REVUE DE STATISTIQUE APPLIQUÉE

A. CIAMPI
J. THIFFAULT
U. SAGMAN

Évaluation de classifications par le critère d'Akaike et la validation croisée

Revue de statistique appliquée, tome 36, n° 3 (1988), p. 33-50
<http://www.numdam.org/item?id=RSA_1988__36_3_33_0>

Rev. Statistique Appliquée, 1988, **XXXVI** (3), 33-50

ÉVALUATION DE CLASSIFICATIONS PAR LE CRITÈRE D'AKAIKE ET LA VALIDATION CROISÉE

A. CIAMPI (1); J. THIFFAULT (1); U. SAGMAN (2)

(1) Hôpital de Montréal pour Enfants,
4060 Ste-Catherine ouest, Montréal, Québec, Canada, H3Z 2Z3
(2) Ontario Cancer Institute,
500 rue Sherbourne, Toronto, Ontario, Canada, M4X 1K9

RÉSUMÉ

Les méthodes de classification représentent de puissants outils pour la formulation d'hypothèses. Il semble donc important de développer un cadre théorique pour jauger l'adéquation des hypothèses correspondant à différentes classifications. Nous proposons une correspondance entre classification et modèle statistique. La statistique classique suggère alors d'utiliser la vraisemblance comme mesure d'adéquation. Nous montrons ici que cette mesure est biaisée et que le biais peut être réduit à l'aide de la validation croisée. Pour des grands ensembles de données, cela est équivalent à substituer à la vraisemblance le critère d'information d'Akaike. La méthode de classification **SEGMAG** et utilisée comme exemple de l'application du schémé proposé.

Mots clés : *Données de survie, Formulation d'hypothèses, Plausibilité d'une classification.*

ABSTRACT

Classification techniques are frequently used as methods of hypothesis generation. It is therefore important to develop a framework to measure the relative plausibility of hypotheses corresponding to different classifications. In this paper, we outline a framework relating classifications to statistical models. Classical statistics recommend the use of likelihood as a measure of plausibility. We will demonstrate here the biasedness of this measure, and the fact that the bias can be reduced with the help of cros-validation. For large data sets, this approach is equivalent to replacing the likelihood by the Akaike Information Criterion. The classification algorithm **SEGMAG** is presented to demonstrate the application of the proposed framework.

Keywords: *Survival data, Hypothesis formulation, Plausibility of a classification.*

1. Introduction et Schéma Général

Les méthodes de classification constituent de puissants outils pour la formulation d'hypothèses basées sur l'examen d'un ensemble de données (Jambu et Lebeaux, 1983; Gordon, 1981). En général, le but d'un algorithme de classification est

d'identifier des classes d'individus qui se ressemblent selon un paramètre spécifié par l'utilisateur et calculé à partir des variables mesurées sur ces individus. En étudiant les valeurs des variables et leurs liaisons selon les classes obtenues, le chercheur arrive souvent à formuler des conjectures sur les phénomènes sous-jacents aux données et sur les processus qui les génèrent. Ce travail d'interprétation est d'autant plus fructueux que la description des classes obtenues est directe. De fait, plusieurs méthodes de classification ont été développées dans le but précis de faciliter cette étape. Par exemple, la méthode des nuées dynamiques a comme caractéristique principale d'aboutir en même temps à la définition des classes et à leur description. Dans cette optique, l'approche logique en classification (Sidi, 1980) apparaît particulièrement prometteuse puisqu'elle permet une interprétation des classes à l'aide d'un langage simple formé de propositions logiques portant sur les variables observées.

Cependant, la formulation d'hypothèses en classification demeure informelle et qualitative. Cela n'est pas nécessairement une limitation; au contraire, ce n'est qu'à travers l'engagement de la créativité de l'analyste qu'on pourra découvrir des structures intéressantes et les mettre en relation avec les connaissances scientifiques pertinentes. Toutefois, il existe une grande variété de méthodes de classification. Pour un même ensemble de données, les résultats varient fréquemment d'une méthode à l'autre, ou dépendent même des options spécifiées pour une méthode donnée. Ce phénomène reflète soit l'instabilité des méthodes considérées, soit une multiplicité réelle d'hypothèses compatibles avec les données. Il apparaît dont important de disposer, comme guide à l'interprétation, d'une mesure d'adéquation relative des différentes hypothèses formulées.

Pour sa part, la statistique inférentielle permet d'attacher une mesure de validité à une hypothèse clairement et rigoureusement spécifiée a priori. La théorie de Neyman-Pearson, chère à la pensée et à la pratique d'un grand nombre de statisticiens en faveur de l'approche dite classique, ne permet que d'accepter ou de rejeter une hypothèse donnée. L'approche fishérienne, pour sa part, est un peu plus flexible; selon Efron (1982a), elle permet de considérer un modèle statistique comme un *résumé* du contenu informatif d'un ensemble de données. Le modèle statistique devient alors moins l'expression d'une hypothèse précise qu'un outil pour *lisser* les données, tel un filtre capable de mettre en relief les structures les plus importantes tout en éliminant les moindres détails. Dans cet article, nous montrerons que la synthèse de ce dernier point de vue avec les méthodes de classification fournit un puissant outil à la formulation d'hypothèses statistiques.

Nous nous limiterons à la situation suivante. Pour chaque individu d'une population donnée, on mesure un certain nombre de variables. A partir de ces données, on veut arriver à formuler des hypothèses sur le lien entre, d'une part, un parmètre d'intérêt particulier, $\underline{\gamma}$, qu'on appelle *critère* et de l'autre un vecteur de caractéristiques de l'individu, $\underline{z}$, qu'on appellera *variables de prédiction ou prédicteurs*. Le critère $\underline{\gamma}$ sera une fonction d'un autre vecteur de caractéristiques de l'individu, $\underline{u}$, appelées *variables réponse*. On atteindra notre but en formant des classes d'individus caractérisées à l'aide des prédicteurs et telles que, à l'intérieur de chaque classe, le critère soit homogène. Un tel résultat est d'interprétation immédiate, en ce qu'il montre que certaines configurations des prédicteurs déterminent la valeur de $\underline{\gamma}$ à une variation résiduelle près, laquelle pourra être interprétée soit comme un bruit,

soit comme liée à des variables inconnues ou non mesurées.

Tel qu'indiqué dans la figure 1, le point de départ de l'analyse est une matrice de données X partagée en deux blocs U et Z. A chaque individu on associe une ligne $\underline{x} = (\underline{u} \mid \underline{z})$ de la matrice X, dont les éléments constituent les valeurs que le vecteur $\underline{x}$ prend pour cet individu. Remarquons que le bloc Z peut être vide; en revanche, une variable peut apparaître à la fois comme variable réponse et de prédiction.

