MATHÉMATIQUES ET SCIENCES HUMAINES

BADIH GHATTAS

Prévision par arbres de classification

Mathématiques et sciences humaines, tome 146 (1999), p. 31-49

<http://www.numdam.org/item?id=MSH_1999__146__31_0>

Math. Inf. Sci. hum., (37e année, n° 146, *1999*, pp. 31-49)

PRÉVISION PAR ARBRES DE CLASSIFICATION

Badih GHATTAS[1]

RÉSUMÉ — *Après une présentation de la construction de prédicteurs par arbre de classification, nous nous intéressons à l'instabilité de cette méthode et proposons une méthodologie dans laquelle intervient le bootstrap. Une étude empirique détaillée illustre ce travail.*

MOTS-CLÉS — Arbres de classification, CART, agrégation de classificateurs, bootstrap, prévision, ozone.

SUMMARY — Prediction Using Classification Trees

Following the tree method classification, we focus on the instability of the method and suggest a technique where the bootstrap is used. A detailed empirical study is illustrated in this paper.

KEY WORDS — Classification trees, aggregating classifiers, bootstrap, bagging, prediction, ozone.

1. INTRODUCTION

Les méthodes de classification et de régression par arbres (CART) dues à Breiman *et al.* (1984) ont renouvelé la panoplie de méthodes dites de *segmentation* (Messenger *et al.*, 1972). Leur but est de partager un échantillon de données de façon progressive à l'aide de règles binaires et de visualiser les résultats à l'aide d'un arbre. L'innovation méthodologique produite par CART vient de l'algorithme de l'*élagage* des arbres conduisant à la réalisation d'arbres sans introduction de règles arbitraires d'arrêt de la procédure de construction. La méthode présentée dans le cadre de ce travail est celle de la *classification par arbre* et le domaine d'application est celui de la prévision des pics de pollution par l'ozone. L'intérêt de cette méthode est multiple. Elle est simple, permet d'utiliser des variables explicatives aussi bien quantitatives que qualitatives et ne limite pas leur nombre. De plus les arbres sont à la fois des outils de classification (but de l'opération) mais aussi de description grâce à l'étiquetage de leurs noeuds. Son inconvénient majeur, mis en évidence par Breiman (1996a) est l'*instabilité* traduite par le fait qu'une légère modification de l'échantillon d'apprentissage peut avoir un effet important sur l'arbre. Une méthode de stabilisation par agrégation d'arbres construits sur des *échantillons bootstrap* de

[1] GREQAM, 2 rue de la Charité, 13002 Marseille. e-mail : ghattas@chess.cnrs-mrs.fr

l'échantillon de référence proposée par Breiman (1996b) est présentée et mise en oeuvre.

L'étude empirique repose sur un travail contractuel réalisé pour l'association pour la surveillance de la qualité de l'air de l'Est des Bouches-du-Rhône, du Var et du Vaucluse (AIRMARAIX). Le travail de prévision concerne en fait seize stations du réseau. Nous concentrons notre attention sur une seule d'entre elles et il s'agira ici de prévoir le maximum d'ozone du jour j (situé en général l'après midi) à 06:00 UTC (i.e. 08:00h localement en été) à l'aide des variables météorologiques et de pollution. Les approches statistiques de ce problème de prévision en sont à leur début en France (Bel *et al.*, 1997) et sont nombreuses dans les pays anglo-saxons (Sheifinger *et al.* (1996), Ryan *et al.* et Burrows *et al.* (1994) par exemple).

Le plan de l'article est le suivant : après une présentation relativement détaillée de la méthode et celle d'un algorithme de stabilisation par agrégation, nous illustrons ces propos en présentant des résultats de prévisions obtenus sur la station de mesure de l'ozone située dans la ville de Vitrolles.

2. Classification et erreur de prévision

2.1. Généralités

Soit $\mathcal{E} = (\mathbf{x}_n, j_n)_{1 \leq n \leq N}$ un échantillon de taille N où les j_n sont les observations de la variable aléatoire expliquée Y à valeurs dans un ensemble $C = \{1, 2, ..., J\}$ et $\mathbf{x}_n = \{x_{n_1}, x_{n_2}, ..., x_{n_d}\}$ les observations d'une suite $\mathbf{X}$ de variables aléatoires explicatives à valeurs dans $\mathcal{X} = \prod_{i=1}^{d} \mathcal{X}_i$ où $\mathcal{X}_i$ est soit un ensemble sans structure (associé à une variable qualitative) soit un ensemble totalement ordonné (associé à une variable ordonnée) soit $\mathbb{R}$ (associé à une variable numérique).

Le problème ici est de construire à l'aide de $\mathcal{E}$, une procédure de classification notée $y(., \mathcal{E})$ définie sur $\mathcal{X}$ à valeurs dans C, qui va induire une partition $(B_j)_{j \in C}$ où $B_j = \{\mathbf{x} \in \mathcal{X} \ ; \ y(\mathbf{x}, \mathcal{E}) = j\}$. Elle permettra ensuite de procéder à des prévisions au sens où, ayant observé $\mathbf{x} \in X$, le calcul de $y(\mathbf{x}, \mathcal{E})$ affectera cette observation à une classe. Il s'agira alors de mesurer les performances de cette procédure.

Afin de préciser le concept de performance, introduisons sans la rigueur habituelle un modèle probabiliste sur $\mathcal{X} \times C$, où $P(B, j)$ est la probabilité que $\mathbf{X} \in B$, borélien de $\mathcal{X}$ et que $Y = j \in C$, on peut alors supposer que l'échantillon $\mathcal{E}$ est tiré suivant P.

2.2. Estimation de l'erreur de prévision

L'*erreur de prévision* ou *taux de fausse classification* est évaluée par la probabilité qu'une observation de $\mathcal{E}$, *échantillon d'apprentissage*, soit mal classée par $y(., \mathcal{E})$, soit :

$$\tau(y) = P(y(\mathbf{X}, \mathcal{E}) \neq Y) \tag{1}$$

L'estimation de cette erreur de prévision est à la base de la construction de la procédure de classification, du prédicteur associé et de ses performances. Plusieurs estimateurs sont proposés.

2.2.1 *Par substitution*

C'est la proportion des observations de $\mathcal{E}$ mal classées par $y(.,\mathcal{E})$, soit :

$$\widehat{\tau}(y) = \frac{1}{N} \sum_{(\mathbf{x}_n, j_n) \in \mathcal{E}} I\left(y(\mathbf{x}_n, \mathcal{E}) \neq j_n\right) \tag{2}$$

où I est la fonction indicatrice.

Les performances sont mesurées à l'aide de $\mathcal{E}$ à partir duquel la procédure de classification a été élaborée. Les performances de $y(.,\mathcal{E})$ seront surestimées.

