Prévisions de concentrations d'ozone

Comparaison de différentes méthodes statistiques de type « boîte noire »

Alain Rakotomamonjy* — **Komi Gasso*** — **Stéphane Canu***
Patrick Vannoorenberghe**

Laboratoire PSI, FRE 2645 CNRS
* *INSA Rouen*
11, Avenue de l'Université
F-76801 Saint-Etienne du Rouvray Cedex
{arakotom, kgasso, scanu}@insa-rouen.fr

** *Université de Rouen*
UFR Sciences et Techniques
F-76821 Mont-Saint-Aignan Cedex
patrick.vannoorenberghe@univ-rouen.fr

RÉSUMÉ. *Dans cet article, on s'intéresse à la prévision, à échéance de 24 h, des concentrations maximales journalières d'ozone en utilisant des approches de type boîte noire. Ces approches présentent l'intérêt de ne nécessiter que des données mesurables par les AASQA (Associations agréées de surveillance de la qualité de l'air) et d'être transposables d'un site à un autre, ce qui leur confère un caractère générique. L'article réalise une étude comparative de l'application de quatres méthodes d'apprentissage statistique à ce problème à savoir les arbres de décision, les réseaux de neurones, la régression parcimonieuse et les SVR (Support vector regression). Les résultats obtenus ainsi que leurs commentaires sont présentés. Des conclusions et perspectives finissent l'article.*

ABSTRACT. *The paper investigates the application of black box modelling to the prediction of the daily maxima of ground-ozone level. The main interest of these modelling approaches is their genericity as they are solely based on the available data provided by the Associations of air quality monitoring and they can be transposed from a geographical area to another one. The paper realises a comparative study of four statistical learning approaches, the decisions trees (CART), the neural networks, the least-angle regression (LAR) and the support vector regression (SVR), to the ozone level prediction. The obtained results and their comments are presented. The paper ends with some conclusions.*

RS - JESA - 39/2005. Information et pollution atmosphérique, pages 533 à 552

MOTS-CLÉS : *maxima journaliers, ozone, prévision, approches statistiques, modèles boîte noire.*
KEYWORDS: *Daily maximum, ozone, prediction, statistical approaches, black box modelling.*

1. Introduction

Durant la période estivale, la plupart des grandes ou moyennes métropoles connaissent des épisodes de pollution par l'ozone. Dans les basses couches de l'atmosphère, l'ozone est un polluant secondaire issu de réactions chimiques complexes, influencées par les conditions météorologiques, entre les polluants primaires (les oxydes d'azote, les composants organiques volatiles...) et l'oxygène de l'air. Ce polluant a des effets nocifs pour l'homme et l'environnement quand sa concentration atteint des valeurs excessives. Pour ces raisons, des directives européennes reprises en France dans la loi sur la qualité de l'air ont défini les seuils suivants :

– $110\,\mu g/m^3$ (valeur moyenne sur 8 heures), seuil de protection pour la santé,

– $180\,\mu g/m^3$ (valeur moyenne sur 1 heure), seuil d'information de la population,

– $360\,\mu g/m^3$ (valeur moyenne sur 1 heure), seuil d'alerte de la population.

Ces directives ont incité les autorités régionales à se doter de réseaux de surveillance de la qualité de l'air (les AASQA pour Associations agréées pour la surveillance de la qualité de l'air). Les réseaux ont pour missions principales, la mesure, l'analyse des principaux indicateurs de pollution et l'information du public et des autorités préfectorales en cas de dépassement des seuils fixés, ces autorités pouvant déclencher des procédures de réduction des polluants primaires. Il apparaît donc nécessaire que les AASQA se dotent d'outils de prévision des valeurs de la concentration d'ozone.

Différents modèles de prévision ont été développés et se rangent dans deux familles :

– les modèles déterministes ou physiques (Chenevez *et al.*, 2001) (modèle Chimère (Vautard *et al.*, 2001), le système Prev'air (Honoré *et al.*, 2004)...) qui exploitent numériquement la modélisation des réactions physico-chimiques complexes (chimie atmosphérique, transport de masse d'air, conditions météorologiques) gouvernant le phénomène de formation et de destruction d'ozone (Académie des Sciences, 1993, EPA Environmental Protection Agency, 1999). Ils permettent de fournir des estimations des concentrations de polluants à une grande échelle (régionale ou continentale (Honoré *et al.*, 2004)). Mais ces modèles sont généralement complexes et nécessitent des données d'entrée souvent non disponibles dans les AASQA. Par ailleurs, leur calibrage et la validation de leurs résultats à une échelle très locale ne sont pas souvent aisés.

– les modèles de type « boîte noire » : ces modèles ont l'avantage de s'appuyer sur les données mesurables par les AASQA. Cette forme de modélisation est basée sur des approches statistiques généralement bien établies dans la littérature (Vapnik, 1998). L'autre intérêt de ces modèles est leur caractère générique en ce sens qu'ils sont "facilement" transposables d'un site à un autre moyennant quelques réglages et leur mise à jour à partir de nouvelles données est peu coûteuse en temps. Ces raisons font de l'approche « boîte noire », l'approche de modélisation la plus communément utilisée.

Dans cette approche, on distingue les modèles de prévision basés sur des techniques de régression linéaire (Comrie, 1997, Gasso *et al.*, 2000) ou non linéaire (Cobourn *et al.*, 2000, Rakotomamonjy *et al.*, 2001, Mourot *et al.*, 2003, Zolghadri, 2003) ou encore des modèles relevant des techniques de reconnaissance de forme (Ambroise *et al.*, 2000, Canu *et al.*, 2001, Vannoorenberghe *et al.*, 2001). Tous ces modèles (voir (APPetisse, 2001) pour une synthèse de méthodes de prévision de concentration d'ozone) ont donné des résultats encourageants voire satisfaisants et il est difficile d'exhiber la supériorité d'une méthode particulière.

L'article est consacré à l'étude comparative de l'application de quelques méthodes d'apprentissage statistique à la prévision des concentrations maximales journalières d'ozone. Afin de permettre aux AASQA d'informer les autorités préfectorales en vue d'éventuelles mesures de restriction, il est nécessaire de fournir une prévision du niveau d'ozone à échéance de 24 h (prévision au jour J pour le jour $J + 1$). Vu la rareté et l'hétérogénéité des épisodes de pollution aigüe dans les données traitées, le travail a été axé plutôt sur la prédiction de la valeur du taux maximal journalier d'ozone que sur la prévision de dépassement des seuils fixés par la législation. Les techniques étudiées dans l'article sont des approches classiques pour cette application (APPetisse, 2001) :

– la régression linéaire simple dont les performances serviront de référence,

– les arbres de décision CART (Breiman *et al.*, 1984), basés sur des méthodes d'induction et qui décomposent un problème de classification ou de régression en une suite de tests imbriqués portant sur des variables explicatives,

– les réseaux de neurones artificiels (RNA), en particulier le perceptron multicouche (Haykin, 1999) qui est l'une des techniques de modélisation les plus connues,

et les outils récents de *« machine learning »* :

– la régression parcimonieuse (Efron *et al.*, 2003) : c'est une technique d'estimation fonctionnelle basée sur la méthode des noyaux et qui repose sur l'optimisation d'un critère quadratique (risque empirique) avec une pénalisation ℓ_1 (régularisation dite lasso),

– les SVR (*Support Vector Regression*) : c'est une extension des SVM (*Support Vector Machines*) à la régression. Méthode émergente de ces dernières années, les SVM sont un algorithme de classification et de régression basés sur une théorie d'apprentissage statistique solide (Vapnik, 1998). Dans le contexte de la régression, l'objectif est de réaliser une estimation fonctionnelle la plus régulière possible et qui présente au plus une déviation maximale (fixée) par rapport à la variable à expliquer. Inversement à la régression parcimonieuse, les SVR reposent sur l'optimisation d'un critère ℓ_1 (erreur absolue) avec une pénalisation quadratique (régularisation ridge).

