

REVUE DE STATISTIQUE APPLIQUÉE

B. GHATTAS

Prévisions des pics d'ozone par arbres de régression, simples et agrégés par Bootstrap

Revue de statistique appliquée, tome 47, n° 2 (1999), p. 61-80

http://www.numdam.org/item?id=RSA_1999__47_2_61_0

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

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

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

PRÉVISIONS DES PICS D'OZONE PAR ARBRES DE RÉGRESSION, SIMPLES ET AGRÉGÉS PAR BOOTSTRAP

B. Ghattas

GREQAM – Université de la Méditerranée
2, rue de la Charité, 13002 Marseille
ghattas@ehess.cnrs-mrs.fr

RÉSUMÉ

Cet article fournit quelques résultats de prévisions de pics de pollution d'ozone pour la station de mesure de Vitrolles, située dans l'Aire Métropolitaine Marseillaise. Nous présentons d'abord la méthode régression par arbre et une idée de stabilisation par Agrégation par Bootstrap. Ces méthodes sont ensuite utilisées pour résoudre le problème de prévision du maximum d'ozone de la journée. Les résultats obtenus semblent être de qualité comparable à ceux des équipes internationales.

Mots-clés : arbres de régression, agrégation par bootstrap, prévision, pollution atmosphérique, CART.

ABSTRACT

This paper gives several results of the forecasting of severe ozone episodes for one of the Marseille's Region measurements points, Vitrolles. We first present the Tree Regression method and one idea of stabilization by using bootstrap agregation. These methods are used to resolve the problem of forecasting daily maximum ozone. The quality of the obtained results seems to be comparable to those of other international teams.

Keywords : regression trees, bootstrap agregating, forecasting, atmospheric pollution, CART.

1. Introduction

Cette étude réalisée dans le cadre d'une collaboration contractuelle avec AIRMARAIX¹, concerne la prévision des concentrations maximales quotidiennes de l'ozone sur l'Aire Métropolitaine Marseillaise.

¹ Association pour la surveillance de la qualité de l'air de l'Est des Bouches du Rhône, du Var et du Vaucluse.

Le travail de prévision de l'ozone concerne seize stations des réseaux AIR-MARAIX et AIRFOBEP²; il s'agit dans un premier temps de prévoir le niveau du maximum d'ozone du jour j (situé en général l'après midi) à 06 h UTC (*i.e.* 8 h localement en été), puis dans un deuxième temps la veille.

Ces prévisions permettront de déclencher les procédures d'information des autorités administratives et du public, et des décisions concernant les émetteurs des précurseurs (industries et circulation automobile) et le public (activités scolaires en plein air, personnes âgées par exemple).

Par opposition à des méthodes de modélisation prenant en compte des phénomènes complexes à différentes échelles de temps et d'espace, les approches statistiques de ce problème sont plus légères, ponctuelles au niveau spatial et concernent évidemment le court terme. Quelques travaux cités en bibliographie illustrent ce propos et montrent la diversité des méthodes (Bel *et al.* (1997), Eder *et al.* (1994), Ryan (1995), Sheifinger *et al.* (1996)).

Les données utilisées, décrites ultérieurement dans le détail sont des données de concentrations de polluants (ozone, dioxyde d'azote et dioxyde de soufre) et météorologiques (température, vent, gradient thermique vertical, nébulosité, insolation, humidité et rayonnement global). La période de l'étude concerne les jours du début avril à la fin septembre, de 1992 à 1996. Les résultats présentés ici portent essentiellement sur une seule station, celle de Vitrolles.

La méthode présentée au cours de cet article est la régression non paramétrique par arbre. C'est une des composantes des méthodes CART introduites dans Breiman *et al.* (1984), présentées ensuite dans Devroye *et al.* (1996) et dans Venables *et al.* (1990). Dans un contexte semblable mais concernant aussi d'autres polluants atmosphériques, Bel *et al.* (1997) utilisent un modèle non paramétrique général et une spécialisation additive. La variante classification des méthodes CART a été présentée par Gueguen-Nakache (1988); elle constitue l'un des éléments de la panoplie des méthodes mises en œuvre pour nos études des pics de pollution.

L'intérêt de cette méthode de régression non paramétrique est multiple. D'abord elle est simple, additive, prend en compte l'interaction entre les variables explicatives (quantitatives ou qualitatives) et ne limite pas le nombre de ces variables. Ensuite elle œuvre simultanément sur le plan descriptif (étiquetage de nœuds de l'arbre) et décisionnel (construction de règles de prévision). Elle connaît cependant un inconvénient majeur : l'*instabilité* (Breiman, 1996a) au sens où un petit changement dans l'échantillon d'apprentissage peut avoir un effet important sur le prédicteur.

Le plan de l'article est le suivant : après cette introduction, on présente les principales idées de la régression par arbre et une idée de stabilisation par agrégation par bootstrap. Ensuite sont présentés des résultats obtenus dans l'étude de prévision du maximum de concentrations d'ozone à Vitrolles.

² Association pour la surveillance de la qualité de l'air de la zone Fos – Etang de Berre.

2. Méthodologie

2.1 Généralités

Soit $\mathcal{E} = (\mathbf{x}_i, y_i)_{1 \leq i \leq N}$ un échantillon d'apprentissage de taille N où les y_i sont les observations de la variable expliquée Y et les $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{id})$ sont les observations du vecteur des variables explicatives \mathbf{X} à valeurs dans R^d .

La fonction de régression ou prédicteur

$$y(\mathbf{x}, \mathcal{E}) = E[Y | \mathbf{X} = \mathbf{x}]$$

est de la forme

$$y(\mathbf{x}, \mathcal{E}) = \sum_{j=1}^p c_j I(\mathbf{x} \in F_j) \quad (1)$$

les c_j sont des constantes, les F_j des polytopes de R^d dont les côtés sont parallèles aux axes de coordonnées constituant une partition de R^d , et I est la fonction indicatrice.

La surface de régression est constante par morceaux; c'est un régressogramme pour lequel il s'agit d'estimer les c_j mais aussi de déterminer les classes F_j de la partition ainsi que p .

Le modèle est *additif et non paramétrique*. Les fonctions à estimer sont simples, mais prennent en compte les interactions entre variables explicatives.

Lorsque les F_j sont connus, le meilleur estimateur dans le sens des moindres carrées des c_j est :

$$\hat{c}_j = \frac{1}{|\gamma(F_j)|} \sum_{i \in \gamma(F_j)} y_i \quad (2)$$

où $\gamma(F_j) = \{i, 1 \leq i \leq N; \mathbf{x}_i \in F_j\}$ et $|\gamma(F_j)|$ est le nombre d'éléments de $\gamma(F_j)$; c'est la moyenne de y dans la classe F_j .

