PROCEDURE MANUELLE POUR LA CONSTRUCTION D'ARBRES DE REGRESSION SOUS S+

Badih GHATTAS

GREQAM - Université de la Méditerranée

Ghattas@lumimath.univ-mrs.fr

1. Introduction

Les techniques de régression et de classification par arbres (CART, Breiman *et al.* (1984)) constituent un outil de modélisation et de prévision simple et de plus en plus répandu par la diversité des applications qui lui font appel. Comme beaucoup de méthodes statistiques, ces techniques sont basées sur l'estimation d'une espérance conditionnelle (dans le cas des arbres de régression) en minimisant une erreur quadratique, éventuellement pénalisée. Le logiciel Splus, par exemple, dispose des outils nécessaires à la mise en œuvre de ces techniques.

Le modèle obtenu avec ces méthodes peut être visualisé par un arbre (cf. figure 1). Selon les applications considérées, l'arbre *optimal* n'est pas satisfaisant : sa taille peut être soit trop petite soit trop grande, ou les règles binaires figurant aux nœuds de l'arbre peuvent être incohérentes à la réalité physique dont sont issues les données.

Dans un contexte particulier, qui est celui de données issues de l'environnement (la pollution atmosphérique), nous proposons une procédure dite *manuelle* de construction d'arbres de régression et de classification.

Cette procédure permet de construire des arbres en combinant le critère quadratique usuel, et d'autres critères qui sont souvent subjectifs et qui dépendent du problème considéré: la structure de l'arbre, le contenu des feuilles et l'objectif de la modélisation.

Après un bref rappel de ces techniques, nous décrivons cette procédure en l'illustrant par la prévision de concentrations maximales quotidiennes de l'ozone. Nous montrons d'autre part qu'une telle procédure "subjective" peut s'avérer plus performante que la procédure classique.

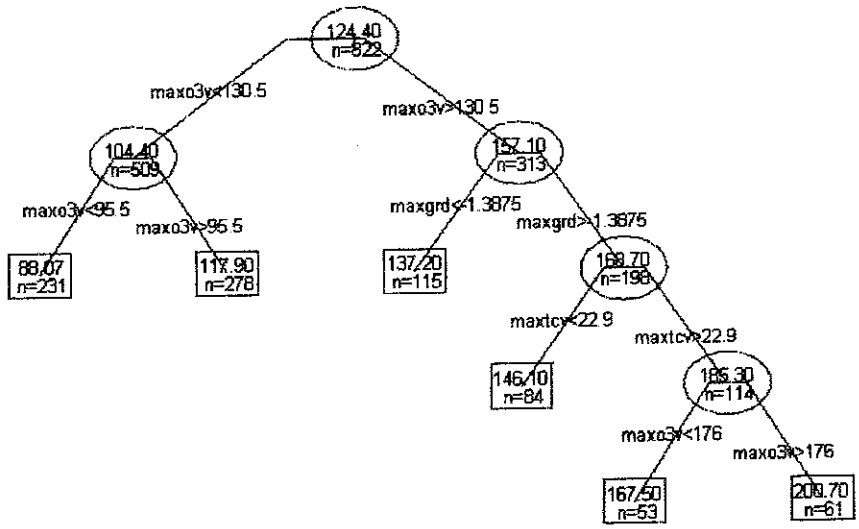
2. Les arbres de régression ou de classification par l'exemple

2.1 Introduction

Les méthodes CART ont été popularisées par Breiman *et al.* (1984); en langue française Ghattas (1999) décrit la méthode de régression par arbre et l'illustre à l'aide d'exemples empruntés à la prévision de la pollution par l'ozone.

Ce sont des techniques de discrimination et de prévision simples et performantes mais aussi des outils de statistiques descriptives très intéressants. La visualisation du modèle à l'aide d'un arbre binaire permet la mise en évidence des variables explicatives actives de la variable à prévoir, ici le maximum quotidien de l'ozone (maxo3) dans une station de mesure du réseau.

