JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

MICHEL AUBERGER

Un modèle simple permettant de prévoir les comportements financiers d'une population multidimensionnelle

Journal de la société statistique de Paris, tome 137, n° 3 (1996), p. 79-102

<http://www.numdam.org/item?id=JSFS_1996__137_3_79_0>

UN MODÈLE SIMPLE PERMETTANT DE PRÉVOIR LES COMPORTEMENTS FINANCIERS D'UNE POPULATION MULTIDIMENSIONNELLE

Michel AUBERGER
Ingénieur Civil des Mines
Docteur en Sciences Economiques

Résumé

L'objet de cet article est d'élaborer une synthèse de travaux empiriques de scoring effectués dans le domaine du crédit à la consommation et de montrer qu'un modèle simple, fondé sur un double effet Guttman, permet d'expliquer les résultats obtenus. Ce modèle est par ailleurs compatible avec les résultats d'enquêtes plus générales sur le comportement financier des ménages. On en déduit qu'il est possible d'étendre le champ d'application de ces méthodes à d'autres comportements financiers ; un exemple est donné dans le secteur de l'assurance-dommages.

I. INTRODUCTION

On désigne généralement sous le terme de scoring les méthodes statistiques permettant, à partir d'analyses de données individuelles provenant de fichiers ou d'enquêtes, de déterminer pour chaque individu un score pouvant servir d'indicateur prévisionnel du phénomène étudié.

Ces méthodes sont largement utilisées dans le domaine du marketing, notamment pour cibler ou segmenter, et elles ont été étendues progressivement, depuis vingt-cinq ans environ, dans le domaine des comportements financiers : analyse du risque de crédit ("credit scoring"), analyse du potentiel de fidélisation ("scoring de comportement"), analyse de la sensibilité à une action commerciale, à une action de recouvrement ou à une tarification.

Elles ont largement bénéficié du développement des méthodes d'analyse de données, mais ont jusqu'à présent peu fait l'objet de publications, ce qui a nui à l'approfondissement de la théorie sous-jacente et à l'extension de leur application dans la gestion des entreprises.

Journal de la Société de Statistique de Paris, tome 137, n° 3, 1996.

Pour aborder la théorie des méthodes de scoring, il convient, nous semble-t-il, de distinguer trois niveaux d'analyse :

- analyse descriptive des comportements individuels à partir d'un fichier ou d'une enquête, qui peut être réalisée à l'aide d'une analyse factorielle des correspondances multiples sur variables quantitatives ou qualitatives, à partir d'un tableau logique ou d'un tableau de Burt [1] ;
- calcul d'un indicateur prévisionnel du phénomène étudié à l'aide d'un score (codage des caractéristiques de l'individu), effectué à partir d'un tableau de contingence ;
- intégration du score individuel dans un système décisionnel (octroi de crédit, gestion du recouvrement, fidélisation, tarification) et analyse des écarts par rapport aux prévisions.

Ces trois niveaux sont interdépendants et leurs résultats obtenus globalement dépendent essentiellement de leur cohérence.

Cet article est consacré au premier niveau, pour lequel nous avons des représentations géométriques simples et qui sert de base au calcul du score.

Le deuxième niveau a gagné en efficacité grâce à l'introduction de méthodes prenant en compte de façon plus satisfaisante les variables qualitatives [2], les relations non linéaires [3] et l'évolution temporelle [4].

Le troisième niveau dépend principalement de l'histoire de l'entreprise, de son implantation sur ses marchés et de sa volonté stratégique de modifier son organisation pour mieux répondre aux besoins de sa clientèle. Le développement de l'informatique a permis aux scorings de s'implanter, en facilitant la décentralisation des décisions et le management de l'organisation. Des gains de productivité ont pu être réalisés.

Les pouvoirs publics souhaitent aujourd'hui, dans une conjoncture difficile, restaurer la rentabilité des banques, en réduisant notamment le coût de leur risque de crédit, afin de faire baisser les taux d'intérêt de façon permanente. Cet objectif est réalisable par la généralisation de l'utilisation des scorings dans les banques.

De même les compagnies d'assurances souhaitant améliorer leur rentabilité peuvent utiliser des scorings pour réduire le coût de leurs sinistres qui, en assurance-dommages, représente environ quatre-vingt pour cent du montant des primes versées.

C'est pourquoi nous avons choisi d'orienter notre exposé sur les comportements financiers :

- nous dégagerons une synthèse des travaux réalisés dans le domaine du crédit à la consommation depuis une vingtaine d'années, en tenant compte de travaux non publiés et de résultats communiqués oralement ;
- nous montrerons que les résultats obtenus peuvent s'expliquer par un modèle simple ;
- nous vérifierons la généralité de cette approche en l'appliquant au domaine de l'assurance-dommages.

II. Synthèse des analyses effectuées dans le domaine du crédit à la consommation

Le premier "credit scoring" utilisé en France l'a été dans le domaine du crédit à la consommation [5] [6]. Il permettait de prédire le comportement de remboursement d'un emprunteur à court terme, ayant acheté un bien de consommation durable par l'intermédiaire d'un réseau de distribution. Depuis, de nouvelles analyses ont permis d'améliorer progressivement l'efficacité du scoring dans le cadre d'un phénomène d'apprentissage et de prise en compte de plus en plus précise dans le système de décision à court et à long termes.

Nous allons passer en revue les étapes de ce processus et examiner les résultats obtenus au cours de chacune d'entre elles : l'utilisation du scoring a été banalisée dans les circuits de décision ; la sélection des emprunteurs, par un système sûr et stable, a conduit à la création du credit revolving et de la carte de crédit. Le scoring de sélection a évolué vers un scoring "de comportement" ou de fidélisation des bons clients [7].

A – Le risque du crédit à la consommation

1. L'analyse des premiers retards de remboursement (AFCM n°1)

Le principe de base du credit scoring est le suivant : il existe un comportement face au remboursement d'un crédit à long terme, qui peut être diagnostiqué à partir des caractéristiques initiales de l'emprunteur et de l'emprunt et qui se révèle dès les premières mensualités de remboursement.

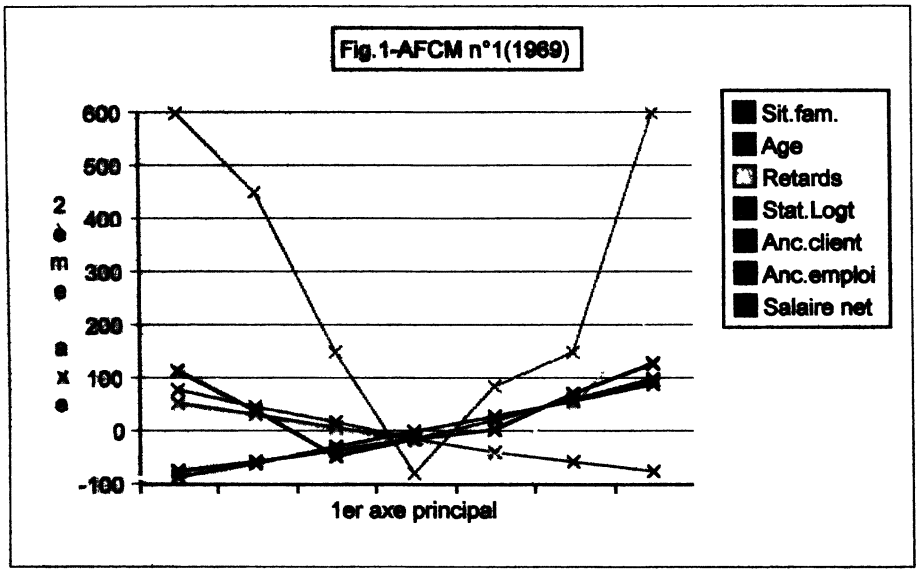
Sa justification essentielle est la suivante : le comportement d'un emprunteur est lié à la gestion de son budget. Or les dépenses à une date t dépendent des dépenses à la date $t-1$. Elles se répartissent proportionnellement au revenu selon des coefficients dus aux habitudes de consommation ou aux "styles de vie", eux-mêmes liés aux caractéristiques de l'emprunteur.