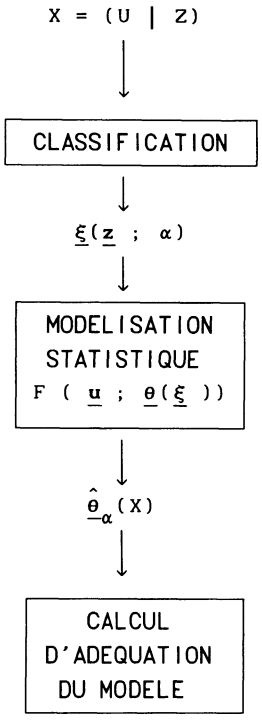

Figure 1 : Schéma général

Le bloc (1) de la figure 1 représente la mise en œuvre d'un algorithme de classification quelconque. Les choix faits au cours de la définition de l'algorithme sont représentés par le vecteur α des *paramètres de la classification*. La sortie de l'algorithme est la matrice $(U \mid \Xi)$, où le bloc Ξ, qui a pris la place de Z, est en général beaucoup plus petit que ce dernier; ses lignes représentent les valeurs prises sur l'ensemble des individus par le vecteur des *variables indicatrices* des classes obtenues, soit $\underline{\xi} = \underline{\xi}(\underline{z}; \alpha)$. Le résultat de l'algorithme de classification peut donc être vu comme la construction d'une application non-linéaire $\underline{z} \longrightarrow \underline{\xi}$ ayant pour effet une importante réduction des données.

Exemple 1 : Classification hiérarchique

Tout algorithme de classification ascendante hiérarchique aboutit à la construction d'un arbre d'agrégation. Dans la plupart des cas, le critère est tout simplement le vecteur des moyennes des variables mesurées sur les individus et il n'y a pas de prédicteurs : $\underline{x} = \underline{u}$. Dans ce cas, le vecteur $\underline{\xi} = \underline{\xi}(\alpha)$ ne dépend que d'un paramètre scalaire α qui indique le palier auquel l'arbre d'agrégation a été coupé. Par ailleurs, si l'on fait varier la règle de construction de l'arbre et qu'on considère, pour chaque hiérarchie, différents paliers de coupure, le paramètre de la classification sera bidimensionnel : $\alpha = (k,p)$, où k indique le choix de la règle de construction et p le palier de coupure

Exemple 2 : Régression locale

L'algorithme de régression locale (Diday, 1980) est un cas particulier de la méthode des nuées dynamiques. Il permet de classifier les individus en groupes qui sont homogènes du point de vue du vecteur des coefficients de la régression d'une variable scalaire y sur un vecteur $\underline{x}$. Pour établir le lien entre cet exemple et notre schéma, notons que $\underline{u} = (y,\underline{x})$ et que notre γ est le vecteur de régression. Dans l'approche de Diday (1980), il n'y a pas de prédicteurs. Nous montrerons que dans l'algorithme présenté dans la section 3, il est possible de traiter aussi la situation où des variables $\underline{z}$ servent à prédire γ.

Pour chaque choix du vecteur des paramètres de la classification, α, les données suggèrent une hypothèse représentée par le vecteur des variables indicatrices des classes, $\underline{\xi}(\underline{z}\,;\,\alpha)$. Nous noterons ce vecteur sous forme abrégée, $\underline{\xi}_\alpha$, et par un abus de langage qui ne devrait pas provoquer de confusion, $\underline{\xi}_\alpha$ désignera soit la partition, soit l'hypothèse représentée par les variables indicatrices. Sans aucune perte de généralité, nous supposerons que la famille $\left\{\underline{\xi}_\alpha\right\}_\alpha$ renferme l'hypothèse $\underline{\xi}_0$ selon laquelle l'ensemble des individus est homogène du point de vue du paramètre (critère) $\underline{\gamma}$.

Le bloc (2) de la figure 1 représente l'étape qui consiste à associer à chaque hypothèse $\underline{\xi}_\alpha$ un modèle statistique, ou plus précisément une fonction de répartition $F\left\{\underline{u}\,;\,\underline{\theta}\mid\underline{\xi}_\alpha\right\}$, connue à un paramètre près, soit $\underline{\theta}$, qui sera estimé à partir des données. C'est au niveau de l'explicitation de l'application

$$\underline{\xi}_\alpha \quad \longrightarrow \quad F\left\{\underline{u}\,;\,\underline{\theta}\mid\underline{\xi}_\alpha\right\} \tag{1.1}$$

que l'analyste fera appel à la modélisation statistique.

Par souci de simplicité, nous nous limiterons dans cet article à la situation où

$$F\left\{\underline{u}\,;\,\underline{\theta}\mid\underline{\xi}_\alpha\right\} = F\left\{\underline{u}\,;\,\underline{\theta}(\underline{\xi}_\alpha)\right\} \tag{1.2}$$

et

$$\underline{\theta}(\underline{\xi}_\alpha) = \left[\underline{\theta}_0,\,\underline{\gamma}(\underline{\xi}_\alpha)\right]\,. \tag{1.3}$$

Ici $\underline{\theta}$ est partagé en deux sous-vecteurs; le premier, $\underline{\theta}_0$, qui ne dépend pas de $\underline{\xi}_\alpha$, et l'autre, $\underline{\gamma}(\underline{\xi}_\alpha)$, qui représente le critère et sa dépendance par rapport aux classes de la partition $\underline{\xi}_\alpha$. Nous supposerons aussi $\underline{\gamma}(\underline{\xi}_\alpha)$ de forme linéaire :

$$\underline{\gamma}(\underline{\xi}_\alpha) = \Gamma \underline{\xi}_\alpha , \tag{1.4}$$

où Γ est une matrice de paramètres. On a ainsi introduit un modèle d'analyse de variance multidimensionnelle généralisée qui sera assez souple pour accomoder plusieurs situations rencontrées dans la pratique.

Exemple 3 : Analyse de variance

Dans ce cas, supposons que $\underline{u}$ est une variable aléatoire dont la distribution est normale multivariée. Supposons également que la matrice de covariance Σ de $\underline{u}$ soit homogène sur toute la population, mais que $\gamma = E(\underline{u})$ soit susceptible de varier d'une classe à l'autre. On pourra alors interpréter $\underline{\theta}_0$ de (1.3) comme le vecteur des paramètres nécessaires pour définir Σ et (1.4) comme l'explicitation de la dépendance de l'espérance de $\underline{u}$ par rapport aux classes obtenues. L'hypothèse $\underline{\xi}_0$ correspond à l'hypothèse de l'homogénéité globale, soit à l'hypothèse "nulle" de l'analyse de variance classique.

