No 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2005 : Défis méthodologiques reliés aux besoins futurs d'information

2005

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2005 de Statistique Canada
Défis méthodologiques pour les besoins futurs d'information

IMPUTATION DE DISTRIBUTIONS DANS LES DONNÉES FISCALES ADMINISTRATIVES

Rong Huang, Dominique Ladiray[1]

ABSTRACT

Les données fiscales administratives sont largement utilisées à Statistique Canada pour diminuer la taille des échantillons des enquêtes et alléger ainsi la charge de réponse des entreprises. Ces données, en principe exhaustives, présentent les défauts habituels - non réponse partielle, valeurs atypiques etc. - et doivent être analysées et redressées.
Ainsi les ventilations des totaux, souvent facultatives, doivent être imputées. La procédure d'imputation par le ratio actuelle est basée sur un découpage a priori de la population des entreprises répondantes en groupes définis par le code activité et la taille de l'entreprise. Nous proposons une méthodologie alternative laissant aux données le soin de définir des classes homogènes, déterminées par une classification ascendante hiérarchique sur les valeurs des détails observées. Le problème est ensuite d'affecter une entreprise non répondante à l'une de ces classes. Plusieurs procédures d'affectation sont comparées, sur données brutes ou discrétisées : analyses discriminantes paramétrique et non-paramétrique, modèles log-linéaires etc. Les règles d'affectation obtenues sont évaluées et comparées par simulation puis sont validées sur les données de l'année précédente.

MOTS CLÉS : Données administratives, imputation par le ratio, classification ascendante hiérarchique, analyse discriminante non-paramétrique.

1. INTRODUCTION

Statistique Canada s'est résolument engagé depuis plusieurs années dans l'utilisation de données fiscales administratives pour diminuer la taille des échantillons des enquêtes et alléger ainsi la charge de réponse des entreprises. Les données fiscales, transmises par l'Agence du Revenu du Canada, sont en principe exhaustives. Mais elles présentent les défauts habituels - non réponse partielle, valeurs atypiques etc. - et doivent être analysées et redressées avant d'être utilisables par les divisions clientes. En particulier, ces déclarations portent en général plus d'attention aux totaux qu'aux ventilations, souvent facultatives. Lorsque qu'un poste de la déclaration ne comporte que le total (appelé « générique ») de la recette ou de la dépense, une procédure d'imputation est alors mise en œuvre pour estimer la répartition en sous-postes (appelés « détails »).
La procédure actuelle est basée sur un découpage a priori de la population des entreprises répondantes en groupes définis par le code activité (code SCIAN) et par la taille de l'entreprise. La distribution marginale par détails des entreprises répondantes à l'intérieur d'une classe, est alors utilisée pour les entreprises de la même classe qui ne fournissent pas les détails. Cette estimation par le ratio repose sur une hypothèse forte : les entreprises d'une classe donnée sont supposées avoir des comportements très voisins pour que la même répartition moyenne s'applique à toutes ce qui peut aboutir à des imputations étranges.
Nous proposons une méthodologie alternative laissant aux données le soin de définir les classes qui sont déterminées par une classification ascendante hiérarchique sur les valeurs des détails observées. Le problème est ensuite d'affecter une entreprise non répondante à l'une de ces classes homogènes par construction. Plusieurs procédures d'affectation, basées sur des variables explicatives disponibles dans la déclaration fiscale, sont comparées, sur données brutes ou discrétisées : analyses discriminantes paramétrique et non-paramétrique, modèles log-linéaires etc.
Les règles d'affectation ainsi obtenues sont évaluées et comparées par simulation sur données réelles puis validées sur les données de l'année précédente.