Le prédicteur, lui aussi constant par morceaux est alors obtenu en remplaçant les c_j par les \hat{c}_j obtenus en (2) soit

$$\hat{y}(\mathbf{x}, \mathcal{E}) = \sum_{j=1}^p \hat{c}_j I(\mathbf{x} \in F_j) \quad (3)$$

Le problème est donc de déterminer les F_j . La méthode *séquentielle* de Régression par Arbre proposée pour construire la partition $(F_j)_{1 \leq j \leq p}$ est due à Breiman *et al.* (1984).

2.2 Estimation séquentielle du modèle

Supposons d'abord construit un estimateur $\hat{y}(\cdot, \mathcal{E})$ de $y(\cdot, \mathcal{E})$

2.2.1 Erreur quadratique et estimateurs

Le coût de la régression est évalué par son erreur quadratique

$$R(y) = E(Y - y(\mathbf{x}, \mathcal{E}))^2 \quad (4)$$

L'estimation de cette erreur quadratique est à la base de l'estimation du modèle et de la mesure de la qualité du prédicteur. Plusieurs procédures d'estimation sont proposées et utilisées à différentes étapes de la méthodologie.

2.2.1.1 L'estimation par substitution

$$\hat{R}(y) = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}(\mathbf{x}_i, \mathcal{E}))^2 \quad (5)$$

Cet estimateur est à la base de la construction de \hat{y} mais il est bien connu qu'un tel estimateur est trop optimiste pour mesurer sa qualité en particulier lorsque p est grand. Deux autres estimateurs sont proposés.

2.2.1.2 Par échantillon témoin

Un premier estimateur pour R est classiquement utilisé pour l'étude de la validation, c'est celui obtenu à l'aide d'un échantillon témoin. Supposons que l'on dispose d'un second échantillon $\mathcal{T} = (\mathbf{x}_i, y_i)_{1 \leq i \leq N'}$, une fois \hat{y} obtenu à partir de \mathcal{E} on estime $R(y)$ par :

$$\hat{R}^{et}(y) = \frac{1}{|\mathcal{T}|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{T}} (y_i - \hat{y}(\mathbf{x}_i, \mathcal{E}))^2 \quad (6)$$

Un tel estimateur élimine le biais produit par l'estimateur par substitution, mais nécessite un grand volume de données.

2.2.1.3 Par validation croisée

Le second estimateur du risque utilisé est celui obtenu par validation croisée, qui lui aussi limite le biais d'optimisme et ne nécessite pas d'échantillon témoin. Supposons \mathcal{E} partitionné aléatoirement en K sous échantillons de tailles voisines, soient $(\mathcal{E}_k)_{1 \leq k \leq K}$. Notons $\mathcal{E}^{(k)} = \mathcal{E} - \mathcal{E}_k$, $1 \leq k \leq K$. Sur chacun des échantillons $\mathcal{E}^{(k)}$ estimons $y(\mathbf{x}, \mathcal{E}^{(k)})$ par $\hat{y}(\mathbf{x}, \mathcal{E}^{(k)})$, \mathcal{E}_k jouant le rôle d'échantillon témoin pour $\mathcal{E}^{(k)}$. L'estimateur de $R(y)$ par validation croisée est alors :

$$\hat{R}^{vc}(y) = \frac{1}{N} \sum_{k=1}^K \sum_{(\mathbf{x}_i, y_i) \in \mathcal{E}_k} (y_i - \hat{y}(\mathbf{x}_i, \mathcal{E}^{(k)}))^2 \quad (7)$$

2.2.2 Estimation initiale du modèle (1)

Présentons la construction séquentielle du prédicteur. Il s'agit de déterminer une partition de R^d en p classes, où p n'est pas fixé *a priori*.

Soit $s \in R$ un seuil et $\mathbf{x} = (x_1, x_2, \dots, x_d) \in R^d$. On peut construire une partition de R^d en deux classes, soit :

$$t_g(l, s) = \{\mathbf{x} \in \mathcal{E}; x_l < s\}$$

$$t_d(l, s) = \{\mathbf{x} \in \mathcal{E}; x_l \geq s\}$$

$$\gamma(t_g(l, s)) = \{i; x_i \in t_g(l, s), 1 \leq i \leq N\}$$

$$\gamma(t_d(l, s)) = \{i; x_i \in t_d(l, s), 1 \leq i \leq N\}$$

$$\hat{R}_d(l, s) = \sum_{i \in \gamma(t_d(l, s))} (y_i - \bar{y}_{t_d(l, s)})^2 \text{ et } \hat{R}_g(l, s) = \sum_{i \in \gamma(t_g(l, s))} (y_i - \bar{y}_{t_g(l, s)})^2$$

où

$$\bar{y}_{t_d(l, s)} = \frac{1}{|\gamma(t_d(l, s))|} \sum_{i \in \gamma(t_d(l, s))} y_i$$

et $\gamma(t_d(l, s))$ est l'ensemble des indices des observations appartenant à $t_d(l, s)$. La moyenne $\bar{y}_{t_g(l, s)}$ se définit de façon analogue. $\hat{R}_d(l, s)$ ³ (resp. $\hat{R}_g(l, s)$) est la somme des carrés des écarts à la moyenne (SCEM) de $t_d(l, s)$ (resp. $t_g(l, s)$).

Il s'agit de déterminer le couple (l^*, s^*) qui minimise la somme des SCEM de t_d et t_g

$$(l^*, s^*) = \arg \text{Min}_{\substack{l \leq d \\ s \in R}} \hat{R}_d(l, s) + \hat{R}_g(l, s) \quad (8)$$

Le couple (l^*, s^*) correspond à un seuil s^* de partage de \mathcal{E} à l'aide de la variable x_{l^*} (règle binaire). Notons alors $t_g = t_g(l^*, s^*)$ et $t_d = t_d(l^*, s^*)$ où (t_g, t_d) constitue une partition de \mathcal{E} optimale vis-à-vis du critère (8). Notons $t_1 = \mathcal{E}$, $t_2 = t_g$, $t_3 = t_d$ et appelons t_g et t_d les *descendants* de t_1 . Réalisons sur t_2 et sur t_3 la même procédure que sur t_1 . On obtient une nouvelle partition optimale de \mathcal{E} en quatre classes (t_4, t_5, t_6, t_7) et ainsi de suite.

Cette procédure de construction de partitions s'arrête naturellement lorsqu'il n'y a plus qu'un seul élément par classe, ou lorsque dans chaque classe les mesures des observations sont identiques.