Présentons les idées de la régression par arbre en se référant à un exemple simplifié. La variable à prévoir est évidemment le maximum de l'ozone (maxo3), les variables explicatives sont réduites ici à trois: le maximum d'ozone de la veille (maxo3v), la température maximale de la veille (maxtcv), et le maximum du gradient vertical observé entre 00h et 06h (maxgrd). Un exemple d'arbre construit pour la station de Vitrolles avec ces variables est donné dans la figure 1.



Remarque : Notons que ces variables explicatives sont toutes quantitatives ; la méthode autorise aussi le traitement de variables qualitatives.

Figure 1 : Arbre à 6 feuilles. La variable expliquée est "maxo3", les variables explicatives sont "maxo3v", "maxtcv" et "maxgrd". Pour chaque nœud (encerclé) et chaque feuille (encadrée), on indique la moyenne de la variable expliquée pour les observations dans ce nœud, ainsi que le nombre d'observations.

2.2 L'arbre et son utilisation

Nous présentons ensuite très succinctement la méthode de construction. Pour le moment, la question est : comment ça marche ?

Aujourd'hui à 06h TU (Temps Universel), nous devons prévoir le maxo3 et nous disposons des observations des trois variables quantitatives : maxo3v= 140 $\mu g/m^3$, maxtc= 25°C, maxgrd= -1.

A la racine de l'arbre on dispose de toutes les observations (n=822 journées) et la moyenne du maximum quotidien de l'ozone pour ces journées est de 124.4 $\mu g/m^3$. La première coupure de l'arbre est sur le "maxo3v". Le maximum de l'ozone la veille étant supérieur à 130.5$\mu g/m^3$ on s'achemine dans la branche droite de l'arbre (si le maximum de l'ozone la veille était inférieur à 130.5$\mu g/m^3$ on serait aller dans la branche de gauche). Le maximum du gradient de la nuit valant −1, on poursuit vers la droite (maxgrd >-1.3875), puis vers la droite (maxtcv=25°C > 22.9°C), puis vers la gauche (maxo3v<176$\mu g/m^3$). La feuille (entourée d'un rectangle) dans laquelle on se retrouve indique une valeur 167.5 qui est la valeur maximale prévue de l'ozone pour le jour courant. Le nombre de journées qui se retrouvent dans la même situation vis à vis des variables considérées (et donc dans la même feuille) est de n=53.

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

2.3 Idée de la construction de l'arbre

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

Dans l'ellipse du haut de l'arbre - *la racine* - se situent toutes les observations de l'échantillon d'apprentissage. La première division de l'échantillon en deux classes, (maxo3v $\leq$130.5 et maxo3v >130.5) est celle qui, de toutes les divisions possibles du même type obtenues à l'aide de chacune des trois variables explicatives, minimise le critère de la somme des variances de chacune des deux classes réalisées. Donc la division de l'échantillon en deux classes dépend d'une variable (ici maxo3v) et d'un seuil sur cette variable (ici 130.5$\mu g/m^3$).

Chacune de ces classes constitue un nœud de l'arbre (représenté par une ellipse sur la figure 1); sur chacun de ces nœuds la même procédure de division binaire est appliquée et conduit à d'autres divisions. On a ensuite recours à une règle d'arrêt de la procédure (détaillée dans Ghattas

1999) qui conduit à la réalisation d'un arbre terminal dont le nombre de feuilles n'est pas trop grand. Les nœuds extrêmes de l'arbre, sont appelés *feuilles* de l'arbre et sont symbolisés par des rectangles sur la même figure. Leur rôle dans le calcul de la prévision a été vu précédemment.

Un algorithme *d'élagage* peut être utilisé ensuite pour sélectionner à partir l'arbre obtenu un sous arbre "optimal".

Pour plus de détails concernant la procédure de construction, voir Breiman *et al.* (1984) ou Ghattas (1999).

