

R. ROUSSEAU

B. AUGEREAU

A. DAVER

D. LEGUAY

## **Méthodologie de la régression et de la prédiction fondée sur la classification automatique**

*Les cahiers de l'analyse des données*, tome 16, n° 4 (1991),  
p. 479-488

[http://www.numdam.org/item?id=CAD\\_1991\\_\\_16\\_4\\_479\\_0](http://www.numdam.org/item?id=CAD_1991__16_4_479_0)

© Les cahiers de l'analyse des données, Dunod, 1991, tous droits réservés.

L'accès aux archives de la revue « Les cahiers de l'analyse des données » implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme  
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

# MÉTHODOLOGIE DE LA RÉGRESSION ET DE LA PRÉDICTION FONDÉE SUR LA CLASSIFICATION AUTOMATIQUE [MÉTH. PRÉD.]

R. ROUSSEAU\*, B. AUGEREAU\*\*  
A. DAVER\*\*\*, D. LEGUAY\*\*\*\*

## 1 Problèmes et méthodes

### 1.1 Une situation typique

Beaucoup d'études pratiques suivent le schéma suivant: un ensemble I d'individus est décrit par un ensemble J de variables qu'on peut partager en un ensemble J1 de variables explicatives et un ensemble J2 de variables à expliquer. Le problème est de trouver et d'explicitier des relations (causales ou non) entre les variables de J1 et celles de J2.

Si J2 se réduit à une seule variable, j2, plusieurs méthodes classiques de prévision sont applicables selon le type de la variable j2 et selon les types de variables de J1, dont les suivantes : régression si j2 ainsi que toutes les variables de J1 sont quantitatives, analyse de variance si j2 est quantitative et si toutes les variables de J1 sont qualitatives (on peut se ramener à ce cas, par découpage de variables en modalités, si certaines variables de J1 sont quantitatives), analyse factorielle discriminante si j2 est qualitative et toutes les variables de J1 sont quantitatives, discrimination barycentrique si j2 et toutes les variables de J1 sont qualitatives (on peut, de même, se ramener à ce cas si certaines variables de J1 sont quantitatives)...

Si J2 n'est pas réduit à une seule variable, on peut utiliser les méthodes factorielles pour étudier globalement les interactions entre J1 et J2 (analyse canonique, Analyse en Composantes Principales, Analyse de Correspondance,

---

(\*) Vice-Recteur de l'Université Catholique de l'Ouest, Angers.

(\*\*) Diplômé de l'Institut de Mathématiques Appliquées (Université Catholique de l'Ouest).

(\*\*\*) Médecin, Centre Paul Papin, Angers.

(\*\*\*\*) Médecin, Centre Hospitalier Spécialisé, Saint Gemmes-sur-Loire.

..., selon les cas); et aussi, pour chacune des variables de J2, les méthodes de prévision précédemment citées .

Cependant, un grand nombre de variables de J1 ou de J2 ou un nombre important de modalités d'une variable de J2 peuvent être rédhibitoires selon les méthodes utilisées. L'objet de cet article est de présenter, sur trois exemples, l'expérience que nous avons acquise d'une méthodologie susceptible de s'appliquer dans tous les cas et permettant de traiter simultanément, s'il y a lieu, toutes les variables de J2.

### **1.2 Méthodologie fondée sur la classification automatique**

On effectue une classification des individus de I, soit d'après les variables explicatives de J1, soit d'après les variables à expliquer de J2 . En construisant des classes homogènes quant à l'ensemble des variables explicatives, on peut attendre que les individus d'une même classe aient un comportement voisin vis-à-vis des variables à expliquer. De même, si les classes ont été créées d'après l'ensemble des variables à expliquer, on peut espérer que les individus d'une même classe avaient des caractéristiques communes pour les variables explicatives.

Ayant retenu une partition extraite de la classification effectuée, on crée une nouvelle variable qualitative sur I, à savoir l'appartenance pour un individu à l'une des classes de cette partition. En d'autre termes, l'ensemble des classes retenues pour la partition constitue l'ensemble des modalités de la variable ainsi créée.

Ensuite, on cherche les liaisons ou relations entre cette nouvelle variable et les variables à expliquer dans le premier cas, ou les variables explicatives dans le second. Plusieurs méthodes sont possibles: méthodes élémentaires (tris croisés, tests, etc...), méthodes descriptives multidimensionnelles (Analyse de Correspondance des variables découpées en classes, etc...), méthodes discriminantes (analyse factorielle discriminante, discrimination barycentrique, etc...); ou d'autres méthodes statistiques selon les données (courbes de survie, régression, ...).

