fournir au décideur un indice synthétique sur les conditions météorologiques prévues, au sens de la possibilité d'occurrence d'un pic de pollution sur l'agglomération.

2. Les données utilisées

Tout d'abord, il faut reconstituer l'historique de la pollution par l'ozone et des conditions météorologiques associées, autrement dit constituer une base de données rassemblant les mesures de pollution de l'ASPA et de l'UMEG, ainsi que les données météorologiques issues des stations au sol de Météo-France, de l'ASPA, de l'UMEG et du modèle numérique opérationnel ARPEGE de Météo-France.

Le volet météorologique de cette base de données comprend les mesures horaires de température et de vent de quatre stations météorologiques, dont la station de référence de l'aéroport de Strasbourg-Entzheim. Ces observations au sol sont complétées par les champs de vent et de température analysés par le modèle numérique ARPEGE entre le sol et 1 500 mètres à 12 heures sur dix niveaux d'altitude et pour les quatre points de grille les plus proches de Strasbourg. Destinée à l'initialisation du modèle ARPEGE, l'analyse intègre un certain nombre d'observations météorologiques venant complé-

(*) Météo-France (Direction Interrégionale Nord-Est).

ter une ébauche fournie par la dernière prévision du modèle. Cette analyse a permis de calculer la hauteur de la couche de mélange à 12 heures. Des paramètres mal prévus actuellement, le rayonnement solaire ou la nébulosité par exemple, n'ont pas été pris en compte.

Le volet chimique de la base de données rassemble des mesures d'ozone et d'oxydes d'azote exprimées en valeur horaire. Le réseau de capteurs dont sont issues ces mesures s'est progressivement mis en place. Depuis le mois de juillet 1995, il se compose de quatre capteurs ; trois d'entre eux sont localisés en périphérie (à Schiltigheim, Kehl Hafen et Kehl Sud), le quatrième est situé au cœur de la ville de Strasbourg (Place Kleber). Les variations temporelles de l'ozone peuvent se résumer en deux grands cycles : un cycle annuel où les concentrations les plus fortes sont observées en été, et un cycle diurne avec un maximum en début d'après-midi. Les pics de pollution se cantonnent à la période comprise entre avril et septembre (figure 1).

Les variations spatiales de l'ozone sont parfois importantes. Si l'on considère le maximum quotidien d'ozone pour chaque capteur (en moyenne horaire), l'écart-type est de l'ordre de $8 \mu\text{g}/\text{m}^3$ et peut atteindre les 30 à $40 \mu\text{g}/\text{m}^3$ dans les cas de forte pollution. Les mesures effectuées en périphérie sont en moyenne plus élevées qu'au centre-ville, mais lorsque l'on s'intéresse aux journées les plus polluées, ce résultat est souvent mis en défaut.

Compte tenu de l'objectif de départ, la variable à expliquer est la concentration maximale horaire d'ozone mesurée sur l'ensemble de l'agglomération. Pour garantir l'homogénéité des séries de mesures d'ozone et pour tenir compte des modifications intervenues dans le plan de circulation de Strasbourg en 1994, l'étude statistique se limite aux données des étés 1995, 1996

et 1997 (« été » désigne ici une période de six mois s'étalant d'avril à septembre).

3. La modélisation statistique

Certains paramètres météorologiques comme la température ou la vitesse du vent à différentes altitudes sont fortement corrélés avec la concentration maximale en ozone mesurée. La modélisation statistique a pour objet de combiner ces paramètres pour en apprécier les effets conjoints. La méthode statistique retenue est l'analyse discriminante linéaire en deux classes.

3.1. Principe de l'analyse discriminante

Considérons la population constituée par les journées d'un été. Chaque journée est définie par un ensemble de paramètres météorologiques (températures minimale et maximale, vent, gradient vertical de température, hauteur de la couche de mélange, etc.) et par le maximum d'ozone du jour.

On se fixe un seuil S_{ozone} d'ozone qui permet de scinder la population en deux classes :

- la classe A des jours où le maximum d'ozone est supérieur à S_{ozone} (les journées dites « polluées »),
- la classe NA des jours où le maximum d'ozone est inférieur à S_{ozone} (les journées dites « non polluées »).