Le dernier bloc du diagramme, (3), représente l'étape du calcul d'une mesure d'adéquation de chaque hypothèse $\underline{\xi}_\alpha$, soit de la distribution (1.2). La statistique inférentielle classique suggère d'utiliser la *vraisemblance* $V(\underline{\theta}_\alpha \mid X)$ comme base pour la définition de cette mesure. Nous remettons le développement de ces considérations à la prochaine section. Dans la section 3, nous décrirons à titre illustratif une méthode de classification qui intègre les stratégies classiques de segmentation et d'agrégation. Cette méthode développée récemment par Ciampi *et al.* (1987) a été conçue dans l'esprit du schéma général énoncé ci-haut. Enfin, la section 4 propose une exemple d'analyse de données cliniques visant à obtenir une classification pronostique.

2. Évaluation des classifications et choix de modèle

Pour simplifier la notation, posons

$$\underline{\theta}_\alpha = \underline{\theta}\left(\underline{\xi}_\alpha\right) \qquad \text{et} \qquad V_\alpha = V(\underline{\theta}_\alpha \mid X) ,$$

où V_α dénote la vraisemblance de la classification $\underline{\xi}_\alpha$ basée sur les données X.

Considérons ensuite la statistique

$$\Lambda_\alpha = 2 \ln \left\{ \frac{V_\alpha}{V_0} \right\} . \tag{2.1}$$

On sait (Cox & Hinkley, 1974) que Λ_α constitue une mesure raisonnable de l'adéquation de la classification $\underline{\xi}_\alpha$, puisqu'elle exprime l'adéquation de la structure

décrite par la partition $\underline{\xi}_\alpha$ vis-à-vis de l'absence de structure décrite par $\underline{\xi}_0$. De même, étant données deux hypothèses $\underline{\xi}_{\alpha_1}$ et $\underline{\xi}_{\alpha_2}$, l'adéquation relative de la première vis-à-vis de la seconde peut être convenablement mesurée par

$$\Lambda_{\alpha_1:\alpha_2} = \Lambda_{\alpha_1} - \Lambda_{\alpha_2} = 2 \ln \left\{ \frac{V_{\alpha_1}}{V_{\alpha_2}} \right\} \tag{2.2}$$

Or, pour calculer Λ_α, il faudrait connaître le paramètre $\underline{\theta}_\alpha$, supposé inconnu, et on serait alors tenté de substituer à $\underline{\theta}_\alpha$ son estimateur du maximum de vraisemblance (EMV), $\widehat{\underline{\theta}}_\alpha(X)$, donc d'estimer Λ_α par

$$\widehat{\Lambda}_\alpha = 2 \ln \left\{ \frac{\widehat{V}_\alpha}{\widehat{V}_0} \right\}, \tag{2.3}$$

où $\widehat{V}_\alpha = V\left(\widehat{\underline{\theta}}_\alpha(X) \mid X\right)$. Cependant cette mesure serait biaisée puisqu'elle favoriserait les partitions ayant un grand nombre de classes. Par exemple, si la partition $\underline{\xi}_{\alpha_1}$ est une partition incluse dans $\underline{\xi}_{\alpha_2}$, on a par définition, $\Lambda_{\alpha_1:\alpha_2} \geq 0$. Le biais est une conséquence du fait que les mêmes données sont utilisées à la fois pour estimer $\underline{\theta}_\alpha$ et pour évaluer la classification correspondante. Le modèle $\widehat{\underline{\theta}}_\alpha(X)$ obtenu est alors trop "proche" des données observées X pour permettre d'ajuster d'autres données potentielles engendrées par le même phénomène. On aurait donc un prix à payer pour l'avantage apparent d'un modèle $\underline{\theta}_\alpha$ qui ajuste bien les données X; le prix serait une perte de reproductibilité dont l'importance est difficile à jauger et une impossibilité d'aboutir à une extension des résultats au-delà des individus étudiés.

Pour remédier à cette difficulté, plusieurs techniques de rééchantillonage ont été proposées, telles que le *bootstrap*, le *jackknife* (Efron, 1982b) et la *validation croisée* (Stone, 1974), dont l'application à l'analyse factorielle a été récemment étudiée en détail par Junca (1985). Pour une raison d'espace, nous ne discuterons que de la dernière, tout en remarquant que celle-ci est équivalente aux deux autres pour des ensembles de données de grande taille.

Soit $X^{(j)}$ la matrice obtenue à partir de X en supprimant la ligne $\underline{x}_j$, et soit $\widehat{\underline{\theta}}_\alpha^{(j)} = \widehat{\underline{\theta}}_\alpha\left(X^{(j)}\right)$, l'EMV de $\underline{\theta}_\alpha$ qui réalise le maximum de la fonction de vraisemblance $V\left(\underline{\theta}_\alpha \mid X^{(j)}\right)$. Pour chaque j, posons

$$\widehat{\Lambda}_\alpha^{(j)} = 2 \ln \left\{ \frac{V\left(\widehat{\underline{\theta}}_\alpha^{(j)} \mid X^{(j)}\right)}{V\left(\widehat{\underline{\theta}}_0^{(j)} \mid X^{(j)}\right)} \right\} \tag{2.4}$$

qui correspond à une estimation de (2.1) basée sur le sous-ensemble de X ne contenant pas le j$^{\text{ème}}$ individu.

Dans l'approche de la validation croisée, on prend la moyenne

$$\overline{\Lambda}_\alpha = \frac{1}{N} \sum_j \widehat{\Lambda}_\alpha^{(j)} \tag{2.5}$$

comme mesure d'adéquation pour $\underline{\xi}_\alpha$, où N dénote le nombre total d'observations; l'écart-type

$$\mathrm{ERT}_\alpha = \sqrt{\frac{1}{N(N-1)} \sum_j \left(\hat{\Lambda}_\alpha^{(j)} - \overline{\Lambda}_\alpha\right)^2} \tag{2.6}$$

est utilisée comme mesure de la dispersion de $\overline{\Lambda}_\alpha$.