Il existe donc bien une justification économique du comportement de l'emprunteur, connue de longue date par les établissements spécialisés de crédit et matérialisée par le questionnaire que doit remplir tout emprunteur au moment de l'octroi de son crédit.

Compte tenu de la difficulté de suivre, sur une période suffisamment longue, les retards de remboursement qui, au niveau du contentieux, peuvent s'échelonner sur plusieurs années, seuls les retards des premiers mois ont fait, au début, l'objet de mesures attentives et figuraient dans les fichiers.

Dès 1969, des analyses factorielles des correspondances multiples appliquées à des échantillons représentatifs de la clientèle du crédit à la consommation ont mis en évidence un schéma géométrique simple pour les centres de gravité des populations de dossiers. Ils se répartissent selon deux axes : un axe 0-1

pour les incidents légers (0 = centre de gravité des emprunteurs à jour, 1 = centre de gravité des emprunteurs ayant une mensualité de retard) et un axe 0-2 pour les incidents graves (2 = centre de gravité des emprunteurs ayant deux mensualités de retard) (Fig.1).

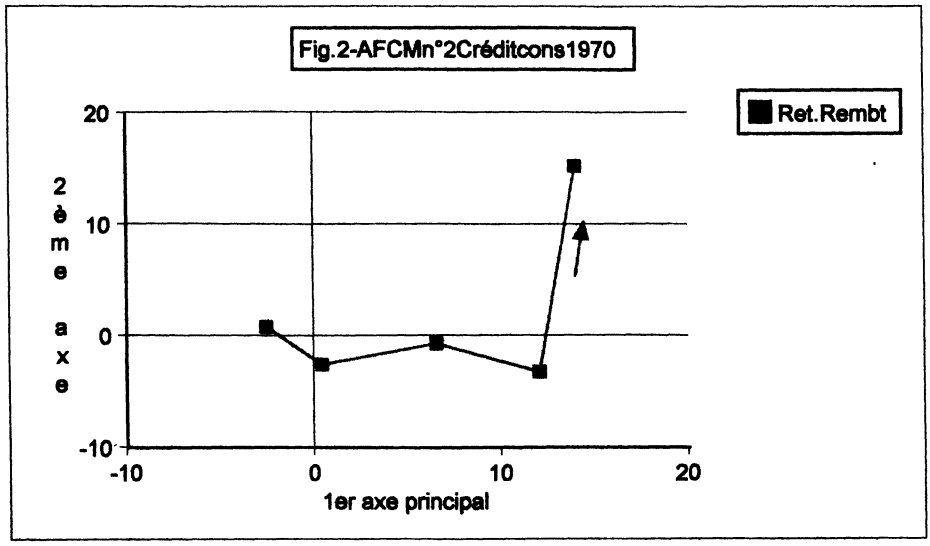


On découvre facilement que les variables les plus influentes sur le risque de crédit sont : l'âge, l'ancienneté dans l'emploi et l'ancienneté dans l'établissement de crédit, le statut d'occupation du logement et la situation de famille. Sur ces variables, on constate que les emprunteurs les plus stables sont aussi ceux qui sont le moins en retard.

Tout emprunteur cumulant plusieurs facteurs d'instabilité (par exemple jeunes, locataires et célibataires) a plus de difficultés à rembourser qu'un autre. Un corollaire de ce premier résultat est qu'une bonne pondération de ces variables dans la prise de décision, lors de l'octroi du crédit, doit conduire à une réduction significative des emprunteurs coûteux en gestion et en capital.

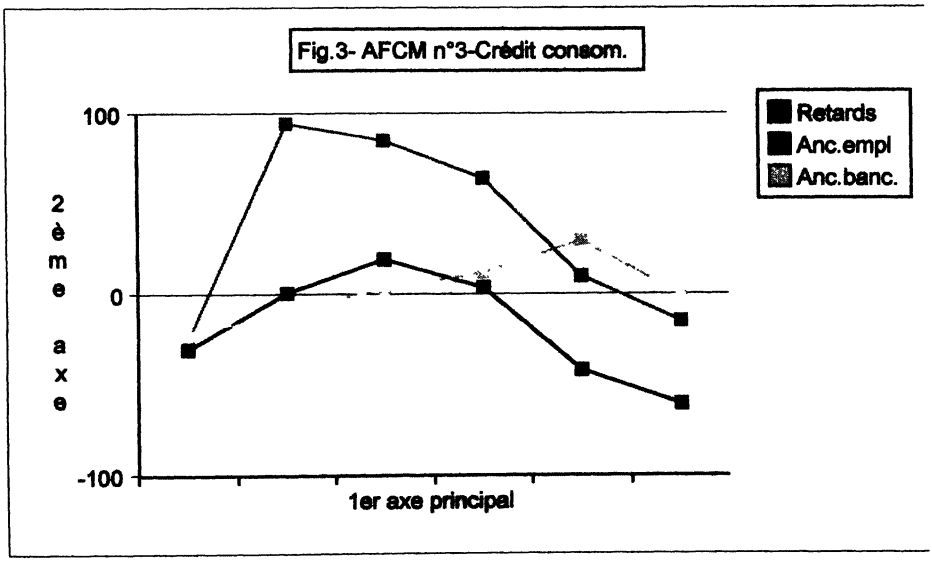
2. La prise en compte des retards graves et des dossiers refusés (AFCM n°2 et 3)

La prise en compte d'incidents plus graves (AFCM n°2)) ne modifie pas sensiblement le schéma précédent (Fig. 2). Le premier axe représente l'axe des incidents légers, ordonnés dans le sens de leur gravité, le deuxième axe, celui des incidents graves.



La prise en compte des dossiers refusés (AFCM n° 3) inverse les deux axes (Fig. 3). Le premier axe est celui des dossiers refusés (supposés représenter des incidents très graves, voire des pertes finales pour la majorité d'entre eux), le second est ordonné selon la gravité des incidents et le niveau de recouvrement.

L'ancienneté dans l'emploi et l'ancienneté bancaire sont des variables déterminantes pour le premier axe.



Le résultat de ces analyses est important : le phénomène "retard de remboursement" est, par nature, ordonné et cet ordre est lié au coût de traitement du recouvrement et au coût des pertes en capital.

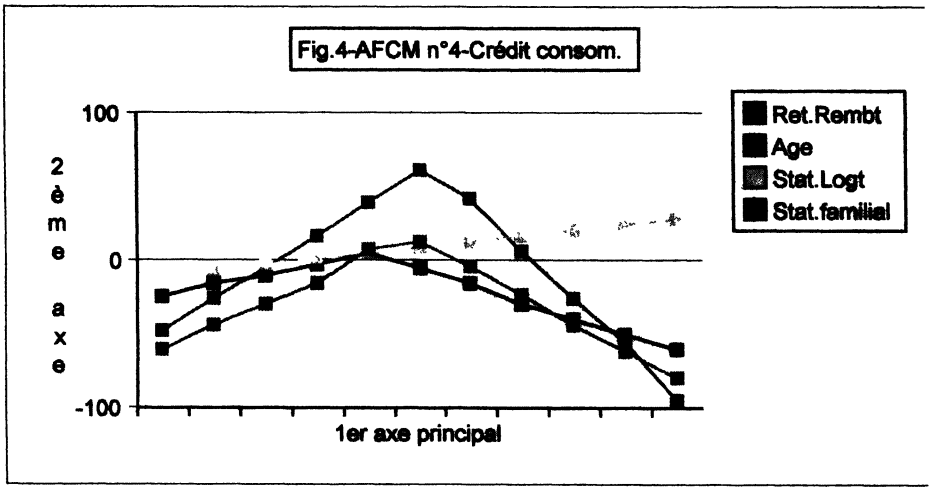
Il est fondamental de bien représenter l'ensemble des incidents de gravité croissante, pour être certain que la représentation dans le plan principal est bonne, et notamment que le premier axe oppose les incidents les plus graves (ou à défaut les dossiers refusés) aux dossiers ne comportant aucun retard de remboursement.