2.2.2 *Par échantillon témoin*

Supposons que l'on dispose d'un second échantillon $\mathcal{T}$ (appelé *échantillon témoin*) de taille N'. Cet échantillon va être utilisé pour mesurer les performances de $y(.,\mathcal{E})$ construit à partir de $\mathcal{E}$.

$$\widehat{\tau}^{et}(y) = \frac{1}{N'} \sum_{(\mathbf{x}_n, j_n) \in \mathcal{T}} I\left(y(\mathbf{x}_n, \mathcal{E}) \neq j_n\right) \tag{3}$$

c'est la proportion des observations de $\mathcal{T}$ mal classées par $y(.,\mathcal{E})$.
La mesure des performances est moins optimiste, mais nécessite un plus grand volume de données.

2.2.3 *Par validation croisée*

Supposons $\mathcal{E}$ divisé aléatoirement en K sous ensembles $(\mathcal{E}_k)_{1 \leq k \leq K}$ disjoints de cardinaux voisins, et définissons K nouveaux échantillons d'apprentissage :

$$\mathcal{E}^k = \mathcal{E} - \mathcal{E}_k \qquad 1 \leq k \leq K$$

Sur chaque échantillon $\mathcal{E}^k$ construisons une procédure de classification

$$y^k(.) = y\left(., \mathcal{E}^k\right)$$

et mesurons la qualité de cette procédure à l'aide de $\mathcal{E}_k$ comme échantillon témoin

$$\widehat{\tau}^{vc}(y) = \frac{1}{N} \sum_{k=1}^{K} \sum_{(\mathbf{x}_n, j_n) \in \mathcal{E}_k} I\left(y\left(\mathbf{x}_n, \mathcal{E}^k\right) \neq j_n\right) \tag{4}$$

Cet estimateur limite aussi le biais d'estimation et n'est pas trop gourmand en données.

REMARQUE 1 On note que : $E\left(\widehat{\tau}(y)\right) = E\left(\widehat{\tau}^{vc}(y)\right) = E\left(\widehat{\tau}^{et}(y)\right) = \tau(y)$. Ce sont tous des estimateurs sans biais de $\tau(y)$.

REMARQUE 2 L'estimation de τ se ramène à celle de $P\left(y(\mathbf{X}) \neq Y\right)$, i.e. au paramètre d'une loi de Bernoulli , à l'aide de $\mathcal{E}$ échantillon de taille N. C'est à partir de l'estimation de P par substitution que va être réalisée une procédure initiale de classification.

3. Construction d'un arbre de classification

3.1. Quelques définitions et notations

- La procédure de classification $y(.,\mathcal{E})$ attribue à tout élément $\mathbf{x} \in \mathcal{X}$ une classe d'appartenance $j \in C$.
- La probabilité a priori de la classe j sera ici définie par $\pi_j = \frac{N_j}{N}$ où $N_j = card\{j_n \mid j_n = j\}$.
- Etant donné $t \subset \mathcal{X}$ on note $N(t)$ le cardinal de l'ensemble des $(\mathbf{x}_n, j_n) \in \mathcal{E}$ tels que $\mathbf{x}_n \in t$.
- De même $N_j(t)$ est le cardinal de l'ensemble des $(\mathbf{x}_n, j_n) \in \mathcal{E}$ tels que $j_n = j$ sachant que $\mathbf{x}_n \in t$.
- Un estimateur de $P(j,t)$, par substitution, noté $p(j,t)$ est défini[2] par : $p(j,t) = \pi_j \frac{N_j(t)}{N(t)}$.
- Un estimateur de $P(t)$, par substitution, noté $p(t)$ est défini par : $p(t) = \sum_{j=1}^{J} p(j,t)$.
- Enfin $\mathrm{P}(j \mid t)$, probabilité a posteriori dans t de la classe j, est estimé par $\frac{p(j,t)}{p(t)}$ qui, avec la définition de π_j est égal à $\frac{N_j(t)}{N(t)}$.

3.2. Réalisation d'une procédure initiale

3.2.1 *Objectif*

Il s'agit de construire une partition de $\mathcal{X}$ en q classes, *où q n'est pas fixé à priori.* La méthode proposée consiste à construire une suite croissante de partitions de $\mathcal{X}$, le passage d'une partition à la suivante étant obtenu par l'utilisation de *règles de division binaire* du type *:*

$$\mathbf{x} \in t \ ? \quad \text{pour } t \subset \mathcal{X}.$$

Précisons ces règles ; elles ne dépendent que d'*une seule variable* x_l , $1 \leq l \leq d$ et d'un *seuil* concernant cette variable :

- $x_l \leq \mu$, $\mu \in \mathbb{R}$, pour une variable quantitative ou totalement ordonnée. Dans ce cas si M valeurs distinctes de x_l sont observées, le cardinal de l'ensemble D des divisions possibles est égal à $M-1$.
- $x_l \in \mu$ où μ est un sous-ensemble de $\{\mu_1, \mu_2, ...\mu_M\}$ où les μ_m sont les différentes modalités d'une variable qualitative. Dans le cas d'une variable qualitative, l'ensemble D des divisions binaires possibles est de cardinal : $2^{M-1}-1$.

Dans cette procédure de construction, au départ, on dispose de l'ensemble $\mathcal{X}$ appelé *racine*, qui à la première étape de l'algorithme est divisé en deux sous ensembles disjoints non vides, ses *descendants* notés t_1 et t_2, de réunion $\mathcal{X}$, appelés *noeuds.* Chacun des noeuds t_1 et t_2,s'il a au moins deux éléments, est à son tour divisé comme précédemment et l'on obtient une nouvelle partition de $\mathcal{X}$ en classes (ou

[2] Par la suite p sera systématiquement l'estimateur de P.

noeuds) t_3, t_4, t_5, t_6.....à la fin de la procédure on dispose d'une partition de $\mathcal{X}$ (la plus fine) en q classes appelées *feuilles* ou *noeuds terminaux*. Cette suite croissante de partitions obtenue par divisions binaires successives peut-être visualisée par un arbre binaire appelé *arbre de classification* et noté A. On note $\widetilde{A}$ l'ensemble des feuilles de A et A^t l'ensemble de tous les descendants du noeud t de l'arbre A. Nous noterons à partir de maintenant $\boldsymbol{\tau}(A)$, $\widehat{\boldsymbol{\tau}}(A)$, $\widehat{\boldsymbol{\tau}}^{et}(A)$, $\widehat{\boldsymbol{\tau}}^{vc}(A)$, remplaçant y par A pour les erreurs de prévision et leurs estimations.

3.2.2 *Critère d'hétérogénéité*

La qualité d'une règle de division binaire sera mesurée à l'aide d'un critère dit d'*hétérogénéité.*

DEFINITION 1 Une fonction d'hétérogénéité est une fonction réelle h définie sur un ensemble de probabilités discrètes, sur un ensemble fini :

$$h : (p_1, p_2, ..., p_J) \rightarrow h(p_1, p_2, ..., p_J)$$

symétrique en $p_1, p_2, ..., p_J$ et telle que :

1. Le maximum de h est atteint pour l'équiprobabilité :
$$\arg\max h(p_1, p_2, ..., p_J) = \left(\frac{1}{J}, \frac{1}{J}, ..., \frac{1}{J}\right)$$
2. Le minimum de h est atteint pour les "dirac" :
$$\arg\min h(p_1, p_2, ..., p_J) \in \{e_1, e_2, ..., e_J\}$$
où e_j est le $j^{ème}$ élément de la base canonique de $\mathbb{R}^J$. Cette fonction d'hétérogénéité est introduite comme outil de mesure de la qualité de la division d'un noeud t en ses descendants t_g et t_d.