On peut *visualiser* cette suite emboîtée de partitions par un arbre binaire A avec les conventions suivantes : $t_1 = \mathcal{E}$ est la *racine*, les t_i sont les *nœuds*, les nœuds extrêmes *i.e.*, ceux de la partition la plus fine, sont les *feuilles*, \tilde{A} est l'ensemble des

³ Suivant le cas \hat{R} désignera soit une variance (cf. formule (5) ou (9) par exemple), soit comme c'est le cas ici, une somme des carrés des écarts à la moyenne. Les deux notations qui, compte tenu du contexte, ne prêtent pas à confusion, ont été gardées pour ne pas alourdir le texte inutilement.

feuilles de A et A^t est la *branche* de A issue du nœud t , *i.e.* l'ensemble de tous les descendants du nœud t . Les feuilles correspondent aux F_j du modèle (1).

Sur le plan du calcul, notons que pour N observations de chaque variable nous avons au plus $(N - 1)$ possibilités de seuils. C'est donc parmi $(N - 1)d$ couples (variable, seuil) que sera déterminé (l^*, s^*) .

La figure 1 montre l'exemple d'un arbre à six feuilles.

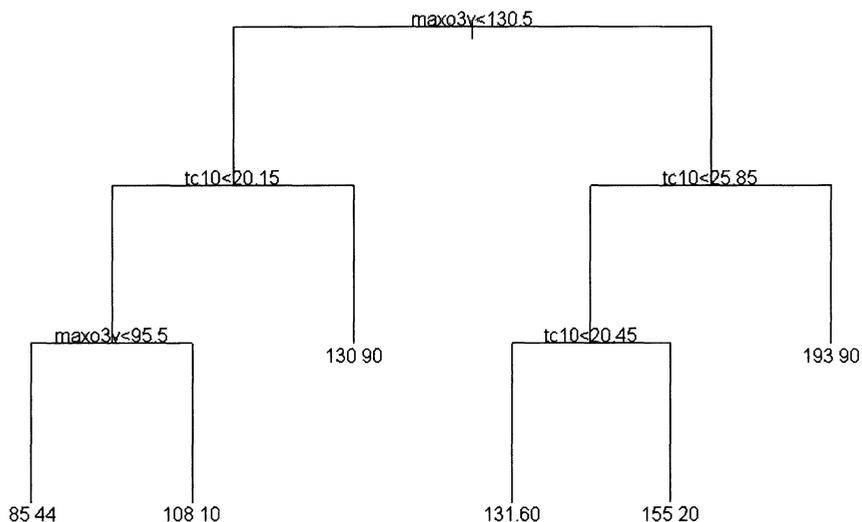


FIGURE 1

*variable expliquée, le maximum d'ozone du jour j («maxo3»)
variables explicatives, le maximum d'ozone du jour $j-1$ («maxo3v»)
et la température du jour j à 10h («tc10»).*

On rend compte du maximum d'ozone du jour j («maxo3») par les deux variables explicatives («maxo3v») (maximum d'ozone de la veille) et («tc10») (températures à 10 h T.U. du jour j). Aux feuilles apparaissent les valeurs moyennes du maximum d'ozone dans la feuille. La variable de la première règle est le maximum d'ozone de la veille («maxo3v») avec le seuil de $130.5 \mu\text{g}/\text{m}^3$. La branche de gauche d'un nœud correspond à la réalisation de la règle binaire au niveau de ce nœud. La branche de droite correspond à sa non réalisation.

Remarques concernant le critère (8)

Le critère présenté et utilisé ici peut être remplacé par d'autres critères souvent plus difficiles à mettre en œuvre, mais intéressants pour certaines applications. Voici quelques exemples :

- Breiman *et al.* (1984) remplace la somme des variances intra-classes par celle des écarts absolus moyens et les c_j par les médianes intra-classes. Les prédicteurs obtenus sont plus robustes vis-à-vis des données aberrantes.
- Chaque nœud est partagé à l'aide d'un critère portant sur les signes des résidus d'une régression polynomiale de y sur les variables explicatives (Chaudhuri *et al.* (1995))

D'autres extensions ont été publiées. Nous citerons Loh *et al.* (1988) et Ciampi (1991) par exemple.

2.2.3 Quel arbre ?

Si la procédure d'arrêt de la construction de l'arbre est celle décrite précédemment 2.2.2, l'arbre obtenu est noté A_{\max} , car il contient un nombre de feuilles maximum. Un tel arbre ne présente aucun intérêt pratique, car il a autant de feuilles que d'observations.

Une fois obtenu un arbre, comment l'utiliser ?

- 1 Sur le plan descriptif, chaque nœud est étiqueté par une variable et un seuil. Par exemple pour le premier nœud «la température maximale du jour j » est la variable et le seuil est 20.15°C .
- 2 Sur le plan prédictif, supposons que l'on dispose aujourd'hui de l'observation des variables explicatives utilisées dans le modèle. Cette observation va cheminer dans l'arbre à l'aide des règles précédentes pour se retrouver dans une feuille F_j . La valeur prévue pour y est alors \hat{c}_j .

Tenant compte de l'introduction de la notion d'arbre, nous appellerons maintenant l'erreur (quadratique) de l'arbre, et nous écrivons $\hat{R}(A)$ au lieu de $\hat{R}(y)$. On vérifie facilement que :

$$\hat{R}(A) = \frac{1}{N} \sum_{t \in \hat{A}} \hat{R}(t) \quad (9)$$

où
$$\hat{R}(t) = \sum_{x_i \in t} (y_i - \hat{y}(x_i, \mathcal{E}))^2$$

La variation de la qualité de l'arbre dûe au partage d'un nœud t est donnée par :

$$\Delta \hat{R}(s, t) = \frac{1}{N} \left[\hat{R}(t) - \hat{R}(t_g) - \hat{R}(t_d) \right] \quad (10)$$

Par construction $\Delta \hat{R}(s, t) > 0$. Donc $\hat{R}(A)$ décroît lorsque le nombre de feuilles croît, et en particulier $\hat{R}(A_{\max}) = 0$.

Comment réduire la taille de A_{\max} ?

Par exemple en arrêtant la procédure séquentielle :

- lorsque pour tout nœud $\Delta \hat{R}(s, t) \leq \lambda$.
- lorsque le nombre d'observations de chaque nœud devient inférieur à ν .

Si λ ou ν sont petits, le nombre de feuilles sera grand, et petit dans le cas contraire. Dans les deux cas l'estimation de $y(\mathbf{x})$ ne sera pas satisfaisante.

