20

Méthodologie statistique pour la discrimination et le classement

Application au ciblage des interventions nutritionnelles

Pierre TRAISSAC (1)

On situe le ciblage dans le cadre des méthodes statistiques de discrimination et de classement. Un certain nombre de points communs aux problèmes de discrimination et de classement sont abordés. Sont présentées ensuite trois méthodes statistiques utilisables dans l'optique du ciblage des interventions : analyse discriminante, régression logistique et arbres de discrimination (CART). Un effort particulier est fait pour expliciter certains points techniques fondamentaux. Les exemples qui sous tendent l'exposé sont basés sur des problèmes de ciblage nutritionnel. La comparaison des méthodes sur des données de terrain est présentée.

1. Discrimination et classement dans le cadre du ciblage

Du point de vue des techniques statistiques à mettre en œuvre, le ciblage peut se placer dans le cadre général des problèmes de discrimina-

(1) Laboratoire de nutrition tropicale, ORSTOM, BP 5045, 34032 Montpellier, France.

tion et de classement. D'après le dictionnaire, la discrimination est l'action d'établir une différence. Il s'agit en effet d'établir une différence sur la base d'un certain nombre de critères, entre les individus qui doivent être bénéficiaires d'une intervention et ceux pour lesquels cette intervention ne se justifie pas.

On considère la population des unités statistiques (notées u.s. dans la suite) sur laquelle on désire faire opérer le ciblage. Il est clair que ces unités statistiques pourront être différentes suivant la nature du ciblage envisagé : individus (enfants ou adultes), familles, ménages, villages, etc. Il est important de bien faire apparaître quel type d'unité statistique on désire cibler. Dans les exemples que nous présenterons les unités statistiques sur lesquelles seront basées le ciblage sont des individus. Elles pourraient être des ménages ou des communes à condition qu'il soit possible d'en donner des caractérisations satisfaisantes, en particulier en termes nutritionnels.

On a une partition de cette population en plusieurs sous-populations. Cette partition est en général dichotomique, mais pas nécessairement. Si l'on est intéressé à cibler des enfants préscolaires à risque de maigreur par exemple, on considérera la partition en deux groupes : « maigre » et « non maigre » (poids-taille < -2 e.t. et poids-taille $> = -2$ e.t.). Le même type de partition peut-être associé au retard de taille : « retardé en taille » et « non retardé en taille » (taille-âge < -2 et taille-âge $> = -2$ respectivement). Pour ce qui concerne les adultes, on peut considérer la partition en deux groupes : risque de déficience chronique en énergie (noté « maigre » dans la suite), correspondant à un indice de masse corporelle (IMC) inférieur à 18,5, et ceux dont l'IMC est supérieur ou égal à 18,5 (noté « non maigre » dans la suite). On peut envisager des partitions plus fines ; par exemple, dans le cas des femmes adultes, considérer les classes « maigre », « normal », « obèse » définies par les valeurs standard d'IMC 18,5 et 25. Dans la suite, nous nous limiterons au cas dichotomique car c'est celui qui semble le plus logique en première approche dans une optique de ciblage.

On considère un certain nombre de descripteurs des unités statistiques considérées. Par exemple, dans le cas des femmes adultes, on peut disposer d'informations telles que l'âge, le niveau d'éducation, l'activité physique, la parenté avec le chef de ménage, des descripteurs socio-économiques du ménage (logement, revenus, caractéristiques du chef de ménage etc.). Ces descripteurs, aussi appelés variables explicatives peuvent être de différents types : quantitatives continues (IMC, âge), quantitatives discrètes (nombre de personnes dans le ménage d'appartenance), qualitatives (présence/absence d'électricité, état de l'habitat : mauvais, moyen, bon).

La question qui se pose est la suivante : à partir de ces descripteurs peut-on expliquer et/ou prédire de façon satisfaisante l'appartenance des unités statistiques aux différentes sous-populations. Dans l'exemple des femmes adultes, la question sera : peut-on expliquer la maigreur d'une femme par les valeurs des différentes variables caractérisant son activité, son niveau d'éducation, le ménage auxquel elle appartient ? Pratiquement cela se passe de la façon suivante : on dispose d'un échantillon d'u.s. extrait de la population à laquelle on s'intéresse (individus adultes par exemple). Pour les unités statistiques de cet échantillon, on connaît à la fois la sous-population d'appartenance (maigre ou non maigre par exemple) et les valeurs prises par les variables explicatives.

Dans une première étape, à l'aide de différentes méthodes statistiques que nous évoquerons plus loin, on essaie d'expliquer l'appartenance des u.s. aux différentes sous-populations par les valeurs prises par les variables explicatives. Cette première étape a un intérêt intrinsèque : elle permet de mettre en évidence les variables explicatives qui sont importantes pour discriminer les sous-populations, de caractériser ces sous-populations par certaines valeurs de ces variables. Nous appellerons cette première étape, phase d'apprentissage ou de discrimination.

Dans un deuxième temps, si les variables explicatives ont un bon potentiel de discrimination, on peut envisager d'utiliser les relations mises en évidence pour prévoir l'appartenance à une des sous-populations d'une u.s. pour laquelle on connaît seulement les valeurs prises par les descripteurs (phase de classement). Dans la suite, nous nous restreindrons au cas où, pour une u.s. donnée, le résultat du classement est l'appartenance à l'un ou l'autre des deux groupes. On exclut donc la possibilité de ne l'affecter à aucun des deux groupes par manque d'information. Cette possibilité existe dans certaines des méthodes que nous présenterons, mais nous ne l'évoquerons pas dans un but de simplification de l'exposé.

Il est clair qu'il faut avoir solidement validé la règle d'affectation d'un individu à une sous-population avant d'utiliser celle-ci de façon opérationnelle. Nous reviendrons sur ce point par la suite.

2. Considérations communes aux diverses méthodes

Avant de rentrer davantage dans le détail des différentes méthodes utilisables, nous présentons un certain nombre de considérations générales communes à ces problèmes de discrimination et de classement.

Ajustement

Dans la première partie de la démarche (cf. ci-dessus), on dispose d'un échantillon classiquement appelé « échantillon de base » à partir duquel on va chercher à ajuster un modèle décrivant « au mieux », au sens d'un certain critère, la relation entre l'appartenance des u.s. aux sous-populations et les valeurs prises par ces u.s. pour les différentes variables explicatives.

La spécification mathématique du modèle dépend de la méthode statistique utilisée, mais un principe général est le suivant : pour un modèle donné, lorsque l'on dispose des valeurs prises par une u.s. pour les différentes variables explicatives, on peut calculer la sous-population d'appartenance telle que prédite par le modèle. Lorsqu'on a fait le choix d'une méthode statistique (c'est-à-dire pour une forme de modèle donnée) l'algorithme de calcul va chercher le modèle qui ajuste au mieux les données obervées, soit un modèle tel que les sous-populations d'appartenance prédites pour les différentes u.s. soient globalement proches des sous-populations d'appartenance effectives. Il est en effet clair qu'il n'existe pas de relation « parfaite » entre les descripteurs utilisés et le caractère à prédire donc pas de concordance absolue entre la valeur obtenue par le modèle et la valeur effectivement observée. Dans l'exemple des femmes adultes, un certain nombre de femmes seront classées maigres par le modèle alors qu'elles sont non maigres en réalité et vice versa.

Chaque méthode optimise un critère d'ajustement des valeurs prédites aux valeurs observées qui lui est propre. Néanmoins le produit final que l'on attend de chacune de ces méthodes est une règle de classement, permettant d'affecter une u.s. à une des sous-populations. Il nous paraît important de rappeler un certain nombre de concepts généraux pour juger de la performance d'une règle de classement, indépendamment de la méthode particulière qui a servi à la construire.

Matrice de concordance, spécificité, sensibilité

Restons dans le cadre de l'exemple où l'on cherche à prédire la maigreur des femmes adultes. Supposons qu'on a ajusté un modèle sur un échantillon observé et donc que l'on dispose d'une règle de classement qui permet de prédire le caractère maigre d'une adulte. Considérons le tableau suivant appelé matrice de décision ou matrice de concordance,

construit après application de la règle aux u.s. de l'échantillon de base :

Tableau 1

Matrice de concordance

	Observée Maigre (M)	Observée Non maigre (N)
Prédite Maigre (R+)	a	b
Prédite Non maigre (R-)	c	d

Si la règle d'affectation était parfaite (i.e. si la connaissance des variables explicatives lui permettait de prédire à coup sûr la maigreur), on aurait c = b = 0. Classiquement a est noté VP, pour nombre de vrais positifs (femmes prédites correctement maigres par la règle), b FP (faux positifs), c FN (faux négatifs) et d VN (vrais négatifs).

• La sensibilité de la règle est définie comme la probabilité de détecter la maigreur :

(Se = P(R + /M)). Elle est estimée par $\hat{Se} = \frac{a}{(a+c)} = \frac{VP}{(VP + FN)}$

(taux de vrais positifs).

• La spécificité de la règle est la probabilité de classer correctement un individu non maigre, soit :

Sp = P(R - /N), estimée par $\hat{Sp} = \frac{d}{(d+b)} = \frac{VN}{VN/(FP+VN).}$

Une règle de classement de bonne qualité doit avoir une sensibilité et une spécificité élevée, c'est-à-dire une valeur du rapport $\frac{Se}{1 - sp} = k$ élevée (un individu a k fois plus de chances d'être déclaré maigre par la règle s'il est effectivement maigre).

Soit l'exemple dans lequel on essaie de prévoir la maigreur (IMC < 18,5) dans une population de femmes de plus de 18 ans. La règle de classement aboutit au tableau suivant :

Tableau 2

Matrice de concordance (exemple)

	Observée Maigre (M)	Observée Non maigre (N)
Prédite Maigre (R+)	25	209
Prédite Non maigre (R–)	15	353
	40	562

Sur l'échantillon, la prévalence de maigreur observée est $\hat{p} = \frac{40}{602} = 0{,}07$. L'estimation de la sensibilité est $\hat{Se} = \frac{25}{40} = 0{,}62$. La spécificité calculée sur l'échantillon est $\hat{Sp} = \frac{353}{562} = 0{,}63$.

La valeur du rapport k, estimée par $\hat{k} = \frac{0{,}62}{1 - 0{,}63} = 2{,}30$ indique qu'une femme a deux fois plus de chances d'être classée comme maigre par la règle si elle est effectivement maigre que si elle est de corpulence normale ou obèse.