On connaît une très bonne approximation pour $\overline{\Lambda_\alpha}$ qui évite les calculs assez laborieux requis par (2.4) et (2.5). Elle est donnée (Stone, 1977; Ciampi *et al.*, 1986) par

$$N\overline{\Lambda}_\alpha \cong \hat{\Lambda}_\alpha - 2(p_\alpha - p_0), \tag{2.7}$$

ou p_α et p_0 sont les nombres de paramètres requis pour spécifier $\underline{\theta}_\alpha$ et $\underline{\theta}_0$ respectivement. Or (Akaike, 1974),

$$\hat{\Lambda}_\alpha - 2(p_\alpha - p_0), = \mathrm{AIC}\left(\hat{\theta}_0\right) - \mathrm{AIC}\left(\hat{\theta}_\alpha\right) \tag{2.8}$$

où AIC dénote le Critère d'Information d'Akaike défini comme suit :

$$\mathrm{AIC}\left(\hat{\theta}_\alpha\right) = -2 \ln\left(\hat{V}_\alpha\right) + 2p_\alpha, \tag{2.9}$$

ce qui implique que la classification la plus adéquate, c'est-à-dire $\underline{\xi}_{\alpha^*}$ telle que $\overline{\Lambda}_{\alpha^*} \geq \overline{\Lambda}_\alpha$ pour tout α, est aussi la classification correspondant à l'AIC minimum.

L'avantage des formules (2.4)-(2.6) sur (2.8)-(2.9) est qu'on peut estimer la dispersion de $\overline{\Lambda}_\alpha$, ce qui peut s'avérer très utile quand l'analyste veut identifier non pas une seule classification, mais plusieurs qui seraient compatibles avec les données. En outre, si pour une partition $\underline{\xi}_\alpha$ on a

$$\frac{1}{N}\left|\overline{\Lambda}_\alpha - \overline{\Lambda}_{\alpha^*}\right| \leq \mathrm{ERT}_{\alpha^*}$$

on pourra considérer $\underline{\xi}_\alpha$ comme aussi adéquate que ξ_{α^*} et la préférer à cette dernière si elle compte moins de classes terminales.

Sur la base de ces considérations, nous proposons la procédure suivante pour évaluer les classifications de la famille $\left\{\underline{\xi}_\alpha\right\}_\alpha$ et pour identifier celles qui sont compatibles avec les données :

1. Pour chaque $\underline{\xi}_\alpha$, calculer $\hat{V}_\alpha$ et AIC $(\hat{\theta}_\alpha)$;
2. Choisir α^* qui minimise l'AIC;
3. Calculer $\underline{\Lambda}_{\alpha^*}$ et ERT_{α^*};
4. **Règle d'1 Γ ERT** : retenir comme partitions adéquates toutes celles qui satisfont l'inégalité

$$\left|\mathrm{AIC}\left(\hat{\underline{\theta}}_\alpha\right) - \mathrm{AIC}\left(\hat{\underline{\theta}}_{\alpha^*}\right)\right| \leq \mathrm{ERT}_{\alpha^*} \tag{2.10}$$

5. S'il faut choisir une seule partition pour représenter les données, donner préférence à la plus simple.

Le critère de simplicité peut être établi formellement comme étant le choix systématique de la partition ayant le plus petit nombre de classes terminales. Une autre approche, dans le cas où les prédicteurs jouent un rôle important, consiste à privilégier les partitions dont la description en termes des prédicteurs est la plus simple possible. Par exemple, dans le cadre de la méthode de la section suivante, nous montrerons qu'il est possible d'ajuster un paramètre de *calibration* de manière à considérer des classes définies par des propositions booléennes portant sur plusieurs variables à la fois, ces propositions étant d'autant plus simples que le paramètre de calibration est large.

Il peut toutefois s'avérer avantageux de ne pas imposer le critère de simplicité, ceci afin de permettre au spécialiste du phénomène étudié de choisir lui-même la partition dont le bien-fondé biologique ou scientifique est le plus grand.

L'importance de la règle décrite ci-dessus réside dans le fait qu'elle permet de réduire considérablement la multiplicité des modèles acceptables. Par la suite, les échanges conversationnels entre l'analyste et le spécialiste devraient mener soit au choix d'un modèle unique pour représenter les données, soit à la conclusion qu'un certain nombre d'hypothèses différentes sont compatibles avec les données.

3. SEGMAG (RECPAM) : une méthode de classification pour données médicales

La méthode de classification **SEGMAG** (**RECPAM** en anglais), développée récemment par Ciampi *et al.* (1987), intègre les stratégies de **SEGM**emtation et d'**AG**régation de Breiman *et al.* (1984) dans un contexte biomédical. Elle tient donc compte de trois situations rencontrées en biostatistique : (1) variable réponse censurée, (2) variable réponse nominale, (3) analyse par sous-groupe de la régression d'une variable censurée. Nous ne décrirons que brièvement l'algorithme utilisé dans **SEGMAG**; on trouvera dans Ciampi *et al.* (1987) une représentation plus détaillée. Il importe toutefois de mentionner les principaux avantages qui font que cette méthode est bien adaptée à l'étude de données cliniques. D'une part, la technique de partitionnement utilisée dans **SEGMAG** aboutit à la formation de classes dont la description est d'interprétation clinique immédiate. On ne saurait trop insister sur ce point puisqu'il favorise le dialogue statisticien/clinicien essentiel au progrès de la science. D'autre part, certaines adaptations des techniques de régression aux données médicales négligent de considérer l'interaction possible entre les variables de prédiction. Comme nous le montrerons dans cette section, l'algorithme de **SEGMAG** tient compte de telles interactions et permet même à l'analyste d'en contrôler le degré de complexité.

Comme dans tout algorithme de classification, la notion de dissimilarité entre deux populations est cruciale. Considérons deux sous-populations P_1 et P_2 de la population P, qu'on désire comparer du point de vue du paramètre $\underline{\theta}_\alpha$ (contenant le critère), tel que décrit dans la section précédente, et la matrice de données correspondante $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$. On peut définir la dissimilarité entre P_1 et P_2 (par

rapport à $\underline{\theta}_\alpha$) sur la base du rapport de vraisemblance

$$d(P_1, P_2) = 2 \ln \left\{ \frac{V\left(\hat{\underline{\theta}}_1(X_1) \mid X_1\right) V\left(\hat{\underline{\theta}}_2(X_2) \mid X_2\right)}{V\left(\hat{\underline{\theta}}(X) \mid X\right)} \right\},$$

où $\hat{\underline{\theta}}_j(X_j)$ représente l'estimateur du maximum de vraisemblance de $\underline{\theta}_j$ estimé sur le sous-ensemble X_j correspondant à la sous-population P_j.

Étant donnée cette définition de la dissimilarité entre deux sous-populations, l'algorithme de **SEGMAG** partitionne les individus de telle sorte qu'on maximise, à chaque étape du partitionnement, la dissimilarité entre les deux sous-populations obtenues.