L'article est structuré de la manière suivante : la section 2 pose la problématique de l'application traitée ; elle présente également les données traitées et la méthodologie suivie. La section 3 décrit les fondements théoriques et les problématiques d'apprentissage des méthodes statistiques. La partie suivante est consacrée à la présentation des résultats obtenus. Les commentaires de ces résultats amèneront dans la dernière partie de l'article à formuler des conclusions générales et des perspectives sur l'utilisation des modèles boîte noire dans les AASQA pour la surveillance de la qualité de l'air.

2. Présentation du problème

2.1. *Données*

L'étude comparative a été menée sur des données issues de la station de mesure périurbaine de Brabois à Nancy (région lorraine). Cette station est gérée par AIRLOR (Association lorraine pour la surveillance de la qualité de l'air). C'est l'un des sites ayant enregistré les concentrations les plus fortes en ozone de l'agglomération nancéienne. De plus, il dispose de mesures de paramètres météorologiques pouvant servir à l'étude du phénomène. Les données portent sur 5 années (1995 à 1999) et comportent des mesures de polluants (ozone O_3, monoxyde d'azote NO, dioxyde d'azote NO_2) et de paramètres météorologiques (température, humidité relative, vitesse du vent, direction du vent, pression et rayonnement). Ces données sont relevées au pas de temps de 15 min ; elles sont ensuite transformées en données horaires par moyennage des mesures quart-horaires. Les épisodes de pollution par l'ozone se produisent essentiellement durant la période allant du printemps à la fin de l'été. Nous avons considéré par conséquent la plage du 1er avril au 30 septembre de chaque année pour l'étude, soit environ 180 jours par an.

2.2. *Variables utilisées*

Variable à expliquer y **-** Dans cette étude, nous avons opté pour la prédiction de la concentration maximale journalière d'ozone à échéance de 24 h (jour J pour le jour $J + 1$). Bien que les directives législatives s'intéressent à un dépassement de seuil et donc plûtot à une catégorisation de la prévision (dépassement d'un seuil ou pas), il nous semble plus judicieux de fournir la prédiction de la concentration en ozone car c'est une démarche qui permet plus de flexibilité dans la décision. Par exemple, une information concernant l'intervalle de confiance de la prédiction peut être fournie et cet intervalle peut servir aux AASQA pour décider un dépassement d'un seuil fixé ou non.

Pour prédire la concentration maximale en ozone, il est nécessaire d'extraire pour chaque journée le maximum de la concentration d'ozone en période diurne. Une étude préalable des données a montré que les maxima d'ozone sont enregistrés majoritairement durant la plage horaire entre 12 h et 17 h TU. Les AASQA ayant la contrainte de fournir l'estimation de la concentration d'ozone au plus tard à 15 h TU, le niveau maximal d'ozone est pris dans la plage 12 h - 15 h TU. La figure 1 montre les maxima d'ozone sur les 5 années. Le tracé des boîtes à moustache (Saporta, 1990) de la concentration maximale journalière d'ozone pour les différentes années (figure 2) montre que la valeur médiane de la concentration d'ozone est relativement faible. Cependant, dans un objectif de protection de la santé, le point qui nous intéresse est essentiellement la prédiction des fortes concentrations d'ozone (au-delà de 130 $\mu g/m^3$) ce qui au vu des données s'apparente à la prédiction d'événements rares. On note également que les très fortes concentrations en ozone, présentes surtout pour l'année 1998, sont des événements exceptionnels donc difficilement prévisibles.

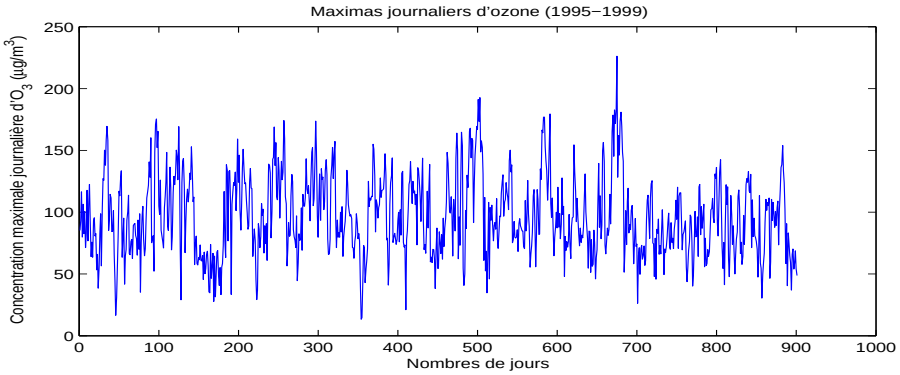


Figure 1. *Evolution des concentrations maximales journalières d'ozone (1995-1999)*

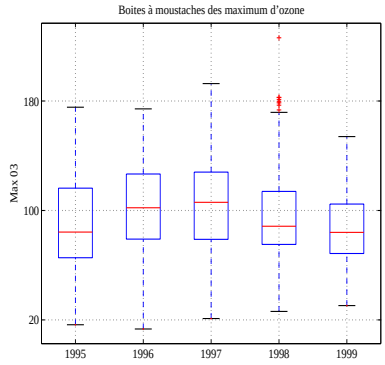


Figure 2. *Boîtes à moustache des maxima journaliers d'ozone année par année*

Variable explicatives - L'étude des travaux disponibles dans la littérature (EPA Environmental Protection Agency, 1999, APPetisse, 2001) et l'analyse phénoménologique révèlent que le phénomène de pollution par l'ozone est fortement influencé par les niveaux des polluants primaires (NOx, COV) mais aussi par les paramètres météorologiques comme le rayonnement (nécessaire au processus de production d'ozone), la température, l'humidité relative ou la vitesse du vent (qui rend compte de la dispersion des polluants). Pour réaliser une prévision à échéance de 24 h, de façon similaire à la synthèse de y, on génère à partir des données brutes, des variables explicatives pertinentes résumant les conditions météorologiques et de pollution sur chaque journée et sur les journées précédentes afin de tenir compte de la dynamique d'évolution du phénomène sur plusieurs journées.

L'analyse de l'autocorrélation des maxima d'ozone (non présentée dans l'article) montre une diminution lente de la fonction d'autocorrélation (par exemple la corrélation entre $O_3(J+1)$ et $O_3(J)$ vaut 0,93). Ceci nous a conduit à élaborer un modèle de persistance linéaire dont la meilleure structure est d'ordre 2. Bien évidemment, ce modèle simple est peu performant mais a le mérite de suggérer pour les conditions de pollution, les niveaux d'ozone des jours J et $J-1$.

En ce qui concerne les conditions météorologiques, les variables candidates les plus couramment utilisées sont la température et/ou l'humidité relative, la vitesse et la direction du vent et enfin le rayonnement du jour $J+1$, de la journée courante J et des journées précédentes (pour prendre en compte l'évolution des conditions météorologiques jusqu'au jour $J+1$). Ce choix initial des variables explicatives a été corroboré par une analyse bivariée entre elles et y. Compte tenu de ces constats et des contraintes d'exploitation des AASQA, nous avons généré 29 variables explicatives telles que décrites dans le tableau 1.

