IMPORTANCE DES VARIABLES DANS LES METHODES CART

Badih GHATTAS

GREQAM - Université de la Méditerranée

Ghattas@lumimath.univ-mrs.fr

1. Introduction

Un arbre de régression ou de classification A obtenu par les méthodes CART (Breiman *et al*, 1984) permet de visualiser des variables dites *actives* qui participent directement à sa construction, et donc à la procédure de discrimination et de prévision correspondante.

Cela dit certaines variables explicatives qui ne jouent plus aucun rôle lorsque l'arbre est construit ont pu être pour plusieurs nœuds *concurrentes* des variables actives. La connaissance de ces variables concurrentes est utile à différents niveaux et peut permettre d'établir une *hiérarchie* de l'ensemble des variables explicatives. Cette hiérarchie peut servir pour mettre en œuvre d'autres méthodes statistiques avec un nombre de variables réduit.

L'exemple présenté sur la figure 1 est un arbre de régression pour la prévision quotidienne du maximum de l'ozone dans la station d'Istres des Bouches du Rhône. Le tableau 1 donne les 10 variables les plus importantes obtenus par la hiérarchie établie sur cet arbre, ainsi qu'un *indice d'importance* sur une échelle de 0 à 100. La valeur 100 est attribuée à la variable la plus importante ayant servi à la construction de ce modèle.

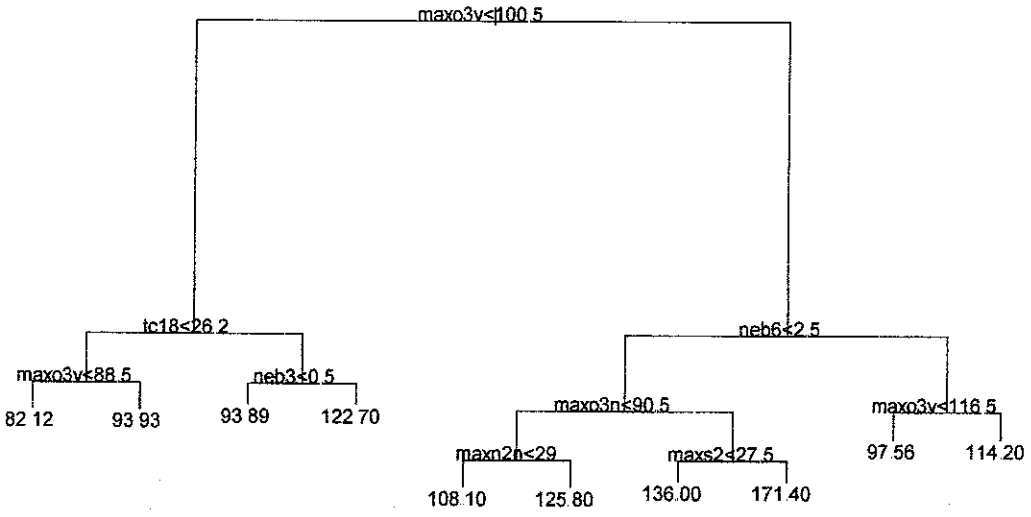

Figure 1 : Les variables actives sont : le maximum de l'ozone la veille "maxo3v" et la nuit "maxo3n" (exprimés en $\mu g/m^3$), la température à 6h et à 18h "tc6" et "tc18" (exprimés en degrés Celsius), la nébulosité à 3h et à 6h "neb3"et "neb6" (ayant neuf niveaux ordonnés de 1 à 9), le maximum et le minimum de NO2 la nuit "maxn2"-"minn2n" et le maximum de SO2 la nuit "maxs2" (exprimés tous deux en $\mu g/m^3$).

maxo3v	maxgrd	dv18	dv6	dv15	vv9	neb6	tc12	minn2n	neb3
100	69	69	61	58	49	48	48	47	47

Tableau 1 : les 10 variables les plus importantes ainsi que leur indice d'importance parmi celles ayant servi à la construction de l'arbre ci-dessus

Comme l'indique la figure ci-dessus, certaines variables comme par exemple le gradient maximal de la nuit (« *maxgrd* ») et la direction du vent à 18 heures (« *dv18*») ne sont pas actives. Cependant elles sont classées en deuxième et en troisième position dans l'ordre d'importance.