3. La procédure manuelle

Nous proposons une construction *interactive* de l'arbre de prévision :

A chaque étape de la construction interactive de l'arbre :

- Chaque feuille de l'arbre peut être partagée en deux (le partage sera optimal par rapport à la somme des variances des deux nouvelles feuilles), et chaque branche de l'arbre peut être élaguée manuellement.
- Deux graphiques sont présentés dans deux fenêtres : celui de l'arbre courant (figures 2a, 3a), et celui des nuages des points croisant les valeurs observées contre les valeurs prévues de la variable expliquée (figures 2b et 3b). A tout moment l'utilisateur peut basculer d'un graphique à un autre.

- On construit un arbre à 2 feuilles.
- On procède aux partages successifs des feuilles de la manière suivante.
  On choisit la feuille pour laquelle la moyenne de la variable expliquée est maximale, et on la partage en la sélectionnant par un "double clique". *Le but étant d'isoler en premier les observations pour lesquelles le maximum de l'ozone (la variable expliquée) est élevé. Ces observations sont rares. Dès que l'on dispose d'une feuille dont la moyenne indiquée dépasse le seuil 280* $\mu g/m^3$ (seuil d'alerte), *on se base sur le deuxième graphique, croisant les observations et les prévisions. La dispersion d'une feuille devient le critère de base d'autant plus que le niveau prévu dans cette feuille est grand.*
- On arrête cette procédure lorsque les distributions des feuilles d'intérêt sont homogènes.

Une illustration de cette procédure est donnée avec les figures 2 et 3.

La figure 2a présente un arbre à 4 feuilles construit sur des données de la station de Vitrolles. La figure 2b présente le nuage des points croisant en abcisse les valeurs observées et en

ordonnée les valeurs prévues de la variable dépendante, le maximum quotidien de l'ozone. On note sur ce graphique le nombre d'observations (Nobs) le nombre de variables (Nvar), l'écart type (ET) des résidus et le nombre de feuilles (nf). La grille tracée indique des seuils souvent associés à ceux fournis par les réglementations des niveaux d'ozone en terme de prévention ou d'alerte à la pollution (130µg/m^3 - protection des végétaux,180µg/m^3 -seuil d'information du public, 280µg/m^3 - seuil d'alerte).