L'hétérogénéité d'un noeud t d'un arbre A est définie par :

$$hét(t) = h(p(1 \mid t), p(2 \mid t), ..., p(J \mid t))$$

où $p(j \mid t)$ est la probabilité d'être dans la classe j sachant qu'on est dans le noeud t.
On définit de même l'hétérogénéité d'un arbre A (ici il s'agit d'un arbre en cours de développement) par :

$$Hét(A) = \sum_{t \in \widetilde{A}} Hét(t) \quad \text{où} \quad Hét(t) = p(t) hét(t) \tag{5}$$

Une règle de division δ d'un noeud t conduit à une proportion $p_g = \frac{p(t_g)}{p(t)}$ des observations situées dans t dans t_g et $p_d = \frac{p(t_d)}{p(t)}$ dans t_d. On définit de même la variation d'hétérogénéité entraînée par une règle de division δ par :

$$\partial hét(\delta, t) = hét(t) - p_g hét(t_g) - p_d hét(t_d) \tag{6}$$

Cette règle de division conduit-elle a une décroissance de l'hétérogénéité, i.e. à :

$$\partial hét(\delta, t) \geq 0 \ ?$$

Si la fonction h est strictement concave, la réponse est oui.

Deux fonctions d'hétérogénéité, possédant cette propriété sont traditionnellement utilisées (pour leur comparaison, voir Breiman (1996c)) :

$$hét(t) = -\sum_{j=1}^{J} p(j \mid t) \ln(p(j \mid t)) \text{ (dérivée de l'entropie de Shannon)} \quad (7)$$

$$hét(t) = \sum_{j \neq k} p(j \mid t) p(k \mid t) \text{ (dérivée de l'indice d'inégalité de Gini)} \quad (8)$$

3.2.3 *Division d'un noeud t*

La division optimale d'un noeud t quelconque, notée δ_t^*, est définie par :

$$\delta_t^* = \arg \max_{\delta \in D} \partial hét(\delta, t) \quad (9)$$

$\arg \max_{\delta \in D} \partial hét(\delta, t)$ désigne la valeur de δ qui rend maximum $\partial hét(\delta, t)$.
Soit $t \in \tilde{A}$ divisé par la règle δ en (t_g, t_d) ; on obtient un arbre A' plus fin que l'arbre A et :

$$Hét(A') = \sum_{s \in \tilde{A} - \{t\}} Hét(s) + Hét(t_g) + Hét(t_d)$$

La variation de l'hétérogénéité de l'arbre A due à la division du noeud t est donnée par :

$$\begin{aligned} Hét(A) - Hét(A') &= Hét(t) - Hét(t_g) - Hét(t_d) \\ &= \partial Hét(\delta, t) \\ &= p(t) \partial hét(\delta, t) \end{aligned}$$

Donc maximiser la décroissance de l'hétérogénéité par division du noeud t équivaut à maximiser la décroissance de l'hétérogénéité de l'arbre A. A chaque étape de la procédure, un noeud donne naissance à deux descendants de telle sorte que la diminution d'hétérogénéité, lors du passage de l'arbre A à l'arbre A' soit maximale.

La figure figure 1 donne un exemple d'un arbre de classification : les variables ayant servi à sa construction sont décrites dans le paragraphe 5.1. La variable expliquée est le maximum quotidien de l'ozone observé à Vitrolles ; elle a été codée en variable qualitative à 4 modalités par rapport aux seuils 0, 130, 180 et 280 $\mu g/m^3$. Dans les noeuds terminaux (les feuilles) on indique la classe majoritaire du noeud.

La première règle (à la racine) est basée sur la température à 10h du matin ("tc10") avec le seuil de 20.35°. Les deux règles suivantes sont basées respectivement sur le maximum d'ozone observé la veille ("maxo3v") et la vitesse du vent à 10 heures ("vv10"). On remarque que dans la branche de gauche de l'arbre on a des niveaux bas de l'ozone (1 ou 2), alors que les niveaux élevés (3) se trouvent dans les feuilles de la branche de droite.

Il s'agit d'un arbre dont la taille optimale a été obtenue par 30 validations croisées (la procédure est détaillé au paragraphe 3.5) sur des échantillons stratifiés[3]. Pour faciliter l'examen complet de l'arbre de la figure figure 1 voici les symboles utilisés pour noter les variables : l'insolation par "ins", la nébulosité par "neb", la direction du vent par "dv" et le gradient thermique par "grad".

[3] La distribution des observations de la variable expliquée dans les échantillons témoins utilisés est très voisine de celle de l'échantillon de construction.

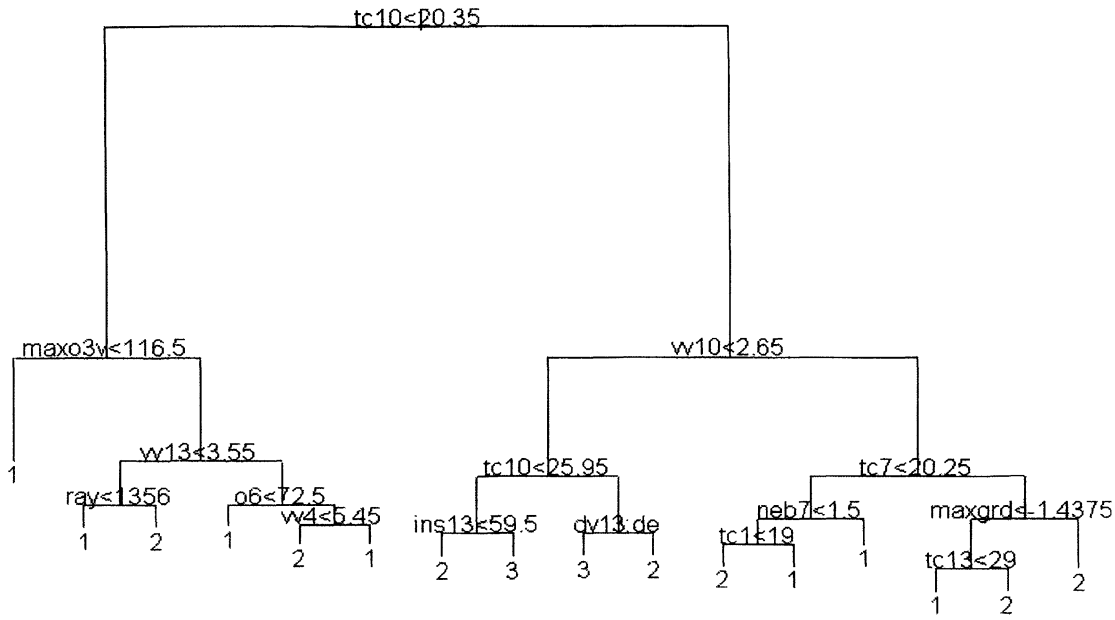

Figure 1 : Arbre de classification de taille optimale. Variable expliquée : l'ozone (qualitative à 4 modalités).