On peut également effectuer les deux classifications sur l'ensemble I des individus, à savoir celle fondée sur les variables explicatives et celle fondée sur les variables à expliquer, puis croiser ces deux classifications. On peut encore effectuer des classifications qui ne sont fondées que sur un sous-ensemble de J1 ou de J2.

L'avantage de cette méthode est de dégager des groupes d'individus qui ont soit des caractéristiques basales voisines soit des comportements voisins (que l'on devra tenter de caractériser), sur lesquels on pourra étudier, séparément ou non, les liaisons ou relations avec l'autre ensemble de variables qui n'a pas servi à la constitution de ces groupes. Ainsi, dans le premier cas, on pourra

éventuellement trouver et expliquer des facteurs significatifs intéressant les différences de comportement des classes ou un comportement spécifique d'une classe; ou bien, dans le second cas, des différences de caractéristiques entre classes ou un ensemble de caractéristiques spécifiques d'une classe.

Deux exemples d'analyses de données cliniques illustreront cette démarche:

au §2.1, exemple 1: une étude sur le cancer du sein non métastatique.

au §2.2, exemple 2: évolution psychopathologique de patients ayant présenté un premier épisode délirant

Une variante de la méthodologie déjà décrite consiste à segmenter, au préalable, I en des sous-ensembles définis, (soit par des études statistiques, soit en s'en remettant au jugement d'experts), d'après une ou plusieurs variables de J1 ou de J2, non prises en compte par la suite. Et c'est à chacun de ces sous-ensembles de I qu'on applique la méthode proposée, (sans prendre en compte les données relatives aux autres sous-ensembles).

Dans l'exemple 3, qui fait l'objet du §3, il s'agit d'une estimation de devis par une entreprise industrielle. Il n'y a dans J2 qu'une seule variable, qui est le montant du devis. Des experts ont déterminé six sous-populations de I; et l'on travaille ensuite uniquement, pour effectuer séparément la classification de chaque sous-population, d'après un croisement de deux autres variables de J1 qui jouent un rôle essentiel dans la politique commerciale.

## **2 Analyse de données cliniques**

### **2.1 Exemple 1: Étude sur le cancer du sein non métastatique**

Notre travail a fait l'objet d'une communication de A. DAVER et R. ROUSSEAU au 12-th Breast Cancer Symposium tenu à San Antonio (Texas, USA) les 8 et 9 décembre 1989.

La présente étude, multicentrique et multiparamétrique, porte sur une population de malades atteintes de cancer du sein non métastatique. Au cours d'une période allant d'octobre 1981 à octobre 1989, 1047 prélèvements, provenant de 30 hôpitaux et cliniques privées, ont été collectés après chirurgie et congelés dans l'azote liquide.

768 dossiers ont permis de rechercher des corrélations entre de nouveaux paramètres biologiques déterminés par une technique de cytométrie en flux (contenu en ADN, pourcentage de cellules en phase de synthèse) et les facteurs pronostiques clinico-biologiques classiques (l'état hormonal, le stade UICC, la taille histologique de la tumeur, les dosages de récepteurs hormonaux, l'envahissement ganglionnaire, le grade et l'histologie).

354 dossiers complets ont été retenus pour une étude de la survie globale et de la survie sans événement.

I représente un ensemble de 768 patientes atteintes d'un cancer du sein non métastatique.

J1 est un ensemble de variables pronostiques relevées lors de la première consultation; l'ensemble hétérogène J1 se compose de 15 variables quantitatives et de 10 variables qualitatives énumérées ci-après:

âge de la patiente  
 nombre de ganglions envahis  
 taille chirurgicale de la tumeur  
 index d'ADN  
 pourcentage de cellules supérieures à 2n  
 pourcentage de cellules en phase de synthèse  
 pourcentage de cellules en phase de synthèse et mitose et prémitose  
 poids de la tumeur  
 quantité d'ADN  
 nombre de fentomoles de récepteur estradiol par gramme de tumeur  
 nombre de fentomoles de récepteur estradiol par mg de protéines  
 nombre de fentomoles de récepteur estradiol par milligramme d'ADN  
 nombre de fentomoles de récepteur progestérone par gr. de tumeur  
 nombre de fentomoles de récepteur progestérone par mg de protéines  
 nombre de fentomoles de récepteur progestérone par mg d'ADN

ménopause  
 localisation de la tumeur  
 quadrant  
 grade  
 taille clinique  
 adénopathies régionales  
 métastases à distance  
 histologie  
 récepteur estradiol + ou - (selon que la valeur observée est supérieure ou inférieure à la valeur moyenne)  
 récepteur progestérone + ou - (comme pour le récepteur estradiol)