On doit aussi définir une classe $\mathcal{C}$ de propositions telle que toute proposition $Q \in \mathcal{C}$ induit une partition de la population P en deux segments disjoints, P_1 et P_2. Soit le vecteur $\underline{z}$ des variables de prédiction d'un individu, on considèrera que la sous-classe $\mathcal{C}_S$ composée de propositions de la forme

$$Q(\underline{z}) = \{ z_j \mid z_j \in A_j \}, \tag{3.1}$$

appelées *propositions simples*, où z_j est un élément du vecteur $\underline{z}$ et A_j représente un sous-ensemble des modalités possibles de la variable z_j. Tous les individus dont la valeur de z_j appartient au sous-ensemble A_j défini par $Q(\underline{z})$ sont affectés à P_1 tandis que les autres forment le segment P_2.

La proposition $Q_1 = \{\text{Age} > 65\}$, que l'on retrouve au premier palier de l'arbre de la figure 2(a), constitue un exemple de proposition simple.

Afin de tenir compte des interactions possibles entre variables de prédiction, on étendra la recherche d'une partition acceptable à la sous-classe $\mathcal{C}_B$ composée de propositions booléennes telles que

$$Q(\underline{z}) = \{ [z_{j_1} \mid z_{j_1} \in A_{j_1}] \wedge [z_{j_2} \mid z_{j_2} \in A_{j_2}] \wedge \ldots \wedge [z_{j_k} \mid z_{j_k} \in A_{j_k}] \}.$$

L'arbre illustré à la figure 2(b) montre le résultat d'une analyse tenant compte des interactions possibles entre les variables; on y observe une partition induite par la propostion booléenne suivante

$$Q_2 = \{[\text{AGE} > 65] \wedge [\text{SEXE} = M]\}.$$

Dans ce cas, tous les hommes âgés de 65 ans et plus seront assignés à la sous-population P_1, tandis que les autres formeront le segment P_2.

Un paramètre de calibration, $0 \leq \beta \leq 1$, permet de contrôler la complexité des propositions considérées. Une valeur de β s'approchant de 1 restreint la recherche à des propositions simples telle qu'illustrée à la figure 2(a). En diminuant la valeur de β, on permet à un nombre de variables croissant de définir une partition, tel que montré dans la figure 2(b). La définition précise du paramètre β est donnée dans Ciampi *et al.* (1987).

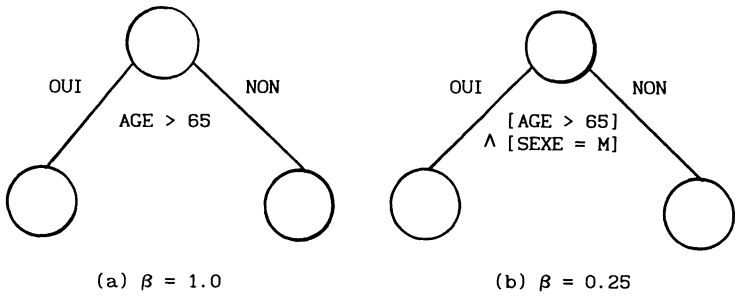

Figure 2 : Effet du paramètre de calibration β

A chaque étape de la construction de l'arbre, on considère toutes les propositions $Q_j \in \mathcal{C}$, et pour chacune on calcule la dissimilarité $\widehat{\Lambda}_j$ entre les deux sous-populations qu'elle induit. La *meilleure* proposition est celle qui maximise le rapport de vraisemblance $\widehat{\Lambda}_j$, c'est-à-dire celle qui maximise la dissimilarité entre les deux sous-populations obtenues.

Au point de départ de l'analyse, on considère tous les individus comme membres d'une même population P. On dispose d'autre part d'une description multidimensionnelle de chaque individu, sous la forme d'un vecteur de variables nominales. (Dans le cas de données continues, on doit avoir recours à un recodage préalabe par catégories). L'algorithme cherche d'abord la *meilleure* proposition Q_1, c'est-à-dire celle qui sépare le mieux la population entière P en deux sous-populations distinctes, P_1 et P_2. Par convention, P_1 regroupe l'ensemble des individus satisfaisant la proposition Q_1.

Le même processus de segmentation est ensuite appliqué aux deux segments obtenus P_1 et P_2, et ainsi de suite jusqu'à ce que la population de chaque segment devienne plus petite qu'un effectif minimun fixé à l'avance. Il sera cependant possible d'établir un seuil d'arrêt α, qui contrôlera le nombre de segments terminaux de l'arbre. Dans ce cas, une partition (P_1, P_2) induite par une proposition Q_j sera admissible dans la mesure où la dissimilarité entre P_1 et P_2 sera significative au niveau nominal α; sinon le processus de segmentation se terminera. Le critère de simplicité énoncé à la section précédente amènera donc l'analyste à privilégier des valeurs moindres d'α et un β élevé.

Il faut remarquer que la méthode de segmentation utilisée dans **SEGMAG** ne constitue pas une méthode globalement optimale, c'est-à-dire que l'arbre final $\underset{\sim}{\xi}_\alpha$ ne maximise pas nécessairement la vraisemblance sur l'ensemble de tous les choix possibles de $\underset{\sim}{\xi}_\alpha$.

De par la nature même de l'algorithme de segmentation, il est possible que des segments terminaux provenant d'*ancêtres* différents soient identiques du point de vue du critère. L'algorithme d'agrégation de **SEGMAG** permet de corriger cette situation en joignant ensemble les segments qui se ressemblent le plus du point de vue du critère. Un paramètre α' contrôle l'agrégation de la même façon que pour

l'algorithme de segmentation. Il est à noter que le paramètre α' peut être défini comme étant égal à α, mais il peut prendre une valeur différente de celle utilisée pour la segmentation.

4. Un exemple d'application

Dans cette section, nous allons illustrer l'utilisation des procédures discutées dans la section précédente pour sélectionner une classification à l'intérieur d'une famille de classifications possibles, obtenues en faisant varier les paramètres de calibration. L'exemple concerne l'analyse d'un fichier de données cliniques composé de 614 patients atteints du cancer du poumon de type "petites cellules". Ces patients ont été traités à l'Hôpital Général de Toronto ainsi qu'à l'Hôpital Princess Margaret, entre 1976 et 1986. La liste des variables sélectionnées pour l'analyse et leurs modalités sont données dans le tableau I, avec leurs effectifs respectifs. A cette liste de variables prédictrices (facteurs pronostiques) formant le vecteur $\underline{z}$ s'ajoute, pour chaque individu, le vecteur $\underline{u}$ composé du temps de survie en jours, t, et d'un indicateur, δ, défini comme suit

$$\delta = \begin{cases} 0 & \text{si le patient est toujours vivant} \\ 1 & \text{si le patient est décédé} \end{cases}$$

On se pose alors les questions suivantes.