3.3. Règles d'arrêt de la procédure de construction

La construction de partitions emboîtées s'arrête naturellement lorsque chaque classe contient un seul élément ou plusieurs auxquels sont associées les mêmes observations de la variable composite $\mathbf{X}$. D'autres règles d'arrêt sont possibles :

- lorsque pour tout noeud : $\max_{\delta \in D} \partial Hét(\delta, t) < \delta$, $\delta \in \mathbb{R}_+^*$.
- lorsque le nombre d'observations de chaque noeud est inférieur à $\nu \in \mathbb{N}^*$.

Si δ et ν (*arbitrairement* choisis) sont petits, le nombre de feuilles sera grand ; il sera petit dans le cas contraire.
L'arbre ainsi obtenu est noté $A_{\max}$. *La règle d'affectation d'une classe à un noeud terminal* t est alors la suivante :

$$y(\mathbf{x}, \mathcal{E}) = i_0(t) \text{ où } i_0(t) = \arg \max_{j \in C} p(j \mid t)$$

C'est celle qui maximise la probabilité a posteriori, dans le noeud t, de la classe j ; on l'appelle règle de Bayes. L'estimation par substitution du taux de fausse classification d'une observation tombant dans un noeud t est :

$$r(t) = 1 - \max_{j \in C} p(j \mid t) \tag{10}$$

et le taux de fausse classification global, estimé par substitution, prend (si l'on pose : $\widehat{\tau}(t) = r(t)\, p(t)$) la forme :

$$\widehat{\tau}(A) = \sum_{t \in \tilde{A}} \widehat{\tau}(t)$$

L'arbre $A_{\max}$ présente un double inconvénient : selon la règle d'arrêt utilisée, il peut avoir un grand nombre de feuilles ; dans ce cas il est illisible et inutilisable dans la pratique. D'autre part la performance d'un tel arbre est maximale sur l'échantillon d'apprentissage mais est très mauvaise sur un échantillon témoin.

Finalement Breiman *et al.* (1984) conjuguent validation croisée (définie en 2.2.3 ci-dessus) et procédure d'*élagage* (qui sera définie en 3.5 ci-dessous) des arbres afin de proposer une solution satisfaisante à ce problème.

3.4. Introduction de coûts de mauvais classement

Il est parfois réaliste de pénaliser par des coûts et à divers degrés une fausse classification.

DEFINITION 2 Un coût de mauvais classement est une fonction $\Gamma : C \times C \rightarrow \mathbb{R}_+$ telle que :

$$\Gamma(i \mid j) \geq 0 \quad et \quad \Gamma(i \mid i) = 0 \tag{11}$$

DEFINITION 3 ($\Gamma(i \mid j)$ est le coût de prévoir dans la classe i une observation appartenant à la classe j).

Le modèle probabiliste permet de définir un coût moyen de fausse classification. Posons :

$$P_y(i \mid j) = P(y(\mathbf{x},\mathcal{E}) = i \mid Y = j) \tag{12}$$

(c'est la probabilité qu'une observation de la $j^{ème}$ classe soit rangée par $y(.,\mathcal{E})$ dans la $i^{ème}$), et

$$\tau_y(j) = \sum_i \Gamma(i \mid j) P_y(i \mid j) \tag{13}$$

(C'est le coût moyen de fausse classification pour la classe j).

$$\tau_y = \tau(A) = \sum_j \pi(j)\tau_y(j) \tag{14}$$

$$= \frac{1}{N} \sum_j N_j \tau_y(j) \tag{15}$$

(C'est le coût moyen de fausse classification par $y(.,\mathcal{E})$ ou par l'arbre A ; on notera maintenant P_A et τ_A au lieu de P_y et τ_y).

Dans ce contexte la règle d'affectation d'une observation tombant dans un noeud terminal t à une classe est :

$$y(\mathbf{x},\mathcal{E}) = i_0(t)$$

où

$$i_0(t) = \arg\min_{i \in C} \sum_{j \in C} \Gamma(i \mid j) p(j \mid t) \tag{16}$$

L'estimation par substitution du taux de fausse classification dans un noeud t est :

$$r(t) = \min_{i \in C} \sum_{j \in C} \Gamma(i \mid j)\, p(j \mid t) \tag{17}$$

(Dans le cas où les coûts sont unitaires, i.e. pour $i \neq j$, $\Gamma(i \mid j) = 1$ alors $\sum_{j \in C} \Gamma(i \mid j)\, p(j \mid t) = 1 - p(i \mid t)$ et on retrouve le résultat obtenu en l'absence de fonction de coût).

Posons $\boldsymbol{\tau}(t) = r(t)p(t)$ ce qui conduit à un taux global de fausse classification par substitution pour l'arbre A :

$$\widehat{\boldsymbol{\tau}}(A) = \sum_{t \in \widetilde{A}} \widehat{\boldsymbol{\tau}}(t) \tag{18}$$

REMARQUE 3 Une propriété importante de $\widehat{\boldsymbol{\tau}}$ est que si A' est un arbre obtenu a partir de A par partage d'un quelconque noeud terminal alors

$$\widehat{\boldsymbol{\tau}}(A') \leq \widehat{\boldsymbol{\tau}}(A) \tag{19}$$

En effet :

$$\begin{aligned} \widehat{\boldsymbol{\tau}}(t) &= \sum_{j \in C} \Gamma(i_0(t) \mid j)\, p(j,t) \\ &= \sum_{j \in C} \Gamma(i_0(t) \mid j)\, [p(j,t_d) + p(j,t_g)] \end{aligned}$$

or :

$$\begin{aligned} \widehat{\boldsymbol{\tau}}(t) - \widehat{\boldsymbol{\tau}}(t_g) - \widehat{\boldsymbol{\tau}}(t_d) &= \sum_{j \in C} \Gamma(i_0(t) \mid j)\, p(j,t_d) - \min_i \sum_{j \in C} \Gamma(i \mid j)\, p(j,t_d) \\ &\quad + \sum_{j \in C} \Gamma(i_0(t) \mid j)\, p(j,t_g) - \min_i \sum_{j \in C} \Gamma(i \mid j)\, p(j,t_g) \end{aligned} \tag{20}$$

qui est non négatif, et nul si et seulement si : $i_0(t) = i_0(t_d) = i_0(t_g)$.

REMARQUE 4 La propriété

$$\widehat{\boldsymbol{\tau}}(t) - \widehat{\boldsymbol{\tau}}(t_g) - \widehat{\boldsymbol{\tau}}(t_d) \geq 0 \tag{21}$$

demeure vraie en l'absence de fonction de coût. Il en est de même de (19).