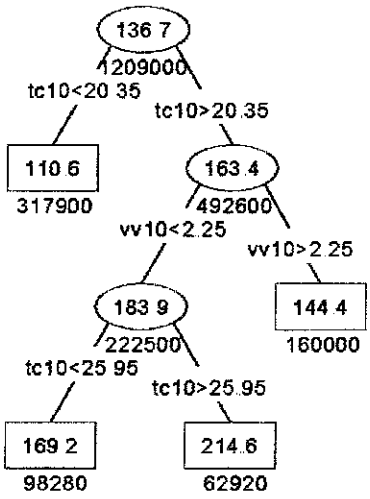


Figure 2a

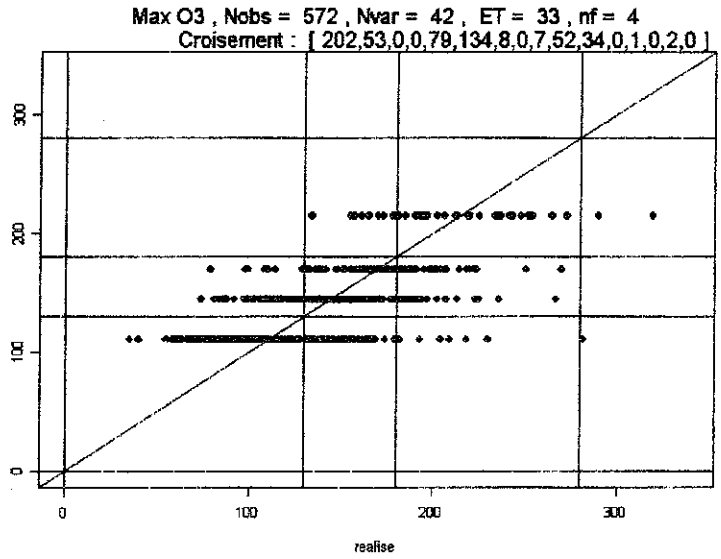


Figure 2b

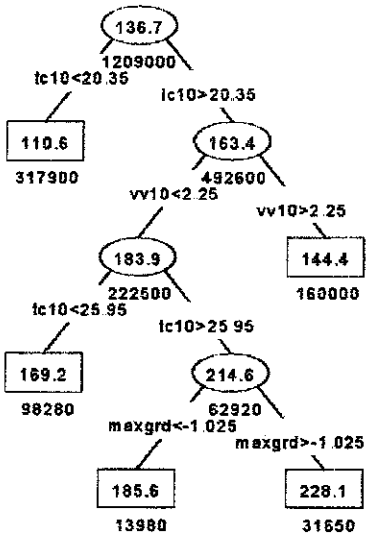


Figure 3a

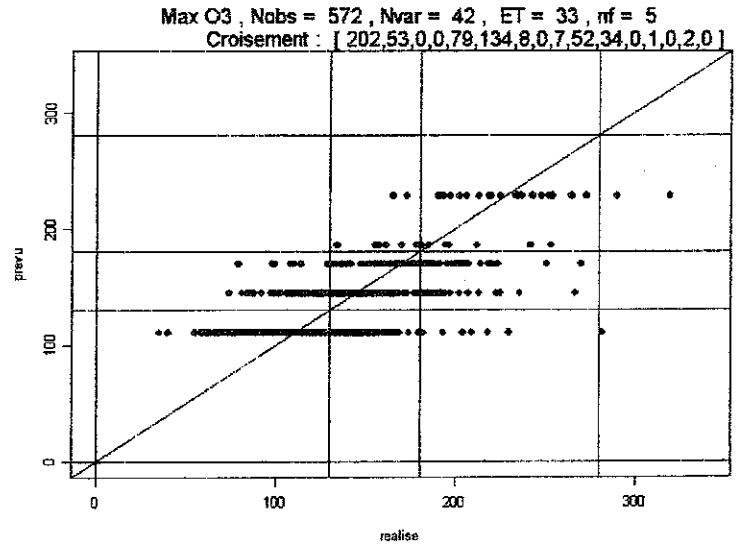


Figure 3b

Les lignes de points horizontales apparaissant sur la figure 2b correspondent aux feuilles de l'arbre et se situent aux niveaux indiqués dans les feuilles ; 110.6, 169.2, 214.6, 144.4. Ce graphique permet de faire le choix de la suite de la procédure de construction de l'arbre. La hauteur

des lignes horizontales indique les valeurs prévues dans les feuilles, et leur dispersion est celle des valeurs observées dans les feuilles. On cherchera systématiquement à limiter la dispersion des feuilles, et en priorité à s'assurer que certaines feuilles présentent un niveau élevé de l'ozone, en particulier dépassant le seuil de 180µg/m3.

Dans cet exemple on a choisit de partager la feuille dont le niveau est le plus élevé et on a double cliqué dans la figure 2a sur cette feuille. Le résultat est donné dans la figure 3a pour le nouvel arbre obtenu, et dans la figure 3b son évaluation mise à jour.

4. Comparaison d'un arbre "manuel" et d'un arbre optimal obtenu par validation croisé

La comparaison des arbres sera faite avec des tableaux de croisement des valeurs observées (en ligne) avec les valeurs prévues (en colonne). Ces tableaux ont la forme suivante :

		Niveau prévu			
		0	**1**	**2**	**3**
Niveau réalisé	**0**	a_{00}	a_{01}	a_{02}	a_{03}
	1	a_{10}	a_{11}	a_{12}	a_{13}
	2	a_{20}	a_{21}	a_{22}	a_{23}
	3	a_{30}	a_{31}	a_{32}	a_{33}

a_{ij} indique le nombre de fois où l'on a observé le niveau *i* pour le maximum de l'ozone alors qu'on a prévu le niveau *j*. Ces niveaux correspondent à un découpage du maximum quotidien de l'ozone à partir des seuils réglementaires (0 : [0,130[, 1 :[130,180[, 2 : [180,280[, 3 : >=280).