[1] Rong Huang, Statistique Canada, Division des Méthodes d'Enquêtes Entreprises, Édifice R.H. Coats, 120 Parkdale Avenue, Ottawa, Ontario, Canada K1A0T6 (courriel : rong.huang@statcan.ca); Dominique Ladiray , INSEE, Département des comptes nationaux, 15 boulevard Gabriel Péri, 92240 Malakoff, France (courriel : dominique.ladiray@insee.fr)

2. MÉTHODE ACTUELLE : PRÉSENTATION ET PROBLÈMES

La déclaration fiscale remplie par les entreprises (GIFI) contient environ 685 variables, certaines représentant des totaux (génériques) et d'autres des détails. Le tableau montre un exemple de la structure de la déclaration. Le *bloc* est composé d'un générique (code 2700 en gris) et 6 *détails* (codes 2701 à 2706). Le déclarant est libre de reporter dans sa déclaration le niveau de détail qu'il souhaite. S'il n'est pas sûr du code détail, il peut choisir d'inclure le montant concerné dans la partie générique. En conséquence, un déclarant peut choisir de remplir un code générique, un autre les codes détails du même générique et un troisième à la fois le générique et des détails.

Tableau 1 : Générique et détails

BT2680	**miscellaneous taxes payable block total amount**
	SHORT TERM DEBT FLDS 2700 TO 2706
2700	short term loan and debt amount
2701	current Canadian bank loan amount
2702	current security sold short liability amount
2703	current security sold under repurchase agreement amount
2704	gold and silver certificate amount
2705	items in transit and check amount
2706	current lien note payable amount

L'idée de la méthode d'imputation utilisée est assez simple : la valeur générique est répartie selon la distribution en détails observée sur l'ensemble des déclarations des entreprises ne reportant que des détails[2].
Le tableau représentant la répartition des montants déclarés par les entreprises d'un même sous-groupe est une table de contingence. Notons cette table (n_{ij}), $i = 1 \cdots r$, $j = 1 \cdots c$. n_{ij} représente la valeur reportée par l'entreprise *i* pour le détail *j*.

On a $n = \sum_i \sum_j n_{ij}$ et la valeur totale reportée par l'entreprise i est $n_{i.} = \sum_{j=1}^{j=c} n_{ij}$. Dans le sous-groupe, le montant total reporté pour le détail *j* est $n_{.j} = \sum_{i=1}^{i=r} n_{ij}$.

L'idée de base de la procédure d'allocation utilisée actuellement est d'utiliser la distribution marginale observée sur les entreprises ne reportant que des détails, soit la distribution : $\left\{\frac{n_{.1}}{n}, \frac{n_{.2}}{n}, \ldots, \frac{n_{.c}}{n}\right\} = \{f_{.1}, f_{.2}, \cdots, f_{.c}\}$. Il est ainsi implicitement supposé que les distributions des diverses entreprises sont sensiblement les mêmes, ce qui fait de la distribution marginale un bon estimateur des distributions individuelles.
Mesurer « l'homogénéité d'un sous-groupe » (c'est à dire la similarité des distributions lignes) peut se faire en utilisant la distance du χ^2 (voir par exemple Saporta, 1990) : $d^2 = n \sum_i \sum_j \frac{\left(f_{ij} - f_{i.} f_{.j}\right)^2}{f_{i.} f_{.j}}$

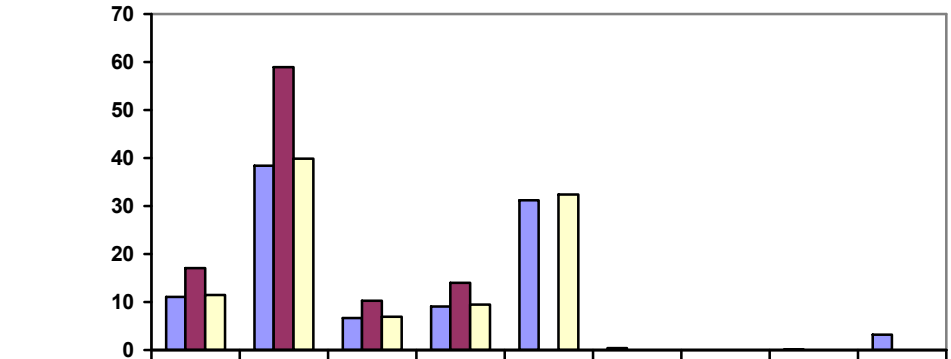
d^2 varie entre 0 et $n \inf(r-1, c-1)$. 0 correspond à l'indépendance totale entre les lignes et les colonnes de la table et donc à l'égalité des distributions en détails. La valeur maximale $n \inf(r-1, c-1)$ traduit une relation fonctionnelle parfaite entre les lignes et les colonnes. Il est malheureusement difficile de comparer les valeurs de la statistique d^2 pour deux tables différentes puisque *n, r* et *c* sont en général différents. Un grand nombre