3.5. L'ÉLAGAGE

3.5.1 *Principes et définitions*

- *Elaguer* un arbre A à partir d'un noeud t consiste à supprimer de A l'ensemble A^t (la branche de A issue de t). D'après (21) on déduit immédiatement que :

$$\widehat{\tau}(t) \geq \widehat{\tau}(A^t) \tag{22}$$

On note $A - A^t$ l'arbre élagué au noeud 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 $\widehat{\tau}(A) \leq \widehat{\tau}(A')$. Une idée consiste alors à *pénaliser* le fait d'avoir un trop grand nombre de feuilles par l'introduction d'un *nouveau critère*, soit par exemple :

$$\widehat{\tau}_\alpha(A) = \widehat{\tau}(A) + \alpha Card(\tilde{A}) \tag{23}$$

où $\alpha \geq 0$ et $Card(\tilde{A})$ est le nombre de feuilles de l'arbre A. Le terme $\alpha Card(\tilde{A})$ peut être interprété comme un coût de complexité et l'erreur pénalisée $\widehat{\tau}_\alpha(A)$ est à rapprocher des critères d'information introduits par Akaïke (1974).
On remarque que :
- Si $\alpha = 0$ alors $\widehat{\tau}_0(A) = \widehat{\tau}(A)$.
- Si t est un noeud alors $\widehat{\tau}_\alpha(t) = \widehat{\tau}(t) + \alpha$.

3.5.2 *L'algorithme d'élagage*

Posons $\alpha = 0$. Soit A_0 le plus petit sous arbre de $A_{\max}$ tel que $\widehat{\tau}(A_0) = \widehat{\tau}(A_{\max})$; l'arbre A_0 est obtenu par l'élagage de $A_{\max}$ aux différents noeuds t tel que :

$$\widehat{\tau}(t) = \widehat{\tau}(t_g) + \widehat{\tau}(t_d)$$

D'après (22) et par définition de A_0, pour tout noeud t de A_0, $\widehat{\tau}(t) > \widehat{\tau}(A_0^t)$.
Alors pour tout noeud t de A_0 l'inégalité $\widehat{\tau}_\alpha(t) > \widehat{\tau}_\alpha(A_0^t)$ est vraie pour $\alpha = 0$ et demeure vraie tant que α reste inférieur à la valeur :

$$c(t, A_0^t) = \frac{\widehat{\tau}(t) - \widehat{\tau}(A_0^t)}{Card(\tilde{A}_0^t) - 1} \tag{24}$$

Quand α atteint cette valeur $\widehat{\tau}_\alpha(t) = \widehat{\tau}_\alpha(A_0^t)$.
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 noeuds 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

$$\widehat{\tau}(A_0 - A_0^t) - \widehat{\tau}(A_0)$$

est " petite ". Ce qui réalise le compromis souhaité lors de l'introduction du critère $\widehat{\tau}_\alpha$.
On procède de même sur A_1 comme sur A_0,on détermine $\alpha_2 > \alpha_1$ (par construction) tel que :

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

et on élague A_1 à tous les noeuds 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 $A_0 > A_1 >_{...} > A_L$ où A_L est la racine de l'arbre A_0, et une suite croissante $(\alpha_l)_{1 \leq l \leq L}$ de coefficients de pénalisations.
De plus, par construction si $\alpha \in [\alpha_l, \alpha_{l+1}[$, pour tout noeud t de A_l :

$$\widehat{\tau}_\alpha(t) > \widehat{\tau}_\alpha(A_l^t)$$

Nous sommes ramenés à déterminer dans la suite (A_l) un arbre optimal, vis à vis du critère de taux de fausse classification.

3.5.3 *Choix d'un arbre optimal*

3.5.3.1 *Par échantillon témoin*

La règle utilisée est ici :

$$\widehat{\tau}^{et}(A_{l_o}) = \min_{1 \leq l \leq L} \widehat{\tau}^{et}(A_l) \tag{25}$$

où $\widehat{\tau}^{et}(A)$ est un estimateur par échantillon témoin de $\tau(A)$.

On rappelle que l'échantillon témoin $\mathcal{T}$ est de taille N' et on note N'_j le nombre d'observations de $\mathcal{T}$ dans la classe j et N'_{ij} celui des observations de la classe j rangées en i par l'arbre A. Le taux de fausse classification de l'arbre A est donné par :

$$\widehat{\tau}^{et}(j) = \sum_i \Gamma(i \mid j) \frac{N'_{ij}}{N'_j}$$

$$\widehat{\tau}^{et}(A) = \frac{1}{N'} \sum_{i,j} \Gamma(i \mid j) N'_{ij} \tag{26}$$

Dans le cas de coûts unitaires, $\widehat{\tau}^{et}(j)$ est la proportion des éléments de la classe j, dans $\mathcal{T}$, qui sont mal classés et $\widehat{\tau}^{et}(A)$ la proportion des éléments de $\mathcal{T}$ mal classés par A.

3.5.3.2 *Par validation croisée*

Il s'agit d'estimer (14) par validation croisée. Supposons comme en 2.2.3 l'échantillon d'apprentissage $\mathcal{E}$ divisé aléatoirement en K sous ensembles $(\mathcal{E}_k)_{1 \leq k \leq K}$ disjoints de cardinaux voisins, et définissons K nouveaux échantillons d'apprentissage :

$$\mathcal{E}^k = \mathcal{E} - \mathcal{E}_k \qquad 1 \leq k \leq K$$

Si K est assez grand $\mathcal{E}^k$ contient une très grande part des éléments de $\mathcal{E}$.

A l'aide de l'algorithme de l'élagage (3.5.2) construisons sur $\mathcal{E}$ (resp. sur chaque $\mathcal{E}^k$) une suite d'arbres $A_l = A_l(\alpha_l)$ (resp. $A_l^k = A_l^k(\alpha_l^k)$) de coût pénalisé minimum avec $1 \leq l \leq L$. Les éléments de $\mathcal{E}$ situés dans $\mathcal{E}_k$ n'ont pas contribué à la construction de $\left(A_l^k\right)_l$. Donc $\mathcal{E}_k$ peut être utilisé comme échantillon témoin pour évaluer les arbres A_l^k.

Notons N_j (resp. N_j^k) le nombre d'éléments de $\mathcal{E}_k$ (resp. de $\mathcal{E}_k$) appartenant à la classe j, et $N_{ij}(\alpha)$ (resp. $N_{ij}^k(\alpha)$) celui des éléments de la classe j rangés dans la classe i par $A(\alpha)$ (resp. $A^k(\alpha)$). On remarque que $N_{ij}(\alpha) = \sum_k N_{ij}^k(\alpha)$. La probabilité qu'une observation située dans la classe j soit rangée dans la classe i par $A(\alpha)$ (resp. $A^k(\alpha)$) est notée $P_{A(\alpha)}(i \mid j)$ (resp. $P_{A^k(\alpha)}(i \mid j)$).