Arbre obtenu par validation croisée

La recherche d'un arbre de taille optimale est faite par validation croisée, sur vingt échantillons stratifiés[1]. L'arbre ainsi obtenu possède 15 feuilles (Figure 4). On indique au niveau des feuilles, la moyenne des observations de la variable expliquée (en encadré) ainsi que la déviance (somme des carrés des écarts à la moyenne).

La branche de gauche d'un nœud correspond à la réalisation de la règle binaire au niveau de ce nœud (ex : tc10 <20.95°). La branche de droite correspond à sa non réalisation (ex : tc10 > 20.95°).

Les variables apparaissant aux premiers nœuds sont : la température prévue à 10h du matin, le maximum d'ozone de la veille et la vitesse du vent prévue à 10h du matin.

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.

---

[1] 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

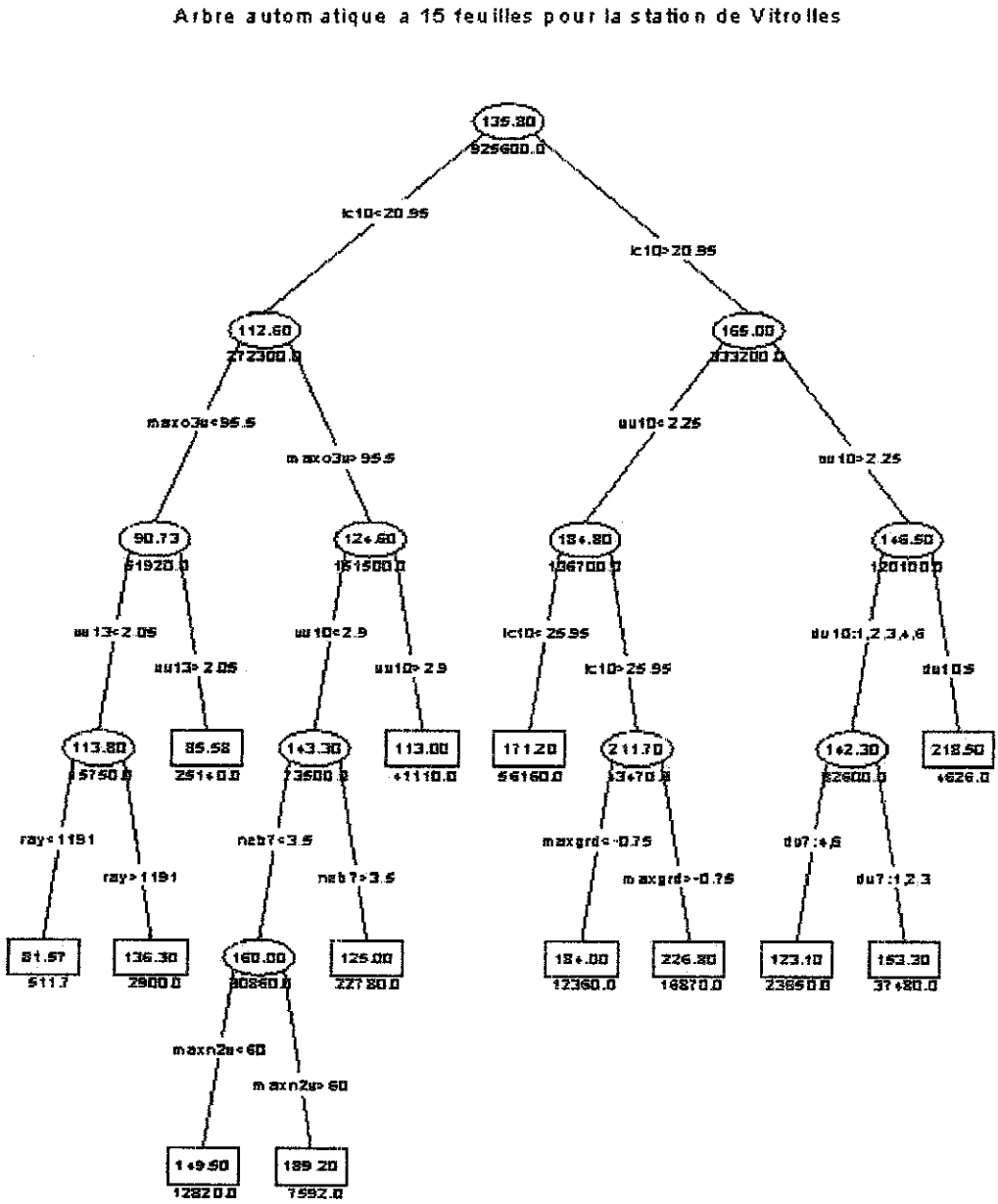


Figure 4 : arbre de validation croisée, 15 feuilles.

échantillon de construction

Niveau prévu

Niveau réalisé	0	1	2	3
0	194	18	0	0
1	59	113	10	0
2	3	36	38	0
3	0	0	1	0

TS = 71.22 σ = 23.65

échantillon témoin

Niveau prévu

Niveau réalisé	0	1	2	3
0	33	9	1	0
1	13	23	3	0
2	1	7	8	0
3	0	0	2	0

TS = 64.18 σ =34.73

Tableau 1 : évaluation de l'arbre obtenu par validation croisée

Le tableau d'alerte associé à cet arbre (tableau 1) donne les valeurs du threat score (TS) pour 130µg/m^3 et de l'écart type des résidus d'une part pour l'échantillon d'apprentissage (tableau 1a), et d'autre part pour l'échantillon témoin (tableau 1b).

Le threat score est un indice donnant la qualité de la prévision des valeurs élevées de la variable expliquée. Pour un tableau comme ci dessus noté $\{a_{ij}\}_{0<i,j<3}$ le threat score est calculé de la manière suivante :