A la lecture de ce tableau, on constate que pour prédire la concentration d'ozone au jour $J+1$, nous utiliserons des variables explicatives représentatives des conditions météorologiques sur le jour $J+1$ (x_3, x_6 et x_9 par exemple). La prédiction étant réalisée au jour J, les mesures de ces variables ne sont pas disponibles, ce qui impose d'utiliser les prévisions de Météo-France. Ces prévisions n'étant pas disponibles pour les 5 années d'étude, nous avons directement utilisé les mesures du réseau AIRLOR. En situation d'exploitation des modèles, ces mesures devront être remplacées par les prévisions de Météo-France.

2.3. *Méthodologie*

Les variables explicatives du tableau 1 sont les entrées candidates des différents modèles statistiques. La dimension de l'espace des variables explicatives étant relativement importante ($d = 29$), une sélection de variables est réalisée afin d'identifier des modèles parcimonieux ayant de bonnes capacités de généralisation (Hastie *et al.*, 2001).

Pour évaluer la robustesse des différentes méthodes, nous avons comparé leurs performances en utilisant une méthode de validation par année c'est-à-dire que sur les 5 années, les données de 4 années servent à l'apprentissage (sélection de variables d'entrée, détermination de la structure et estimation des paramètres) des différents modèles alors que la dernière année est utilisée comme année de test. Cette approche permet d'avoir une évaluation robuste des méthodes mais a le désavantage de ne pas respecter la chronologie de la série temporelle. Précisons que sur les données des 4 années servant à l'apprentissage, chaque méthode divise le jeu de données en deux parties : un jeu de données pour l'estimation des paramètres et un jeu de validation croisée pour l'identification des paramètres structuraux (par exemple variables d'entrées, hyper paramètres des méthodes à noyaux, nombre de couches cachées du réseaux de neurones).

Variables	Notation
Valeur maximale d'ozone du jour $J+1$ entre 12h et 15h	y
Valeur maximale d'ozone du jour J entre 12h et 15h	x_1
Valeur maximale d'ozone du jour $J-1$ entre 12h et 17h	x_2
Valeur maximale de la tempérare de $J+1$ et J entre 12h et 15h	x_3, x_4
Valeur maximale de la tempérare de $J-1$ entre 12h et 17h	x_5
Valeur minimale de la tempérare entre 0h et 5h de $J+1$, J et $J-1$	x_6, x_7, x_8
Amplide thermique $T_{max} - T_{min}$ de $J+1$, J et $J-1$	x_9, x_{10}, x_{11}
Rayonnement cumulé sur $J+1$, J et $J-1$	x_{12}, x_{13}, x_{14}
Valeur maximale du rayonnement de $J+1$ et J entre 12h et 15h	x_{15}, x_{16}
Valeur maximale du rayonnement de $J-1$ entre 12h et 17h	x_{17}
Valeur maximale de la vitesse du vent de $J+1$ et J entre 12h et 15h	x_{18}, x_{19}
Valeur maximale de la vitesse du vent de $J-1$ entre 12h et 17h	x_{20}
Valeur moyenne de la vitesse du vent de $J+1$ et J entre 12h et 15h	x_{21}, x_{22}
Valeur moyenne de la vitesse du vent de $J-1$ entre 12h et 17h	x_{23}
Valeur maximale de l'humidité relative de $J+1$ et J entre 10h et 15h	x_{24}, x_{25}
Valeur maximale de l'humidité relative de $J-1$ entre 10h et 17h	x_{26}
Valeur moyenne de l'humidité relative de $J+1$ et J entre 10h et 15h	x_{27}, x_{28}
Valeur moyenne de l'humidité relative de $J-1$ entre 10h et 17h	x_{29}

Tableau 1. *Variable à expliquer et variables explicatives (les heures sont données en temps universel)*

2.4. *Critères de performances des méthodes*

L'évaluation des performances des différents modèles est faite sur la base de critères calculés sur les données de test. Les critères choisis sont les erreurs quadratiques moyennes suivantes :

$$J_1 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad \text{et} \quad J_2 = \frac{1}{|i : y_i \geq 130|}\sum_{i:y_i \geq 130}^{n}(y_i - \hat{y}_i)^2$$

où $\hat{y}_i$ représente l'estimation du taux d'ozone et n le nombre d'échantillons du jeu de test. J_1 permet d'avoir une mesure globale de la qualité de généralisation des modèles alors que J_2 quantifie les performances pour les fortes valeurs d'ozone. Le seuil de 130 $\mu g/m^3$ a été choisi compte tenu de l'analyse faite sur la distribution des maximas d'ozone (voir section 2.2). La technique SVR étant basée sur l'optimisation d'un cri-

tère ℓ_1, deux autres critères de performances basés sur les erreurs absolues moyennes sont considérés :

$$J_3 = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad \text{et} \quad J_4 = \frac{1}{|i : y_i \geq 130|}\sum_{i:y_i \geq 130}^{n}|y_i - \hat{y}_i|$$

Les autorités publiques ayant le devoir de prévenir la population lors de dépassements de seuils, nous nous proposons de relever pour différentes valeurs seuils (respectivement $120\,\mu g/m^3$, $130\,\mu g/m^3$, $140\,\mu g/m^3$, $150\,\mu g/m^3$, $160\,\mu g/m^3$, $170\,\mu g/m^3$ et $180\,\mu g/m^3$), les performances des différentes méthodes quant aux dépassements de ces seuils sous la forme de matrice de confusion pour chaque seuil et chaque algorithme ainsi que le rapport :

$$\frac{bd}{bd + fa + md}$$

bd, fa et md sont respectivement le nombre de détection correcte de dépassements, de fausses alarmes et de manques à la détection. Pour une technique prédisant correctement tous les dépassements de seuil sans fausse alarme, ce rapport vaut 1.

3. Algorithmes et méthodes

Notre objectif étant de proposer un modèle non-paramétrique de la concentration d'ozone en fonction de différentes variables météorologiques et de l'historique de la concentration en ozone, nous allons nous intéresser plus particulièrement aux méthodes d'estimations fonctionnelles utilisées en apprentissage statistique.

3.1. *Estimations fonctionnelles et modèles boîtes noires*

Le but de ce paragraphe est d'introduire différentes méthodes de modélisation. Le cadre théorique est le suivant : nous cherchons à estimer la relation de dépendance entre les entrées $x \in \mathbb{R}^d$ d'un système et la sortie y de ce même système à partir d'un ensemble de données expérimentales mesurées $\{x_i, y_i\}_{i=1,\cdots,n}$. Ces données sont généralement considérées comme provenant de l'échantillonnage de $\mathcal{X}$ et $\mathcal{Y}$ suivant une distribution $P(X, Y)$. Ainsi, notre objectif serait de trouver une fonction f d'un espace d'hypothèses $\mathcal{H}$ minimisant le risque :

$$R[f] = \int_{\mathcal{X}\times\mathcal{Y}} L(y, f(x))dP(X, Y)$$

où $L(\cdot, \cdot)$ est une fonction de coût pénalisant les différences entre la valeur réelle y et la prédiction $f(x)$ associée à un vecteur x. Comme typiquement $P(X, Y)$ est une loi

inconnue, l'approche classique pour obtenir $f(x)$ est de minimiser un risque régularisé basé sur les échantillons $\{x_i, y_i\}_{i=1,\cdots,n}$:

$$\min_{f\in\mathcal{H}} R_{reg}[f] \quad \text{avec } R_{reg}[f] = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda\Omega(f) \qquad [1]$$

où $\Omega(f)$ est une fonctionnelle mesurant la régularité de f dans l'espace $\mathcal{H}$. Ce cadre permet d'unifier une grande partie des algorithmes d'estimation fonctionnelle utilisés en apprentissage statistique.