Une estimation de $P_{A^k(\alpha)}(i \mid j)$ est donnée par $p_{A^k(\alpha)}(i \mid j) = \frac{N_{ij}^k(\alpha)}{N_j^k}$ et celle de $P_{A(\alpha)}(i \mid j)$ par validation croisée est donnée par :

$$p_{A(\alpha)}^{vc}(i \mid j) = \sum_{k=1}^{K} \frac{N_j^k}{N_j} p_{A^k(\alpha)}(i \mid j) = \frac{N_{ij}(\alpha)}{N_j}$$

où $N_{ij}(\alpha) = \sum_k N_{ij}^k(\alpha)$.

L'estimation par validation croisée du taux de fausse classification d'un élément d'une classe j par $A(\alpha)$ est :

$$\widehat{\tau}_j^{vc}(A(\alpha)) = \sum_i \Gamma(i \mid j)\, p_{A(\alpha)}^{vc}(i \mid j)$$

et celle recherchée, du taux de fausse classification associé à $A(\alpha)$:

$$\widehat{\tau}^{vc}(A(\alpha)) = \frac{1}{N} \sum_j \sum_i \Gamma(i \mid j)\, N_{ij}(\alpha) \tag{27}$$

(Dans le cas où $\Gamma(i \mid j) = 1$ pour $i \neq j$ on trouve la somme des proportions des observations mal classées de $\mathcal{E}$ vu comme réunion disjointe des $\mathcal{E}_k$, par les $A^k(\alpha)$).

Pour $1 \leq l \leq L$ posons $\alpha_l' = \sqrt[2]{\alpha_l \alpha_{l+1}}$ la moyenne géométrique[4] de α_l et α_{l+1}, c'est-à-dire une valeur de α comprise entre α_l et α_{l+1} ; alors l'erreur de prévision de l'arbre $A(\alpha_l)$ estimée par validation croisée est par définition :

$$\widehat{\tau}^{vc}(A_l) = \widehat{\tau}^{vc}(A(\alpha_l')) \tag{28}$$

où le membre de droite est défini en (27). Donc $\widehat{\tau}^{vc}(A_l)$ est l'estimateur obtenu en utilisant chaque $\mathcal{E}_k$ comme échantillon témoin pour chaque $A^k(\alpha_l')$.

La règle de sélection du meilleur arbre de cette suite est :

$$\widehat{\tau}^{vc}(A_{l_0}) = \min_{1 \leq l \leq L} \widehat{\tau}^{vc}(A_l) \tag{29}$$

A_{l_0} est cet arbre et l'erreur de prévision est : $\widehat{\tau}^{vc}(A_{l_0})$.

[4] On aurait tout aussi bien pu prendre la moyenne arithmétique (ou toute autre "valeur centrale") et poser $\alpha_l' = \frac{\alpha_l + \alpha_{l+1}}{2}$

4. INSTABILITÉ ET AGRÉGATION PAR BOOTSTRAP

4.1. INSTABILITÉ DE L'ARBRE DE CLASSIFICATION

L'instabilité des méthodes de classification par arbre a été remarquée et étudiée dans Breiman (1996a) ; nous allons la mettre en évidence dans le cadre de notre étude empirique.

Nous avons construit 10 échantillons bootstrap (c'est-à-dire 10 échantillons de 822 observations obtenus par tirage *avec remise* dans l'échantillon de base, ayant lui aussi 822 observations) à partir de 822 observations de Vitrolles et sur chaque échantillon nous avons examiné les 10 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 : le tableau ci-dessous indique pour les 10 arbres (en colonne) le nom des variables apparaissant au niveau des trois premiers noeuds (en ligne) de chaque arbre.

" tcl0"	"maxo3v"	"maxo3v"	" tcl0"	" tcl0"	"maxo3v"	"maxo3v"	"tcl0"	"maxo3v"	" tcl0"
" neb10"	" tcl0"	" tcl0"	"maxo3v"	"maxo3v"	" tcl3"	" dv16"	"vvl3"	" tcl0"	"maxo3v"
" dvl3"	" vvl3"	" vv10"	" dv16"	" vv10"	" vvl3"	" tc7"	" ray"	" dv10"	" vv10"

Tableau 1 : Les variables figurant aux trois premiers noeuds de 10 arbres construits sur des échantillons bootstrap

Deux variables apparaissent au niveau de la racine : "tcl0" (température à 10 heures) pour 5 arbres et "maxo3v" (maximum d'ozone de la veille) pour les cinq autres. Le nombre de variables différentes augmente avec la profondeur de l'arbre. L'examen des noeuds suivants accentue la mise en évidence de cette instabilité même si les coupures sont interprétables.

Cette instabilité se manifeste aussi au niveau des prévisions données par ces arbres. Pour montrer cela, on effectue la prévision sur l'ensemble des observations de l'échantillon de construction avec les 10 arbres construits. Le tableau ci-dessous donne la discordance[5] entre les prévisions faites par les différents arbres.

La discordance entre les prévisions données par deux arbres construits à partir de deux échantillons bootstrap peut atteindre les 37% ; elle est cependant liée à l'importance de la différence entre les distributions des échantillons bootstrap.

4.2. UN ALGORITHME D'AGRÉGATION PAR BOOTSTRAP

L'idée ici est d'engendrer K échantillons bootstrap à partir de l'échantillon de construction de base, de construire un prédicteur (arbre de classification) optimal $(\hat{y}_k(,\mathcal{E}))$ sur chaque échantillon, et d'agréger ces prédicteurs ; chaque prédicteur ayant attribué une classe à chaque observation, la classe finale affectée par le prédicteur agrégé est la classe majoritairement attribuée par les K prédicteurs.

Afin de mettre en oeuvre cette idée et de la tester sur plusieurs échantillons on procède de la manière suivante :

1. L'ensemble des données est divisé aléatoirement en deux parties inégales.

[5] C'est la proportion d'observations pour lesquelles les prévisions sont différentes.

0	24.8	27.3	24.7	22.8	28.1	26.3	31.6	31.4	20.3
	0	21.4	29.4	28.8	31.4	14.6	32.4	23.8	23.7
		0	23.8	24.2	33.9	27.4	33.6	21.4	22.1
			0	6.6	29.7	26.3	36.6	24.8	14.6
				0	27.5	24.7	33.8	27.5	12.8
					0	28.4	15.8	31.4	20.2
						0	31.8	25.9	21.1
							0	32.9	25.6
								0	24.3
									0

Tableau 2 : Discordances exprimées en pourcentage, entre les prévisions données par les 10 arbres, pris deux à deux, construits sur des échantillons bootstrap

(a) $N_1\%$ de cet ensemble constitue l'échantillon d'apprentissage $\mathcal{E} = (\mathbf{x}_n, j_n)$
(b) $N_2\%$ de cet ensemble constitue l'échantillon témoin $\mathcal{T} = (\mathbf{x}_n, j_n)$. $N_1 + N_2 = 100$.