J2 est un ensemble de variables discrètes caractérisant la durée et les conditions de la survie éventuelle (avec ou sans événement) sur une période allant de 1 à 7 ans.

On a de plus un ensemble J3 constitué de variables de traitement thérapeutique différents, pris seuls ou en association selon des protocoles définis : radiothérapie, chimiothérapie, hormonothérapie, chirurgie. On note cet ensemble J3 pour le différencier de l'ensemble J1 des variables pronostiques, même si on peut l'inclure dans J1 vis-à-vis de l'ensemble J2.

L'ensemble I des patientes peut être considéré comme un ensemble représentatif, quant aux variables pronostiques généralement admises, pour l'étude du cancer du sein.

Une ACPN (composantes principales normées), sur les 354 patientes sans données manquantes et sur les 15 variables quantitatives, a montré l'existence de quatre groupes de variables non ou peu corrélés entre eux, déterminés par les quatre premiers axes expliquant 57% de l'inertie totale :

groupe des récepteurs progestérone

groupe des récepteurs estradiol

groupe de prolifération cellulaire (pourcentages de cellules, index d'ADN)

groupe d'importance de la tumeur (nombre de ganglions, taille, masse)

La variable âge n'intervient que par sa liaison avec le récepteur estradiol. Une autre ACPN, sur toutes les patientes en donnant aux valeurs manquantes les moyennes observées, donne les mêmes résultats, à l'ordre des axes près.

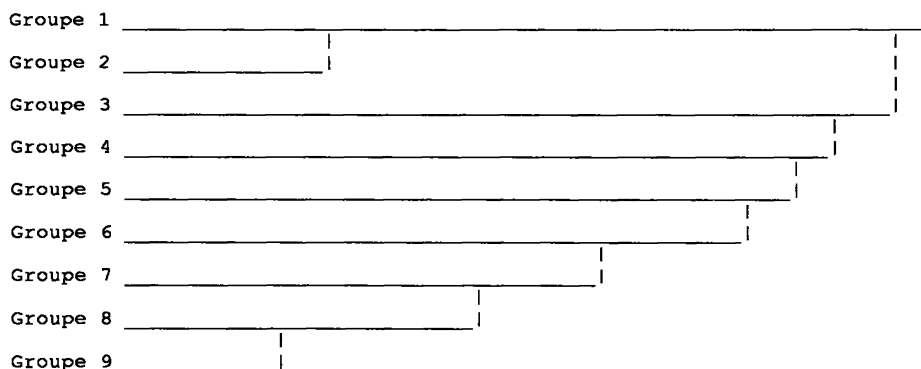
Une Analyse Factorielle de Correspondance a été faite ensuite sur les principales variables de pronostic, découpées en classes: ménopause, grade, combinaison des récepteurs estradiol et progestérone (++, +-, -+, --), envahissement ganglionnaire, index d'ADN et pourcentage de cellules en phase de synthèse (reprenant les quatre groupes déterminés par l'ACPN et ajoutant deux variables qualitatives), les autres variables étant mises en supplémentaire.

Le découpage en classes des variables quantitatives a été fait sur des critères médicaux et ne donne donc pas des classes d'effectifs égaux. L'analyse a porté sur le fichier des 331 patientes, sans données manquantes.

Les quatre premiers axes expliquent 58% de l'inertie du tableau de BURT et montrent des liaisons fortes ou des non-liaisons entre modalités des six variables principales. L'axe 1 a été interprété comme axe de gravité de la maladie ... et l'axe 4 comme l'axe des envahissements ganglionnaires extrêmes, ce qui a été une surprise (comme pour l'ACPN) car le nombre de ganglions envahis est un critère important pour le choix de la thérapie.