3. Les limites de la méthode dues aux dossiers refusés (AFCM n°4)

Un dossier refusé peut l'être plus ou moins aléatoirement, ou sur des critères qui ne peuvent pas être pris en compte dans l'analyse.
C'est ainsi que dans les analyses 1 et 2 les dossiers refusés n'avaient pas été retenus, car on avait supposé que les refus n'étaient pas entièrement justifiés, donc étaient peu représentatifs des très mauvais dossiers.

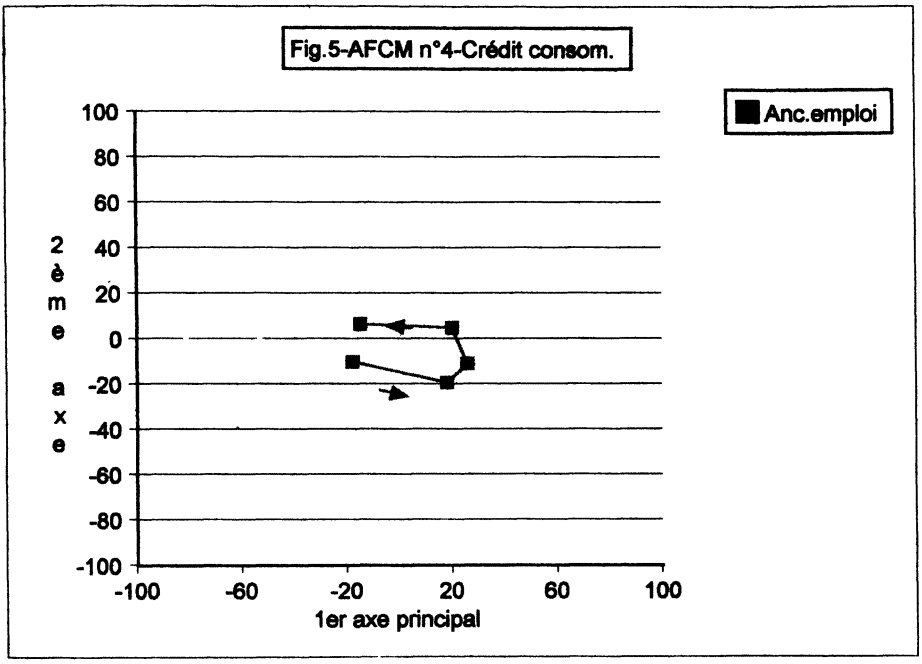
Si l'on utilise un scoring pour refuser les dossiers, on peut vérifier que les dossiers refusés sont de plus mauvaise qualité : les centres de gravité des dossiers sont bien ordonnés selon leur qualité et les dossiers refusés figurent en queue de classement.

L'AFCM n° 4 correspond à une situation dans laquelle les méthodes de recouvrement sont efficaces et les refus effectués avec un scoring de bonne qualité.



Le schéma (Fig. 4) est le même que celui de la fig. 3, mais on constate que les variables usuelles (âge, statut d'occupation du logement, statut familial) continuent à discriminer correctement les dossiers acceptés des dossiers refusés, tandis que des variables importantes, comme l'ancienneté dans l'emploi, ne sont plus aussi discriminantes, car elles sont déjà très utilisées dans la sélection.

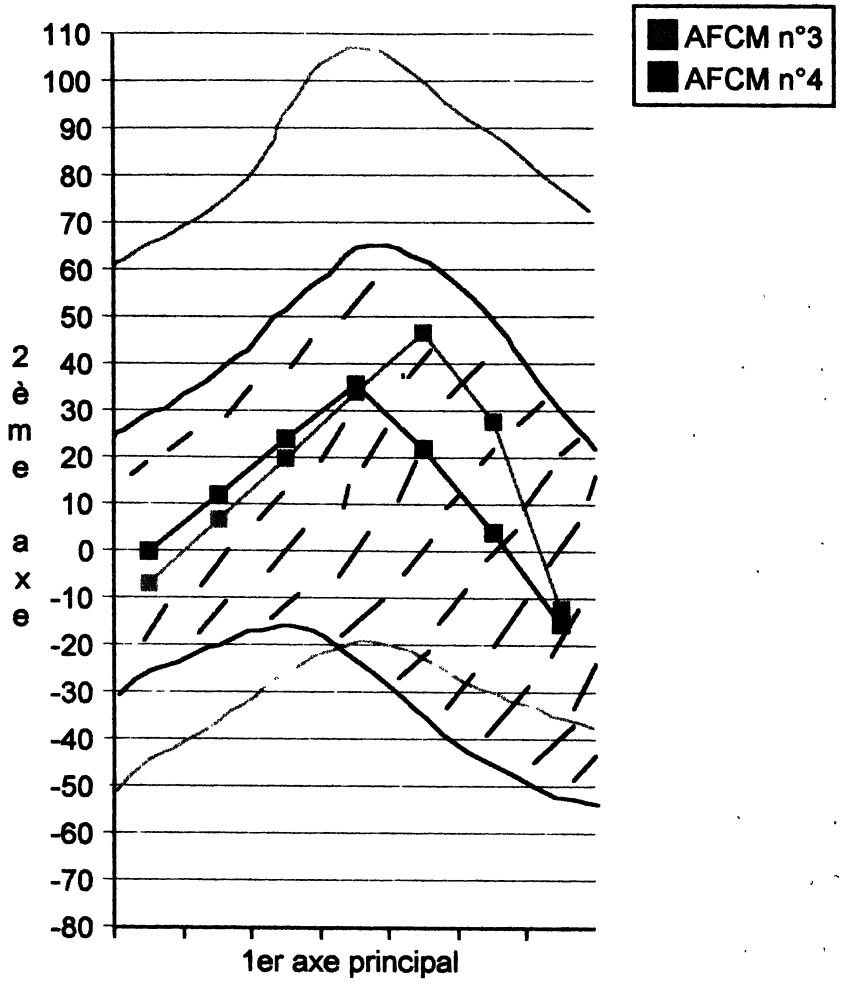
Par exemple, les dossiers d'emprunteurs de faible ancienneté dans l'emploi, qui ont été acceptés, sont bons, sur la base de leurs autres caractéristiques (Fig. 5).



On voit ainsi apparaître une limite dans l'efficacité de la méthode, dès lors qu'elle est décisionnelle et fondée sur l'exploitation de fichiers internes à l'établissement.

Toutefois, si on compare le nuage de points-individus de cette analyse à celui de l'AFCM n° 3 (Fig. 6), on constate qu'il reste une marge de discrimination pour les dossiers refusés dans l'analyse no 4, alors que cette marge se trouve plutôt chez les dossiers en retard dans l'analyse n° 3.

