Imputation de distributions dans les données fiscales administratives

Rong HUANG (*), Dominique LADIRAY (**)

(*) Statistique Canada, Division des Méthodes d'Enquêtes Entreprises
(**) INSEE, Département des Comptes Nationaux

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, qui sont transmises régulièrement 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, les déclarations fiscales 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 mis 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 n'est malheureusement pas le cas et, de plus, l'utilisation d'une répartition moyenne peut aboutir à des imputations étranges : telle entreprise de restauration rapide s'est ainsi vue attribuer à tort des dépenses en boissons alcoolisées.

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 fiscales, 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. Enfin, l'efficacité globale des différentes procédures d'affectation associées elles-mêmes à des méthodes d'imputation différentes (ratio ou donneur) sont validées sur les données de l'année précédente.

1. La méthode actuelle d'allocation des valeurs génériques aux détails : 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'allocation des génériques aux détails ("Generic to Detail Allocation" par la suite *GDA*) est une répartition automatique du montant reporté en générique aux détails correspondants. Toute valeur reportée dans le champ 2700 sera automatiquement réallouée aux détails 2701 à 2706 selon un algorithme prédéfini (voir [1], [2]) résumé ci-après.

1.1. L'algorithme de base

L'idée est assez simple : la valeur générique est répartie selon la distribution en détails observée sur les déclarations des entreprises ne reportant que des détails. Pour un bloc (avec m détails) et un sous-groupe (Activité x Taille) donnés, supposons que nous avons n entreprises ne reportant que des détails et soient y_{ij} la valeur reportée par l'entreprise i ($i = 1, \cdots, n$) pour le détail j ($j = 1, \cdots, m$), et $x_i = \sum_j y_{ij}$ la somme des détails (le « générique »).

Les coefficients de répartition du générique aux détails ont été calculés, pour les données des années 2000 et 2001, par la formule suivante :

$$\hat{\beta}_j = \frac{\sum_i y_{ij}}{\sum_i x_i}.$$

Si on suppose que les valeurs reportées sont non négatives, les ratios $\hat{\beta}_j$ représentent la distribution marginale des détails dans le sous-groupe considéré, distribution calculée sur les entreprises ne reportant que des détails.

Le modèle statistique associé, qui suppose que toutes les entreprises d'un même sous-groupe sont indépendantes, est le suivant :

$$y_{ij} = \beta_j x_i + \varepsilon_{ij} \quad i = 1, \cdots, n \quad j = 1, \cdots, m \quad \varepsilon_{ij} \sim (0, \sigma_j{}^2 x_i) \tag{1}$$

Si les y_{ij} ne sont pas négatifs, alors les $\hat{\beta}_j$ sont non positifs ou nuls et leur somme est égale à 1.

Quelques compléments sur l'algorithme réellement utilisé :

L'algorithme de calcul des ratios GDA utilisé est en pratique un peu plus complexe (voir [3]).

- Les sous-groupes sont définis par bloc (18 blocs), code activité (25 codes NAICS) et taille (3 tailles). Le code taille utilisé est défini à partir d'un « revenu total ajusté ». Une entreprise est qualifiée de « grande » si son revenu est supérieur à 25 millions de dollars canadiens ; elle « petite » si son revenu est inférieur à 5 millions de dollars canadiens et « moyenne » dans les autres cas.
- Les ratios ne sont calculés que pour des sous-groupes d'au moins 25 entreprises ne reportant que des détails. Dans le cas contraire, les sous-groupes sont agrégés (par taille puis par code activité) avant de calculer les coefficients.
- Dans un sous-groupe, seuls les détails ayant été reportés par au moins 10% des entreprises sont pris en compte.

1.2. Mesurer l'homogénéité d'un sous-groupe

Cette estimation par le ratio repose sur une hypothèse forte : les entreprises d'un sous-groupe donné sont supposées avoir des comportements très voisins pour que la même répartition moyenne s'applique à toutes.

Le montre les déclarations de quelques entreprises, parmi celles ne reportant que des détails, du sous-groupe 8040-21-1 (bloc 8040[1], code activité 21, taille 1). Si les détails 1 et 10 apparaissent « dominants », les répartitions par détails semblent assez différentes.