Une présentation des concepts de sensibilité, spécificité, valeurs prédictives dans le contexte du ciblage est également faite dans le papier de B. Maire *et al.* « Le ciblage dans les politiques et programmes nutritionnels ».

Courbe ROC

Très souvent, la règle découle du choix d'une valeur seuil sur une grandeur continue G. La règle est de la forme si G < a alors l'individu est classé « maigre » sinon il est classé « non maigre ». C'est le cas lorsque le résultat de la méthode statistique utilisée est une probabilité de maigreur par exemple. Le choix de la valeur seuil détermine la sensibilité et la spécificité du test (qui sont antagonistes de par leur définition même). On représente en général les résultats associés aux différents seuils de coupure par une courbe dite courbe ROC (Receiver Operating Carateristic Curve). On porte en abscisse le taux de faux positifs (1 – Sp) et en ordonnée le taux de vrais positifs Se.

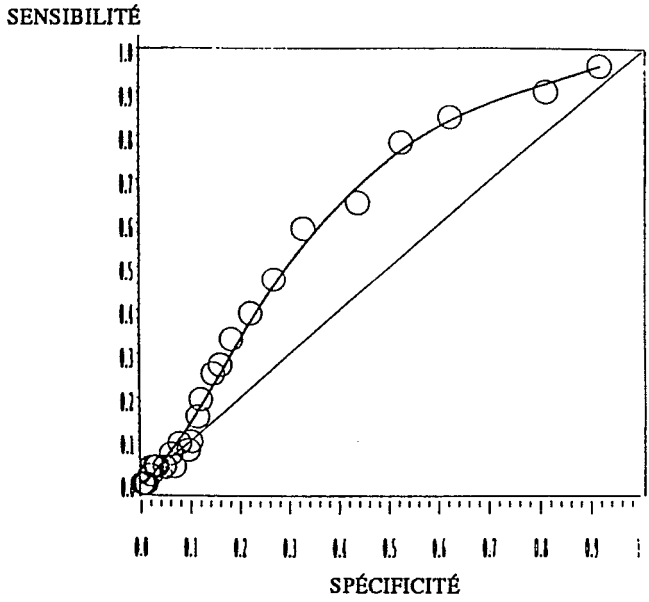
Le choix de la valeur seuil dépend de ce que l'on veut privilégier (Sp ou Se) et donc en fait des « coûts » — au sens général du terme – associés aux mauvaises décisions. Supposons qu'on sache évaluer ces « coûts » : soient C(FP) et C(FN) les coûts associés au mauvais classement d'individus non maigres et maigres respectivement (les coûts associés aux décisions correctes sont pris égaux à 0). Si P(M) est la proportion de maigres dans la population, le coût global lié aux mauvais classements est $\frac{1}{2}$:

$$C = P(M)\, C(FN)\, (1 - Se) + (1 - P(M))\, C(FP)\, (1 - Sp).$$

On peut montrer que le choix optimal (coût global minimum) est le point de la courbe ROC dont la pente vaut $\frac{C(FP)}{C(FN)} \times \frac{1 - P(M)}{P(M)}$

Graphique 1

Exemple de courbe ROC



Si les coûts sont égaux le choix optimal est le point de la courbe dont la pente est $\frac{1 - P(M)}{P(M)}$. Cette formule est assez intuitive ; en effet, elle nous indique que si la maigreur est rare (rapport élevé), il ne faut pas se placer dans le haut de la courbe (c'est-à-dire vouloir détecter tous les cas) car cela conduirait à un nombre très grand de faux positifs. Néanmoins si l'on pense qu'il est très important de détecter les maigres et pas grave d'intervenir à tort chez les non-maigres, on choisira des valeurs de coûts CFP faible et CFN élevé soit $\frac{CFP}{CFN}$ petit. La formule nous incite alors à choisir un seuil de coupure qui nous place sur la courbe ROC au point de pente $\frac{C(FP)}{C(FN)} \times \frac{1 - P(M)}{P(M)}$. Ce point se trouvera alors déplacé vers le haut de la courbe (on privilégie la sensibilité) par rapport au cas précédent (coûts égaux).

Valeurs prédictives

On définit les valeurs prédictives de la règle :

• Valeur prédictive positive VPP = P (M/R +) : probabilité d'être effectivement maigre lorsque classé comme tel par la règle. C'est la capacité de la règle à prédire correctement la maigreur.

• Valeur prédictive négative VPN = P (N/R -) : probabilité d'être effectivement non-maigre lorsque classé comme tel par la règle.

Ces probabilités VPP et VPN sont aussi appelées probabilités *a posteriori* (i.e. après application de la règle) par opposition à P (M) et P (N) = 1 - P (M) qui sont les probabilités d'être maigre ou non maigre *a priori* (i.e. avant application de la règle). On considère également le rapport k' = P (M/R +)/P (M/R -), risque relatif de maigreur associé au classement « maigre » par la règle : un individu a k' fois plus de chances d'être effectivement maigre si il est classé comme tel par la règle que s'il ne l'est pas. (k est estimé sur l'échantillon par $\hat{k}'$ = VPP/(1 - VPN)).

De plus, on peut montrer les résultats suivants :

$$VPP = \frac{P(M)\,Se}{P(M)\,Se + (1 - P(M))\,(1 - Sp)}$$

et

$$VPN = \frac{(1 - P(M))\,Sp}{P(M)\,(1 - Se) + (1 - P(M))\,Sp}$$

Pour Se et Sp fixées, quand la prévalence de maigreur dans la population croît de 0 à 1 la valeur de VPP croît. Elle sera donc plus faible si la maigreur est rare. A contrario, la VPN sera élevée pour cette population. Reprenons l'exemple précédent ($\hat{Se}$ = 0,62 ; $\hat{Sp}$ = 0,63). Le tableau suivant donne les valeurs de VPP, VPN et $\hat{k}'$ pour différentes valeurs de P (M) (prévalence de maigreur dans la population étudiée).

Tableau 3

Valeurs de VPP, VPN et k' (exemple)

P	VPP	VPN	$\hat{k}'$
0,05	0,08	0,97	2,64
0,10	0,16	0,94	2,50
0,20	0,29	0,87	2,25
0,40	0,53	0,71	1,83

On voit que pour une sensibilité et une spécificité données, l'information supplémentaire apportée par la règle pour le classement des individus dépend grandement de la probabilité de maigreur *a priori* (probabilité de maigreur dans la population).

Estimation « honnête » de la qualité de la règle

Indépendamment du fait que la règle provienne du découpage d'un indice continu ou non, une bonne mesure de la qualité de la règle est le coût global :

$$C = P(M)\, C(FN)\, (1 - Se) + (1 - P(M))\, C(FP)\, (1 - Sp)$$

Si les coûts sont égaux :

$$C = P(M)\, (1 - Se) + (1 - P(M))\, (1 - Sp).$$

C est alors parfois appelé taux d'erreur de classement.

Une considération importante est la qualité de l'estimation de l'erreur de classement associée à la règle de décision. Comme c'est souvent le cas en statistique, on ne dispose pas de la vraie valeur de cette quantité : en effet sa connaissance nécessiterait d'observer la totalité de la population. On estime (2) le taux d'erreur de classement à partir des résultats observés sur l'échantillon de base par

$$\hat{C} = P(M)\, \frac{c}{a+c} + (1 - P(M))\, \frac{d}{d+b}$$

où P (M) est la probabilité d'être maigre dans la population qui peut éventuellement être estimée sur l'échantillon par $\frac{a+c}{a+b+c+d}$ si l'on pense que la proportion de maigres dans l'échantillon est une bonne image de ce qui se passe dans la population (cas des études prospectives ou transversales par exemple). Dans le cas contraire se pose le problème du choix des probabilités *a priori*. Il est clair que le choix des probabilités *a priori* et celui des coûts éventuellement associés aux mauvais classements influe grandement sur l'estimation de C (ceci pour un taux de mal classés

(2) Pour obtenir de l'information sur une quantité inconnue, on utilise ce qu'on appelle un estimateur de cette quantité. Un estimateur est une valeur calculée sur l'échantillon dont on pense qu'elle reflète de façon satisfaisante la vraie valeur inconnue. Par exemple, si l'on veut connaître la valeur de la prévalence de maigreur (inconnue) dans une population, on prélève un échantillon et l'on utilise la proportion de maigres sur l'échantillon comme estimateur de cette prévalence.

identique sur l'échantillon). La valeur $\hat{C}$ définie ci-dessus est une estimation possible de l'erreur de classement associée à la règle considérée.

On montre que cet estimateur $\hat{C}$ a un biais (3) négatif, c'est-à-dire qu'il sous-estime systématiquement le taux d'erreur de classement vrai et donne des taux d'erreurs de classement optimistes (trop faibles). Cela se conçoit bien puisqu'on évalue la qualité de la règle d'affectation à partir des mêmes observations qui ont servi à la construire. Il existe plusieurs possibilités pour éliminer ce biais. Nous parlerons de la méthode de l'échantillon test et de la validation croisée.

Échantillon test

L'idée est d'utiliser la règle d'affectation pour prédire l'appartenance à l'une ou l'autre des sous-populations pour des u.s. :

- dont on connaît à la fois les valeurs prises pour les descripteurs et l'appartenance aux sous-populations (comme les u.s. de l'échantillon de base) ;
- mais n'ayant pas servi à l'élaboration de la règle (au contraire des u.s. de l'échantillon de base).

L'échantillon ainsi constitué est appelé *échantillon test*, tandis que les u.s. ayant servi à caler le modèle et donc à élaborer la règle constituent *l'échantillon de base*.

L'idéal est d'utiliser un échantillon test complètement indépendant de l'échantillon de base. Cela peut impliquer, pour valider une règle, d'acquérir sur le terrain un nouvel échantillon. Cela paraît nécessaire dans une optique de ciblage pour valider une règle de décision et cela d'autant plus si l'on veut vérifier la robustesse de cette règle vis-à-vis de facteurs tel que pays, milieu, ethnie, etc.