$$TS = 100 * \frac{\sum_{i,j>0} a_{ij}}{\sum_{i\,j} a_{ij} - a00}$$

Cet indice est particulièrement intéressant dans notre application car il est indicateur de bonne prévision des niveaux élevés de la variable expliquée. Plus il est élevé, meilleure est la prévision de ces niveaux. Son inconvénient est sa variabilité importante quand on dispose de peu d'observations de ce type. Il est donc préférable d'observer de près les tableaux de croisement.

On peut noter que le threat score est plus élevé pour l'échantillon d'apprentissage que pour l'échantillon témoin. C'est l'inverse pour l'écart type.

Arbre manuel

Nous présentons **un** arbre à 15 feuilles dans la figure 5. Il a été obtenu par la procédure décrite dans le paragraphe 3.

Les tableaux de croisement sont présentés pour l'échantillon d'apprentissage (tableau 2a) puis pour l'échantillon témoin.

Notons d'abord les différences dans la structure des arbres. Nous pouvons nous référer aux nœuds de l'arbre par les valeurs indiquées du maximum d'ozone (elles sont toutes différentes).

Trois branches de l'arbre de validation croisée (figure 4) n'existent pas dans l'arbre manuel. Ce sont celles issues des nœuds indiquant les valeurs 90.73, 160, et 142.30. Ces niveaux sont faibles ou moyens.

Par contre deux feuilles de l'arbre de validation croisée ont été développées dans l'arbre manuel. Ce sont celles indiquant les valeurs 171.20 et 226.80. Le développement de ces feuilles dans l'arbre manuel a permis de faire apparaître des nouvelles feuilles avec des valeurs relativement élevées des maximum de l'ozone : 227.4, 281.5,235, et 184.30.