2. Un arbre de classification A est construit avec validation croisée à partir de $\mathcal{E}$. Le taux de fausse classification est estimé à partir de l'échantillon témoin.

$$\widehat{\tau}^{et}(A) = \frac{1}{Card\,(\mathcal{T})} \sum_{(\mathbf{x}_n, j_n) \in \mathcal{T}} I\left(j_n \neq \hat{y}(\mathbf{x}_n, \mathcal{E})\right)$$

3. On construit K échantillons bootstrap $\left(\mathcal{E}_B^k\right)_{1 \leq k \leq K}$ 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

$$\hat{y}^B(\mathbf{x}_n) = \arg\max_j Card\left\{\hat{y}(\mathbf{x}_n, \mathcal{E}_B^k) = j\right\}$$

les $\mathbf{x}_n$ sont dans $\mathcal{T}$, $\hat{y}^B(\mathbf{x}_n)$ prédit la classe i correspondante, et

$$\widehat{\tau}^B = \frac{1}{Card\,(\mathcal{T})} \sum_{(\mathbf{x}_n, j_n) \in \mathcal{T}} I\left(j_n \neq \hat{y}^B(\mathbf{x}_n, \mathcal{E})\right)$$

est l'estimateur du taux de fausse classification 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 $\widehat{\tau}^{et}$ et $\widehat{\tau}^{B_i}$ à partir desquels on calcule :

$$\overline{\widehat{\tau}^{et}} = \frac{1}{H} \sum_{h=1}^{H} \widehat{\tau}^{et}(A_h) \quad \text{et} \quad \overline{\widehat{\tau}^{B}} = \frac{1}{H} \sum_{h=1}^{H} \widehat{\tau}^{B_h}$$

4.3. Nombre d'échantillons bootstrap à utiliser

Pour justifier le nombre d'échantillons bootstrap utilisés dans l'étude empirique décrite ici, nous avons réalisé l'expérience suivante : construction de 1000 échantillons bootstrap et des 1000 prédicteurs construits par agrégations successives. Les résultats numériques obtenus et visualisés sur la figure figure 2, semblent indiquer que l'agrégation d'environ 400 arbres permet d'obtenir une réduction de la qualité de prévision constante. Dans notre application, la qualité des résultats obtenus avec cinquante arbres nous paraît satisfaisante.

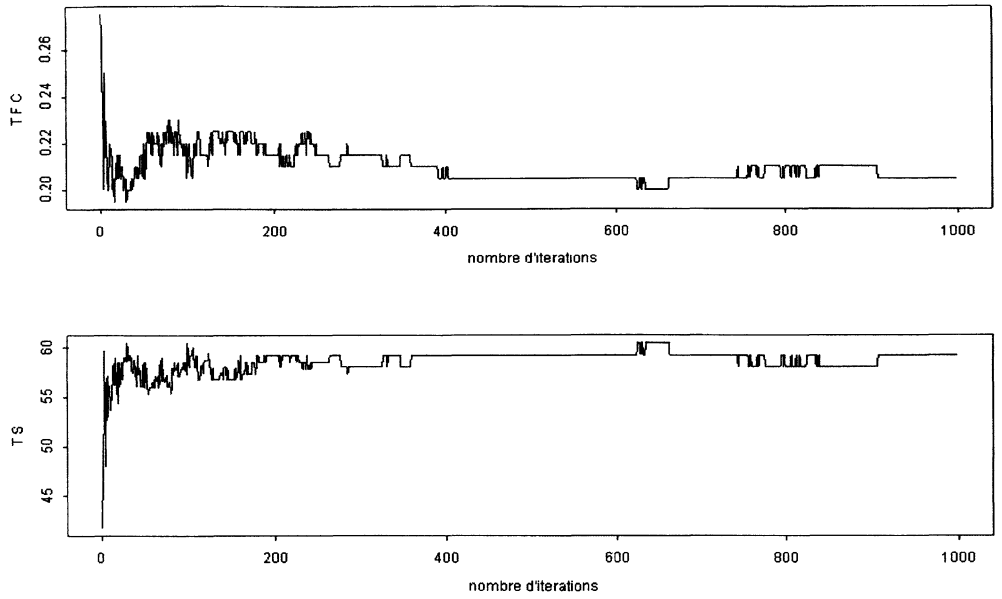

Figure 2 : Evolution de la qualité de prévision en fonction du nombre d'arbres agrégés. T.F.C.=Taux de Fausse Classification, T.S.=Threat Score (cf. paragraphe 5.2 ci-dessous).

5. PRÉDICTION DU MAXIMUM D'OZONE

5.1. DONNÉES

La variable expliquée est le maximum d'ozone du jour j notée "*maxO3*" ; c'est celle que l'on cherche à prévoir. Les périodes d'observations et de prévisions sont estivales ; plus précisément du 01 Avril au 30 Septembre, de 1992 à 1997. La station de mesure est située dans la ville de Vitrolles.
On dispose de 822 observations et de 41 variables explicatives :

- d'une part des variables concernant la *pollution* : le maximum d'ozone du jour $j-1$, l'ozone du jour j à 1h, 4h, 6h, le maximum des observations du dioxyde de soufre durant la nuit, le maximum du dioxyde d'azote de jour $j-1$ entre 15h et 24h et le dioxyde d'azote du jour j à 1h, 4h et 6h.
- d'autre part des variables *météorologiques* trihoraires du jour j de 1h à 16h. Il s'agit d'abord de : la nébulosité, la température, la vitesse du vent, la direction du vent. Nous avons aussi tenu compte des variables suivantes : l'insolation trihoraire de 7h à 16h, l'humidité minimale du jour j, le maximum du gradient thermique vertical du jour j mesuré entre 1h et 6h, le maximum des températures du jour $j-1$.

Dans cette phase préliminaire nous utilisions 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)).

5.2. Evaluation des résultats

Quatre niveaux de pollution ou d'alerte en ozone sont considérés ici en fonction des valeurs observées. Ces niveaux d'alerte sont : niveau 0 : $0 \leq O3 < 130\mu g/m^3$, niveau 1 : $130 \leq O3 < 180\mu g/m^3$, niveau 2 : $180 \leq O3 < 280\mu g/m^3$ (seuil d'information de la population), niveau 3 : $280 \leq O3$ (seuil d'alerte à la pollution).

Les résultats obtenus par les modèles seront résumés dans un tableau de dépendance (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}

Tableau 3 : Tableau de croisement des prévisions et des observations

a_{ij} représente le nombre d'observations de niveau i prévu au niveau j. A partir de ce tableau on calcule le *Threat Score,* critère largement utilisé en météorologie et en pollution pour évaluer la qualité des prévisions. Supposons qu'on dispose du tableau de dépendance (2×2) suivant

bien prévus $C1$	*surestimés*
sous − estimés	*bien prévus* $C2$

Tableau 4 : Tableau de croisement pour le calcul du threat score

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

Le Threat Score noté TS est défini par :

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

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 ($130\mu g/m^3$), le threat score s'écrit alors

$$TS = \frac{\sum_{1\leq i,j\leq 3} a_{ij}}{\sum_{0\leq i,j\leq 3} a_{ij} - a_{00}} \tag{30}$$

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

5.3. Arbre de classification et arbres agrégés : comparaison

Dans l'étude empirique présentée aucune fonction de coût n'a été introduite. L'échantillon d'apprentissage $\mathcal{E}$ comportera toujours 622 observations et l'échantillon témoin $\mathcal{T}$, en comportera 200. Ces divisions de l'ensemble des données seront réalisées aléatoirement et avec stratification. On génère K=50 échantillons bootstrap

de $\mathcal{E}$. On construit $\hat{y}^B$ et $\widehat{\tau}^B$ comme en 4.2. Les performances de $\hat{y}^B$ sont résumées dans le Tableau 5.