Pour une première étape de validation de la règle, si l'on ne dispose pas de ce deuxième échantillon, on peut procéder de la façon suivante : considérons l'ensemble des unités statistiques de l'échantillon pour lesquelles on connaît les valeurs des descripteurs et la sous-population d'appartenance. Si l'on dispose d'un nombre assez important d'unités sta-

(3) Une qualité importante d'un estimateur est son absence de biais. Un estimateur est dit sans biais si, lorsqu'on répète un grand nombre de fois le tirage d'un échantillon, la moyenne des différentes valeurs obtenues pour cet estimateur coïncide avec la vraie valeur du paramètre. La proportion de maigres sur l'échantillon par exemple est un estimateur sans biais de la proportion de maigres dans la population : si l'on répète un grand nombre de fois le tirage d'un échantillon de même taille, on obtiendra des valeurs différentes de la proportion de maigres sur l'échantillon mais qui en moyenne seront centrées sur la vraie valeur.

tistiques dans l'échantillon, on peut envisager d'en tirer au hasard un certain nombre (par exemple 30 %) que l'on met de côté qui constitueront « l'échantillon test ». Avec les 70 % restants, on ajuste le modèle et on élabore la règle d'affectation. Lorsque la règle d'affectation est construite, on utilise celle-ci pour prédire la sous-population d'appartenance pour les individus de l'échantillon test et l'on compare les résultats fournis par le modèle et la réalité observée. On utilise les mêmes expressions que précédemment pour estimer le taux d'erreur de classement, sensibilité, etc.

La différence réside dans le fait que les valeurs sont estimées à partir de l'échantillon test (c'est-à-dire sur des u.s. n'ayant pas servi à la construction du modèle). L'estimateur ainsi obtenu pour le taux d'erreur de classement est non biaisé. Il donne une estimation plus « honnête » du taux d'erreur vrai liée à l'utilisation du modèle. La valeur obtenue est en général plus élevée que celle obtenue sur l'échantillon de base.

La situation idéale est celle où l'on dispose d'un échantillon indépendant du premier. La méthode de tirage au hasard d'un échantillon test dans l'échantillon de base est intéressante si l'on dispose d'un nombre important d'u.s. dans l'échantillon de base. Dans le cas contraire elle présente l'inconvénient de réduire le nombre d'u.s. à partir desquelles on élabore la règle d'affectation. Une solution possible est alors l'utilisation de la validation croisée.

Validation croisée

On divise de façon aléatoire l'échantillon de base en un certain nombre de sous-ensembles de taille égale (par exemple 10, on parle alors de validation croisée d'ordre 10). Soit V_i un de ces sous-ensembles : on considère comme échantillon de base les u.s. n'en faisant pas partie et comme échantillon test le sous-ensemble lui-même. On élabore la règle de décision sur l'échantillon de base et on estime le taux d'erreur de classement par la méthode de l'échantillon test décrite ci-dessus. Soit $\hat{C}(V\{_i\})$ ce taux. On répète cela 10 fois (autant de fois que de sous-ensembles). L'estimateur final du taux d'erreur de classement utilisé est la moyenne des 10 taux d'erreurs de classement, soit

$$\hat{C}_{vc} = \frac{1}{10} \sum_i \hat{C}(V_i)$$

qui est appelé estimateur du taux d'erreur de classement par validation croisée d'ordre 10.

L'avantage de cette façon de faire est que toutes les u.s. participent à l'estimation du taux d'erreur. C'est cette méthode qui est préconisée dans

le cas ou la taille de l'échantillon est trop réduite pour utiliser la méthode de l'échantillon test. Il n'y pas d'inconvénient majeur à l'utiliser sur des échantillons de très grande taille à part les temps de calcul.

Compromis entre simplicité et performance

Un point important dans une optique de ciblage opérationnel est la simplicité d'emploi du modèle utilisé. Néanmoins, on ne doit pas privilégier la simplicité du modèle au détriment de la performance de la règle de classement. Il s'agit donc de trouver un compromis entre la simplicité du modèle et la qualité de la règle de classement. Nous retiendrons trois éléments pouvant influencer ces deux caractéristiques antagonistes : le nombre de variables entrant dans la construction de la règle de classement, la spécification mathématique du modèle et la plus ou moins grande facilité de collecte des descripteurs sur le terrain.

Nombre de variables

De manière générale, pour un certain nombre de méthodes statistiques existent des algorithmes dits « pas à pas » ou « stepwise ». Dans un algorithme de type ascendant, par exemple, les variables sont prises en compte une à une dans le modèle de façon progressive : l'algorithme choisit en premier la variable permettant la meilleure discrimination (au sens du modèle utilisé). Puis une autre variable est choisie en tenant compte du fait que la première est déjà dans le modèle (on s'intéresse à l'information supplémentaire apportée par la variable candidate à l'entrée). Lorsque ces deux variables sont dans le modèle, une troisième variable est éventuellement choisie sur les mêmes critères, etc. La sélection s'arrête lorsque l'information supplémentaire apportée pour la discrimination par la prochaine variable candidate à l'entrée est inférieure à un seuil fixé. Cette façon de procéder présente des avantages évidents :

- sélection d'un modèle présentant un bon compromis simplicité/performance (pour ce qui concerne la discrimination) par élimination des variables redondantes et/ou des variables non discriminantes ;
- hiérarchisation des variables : l'ordre d'introduction des variables dans le modèle peut être un indice de leur importance relative dans la discrimination des populations (néanmoins si une variable V est déjà dans le modèle, une variable très discriminante mais très corrélée avec V ne sera sans doute pas sélectionnée bien qu'intrinsèquement elle apporte beaucoup d'information sur la discrimination).

Il existe également des méthodes descendantes. Au départ, toutes les variables sont présentes dans le modèle. Sont éliminées au fur et mesure les variables dont le retrait n'influe pas de façon significative sur la discrimination. Certains algorithmes combinent les deux techniques précédentes.

Forme du modèle

Pour un nombre de variables équivalent, la spécification mathématique plus ou moins simple de la règle de classement peut également influer sur le choix entre deux modèles dans une optique de ciblage. Cette règle de classement peut résulter de l'évaluation d'une expression algébrique, être basée sur un arbre de décision, etc. Dans une optique opérationnelle, à performances égales, on préférera utiliser une règle de classement plus intuitive ou plus facile d'emploi pour les personnels de terrain.

Facilité de collecte des valeurs des descripteurs

Un autre facteur influant sur la simplicité d'usage de la règle est la facilité d'obtention des valeurs des descripteurs mais ceci est un problème extra-statistique qui sort du cadre de notre présentation.

3. Les méthodes statistiques

Dans cette partie nous présentons trois méthodes statistiques parmi celles les plus classiquement utilisées dans les problèmes de discrimination et classement :

- l'analyse discriminante,
- la régression logistique,
- la segmentation ou arbres de discrimination (méthode CART).

Notre but n'est pas de présenter ces méthodes dans le cadre mathématique le plus général, mais de préciser certains points qui, bien qu'un peu techniques, sont importants pour la compréhension des résultats fournis par ces méthodes.

L'analyse discriminante

C'est une des méthodes les plus anciennement utilisées dans les problèmes de discrimination et de classement. Sous le vocable analyse discriminante sont regroupées un certain nombre de méthodes assez différentes dans leurs fondements mathématiques et les hypothèses nécessaires à leur validité. Nous présentons l'analyse factorielle discriminante linéaire. Cette technique appartient à la famille des méthodes factorielles au même titre que l'analyse en composantes principales, l'analyse des correspondances, etc.

Lors de la phase d'apprentissage, on recherche une combinaison linéaire des variables de départ, c'est-à-dire une variable de la forme $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_j X_j + \ldots$ où les X_i sont les descripteurs des u.s. Cette combinaison linéaire est construite sous contrainte de maximiser un critère de séparation des deux groupes : par exemple la quantité

$$\frac{\textit{Variance inter}}{\textit{Variance totale}}$$ (4) qui est appelée pouvoir discriminant de la nouvelle variable. Cette quantité est théoriquement comprise entre 0 et 1. La valeur 0 correspond à l'absence complète de discrimination (la connaissance des variables n'apporte aucune information sur l'appartenance aux sous-populations). La valeur 1 correspond à une discrimination parfaite (tous les individus d'un même groupe ont mêmes valeurs pour tous les descripteurs).

En pratique cette valeur est comprise entre les deux extrêmes. Plus elle est élevée, plus les variables explicatives sont un bon potentiel d'explication de l'appartenance aux groupes. La nouvelle variable ainsi construite est appelée *axe discriminant.*

Un des résultats de l'analyse discriminante est de fournir, pour chacun des groupes, une fonction discriminante de la forme $\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_j X_j + \ldots$ où les X_i sont les variables retenues dans le modèle. Ces fonctions ont la propriété de prédire au mieux la sous-population d'appartenance des u.s. de la façon suivante : on calcule pour chaque u.s. les valeurs $\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_j x_j + \ldots$ à partir des valeurs x_i qu'elle prend pour les variables et ceci pour chacune des fonctions discriminantes. L'u.s. est affectée ensuite au groupe pour lequel sa valeur calculée

(4) Variance intergroupe : variance de cette nouvelle variable calculée entre les moyennes des groupes. Variance totale : variance de cette nouvelle variable calculée entre toutes les u.s.

est la plus grande. Géométriquement, cela revient à classer une u.s. dans le groupe pour lequel la distance calculée sur l'axe discriminant entre le point moyen du groupe et l'u.s. est la plus faible.

On utilise en général une méthode pas à pas pour le choix des variables rentrant dans la construction des fonctions discriminantes. Le résultat de l'application des fonctions discriminantes prend ses valeurs dans un intervalle continu ; se pose donc le problème du choix d'un seuil de coupure. D'autre part la validation de la règle de classement par échantillon test ou validation croisée est recommandée.

Cette méthode est initialement conçue pour des variables quantitatives. C'est une des plus performantes si la distribution des variables explicatives est proche de la multinormalité. Dans ce cas on peut également donner une interprétation probabiliste de la règle géométrique d'affectation. En fait, cette hypothèse est rarement vérifiée en pratique. D'autre part, dans une optique de ciblage on aura souvent un mélange de descripteurs quantitatifs (pas nécessairement continus) et qualitatifs. Des étapes préliminaires de codage des données qualitatives peuvent permettre de se ramener au cas quantitatif (analyse des correspondances par exemple) mais ne facilitent pas l'interprétation finale. Une possibilité plus simple est de réaliser un codage disjonctif des variables qualitatives. Soit par exemple la variable état de l'habitat qui prend les valeurs « bon », « moyen », « mauvais ». On recodera en trois variables Vbon, Vmoyen, Vmauvais qui prendront les valeurs 1 ou 0.