Une CAH effectuée sur ces six variables, (distance du Chi2 et critère de variance) a défini nettement 9 groupes de patientes, qui sont ici simplement désignés par un chiffre de 1 à 9. Sur le schéma publié ici, on voit, au sommet de la hiérarchie, une dichotomie entre 1 ∪ 2, d'une part, et 3 ∪ 4 ∪ 5 ∪ 6 ∪ 7 ∪ 8 ∪ 9, d'autre part.

Il est à noter que chacun des groupes est caractérisé par une et une seule modalité des six variables.



Le groupe 1  $\cup$  2 est un groupe de très bon pronostic. Une fois fait le partage entre les groupes 1  $\cup$  2, d'une part, et 3 $\cup$ 4 $\cup$ 5 $\cup$ 6 $\cup$ 7 $\cup$ 8 $\cup$ 9, d'autre part, la lecture descendante de l'arbre est très simple.

En effet, on constate (sauf pour seulement 7 des 331 patientes) que l'appartenance aux groupes 3 à 9 se fait sur les modalités caractéristiques de ces groupes dans l'ordre 3-4-5-6-7-8-9: c'est-à-dire que, parmi les patientes du groupe 3 $\cup$ 4 $\cup$ 5 $\cup$ 6 $\cup$ 7 $\cup$ 8 $\cup$ 9, sont dans le groupe 3 celles qui possèdent la modalité caractéristique du groupe 3; sont dans le groupe 4 celles qui ne sont pas dans le groupe 3 et qui possèdent la modalité caractéristique du groupe 4; ...etc...

La CAH une fois faite, des courbes de survie globale et des courbes de survie sans évènements ont été faites sur les 9 groupes de la CAH en se restreignant aux patientes dont on connaissait les données de survie. Les courbes de survie trouvées sont très différentes et, semble-t-il, assez caractéristiques, mais n'ont pu être comparées significativement entre elles du fait de faibles effectifs. Néanmoins, le monde médical international y a vu un intérêt certain quant à la démarche et aux résultats trouvés.

De même, le faible effectif n'a pas permis de savoir l'influence de la thérapie sur la survie dans chacun des groupes. Mais cette méthodologie permet de le faire si l'on part d'un nombre suffisant de patientes. En effet, la classification permet la constitution de groupes homogènes de patientes basée uniquement sur les variables pronostiques établies lors de la première consultation et sans a priori.

Placer ensuite une nouvelle patiente dans l'un de ces groupes est possible par des méthodes d'affectation d'un élément supplémentaire à l'un des groupes de la classification obtenue. Si on connaissait la ou les meilleures thérapies pour chacun de ces groupes, la patiente recevrait ainsi les meilleurs soins. Cette méthodologie peut donc permettre d'améliorer le protocole de thérapie, à

condition de constituer un corpus suffisant de dossiers pour déterminer les thérapies convenant le mieux à chacun des groupes de la classification.

Une autre étude a été faite en découpant en classes les variables quantitatives récepteur estradiol et récepteur progestérone, au lieu de considérer la combinaison de ces deux récepteurs (++,+,-,+,-). Les résultats obtenus sont très similaires à ceux précédemment énoncés.

Il apparaît, sur ce premier exemple, que la méthodologie proposée offre une nouvelle approche préférable à celle classiquement utilisée par les médecins: constitution de groupes *a priori*, puis utilisation de tests statistiques (en général de Cox) pour en étudier la significativité. En plus des limites inhérentes à ces méthodes, l'approche classique se trouve souvent confrontée à un problème combinatoire presque insoluble du fait du grand nombre des variables pronostiques.

## 2.2 Exemple 2: Évolution psychopathologique de patients ayant présenté un premier épisode délirant

Il s'agit d'un travail en cours avec le Dr Denis LEGUAY du C.H.S. de Saint-Gemmes (Maine et Loire). L'objet de l'étude est d'apprécier la pertinence des stratégies thérapeutiques et d'aides sociales mises en œuvre pour traiter un patient après un premier épisode délirant, inaugurant bien souvent une longue pathologie. On cherchera donc à objectiver l'impact éventuel sur l'état du patient de chacune des mesures prises.

Ces mesures peuvent être d'ordre thérapeutique: nature du traitement (psychothérapeutique, chimiothérapeutique, de réadaptation, ...), modalités du soin (hospitalisation, consultation ambulatoire, visites à domicile, ...); et varier d'intensité (fréquence des hospitalisations, fréquence des consultations, ...). Elles peuvent être également d'ordre social: mesures d'assistance juridique (tutelle, curatelle, ...), mesures d'assistance financière (pension d'invalidité, allocation pour adulte handicapé, ...).