3.2. *Les espaces d'hypothèses* $\mathcal{H}$

En fonction des différentes méthodes d'estimation fonctionnelle, la forme et la nature des espaces d'hypothèses auxquels appartient la fonction d'estimation $f(x)$ diffèrent. Tout d'abord, ces espaces peuvent être engendrés par une famille génératrice $\{\phi_i(x)\}_{i=1,\cdots,m}$ ne dépendant que de x. Cette famille peut être définie *a priori* par l'utilisateur ou alors peut être constituée par les fonctions de base d'un espace donné. Par exemple, l'ensemble des fonctions $\{\phi_i(x)\}$ peut être l'ensemble $\{1, x, x^2, \cdots, x^m\}$ des fonctions monomiales ou alors une base d'ondelettes de l'espace $L_2(\mathbb{R}^d)$. Dans ce cas, l'espace d'hypothèses s'écrit :

$$\mathcal{H} = \{f : f(x) = \sum_{i=1}^{m} \alpha_i\phi_i(x), \alpha \in \mathbb{R}^m, \alpha = [\alpha_i]_{i=1\cdots m}\}$$

Pour certaines méthodes d'apprentissage, si l'espace d'hypothèses est un espace de Hilbert à noyaux reproduisants (Scholkopf *et al.*, 2001b), alors la fonction d'estimation peut être décrite en fonction d'un noyau évalué aux points d'apprentissage $\{x_i\}$. En effet pour ces espaces, il existe une fonction symétrique définie positive de deux variables $\phi(\cdot,\cdot)$ telle que :

$$\forall f \in \mathcal{H}, f(x) = \langle f, \phi(x,\cdot)\rangle_{\mathcal{H}}$$

Dans ce cas, d'après le théorème de la représentation généralisé de Scholkopf (Scholkopf *et al.*, 2001a), la solution du problème défini à l'équation (1) s'écrit :

$$\mathcal{H} = \{f : f(x) = \sum_{i=1}^{n} \alpha_i\phi(x, x_i)\}$$

On notera que dans ce cas, la fonction de régression $f(x)$ est la combinaison linéaire de n noyaux centrés sur les données d'apprentissage tandis que si l'espace d'hypothèses est décrit en fonction des élements de sa base alors $f(x)$ peut être écrite comme étant la combinaison linéaire d'un très grand nombre (voire une infinité) de fonctions.

Une dernière manière pour décrire l'espace d' hypothèses $\mathcal{H}$ est d'utiliser des fonctions génératrices de cet espace dépendant des données. Dans ce cas, $\mathcal{H}$ devient :

$$\mathcal{H} = \{f : f(x) = \sum_{i=1}^{m} \alpha_i\phi_i(x, w_i)\}$$

où chaque $\phi_i(x, w_i)$ est donc une fonction de x et d'un paramètre w_i lui même dépendant des données. L'avantage de cette description par rapport à la description par base est la flexibilité introduite par le paramètre w et cette flexibilité permet d'être plus parcimonieux dans le sens où pour un même espace $\mathcal{H}$, moins de fonctions génératrices sont nécessaires dans ce cas que dans le cas de l'utilisation des fonctions de base.

3.3. *Les fonctions de coût*

Pour la construction de modèle de régression, deux types de fonctions de coût sont usuellement utilisés : la fonction de coût quadratique L_2 et la fonction dite ε-insensible L_ε dont les equations sont respectivement :

$$L(y, f(x)) = (y - f(x))^2$$

$$L(y, f(x)) = |y - f(x)|_\varepsilon = \max(0, |y - f(x)| - \varepsilon)$$

3.4. *Les termes de régularisation*

Il existe typiquement trois types de termes de régularisation dans les méthodes de régression. La plus commune est celle utilisant une pénalisation quadratique. Généralement, dans ce cas le terme régularisant s'écrit :

$$\Omega(f) = \sum_i \alpha_i^2 \quad \text{ou} \quad \Omega(f) = \sum_{i,j} \alpha_i \alpha_j \phi(x_i, x_j)$$

en fonction du choix de $\mathcal{H}$. Ces termes régularisants s'apparentent à la norme quadratique $\|f\|^2$ de la fonction $f(x)$ dans $\mathcal{H}$ et ont plutôt tendance à pénaliser de fortes valeurs pour les paramètres α_i. Pour favoriser la parcimonie de la solution $f(x)$, il est également fréquent de choisir un terme régularisant de type ℓ_1 :

$$\Omega(f) = \sum_i |\alpha_i|$$

De par la forme des courbes iso-valeurs de la norme ℓ_1 par rapport à ceux de la norme ℓ_2, le fait d'utiliser ce régularisateur a tendance à mettre à 0 les composantes α_i.

Un troisième type de terme régularisant peut être utilisé dans les méthodes d'apprentissage :

$$\Omega(f) = \text{card}\{\alpha_i \neq 0\} = \|\alpha\|_0$$

Ce terme, nommé par abus de langage norme ℓ_0 s'apparente plutôt à un terme de mesure de la parcimonie. Ce terme est implicitement utilisé dans les méthodes itératives d'approximation comme le *matching pursuit* (Mallat *et al.*, 1993) et permet de limiter le nombre de composantes dans le modèle.

Modèles	$\mathcal{H}$	$L(y, f(x))$	$\Omega(f)$
Linéaire	$\phi_i(x) = I$	L_2	0
Regression rigde	$\phi_i(x)$	L_2	$\sum_i \alpha_i^2$
Regression rigde parcimonieuse	$\phi_i(x)$	L_2	$\sum_i \|\alpha_i\|$
Réseaux de neurones	$\phi_i(x, w_i)$	L_2	$\sum_i \alpha_i^2 + \sum_i \|w_i\|^2$
Regression splines	$\phi(x, x_i)$	L_2	$\sum_{i,j} \alpha_i \alpha_j \phi(x_i, x_j)$
Splines parcimonieuse (LAR)	$\phi(x, x_i)$	L_2	$\sum_i \|\alpha_i\|$
SV en régression (SVR)	$\phi(x, x_i)$	L_ε	$\sum_{i,j} \alpha_i \alpha_j \phi(x_i, x_j)$
SVR parcimonieux	$\phi(x, x_i)$	L_ε	$\sum_i \|\alpha_i\|$
CART	$\phi_i(x, w_i)$	L_2	$\|w\|_0$

Tableau 2. *Récapitulatif des différents modèles en fonction de leur espace d'hypothèses, de leur fonction de coût et du terme régularisant*

3.5. *Les différents modèles*

Le tableau 2 regroupe les différentes méthodes présentées dans cette partie en fonction de $\mathcal{H}$, de la fonction de coût et du terme de régularisation.

3.5.1. *Modèles linéaires*

Pour les problèmes de régression le modèle de référence est le modèle linéaire. Dans ce cas l'espace de recherche $\mathcal{H}$ est l'ensemble des fonctions linéaires avec $\phi_i(x) = x$. La solution du problème d'estimation est obtenue grâce à la résolution d'un système linéaire.