Finalement Breiman *et al.* (1984) proposent de conjuguer validation croisée et procédure d'élagage des arbres pour résoudre ce problème avec un certain succès.

2.3 Procédure d'élagage

2.3.1 les principes et les définitions

- *Elaguer* un arbre A à partir d'un nœud t consiste à supprimer de A l'ensemble A^t (la branche de A issue de t). On note $A - A^t$ l'arbre élagué au nœud t . Tout arbre A' obtenu par élagage de l'arbre A est un *sous-arbre* de A et l'on note $A' < A$, et A' est dit *emboîté* dans A .
- La recherche d'un sous-arbre optimal pourrait consister à considérer tous les sous-arbres de A_{\max} et à les comparer à l'aide d'un échantillon témoin. Le nombre (fini) de sous-arbres est en général très grand, ce qui rend difficiles les calculs. De plus l'arbre optimal ainsi obtenu serait optimal vis-à-vis d'un échantillon témoin, deuxième difficulté.
- On note que si $A' < A$ alors $\widehat{R}(A) \leq \widehat{R}(A')$. Une idée consiste à pénaliser un trop grand nombre de feuilles par l'introduction d'un nouveau critère, soit :

$$\widehat{R}_\alpha(A) = \widehat{R}(A) + \alpha \left| \widetilde{A} \right| \quad (11)$$

où $\alpha \geq 0$ et $\left| \widetilde{A} \right|$ est le nombre de feuilles de l'arbre A . Le terme $\alpha \left| \widetilde{A} \right|$ peut être interprété comme un coût de complexité et l'erreur pénalisée $\widehat{R}_\alpha(A)$ est à rapprocher des critères d'information introduits par Akaike.

On remarque que :

- Si $\alpha = 0$ alors $\widehat{R}_0(A) = \widehat{R}(A)$
- Si t est un nœud alors $\widehat{R}_\alpha(t) = \widehat{R}(t) + \alpha$.

2.3.2 L'algorithme d'élagage

Posons $\alpha = 0$. Soit A_0 le plus petit sous arbre de A_{\max} tel que $\widehat{R}(A_0) = \widehat{R}(A_{\max})$. L'arbre A_0 est obtenu par l'élagage de A_{\max} aux différents nœuds t tels que : $\widehat{R}(t) = \widehat{R}(t_g) + \widehat{R}(t_d)$.

Alors pour tout nœud t de A_0 l'inégalité $\widehat{R}_\alpha(t) \geq \widehat{R}_\alpha(A_0^t)$ est vraie pour $\alpha = 0$ et demeure vraie tant que α reste inférieur à la valeur

$$c(t, A_0^t) = \frac{\widehat{R}(t) - \widehat{R}(A_0^t)}{\left| \widetilde{A}_0^t \right| - 1} \quad (12)$$

Quand α atteint cette valeur $\hat{R}_\alpha(t) = \hat{R}_\alpha(A_0^t)$.

Posons

$$\alpha_1 = \min_{t \in A_0} c(t, A_0^t)$$

et construisons l'arbre $A_1 = A(\alpha_1)$ obtenu en élaguant A_0 à tous les nœuds minimisant $c(t, A_0^t)$. Par cette procédure on supprime par élagage les branches A_0^t de l'arbre telles que la variation de la qualité de l'arbre par cet élagage $\hat{R}(A_0 - A_0^t) - \hat{R}(A_0)$ est «petite». Ce qui réalise le compromis souhaité lors de l'introduction du critère \hat{R}_α .

On procède de même sur A_1 comme sur A_0 , on détermine $\alpha_2 > \alpha_1$ tel que

$$\alpha_2 = \min_{t \in A_1} c(t, A_1^t)$$

et on élague A_1 à tous les nœuds minimisant $c(t, A_1^t)$ pour obtenir $A_2 = A(\alpha_2)$.

Notons $A_l = A(\alpha_l)$. Par cette procédure on construit une suite finie d'arbres emboîtés $A_0 > A_1 > \dots > A_L$ où A_L est la racine de l'arbre A_0 , et une suite croissante α_l $1 \leq l \leq L$ de coefficients de pénalisation.

De plus, par construction si $\alpha \in [\alpha_l, \alpha_{l+1}[$, pour tout nœud t de A_l :

$$\hat{R}_\alpha(t) > \hat{R}_\alpha(A_l^t) .$$

Nous sommes ramenés à déterminer dans la suite (A_l) un arbre optimal, vis-à-vis du critère d'erreur quadratique.

2.4 Choix d'un bon arbre

2.4.1 Par échantillon témoin

Ayant construit la suite A_l $0 \leq l \leq L$ d'arbres suivant la procédure d'élagage sur l'échantillon d'apprentissage \mathcal{E} , si nous notons \hat{y}^l les prédicteurs associés à chacun des A_l , on choisit l'arbre A_{l_0} tel que :

$$\alpha_{l_0} = \text{Arg min} \frac{1}{N'} \sum_{(\mathbf{x}_n, y_n) \in \mathcal{T}} (y_n - \hat{y}^l(\mathbf{x}_n, \mathcal{E}))^2 \quad (13)$$

où \mathcal{T} est l'échantillon témoin de taille N' .

2.4.2 Par validation croisée

Reprenons les définitions et notations de 2.2.1. Pour chaque k tel que $1 \leq k \leq K$ construisons sur le sous échantillon $\mathcal{E}^{(k)}$ une suite d'arbres $A^k(\alpha_l^k)$ (comme en 2.3) de coût pénalisé minimum pour α . Procédons à la même construction sur la totalité

de l'échantillon d'apprentissage \mathcal{E} . Notons (A_l) la suite d'arbres obtenus et (α_l) les coûts de complexité correspondants.

Posons $\alpha'_l = \sqrt{\alpha_l \alpha_{l+1}}$ la moyenne géométrique de α_l et α_{l+1} . On sait que $A(\alpha'_l) = A(\alpha_l)$ (2.3.2). L'estimateur de l'erreur pour l'arbre $A(\alpha'_l)$ obtenu par validation croisée est donné comme en (7) par :

$$\hat{R}^{vc}(A(\alpha'_l)) = \frac{1}{N} \sum_{k=1}^K \sum_{(\mathbf{x}_n, y_n) \in \mathcal{E}_k} \left(y_n - \hat{y}'(\mathbf{x}_n, \mathcal{E}^{(k)}) \right)^2 \quad (14)$$

où $\hat{y}'(\mathbf{x}_n, \mathcal{E}^k)$ est l'estimateur de $y(\mathbf{x}_n)$ associé à l'arbre $A^k(\alpha'_l)$.