Fig.6-Nuages de points

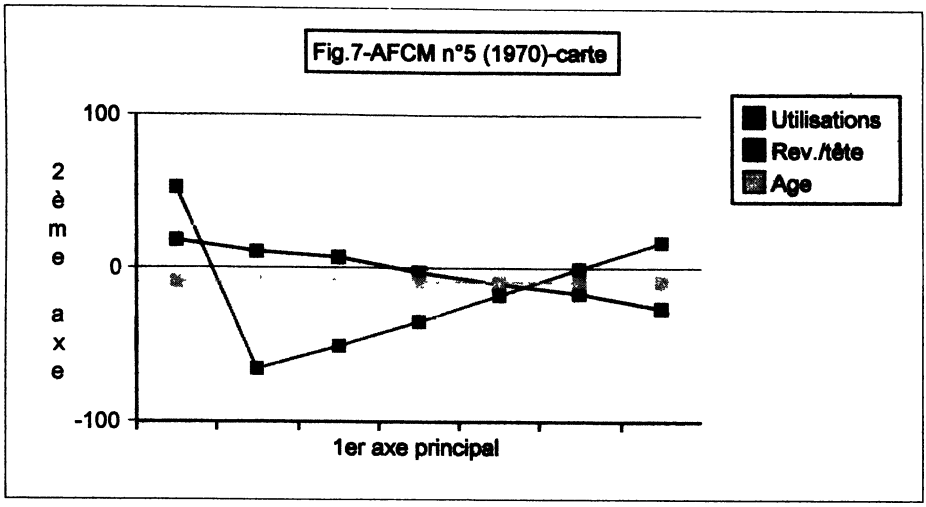


B – Les utilisations d'une carte de crédit

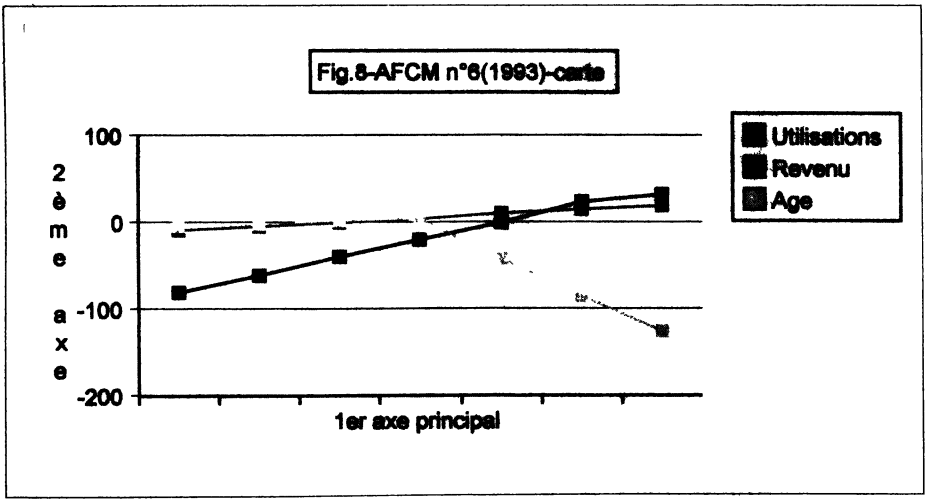
Les méthodes de scoring ont ouvert la porte à la diffusion de cartes de crédit, dont les porteurs sont sélectionnés, à l'ouverture de la carte, sur la base de leurs caractéristiques personnelles, et dont le découvert accordé peut également être fixé par le scoring.

Deux analyses effectuées indépendamment l'une de l'autre permettent de discriminer les forts utilisateurs de crédit de ceux qui ne l'utilisent pas ou peu.

Dans la première (Fig. 7), le schéma des centres de gravité des individus plus ou moins utilisateurs de crédit, est bien ordonné, comme dans les analyses de risque. Le premier axe représente plutôt le revenu.



Dans la seconde (Fig. 8), on observe surtout une relation entre l'âge et le revenu, analogue à un effet Guttman et dénotant l'importance de la contrainte de revenu dans le recours au crédit. Le premier axe est le revenu.



L'ensemble des analyses A et B peut être synthétisé dans les deux schémas suivants.

Fig.9

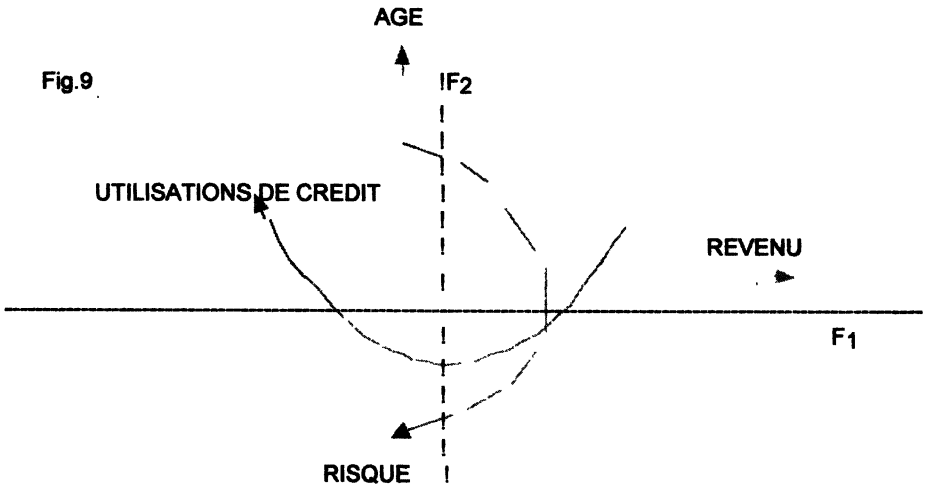
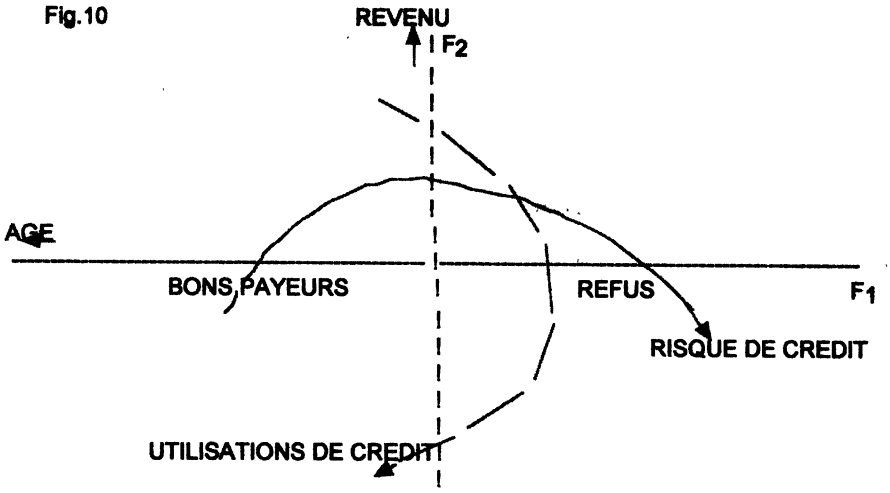


Fig.10



III. Un modèle simple explique les résultats obtenus

Comme nous l'avons observé sur les graphiques, un efet Guttman apparaît clairement, notamment sous la forme d'une liaison du second degré entre l'âge et le revenu.

Cet effet se produit notamment lorsque les caractères (classes d'âge, classes de revenu,...) sont naturellement ordonnés [1]. L'axe 1 oppose les classes de rang faible aux classes de rang fort, tandis que l'axe 2 oppose les classes moyennes aux classes extrêmes. M. VOLLE note par exemple une liaison du second degré entre l'âge et le salaire. L'analyse des correspondances multiples permet de bien traiter ce type de liaison dans lequel les deux variables, qui ne sont pas indépendantes, ont un coefficient de corrélation nul. Le tableau de contingence qui les croise est à diagonale chargée.

Plusieurs modèles sont possibles, qui conduisent à une relation du second degré entre les composantes principales : découpage en classes de deux variables aléatoires, suivant une loi normale à deux dimensions [8], changement d'état de l'état 0 à l'état 1 à une date t, variable aléatoire uniformément répartie sur la durée de l'étude [4], réponses ordonnées à un questionnaire [8].

C'est ce dernier modèle que nous retiendrons comme particulièrement adapté, parce que les profils des individus dans le domaine de la sinistralité et de l'endettement à court terme, sont parfaitement ordonnés et aussi parce qu'il s'applique bien à un ensemble de consommations ou à une accumulation de type patrimonial (exemple retenu parWELLER et ROMNEY [9] : la possession de biens durables par des habitants de la banlieue de Papeete à Tahiti), et à un phénomène comme le rejet ou l'exclusion, où l'on sait que les personnes exclues sont généralement celles qui cumulent plusieurs causes d'exclusion.

On considère un ensemble de questions $x_1, x_2, \cdots, x_m$ auxquelles on ne peut répondre que par oui ou par non. Cet ensemble de questions forme une échelle de Guttman parfaite si tout sujet qui répond oui à x_j répond de même à toute question d'indice supérieur à j. (On dit alors que le profil est parfait ; sinon, il est dit imparfait). Les questions sont ordonnées selon leur degré d'importance pour "expliquer" le phénomène étudié. Les profils extrêmes sur les réponses sont *a priori* les plus rares.