Question 1

Est-ce que la survie, t, dépend des facteurs pronostiques z ?

Quelle est la forme de cette dépendance?

(Classification pronostique).

Question 2

Est-ce que le marqueur tumoral LDH prédit la survie uniformément ou existe-t-il certains groupes de patients à l'intérieur desquels le pouvoir de prédiction de LDH est supérieur?

(Classification selon un coefficient de régression)

Dans le but de simplifier le calcul de la dissimilarité entre deux populations données, nous faisons ici l'hypothèse que la survie, t, suit une fonction de densité exponentielle

$$f(t) = \gamma e^{-\gamma t} ;$$

le paramètre γ joue donc ici le rôle du critère puisqu'il caractérise à lui seul la fonction de densité.

TABLEAU I

Variables utilisées et effectifs correspondants

Variable	Description	Catégories	Effectif
ETEN	Étendue du cancer	1 – limitée 2 – grande	286 328
PERF	État de performance au début de la période d'observation	1 – 0 ou négatif 2 – 2 ou 3	479 135
SEXE	Sexe	1 – femme 2 – homme	178 436
ADH	Marqueur tumoral	0 – manquant 1 – négatif 2 – positif	4 579 31
PHAL	Phosphatase alcalin	0 – manquant 1 – normal 2 – élevé	23 376 215
GB	Globules blancs	0 – manquant 1 – normal 2 – élevé	26 414 174
FOIE	Métastases au foie	0 – manquant 1 – non 2 – oui	33 420 161
SOD	Sodium	0 – manquant 1 – $<$ 135 mmol/L 2 – $\geq$ 135 mmol/L	35 142 437
PLAQ	Plaquettes	0 – manquant 1 – $<$ 150,000/mm^3 2 – $\geq$ 150,000/mm^3	38 15 561
OS	Métastases aux os	0 – manquant 1 – non 2 – oui	40 423 151
NOEU	Atteinte des nœuds lymphatiques	0 – manquant 1 – non 2 – oui	42 506 66

TABLEAU I (suite)

Variable	Description	Catégories	Effectif
STAD	Stade de cancer	0 – manquant 1 – I 2 – II ou III 3 – > III	43 47 196 328
CEA	Marqueur tumoral	0 – manquant 1 – < 6 ng/ml 2 – 6-19.9 ng/ml 3 – > 19.9 ng/ml	345 171 53 45
CERV	Métastases au cerveau	0 – manquant 1 – non 2 – oui	52 529 33
LDH	Marqueur tumoral	0 – manquant 1 – < 175 2 – 175-275 3 – > 275	326 88 106 94
MOEL	Examen de la moëlle osseuse	0 – manquant 1 – non 2 – oui	110 426 78
SITES	Nombre total de sites de métastases	1 – aucun 2 – 1 3 – > 1	286 189 139
AGE	Âge recodé	1 – < 50 2 – 50-55 3 – 55-60 4 – 60-65 5 – > 65	87 93 139 122 173
HEMO	Hémoglobine	0 – manquant 1 – ≤ 12.0 2 – 12.0-13.5 3 – 13.5-14.5 4 – > 14.5	30 116 186 124 158

Soient deux segments P_1, P_2 résultant d'une partition à un niveau donné de l'arbre, on fait dont l'hypothèse que la fonction de densité de t à l'intérieur de chacun d'eux est de la forme

$$f_j(t) = \gamma_j \, e^{-\gamma_j t}, \qquad j = 1,2,$$

respectivement. Le calcul de la dissimilarité entre P_1 et P_2, tel que décrit à la section précédente, équivaut à tester l'hypothèse selon laquelle $\gamma_1 = \gamma_2$.

D'autres modèles plus réalistes et plus complexes que le modèle exponentiel peuvent être utilisés; **SEGMAG** en propose d'ailleurs près d'une dizaine. Le tribut à payer est celui de la lenteur des calculs; certaines améliorations techniques telles que le calcul parallèle et l'accélération des accès mémoire devraient toutefois permettre de résoudre ce problème.

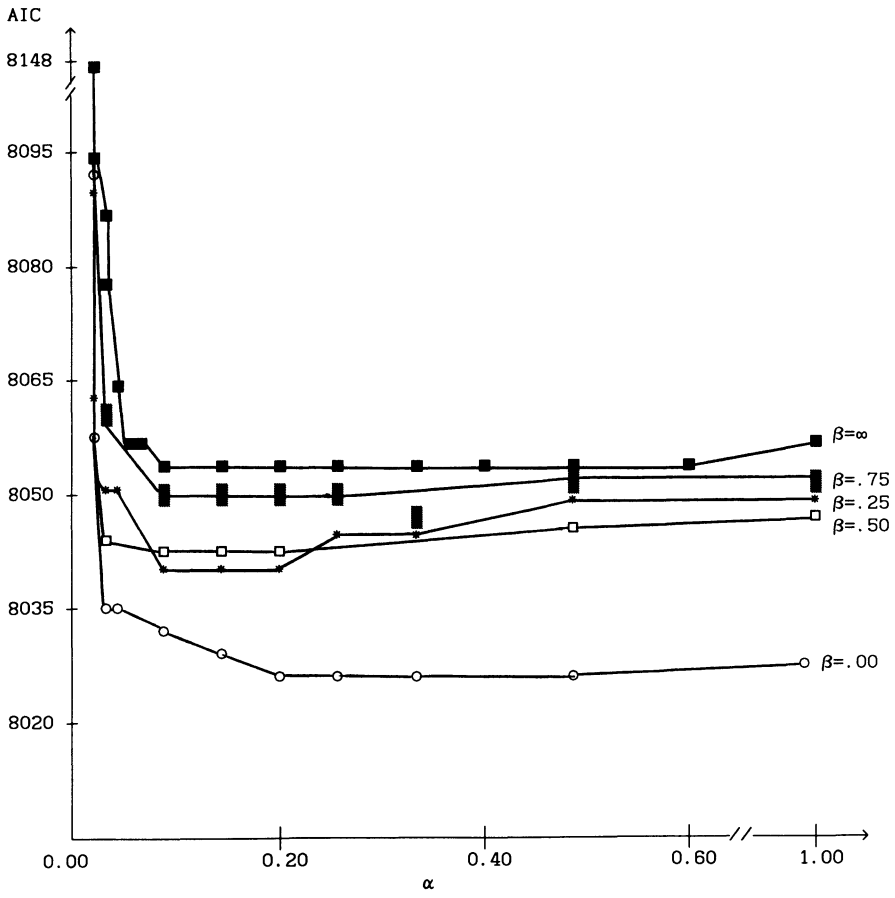

Figure 3 : AIC en fonction de α

Comme le montre la figure 3, la règle d'1 – ERT, décrite à la fin de la section 2, nous amène à choisir l'arbre illustré à la figure 4(a), avec α égal à 0.005 et un β de 0.25. En effet, la partition dont l'AIC atteint son minimum est celle obtenue avec $\alpha = 0.20$ et $\beta = 0.00$. Cependant, la règle d'1 – ERT recommande de retenir comme partitions candidates toutes celles qu'on retrouve à l'intérieur de la bande de hauteur ERT_{α^*}, s'étendant à partir du point d'AIC minimum. Parmi toutes les partitions comprises dans cette bande, on favorise la plus simple, dans ce cas-ci l'arbre obtenu avec $\alpha = 0.005$ et $\beta = 0.25$.