L'arbre optimal vis-à-vis de ce critère est $A(\alpha'_{opt})$ tel que

$$\hat{R}^{vc}(A(\alpha'_{opt})) = \text{Min}_{\alpha'_l} \hat{R}^{vc}(A(\alpha'_l))$$

2.4.3 Par construction manuelle

Nous proposons une construction interactive de l'arbre de prévision obtenue de la manière suivante :

1. On construit un arbre à un petit nombre de feuilles (3 ou 4).
2. On décide de partager les feuilles par examen graphique de la distribution de la variable Y ainsi que de son écart-type dans la feuille.
3. On arrête cette procédure lorsque les distributions des feuilles d'intérêt sont homogènes.
4. Ayant construit «à la main» plusieurs arbres, le choix d'un arbre peut être fait comme en 2.4.1 à l'aide d'un échantillon témoin.

2.5 Agrégation par bootstrap

2.5.1 Généralités

La régression par arbre, comme l'illustrent les exemples, œuvre à la fois sur le plan descriptif et sur le plan décisionnel, ce qui en constitue un des intérêts majeurs.

Cependant, cet avantage est contrebalancé par un inconvénient : l'*instabilité*, mise en évidence et étudiée par Breiman (1996a).

Une idée classique pour améliorer la stabilité des prédicteurs obtenus consiste à les agréger. Si l'on dispose de K échantillon $(\mathcal{E}^{(k)})_{1 \leq k < K}$ de même taille N , et de même loi, l'idée à mettre en œuvre consiste à calculer K prédicteurs à l'aide de K arbres soit $\hat{y}_N(\mathbf{x}, \mathcal{E}^{(k)})$ et d'estimer, par agrégation un nouveau prédicteur soit :

$$\hat{y}^a(\mathbf{x}) = \frac{1}{K} \sum_{k=1}^K \hat{y}(\mathbf{x}, \mathcal{E}^k) \quad (15)$$

Cela dit, en pratique un seul échantillon \mathcal{E} est disponible, et les $(\mathcal{E}^{(k)})_{1 \leq k \leq K}$ sont remplacés par $(\mathcal{E}_B^{(k)})_{1 \leq k \leq K}$ obtenus par bootstrap de \mathcal{E} ; on rappelle que chaque $\mathcal{E}_B^{(k)}$ est un échantillon de taille N tiré au hasard, mais avec remise, dans \mathcal{E} . On construit ensuite un prédicteur agrégé à l'aide de ces échantillons bootstrap, dont l'estimateur est :

$$\hat{y}^B(\mathbf{x}, \mathcal{E}) = \frac{1}{K} \sum_{k=1}^K \hat{y}(\mathbf{x}, \mathcal{E}_B^{(k)}) \quad (16)$$

Cette procédure d'agrégation par bootstrap est proposée par Breiman (1996b) sous le nom «Bagging». Comme le montreront les applications et comme le justifie la théorie, l'agrégation par bootstrap améliore la qualité des prédicteurs, mais il n'existe plus d'arbre associé à la régression.

3. Prédiction du maximum d'ozone

3.1 Généralités

3.1.1 Choix de la station de mesure

L'aire métropolitaine Marseillaise possède actuellement seize points de mesure d'ozone (Ghattas *et al.* (1998)), en fonctionnement depuis deux ans seulement pour les plus récentes.

Même si les travaux sur le sujet portent sur l'ensemble des stations, et qu'une typologie préalable a permis de déterminer deux classes relativement homogènes, c'est sur la station de Vitrolles que portent la plupart des essais méthodologiques que nous présentons, d'autres stations étant cependant utilisées pour valider certains résultats.

3.1.2 Données

La variable expliquée est le maximum d'ozone du jour j : «*maxO3*».

Les variables explicatives sont au nombre de 41. Le nombre de journées est égal à 822.

1) Les variables de pollution

- le maximum d'ozone du jour $j - 1$; «*maxO3v*»
- l'ozone du jour j à 1 h, 4 h, 6 h; «*o1*», «*o4*», «*o6*»
- le maximum des trois mesures du dioxyde de soufre du jour j à 1 h, 4 h, 6 h; «*maxs2*»
- le maximum du dioxyde d'azote du jour $j - 1$ entre 15 h et 24 h; «*maxn2v*»
- le dioxyde d'azote du jour j à 1 h, 4 h et 6 h; «*no1*», «*no4*», «*no6*»

2) Les variables météorologiques :

On utilise les observations trihoraires du jour j de 1 h à 16 h pour :

- Nébulosité; «*neb1*», «*neb4*» ... «*neb16*»
- Températures; «*tc1*», «*tc4*» ... «*tc16*»
- vitesse du vent; «*vv1*», «*vv4*» ... «*vv16*»
- direction du vent; «*dv1*», «*dv4*» ... «*dv16*»

Cette variable est qualitative à 8 modalités correspondant aux intervalles suivants $[0^\circ \dots 20^\circ]$, $[20^\circ \dots 60^\circ]$, $[60^\circ \dots 160^\circ]$, $[160^\circ \dots 220^\circ]$, $[220^\circ \dots 280^\circ]$, $[280^\circ \dots 360^\circ]$

Les autres variables sont :

- Insolation (trihoraire de 7 h à 16 h); «*ins7*», ..., «*ins16*»
- Humidité minimale du jour j ; «*hum*»
- Rayonnement global du jour j ; «*ray*»
- le maximum du gradient thermique vertical du jour j mesuré à 1h, 4h, 6h; «*maxgrd*»
- le maximum des températures du jour $j - 1$; «*maxtcv*»

Dans cette phase préliminaire nous utilisons les données météorologiques observées du jour j en attendant de disposer et d'utiliser les prévisions météorologiques correspondantes afin de procéder à une *prévision parfaite* (Wilks, (1995)).

3.2 Evaluation des résultats

Quatre niveaux de pollution ou d'alerte en ozone sont actuellement considérés en fonction des valeurs observées. Ces niveaux d'alerte sont :

- niveau 0 : $0 < O_3 < 130 \mu \text{ g/m}^3$
- niveau 1 : $130 < O_3 < 180 \mu \text{ g/m}^3$ (seuil d'information de la population)
- niveau 2 : $180 < O_3 < 280 \mu \text{ g/m}^3$
- niveau 3 : $280 < O_3 < 360 \mu \text{ g/m}^3$ (seuil d'alerte à la pollution)

Les résultats obtenus par les modèles seront résumés dans un tableau de contingence (4*4), appelé ici *tableau d'alerte*, croisant les effectifs d'observations (en ligne) avec les effectifs des prévisions (en colonne) pour chacun des quatre niveaux.

		prévu			
		a_{00}	a_{01}	a_{02}	a_{03}
observé		a_{10}	a_{11}	a_{12}	a_{13}
		a_{20}	a_{21}	a_{22}	a_{23}
		a_{30}	a_{31}	a_{32}	a_{33}

a_{ij} représente le nombre d'observations de niveau i prévu au niveau j .