La régression logistique

On code l'appartenance aux sous-groupes par une variable Y dichotomique. Par exemple, dans le cas d'une femme adulte, on codera Y = 1 si elle est à risque de DCE (IMC < 18,5), Y = 0 si elle ne l'est pas. On peut alors se poser le problème de discrimination en terme d'un modèle de régression linéaire liant la variable Y aux variables explicatives X_1, X_2, ...X_p. Néanmoins la valeur à prédire étant comprise entre 0 et 1, certaines contraintes sont imposées sur la forme de l'équation de régression de façon à ce que les valeurs prédites par le modèle restent dans cet intervalle (qui définit en fait une probabilité : Y = P (Maigre/X_1, X_2,..., X_p). Une possibilité est de chercher un modèle de la forme :

$$\text{Logit}(Y) = \log\left(\frac{Y}{1-Y}\right) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i$$

ou ce qui est équivalent.

$$f(x) = \frac{e^x}{1+e^x}$$

La fonction $Y = \frac{e^{\beta 0} + \sum \beta_i X_i}{1 + e^{\beta 0} + \sum \beta_i X_i}$ est classiquement appelée fonction logistique, d'où le nom de la méthode. Un certain nombre de considérations la font préférer à d'autres, notamment l'interprétation des paramètres (voir plus loin). Un des intérêts de la régression logistique est la possibilité d'utiliser tous types de variables explicatives (quantitatives continues, quantitatives discrètes, ordinales, qualitatives).

A partir de l'échantillon de base, les paramètres i inconnus, sont estimés par un algorithme maximisant la vraisemblance. L'interprétation de la valeur des paramètres est la suivante : soit i l'estimation du coefficient de la variable X_i dans le modèle. On peut montrer que $\phi = e^{\beta i(b-a)}$ est l'estimation de l'odds-ratio associé à la modification du risque (par exemple de maigreur dans notre cas) lorsque la valeur de X_i passe de a à b, effet ajusté pour les autres variables dans le modèle.

Dans le cas d'une variable qualitative à deux modalités (présence ou absence d'eau courante par exemple), on code cette variable en 1 : présence d'eau courante, 0 : si absence. La modification du risque de maigreur lié à la présence d'eau courante est mesurée en terme d'odds-ratio par $\phi = e^{\beta i(1-0)} = e^{\beta i}$, si βi est l'estimation du coefficient de la variable dans le modèle.

Dans le cas de variables qualitatives à plus de deux modalités, le choix d'une catégorie de référence et un codage approprié permettent de se ramener au cas précédent. Cette interprétation a un intérêt majeur dans une optique explicative.

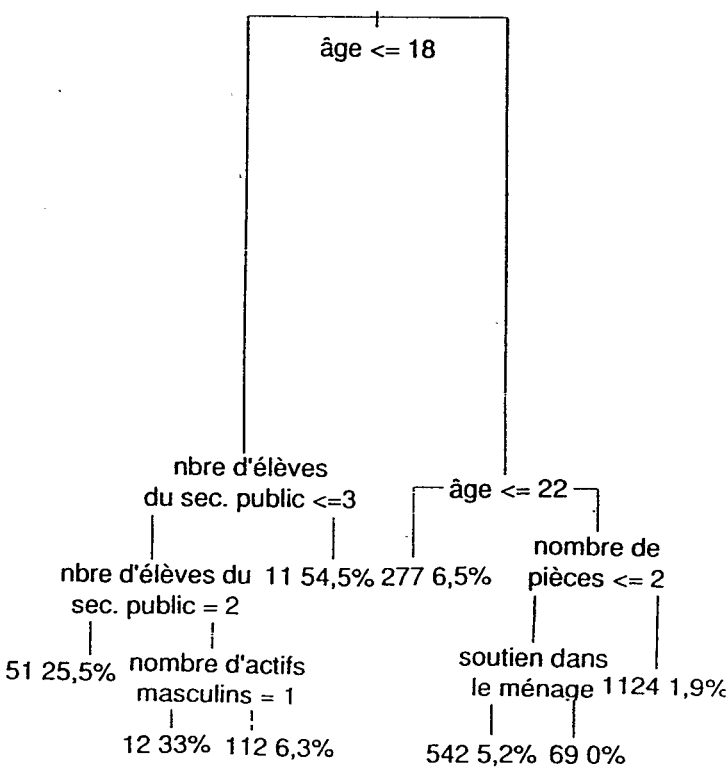
Pour ce qui concerne la prédiction : la fonction $\hat{Y} = \frac{e^{\beta 0} + \sum \beta_i X_i}{1 + e^{\beta 0} + \sum \beta_i X_i}$ permet de calculer une valeur Y ε[0, 1] dès que l'on connaît la valeur prise par l'u.s. pour les variables explicatives X_1, X_2, ...X_p. Ce que l'on obtient en fait est une estimation de la probabilité de présenter le caractère étudié (par exemple maigreur) étant donné la valeur des variables explicatives. Le choix d'un seuil permet ensuite d'affecter l'u.s. à l'un des deux groupes. On choisit une valeur seuil a comprise entre 0 et 1 ; si $Y \geq a$, l'individu est classé « maigre », sinon l'individu est classé dans « non maigre ». Le choix du seuil est en général fait après examen d'une courbe ROC et/ou par minimisation d'une fonction de coût (cf. paragraphe 2), ou encore de manière pragmatique en fonction du nombre de ménages ou d'individus sur lesquels on a les moyens d'intervenir.

La segmentation (CART)

La segmentation désigne généralement les méthodes de discrimination qui construisent des arbres de décision binaires. Cette méthode de discrimination est relativement ancienne dans son principe. Des développements méthodologiques relativement récents (cf. *Classification And Regression Trees* par Breiman et *al.*) ainsi que la banalisation de moyens de calculs performants permettant de les mettre en œuvre ont donné une nouvelle jeunesse aux méthodes de segmentation. Les performances sont la plupart du temps égales ou supérieures à celles des méthodes de discrimination classiques (décrites ci-dessus) avec l'avantage de la lisibilité immédiate de la règle d'affectation (arbre de décision).

Graphique 2

Exemple d'arbre de discrimination



Un exemple d'un tel arbre figure sur le graphique 2 : dans cet exemple on cherche à prédire la maigreur (IMC < 18) d'individus adultes (âge ≥ 17 ans) à partir d'un certain nombre de variables (âge, parenté avec le chef de ménage, descripteurs socio-économiques du ménage et du chef de ménage...). Cet arbre est constitué de « nœuds » reliés par des segments de droites.

Principe de construction de l'arbre

Deux éléments essentiels dans la construction de tels arbres sont la notion de coupure associée à une variable et celle d'impureté d'un nœud.

– *Coupure* : étant donné une variable continue X on appelle coupure, une question binaire du type X < c (par exemple âge < 18) qui permet de répartir sans ambiguïté les u.s. en deux groupes : celles pour lequelles X < c et les u.s. telles que X > = c. Si la variable X est qualitative soit par exemple l'état de l'habitat avec les modalités bon, moyen, mauvais, une coupure sera définie par un sous-ensemble des modalités : par exemple bon ou moyen. Cette coupure définit une répartition des u.s. en deux groupes : celles pour lesquelles l'état de l'habitat est bon ou moyen, et les autres (état de l'habitat mauvais).

– *Impureté d'un nœud* : un nœud est d'autant plus pur qu'il contient une classe largement majoritaire. L'impureté d'un nœud est maximale si la répartition des 2 groupes dans le nœud est 50 %, 50 %. Elle est minimale si la répartition est 100 %, 0 % ou 0 %, 100 %. Différents choix sont possibles pour l'expression mathématique de cette impureté. Il semble que ce choix n'ait en général pas grande influence sur les résultats.

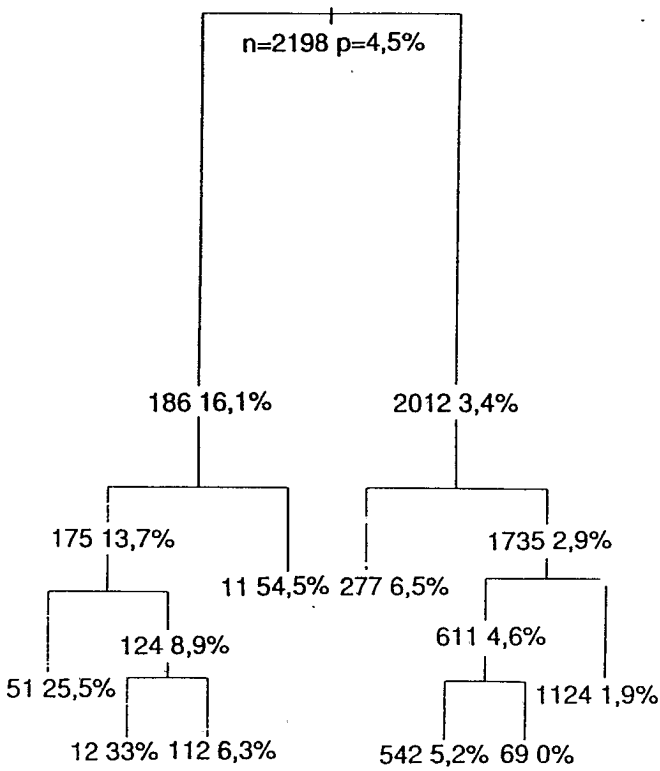
A partir de ces éléments, l'algorithme fonctionne de la façon suivante : à l'étape 1, on dispose de l'ensemble des u.s. de l'échantillon de base qui constitue le premier nœud de l'arbre. On recherche alors la coupure qui permet de scinder en deux l'ensemble des u.s. de façon à obtenir une réduction maximale de l'impureté. Soit i(t) l'impureté d'un nœud t et tg et td respectivement les nœuds gauche et droite créés par la coupure. La réduction d'impureté est $\Delta_i = i(t) - p(tg)i(tg) - p(td)i(td)$ où p(tg) et p(td) sont les proportions d'u.s. du nœud t qui tombent dans tg et td. La recherche se fait de manière exhaustive par examen de toutes les variables explicatives et balayage de l'ensemble des valeurs prises par chacune de ces variables pour détermination de la coupure permettant une réduction maximale de l'impureté. Sur l'exemple présenté, la coupure optimale « âge ≤ 18 » scinde l'ensemble des u.s. (n = 2198, p = 0,045) en deux sous-groupes (n (tg) = 186 et n (td) = 2012) dans lesquels la proportion de maigres est de 0,161 et 0,034 respectivement. Le graphique 3 représente le

même arbre que ci-dessus mais sur lequel on a fait apparaître pour chaque nœud l'effectif et la proportion de maigres.