Le but de ce papier est de décrire la procédure qui permet d'établir une hiérarchie sur les variables explicatives pour les modèles de régression (ou de classification) par arbre. Le rôle des variables importantes qui n'apparaissent pas sur l'arbre sera examiné et on examinera la stabilité d'une telle procédure vis à vis de perturbation sur l'échantillon d'apprentissage.

Notations

On dispose d'un échantillon de données $E = (\boldsymbol{x}_n, j_n)_{1 \le n \le N}$ où $\boldsymbol{x}_n = (x_n^1, x_n^2, \ldots, x_n^q)$ correspond à une réalisation de la variable aléatoire $\boldsymbol{X} = (X^1, X^2, \ldots, X^q)$ à valeurs dans R^q et j_n à celle d'une variable Y qui peut être réelle ou qualitative à J modalités. On notera x^m une réalisation quelconque de la $m^{\text{ième}}$ composante du vecteur X. Soit $t \subset E$ un nœud de l'arbre A construit sur l'échantillon d'apprentissage E. Dans le cas où la variable expliquée est *qualitative* on introduit :

- La probabilité a priori de la classe j sera ici définie par $\pi_j = \dfrac{N_j}{N}$, où $N_j = Card\{j_n ; j_n = j\}$.
- Etant donné $t \subset E$ on note $N(t)$ le cardinal de l'ensemble t.
- De même $N_j(t)$ est le cardinal de l'ensemble $\{(\boldsymbol{x}_n, j_n) \in t ; j_n = j\}$.
- Un estimateur de $P(j,t)$ (qui est la probabilité qu'une observation appartienne à t et qu'elle ait pour classe j) par substitution, noté $p(j,t)$ est défini[1] par : $p(j,t) = \pi_j \dfrac{N_j(t)}{N_j}$.

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

- Un estimateur de $P(t)$ (la probabilité d'appartenir au nœud t) par substitution, noté $p(t)$ est défini par : $p(t)=\sum_{j=1}^{J} p(j,t)$.
- Enfin, $P(j\mid t)$ la probabilité a posteriori qu'une observation ait la classe j sachant qu'elle appartient à t, est estimé par $\frac{p(j,t)}{p(t)}$ qui avec la définition de π_j, est égal à $\frac{N_j(t)}{N(t)}$.

2. Construction d'un arbre de régression ou de classification

Un arbre de régression ou de classification est construit avec une procédure itérative. Dans cette procédure on commence par rechercher une *règle de division binaire* $d=d(x^m,s)$ du type $x^m \le s$ ($s\in R$ si X^m est quantitative) ou $x^m \in S$ (où S est un sous-ensemble de l'ensemble des modalités de X, si X est qualitative) permettant de partager l'ensemble des observations initiales (E noté aussi t_0 et dit *racine* de l'arbre) en deux sous-ensembles t_g et t_d (dits *nœuds descendants* de t_0 de l'arbre). Parmi tous les partages possibles explorés sur toutes les variables explicatives et tous les seuils, on retient celui qui maximise un des deux indices suivants :

* $\Delta\hat{R}(d,t)=\hat{R}(t)-\hat{R}(t_g)-\hat{R}(t_d)$ si Y est *quantitative*

avec

$$\hat{R}(t)=\frac{1}{Card(t)}\sum_{i;(x_i,j_i)\in t}(y_i-\bar{y}_t)^2 \text{ et } \bar{y}_t=\frac{1}{Card(t)}\sum_{i;(x_i,j_i)\in t} y_i$$

ou

** $\Delta\hat{h}(d,t)=\hat{h}(t)-p_g\hat{h}(t_g)-p_d\hat{h}(t_d)$ si Y est *qualitative*

avec $\hat{h}(t)$ une fonction dite d'*hétérogénéité* qui peut être

soit $\hat{h}(t)=-\sum_{j=1}^{J} p(j\mid t)\log(p(j\mid t))$ (dérivée de l'entropie de Shannon)

soit $\hat{h}(t)=-\sum_{j\neq k} p(j\mid t)p(k\mid t)$ (dérivée de l'indice d'inégalité de Gini)

où $p(j\mid t)=\frac{N_j(t)}{N(t)}$ est une estimation de la probabilité a posteriori dans t de la classe j.