Mesurer l'influence de ces dispositions suppose que l'on puisse la discriminer des autres influences que l'enquête devra identifier. Les événements susceptibles d'avoir un retentissement sur l'évolution de la pathologie seront donc également pris en compte. Leur recensement s'opérera à partir d'une sélection reprise d'enquêtes internationales. L'évolution de la pathologie sera, quant à elle, mesurée à l'aide d'échelles de psychopathologie validées (système AMDP).

Il s'agit donc d'une enquête qui s'inscrit dans le cadre général de l'évaluation des soins.

Il représente un ensemble d'environ 300 patients ayant présenté un premier épisode délirant.



J1 est un ensemble de variables évènementielles et thérapeutiques mesurées en 6 examens répartis sur 30 mois.

J2 est un ensemble de variables d'évolution psychopathologiques (6 examens sur 30 mois).

La méthodologie proposée consistera à effectuer une classification de patients sur l'ensemble des variables psychopathologiques afin d'obtenir des groupes de patients d'évolutions psychopathologiques voisines. Ensuite, pour chacun de ces groupes, on recherchera et on étudiera les liaisons avec les évènements ou les thérapies pour savoir si celles-ci ont une incidence significative sur l'évolution psychopathologique. Ceci ne permettra pas de faire une prévision comme dans l'exemple 1, mais permettra de mieux comprendre les types et les risques potentiels d'évolution psychopathologique.

On pourra aussi, selon la méthodologie proposée dans cet article, constituer des groupes de patients basés sur les variables évènementielles et/ou thérapeutiques. Cependant, le nombre de ces variables, très grand par rapport au nombre de patients, laisse craindre que les groupes obtenus ne soient pas suffisamment homogènes pour pouvoir tirer des conclusions quant à leur évolution psychopathologique. Cela sera peut être néanmoins possible après avoir trouvé les évènements ou les thérapies les plus liés à l'évolution psychopathologique. Il est à noter aussi que l'aspect prévisionnel est exclu, comme dans la démarche précédente, du fait de la non-connaissance à la première consultation d'évènements pouvant survenir ensuite.

D'autres méthodes classiques de traitement sont envisagées, en particulier la constitution de groupes-types de patients d'une évolution donnée (évènementielle et/ou thérapeutique) pour en étudier l'évolution psychopathologique en les comparant au reste de la population des patients. On retrouve dans cette démarche celle classique donnée dans l'exemple 1 et la même différence avec la méthodologie proposée dans cet article.

### **3 Estimation de devis par une entreprise industrielle**

Ce troisième exemple a fait l'objet du stage de maîtrise MASS de Bernard AUGEREAU, étudiant à l'Institut de Mathématiques Appliquées d'Angers, en février-mars-avril 1989. Pour des raisons de confidentialité, les données sont décrites ici sous une forme modifiée; et de ce fait elles ont perdu un peu de leur caractère concret.

Une Société de l'Ouest de la France, filiale d'un groupe, fabrique et commercialise, notamment, le produit A. Elle exporte 55% de son chiffre d'affaires dans plus de 20 pays et dans des secteurs d'activités très divers (bâtiment et travaux publics, manutention et levage, industrie forestière, ...). La concurrence dans ces secteurs est très sévère, et le problème crucial est de déterminer le prix de vente à proposer à une commande donnée. Le meilleur prix

n'est pas nécessairement le plus élevé possible (car avec un tel prix, le marché risque d'être perdu); ni le plus faible (car la marge bénéficiaire risque d'être insuffisante; et un prix trop bas vis-à-vis de celui des concurrents, rend le client méfiant, d'où une perte de crédit); il doit être le plus adapté aux possibilités de la Société et de ses concurrents.

La connaissance du marché montre que le prix du produit A dépend des caractéristiques mêmes de la pièce (poids, matière, phase de vie de la pièce, ...), du client (pays, secteur d'activités, type du client, historique dans la société, quantité commandée, ...) , de facteurs de production de la Société et de facteurs concernant la concurrence.