[2] L'algorithme de calcul des ratios GDA utilisé est en pratique un peu plus complexe (Rondeau, 2003). En particulier, les ratios ne sont calculés que pour des sous-groupes d'au moins 25 entreprises ne reportant que des détails ; de plus, dans un sous-groupe, seuls les détails ayant été reportés par au moins 10% des entreprises sont pris en compte (« détails fréquents »).

d'indicateurs variant entre 0 et 1 ont été proposés dans la littérature (voir Agresti, 1986 ; Goodman & Kruskal, 1979 ; SAS Institute, 2001) : coefficient P de Pearson, coefficient V de Cramer, τ_b de Kendall etc.

L'homogénéité des sous-groupes obtenus avec les différentes classifications habituellement utilisées (12 classifications ont été étudiées) a été mesurée avec les 3 indicateurs. Pour une classification du code activité en 2 positions, les 3 indicateurs sont en moyenne égaux à 0.8, ce qui traduit des distributions par détails très différentes.
La figure 1 montre, pour un bloc donné (bloc 9040), les distributions marginales réelle et celles calculées à partir des « détails fréquents » (reportés par au moins 10% des entreprises du groupe) et des « détails importants » (détails les plus fréquents dont la valeur cumulée est supérieure à 90% de la somme totale des détails).

Figure 1 : Comparaison des distributions marginales calculées avec tous les détails, les détails fréquents et les détails importants pour le bloc 9130.



	1	2	3	4	5	6	7	8	9
Total	11,07	38,33	6,65	9,07	31,15	0,37	0,03	0,1	3,22
Fréquents	16,99	58,87	10,21	13,93	0	0	0	0	0
Importants	11,49	39,82	6,91	9,42	32,36	0	0	0	0

Pour le bloc 9130, la distribution des « détails fréquents » ne prend pas en compte le détail 5 qui représente environ 30% du total tout en sur-estimant le détail 2. L'utilisation des « détails importants » conduit à éliminer les détails 6 à 9 qui, au total, représentent en valeur moins de 4%.

Dans ces conditions, comment peut on améliorer la qualité du redressement lorsque les détails ne sont pas fournis par l'entreprise ? Plusieurs méthodes sont a priori envisageables.

1. Tout d'abord, on peut penser à utiliser une imputation par donneur en utilisant donc la distribution d'une entreprise n'ayant reportée que des détails pour imputer les détails d'une entreprise n'ayant reportée qu'une valeur générique. On se heurte à deux difficultés :
   - La première tient au fait que les sous-groupes peuvent être d'effectifs assez faibles. On court donc le risque de recourir souvent au même donneur.
   - La seconde tient au comportement même de réponse des entreprises. Comme le montre le **Error! Reference source not found.**, les distributions peuvent être radicalement différentes à l'intérieur d'un même bloc. On court donc le risque, en sélectionnant un « mauvais » donneur, d'imputer une distribution plus lointaine de la distribution réelle que la distribution moyenne elle même.
2. Le plan comptable normalisé (COA, « Chart of Accounts ») mis au point par Statistique Canada comporte moins de détails que la déclaration financière et ce sont ces détails qui sont finalement importants pour les divisions clientes. En réduisant le nombre de détails, on augmente en principe l'homogénéité des distributions et donc la pertinence de la distribution moyenne dans un même sous-groupe.
3. Une autre possibilité, celle développée dans la suite de ce travail, vise à constituer automatiquement des sous-groupes dans lesquels les distributions en détails seront par construction similaires. Ces sous-groupes d'entreprises ne reportant que des détails sont définis par une classification automatique faite sur les distributions observées. Une analyse discriminante permet ensuite de repérer automatiquement la classe à laquelle appartient une entreprise ne reportant qu'une valeur générique. La répartition de cette valeur générique en détails se fait alors à partir de la distribution estimée sur la classe, par exemple en utilisant la distribution marginale.