3.5.2. *Régressions ridge non linéaires*

Pour ce type d'estimateur, la fonction de coût reste une fonction quadratique. Cependant, le fait de choisir les fonctions $\phi_i(x)$ comme des fonctions non linéaires augmente la complexité de $\mathcal{H}$ et par conséquent, l'apprentissage du modèle nécessite un terme régularisant qui peut être $\|\alpha\|_2^2$ ou $\|\alpha\|_1$. Ici le choix du terme régularisant a deux conséquences : le premier d'un point de vue conceptuel et le deuxième d'un point de vue algorithmique. D'un coté, le choix d'un terme régularisant de type $\|\alpha\|_2^2$ a pour effet de pénaliser les fortes valeurs de α_i et a donc tendance à atténuer ces valeurs. De par la nature quadratique du régularisateur, la solution du problème s'obtient par la résolution d'un système linéaire dont la dimension dépend de l'ensemble ϕ_i. D'un autre côté, choisir $\|\alpha\|_1$ permet d'avoir un modèle parcimonieux et donc plus aisément interprétable. Notons que dans le cas du modèle linéaire, ce régularisateur a pour effet de sélectionner les variables pertinentes. Cependant, d'un point de vue algorithmique la résolution d'un tel problème est plus complexe et nécessite la résolution d'un problème d'optimisation quadratique et convexe ou l'utilisation d'un algorithme itératif tel que le LAR (*Least Angle Regression* (Efron *et al.*, 2003)).

3.5.3. *Réseaux de neurones artificiels (RNA)*

Les réseaux de neurones constituent une des méthodes de régression non linéaire les plus utilisées. Pour ces estimateurs l'espace de recherche $\mathcal{H}$ est la combinaison linéaire de plusieurs fonctions sigmoidales ϕ_i dont les entrées sont, elles-mêmes, des combinaisons linéaires des exemples x. Il faut dans ce cas aussi régulariser le problème pour obtenir une bonne solution. Les termes régularisants les plus fréquents sont des termes pénalisant quadratiquement les coefficients de pondération des fonctions et des entrées. Contrairement à la régression rigde non linéaire, l'identification des coefficients w_i et α_i est pour les réseaux de neurones un problème difficile. Ce problème de minimisation non linéaire et non-convexe est usuellement résolu à travers une méthode de descente de gradient garantissant une solution localement optimale.

3.5.4. *CART*

Une autre méthode de régression courante sont les arbres de régression (Breiman *et al.*, 1984). Le principe des arbres de régression est le partitionnement de l'espace des entrées et ensuite d'estimer un modèle très simple sur chaque partition de cet espace. La dénomination d'arbre de régression vient du fait que la partition se fait généralement de façon binaire et récursive. Pour chaque partition $\mathcal{A}_i$, typiquement le modèle de regréssion ϕ_i est une fonction constante ou une fonction linéaire optimisée au sens d'une fonction coût quadratique. Les arbres de régression peuvent être vus comme la solution d'un problème de minimisation d'une fonction de coût régularisée où le terme régularisant est le cardinal de la partition. Le problème d'optimisation devient également un problème plus complexe dans le sens où l'estimation f dépend essentiellement de la partition $\{A_i\}$ et ce partitionement dépend d'un critère typiquement quadratique. C'est en ce sens que CART peut être considéré comme une méthode d'apprentissage régularisé où l'espace d'hypothèses est décrit par des fonctions dépendantes des données (le partitionnement noté w dans le tableau 2) et où le terme régularisant minimise la norme ℓ_0 de cette partition (Scott *et al.*, 2003).

3.5.5. *Régressions splines*

La régression splines diffère de la régression non linéaire présentée ci-dessus du fait que l'espace de recherche est défini simplement par un noyau $\phi(x, y)$. Grâce à cette propriété, il est possible de montrer que la solution du problème de minimisation appartient à un sous-espace de $\mathcal{H}$ qui est, en fait, l'ensemble des fonctions engendrées par les combinaisons linéaires des noyaux évalués aux points x_i. Le contrôle de la complexité du modèle est réalisé à travers le choix du régularisateur, qui peut être soit la norme quadratique de f ou $\|\alpha\|_1$ si l'objectif est d'avoir un modèle parcimonieux. Comme pour la régression ridge, le principal intérêt des termes quadratiques réside dans la simplicité du problème de minimisation associé. En effet, dans ce cas, les coefficients α s'obtiennent par la résolution d'un système linéaire. Pour une régularisation de type norme ℓ_1, la solution du problème suppose la résolution d'un problème quadratique convexe ou l'utilisation d'un algorithme itératif comme le le LAR.

3.5.6. *Séparateurs à vaste marge pour la régression (SVR)*

Les Séparateurs à vaste marge pour la régression (SVR) s'apparentent à la régression splines dans le sens où l'espace de recherche est défini à travers un noyau et le terme régularisant reste la norme quadratique de la fonction f dans $\mathcal{H}$. La différence majeure entre ces deux approches réside dans le choix de la fonction de coût L_ε pour des séparateurs à vaste marge. De par la nature de cette fonction coût (de type valeur absolue), la solution unique du problème est donnée par la résolution d'un problème quadratique convexe sous contraintes. Bien que le problème d'optimisation soit difficile, les complexités empiriques de ces algorithmes peuvent être plus faibles que celles de la résolution d'un système linéaire. Une autre propriété des SVR, toujours due à la fonction de coût, est la parcimonie de la fonction de regréssion obtenue.

3.5.7. *Discussions sur les modèles*

Comme nous l'avons dit précedemment, cet article propose une comparaison théorique de plusieurs méthodes d'estimation fonctionnelle utilisables pour la prédiction de la concentration en ozone. Cette comparaison se justifie par le *no free lunch* thèoreme (Hastie *et al.,* 2001) qui dispose qu'aucune méthode de régression n'est pas plus performante que n'importe quelle autre méthode pour n'importe quel problème. La comparaison proposée ci-dessus donne un aperçu théorique des avantages et inconvénients des différentes méthodes essentiellement en termes d'interprétabilité du modèle (linéaire, non linéaire ou noyau), en termes de parcimonie de la solution (norme ℓ_1 ou norme ℓ_2 du terme régularisant) et en termes de résolution algorithmique du problème posé (système linéaire, optimisation convexe, optimisation non linéaire et non convexe). Dans ce cas encore, il n'est pas possible de déterminer le meilleur modèle car pour chacune des méthodes une propriété est contrebalancée par un désavantage (par exemple la flexibilité des réseaux de neurones engendre un problème d'optimisation non convexe). Seule une comparaison empirique des modèles permet donc de déterminer quelles sont les méthodes les mieux adaptées à la prédiction de la concentration maximale d'ozone sur les données utilisées. Les résultats peuvent s'inverser en considérant des données d'autres sites.