Une comparaison empirique de ces deux critères est donnée dans Breiman (1996c).

Dans les deux cas (régression ou classification) on parle de *déviance* pour les critères $\hat{R}(t)$ et $\hat{h}(t)$. *La règle optimale de division de l'ensemble* t_0 *est celle qui minimise la somme des déviances intra classes des nœuds descendants*.

Une fois cette règle obtenue et le partage effectué on recommence la même procédure de partage appliquée aux nœuds t_g et t_d. On itére cette procédure jusqu'à par exemple, obtenir des sous-ensembles ayant très peu d'observations.

On obtient ainsi un arbre, auquel on peut appliquer un algorithme d'*élagage*, permettant de rechercher dans cet arbre un sous arbre optimal vis à vis d'un *critère pénalisé* (la déviance par exemple).

Pour plus de détails concernant la procédure de construction, voir Breiman *et al.* (1984) ou Ghattas (1999a) pour la version régression et Gueguen *et al.* (1988) pour la version classification.

3. Divisions de substitution et leur utilisation

3.1. Division concurrente

La division d^* utilisée au niveau d'un nœud t d'un arbre A est une division optimale dans le sens où c'est celle qui maximise la décroissance de la déviance.

La division qui réalise le *deuxième maximum* de ce critère est la première *division concurrente* à d^*. Elle peut être réalisée sur une autre variable ou sur la même que d^* avec un autre seuil et elle peut causer une diminution de ce critère très voisine de celle réalisée par la division d^*.

3.2. Division de substitution

Dans ce paragraphe ainsi que dans le paragraphe suivant nous noterons une division d d'un nœud t, $d = d(t)$. De même d^*, d_m, et $\tilde{d}_m$ désignent respectivement $d^*(t)$, $d_m(t)$ et $\tilde{d}_m(t)$.

Soient t le nœud d'un arbre A, et d^* la meilleure division de t en les descendants t_g et t_d. Soit X^m $1 \le m \le q$ une variable réelle quelconque, D_m l'ensemble de ses divisions possibles (*e.g.* $x^m \le \alpha$) et D_m^c l'ensemble des divisions complémentaires (*e.g.* $x^m > \alpha$). Pour toute division

$d_m \in D_m \cup D_m^c$ du nœud t selon ses descendants t_d et t'_g on note $N_j(t'_g \cap t_g)$ le nombre d'éléments de la classe j de t envoyés sur le descendant de gauche par d_m et d^*; ils sont donc dans $t_g \cap t_g$.

Estimons la probabilité $P(t_g \cap t_g)$ qu'une observation de t aille à gauche de t pour les deux divisions d_m et d^*, par :

$$p(t_g \cap t_g) = \sum_{j=1}^{J} \pi_j \frac{N_j(t_g \cap t_g)}{N_j}$$

La probabilité que les divisions d_m et d^* envoient une observation de t vers leur descendant de gauche est :

$$p_{gg}(d_m, d^*) = \frac{p(t'_g \cap t_g)}{p(t)}$$

On définit de la même manière $p_{dd}(d_m, d^*)$.

Nous définissons alors la probabilité que la division d_m prédise correctement la meilleure division d^* par :

$$p(d_m, d^*) = p_{gg}(d_m, d^*) + p_{dd}(d_m, d^*)$$

Définition

Une division $\tilde{d}_m \in D_m \cup D^c{}_m$ basée sur la variable X^m est appelée *division de substitution* pour d^* si :

$$p(\tilde{d}_m, d^*) = \max_{d_m \in D_m \cup D^c{}_m} p(d_m, d^*)$$

$\tilde{d}_m$ est parmi les divisions basées sur X^m, celle qui peut prédire au mieux le partage effectué par d^*.