3. UNE MÉTHODE ALTERNATIVE

Le principe de la méthode proposée est assez simple et se résume en 3 étapes :

- Définir des classes de distributions homogènes, à partir des entreprises ne reportant que des détails dans leur déclaration.
- Déterminer la classe à laquelle appartient une entreprise ne reportant qu'un générique pour le bloc considéré. Cette « inférence » repose sur les variables disponibles dans la réponse de l'entreprise c'est à dire les génériques (somme des détails) ou les détails des autres blocs et certaines caractéristiques de l'entreprise (activité, taille).
- Estimer la répartition par détail en utilisant par exemple la méthode actuelle, c'est à dire la distribution marginale estimée à partir des entreprises de la classe.

Si, dans ses grandes lignes, la méthode paraît simple, il reste pour la mettre en application à fixer de nombreux paramètres dans le but de déterminer la meilleure stratégie.

3.1 Recherche de la meilleure stratégie

Les données

La présence dans les données de valeurs atypiques pourrait affecter les résultats de l'analyse, notamment parce que nous faisons parfois appel à des procédures non-robustes basées sur la moyenne (calcul des distributions marginales par exemple). Dans la recherche de la meilleure méthode à employer, par exemple au moment de la prédiction de la classe, nous avons donc a priori travaillé d'une part sur des données discrétisées à partir des quantiles de la distribution en 5, 10, 15, 20, 25 et 30 groupes et d'autre part sur les données brutes.

Classification

La classification a été faite, dans tous les cas, sur les données brutes de chaque bloc à partir du tableau de contingence des montants (entreprises en lignes, détails en colonne). Deux types de classification ont été envisagés :

1. Une classification ascendante hiérarchique (CAH) utilisant une métrique du χ^2 (adaptée à ce type de données) et la stratégie de Ward[3]. Cette méthode pose plusieurs problèmes :
   - Le nombre d'entreprises peut être assez grand dans un bloc donné (plusieurs dizaines de milliers) et une CAH peut alors difficilement être envisagée. Une première classification non hiérarchique a été réalisée[4] pour déterminer 500 classes sur les centres desquels on a fait une CAH.
   - Le nombre de classes doit être précisé a priori. Dans ce cas, nous avons choisi un nombre de classes lié au nombre de variables. Par exemple, dans le cas du bloc 8040, on a trois variables COA et on définira donc une classification en 3 ou 4 classes.
2. Une classification « a priori », justifiée par le type de comportement de réponse des entreprises qui ont tendance à ne reporter qu'un faible nombre de détails. On a donc défini des classes, nommées « AttracteurXX », en affectant à la classe *i* les entreprises affectant plus de XX% du montant total au détail *i*. Pour le bloc 8040, on a ainsi défini les variables Attracteur80, Attracteur90, Attracteur95. Dans chaque cas, une classe supplémentaire regroupe toutes les entreprises pour lesquelles aucun détail renseigné n'égale les XX% requis.

Sélection des variables explicatives

La déclaration financière d'une entreprise contient environ 300 variables qui peuvent a priori servir de variables explicatives pour déterminer la classe à laquelle une entreprise appartient. Une première sélection de variables « potentiellement explicatives » a été faite sur cet ensemble de variables en utilisant une analyse discriminante paramétrique pas à pas (stepwise) et une régression logistique[5] pas à pas. Les deux listes de variables candidates ont alors été fusionnées pour donner une liste plus réduite de variables explicatives.