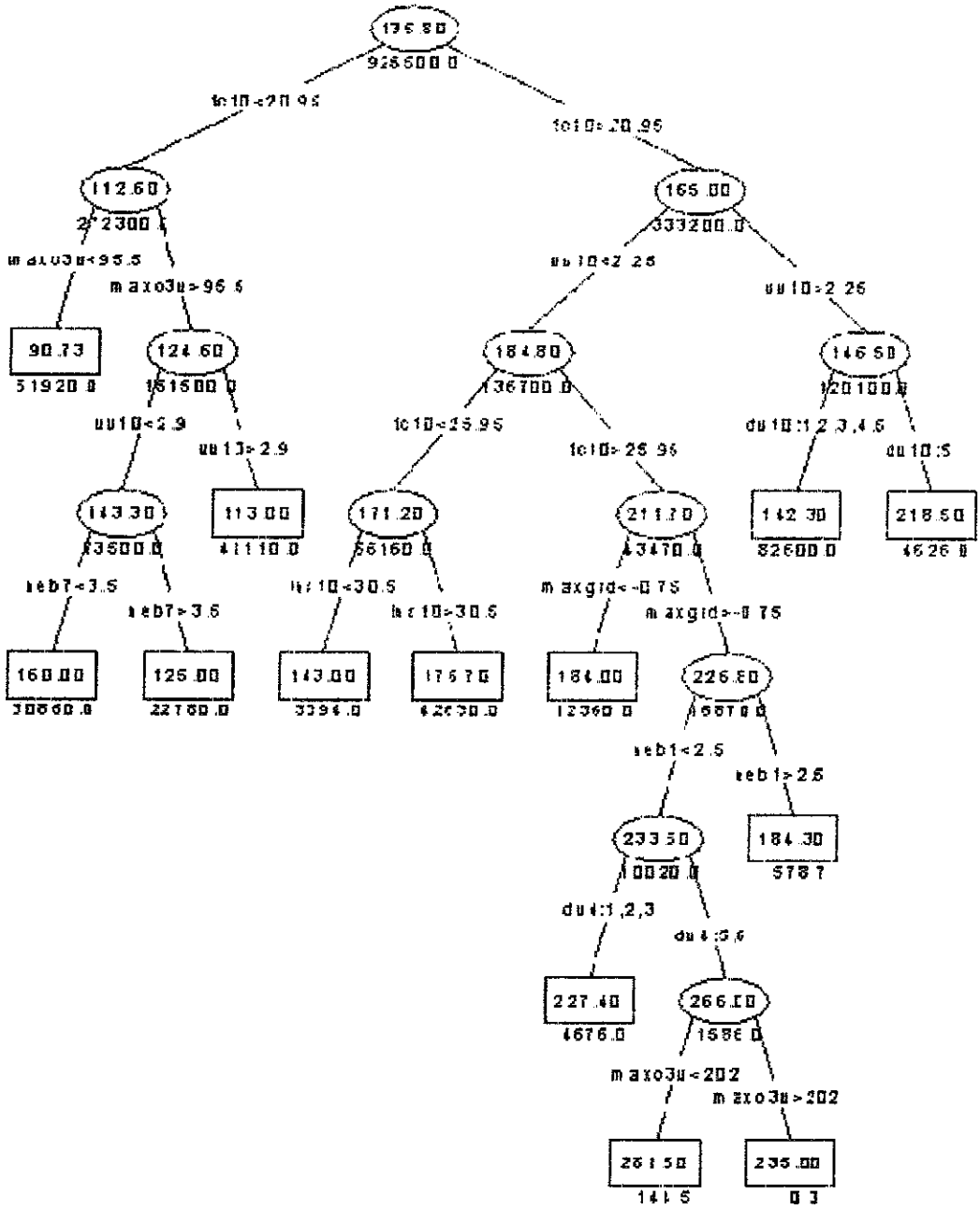


Figure 5 : arbre manuel à 15 feuilles

échantillon de construction

Niveau prévu

Niveau réalisé	**0**	**1**	**2**	**3**
0	175	37	0	0
1	53	122	7	0
2	1	44	31	1
3	0	0	0	1

TS = 69.36 σ = 25.10

Tableau 2a

échantillon témoin

Niveau prévu

Niveau réalisé	**0**	**1**	**2**	**3**
0	31	12	0	0
1	10	27	2	0
2	0	8	8	0
3	0	1	1	0

TS = 68.12 σ = 33.61

Tableau 2b

Tableau 2 : évaluation de l'arbre manuel

Le tableau 2 donne l'évaluation de l'arbre manuel pour l'échantillon d'apprentissage (tableau 2a) et pour l'échantillon témoin (tableau 2b).