J.-P. BENZECRI [10] a donné une solution analytique lorsque le nombre de questions m tend vers l'infini. Les individus sont indicés par une valeur $x(0,1)$ et les variables réponses par une valeur $y(0,2)$, les oui sur $(0,1)$ et les non sur $(1,2)$.

Les deux premières composantes principales sont alors :

$$F_1(x) = \sqrt{3/2}\, x \qquad \lambda_1 = 1/2$$

$$F_2(x) = \sqrt{5/6}\,(3x^2 - 1) \qquad \lambda_2 = 1/6$$

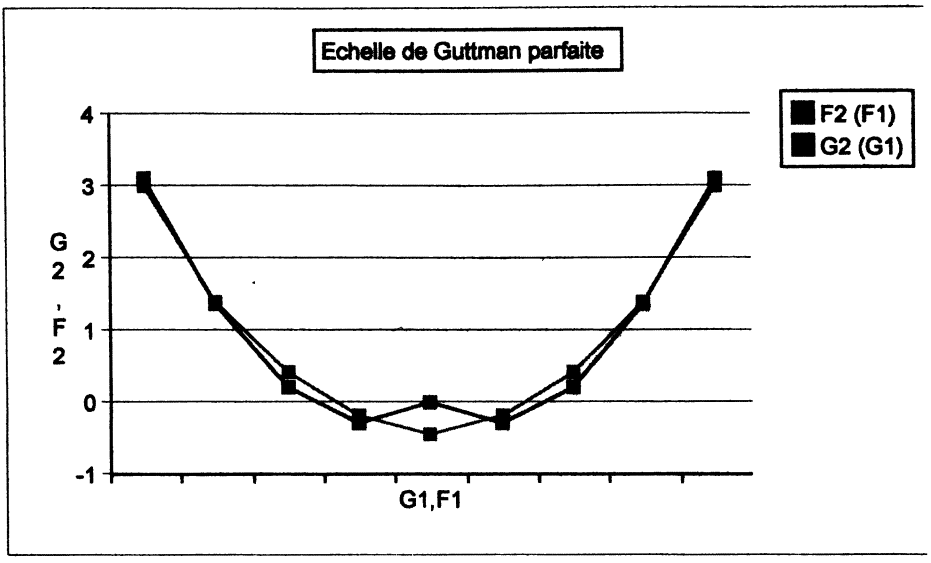
D'où $\quad F_2 = \sqrt{5/24}\,(2{F_1}^2 - 1)$

Les relations de transition conduisent à la relation :

$$G_2 = \sqrt{5}\left(2\frac{{G_1}^2}{3} - \frac{|G_1|}{\sqrt{3}}\right)$$

La représentation graphique des fonctions $F_2(F_1)$ et $G_2(G_1)$ est la suivante :

Fig. 11 : Echelle de Guttman parfaite.
Cas continu, probabilité uniforme sur les profils.



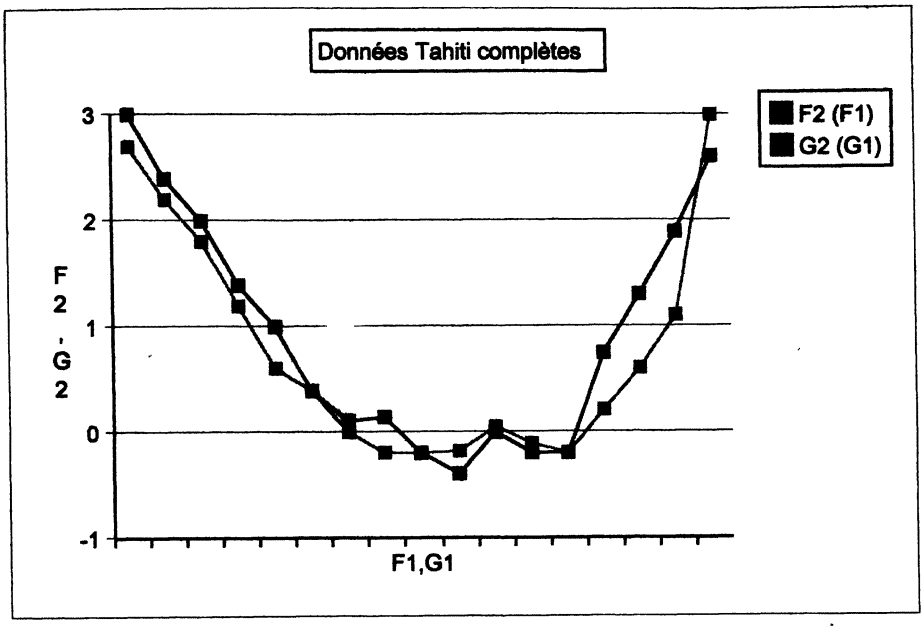
Source : M. TENENHAUS - Méthodes statistiques en gestion - DUNOD (1994).

Un exemple concret a été analysé par J.C. Weller et A.K. Romney avec l'analyse des correspondances multiples, et résumé par M. Tenenhaus [8] dans le cas où le nombre de questions est égal à 7 (et le nombre de profils égal à 8). On constate que :

– le nombre de questions au-delà de 7 ne joue plus beaucoup dans le cas de profils équiprobables : les courbes $F_2(F_1)$ et de $G_2(G_1)$ ainsi que les ratios entre valeurs propres sont quasiment identiques pour $n = 7$ et $n \infty$. Toutefois la parabole représentative des courbes a tendance à s'évaser lorsque le nombre de questions croît, ce qui signifie que le premier axe devient relativement plus important. Cette remarque est cohérente avec le fait que ce que l'on étudie est une variable ordonnée dont les valeurs extrêmes sont les plus éloignées sur le premier axe, tandis que le second axe ordonne plutôt les valeurs intermédiaires.

– les courbes $F_2(F_1)$ et $G_2(G_1)$ restent parallèles quels que soient le nombre des questions posées, la répartition de la population entre les profils et le nombre de profils imparfaits. Elles sont approximativement des paraboles et les centres de gravité des profils restent à l'intérieur du profil convexe (Fig. 12) :

Fig. 12 – Effet Guttman - Données Tahiti complètes



Source : M. TENENHAUS- Méthodes statistiques en gestion - DUNOD (1994).

On peut en déduire que, lorsque la conjoncture varie (par exemple l'environnement socio-économique), ces schémas sont relativement stables. En effet, si la répartition des effectifs entre profils et le pourcentage de profils imparfaits peuvent varier, l'ordre des questions qui conduit au schéma parfait de la parabole reste stable par hypothèse. La variation de conjoncture peut conduire à des données moins structurées et à des relations plus floues, mais le modèle reste stable s'il continue à s'appuyer sur des questions ordonnées.

IV. Une application du modèle dans le secteur de l'assurance

Le modèle précédent peut se résumer en une phrase : une population multidimensionnelle peut être décrite simplement dans un espace à deux dimensions (75 % de l'information) par les réponses à des "questionnaires", si les questions sont ordonnées et correspondent à des phénomènes extrêmes sur chacun des deux axes principaux : risque, sécurité sur le premier axe, revenu, endettement, épargne sur le second.

Nous n'avons considéré que des ménages ; il reste à analyser dans quels autres cas on peut considérer qu'une population multidimensionnelle, par exemple de petites et moyennes entreprises, se comporte de la même façon.

On peut ensuite prévoir les comportements financiers de ces "individus" à l'aide d'une méthode de scoring pour des comportements tels que la sinistralité (crédit, assurance-dommages) et l'accumulation (épargne ou endettement en vue d'une accumulation patrimoniale).

Cette prévision est d'autant plus stable dans le temps que les profils des répondants sont équiprobables et parfaits.

Elle est indépendante de la stratégie adoptée par l'établissement et de sa part de marché.

Nous allons appliquer ce modèle à l'assurance pour prévoir la sinistralité en assurance-dommages (auto et habitation) et le budget annuel du ménage en assurances. Nous vérifierons que la sinistralité en assurance-dommages est fortement corrélée à la sinistralité du ménage dans le secteur du crédit à la consommation.