On peut alors comparer le résultat à celui obtenu à partir de l'arbre de classification construit par validation croisée sur le même échantillon $\mathcal{E}$ et présenté par la figure 1. Les performances de cet arbre sont présentées dans le Tableau 6.
Une comparaison des performances sur ce seul exemple semble montrer l'intérêt de cette procédure d'agrégation. Les gains relatifs sont : -35.8% pour le taux de fausse classification (calculée sur l'échantillon témoin), et +40.4% pour le threat score.

114	6	0	0
11	43	5	0
0	10	9	0
0	0	2	0

Tableau 5 : Prévisions par arbres agrégés : TS=80.23%, $\tau^B = 17\%$

116	4	0	0
29	24	6	0
3	9	7	0
0	1	1	0

Tableau 6 : Prévisions par un arbre construit par validation croisée : TS=57.14%, $\tau^{et} = 26.5\%$

5.4. Validation des résultats

Dans le paragraphe précédent, nous avons montré sur un seul échantillon témoin que la procédure d'agrégation permet d'obtenir un gain dans la qualité des prévisions. Pour tenter de confirmer ce résultat, nous avons mis en oeuvre pour une valeur de $H = 100$ l'étape 4 de l'algorithme présenté en 4.2. Pour chaque itération nous calculons donc le taux de fausse classification moyen 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 (taux de fausse classification et threat score) obtenus sur ces 100 essais, sans agrégation (τ^{et} et TS^{et}) et avec agrégation (τ^B et TS^B), permettent de comparer les deux procédures.

Cet essai est réalisé sur deux stations de mesure, Vitrolles (VTRL), Rognac (RBRT). Les résultats sont présentés dans le tableau 7. Les gains relatifs présentés sont calculés par rapport à la moyenne (GAIN(%)).

L'amélioration relative de la qualité de prévision par agrégation en terme du taux de fausse classification est d'au moins 17%. L'augmentation relative du threat score est de 14.3% à 17.4% selon les stations.

Dans tous les cas l'écart-type du taux de fausse classification et du threat score, est plus faible pour les prédicteurs agrégés que pour les prédicteurs simples. Ceci montre le gain en stabilité obtenu suite à l'agrégation des arbres.

6. Conclusion

Les arbres de classification constituent un outil descriptif facilement utilisable pour la modélisation et la prévision. Leur instabilité peut être réduite grâce à

Station		τ^{et}	τ^{B}	TS^{et}	TS^{B}
VTRL	moyenne	28.6	22.5	57.9	66.2
	Ecart type	3.23	2.6	6.1	5.1
	GAIN(%)		**21.3**		**-14.3**
RBRT	moyenne	27.9	23.1	45.5	53.4
	Ecart type	2.8	2.3	5.7	4.7
	GAIN(%)		**17.2**		**-17.4**

Tableau 7 : Comparaisons du taux de fausse classification et du threat score moyens sans et avec agrégation par bootstrap sur deux stations des Bouches-du-Rhône. Le gain montre l'évolution relative des deux indices. Il est positif pour le taux de fausse classification (diminution) et négatif pour le threat score (augmentation).

une procédure de combinaison d'arbres construits sur des échantillons bootstrap de l'échantillon de base. De plus de cette procédure résulte un outil de prévisions plus performant. Nous l'avons montré ici dans le cadre de la prévision de l'ozone.

Il est vrai que l'agrégation d'arbres fait perdre l'aspect visuel de cette technique. Mais il est toujours possible de conserver l'arbre construit sur les données de base comme outil descriptif et utiliser les arbres agrégés pour la prévision. C'est ainsi que seront utilisées ces techniques pour la prévision de l'ozone par AIRMARAIX au cours de l'été 1999.

Remerciements : J'adresse mes remerciements à Claude Deniau (Université de la Méditerranée, Aix-Marseille II) et Georges Oppenheim (Université de Marne la Vallée) pour tout le temps qu'ils m'ont consacré.

Les programmes utilisés pour l'étude empirique, ont été écrits avec le logiciel S+.

BIBLIOGRAPHIE

[1] AKAIKE H., "A now look at the statistical model identification", *I.E.E.E. Transaction of Automatic Control. A.C.* (19), 1974, 716-723.

[2] 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., "Prévisions des pointes de pollution dans la région parisienne, O3 et NO2 : Phase opérationnelle", *Rapport de contrat de recherche Airparif*, 1997.

[3] BREIMAN L., FRIEDMAN J.H., OLSHEN R., STONE C.J., *Classification And Regression Trees*, Belmont CA., Wadsworth, 1984.

[4] BREIMAN L., "Heuristic of instability and stabilization in model selection", *The Annals of Statistics,* Vol. 24, N°6,1996a, 2350-2383.

[5] BREIMAN L., "Bagging Predictors", *Machine Learning*, 24, 1996b, 123-140.

[6] BREIMAN L., "Technical Note : Some Properties of Splitting Criteria", *Machine Learning*, 1996c, 24, 41-47.

[7] BURROWS W., BENJAMIN M., BEAUCHAMP S., LORD E.R., McCOLLOR D., THOMSON B., "CART Decision-Tree Statistical Analysis and Prediction of Summer Season Maximum Surface Ozone for the Vancouver, Montreal, and Atlantic Regions of Canada", *Journal of applied meteorology*, Vol. 34, 1994, 1848-1862.

[8] GHATTAS B., DENIAU C., "Essai de classification des stations de mesures d'ozone des réseaux AIRMARAIX-AIRFOBEP", *contrat* GREQAM-AIRMARAIX, 1998.

[9] MESSENGER R.C., MANDELL M.L., "A model search technique for predictive nominal scale multivariate analysis", *Journal of American Statistical Association* N° 67, 1972, 768-772.

[10] RYAN W.F., "Forecasting severe ozone episodes in the Baltimore metropolitan area", *Atmospheric Environment*, 29 (17), 1995, 2387-2398.

[11] SHEIFINGER H., STOHL A., KROMP-KOLB H., SPANGL W., "A statistical method for predicting daily maximum ozone concentrations", *Gefahrstoffe-Reinhaltung der Luft*, Vol. 56, 1996, 133-137.

[12] WILKS D.S., *Statistical Methods in Atmospheric Sciences : an introduction*, Academic Press, 1995.