Une journée de la population est maintenant définie par ses paramètres météorologiques et par la classe (A ou NA) à laquelle elle appartient. L'analyse discriminante consiste à trouver, dans l'espace des prédictors (qui sont ici les paramètres météorologiques), la surface qui sépare au mieux les classes A et NA. Cette surface servira ensuite en prévision, ce sera la règle décisionnelle pour affecter une nouvelle journée

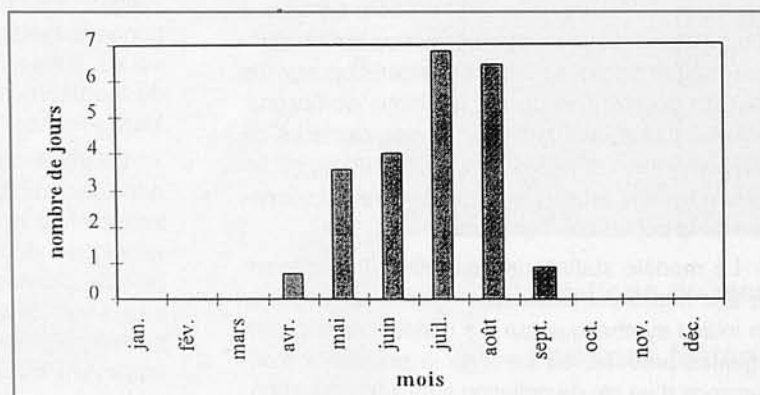


Fig. 1.

Nombre moyen de jours où la concentration d'ozone a dépassé $180 \mu\text{g}/\text{m}^3$ en valeur horaire à Strasbourg (période 1991-1997).

Mean number of days where ozone concentration exceed $180 \mu\text{g}/\text{m}^3$ in Strasbourg (1991-1997).

dans la classe des journées « polluées » (A) ou des journées « non polluées » (NA). En discrimination linéaire, l'équation de la surface discriminante est de la forme $a_1x_1 + \dots + a_nx_n = D$ où les x_i sont les valeurs des prédicteurs météorologiques et les a_i des coefficients optimaux à déterminer. D désigne un niveau de discrimination qui dépend du rapport entre le coût des fausses alertes et celui des non-détections. Ces coûts sont supposés égaux ici.

La figure 2 illustre cette méthode dans le cas simple où l'on utilise deux prédicteurs : la température maximale et la vitesse du vent moyenne. Les données sont celles de l'été 1995, le seuil d'ozone est fixé à $160 \mu\text{g}/\text{m}^3$. Chaque jour est pointé dans l'espace des prédicteurs (un plan ici) en distinguant la classe à laquelle appartient le jour (A ou NA). Dans cet exemple, la courbe discriminante est une droite.

3.2. Construction des modèles et résultats

La construction d'un modèle commence par la définition du prédicand ou variable à prévoir, puis vient le choix des prédicteurs, et enfin le calcul des coefficients de l'équation discriminante. Ces coefficients sont calculés à partir d'un fichier d'apprentissage, la qualité de la discrimination est ensuite évaluée sur un autre fichier appelé fichier test. Les données de 1995 et 1996 ont servi à l'apprentissage, celles de 1997 au test.

Le prédicand est défini de la manière suivante : il y a dépassement du seuil S_{ozone} lorsqu'un des capteurs de l'agglomération au moins enregistre une concentration moyenne horaire en ozone supérieure ou égale à S_{ozone} . Plusieurs seuils S_{ozone} sont testés : 160, 170 et $180 \mu\text{g}/\text{m}^3$.

Ceci permet de séparer les journées en deux classes : « A » lorsque le seuil S_{ozone} est atteint ou dépassé, « NA » sinon.

Il faut remarquer que compte tenu du déséquilibre entre les cas A et NA (quel que soit le seuil, le nombre de cas NA est très supérieur au nombre de cas A), il est préférable de faire un tri dans le fichier d'apprentissage initial. En ajoutant comme condition une température maximale supérieure à 25°C , on élimine des journées peu polluées, faiblement informatives, tout en conservant les journées où la concentration en ozone a dépassé $180 \mu\text{g}/\text{m}^3$ (seuil d'information de la population) pour au moins un capteur. Le « nouveau » fichier d'apprentissage contient 101 jours. Moins dissymétrique que le fichier initial, il est de taille suffisante pour caler une équation discriminante.