A – Les données de base

Aucune compagnie d'assurances n'ayant une part de marché permettant de dire que tous les profils sont représentés dans sa clientèle, nous nous sommes fondés sur une enquête annuelle effectuée par la SOFRES à partir d'un échantillon représentatif de la population française, soit environ quinze mille interviews.

La majeure partie des questions concernent l'assurance automobile, l'assurance-habitation et l'assurance-vie. Quelques questions concernent plus directement les caractéristiques du ménage interviewé. Ce sont ces questions que nous avons analysées préférentiellement, par exemple :
- l'âge du chef de famille,
- l'ancienneté dans la résidence principale,
- l'ancienneté dans la Compagnie assurant l'automobile,
- l'ancienneté dans la Compagnie d'assurance-vie,
- le coefficient de réduction-majoration, ou bonus-malus, représentant le comportement passé du conducteur pour les sinistres dont il est responsable,
- le statut du logement,
- le niveau de revenu,
- le niveau d'instruction,
- le statut familial,
- la taille de l'agglomération,
- la catégorie socio-professionnelle.

Nous cherchons les deux groupes de variables correspondant aux deux axes de l'AFCM.

Le premier oppose la mobilité à la stabilité et nous vérifions que des variables plus traditionnellement utilisées dans le domaine de l'assurance sont liées à la mobilité et aux transports, comme :
- le nombre annuel de kilomètres parcourus,
- un indicateur de la vitesse du véhicule, donné par sa marque et son type, donc sa puissance,

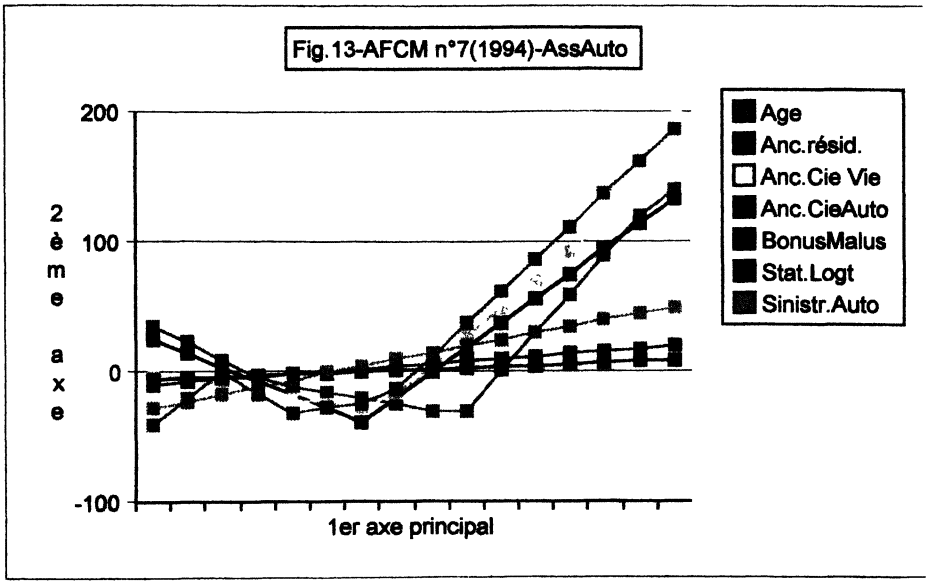
– l'ancienneté du véhicule, donnée soit par l'année de première mise en circulation, soit par l'année d'achat.

Ce sont également des variables figurant implicitement dans le coefficient budgétaire du ménage relatif aux transports, et celui-ci est généralement indicateur d'un comportement de dépense spécifique.

B – Une analyse factorielle des correspondances multiples en assurance-auto

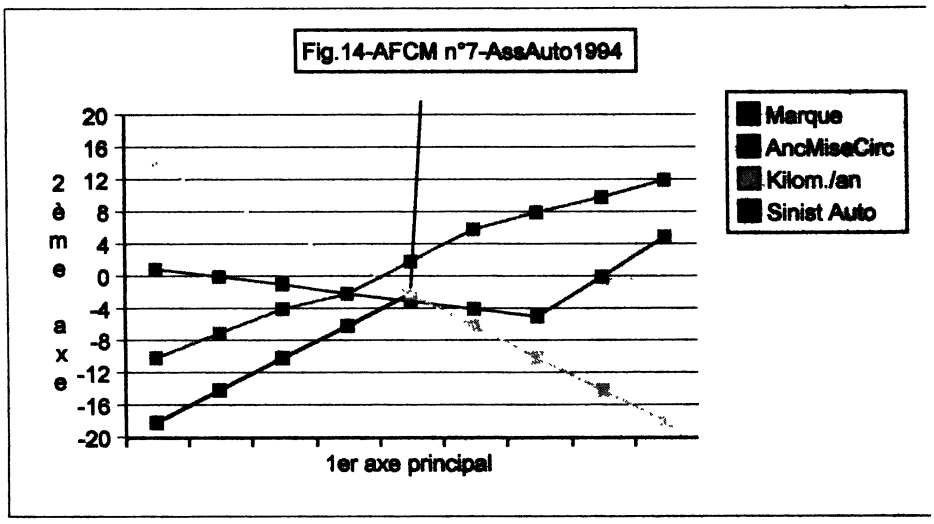
Une analyse (AFCM n° 7) de l'ensemble des variables permet de préciser la constitution des axes :

– le phénomène étudié est ordonné : c'est le nombre de sinistres auto durant les trois dernières années, le conducteur étant responsable ou non. Le problème des dossiers refusés ne se pose pas dans l'enquête puisque l'assurance est obligatoire. Le nombre de conducteurs ayant eu deux sinistres et plus est trop faible (5,1 %) pour que l'on puisse distinguer des classes d'incidents plus graves. On constate que les centres de gravité des trois populations (zéro sinistre, un sinistre, deux sinistres et plus) sont alignés pratiquement le long du premier axe (Fig. 13), ce qui correspond à notre modèle, bien que nous ne retrouvions pas la forme parabolique, qui nécessiterait un échantillon plus important d'interviewés et des classes intermédiaires dans la classe "Deux sinistres et plus", voire un échantillon de dossiers refusés par les compagnies.

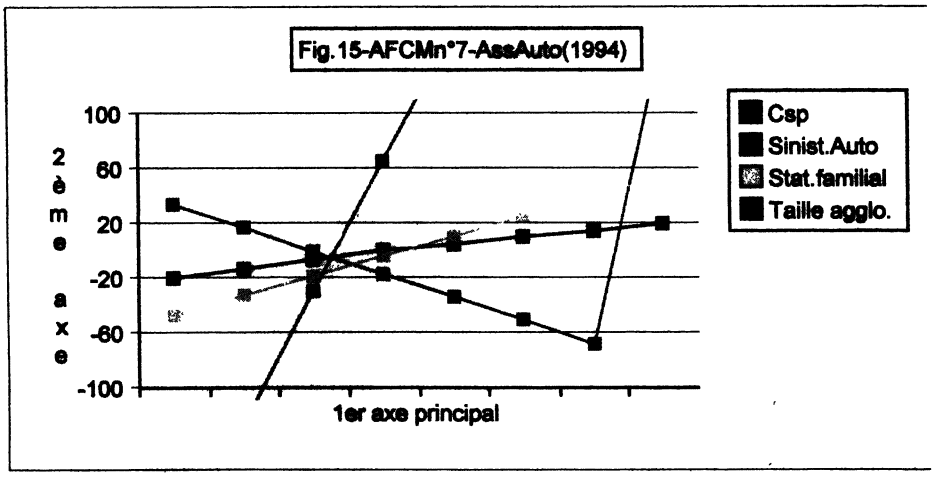


– le premier axe est lié à des variables telles que l'âge, la stabilité et le comportement passé pour les sinistres dont le conducteur est responsable (Fig. 13).

– et également à des variables liées au comportement dans le domaine des transports automobiles telles que le nombre de kilomètres parcourus, la puissance de la voiture ou l'ancienneté du véhicule (Fig. 14) :



– le deuxième axe est lié essentiellement au revenu, à la taille de l'agglomération, et en partie à la catégorie socio-professionnelle et à la situation de famille (Fig. 15) :



On retrouve donc bien toutes les caractéristiques du modèle prévu, avec au moins un effet Guttman fort entre l'âge et le revenu.