L'objet de l'étude était de donner au responsable concerné une aide à la décision en proposant, pour une commande potentielle (ou demande de prix) donnée d'un certain type du produit A, une estimation du prix à proposer, des informations diverses sur le cas traité (marge,...) et si possible des éléments de comparaison avec les concurrents. On s'intéressera uniquement ici à l'estimation du prix de marché (par la Société ou par un concurrent).

I représente l'ensemble des commandes (une commande par type de pièce) évaluées à 15000 sur trois ans d'activité.

J1 est la réunion de l'ensemble J11 des caractéristiques de la pièce, de l'ensemble J12 des caractéristiques du client et de l'ensemble J13 des variables de production de la Société.

J2 est constitué d'une seule variable : le prix unitaire proposé de la pièce.

La méthode définitive n'a été adoptée qu'après de longues discussions, en tenant compte des résultats de quelques études préliminaires. Sur les conseils d'experts, on a retenu une classification des pièces en 6 types, couvrant 90% des cas de prix à déterminer (mis à part ceux dont la détermination est simple). Cette classification, basée essentiellement sur trois variables caractéristiques de la pièce (masse, matériaux, temps de fabrication) faisant partie de l'ensemble J1, a servi de base *a priori* pour notre traitement statistique.

La première étape de ce traitement a été d'effectuer, pour chaque type de pièces, une classification (CAH avec distance du Chi2 et critère de la variance). Celle-ci a été faite sur un tableau d'abondance issu d'un tableau de contingence croisant, d'une part, l'ensemble des couples pays × secteurs d'activités, et, d'autre part, l'ensemble des classes de prix unitaires pratiqués; une case de ce tableau de contingence représentant le nombre de pièces vendues (du type considéré) dans une classe de prix unitaire et dans un couple pays × secteur d'activités donné.

Afin d'éviter de donner trop de poids à une case du fait d'une commande importante d'un seul client, le tableau de contingence a été transformé en tableau

d'abondance: c'est-à-dire qu'on a remplacé les nombres bruts par un indice d'abondance ne prenant que quelques valeurs entières; disons de 0 (nul) à 5 (très élevé). Ainsi, la classification a permis de trouver, pour chacun des types de pièces, des groupes de couples pays-secteurs d'activités homogènes (quant aux prix unitaires pratiqués) qui ont servi de base à l'estimation du prix unitaire d'une nouvelle commande.

Ceci posé, pour une nouvelle commande du produit A, on pratique le même prix ou le prix actualisé lorsque la pièce a déjà été vendue dans les trois années précédentes au même client; sinon, à condition que la pièce rentre dans l'un des 6 types, on s'engage dans la voie décrite ci-après:

si on a un nombre suffisant de commandes passées pour ce type de pièces dans le même pays et le même secteur d'activités, on pratique un prix estimé par régression (pas-à-pas ascendante) selon certaines caractéristiques de la pièce (sous-ensemble de J11), du client (sous-ensemble de J12) et de la Société (sous-ensemble de J13) à partir de cet ensemble de commandes passées.

si ce nombre est insuffisant, on élargit l'ensemble des commandes prises en considération jusqu'à obtenir une base d'estimation suffisante pour appliquer la même régression que précédemment:

on ajoute d'abord les commandes passées pour le même type de pièces, à partir du même secteur d'activités et d'un pays du même groupe pays  $\times$  secteurs d'activités que celui du client.

on considère ensuite, au besoin, pour appliquer la régression, les commandes du même type de pièces, du même pays et d'un secteur d'activités du même groupe pays  $\times$  secteurs d'activités que celui du client.

si le nombre des cas ne suffit pas, on élargit encore l'ensemble précédent en y ajoutant les commandes passées du même type de pièces à partir d'un couple pays  $\times$  secteur d'activités rentrant dans le même groupe pays  $\times$  secteurs d'activités que celui du client.

Si, malgré tous les efforts effectués pour recruter des cas analogues à celui proposé, on n'en a pas trouvé un nombre suffisant, on renonce à faire une estimation statistique, par régression; et le choix du prix dépend de la seule politique commerciale de la Société. Telle est également la seule issue pour une pièce qui n'appartient pas à l'un des 6 types reconnus *a priori*, par exemple pour un prototype.

La méthode a donné de très bonnes estimations tant pour décider du prix à pratiquer pour la Société que pour prévoir celui arrêté par les concurrents (à condition, cela va sans dire, que des données relatives à ceux-ci soient disponibles).

Ceci nous incite à entreprendre d'autres études dans la voie explorée ici.