3.3. mesure d'association de deux divisions

Si d^* est la division sélectionnée au nœud t, elle envoie les observations vers t_g avec une probabilité p_g et vers t_d avec une probabilité p_d.

Si la variable sur laquelle est basée la division d^* n'est pas disponible pour nouvelle observation tombant dans t on voudrait pouvoir acheminer cette observation vers un des deux descendants de t.

Une règle d'acheminement naturelle peut-être t_g si $p_g = \max(p_g, p_d)$; la probabilité de se tromper en utilisant une telle règle d'acheminement est dans ce cas $per_1 = \min(p_g, p_d)$.

Une autre règle d'acheminement pourrait utiliser une division de substitution $\tilde{d}_m$. La probabilité de mal acheminer cette observation est dans ce cas $per_2 = 1 - p(\tilde{d}_m, d^*)$.

Il est préférable d'utiliser une règle de substitution pour l'acheminement de la nouvelle observation si $per_2 < per_1$.

La comparaison relative de ces deux règles d'acheminement est donnée par la mesure d'association entre $\tilde{d}_m$ et d^* :

$$As(\tilde{d}_m, d^*) = \frac{per_1 - per_2}{per_1}$$

Si cette mesure est petite (mais positive) la division de substitution contribue peu dans la réduction de l'erreur d'acheminement. Si elle vaut un, alors la division de substitution prévoit parfaitement le partage effectué par d^*. Si elle est négative, $\tilde{d}_m$ n'a pas d'intérêt car la probabilité de mauvais acheminement suite à son utilisation est plus élevée que celle obtenue par la première règle naturelle.

3.4. Importance des variables

L'importance d'une variable X^m pour l'arbre A est calculée de la manière suivante.

$$I(X^m) = \sum_{t \in A} \Delta\hat{R}(\tilde{d}_m(t), t) \qquad \text{pour la régression}$$

$$I(X^m) = \sum_{t \in A} \Delta\hat{h}(\tilde{d}_m(t), t) \qquad \text{pour la classification}$$

$\tilde{d}_m(t)$ est la division de substitution au nœud t basée sur la $m^{ième}$ variable.

Cette mesure dépend de l'arbre à partir duquel elle est calculée. C'est la somme des diminutions de la déviance provoquées à chaque nœud t de l'arbre, si l'on remplaçait pour chaque nœud la division optimale par la division de substitution $\tilde{d}_m(t)$, basée sur la variable X^m.

En général on ramène l'importance des variables à l'intervalle [0..100], où la variable la plus importante aura pour indice 100 (les importances obtenues par les formules ci-dessus sont divisées par leur maximum et multipliées par cent).

Un exemple d'utilisation de la hiérarchie des variables est donnée dans Casmassi (1998). Dans cet exemple la hiérarchie des variables a servi à sélectionner des paramètres parmi un grand nombre de variables explicatives. Ces variables pourraient être utilisées ensuite pour une régression non paramétrique ou une classification.

3.5. Variables cachées

Une variable peut être assez importante sans être active. Ceci est illustré sur la figure 1, présentant un arbre à 10 feuilles construit sur les données de la station d'Istres (*Deniau et al.* 1998).

On remarque que les variables « *maxgrd* » et « *dv18* »[2] d'ordre d'importance respectif 2 et 3 ne sont pas actives. Afin de comprendre l'importance de ces variables pour la construction de l'arbre présenté, nous avons observé à chaque nœud de l'arbre, les divisions concurrentes basées sur ces variables *cachées*. Le tableau 2 donne pour chaque nœud de l'arbre (trois lignes par nœud, l'une pour la division apparaissant sur l'arbre, et les deux pour les divisions de substitution basées sur les variables cachées) les informations suivantes :

- la variable et le seuil composant les règles de division.
- l'ordre de la division en tant que division concurrente, donc par rapport à la diminution de la déviance qu'elle provoque.
- la mesure d'association entre chaque division et la division apparaissant au nœud
- la diminution de la déviance due à cette division.