Tableau 2 : Déclarations de quelques entreprises du sous-groupe 8040-21-1 (Bloc 8040, code NAICS 21, taille 1).

(entreprises ne reportant que des détails).

	Det1	Det2	Det3	Det4	Det5	Det6	Det7	Det8	Det9	Det10	Det11	Det12	Det13
1	1057238	0	0	0	0	0	0	0	0	0	0	0	0
2	59931	0	0	0	0	0	0	0	0	0	0	0	0
3	195333	0	0	0	0	0	0	0	0	0	0	0	0
4	46667	0	0	0	0	0	0	0	0	0	0	0	0
5	167059	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	6511	0	0	0
7	0	0	0	0	0	0	26991	0	0	0	0	0	0
8	226879	0	0	0	2327	0	0	0	0	1240920	0	0	0
9	53244	0	0	0	0	0	0	0	0	47745	0	0	0
10	1361890	0	0	0	0	0	0	0	0	173008	0	0	0
11	394683	0	0	0	0	0	0	0	0	70172	0	0	0
12	2614395	0	0	0	0	0	0	0	0	0	0	0	0
13	25522	0	0	0	0	0	0	0	0	0	0	0	0
14	7025	0	0	0	0	0	0	0	0	0	0	0	0
15	1130454	0	0	0	0	0	0	0	0	46491	0	0	24609
16	120878	0	0	0	0	0	0	0	0	0	0	0	0
17	152903	0	0	0	0	0	0	0	0	0	0	0	0
18	143782	0	0	0	0	0	0	0	0	0	0	0	0
19	219158	0	0	0	0	0	0	0	0	76357	0	0	0
20	229099	0	0	8587	0	0	0	0	0	9241	0	0	0
21	59945	0	0	5057	0	0	0	0	0	0	0	0	0

Le tableau représentant la répartition des montants déclarés par les entreprises d'un même sous-groupe est une table de contingence dont l'homogénéité peut être mesurée à l'aide d'indicateurs basés sur la distance du χ^2. Notons cette table $\left(n_{ij}\right),\ i=1\cdots r,\ j=1\cdots c$.

[1] La composition du bloc 8040 est précisée au Tableau 7.

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 :

$$\left\{ \frac{n_{.1}}{n}, \frac{n_{.2}}{n}, \cdots, \frac{n_{.c}}{n} \right\} = \{ f_{.1}, f_{.2}, \cdots, f_{.c} \}.$$

Il est ainsi implicitement supposé que les distributions des diverses entreprises $\left\{ \frac{n_{i1}}{n_{i.}}, \frac{n_{i2}}{n_{i.}}, \cdots, \frac{n_{ic}}{n_{i.}} \right\}$, $i = 1 \cdots r$, sont sensiblement les même, 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 [5]) :

$$d^2 = \sum_i \sum_j \frac{\left(n_{ij} - \frac{n_{i.} n_{.j}}{n} \right)^2}{\frac{n_{i.} n_{.j}}{n}} = n \sum_i \sum_j \frac{\left(f_{ij} - f_{i.} f_{.j} \right)^2}{f_{i.} f_{.j}}$$

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 d'indicateurs variant entre 0 et 1 ont été proposés dans la littérature (voir [1], [1] ou [6]). Par exemples :

Le coefficient de contingence de Pearson : $P = \left(\frac{d^2}{d^2 + n} \right)^{1/2}$

Le coefficient de contingence de Cramer : $V = \left(\frac{d^2}{n \ \inf\{(r-1);(c-1)\}} \right)^{1/2}$

Le τ_b *de Kendall :* $\tau_b(y/x) = \frac{\sum_i \sum_j \frac{n_{ij}^2}{n n_{i.}} - \sum_j \left(\frac{n_{.j}}{n} \right)^2}{1 - \sum_j \left(\frac{n_{.j}}{n} \right)^2}$

Les résultats pour le sous-groupe 8040-21-1 (bloc 8040, code activité 21, taille 1), présentés dans le Tableau 3, montrent que les valeurs des 3 indicateurs (P, V et tau) sont proches de 1 ce qui traduit des distributions par détails assez différentes.