C – Le calcul du scoring pour prévoir la sinistralité auto en assurance-auto

Pour simplifier, nous nous sommes limités à prévoir le nombre d'individus ayant eu deux sinistres et plus au cours des trois dernières années.

Après avoir utilisé une régression logistique pas à pas pour éliminer les variables insuffisamment significatives, nous en avons retenu neuf sur les vingt et une testées, soit dans l'ordre :

- le coefficient de réduction-majoration (CRMAUT),
- la profession de la mère de famille (PROFMA),
- la taille de l'agglomération de résidence (AGGLOM),
- le statut d'occupation du logement (STAOCC),
- le nombre de kilomètres parcourus par an (KILOME),
- l'âge du chef de famille (AGCHEF),
- la date de première mise en circulation du véhicule (MISECI),
- la date d'achat du véhicule (AACHAT),
- la marque et le type du véhicule (MARQUE).

On y ajoute la possession par le ménage d'un répondeur téléphonique (REPOND), liée au revenu et à la taille de l'agglomération, mais également significative pour la sinistralité.

Par hypothèse, notre modèle théorique correspond assez précisément aux fondements du modèle linéaire généralisé (procédure CATMOD de SAS), dans lesquels les variables explicatives sont des variables qualitatives (réponses au questionnaire) croisées entre elles et où la répartition des effectifs selon la variable à expliquer (la sinistralité) suit une loi binomiale (11).

La méthode du maximum de vraisemblance permet d'estimer les paramètres β et ε d'un modèle de calcul du score qui s'écrit :

$$\text{Log } p_i / (1 - p_i) = X_i \beta + \varepsilon$$

p_i est la probabilité d'avoir deux sinistres et plus pour la population i,
X_i est le vecteur des caractéristiques de la population i,
β est un vecteur-colonne de paramètres ou scores,
ε est une constante.

On constate (Tableau n° 1) que toutes les variables retenues sont explicatives. Un indicateur du poids de chacune dans le calcul des scores est donné par la valeur du chideux dans le test du maximum de vraisemblance. Des variables non retenues traditionnellement en assurance, telles que le statut d'occupation du logement, la profession de l'épouse, la possession d'un répondeur, la date de mise en circulation ou la date d'achat du véhicule, peuvent "expliquer" jusqu'à 40 % de la variance.

TABLEAU N° 1

Analyse de la variance selon la méthode du maximum de vraisemblance

Source	Degrés de liberté	Statistique de Wald	Niveau de signification
Constante	1	351,41	0
STAOCC	1	12,92	0,0003
KILOME	1	11,86	0,0006
AGGLOM	2	23,49	0
MARQUE	4	30,76	0
AACHAT	2	24,44	0
AGCHEF	3	8,12	0,0436
CRMAUT	1	26,87	0
REPOND	1	9,89	0,0017
MISECI	2	9,66	0,0008
PROFMA	2	7,17	0,0277
Résidu	3705	2006,20	1

Enfin l'échantillon est scoré avec les poids obtenus pour toutes les modalités des variables et le tableau croisé score*nombre de sinistres sur trois ans permet d'apprécier le caractère prévisionnel du score (Tableaux n^{os} 2 et 3) (en n'oubliant pas que l'échantillon est le même que celui sur lequel ont été calculés les poids, et qu'il ne s'agit pas d'une prévision en grandeur réelle).

TABLEAU N° 2

SCORE * SINISTRALITÉ AUTO (en nombre de ménages)

SCORE	2 Sin. et +	Total classe	% ménages
1	81	462	17,50 %
2	75	650	11,50 %
3	26	326	8,00 %
4	101	1396	7,20 %
5	166	3662	4,50 %
6	65	3227	2,00 %
7	4	439	0,90 %
TOTAL	518	10162	5,10 %

TABLEAU N° 3

Résumé statistique pour le tableau croisé SCORE*SINAUT

STATISTIQUES DE COCHRAN-MANTEL-HAENSZEL

Hypothèse	Degrés de Liberté	Valeur	Probabilité
Corr.non nul	1	282,038	0
Ecarts col.	7	303,884	0
Assoc. gén.	7	303,884	0

Taille effective de l'échantillon : 10 162.

La qualité de la discrimination pourrait être augmentée en ajoutant des variables pertinentes, absentes du sondage et en réduisant le nombre de non-répondants, de façon à accroître la taille de l'échantillon utile.

D – Le budget annuel Assurances d'un ménage

Comme nous l'avons fait pour le crédit à la consommation, nous pouvons chercher à prévoir la dépense annuelle d'assurances par ménage ou "Budget assurances" (comprenant toutes les assurances du ménage, y compris l'assurance-vie).

Cette dépense est évidemment très liée au revenu. Si on suit la même démarche que pour la sinistralité auto, on peut obtenir un scoring Budget permettant de prévoir le budget potentiel de chaque ménage (Tableau n° 4).

TABLEAU N° 4

Analyse de la variance selon la méthode du maximum de vraisemblance

Source	Degrés de liberté	Statistique de Wald	Niveau de signification
Constante	1	309,35	0
REVENU	2	151,14	0
NBAUTO	2	12,27	0,0022
MISECI	2	67,89	0
PROFCH	2	36,70	0
AG1CON	3	8,91	0,0305
ANNEMM	2	30,75	0
CRMAUT	1	15,48	0,0001
AACHAT	2	7,87	0,0195
PIECES	3	24,53	0
ACTICH	2	18,31	0,0001
Résidu	3896	4190,68	0,0005

TABLEAU N° 5

SCORE * BUDGET ASSURANCES (en nombre de ménages)

SCORE	+10000F/an	Total classe	% ménages
1	150	303	49,50 %
2	196	471	41,60 %
3	81	224	36,20 %
4	527	1821	28,90 %
5	600	2535	23,70 %
6	492	2531	19,40 %
7	352	2439	14,40 %
8	51	610	8,40 %
9	21	421	5,00 %
TOTAL	2470	11355	21,80 %

TABLEAU N° 6

Résumé statistique pour le tableau croisé SCORE*BUDGET

STATISTIQUES DE COCHRAN-MANTEL-HAENSZEL

Hypothèse	Degrés de Liberté	Valeur	Probabilité
Corr.non nul	1	539,629	0
Ecarts col.	8	552,812	0
Assoc. gén.	8	552,812	0

Taille effective de l'échantillon : 11 355.

On vérifie que ce scoring, fondé essentiellement sur le revenu, est très largement indépendant de la sinistralité auto (Le coefficient d'association générale entre le score budget et la sinistralité auto est de 40 pour 8 degrés de liberté).

Les phénomènes sous-jacents aux analyses précédentes (variables prédictives du comportement de même nature, traitement analogue de variables qualitatives et non linéaires, indépendance de la sinistralité et des variables liées au revenu) sont tout-à-fait analogues à ceux observés dans le domaine du crédit à la consommation.

Le même modèle théorique simple s'applique, car la description sous-jacente des ménages est indépendante des comportements étudiés.

On en déduit que la population sinistrable en assurance devrait recouper largement celle du crédit.

Une étude, effectuée sur 400 000 dossiers [12] par une société américaine spécialisée en scoring au niveau mondial, Fair Isaac, confirme l'étroite corrélation existant entre les deux populations sinistrées.

E – La sinistralité en multirisques habitation

Le phénomène que l'on cherche à prévoir est la sinistralité sur trois ans (zéro sinistre ou un sinistre et plus) dans le domaine de la multirisques habitation. Dans l'échantillon, plus de 17 % des ménages ont eu au moins un sinistre en trois ans, ce qui est trois fois supérieur à la moyenne des ménages ayant eu au moins deux sinistres auto sur la même période. Il s'agit donc d'un phénomène moins rare et donc plus difficile à prévoir.