Graphique 3

Exemple d'arbre de discrimination

CART : hommes ville



L'algorithme se déroule ensuite de manière récursive : pour chaque nouveau nœud ainsi créé est recherchée la question binaire permettant un découpage en deux nouveaux nœuds et conduisant à une réduction maximale d'impureté. Pour le découpage du nœud de gauche dans le cas de l'exemple, la coupure est associée à la variable nombre d'élèves du secondaire public dans le ménage. Pour celui de droite, c'est la variable âge qui intervient de nouveau, etc. Se pose le problème de l'arrêt du découpage. Nous y reviendrons plus bas.

Prédiction de l'appartenance à une sous-population

On suppose que l'arbre de décision est figé. Il est contitué d'un ensemble de nœuds terminaux et de nœuds non terminaux. L'arbre est entièrement déterminé par la suite de questions qui ont servi à sa construction. Chaque nœud terminal est défini de façon unique par un ensemble de conditions : par exemple (âge ≥ 18 ans) et (âge ≥ 22 ans) et (nombre de pièces du logement ≤ 2) et (soutien financier principal hors du ménage) pour le nœud terminal correspondant à p = 0 %.

La règle d'affectation d'un individu est comme suit : partant du sommet, l'individu descend le long de l'arbre en fonction des valeurs qu'il prend pour chacune des variables définissant l'arbre, jusqu'à ce qu'il arrive dans un nœud terminal. Le choix de coûts de mauvais classement égaux pour chacun des groupes et des probabilités *a priori* proportionnelles aux effectifs observés conduit à l'affecter à la classe majoritaire dans le nœud terminal.

Une autre possibilité est de considérer la proportion de retardés en taille dans le nœud, comme une estimation de la probabilité de retard de taille de l'individu. L'affectation à une des sous-populations est alors liée au choix d'un seuil qui peut par exemple être fait après examen d'une courbe ROC.

Principe de l'obtention d'un arbre de taille adéquate

Un arbre de bonne taille est un arbre fournissant le taux d'erreur de classement le plus petit possible avec le nombre de nœuds le plus réduit possible. Si l'on considère l'échantillon de base, le taux d'erreur le plus petit possible sera obtenu en poursuivant le découpage jusqu'à ce que chaque nœud terminal ne contienne qu'une seule u.s. Avec la règle d'affectation de l'u.s. définie ci-dessus, on a alors un taux de mauvais classement égal à 0. Ce choix n'est néanmoins pas acceptable. Il conduit à des arbres très grands, peu pratiques à utiliser. De plus, l'inconvénient majeur est que pour tout autre échantillon, l'utilisation de cet arbre risque de conduire à un taux d'erreur très important. Il est en effet peu probable de trouver deux échantillons ayant exactement la même structure.

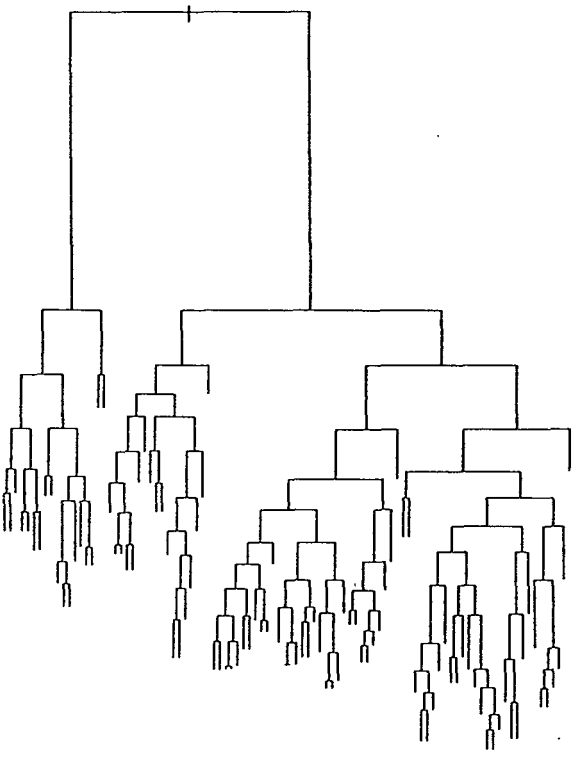
On procède en fait de la façon suivante : on laisse se développer l'arbre assez loin (exemple de règle d'arrêt : impureté d'un nœud égale à 0 ou effectif < = 5). Le graphique 4 montre un tel arbre correspondant à l'exemple présenté. Chacun des nœuds terminaux a un effectif très faible. Par utilisation d'un échantillon test ou par validation croisée, l'arbre est ensuite progressivement élagué par le bas. On obtient ainsi une suite d'arbres emboités que l'on compare pour choisir l'arbre de taille adéquate

en fonction du critère choisi (taux d'erreur de classement par exemple). On peut noter que l'échantillon test (ou la validation croisée) sert à la fois à la construction de la règle et à une meilleure estimation du taux de mauvais classement. Dans le cas des deux autres méthodes présentées ci-dessus, la règle est bâtie uniquement sur l'échantillon de base tandis que l'échantillon test est utilisé *a posteriori* pour donner une meilleure estimation des performances de celle-ci.

Graphique 4

Arbre avant élagage

CART : hommes ville



Remarques diverses

Un des traits principaux de la discrimination par arbre binaire est la *simplicité* de la lecture. Il n'y a pas d'équation algébrique à manipuler ni

de score à calculer pour affecter un nouvel individu à une des sous-populations. On a une règle d'affectation parfaitement lisible même par un non-spécialiste. Ce peut être un avantage déterminant dans une optique de ciblage opérationnel.

Il existe la possibilité d'affecter un individu présentant des données manquantes, basée sur la notion de coupure suppléante : lors de la construction de l'arbre, pour chaque question binaire, on peut rechercher s'il existe d'autres questions (basées sur d'autres variables) qui donnent « presque » la même division du nœud. Une fois que l'arbre est établi, si l'on veut affecter un individu ayant une donnée manquante pour une coupure présente dans l'arbre, on peut éventuellement remplaçer cette question par une de ses questions suppléantes. L'efficacité de cette façon de procéder dépend de la nature des données. Néanmoins c'est une possibilité que ne possède ni l'analyse discriminante « classique » ni la régression logistique. En effet, l'affectation d'un individu nécessite l'évaluation d'une expression algébrique du type $\alpha_0 + \alpha_1.x_1 + \alpha_2.x_2 + \alpha_j.x_j + ...$ qui ne peut être évaluée si une des valeurs est manquante.

Du point de vue statistique, un élément intéressant est la prise en compte automatique des *interactions*. En effet, sur l'arbre présenté en exemple, pour les u.s. vérifiant « âge ≤ 18 » la variable prise en compte par la suite est le nombre d'enfants du ménage élèves dans le secondaire public, alors que pour ceux du nœud de droite (« âge > 18 »), la variable âge semble plus intéressante : l'effet du facteur « nombre d'élèves du secondaire public » dépend donc du niveau d'âge auquel on se place. Ceci est caractéristique d'une situation d'interaction. Elle est détectée de façon implicite par la méthode de construction de l'arbre. Dans la majorité des autres méthodes, la gestion des interactions est assez laborieuse. Dans ce cas de la régression logistique, par exemple, celles-ci doivent être spécifiées et conduisent rapidement à un modèle très complexe dès qu'on dépasse l'ordre 2.

Au vu des caractéristiques des diverses méthodes, il semblerait que, dans une optique de ciblage des interventions, la discrimination par arbre soit une alternative séduisante aux méthodes classiques de régression et d'analyse factorielle discriminante. Il reste à comparer les performances relatives de ces méthodes sur des données de terrain, ce que nous faisons dans le paragraphe suivant.

4. Comparaison des méthodes sur les données de l'enquête nationale de Tunisie en 1990

Les données utilisées dans la suite sont issues de l'enquête nationale sur le budget et la consommation des ménages réalisée par l'Institut national de la statistique en Tunisie en 1990. Cette enquête portait sur 3 852 ménages, totalisant 15 045 individus. Certaines informations ont été recueillies au niveau du ménage (condition de l'habitat et de l'environnement), d'autres à celui de l'individu lui-même (âge, anthropométrie, niveau d'éducation...). Nous ne rentrerons pas dans le détail. Pour plus de précisions sur ces données et le cadre dans lequel elles ont été recueillies, on pourra se reporter au texte de A. Mouelhi, « Évolution de la pauvreté en Tunisie » ainsi qu'à celui de N. Lacourly, « Sélection des bénéficiaires des programmes d'assistance ».

Les unités statistiques que nous avons choisi de cibler, dans le cadre de cet exemple de comparaison de méthodes, sont les individus adultes (âge $\geq$ 17 ans). La partition à expliquer est définie par les deux groupes « maigre » (IMC < 18,5) et « non maigre » (IMC > = 18,5) (5). Dans la mesure où l'indice de masse corporelle ne peut s'interpréter de la même façon suivant le sexe, nous avons procédé à des analyses séparées pour chaque sexe. De plus, les analyses concernant les milieux rural et urbain ont été différenciées, du fait que les variables socio-économiques n'y ont pas la même signification. On a donc procédé à la comparaison des trois méthodes (analyse discriminante, régression logistique et arbre de discrimination CART) sur chacun des sous-ensembles « hommes rural », « hommes ville », « femmes rural » et « femmes ville ».

Les variables potentiellement explicatives de la maigreur et qui ont été introduites dans les analyses sont au nombre d'une cinquantaine, décrivant le niveau socio-économique du ménage (caractéristiques de l'habitat, nombre d'actifs...), ainsi que des caractéristiques propres à l'individu (âge, parenté avec le chef de famille, etc.). Dans le cadre de cet exposé méthodologique, il nous paraîtrait fastidieux d'en donner une liste complète.

---

(5) Il faut souligner que la proportion d'individus à risque de DCE (IMC < 18,5) est faible dans cette population (5,5 %). Ceci explique en grande partie que l'âge prédomine dans les critères de ciblage comme nous le verrons par la suite. Cela ne remet pas en cause l'intérêt des méthodes présentées. Il nous semblait important de les illustrer par un exemple concret. Le choix des données de l'enquête tunisienne a été fait parce qu'elles étaient disponibles, et pour rester dans le thème de l'atelier qui était centré sur les pays du Maghreb.