---

[3] PROC CLUSTER de SAS après transformation des données pour que l'application de la métrique euclidienne sur les données transformées soit équivalente à l'application d'une métrique du χ^2 sur les données brutes ;

[4] PROC FASTCLUS de SAS

[5] En toute rigueur une régression logistique n'est ici pas justifiée puisque le numéro de la classe n'est pas une variable numérique mais nous n'avons ici utilisé cette méthode que comme « méthode de tri » des variables.

Sélection des modèles
Même avec une liste de variables explicatives réduite à quelques dizaines, le nombre potentiel de modèles est énorme. Là encore, nous avons sélectionné automatiquement des modèles potentiels (avec moins de 30 variables explicatives) en utilisant des méthodes de régression logistique et d'analyse discriminante paramétrique.

Evaluation des modèles
Dans cette dernière étape, chaque modèle a été estimé par plusieurs méthodes :

- Analyse discriminante paramétrique
- Analyse discriminante non-paramétrique utilisant la méthode des k plus proches voisins (avec k=15 ou k=20)
- Modèle de régression linéaire adapté à la nature qualitative de la variable à prévoir[6].

Enfin, chaque modèle a été évalué par son taux d'erreur de classement, et des indicateurs de concordance entre la répartition en classes prévue et la répartition réelle.

3.2 Premiers résultats

Une simulation à grande échelle sur les données du bloc 8040 a permis de tirer des enseignements précieux pour la mise en production de la méthode :

- On obtient des taux d'erreurs de classement tout à fait raisonnables, de l'ordre de 15%.
- Les modèles de régression linéaire sur données qualitatives sont souvent longs à estimer, essentiellement parce que le nombre de modalités des variables explicatives est assez important.
- L'analyse discriminante non-paramétrique réalise en général d'excellentes performances et ce avec des modèles comportant peu de variables explicatives.
- L'utilisation de variables explicatives discrétisées se révèle efficace, en particulier si le nombre de groupes est assez élevé. Dans ce cas, on minimise l'effet de valeurs atypiques, tout en conservant un caractère « continu » aux valeurs.

3.3 Evaluation finale de la méthode

Le but final de l'étude est d'imputer les distributions de détails. Il nous reste donc à évaluer la qualité de l'imputation réalisée à partir de la méthode alternative présentée ci-dessus et de la comparer à la qualité de la méthode actuelle.

Le fichier test et la procédure d'évaluation
Pour ce faire, nous utiliserons un « fichier test » comportant un ensemble d'entreprises pour lesquelles la distribution réelle en détails est connue. L'utilisation de tels fichiers tests est fréquente dans d'autres domaines : tests d'algorithmes statistiques (voir NIST, 1998), tests de méthodes de prévisions en séries temporelles (voir Makridakis et Hibon, 2000) etc.
La procédure de test repose ici sur les données des années 2001 et 2002 :

- Les données de l'année 2001 sont utilisées pour calculer les distributions qui seront utilisées pour imputer les données de l'année 2002.
- Pour chaque bloc étudié, on sélectionne les entreprises qui en 2002 n'ont reporté que des détails. Ces valeurs sont agrégées pour définir un générique « fictif » dont la distribution en détails sera alors imputée à l'aide des ratios calculés sur le fichier 2001.
- Les distributions réelles et imputées sont alors comparées : la meilleure méthode sera celle qui fera le moins d'erreur.

La figure 2 résume d'évaluation mis en œuvre.

Les indicateurs statistiques de comparaison
Des indicateurs statistiques doivent être définis pour permettre une comparaison entre des distributions de détails vraies et imputées. Ces indicateurs sont construits au niveau de l'entreprise (niveau micro) et au niveau du bloc (niveau macro).