Parmi l'ensemble des prédicteurs, certains ne sont pas discriminants, ou sont redondants. On utilise donc une méthode de sélection des prédicteurs : la sélection progressive ascendante. Son principe est le suivant : le premier prédicteur x_1 est sélectionné, il s'agit du plus discriminant. Le deuxième prédicteur x_2 est celui qui complète le mieux x_1 , le troisième est celui qui complète le mieux x_1 et x_2 , etc. Cette sélection fournit un ordonnancement de l'ensemble des prédicteurs disponibles. Le nombre n optimal de prédicteurs retenus est un compromis entre informativité (ou pouvoir discriminant, qui augmente avec n) et stabilité du modèle (qui diminue quand n augmente). Une méthode objective, le test de Fischer appliqué à un indice de séparabilité des deux classes de la discrimination (la distance de Mahalanobis ici), permet de vérifier à chaque

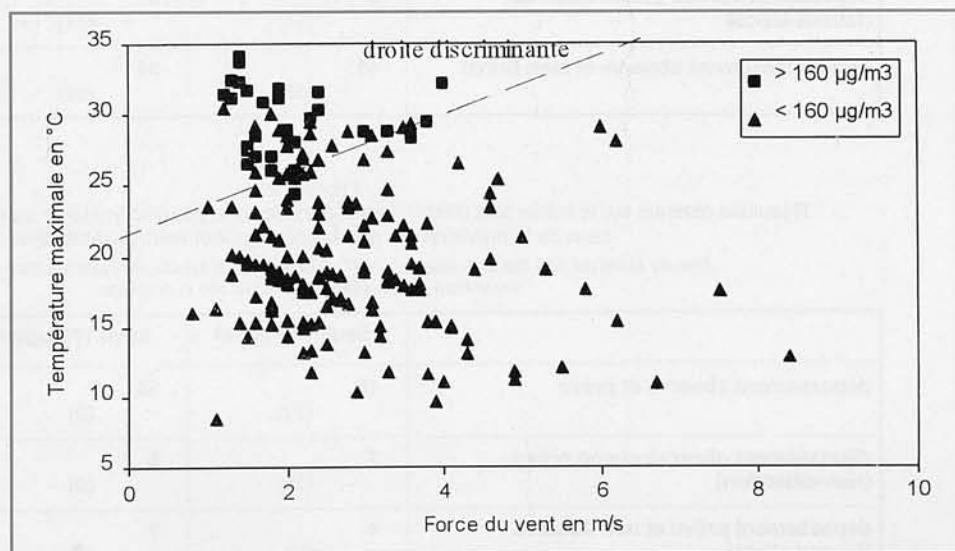


Fig. 2.
Un exemple simple d'analyse discriminante.
A simple example of discriminant analysis.

étape si l'apport d'un nouveau prédicteur est significatif.

Les tableaux 1 et 2 présentent des résultats obtenus à l'aide de modèles à quatre prédicteurs. Les groupes de prédicteurs retenus ne sont pas identiques pour les trois seuils d'ozone : la température maximale quotidienne et la valeur du maximum d'ozone de la veille, toujours sélectionnées, sont complétées par deux prédicteurs parmi les trois suivants : la température minimale quotidienne, la vitesse du vent à 1 500 mètres et la hauteur de la couche de mélange à 12 h. A titre de comparaison figurent entre parenthèses les résultats obtenus par persistance, c'est-à-dire en affectant la journée à la classe des journées polluées lorsque la température est supérieure à 25 °C et la concentration en ozone de la veille supérieure au seuil S_{ozone} . La confrontation des résultats avec ceux obtenus par la prévision par persistance permet de mesurer le gain du modèle statistique par rapport à une prévision « triviale ».

Les résultats présentés ci-dessous ont valeur diagnostique, c'est-à-dire qu'une situation est

décrite à partir d'observations et d'analyses. En mode opérationnel (ou pronostic), les erreurs de prévision sur les paramètres météorologiques d'entrée s'ajoutent aux erreurs de la discrimination.

4. Perspectives

Deux méthodes permettent classiquement de construire les équations d'un modèle de prévision par adaptation statistique : la prévision parfaite et les statistiques de sorties de modèle. En prévision parfaite, les liaisons statistiques sont établies entre le prédictant et les valeurs observées ou analysées des prédicteurs. Les équations ainsi mises au point sont ensuite utilisées avec les valeurs prévues des prédicteurs. Dans les statistiques de sorties de modèle, les équations sont directement calculées à partir des valeurs prévues des prédicteurs. Ils sont ensuite naturellement alimentés par des prévisions. Les deux méthodes ont leurs avantages, elles seront évaluées en pré-opérationnel durant l'été 1998.