Un certain nombre de variables présentes dans le fichier initial ont été éliminées *a priori* car non adéquates dans une optique de ciblage opérationnel. D'autre part, un certain nombre de recodages ont été effectués par rapport aux données qui nous ont été fournies. Du fait de la nature des données de départ et des recodages effectués, on a un mélange de variables quantitatives et qualitatives.

Dans la mesure où les effectifs le permettaient, pour chacun des 4 sous-ensembles définis ci-dessus, on a constitué un échantillon de base et un échantillon test de la façon suivante : l'échantillon test a été construit par tirage au hasard de 30 % des u.s., les 70 % restants constituant l'échantillon de base.

La mise en œuvre informatique des analyses a été faite sur une station de travail fonctionnant sous Unix. La gestion des données, les graphiques et la majeure partie des analyses statistiques (analyse discriminante, régression logistique) ont été réalisées à l'aide du système SAS version 6.07 pour station de travail sous Unix. Nous avons utilisé le macro langage SAS pour automatiser certains traitements. Les analyses par arbre de segmentation (CART) ont été réalisées avec le logiciel Splus Version 3.1 pour Unix.

Afin de ne pas alourdir le texte, nous présentons dans la suite uniquement les résultats concernant le sous-ensemble « hommes ville ». L'objectif est en effet d'illustrer sur un exemple les résultats fournis par les différentes méthodes et non d'analyser dans sa totalité les données de l'enquête tunisienne.

Comparaison des méthodes sur les données « hommes ville »

Dans cette partie, nous comparons les résultats de la règle de classement obtenue pour chacune des trois méthodes statistiques (analyse discriminante, régression logistique et CART) sur les données « hommes ville ». Compte tenu des variables utilisées, la taille de l'échantillon disponible pour ces analyses était de 3 138 u.s. au total, réparties en 2 198 u.s. pour l'échantillon de base (98 maigres et 2 100 non maigres, soit une prévalence de maigreur de 4,5 %) et 940 pour l'échantillon test. Les mêmes échantillon de base et échantillon test ont été utilisés pour les trois types d'analyse.

Pour chacune des trois analyses présentées ci-dessous, la démarche conduisant à la construction de l'outil de ciblage est structurée de façon identique :

– sur l'échantillon de base, ajustement du modèle et sélection des variables les plus explicatives de la maigreur dans le cadre du modèle utilisé ;

– validation de la règle de classement sur l'échantillon test. Construction de la courbe ROC et choix d'un seuil ;

– construction d'un score (ou d'un arbre de décision) pour le ciblage adapté à un usage par des personnels de terrain.

Régression logistique

– Ajustement du modèle et sélection des variables

Sur les u.s. de l'échantillon de base, on ajuste un modèle de régression logistique, avec sélection de variables de type pas à pas ascendant. La variable à expliquer est la maigreur codée en 0 (non maigre) et 1 (maigre). Les variables explicatives candidates à la sélection sont au nombre de 50 environ (cf. ci-dessus).

Les variables retenues sont les suivantes :

- âge < 20 ans (âge1)
- enfant ou petit-enfant du chef de ménage (par 3)
- pas de radio (rad2)
- nombre d'enfants du ménage dans le secondaire public ≥ 2 (spu2)
- activité nulle ou légère (act1)
- pas de machine à laver (mal2)

Le tableau suivant donne les valeurs des paramètres du modèle logistique ainsi que leur interprétation en terme d'odds-ratio :

Tableau 4

Paramètres du modèle logistique. Données tunisiennes

Variable	$\hat{\beta}_i$	p	$\hat{OR}$
constante	$-5,14$		
age1	0,99	0,0001	2,70
par3	0,52	0,04	1,68
rad2	0,58	0,01	1,80
spu2	0,53	0,02	1,70
act1	0,57	0,02	1,77
mal2	1.06	0,03	2,88

L'interprétation est la suivante : par exemple, la valeur de OR = 2,88 pour la variable mal2 signifie qu'un individu appartenant à un ménage ne possèdant pas de machine à laver a environ 3 fois plus de chances d'être maigre qu'un individu dont le ménage en possède une. Il est clair que

cette variable est un indicateur indirect de niveau économique. D'autres variables de ce type sont sans doute susceptibles d'apporter le même genre d'information.

– Validation de la règle, construction de la courbe ROC, choix d'un seuil

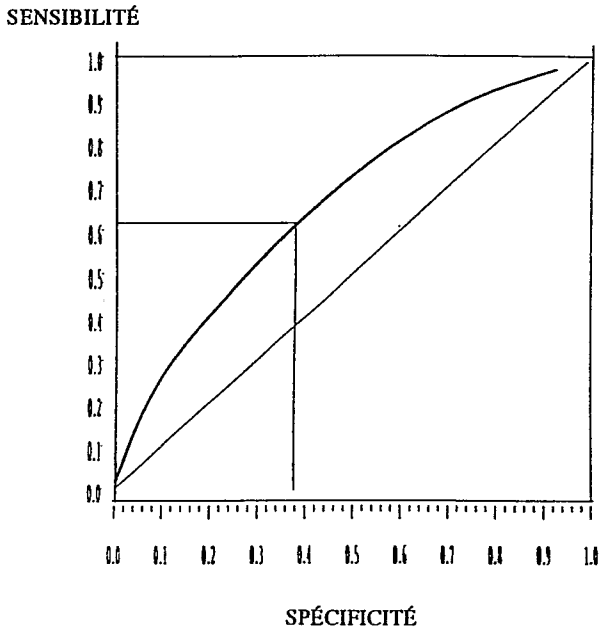
Pour chacun des 940 individus de l'échantillon test, on calcule $\hat{P} = \frac{e^{\beta 0} + \sum \beta_i X_i}{1 + e^{\beta 0} + \sum \beta_i X_i}$. C'est une estimation de sa probabilité de maigreur étant donné les valeurs qu'il prend pour les variables retenues dans le modèle. Toute règle de classement basée sur ce modèle sera de la forme :

- si $P \geq p_0$ l'individu est classé « maigre (DCE) ».
- si $P < p_0$ l'individu est classé « non maigre (normal) ».

La courbe ROC résultant de l'utilisation de l'échantillon test est la suivante :

Graphique 5

Courbe ROC régression logistique. Données tunisiennes



Le choix d'un compromis entre une bonne sensibilité et un taux de faux négatifs relativement faible conduit au choix du point de coordonnées $(1 - \hat{Sp} ; \hat{Se}) = (0,38 ; 0,61)$. Ce choix correspond à une valeur p_0 de 0,04. La table de concordance associée à cette règle de classement est la suivante (avec les notations M : maigre, N : non maigre, R+ : classé maigre par la règle, R- : classé non maigre) :

Tableau 5

Table de concordance. Régression logistique. Données tunisiennes

	M	N	
R+	32	336	368
R−	20	552	572
	52	888	940

Les qualités de la règle sont :

$$\hat{Se} = 0,61 \ \hat{Sp} = 0,62 \ k = 1.62$$
$$\hat{VPP} = 0,09 \ \hat{VPN} = 0,97 \ \hat{k}' = 2.48$$

– Construction du score de ciblage

A partir de transformations simples des coefficients du modèle logistique, on peut construire un score de ciblage facile d'emploi :

Tableau 6

Score de ciblage. Régression logistique. Données tunisiennes

Score de départ	+100
Age < 20 ans	+99
Pas d'activité ou activité légère	+57
Enfant ou petit enfant du chef de ménage	+52
Pas de radio	+58
Pas de machine à laver	+106
Nombre d'enfants du ménage dans le secondaire public ≥ 2	+52
Total	

Ce tableau peut servir de base à l'élaboration d'un questionnaire utilisable par les personnels de terrain en charge de la récolte des informa-

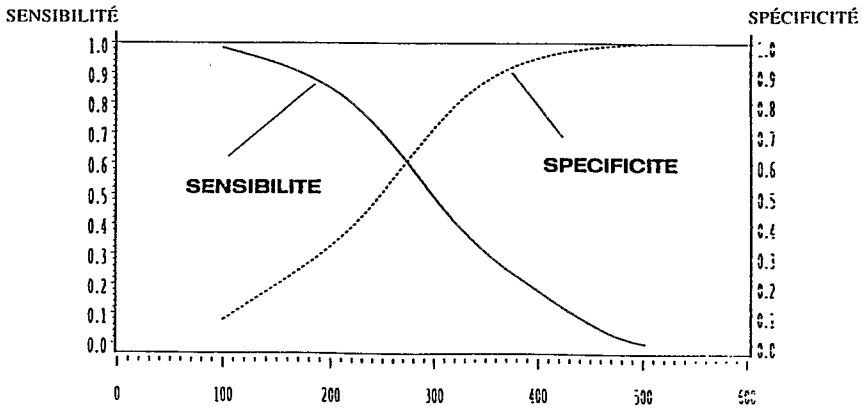
tions pour le ciblage. Avec le choix de spécificité et de sensibilité fait ci-dessus, la règle de classement est :

$$\text{score} \geq 300 \rightarrow \text{risque de maigreur}$$

Si on se place du point de vue du décideur, celui-ci désirera sans doute pouvoir choisir lui même son seuil de ciblage en fonction des ressources dont il dispose et/ou de l'importance relative qu'il attache aux caractéristiques de sensibilité et de spécificité. Les abaques suivantes permettent, à partir du score de ciblage défini ci-dessus, de choisir la valeur seuil en fonction du compromis désiré entre sensibilité et spécificité.

Graphique 6

Abaques pour choix du seuil. Régression logistique. Données tunisiennes



Analyse discriminante

– Sélection des variables les plus discriminantes

Les données utilisées sont identiques à celles décrites dans le paragraphe ci-dessus. Sur les 2 198 u.s. de l'échantillon de base, par une analyse factorielle discriminante pas à pas ascendante, on sélectionne les variables les plus pertinentes pour expliquer la partition en deux groupes « maigre », « non maigre ». Les variables retenues sont les suivantes :

– âge < 20 ans (âge 1)
– radio (rad1)

– nombre d'enfants du ménage dans le secondaire public ≥ 2 (spu2)
– activité nulle ou légère (act1)
– machine à laver (mal2)
– chef de ménage homme (scm1)

Les coefficients de la combinaison linéaire discriminante sont les suivants :

Tableau 7

Coefficients de la combinaison linéaire discriminante
Données Tunisiennes

Variable	β_j
age1	2,14
rad1	−0,72
spu2	+0,71
act1	+0,62
mal1	-0,77
scm1	-0,97

Les coordonnées des points moyens des groupes sur l'axe discriminant $\psi = \sum\beta_j X_j$ sont respectivement pour maigre et non maigre, 0,90 et – 0,04.