3.6. *Sélection de variables*

Dans un problème de classification ou de régression, le choix des variables explicatives peut être primordial pour l'obtention d'une bonne modélisation. Ainsi, des étapes de prétraitement des données visant à extraire cette information s'avèrent souvent nécessaires. Pour notre problème de prédiction de la concentration d'ozone, nous nous intéressons à ce problème de sélection de variables selon le point de vue, communément nommé dans la littérature approche *wrapper*. Soit $x \in \mathbb{R}^d$ un vecteur d'entrée de notre système et $\nu \in \{0,1\}^d$ un vecteur de pondération des variables de x. Notre objectif serait de trouver une fonction $f \in \mathcal{H}$ et un vecteur ν qui minimise le risque :

$$R_\nu[f] = \int_{\mathcal{X}\times\mathcal{Y}} L(y, f(\nu \cdot x)) dP(X,Y)$$

Par cette approche le vecteur ν optimal dénotera des variables explicatives les plus pertinentes pour le problème. Comme $P(X,Y)$ est inconnue, il est préférable de résoudre un problème approximativement équivalent (Lal *et al.*, 2005) :

$$\inf_{\nu\in\{0,1\}^d} G(f^\star,\gamma^\star,\nu,S) \quad \text{sous les contraintes} \quad \left\{ \begin{array}{l} s(\nu)\leq \nu_0 \\ \gamma^\star = T(\mathcal{H},\nu,S) \\ f^\star(x) = f_{\gamma^\star}(x) \end{array} \right. \qquad [2]$$

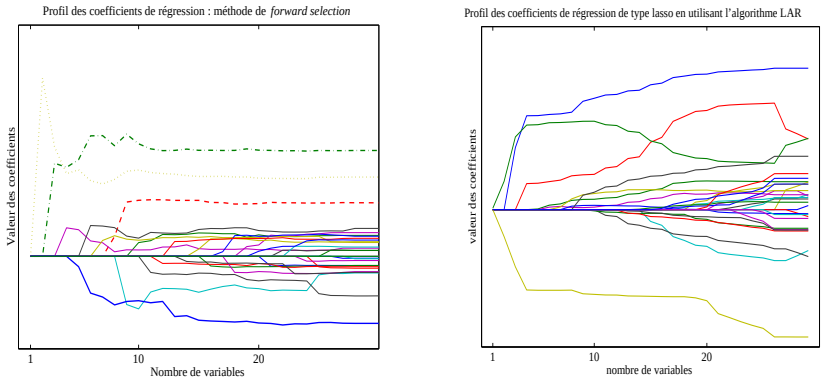
où $f^\star$ et $\gamma^\star$ sont respectivement la fonction de régression et les paramètres de cette fonction obtenus par l'algorithme d'induction T. S est l'ensemble des échantillons de la base d'apprentissage. G est une fonction mesurant la performance de $f^\star$, s une fonction mesurant la parcimonie de ν et ν_0 un paramètre défini par l'utiliseur définissant le niveau de parcimonie desiré pour ν.

De par la nature même des variables à optimiser, ce problème est un problème NP complet et est souvent résolu de façon sous optimale. Ainsi, soit le problème est relaxé en considérant que $\nu \in \mathbb{R}^d$ et en utilisant une méthode d'induction T parcimonieuse (telle qu'une régression linéaire parcimonieuse obtenue par régularisation ℓ_1), soit l'optimisation du critère se fait par énumération des sous-ensembles de variables possibles ou par une méthode heuristique de type *forward-backward*. Dans ces deux cas, le critère G optimisé est un critère d'erreur de validation.

4. Résultats obtenus

Parmi l'ensemble de méthodes présentées ci-dessus, seuls quelques algorithmes ont été étudiés : le modèle linéaire, les réseaux de neurones, les séparateurs à vaste marge pour la régression, les arbres de régression et la régression splines parcimonieuse. Dans toute cette étude il apparaît que la sélection des variables explicatives pertinentes est un des éléments clé de la qualité de la prédiction, et ce pour toutes les méthodes, y compris pour le linéaire. Pour le modèle linéaire, les techniques de sélection de type *forward/backward* et la pénalisation ℓ_1 ont donné des résultats tout à fait analogues, sauf en ce qui concerne le nombre de prédicteurs retenus. La méthode *forward/backward* sélectionne autour de 8 variables alors que la technique basée sur la pénalisation ℓ_1 retient 17 variables.

La figure 3 montre l'évolution du profil des coefficients de régression pour le modèle linéaire obtenu en réalisant une *forward selection* et par la pénalisation ℓ_1 lorsque le nombre de variables explicatives augmente. On se rend compte aisément que les variables les plus significatives sont dans les 2 cas l'ozone de la veille x_1, l'amplitude thermique x_9 (variation de température) et la valeur moyenne de l'humidité relative x_{27} sur la journée. Au vu de cette figure, le modèle de régression linéaire comporterait 7 à 8 variables explicatives.



a) b)

Figure 3. *Profil des coefficients du modèle linéaire (a) et de la régression parcimonieuse (b) lorsque le nombre de variables explicatives augmente. Les quatre variables les plus significatives pour la méthode* forward/backward *sont (par ordre d'entrée dans le modèle)* x_9, x_1 , x_{27} *et* x_6 *. Pour la pénalisation* ℓ_1*, les variables significatives sont (par ordre d'entrée dans le modèle) :* x_9, x_3, x_{27} *et* x_1

Les résultats obtenus en validation croisée pour les modèles linéaires et ceux des modèles non-linéaires présentés dans la section 2 sont décrits dans les tableaux 3 à 9. Une première analyse de ces tableaux nous montre que les performances des différents algorithmes sont plus ou moins équivalentes au sens des erreurs quadratiques moyennes J_1 et J_2 ou des erreurs absolues moyennes J_3 et J_4 excepté celles des arbres de régression. Par ailleurs, on remarque également sur le tableau 3 qu'en fonction de l'année de test, les performances fluctuent et tous les algorithmes atteignent leurs meilleures performances pour l'année de test 1999. Cette année bénéficie des concentrations d'ozone les plus faibles (voir les boîtes à moustaches, figure 2). C'est donc pour cette année que les non-linéarités sont les plus prévisibles. C'est aussi la seule pour laquelle la régression splines parcimonieuse (LAR), les réseaux de neurones et dans une moindre mesure les SVR font mieux que le modèle linéaire. Un test statistique évaluant l'égalité des médianes des erreurs quadratiques pour l'année 1999 (nommé test T de Wilcoxon ou Wilcoxon *sign rank test*) est également reporté dans le tableau (4) (Saporta, 1990). L'analyse de cette table montre que de manière significative tous les modèles testés sauf CART sont plus performants que le modèle linéaire. Par ailleurs, on constate aussi que les performances de la régression splines parcimonieuse sont significativement différentes de celles des autres modèles sauf les réseaux de neurones. Mais les performances de ces modèles sont légèrement inférieures à celles obtenues par un modèle linéaire par morceaux (Gasso *et al.*, 2005). Cela indique qu'il existe des non-linéarités dans le phénomène étudié, mais étant donné le faible nombre de données disponibles, la variabilité et la non stationnarité du phénomène, il est difficile de retrouver ces non linéarités avec les modèles présentés.