Les variables cachées présentent des règles de divisions très compétitives bien placées au niveau des premiers nœuds de l'arbre (les lignes du tableau correspondant à ces nœuds sont ombrées) En particulier aux nœuds numérotés 1 et 6, les deux variables cachées forment les divisions les plus concurrentes.

---

au découpage d'un cercle en 9 secteurs de 40° dans le ses des aiguilles d'une montre, le "a" correspondant au Nord.

noeud	Variable	Ordre	seuils	Association	Déviance
1	maxo3v	1	100,5	1,00	61858,38
	Maxgrd	2	-0,80	0,23	29909,55
	dv18	3	abghi	0,18	29016,73
2	tc18	1	26,2	1,00	9751,37
	dv18	19	a	-0,10	762,46
	Maxgrd	39	3	-0,05	0,01
3	maxo3v	1	88,5	1,00	5198,30
	Maxgrd	2	-0,05	0,05	2139,29
	dv18	8	abcdeghi	0,02	1622,48
4	tc6	1	18,8	1,00	3216,00
	dv18	10	g	0,78	778,89
	Maxgrd	19	-2,15	0,73	312,02
5	neb3	1	0,5	1,00	4116,68
	dv18	6	Fh	0,44	3100,83
	Maxgrd	13	-1,15	0,33	1750,67
6	neb6	1	2,5	1,00	14989,75
	dv18	2	abdghi	0,15	12857,48
	Maxgrd	3	-1,45	0,32	12519,71
7	maxo3n	1	90,5	1,00	9733,18
	Maxgrd	7	2,65	0,00	3598,45
	dv18	8	adefghi	0,07	3235,59
8	maxn2n	1	29	1,00	5221,50
	Maxgrd	2	-1,05	0,48	4452,04
	dv18	8	hi	0,12	2697,36
9	dv6	1	adi	1,00	5715,64
	Maxgrd	5	-0,05	0,82	3097,79
	dv18	22	efi	0,82	1127,52
10	maxo3v	1	116,5	1,00	5377,52
	dv18	5	agh	0,32	3252,79
	Maxgrd	13	-0,85	0,27	974,26
11	dv15	1	ach	1,00	3511,24
	dv18	26	A	0,00	234,51
	Maxgrd	39	-2,55	-0,11	2,12

Tableau 2 : Pour chaque nœud on donne les variables actives et les deux variables cachées ("dv18" et "maxgrd"), avec l'ordre des divisions concurrentes basées sur celles ci, la diminution de la déviance que ces divisions provoquent, et la mesure d'association avec la division basée sur la variable active.
Pour les seuils sur la direction du vent, par exemple le nœud n°9, la première division se lit : "dv6" égale à "a"(Nord) ou "d" (Sud Est) ou "i" (Nord Ouest)

3.6. Traitement des données manquantes

Supposons disponible une nouvelle observation x_l où une des composantes x_l^m par exemple est une donnée manquante et pour laquelle on voudrait prédire la classe j_l. Supposant qu'en acheminant cette observation dans l'arbre on atteint un nœud t avec une division d^* basée sur la composante manquante x_l^m de l'observation. Donc la règle apparaissant au niveau de ce nœud ne permet pas l'acheminement de cette observation vers les nœuds descendants. On peut dans ce cas utiliser une division concurrente à d^* basée sur une autre variable dont l'observation est disponible.

4. Stabilité de l'importance des variables

Diverses études empiriques (Breiman 1996a , Ghattas 1999b) ont montré que les arbres de régression et de classification sont instables par rapport à des légères perturbations des données. Cette instabilité a été mise en évidence par les changements de la structure de l'arbre et par les prévisions qu'il fournit dans Ghattas (1999b).