Tableau 1

Résultats obtenus sur le fichier d'apprentissage (1995 et 1996). Les scores de l'analyse discriminante sont présentés en gras, ceux de la prévision par persistance figurent entre parenthèses.

Results obtained with the learning data set (1995 and 1996). The scores of the discriminant analysis are in bold, those from the persistence forecast are in brackets.

	Seuil 160 µg/m ³	Seuil 170 µg/m ³	Seuil 180 µg/m ³
dépassement observé et prévu	41 (30)	28 (21)	19 (12)
dépassement observé et non prévu (non-détection)	10 (21)	10 (17)	8 (15)
dépassement prévu et non observé (fausse alerte)	9 (14)	9 (11)	5 (11)
non-dépassement observé et bien prévu	41 (36)	54 (52)	69 (63)

Tableau 2.

Résultats obtenus sur le fichier test (1997). Les scores de l'analyse discriminante sont présentés en gras, ceux de la prévision par persistance figurent entre parenthèses.

Results obtained with the test data set (1997). The scores of the discriminant analysis are in bold, those from the persistence forecast are in brackets.

	Seuil 160 µg/m ³	Seuil 170 µg/m ³	Seuil 180 µg/m ³
dépassement observé et prévu	18 (13)	13 (9)	10 (8)
dépassement observé et non prévu (non-détection)	7 (12)	5 (9)	5 (7)
dépassement prévu et non observé (fausse alerte)	4 (7)	7 (7)	3 (5)
non-dépassement observé et bien prévu	31 (28)	35 (35)	42 (40)

Les scores sont encourageants et montrent que l'on peut tirer des conditions météorologiques locales une information utile à la prévision des pics de pollution par l'ozone. La liste des prédictors peut être élargie. La situation météorologique synoptique en particulier est un facteur important dans la description des épisodes de forte pollution, par la prise en compte indirecte des phénomènes de transport à distance. L'analyse discriminante sera combinée à une classification de notre échantillon de situations par type de temps. L'idée est de calculer une équation discriminante par type de temps, après avoir identifié le champ météorologique (géopotential, température...) donnant la classification la plus informative vis-à-vis de la pollution par l'ozone. Ces développements sont en cours de réalisation.

Cependant, il faut avoir à l'esprit que la discrimination entre les cas de dépassement et de non-dépassement ne sera jamais parfaite. L'outil statistique ne peut constituer qu'une aide à la décision. L'analyse discriminante permet de quantifier l'incertitude et de modifier le rapport entre les probabilités de fausse alerte et de non-détection. Dans le cas présent, on peut par exemple bâtir un modèle statistique de prévision qui a une très faible probabilité de « rater » un pic de pollution, mais qui en contrepartie donne un certain nombre de fausses alertes. L'utilisateur connaît ainsi les limites du modèle statistique.

Bibliographie

- DER MEGREDITCHIAN G. Méthodes statistiques de prévision par classes en météorologie. Application au verglas de l'analyse discriminante. *La Météorologie* 5^e série, n° 26, 1973.
- DER MEGREDITCHIAN G. Quelques aspects de la sélection des prédictors. La prévision statistique des phénomènes météorologiques. *La Météorologie* 6^e série, n° 19, 1979.
- DER MEGREDITCHIAN G. La prévision statistique des phénomènes météorologiques. *Météorologie Nationale*, note technique de l'EERM, n° 100, 1981.
- DER MEGREDITCHIAN G. Eléments de l'analyse discriminante - Théorie et pratique. *Météorologie Nationale*, note de l'ENM, 1988.
- DER MEGREDITCHIAN G. Le traitement statistique des données multidimensionnelles, application à la météorologie. Tome 2. *Météo-France*, Cours et manuels, n° 9, 1993.
- LOSSEC B. Situation météorologique et pollution photo-oxydante sur la région parisienne : l'épisode remarquable du 29 juillet au 1^{er} août 1992. *La Météorologie* 8^e série, n° 8, 1994.
- MENIN V. *Prévision d'un indice de pollution en région parisienne relatif aux conditions météorologiques*. DEA Statistiques, Université Paris VI, 1991.
- POTTIER P. Prévision de la température par adaptation statistique du modèle Périodot. *La Météorologie* 7^e série, n° 32, 1990.