Les tableaux 1a et 2a reflètent directement les différences dans la structure des deux arbres. La dernière ligne montre que la seule observation de niveau supérieur à 280 $\mu g/m^3$ est bien prévue par l'arbre manuel mais ne l'est pas par l'arbre de validation croisée.

D'autre part les lignes 2 et 3 montrent que dans l'arbre manuel on a moins de *surestimations* que dans l'arbre de validation croisée pour les niveaux entre 130 $\mu g/m^3$ et 280 $\mu g/m^3$.

L'arbre manuel surestime les observations au-dessous de 130 $\mu g/m^3$ plus que l'arbre de validation croisée. Ces observations nous intéressent moins, et leur sur estimation est d'un seul niveau; elles sont prévues dans la classe 130-180.

L'arbre manuel est globalement plus performant sur l'échantillon témoin que l'arbre de validation croisée (tableau 1b et 2b). Le threat score est plus élevé pour l'arbre manuel et l'écart type est plus faible.

Les observations de plus de 280$\mu g/m^3$ (ligne 4) sont mal prévues par les deux arbres. Elles sont *sous-estimées* d'un niveau par l'arbre de validation croisée, et une des deux observations dans ce cas est sous-estimée de deux niveaux par l'arbre manuel.

Le reste du tableau est nettement à l'avantage de l'arbre manuel. Pour les observations de moins de 130 $\mu g/m^3$ l'arbre manuel n'a aucune surestimation de plus d'un niveau. Pour le niveau 0 (130-180), la deuxième ligne, l'arbre manuel a une meilleure prévision, moins de surestimation et moins de sous-estimation que l'arbre de validation croisée. Pour le niveau 1 (180-280, la troisième ligne) l'arbre manuel n'a de sous-estimation que d'un niveau, alors que l'arbre de validation croisée en a une de deux niveaux.

5. Implémentation avec S+

La mise en œuvre de la procédure manuelle utilise plusieurs fonctions du logiciel S+ comme *tree* - pour la construction d'un arbre, *prune.tree* - pour élaguer un arbre, *graft.tree* - pour greffer une branche à un arbre, *snip.tree* - pour éliminer une branche d'un arbre.

Certaines fonctions de S+ ont été modifiées et adaptées aux besoins de l'application décrite ici. En particulier, *text.tree* - qui place les informations textuelles sur l'arbre et *identify.tree* - qui repère les observations appartenant à un nœud.

D'autres fonctions développées particulièrement pour cette procédure sont aussi utilisées, comme par exemple celle du calcul des threat scores et des tableaux de croisements.

Toutes ces fonctions sont disponibles par simple demande auprès de l'auteur.

Conclusion

Nous avons montré à travers un exemple d'application un moyen d'adapter une technique de modélisation, les arbres de régression, à des cas particuliers. Par une procédure simple dite "manuelle" nous avons intégré des critères subjectifs et propres à l'application en question, aux critères classiques d'ajustement des modèles. Dans le cas de la pollution par l'ozone, nous avons forcé le modèle à une optimisation *globale* mais aussi "*locale*", celle des valeurs élevées de la variable dépendante.

Il en résulte un modèle plus adapté à l'application en question. Les performances sont souvent plus intéressantes surtout dans les régions optimisées mais aussi globalement.

Ce genre de procédure est de plus en plus utilisé pour la prévision de la pollution atmosphérique dans les réseaux de surveillance de la qualité de l'air (Atmo 1998).

Références

Breiman L., Friedman J.H., Olshen R., Stone C.J. (1984) *Classification and Regression Trees*, Wadsworth, Belmont CA.

Ghattas B., (1999), Prévisions des pics d'ozone par arbres de régression simples et agrégés par bootstrap. *Revue de Statistique Appliquée*, XLVII (2), pp.61-80.

XVII^{ièmes} rencontres ATMO 1998, Caen, octobre 1998.