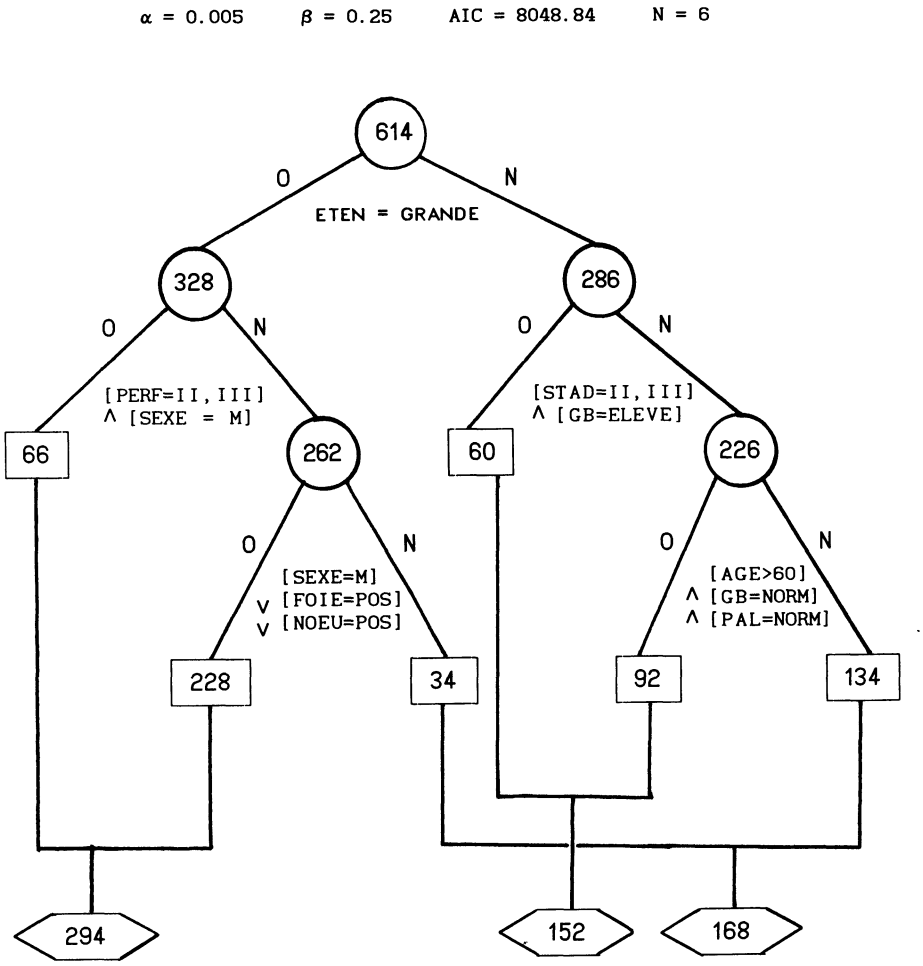

Figure 4(A) : Arbre choisi selon la règle d'1 – *ERT*

L'algorithme d'agrégation appliqué aux six classes terminales, avec la même valeur d'α, conduit à trois groupes dont les courbes de survie, estimées par la méthode de Kaplan-Meier, apparaissent à la figure 4(b). On constate, par exemple, que les hommes pour qui l'étendue du cancer est grande et qui avaient un état de performance de II ou III au début de la période d'observation font partie du groupe d'individus ayant la plus courte durée de survie. Par contre, une petite étendue du cancer et un bas niveau de globules blancs semblent favoriser une longue survie chez les jeunes patients. On peut ainsi caractériser en termes cliniques chaque groupe terminal de l'arbre obtenu et répondre de ce fait à la question 1 formulée précédemment.