Nous nous intéressons ici à la stabilité de l'importance des variables vis à vis de perturbations de l'échantillon de construction. Pour ce faire, nous avons généré 40 échantillons bootstrap des données disponibles à la station de Vitrolles, sur lesquels nous avons construit des arbres de taille fixe (10 feuilles).

Pour ces 40 arbres nous avons calculé l'importance des variables. Sur ces 40 essais, nous avons regardé la position de chacune des 39 variables utilisées, dans la hiérarchie établie à partir des différents arbres. Les résultats sont donnés sous forme de Boxplot (figure 2) construit pour chaque variable à partir des mesures de son importance pour les 40 arbres.

Le maximum d'ozone observé la veille est en tête dans 39 cas sur 40 ; les autres variables importantes : gradient, direction du vent à 18 heures, nébulosité et température ont une dispersion non négligeable. La variable la plus importante maintient sa position malgré la perturbation de données. Pour les variables intermédiaires la hiérarchie semble être instable, alors que les variables les moins importantes semblent conserver leur position (cf. figure 2).

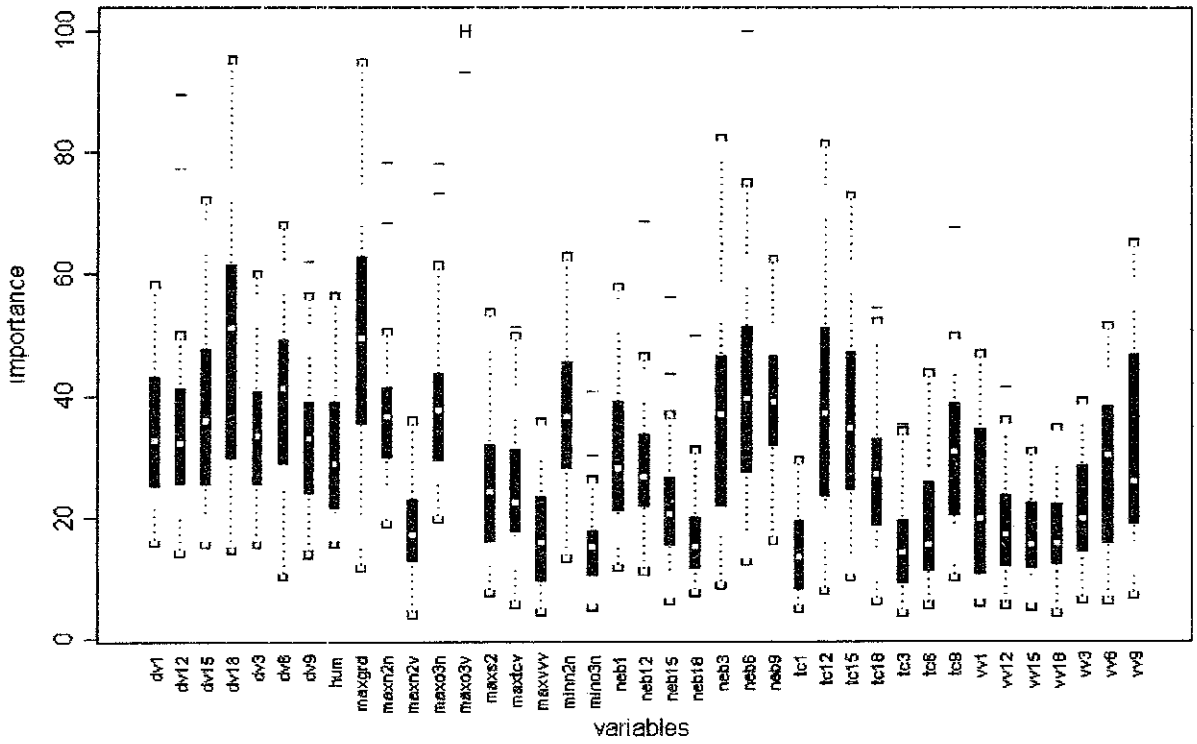

Figure 2 : boxplot des importances de chaque variable, obtenues à partir de 40 arbres construits sur des échantillons bootstrap.