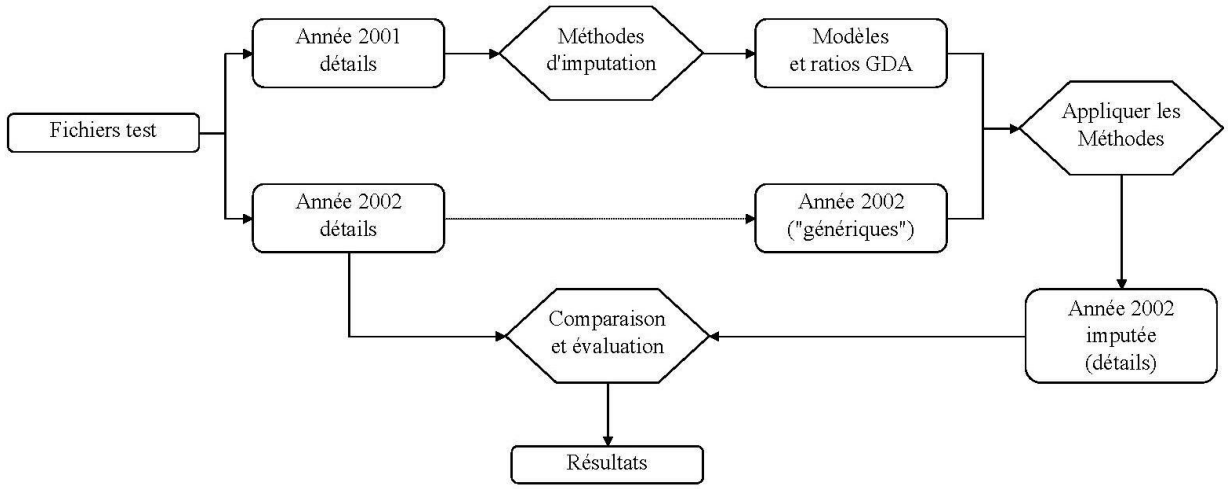
[6] Ce modèle a été estimé avec la PROC CATMOD de SAS.

Au niveau micro :
Pour chaque entreprise, le coefficient de contingence de Pearson a été calculé pour mesurer la distance entre la vraie distribution en détails et la distribution estimée par la méthode d'imputation.
Une autre mesure, le « Micro_pseudo_CV » est définie pour chaque entreprise par :

$$Micro\_pseudo\_CV_j = \sqrt{\sum_i (x_{ij} - \hat{x}_{ij})^2} \Big/ \sum_i x_{ij}, \quad j = 1, \cdots n$$

où x_{ij} est le i ème détail réellement reporté par l'entreprise j et $\hat{x}_{ij}$ la prévision du i ème détail pour l'entreprise j.

Figure 2 : Diagramme du processus d'évaluation des méthodes d'imputation



Au niveau macro
Un certain nombre de mesures peuvent sont définies au niveau du bloc :
La somme des carrés des écarts : $SSE = \sum_i (t_i - \hat{t}_i)^2$ où t_i (respectivement $\hat{t}_i$) désigne le total réel reporté (respectivement le total estimé) pour le détail i.

La somme des carrés des écarts sur les pourcentages de distribution : $SSEP = \sum_i (\frac{t_i}{\hat{t}_i} - 1)^2$

Le « Macro_pseudo_CV » : $Macro\_pseudo\_CV = \sqrt{\sum_i (t_i - \hat{t}_i)^2} \Big/ \sum_i t_i$

3.4 Application de la méthode au bloc 9760

Pour l'année 2001 (resp. 2002), 21782 (resp. 22041) entreprises ont reporté des détails seulement. Ce bloc correspond à 5 détails COA.
Les 21782 entreprises ont été classées en utilisant des classifications automatiques à 2 (Cluster2), 5 (Cluster5) et 15 (Cluster15) classes, et différents « attracteurs » : Attractor60, Attractor70, Attractor90. Une dernière classification ad hoc (Clus_User) a été définie à partir des détails importants.

Homogénéité
Comme le montrent les coefficients de Pearson présentés dans le tableau 2, les groupes obtenus par les différentes classifications utilisées sont plus homogènes que les groupes définis par la méthode actuelle d'imputation.

Précision

Pour ce bloc, nous considérons les 10 meilleurs modèles pour les variables Cluster2, Cluster5, Clus_User et Attractor80, et les 20 meilleurs modèles pour les autres variables de classification (Attractor60, Attractor70 et Cluster15). Ces modèles ont été choisis en fonction le leur taux d'erreur de classement.