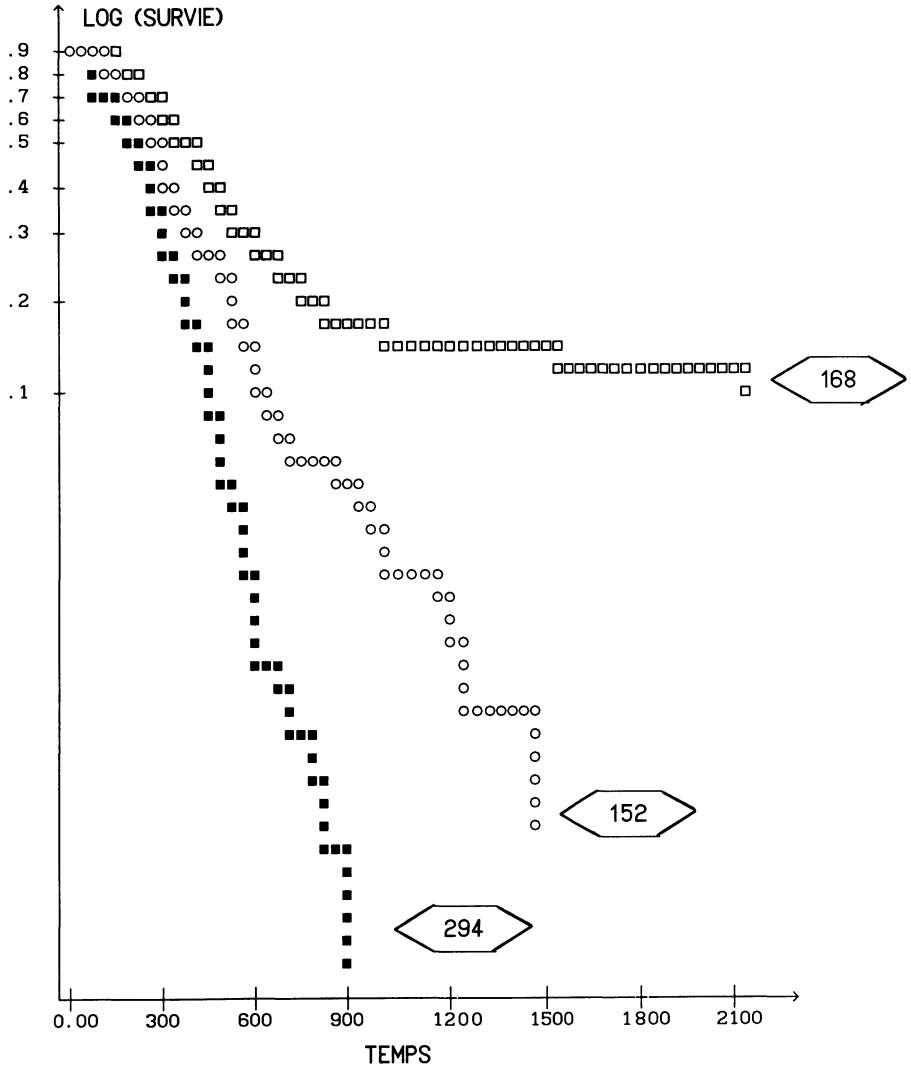

Figure 4(B) : Courbes de survie des trois classes

Dans le contexte de la question 2, on se demande si le pouvoir de prédiction du marqueur tumoral LDH est uniforme sur l'ensemble des individus, ou si au contraire il existe certains groupes de patients pour lesquels la survie est mieux prédite par la variable LDH. Le critère γ servant à mesurer la dissimilarité entre deux groupes devient le coefficient de régression de la survie t sur la variable LDH, dénotée par y.

On définit le risque de décès $\lambda(t)$ comme le rapport

$$\lambda(t) = \frac{f(t)}{S(t)}.$$

où $f(t)$ et $S(t)$ représentent la fonction de densité de t et la fonction de survie, respectivement. Afin de simplifier les calculs nécessaires, on fait l'hypothèse qu'à l'intérieur de chaque groupe P_j, la variable LDH exerce un effet multiplicatif sur le risque de décès spécifique de ce groupe. On a donc

$$\lambda_j(t\,;\,y) = e^{\gamma_j y}\,\lambda_{0j}(t),$$

où $\lambda_j(t\,;\,y)$ dénote le risque de décès pour les individus du groupe P_j ayant la valeur y de LDH et $\lambda_{0j}(t)$ représente le risque spécifique de ce groupe, qu'on suppose inconnu.

La règle d'1 – ERT nous amène à choisir l'arbre avec $\alpha = 0.08$ illustré dans la figure 5 (graphique AIC en fonction de α non illustré). On constate par exemple que le meilleur pouvoir de prédiction de la variable LDH se retrouve dans le groupe composé de femmes ayant moins de 65 ans. En effet, le coefficient de régression γ est égal à 0.881, avec un risque relatif égal à 2.413. Par contre, chez les patients de plus de 65 ans, la variable LDH a un pouvoir de prédiction négligeable puisque le risque relatif n'est que de 1.041.

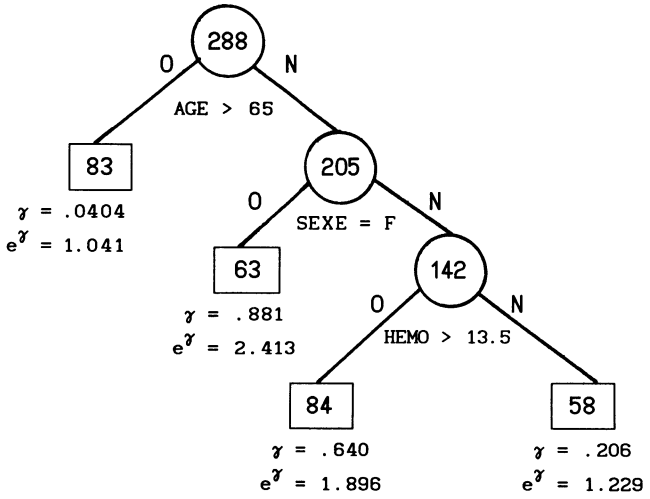

Figure 5. Modèle de COX pour la variable LDH
$\alpha = 0.08$ $\beta = 1.0$ AIC = 296.59 $N = 4$

On pourrait aussi identifier, par le même procédé, des groupes d'individus pour lesquels un traitement quelconque est particulièrement bénéfique. Cette information revêt une importance primordiale lors de la planification d'expériences cliniques ultérieures, puisqu'elle permet d'orienter le choix d'un traitement approprié à chaque patient.

Références

H. AKAIKE (1974). — A new look at the statistical model identification, *IEEE Trans. Automat. Control*, 19, 716-723.

L. BREIMAN, J.H. FRIEDMAN, R.A. OLSHEN et C.J. STONE (1984). — Classification and Regression trees, Belmont, CA : Wadsworth.

A. CIAMPI et J. THIFFAULT(1986). — Recursive Partition in Biostatistics : Criteria for tree selection, COMPSTAT 86, Short Communication and Posters, 47-48.

A. CIAMPI, C.H. CHANG, S. HOGG et S. McKINNEY(1987). — Recursive Partition : A versatile method for exploratory data analysis in Biostatistics, *biostatistics,* I.B. MacNeill et G.J. Umphrey, eds., D. Reidel Publishing Company, 23-50. -50.

D.R.COX et D.V. HINKLEY (1974). — Theoretical Statistics, Chapman & Hall, Londres.

E. DIDAY (1980). — Optimisation en Classification Automatique, (Tomes 1 et 2). INRIA, Le Chesnay.

B. EFRON (1982a). — Maximum Likelihood and Decision Theory, *The Annals of Statistics*, 10, 340-356.

B. EFRON (1982b). — The Jackknife, the Bootstrap and other resampling plans, *SIAM*, Philadelphie.

A.D. GORDON (1981). — Classification : Methods for the exploratory analysis of multivariate data, Chapman & Hall, Londres.

M. JAMBU et M.-O. LEBEAUX (1983). — Cluster Analysis and Data Analysis, North-Holland, Amsterdam.

S. JUNCA (1985). — Outils informatiques pour l'évaluation de la pertinence d'un résultat en analyse de données, Thèse de Doctorat, 3ème cycle, USTL, Montpellier.

J. SIDI (1980). — L'approche logique en classification automatique et reconnaissance des formes, Université Paris VI, Thèse de Doctorat, 3ème cycle.

M. STONE (1974). — Cross-validatory choice and assessment of statistical predictions, *J. Royal Statist. Society B*, 36, 111-147.

M. STONE (1977). — An asymptotic equivalence of choice of model by cross-validation and Akaike's Criterion, *J. Royal Statist. Society B*, 39, 44-47.