Si on s'intéresse à un modèle qui est composé d'agrégation d'arbres par bootstrap (Ghattas 1999a), il est aussi possible d'utiliser la hiérarchie des variables. Dans ce cas, supposons qu'on agrége K arbres de régression ou de classification; on calcule la hiérarchie des variables sur ces K arbres, et pour chaque variable on additionne les importances obtenues dans les différentes hiérarchies. Il suffit ensuite de normer les valeurs obtenues pour avoir une hiérarchie associée au modèle agrégé. Le tableau 3 donne les résultats obtenus pour 40 arbres de régression agrégés construits sur la station d'Istres. La hiérarchie obtenue dans ce cas est plus stable que celle obtenue pour un arbre simple (comme celle par exemple du tableau1).

maxo3v	maxgrd	dv18	dv6	Neb6	neb9	dv15	neb3	dv12	tc12
100	53	47	42	42	40	40	39	38	38

Tableau 3 : hiérarchie des variables pour 40 arbres agrégés,les dix variables les plus importantes avec leur indice d'importance normé.

5. Implémentation dans S+

Le logiciel *S+* offre la possibilité de construire des arbres de régression et de classification (fonction *tree*) , avec un large choix de paramètres et une interface graphique interactive. Il est

possible par exemple de chercher la taille optimale des arbres par validation croisée (*cv.tree*), d'élaguer des arbres de manière interactive (*prune.tree*) et d'analyser le contenu des nœuds et des feuilles par rapport à toutes les variables utilisées (*tile.tree*).

Cependant quelques options ne sont pas encore intégrées dans S+. Il est par exemple impossible d'avoir les divisions de substitution alors que les divisions concurrentes sont disponibles (*burl.tree*). Bien qu'un argument permettant d'intégrer des poids soit disponible dans quelques fonctions, cet argument n'est pas encore utilisé dans tous les calculs fournis par ces fonctions.

Il existe une librairie « *rpart* » réalisée par T. Thernau et E. Atkinson qui complète les fonctions de S+. Elles permettent par exemple d'obtenir les règles de substitution au niveau de chaque nœud de l'arbre, et d'obtenir différents types de tracés d'arbres, ou encore d'utiliser différents critères dans la construction des arbres. Mais elles restent incomplètes, les mesures d'association entre les divisions ainsi que l'importance des variables y manquent

Parmi les fonctions permettant de compléter les outils de manipulation des arbres sous S+ nous proposons la fonction « *impvar* » qui permet d'obtenir directement les divisions de substitution et la hiérarchie des variables. Elle est disponible avec une documentation complète sur simple demande par courrier électronique. Ultérieurement cette fonction sera mise à la disposition du public sur le serveur *Statlib*.

Références

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

Breiman L.(1996a), Heuristic of instability and stabilization in model selection, *Annals of Statistics*, Vol 24, N°6, pp. 2350-2383.

Breiman L. (1996b), Bagging Predictors, *Machine Learning*, 24, 123-140.

Breiman L. (1996c), Technical Note: Some Properties of Splitting Criteria, *Machine Learning*, 24, 41-47.

Casmassi J., (1998) Comparison of meteorological criteria characterizing 1-hour and 8-hour average ozone episodes in the south coast air basin. *Technical report,* South Coast Air Quality Management Department CA.

Deniau C., Ghattas B., (1998) Essais de Classification des stations de mesure de l'ozone des réseaux AIRMARAIX et AIRFOBEP. *Rapport de contrat GREQAM-AIRMARAIX*.

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

Ghattas B., (1999b), Agrégation d'arbres de classification. A paraître dans la *Revue de Statistique Appliquée*.

Gueguen A., Nakache J.-P. (1988), Méthode de discrimination basée sur la construction d'un arbre binaire, *Revue de Statistique Appliquée* XXXVI (1), 19-37.