- **Threat Score**

Ce critère est largement utilisé en météorologie et en pollution pour évaluer la qualité des prévisions. Supposons qu'on dispose du tableau de contingence suivant

bien prévus C1	surestimés
sous-estimé	bien prévus C2

C1 correspond par exemple aux observations de niveau 0, et C2 aux observations de niveau supérieur.

Le Threat Score (TS) est défini par :

$$TS = \frac{\text{bien prévus } C2}{\text{sous-estimés} + \text{bien prévus } C2 + \text{surestimés}}$$

Nous l'utiliserons pour évaluer la qualité de l'arbre construit, sur l'échantillon d'apprentissage et sur un échantillon témoin. Nous présentons des résultats avec le seuil de niveau 0 ($130\mu \text{ g/m}^3$), le threat score s'écrit alors

$$TS = \frac{\sum_{1 \leq i, j \leq 3} a_{ij}}{\sum_{0 \leq i, j \leq 3} a_{ij} - a_{00}} \quad (17)$$

Le sous tableau $(a_{ij})_{1 \leq i, j, \leq 3}$ correspond à la classe C2.

- On pourra comparer les résultats de prévision fournis par les arbres, à ceux obtenus par la *persistance*, qui consiste à prévoir le maximum du jour j comme étant égal au maximum observé du jour $j - 1$.

3.3 Résultats

Pour évaluer la qualité des arbres construits sur la station Vitrolles, nous avons utilisé les observations obtenues durant la même période pour la station Rognac, comme échantillon témoin (685 observations). Cette station est géographiquement voisine de Vitrolles, et les distributions de l'ozone dans les deux stations sont semblables (Ghattas *et al.*, 1998).

3.3.1 Arbre obtenu par validation croisée

La recherche d'un arbre de taille optimale est faite par validation croisée, sur vingt échantillons stratifiés⁴. L'arbre ainsi obtenu possède 15 feuilles (figure 2). On indique au niveau des feuilles, la moyenne des observations de la variable expliquée,

⁴ La stratification permettant d'obtenir des échantillons dont les niveaux 0, 1, 2 et 3 ont des proportions semblables à celles de l'échantillon de base

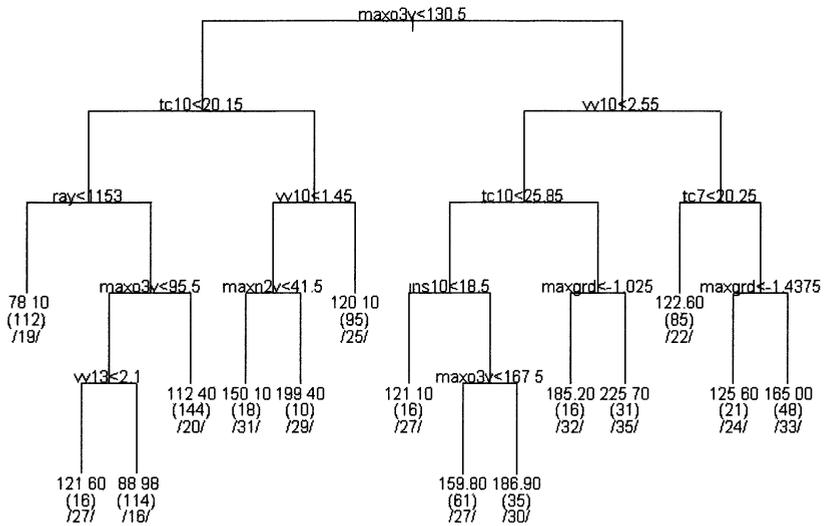


FIGURE 2

Vitrolles, Arbre de validation croisée, 15 feuilles.

La branche de gauche d'un nœud correspond à la réalisation de la règle binaire au niveau de ce nœud.

La branche de droite correspond à sa non réalisation.

le nombre d'observations dans chaque feuille (entre parenthèses) et l'écart type de ces observations (entre //).

Le croisement des valeurs prévues (en ordonnée) contre les valeurs observées (en abscisse) permet d'apprécier la qualité et les erreurs du modèle associé à cet arbre (figure 3). Les lignes verticales et horizontales représentent les niveaux d'alerte décrits ci-dessus.

L'examen des premiers nœuds est intéressant. La variable de la première règle est le maximum d'ozone de la veille. Ceci peut s'expliquer par la relative persistance des conditions météorologiques dans l'aire métropolitaine marseillaise.

Dans l'ensemble les feuilles de la branche de gauche présentent des niveaux faibles de l'ozone et celles de droite des niveaux plus importants. Les deux règles du deuxième niveau de l'arbre sont basées sur la température à 10 h (à gauche) et la vitesse du vent à 10 h (à droite). Le seuil pour la température est de 20,15°, seuil au dessous duquel le niveau d'ozone ne dépasse pas 122 μg/m³. Le seuil de la vitesse du vent est de 2.5 m/s, ce qui est la limite entre un vent faible et un vent modéré; pour des vitesses de vent au dessous de ce seuil, les valeurs de l'ozone sont assez importantes (branche gauche issue de ce nœud).

Le tableau d'alerte associé à cet arbre, les valeurs du threat score pour 130 μg/m³ et l'estimation par substitution puis par échantillon témoin (Rognac) de l'erreur quadratique sont donnés ci-dessous.

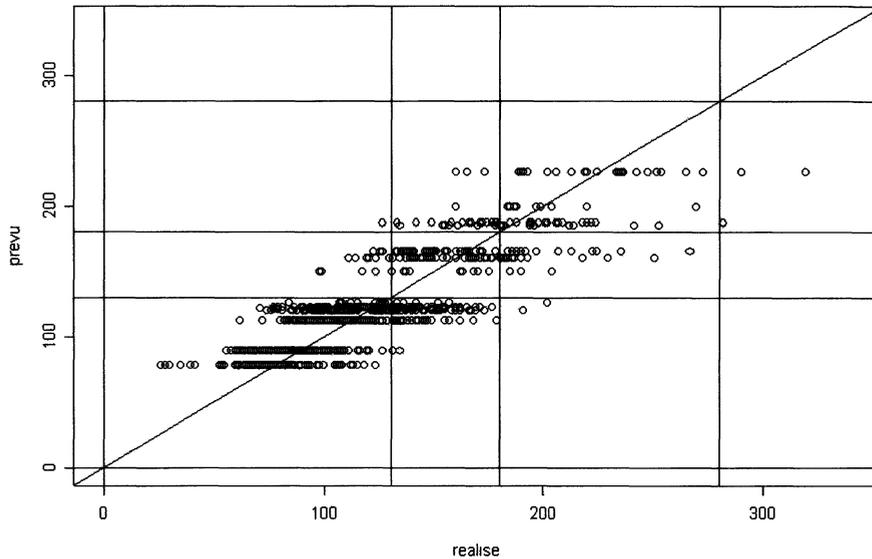


FIGURE 3
 Vitrolles, résultats de l'arbre de validation croisée : prévu * réalisé.