	J_1					
	95	96	97	98	**99**	Tot
SVR	239,0 (394)	274,4 (416)	309,7 (475)	261,5 (695)	**182,6** (279)	253,7 (473)
RNA	314,7 (430)	250,1 (386)	337,9 (531)	259,1 (696)	**168,6** (274)	266.5 (488)
Lin	249,5 (360)	234,4 (349)	308,6 (494)	265,7 (642)	**201,5** (309)	252.2 (449)
LAR	263,6 (418)	288,9 (462)	317,6 (502)	269,7 (722)	**163,1** (248)	261.0 (497)
CART	371,1 (466)	415,2 (615)	397,7 (710)	346,9 (1021)	**284,8** (436)	363,5 (684)

Tableau 3. *J_1 en fonction des modèles et des années. L'écart-type de l'erreur quadratique est donné entre parenthèses*

	RNA	Lin	LAR	CART
SVR	0,3414	0,1775	0,0292	0,0016
RNA	-	0,0715	0,3727	0,0002
Lin	-	-	0,0190	0,8685
LAR	-	-	-	0,0068

Tableau 4. *p-value du test de l'égalité des médianes (Wilcoxon rank sign test) des erreurs quadratiques pour différents modèles pour l'année 1999*

La technique de validation croisée utilisée pour retrouver les hyperparamètres (paramètres des noyaux, nombre de couches cachées des réseaux de neurones...) est elle aussi à mettre en question, car le caractère « universel » des modèles utilisés devrait leur permettre de faire jeu égal avec un modèle linéaire par morceaux, ce qui n'est pas le cas. Une autre explication peut être avancée pour expliquer le caractère de l'année 1999. Il s'agit de l'importance de la chronologie de la série temporelle de la concentration en ozone. En effet pour prévoir 1999 on utilise les données des quatre années précédentes, alors que pour 1998, les données utilisées ne sont pas contigües (1995, 1996, 1997 et 1999). Là encore la qualité du modèle doit être mise en regard du nombre de données disponibles.

Une attention toute particulière doit être apportée à la comparaison des performances pour les fortes valeurs d'ozones. Le tableau 5 montre que les réseaux de neurones présentent les meilleures performances sur ce critère avec 4 des meilleures performances sur les 5 années de test. Toutefois, ces résultats doivent être relativisés par le fait qu'il ne s'agit pas du critère minimisé par nos algorithmes et que les performances du modèle linéaire sont statistiquement comparables. Ces résultats sont par ailleurs corroborés par la qualité des prédictions des dépassements de seuils. En effet, les tableaux 6 et 7 relatifs à ces résultats montrent la bonne tenue des réseaux de neurones. Paradoxalement, pour les seuils extrêmes que nous avons considéré, le modèle linéaire reste le plus performant. Cependant, cela peut s'expliquer par le fait qu'il existe très peu de données d'apprentissage pour les fortes valeurs et donc dans ces régions les modèles flexibles ont tendance à surapprendre. L'utilisation d'un modèle linéaire permet dans ce cas d'éviter cet écueil. Les tableaux 8 et 9 présentent les

résultats obtenus par les différents modèles pour les critères J_3 et J_4. Les conclusions que l'on peut tirer de ces tableaux sont similaires à celles déduites pour les critères J_1 et J_2 à savoir que, hormis le modèle CART, toutes les performances des modèles non linéaires sont significativement plus intéressantes que celles du modèle linéaire que sur l'année 1999.

	J_2					
	95	96	97	98	99	Tot
SVR	346,0 (711)	420,0 (466)	477,5 (509)	842,3 (1505)	87,2 (110)	489,0 (847)
RNA	310,0 (580)	415,6 (489)	410,4 (460)	649,3 (1496)	78,0 (111)	422,5 (810)
Lin	405,2 (612)	376,5 (421)	422,8 (444)	715,2 (1368)	106,8 (133)	446,6 (755)
LAR	362,3 (783)	509,2 (607)	445,3 (482)	795,8 (1559)	129,3 (125)	501,5 (891)
CART	409,1 (627)	572,1 (719)	622,3 (733)	1057,1 (2282)	181,9 (193)	633,4 (1215)

Tableau 5. *J_2 en fonction des modèles et des années*

	Seuils													
Modèles	120		130		140		150		160		170		180	
SVM	160	55	100	50	52	45	18	43	1	32	0	17	0	6
	35	651	30	721	23	781	12	828	3	865	0	884	0	895
RNA	164	51	101	49	56	41	23	38	8	25	2	15	0	6
	44	642	36	715	24	780	14	826	5	863	0	884	0	895
Lin.	163	52	94	56	48	49	20	41	7	26	3	14	0	6
	29	722	29	722	19	785	14	826	6	862	3	881	1	894
LAR	156	59	98	52	51	46	25	36	7	26	1	16	0	6
	35	651	28	723	20	784	13	827	4	864	3	881	0	895
CART	137	78	86	64	67	30	8	53	0	33	0	17	0	6
	36	650	29	722	48	756	11	829	0	868	0	884	0	895

Tableau 6. *Prédictions de dépassements de seuils [bd md ; fa vn] (vn étant le nombre de vrais négatifs)*

5. Conclusions

Dans cet article, des modèles statistiques de type « boîte noire » ont été utilisés pour la prédiction des pics de pollution à l'ozone. Ainsi, nous avons présenté les réseaux de neurones, les SVR, la régression splines parcimonieus et une méthode basée sur les arbres de régression. D'autres techniques telles que les modèles additifs ont aussi été étudiées mais des résultats préliminaires peu prometteurs nous ont amené à ne pas pousser plus en avant notre analyse. Afin de pouvoir faire des comparaisons, un modèle linéaire a aussi été proposé. A l'issue de notre étude, nous constatons que les résultats ne sont pas à la hauteur de nos espérances. Cet échec relatif des modèles statistiques de type « boîte noire » peut être expliqué d'une part, par le manque de variables explicatives pertinentes pour expliquer le phénomène et d'autre part, par le faible nombre d'observations informatives disponibles « indépendantes ».

	Seuils						
Modèles	120	130	140	150	160	170	180
SVR	0,640	0,556	0,433	0,247	0,028	0,000	0,000
RNA	0,633	0,543	0,463	0,307	0,211	0,118	0,000
Lin.	0,647	0,525	0,414	0,267	0,179	0,150	0,000
LAR	0,624	0,551	0,436	0,338	0,189	0,050	0,000
CART	0,546	0,480	0,462	0,111	0,000	0,000	0,000

Tableau 7. $\frac{bd}{bd+fa+md}$ *en validation croisée*

	J3					
Modèles	95	96	97	98	99	Tot
SVR	12,2 (9,5)	13,0 (10,2)	13,6 (11,2)	11,8 (11,1)	**10,9 (8,0)**	12,3 (10,1)
RNA	14,2 (10,7)	12,2 (10,1)	14,2 (11,7)	11,7 (11,1)	**10,3 (7,9)**	12,5 (10,5)
Lin	12,8 (9,3)	12,2 (9,3)	13,2 (11,6)	12,2 (10,9)	**11,4 (8,5)**	12,4 (10,0)
LAR	13,0 (9,8)	13,1 (10,9)	13,8 (11,4)	11,8 (11,4)	**10,1 (7,8)**	12,4 (10,4)
CART	15,7 (11,2)	16,0 (12,6)	15,1 (13,0)	13,5 (12,8)	**13,2 (10,5)**	14,7 (12,1)

Tableau 8. J_3 *en fonction des modèles et des années (l'écart-type de l'erreur absolue moyenne est donné entre parenthèses)*

En effet, lorsque trop peu de données « indépendantes » sont disponibles, la stratégie conservatrice consistant à utiliser un modèle linéaire (avec sélection des variables pertinentes) est la plus payante. Cependant, le phénomène est non linéaire car sur le cas particulier de l'année 1999, les meilleurs résultats sont donnés par un modèle linéaire par morceaux avec seulement deux morceaux. Là encore à cause du faible nombre de données disponibles, l'utilisation d'une stratégie « quasi paramétrique » (linéaire par morceaux) est plus payante que le fonctionnement en aveugle des modèles de type « boîte noire ».