Tableau 2 : Homogénéité des groupes obtenus par les différentes classifications (bloc 9760, coefficient de Pearson)

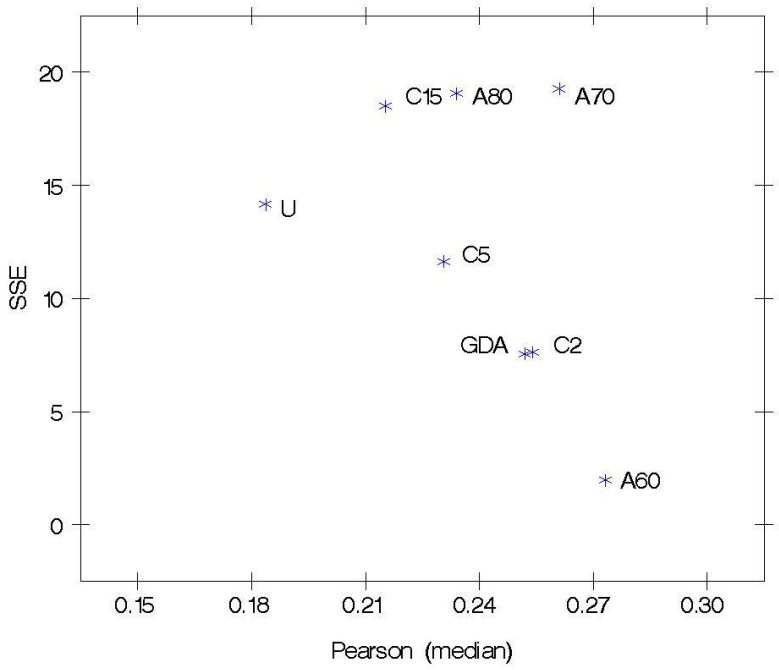
Cluster	Attractor60
1	0.602
2	0.601
3	0.492
4	0.630
5	0.588
6	0.682

Clus_User	
Maximum	0.711
Q3	0.404
Median	0.381
Q1	0.353
Minimum	0.320

Sous-groupes GDA	
Maximum	0.817
Q3	0.811
Median	0.739
Q1	0.707
Minimum	0.583

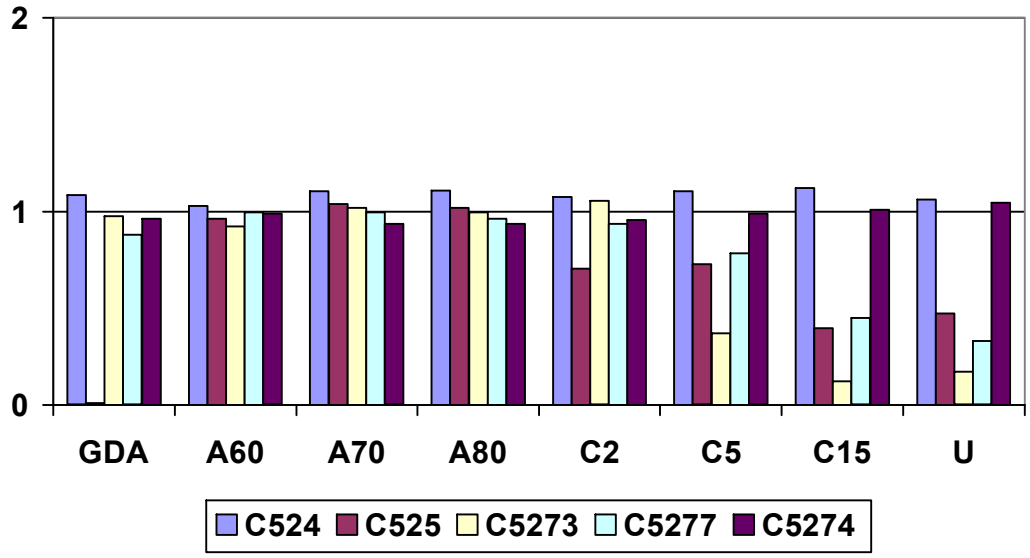
La figure 3 représente les meilleurs modèles des 8 types de variables de classe en fonction de critères de précision précédemment définis : SSE et coefficients de Pearson. La procédure GDA actuelle est moins bonne que les modèles basés sur la variable Attractor60 pour les critères relatifs au niveau macro, et par les modèles basés sur la variable Clus_User pour les critères relatifs au niveau micro. Elle réalise des performances comparables à celles des modèles basés sur la variable Cluster2.