Le tableau 1 donne les résultats pour l'échantillon d'apprentissage et le tableau 2 celui de l'échantillon témoin (Rognac)

TABLEAU 1
 15 feuilles, $TS = 59.9\%$, $\hat{R} = 560.2$

485	16	1	0
116	81	27	0
2	30	61	0
0	0	3	0

TABLEAU 2
 15 feuilles, $TS = 52.5\%$, $\hat{R}^{et} = 776.7$

430	20	9	0
89	71	16	0
3	19	28	0
0	0	0	0

On peut noter que le threat score de niveau 130 est de l'ordre de 60% sur l'échantillon d'apprentissage et de 52% sur l'échantillon témoin.

3.3.2. Instabilité de l'arbre de régression

L'instabilité des méthodes CART a été remarquée et étudiée dans Breiman (1996a); nous allons la mettre en évidence dans le cadre de notre exemple.

Nous avons construit 30 échantillons bootstrap à partir de 822 observations de Vitrolles et sur chaque échantillon nous avons examiné les 30 arbres obtenus par

validation croisée. D'une part la taille de ces arbres est très variable; d'autre part les arbres sont différents à partir de la racine : Pour 15 de ces arbres la variable de partage est le maximum d'ozone de la veille, et pour les 15 autres c'est la température du jour j à 10 h.

Même si les spécialistes de météorologie et de la pollution indiquent que l'une ou l'autre de ces deux variables sont des variables explicatives essentielles, et acceptent l'une ou l'autre de ces premières coupures, l'examen des nœuds suivants accentue la mise en évidence de cette instabilité même si les coupures sont interprétables. L'idée de stabiliser les procédures semble être utile et fait l'objet de la suite de ce travail.

3.4 Mise en œuvre de la procédure d'agrégation par bootstrap

1. L'ensemble des données est divisé aléatoirement en deux parties inégales.
 - a) $N_1\%$ de cet ensemble constitue l'échantillon d'apprentissage $\mathcal{E} = (\mathbf{x}_n, y_n)$
 - b) $N_2\%$ de cet ensemble constitue l'échantillon témoin $\mathcal{T} = (\mathbf{x}_n, y_n)$.
 $N_1 + N_2 = 100$.
2. Un arbre de régression A est construit avec validation croisée à partir de \mathcal{E} . L'erreur de prévision est estimée à partir de l'échantillon témoin.

$$\hat{R}^{et}(A) = \frac{1}{|\mathcal{T}|} \sum_{(\mathbf{x}_n, y_n) \in \mathcal{T}} (y_n - \hat{y}(\mathbf{x}_n, \mathcal{E}))^2$$

3. On construit K échantillons bootstrap $(\mathcal{E}_B^{(k)})_{1 \leq k \leq K}$ à partir de \mathcal{E} et sur chacun de ces échantillons on construit un arbre $(A_k)_{1 \leq k \leq K}$ comme en 2). On calcule

$$\hat{y}^B(\mathbf{x}_i) = \frac{1}{K} \sum_{k=1}^K \hat{y}(\mathbf{x}_i, \mathcal{E}_B^{(k)})$$

les \mathbf{x}_i sont dans \mathcal{T} , $\hat{y}^B(\mathbf{x}_i)$ prédit le y_i correspondant, et

$$\hat{R}^B = \frac{1}{|\mathcal{T}|} \sum_{\mathcal{T}} (y_n - \hat{y}^B(\mathbf{x}_n))^2$$

est l'estimateur de l'erreur de prédiction par agrégation par bootstrap.

4. \mathcal{E} et \mathcal{T} sont reconstruits par exemple H fois en permettant la construction des H arbres $(A_h)_{1 \leq h \leq H}$. Sur les observations de chacun des couples échantillon d'apprentissage \mathcal{E} et échantillon témoin \mathcal{T} on calcule \hat{R}^{et} et \hat{R}^{B_i} à partir

desquels on calcule :

$$\overline{\hat{R}^{et}} = \frac{1}{H} \sum_{i=1}^H \hat{R}^{et}(A_i)$$

$$\overline{\hat{R}^B} = \frac{1}{H} \sum_{i=1}^H \hat{R}^B_i$$

Le paragraphe suivant montre les résultats obtenus pour une seule application de ces étapes sans l'itération décrite dans 4.

3.4.1 Un exemple d'application

Dans l'exemple présenté l'ensemble des données est celui de Vitrolles et l'échantillon d'apprentissage \mathcal{E} comportera toujours 622 observations et l'échantillon témoin \mathcal{T} 200. Ces divisions de l'ensemble des données seront réalisées aléatoirement et avec stratification. On génère $K=30$ échantillons bootstrap de \mathcal{E} . On construit \hat{y}^B et \hat{R}^B comme en 3.4. Les performances de \hat{y}^B sont résumées dans le Tableau 3.

On peut alors comparer le résultat à celui obtenu à partir d'un arbre construit par validation croisée sur le même échantillon \mathcal{E} . Les performances de cet arbre à 25 feuilles sont présentées dans le Tableau 4.

Une comparaison des performances sur ce seul exemple semble montrer l'intérêt de cette procédure d'agrégation. Les gains sont : 31% pour l'erreur quadratique (calculée sur l'échantillon témoin), et 3% pour le threat score.

TABLEAU 3
 $TS = 71.6\%$, $\hat{R}^B = 553.3$

112	9	0	0
16	29	5	0
0	11	17	0
0	0	1	0

TABLEAU 4
 $TS = 69.3\%$, $\hat{R}^{et} = 807.3$

112	8	1	0
18	25	7	0
0	14	14	0
0	0	1	0

3.4.2 Eléments de comparaisons

Pour vérifier la validité des résultats obtenus dans le paragraphe précédent, on procède de la manière suivante : on itère $H = 50$ fois la procédure de construction d'un prédicteur agrégé en changeant l'échantillon témoin (l'étape 4 décrite dans 3.4). A chaque itération on calcule l'erreur quadratique moyenne et le threat score, d'une part pour le prédicteur par validation croisée et d'autre part pour le prédicteur agrégé. La moyenne et l'écart type de ces deux indices (erreur quadratique et threat score)

obtenus sur ces 50 essais, sans agrégation (\widehat{R}^{et} et TS^{et}) et avec agrégation (\widehat{R}^B et TS^B), permettent de comparer les deux procédures.