	J4					
Modèles	95	96	97	98	99	Tot
SVR	14,1 (12,4)	17,0 (11,6)	18,3 (12,1)	23,3 (17,6)	7,7 (5,5)	17,5 (13,5)
RNA	13,3 (11,8)	16,6 (11,9)	16,6 (11,8)	18,4 (17,9)	7,1 (5,6)	15,8 (13,2)
Lin	16,2 (12,2)	16,3 (10,6)	17,0 (11,7)	21,2 (16,5)	8,6 (6,0)	17,0 (12,6)
LAR	14,1 (13,1)	18,2 (13,4)	17,8 (11,5)	21,8 (18,2)	9,6 (6,4)	17,6 (13,9)
CART	15,9 (12,7)	19,3 (14,3)	20,4 (14,5)	25,0 (21,2)	11,5 (7,4)	19,7 (15,7)

Tableau 9. J_4 *en fonction des modèles et des années (l'écart-type de l'erreur absolue moyenne est donné entre parenthèses)*

Cela indique que ces modèles non paramétriques nécessitent pour ce type de problème plus de données pour estimer correctement les non linéarités. La technique de validation croisée mise en œuvre montre ici ses limites à cause du non respect de la chronologie temporelle et de la « non stationnarité » des entrées. Mais il n'est pas aisé de la remplacer. Cela nous amène à tirer quelques conclusions :

– le problème de la prévision du maximum de la concentration en ozone du lendemain est un problème non linéaire et sans doute non stationnaire pour lequel les modèles de type « boîte noire » connaissent des limites du fait du manque de données et des variables explicatives pertinentes pour mieux prévoir le phénomène,

– toutes les méthodes, pour donner des bons résultats, doivent être utilisées avec une technique de sélection de variables,

– la rapidité de certains algorithmes disponibles (LAR, SVR) rend possible l'exploration des hyper paramètres (largeur du tube, du noyau et paramètres de régularisation) liés à la régression parcimonieuse et de ce fait l'exploitation de ces techniques.

Cela indique des pistes de recherche sur le choix automatique de la métrique (les variables explicatives pertinentes et leur importance) et sur la relation entre le contrôle de la complexité de la méthode et la taille de l'échantillon disponible. Enfin, dans le cas précis de l'ozone, les limites des modèles de type « boîte noire » confirment l'intérêt d'une approche mixte permettant la coopération entre modèles physique et statistique.

6. Bibliographie

Académie des Sciences ., Ozone et propriétés oxydantes de la troposphère, Rapport n° 30, Académie des Sciences, 1993.

Ambroise C., Grandvalet Y., « Prediction of ozone peaks by mixture model », 2^{nd} *Int. Conf. on Applications of machine Learning to Ecological Modelling*, Adelaide, Australia, 2000.

APPetisse ., Literature review of statistical approaches to modelling ground ozone level at a point, Report n° IST-99-11764, APPETISE, European Community, Information Society Technology, 2001.

Breiman L., Friedman J., Olshen R., Stone C., *Classification and Regression Trees*, Wadsworth and Brooks-Cole, 1984.

Canu S., Rakotomamonjy A., « Ozone peak and pollution forecasting using support vectors », *Proc. of IFAC Workshop on Environmental modeling*, Yokohama, Japan, 2001.

Chenevez J., Jensen C., « Operational ozone forecasts for the region of Copenhagen by the Danish Meterological Institute », *Atmospheric Environment*, vol. 35, p. 4567-4580, 2001.

Cobourn G., Dolsine L., French M., Hubbard M., « A comparison of nonlinear regression and neural network models for ground-level ozone forecasting », *Journal of the Air & Waste Management Association*, vol. 50, p. 219-226, 2000.

Comrie A., « Comparing neural networks and regression models for ozone forecasting », *Journal of the Air & Waste Management Association*, vol. 47, p. 653-663, 1997.

Efron B., Johnstone I., Hastie T., Tibshirani R., « The least angle regression (LAR) algorithm for solving the Lasso », *Annals of Statistics*, 2003.

EPA Environmental Protection Agency ., Guideline for developing an ozone forecasting program, Report n° EPA-454/R-99-009, United States Environmental Protection Agency, Office of Air Quality Planning and Standards, Research Triangle Park, North Carolina 27711, july, 1999.

Gasso K., Mourot G., J. R., « Prévision des maxima journaliers d'ozone -Approche multimodèle », *Journal Européen des systèmes automatisés*, vol. 39, n°4, 2005.

Gasso K., Mourot G., Ragot J., Lebois D., Bastin E., « Différents aspects du traitement des données de pollution : validation, prédiction, explication », *Actes du Colloque Automatique et Environnement, A & E'00*, Nancy, France, 2000.

Hastie T., Tibshirani R., Friedman J., *The elements of statistical learning : data mining, inference, and prediction*, Springer, 2001.

Haykin S., *Neural networks : a comprehensive foundation*, 2^{nd} edn, Prentice Hall, Upper Saddle River, New Jersey, 1999.

Honoré C., Rouil L., Malherbe L., Bessagnet B., Vautard R., Poisson N., Colosio J., « Le système Prev'Air : cartographie et prévision de la qualité de l'air à grande échelle », *Actes de la Conférence Internationale Francophone d'Automatique CIFA'04*, Douz, Tunisie, 2004.

Lal T., Chapelle O., Weston J., Elisseeff A., *Embedded methods*, Guyon, I., S. Gunn, M. Nikravesh and L. Zadeh edn, Springer, 2005.

Mallat S., Zhang Z., « Matching Pursuit with time-frequency dictionaries », *IEEE Trans Signal Processing*, vol. 41, n° 12, p. 3397-3415, 1993.

Mourot G., Epailly R., Gasso K., Ragot J., « Prévision des pointes d'ozone par l'approche locale MOD », *Actes des 3e colloque STIC & Environnement S& E'03*, Rouen, France, p. 39-40, 2003.

Rakotomamonjy A., Canu S., « Estimation de la concentration en ozone par SVM », *Actes du Colloque Automatique et Environnement, A& E'01*, Saint-Etienne, France, 2001.

Saporta G., *Probabilités, Analyse de données et Statistique*, Editions Technip, 1990.

Scholkopf B., Herbrich R., Smola A., Williamson R., « A Generalized Representer Theorem », *Proceedings of the 14th Annual Conference on Computational Learning Theory*, number 2000-81 in *Lecture Notes in Arti cial Intelligence 2111*, p. 416-426, 2001a.

Scholkopf B., Smola A., *Learning with Kernels*, MIT Press, 2001b.

Scott C., Willett R., Nowak R., « CORT : Classification Or Regression Trees », *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2003.

Vannoorenberghe P., Denoeux T., « Diagnostic de la pollution atmosphérique par une approche RDF utilisant les fonctions de croyance », *Actes du Colloque Automatique et Environnement, A & E'2001*, Saint-Etienne, France, 2001.

Vapnik V., *Statiscal learning theory*, Wiley, 1998.

Vautard R., M. B., Roux J. Gombert D., « Validation of a hybrid forecasting system for the ozone concentrations over the Paris area », *Atmospheric Environment*, vol. 35, p. 2449-2461, 2001.

Zolghadri A., « A combined hard and soft computing approach for ground level ozone monitoring in Bordeaux », *Actes des 3e colloque STIC & Environnement S& E'2003*, Rouen, France, p. 15-21, 2003.