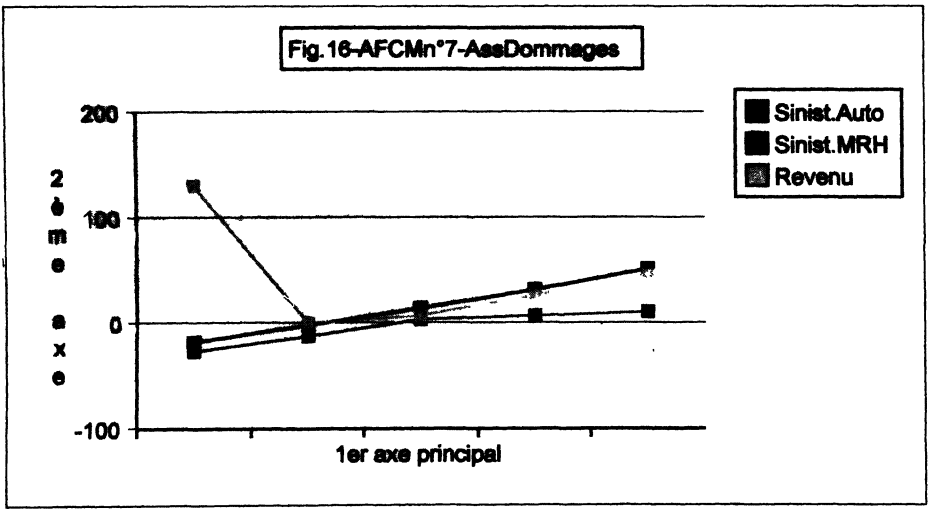
Parmi les variables déjà étudiées en sinistralité auto, seules sept semblent performantes (Tableau n° 7), parmi lesquelles quatre représentent essentiellement le premier axe (statut d'occupation du logement, coefficient de réduction-majoration en assurance auto, possession d'un répondeur ou d'une platine laser) et trois, le second axe (nombre de pièces, revenu du ménage, taille de l'agglomération de résidence).

TABLEAU N° 7

Analyse de la variance selon la méthode du maximum de vraisemblance

Source	Degrés de liberté	Statistique de Wald	Niveau de signification
Constante	1	557,49	0
REVENU	2	13,40	0,0012
PIECES	3	7,10	0,0688
STAOCC	2	20,55	0
CRMAUT	1	8,52	0,0035
PLALAS	1	13,06	0,0003
REPOND	1	18,20	0
AGGLOM	2	23,78	0
Résidu	530	571,24	0,1046

La nature des variables explicatives laisse à penser qu'il pourrait y avoir une corrélation entre la sinistralité auto et la sinistralité habitation, ce qui est vérifiable malgré la faible taille de l'échantillon renseigné (Fig. 16) :



On constate également sur ce graphique qu'un revenu faible pèse sur la sinistralité, mais que le second axe oppose essentiellement les bas revenus aux revenus élevés, critère déterminant pour le budget annuel d'assurances (cf. paragraphe D).

Le calcul du scoring pour prévoir la sinistralité en multirisques habitation, effectué selon une méthode analogue à celle utilisée en sinistralité auto, conduit,

comme on pouvait s'y attendre, à une prévision de la sinistralité sensiblement moins précise que celle obtenue pour la sinistralité auto (Tableaux n^{os} 8 et 9).

TABLEAU N° 8

SCORE * SINISTRALITÉ HABITATION (en nombre de ménages)

SCORE	1 Sin. et +	Total classe	% ménages
1	192	692	27,80 %
2	230	1031	22,30 %
3	413	1972	21,00 %
4	395	2193	18,00 %
5	480	3069	15,60 %
6	279	1941	14,40 %
7	145	1226	11,80 %
TOTAL	2134	12124	17,60 %

TABLEAU N° 9

Résumé statistique pour le tableau croisé SCORE*SINMRH

STATISTIQUES DE COCHRAN-MANTEL-HAENSZEL

Hypothèse	Degré de Liberté	Valeur	Probabilité
Corr.non nul	1	114,263	0
Ecarts col.	8	130,832	0
Assoc. gén.	8	130,832	0

Mais, par rapport au modèle traditionnel, les variables ajoutées (possession d'un répondeur et d'une platine laser, coefficient de réduction-majoration en assurance-auto) "expliquent" environ 40 % de la variance, comme pour la sinistralité auto.

Conclusions

A partir de travaux empiriques effectués dans le domaine du scoring appliqué au crédit à la consommation, nous avons mis en évidence un modèle simple pouvant expliquer les résultats obtenus dans ce domaine et dans celui, plus général, des comportements financiers et des styles de vie. Ce modèle, à deux effets Guttman orthogonaux, permet de positionner les extrémités des deux axes principaux descriptifs de la population, en s'appuyant sur des ensembles de questions "ordonnées", le premier axe étant lié à la sinistralité (score d'acquisition) et le second, au potentiel (score de comportement ou de fidélisation). Un test de ce modèle a été effectué pour prévoir la sinistralité et le potentiel d'un ménage dans le domaine de l'assurance.

Ce début de modélisation permet de concevoir une nouvelle méthode de scoring, applicable notamment dans l'ensemble du secteur Banques-Assurances. Cette nouvelle méthode comprend :

- une enquête approfondie, sur un échantillon représentatif de la population française, comportant des questions appartenant aux deux groupes de variables les plus explicatives des deux axes principaux (notamment de leurs extrémités) ;
- la mise en place de scorings relatifs à différents phénomènes "ordonnés" tels que la sinistralité, le potentiel, la sensibilité aux tarifs, aux actions commerciales et à la communication.

Ces deux phases sont indépendantes de la prise en charge, par les établissements concernés, des lois prévisionnelles ainsi établies, dans leur stratégie et leur planification. Les méthodes de commercialisation, les tarifications, l'offre de produits et les outils de gestion sont modifiés ensuite en fonction de ces lois, dans un processus d'apprentissage.

Une telle méthode permet de contourner différents obstacles à une bonne efficacité, notamment le biais des clientèles répertoriées dans les fichiers des établissements (biais commerciaux par suite d'une faible pénétration, biais décisionnels sur les risques), l'instabilité éventuelle des scorings en fonction de l'environnement, le coût de maintenance d'une équipe scientifique de haut niveau dans chaque établissement.

Elle peut conduire à de nouvelles conditions d'exploitation pour les compagnies d'assurance, comme cela a été le cas dans le passé pour les établissements spécialisés de crédit.

Bibliographie

[1] VOLLE M. (1993) *Analyse des données*, 3e édition, Economica.

[2] ESCOFIER B. et PAGES J.-P. (1990) *Analyses factorielles simples et multiples : objectifs, méthodes et interprétation*, Dunod. Paris.

[3] MASSON M. (1974) "Analyse non linéaire de données", *C. R. Acad. Sc. Paris*, T. 278 (11 mars 1974).

[4] SAPORTA G. (1990) *Probabilités, analyse des données et statistique*, Technip, Paris.

[5] AUBERGER M. (1974) "Le credit scoring", *in Marketing bancaire, marketing financier*, Dalloz.

[6] AUBERGER M. (1977) "Les méthodes modernes d'analyse du risque de crédit : présentation générale", *Vie et Sciences économiques.*

[7] BONIFACE J.-F. (1996) "A bon score, bon crédit", *Banque* n° 567, févr. 1996, p. 28.

[8] TENENHAUS M. (1994) *Méthodes statistiques en gestion*, Dunod.

[9] WELLER J.C. et ROMNEY A.K. (1990) *Metric Scaling : Correspondence Analysis*, Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-075. Sage, Newbury Park, CA.

[10] BENZECRI J.-P. *et alii* (1973) *L'Analyse des données.* Tome I *La Taxinomie.* Tome II *L'Analyse des correspondances*, Dunod. Paris.

[11] TENENHAUS M., LE ROUX Y., GUIMAT C., GONZALES P.L. (1993) "Modèle linéaire généralisé et analyse des correspondances", *R. Statistique appliquée.* 41(2), p. 59-86.

[12] FAIR, ISAAC *et alii* (Eté 1995) "Revue Viewpoints,", *Numéro spécial sur l'Assurance,*Vol. 1, n° 2.