Cet essai est réalisé sur cinq stations de mesure, Vitrolles (VTRL), Rognac (RBRT), Port de Bouc (PDBC), Rousset (ROUSS), et Sausset les Pins (SSLP). Les résultats sont présentés dans le tableau 5. Les gains relatifs présentés sont calculés par rapport à la moyenne (GAIN(%)).

TABLEAU 5

Comparaisons de l'erreur quadratique et du threat score moyens sans et avec agrégation par bootstrap sur cinq stations des Bouches du Rhône. Le gain montre l'évolution relative des deux indices. Il est positif pour l'erreur quadratique moyenne (diminution) et négatif pour le threat score (augmentation).

Station		\widehat{R}^{et}	\widehat{R}^B	TS^{et}	TS^B
VTRL	moyenne	905.6	561.8	57.8	67.7
	Ecart type	112.2	91.4	5.7	4.4
	GAIN(%)		38		- 17.1
RBRT	moyenne	831.5	522.6	49.8	60.2
	Ecart type	96.7	73.6	6	4.9
	GAIN(%)		37.1		- 20.9
ROUSS	moyenne	702.5	468.7	58.3	66.2
	Ecart type	80.3	64	4.6	4.1
	GAIN(%)		33.3		- 13.5
SSLP	moyenne	721.5	482.1	40.6	52.4
	Ecart type	81	71	5.8	5.6
	GAIN(%)		33.2		- 29
PDBC	moyenne	661.6	455.9	55.8	63.5
	Ecart type	98.3	67.5	4.8	4.3
	GAIN(%)		31.1		-13.7

Ce tableau montre que sur l'ensemble des essais réalisés, l'amélioration relative de la qualité de prévision par agrégation en terme de l'erreur quadratique moyenne est d'au moins 30%. L'augmentation relative du threat score est de 13% à 29% selon

les stations, ce qui signifie une meilleure discrimination entre les niveaux d'ozone sur l'ensemble des essais.

Le gain en stabilité par agrégation est mis en évidence par la comparaison des écarts types de ces deux indices.

4. Conclusion

Les avantages de l'utilisation des méthodes CART dans le contexte de la prévision du niveau d'ozone sont évidents : on dispose d'un outil facile à interpréter et à utiliser. Il s'agit d'un outil «descriptif» et de «décision», qui permet de retrouver des phénomènes liés à la pollution.

Une stabilisation du modèle a été examinée à l'aide de l'agrégation par bootstrap. La variance des résidus diminue et le threat score augmente. Le temps de calcul est légèrement augmenté, et on perd la structure d'arbre.

Un tel outil de prévision semble bien adapté aux problèmes de la prévision de l'ozone. De plus la variante classification de CART pour les réponses qualitatives peut être utilisée, avec les mêmes idées de stabilisation par agrégation par bootstrap.

Les résultats obtenus ici sont de bonne qualité. On doit cependant tenir compte que nous avons utilisé les données météorologiques observées du jour j au lieu des prévisions, non encore disponibles.

Remerciements

J'adresse mes remerciements à C. Deniau et G. Oppenheim pour la méthodologie, à C. Genevé, P. Renzi et D. Robin (AIRMARAIX), à D. Brion et B. Sol (Météo France) et à L. Carrié (ADEME), pour les applications

5. Références

- [1] BEL L., BELLANGER L., BONNEAU V., CIUPERCA G., DACUNHA-CASTELLE D., DENIAU C., GHATTAS B., MISITI M., MISITI Y., OPPENHEIM G., POGGI J.-M., TOMASSONE R. (1997), Prévisions des pointes de pollution dans la région parisienne, O₃ et NO₂ : Phase opérationnelle. *Rapport de contrat de recherche Airparif*, Oct., 193 p.
- [2] BREIMAN L., FRIEDMAN J.H., OLSHEN R., STONE C.J. (1984), *Classification And Regression Trees*, Wadsworth, Belmont CA.
- [3] BREIMAN L. (1996a), Heuristic of instability and stabilization in model selection, *The Annals of Statistics*, Vol 24, N°6, 2350–2383.
- [4] BREIMAN L. (1996b), Bagging Predictors, *Machine Learning*, 24, 123–140.
- [5] CHAUDHURI P. (1995), LO W.D., LOH W.Y., YANG C. C. Generalized regression trees, *Statistica Sinica* 5, 641–666.

- [6] CIAMPI A. (1991), Generalized regression trees, *Computational Statistics & Data Analysis* 12, 57–78.
- [7] DEVROYE L., GYÖRFI L., LUGOSI G. (1996) *A probabilistic theory of Pattern Recognition*, Springer Verlag.
- [8] EDER B.K., DAVIS J.M., BLOOMFIELD P. (1994) An automated classification scheme designed to better elucidate the dependence of ozone on meteorology. *Journal of applied Meteorology*, vol. 33, 1182–1198.
- [9] GHATTAS B., DENIAU C. (1998), Essai de classification des stations de mesures d’ozone des réseaux AIRMARAIX-AIRFOBEP, *contrat GREQAM-AIRMARAIX*.
- [10] GUEGUEN A., NAKACHE J.-P. (1988), Méthode de discrimination basée sur la construction d’un arbre binaire, *Revue de Statistique Appliquée* 36, 1, 19-37.
- [11] LOH W.Y., VANICHSETAKUL N. (1988), Tree-structured classification via generalized discriminant analysis, *Journal of the American Statistical Association*, Vol. 83, N°403, p. 715–728.
- [12] RYAN W.F. (1995) Forecasting severe ozone episodes in the Baltimore metropolitan area. *Atmospheric Environment*, 29 (17), 2387–2398.
- [13] SHEIFINGER H., STOHL A., KROMP-KOLB H., SPANGL W. (1996) A statistical method for predicting daily maximum ozone concentrations. *Gefahrstoffe-Reinhalung der Luft*, vol. 56, 133–137.
- [14] VENABLES W.N., RIPLEY B.D. (1994), *Modern Applied Statistics with S+*, Springer Verlag (Statistics and computing).
- [15] WILKS D.S. (1995) *Statistical Methods in Atmospheric Sciences : an introduction*. Academic Press.