Les positions respectives des centres de gravité des groupes sur l'axe discriminant, ainsi que le signe des coefficients des variables, permettent de préciser le sens de leur liaison avec la maigreur : par exemple, le coefficient négatif de la variable radio1 (1 : présence d'une radio dans le ménage, 0 si non) indique que la possession d'une radio par le ménage est plutôt favorable. Néanmoins on n'a pas d'interprétation directe des coefficients en terme de risque relatif comme dans le paragraphe précédent.

– Validation de la règle, contruction de la courbe ROC, choix d'un seuil

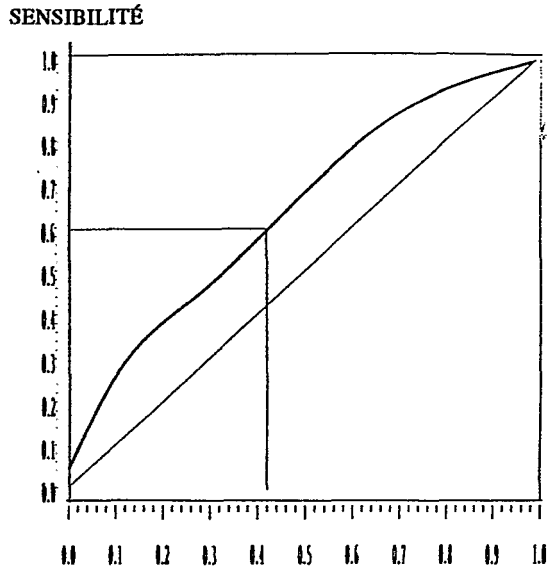
Pour chacun des individus de l'échantillon test, on calcule la valeur $\psi = \sum\beta_j X_j$ correspondant à la coordonnée de sa projection sur l'axe discriminant. Au vu de la position des points moyens sur l'axe discriminant ψ, pour un individu i, une règle de classement sera de la forme :

– si $\psi_i > a$, l'individu est classé « maigre »,
– si $\psi_i < a$, l'individu est classé « non maigre ».

La courbe ROC résultante est la suivante :

Graphique 7

Courbe ROC. Analyse discriminante



On fait le choix du point de coordonnées (0,42 ; 0,60) qui assure un bon compromis sensibilité/spécificité. Les qualités de la règle de classement qui en résulte sont les suivantes (avec les mêmes notations que dans le paragraphe précédent) :

Tableau 8

Matrice de concordance. Analyse discriminante. Données tunisiennes

	M	N	
R+	31	373	404
R–	21	515	536
	52	888	940

D'où :

$$\hat{Se} = 0,60 \; \hat{Sp} = 0,58 \; \hat{k} = 1,41$$
$$\hat{VPP} = 0,08 \; \hat{VPN} = 0,95 \; \hat{k}' = 1,93$$

– Construction du score de ciblage

A partir de transformations affinées des coefficients de la combinaison linéaire discriminante, on construit le score de ciblage :

Tableau 9

Score de ciblage. Analyse discriminante. Hommes ville

Score de départ	200
Age < 20 ans	+214
Radio	-72
Nombre d'enfants du ménage dans le secondaire public ≥ 2	+71
Pas d'activité ou activité légère	+62
Machine à laver	-77
Chef de ménage homme	-97
Total	101

De la même façon que précédemment, cette grille de score peut être directement transcrite sous forme d'un questionnaire utilisable sur le terrain (6). Avec le choix du seuil fait ci-dessus, la règle de classement basée sur le score de ciblage est la suivante :

$$\text{score} \geq 90 \rightarrow \text{risque de maigreur.}$$

De la même façon que pour le score construit à partir du modèle de régression logistique, on pourrait fournir des abaques permettant à un responsable de choisir le seuil et donc la règle de ciblage en fonction du compromis recherché entre sensibilité et spécificité.

Arbre de discrimination (CART)

– Ajustement du modèle et sélection des variables

Sur les individus de l'échantillon de base, on construit un arbre de discrimination binaire par segmentation récursive. L'arbre est ensuite élagué par le bas par utilisation de l'échantillon test pour obtenir l'arbre « opti-

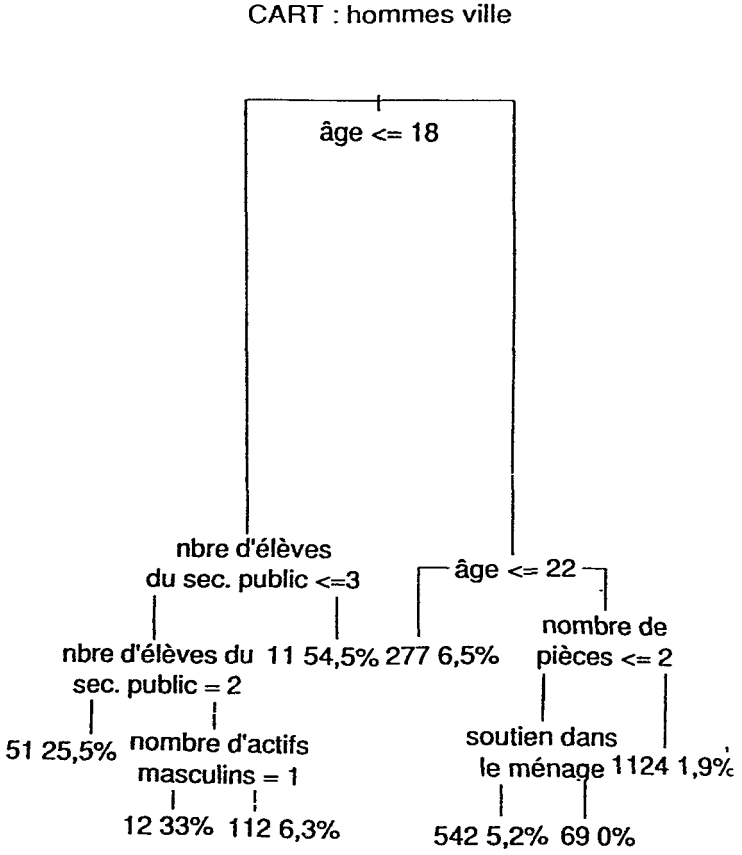
(6) Deux scores obtenus avec la même grille par deux individus différents permettent de comparer leurs risques de maigreur relatifs. Cependant, il n'est pas possible de comparer directement deux scores obtenus pour un même individu avec deux grilles différentes (par exemple celle-ci et celle du paragraphe précédent), dans la mesure où les méthodologies sous-jacentes et donc l'interprétation des scores et des seuils sont différentes.

mal ». Dans cette analyse, de la même façon que dans les deux précédentes, on cherche les variables les plus explicatives de la maigreur dans le cadre d'un certain modèle. La différence est que le modèle a une forme d'« arbre » au lieu d'être basé sur des combinaisons linéaires des variables initiales.

Les variables retenues ainsi que la forme du modèle sont résumés dans l'arbre suivant :

Graphique 8

Arbre de discrimination. Données tunisiennes



L'interprétation est la suivante : pour un individu de sexe masculin vivant en milieu urbain, la variable la plus explicative de la maigreur est l'âge avec la coupure optimale âge ≤ 18 ans. Pour un individu

d'âge ≤ 18 ans, la variable la plus explicative de la maigreur est le nombre de personnes du ménage qui sont scolarisées dans le secondaire public. Pour un individu d'âge > 18 ans, la coupure optimale suivante est basée sur l'age avec un seuil à 22 ans, etc.

Pour chacun des nœuds terminaux est figurée une estimation de la probabilité de maigreur. Par exemple, pour un individu de plus de 22 ans, vivant dans un logement de 2 pièces ou moins et dont le soutien financier est hors du ménage l'estimation de la probabilité de maigreur est de 0. Par contre, pour un individu de 18 ans, vivant dans un ménage comprenant 3 enfants ou plus scolarisés dans le secondaire public, la probabilité de maigreur est estimée à 0,545.

– Validation de la règle, construction de la courbe ROC, choix du seuil

Pour chacun des individus de l'échantillon test, on peut estimer une probabilité de maigreur P, par descente le long de l'arbre comme expliqué ci-dessus. Une règle de classement basée sur ce modèle sera donc de la forme :

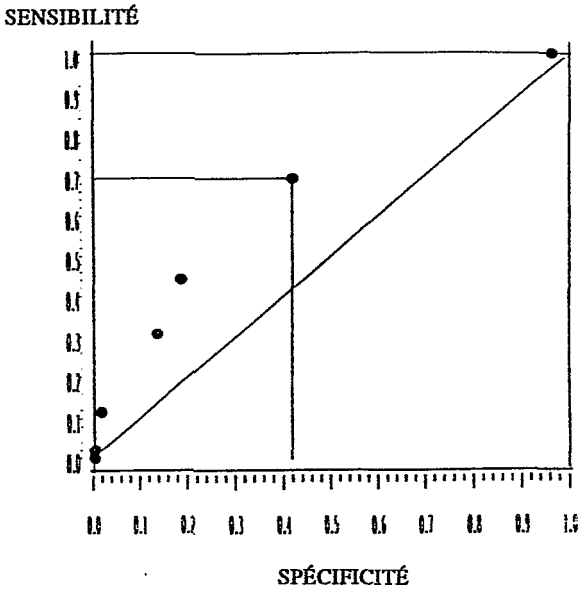
– si $P \geq p_0$, individu classé maigre,

– si $P < p_0$, individu classé non maigre.

La courbe ROC est la suivante :

Graphique 9

Courbe ROC CART. Données tunisiennes



L'examen de cette « courbe » (7) conduit au choix du point de coordonnées (0,42, 0,69). Les qualités de la règle de décision associée à ce choix sont les suivantes :

$$\hat{Se} = 0{,}69\ \hat{Sp} = 0{,}58\ \hat{k} = 1{,}64$$
$$\hat{VPP} = 0{,}09\ \hat{VPN} = 0{,}97\ \hat{k'} = 2{,}83$$

Tableau 10

Matrice de concordance. CART. Données tunisiennes

	M	N	
R+	36	372	408
R–	16	516	532
	52	888	940

– Arbre de décision pour le ciblage

L'arbre de décision qui pourra être utilisé par les personnels de terrain pour le ciblage se déduit immédiatement de l'arbre présenté plus haut et du choix du seuil. Avec le choix de sensibilité et de spécificité fait, l'arbre de décision est celui présenté sur le graphique 10.