Figure 3 : Représentation des meilleurs modèles pour chaque variable de classe en fonction du SSE et de la médiane du coefficient de Pearson (Bloc 9760)



C'est essentiellement la performance des modèles sur la restitution des pourcentages des différents détails qui fera la différence. La figure 4 présente les performances des meilleurs modèles (au sens du SSE) de chaque variable de classe.
La procédure GDA actuelle sous estime fortement le détail C525 et surestime a contrario le détail C524. Les modèles bâtis sur les variables de classe Cluster2, Cluster5, Cluster15 et User_Clus ont des estimations très déséquilibrées des parts de chaque détail.
C'est le modèle basé sur la variable Attractor60 qui paraît le meilleur. C'est celui là que nous sélectionnons, bien qu'il soit assez peu parcimonieux puisqu'il fait intervenir 16 variables explicatives.

Figure 4 : Respect des poids de chaque détail (SSEP) pour les meilleurs modèles de chaque variable de classe.



En résumé, pour le bloc 9760, la meilleure stratégie d'imputation des détails est basée sur une classification a priori de type Attractor60, la classe étant ensuite prédite par une analyse discriminante sur un modèle à 15 variables explicatives (L9370 L9470 L9541 L9662 L9663 L9760 L9799 L9802 L9804 L9811 L9818 L9819 L9820 L9835 et ratio) discrétisées en 30 groupes, et en utilisant une méthode d'estimation non paramétrique.

4. CONCLUSION

La procédure actuelle d'imputation des distributions de détails, basée sur l'utilisation de distributions marginales de sous-groupes définis a priori par le code activité et la taille de l'entreprise, peut être améliorée en créant des classes de distributions homogènes. Cela peut être fait à partir d'une classification automatique (CAH) ou par des procédures ad hoc (de type « Attracteur »). L'utilisation de l'analyse discriminante non paramétrique permet alors d'affecter avec suffisamment de précision une entreprise à une classe en fonction de ses caractéristiques. L'imputation de la distribution des détails par la distribution marginale observée sur la classe donne alors de meilleurs résultats.
La procédure doit cependant être adaptée à chaque bloc puisque la méthode de classification et le modèle optimal peuvent varier d'un cas à l'autre.

RÉFÉRENCES

Agresti, A. (1996), *An Introduction to Categorical Data Analysis*, New York: John Wiley & Sons, Inc.

Goodman, L. et Kruskal, W. (1979), *Measures of Association for Cross-classifications*, Springer-Verlag, New-York.

Helmer, P., Lafontaine-Sorgo, D., Lalande, D. et Thibault, D. (2001), Generic to Detail Allocation: GIFI 1999 Documentation, Statistique Canada.

Rondeau, C. (2003), Specifications of the GDA ratios calculation, Division des données fiscales, Statistique Canada.

Rossiter, P. (1998), Modelling Detailed Business Operating Expenses From ABS Economic Collections, Methodology Advisory Committee Paper, Australian Bureau of Statistics.

Saporta, G. (1990), *Probabilités, Analyse des données et Statistique*, Technip.

SAS Institute (2001), *SAS/STAT User's Guide: PROC FREQ*, North Carolina.

Makridakis, S. et Hibon, M. (2000), "The M3-Competition: results, conclusions and implications", *International Journal of Forecasting*, 16, 451-476.

NIST (1998), "StRD: Statistical Reference Datasets for Assessing the Numerical Accuracy of Statistical Softwares", http://www.nist.gov/itl/div898/strd.