L'utilisation est extrêmement simple et intuitive pour les personnels de terrain. Aucun calcul n'est nécessaire. Il n'y a pas d'expression algébrique à manipuler. Un individu chemine le long de l'arbre en fonction de ses caractéristiques jusqu'à ce qu'il soit affecté à un nœud terminal. Suivant que ce nœud terminal est étiquetté « N » ou « DCE », l'individu est ou n'est pas considéré comme à risque de maigreur.

Caractéristiques et performances relatives des 3 méthodes

La comparaison des différentes méthodes peut s'envisager selon deux points de vue non nécessairement convergents :

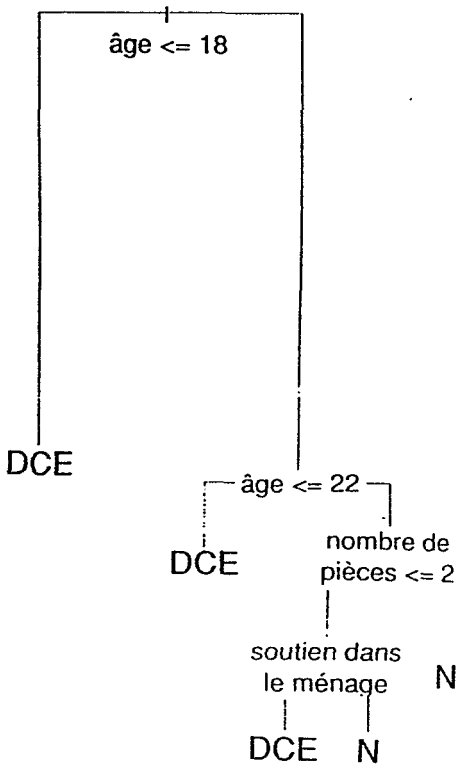
– la lisibilité et l'interprétabilité des résultats fournis, propres à chaque méthode et aux mathématiques qui la sous-tendent ;

(7) Le nombre de valeurs différentes possibles pour la probabilité de maigreur estimée à partir de l'arbre présenté ci-dessus est faible (7), il en est donc de même pour les possibilités de choix du seuil de coupure P_0. A chacun de ces seuils est associé un point de la courbe ROC. Contrairement aux deux cas précédents où nous avions présenté des courbes ROC lissées, il nous semble plus adéquat dans ce cas de faire apparaître les couples (1 – Sp, Se) bruts et non une courbe ajustée qui n'aurait pas grand sens vu le faible nombre de points.

Graphique 10

Arbre de décision pour le ciblage

CART : *hommes ville*



– les performances en terme de ciblage, i.e. pour la prédiction correcte de la maigreur dans le cas de l'exemple traité ici.

Le premier point a été largement abordé lors de la description des méthodes dans les méthodes statistiques. Dans les exemples présentés ci-dessus, les avantages de chaque méthode pour ce qui concerne la lisibilité des résultats apparaissent nettement. Par exemple :

– interprétation des coefficients en terme de risques relatifs dans la régression logistique ;

– modélisation sous forme d'arbre, des relations entre maigreur et variables explicatives dans la méthode CART.

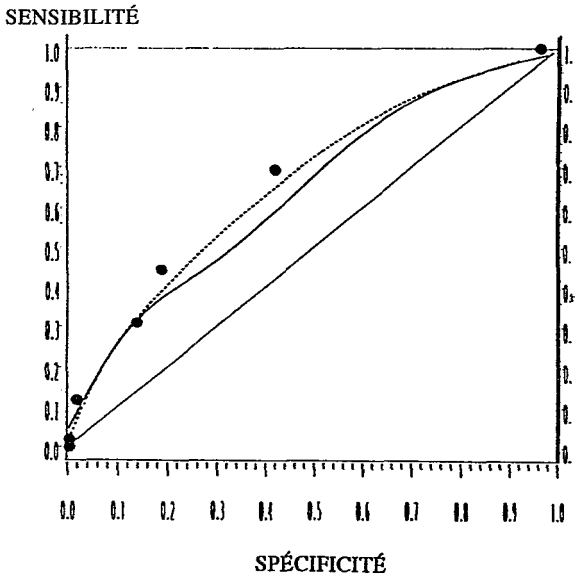
Cependant, rien ne garantit qu'une méthode fournissant des résultats plus lisibles qu'une autre soit meilleure du point de vue de la prédiction du risque de maigreur.

Un certain nombre d'auteurs ont comparé les performances des méthodes de discrimination entre deux populations sur de nombreux jeux de données réelles et simulées. Notre objectif est ici plus modeste, il s'agit simplement de donner une illustration des résultats fournis par ces méthodes sur un jeu de données du type de ceux qui pourraient être utilisés dans une optique de ciblage nutritionnel. Il faudra donc être prudent avant de généraliser les conclusions concernant les performances respectives des différentes méthodes.

Le graphique ci-dessous, fait apparaitre les courbes ROC correspondant aux trois méthodes utilisées sur les données homme ville :

Graphique 11

Comparaison des courbes ROC



Il apparaît que les résultats des trois méthodes en terme de performances sont tout à fait comparables. La méthode de discrimination par arbre binaire (CART) semble donner des résultats aussi bons, sinon meilleurs que l'analyse discriminante ou la régression logistique. Le

même genre de remarque peut être fait pour les trois autres jeux de données « hommes rural », « femmes ville », « femmes rural » (dont nous ne présentons pas les résultats) sur lequels la méthode CART donne des résultats au moins aussi bons que les deux autres.

Un point de vue raisonnable paraît être le suivant : à performances égales, choisir la méthode qui donne la plus grande simplicité de lecture des résultats. Si on adopte ce point de vue, la méthode de discrimination par arbre binaire semble bien placée. En plus d'un certain nombre d'avantages du point de vue de ses qualités « statistiques » (par exemple la prise en compte automatique des interactions, la modélisation des relations entre maigreur et les variables explicatives sous forme d'arbre est un atout. En effet, les arbres de décision tels que ceux présentés ci-dessus sont intuitifs et très faciles d'emploi aussi bien pour un gestionnaire de programme que pour les personnels de terrain.

Néanmoins, les quelques exemples que nous avons analysés ne constituent pas une garantie de la supériorité absolue d'une méthode par rapport à l'autre. Il n'y a certainement pas de méthode universelle.

Conclusion

Nous avons présenté trois méthodes parmi les plus classiquement utilisées pour la discrimination et le classement. Il en existe un certain nombre d'autres, par exemple la discrimination par réseaux de neurones dans le domaine de l'intelligence artificielle.

D'autre part, il est clair que la comparaison de ces méthodes sur d'autres données est nécessaire pour en percevoir mieux les avantages et inconvénients relatifs.

Néanmoins, quelle que soit la méthode utilisée, dans l'optique de la construction d'outils de ciblage opérationnels, il nous paraît important d'insister sur les point suivants :

– aucune méthode aussi performante soit-elle ne peut donner de bons résultats en terme de ciblage si le modèle n'est pas calé sur des données fiables, récentes et représentatives de la population dans laquelle on désire opérer le ciblage ;

– le choix des u.s. à cibler (individus, ménages) et des descripteurs utilisables pour cela est très important. Se pose en particulier le problème de savoir donner des descripteurs nutritionnels satisfaisants au niveau de

l'entité ménage, si ce sont ces derniers que l'on veut cibler au lieu des individus eux mêmes ;

– quelle que soit la méthode utilisée pour le construire, un outil de ciblage doit être évalué suivant des critères objectifs et précis en terme de couverture (sensibilité), taux de ciblage (valeur prédictive positive) etc. (voir le texte de B. Maire *et al.* dans ce même ouvrage) ;

– à performances égales, il faut préférer l'outil le plus simple dans l'expression de ses résultats et le plus intuitif à utiliser sur le terrain ;

– avant utilisation opérationnelle, la validation du critère de ciblage sur le terrain est indispensable. L'utilisation d'un échantillon test extrait des données disponibles, comme nous l'avons fait ici, est une première approche pour réduire le biais dans l'estimation des performances de la règle de classement. Pour la validation du critère de ciblage avant utilisation opérationnelle, la collecte d'un échantillon indépendant du premier paraît indispensable.

Beaucoup reste à faire dans l'évaluation et la mise au point des outils de ciblage. Le but de cet article n'était certes pas de proposer des solutions toutes faites et opérationnelles mais de montrer quelques-unes des possibilités offertes par un certain nombre de méthodes statistiques classiques. Nous espérons qu'il aura été atteint.

Bibliographie

BOUYER J., « La régression logistique en épidémiologie (Parties 1 et 2) », *Revue d'épidémiologie et de santé publique*, 39, 79-87 et 183-196, 1991.

BREIMAN L., FRIEDMAN J.H., OLSHEN R.A., STONE J.C., *Classification and regression trees*, Californie, Wadsworth Inc., 1984.

California Software Inc., *An introduction to CART methodology*, 1985.

CELEUX G., *Analyse discriminante sur variables continues*, INRIA, 1990.

CHAMBERS J.M., HASTIE T.J., *Statistical Models, in* S. Wadsworth and Brooks, 1992.

GREMY F., DAURES J.P., *Éléments de sciences de l'information : épidémiologie de population, épidémiologie clinique*, Faculté de médecine de Montpellier, 1986.

GUEGUEN A., NAKACHE J.P., « Méthode de discrimination basée sur la construction d'un arbre de décision binaire », *Revue de statistique appliquée* XXXVI(1), 19-38, 1988.

HOSMER D.W., LEMESHOW S., *Applied Logistic Regression*, Wiley, 1989.

MARDONES-RESTAT F., JONES G., MARDONES-SANTANDER F., DACHS N., HABICHT J.P., DIAZ M., « Growth failure prediction in Chile », *International Journal of Epidemiology*, 18, n° 4, S44-S49, 1989.

O'GORMAN T.W., WOOLSON R.F., « Variable selection to discriminate between two groups : stepwise logistic regression or stepwise discriminant analysis ? », *The American Statistician*, vol. 45, n° 3, 187-193, 1991.

SAS Institute Inc., *SAS/STAT User's Guide,* Version 6, volumes 1 et 2, 1989.

TOMASSONE R., DANZART M., DAUDIN J.J., MASSON J.-P., *Discrimination et classement,* Masson, 1988.