Tableau 3 : Indicateurs de Pearson, Cramer et Kendall pour les sous-groupes 8040-21.

Taille	n	R	C	d2	P	V	Tau
1	1471966795	1394	13	15931315832	0.957	0.950	0.947
2	1557857891	145	10	10713115428	0.934	0.874	0.897
3	31981769030	118	12	213838929071	0.933	0.780	0.705

1.3. Homogénéité des sous-groupes Bloc-Activité-Taille

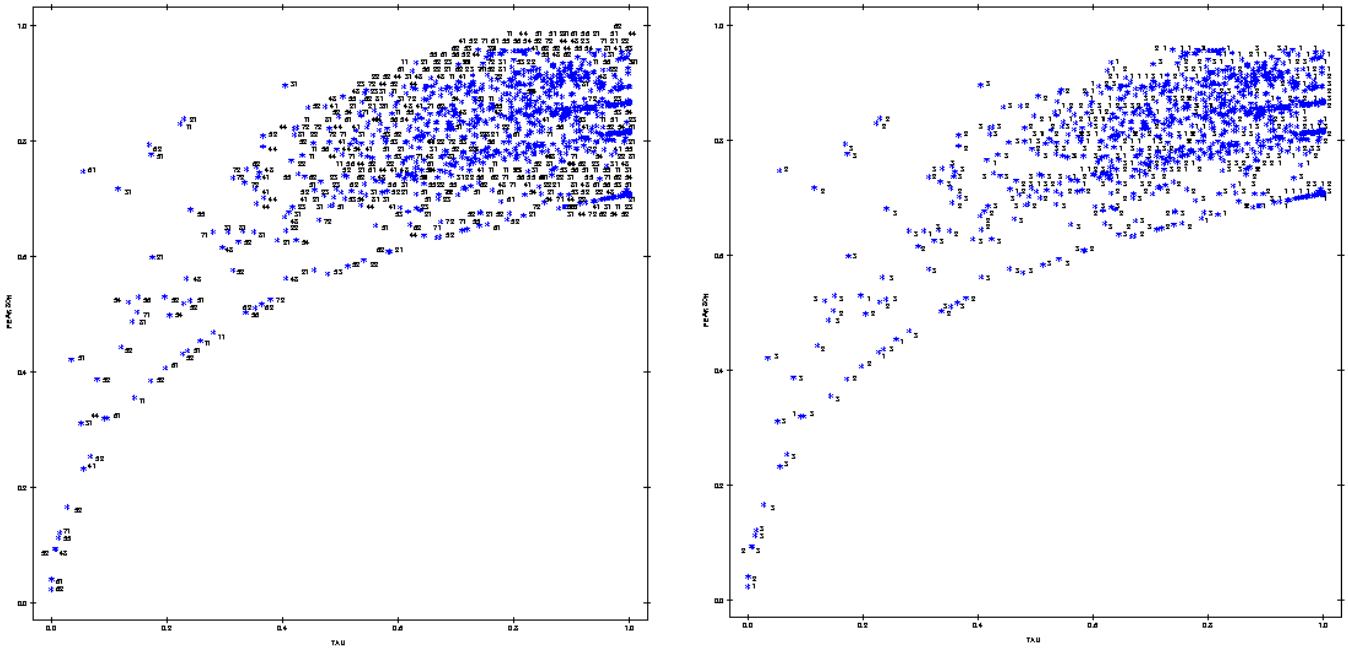
L'homogénéité des sous-groupes obtenus avec les différentes classifications habituellement utilisées (12 classifications ont été étudiées) a été mesurée avec les 3 indicateurs définis précédemment. Les résultats obtenus en utilisant par exemple une classification du code activité en 2 positions, sont présentés dans le Tableau 4 et la **Figure 1**.

Tableau 4 : Quelques statistiques sur les indicateurs de Cramer, Pearson et Kendall pour les sous-groupes Bloc-Activité à 2 positions-Taille.

Nombres	
N	1010
Missing values	110
$r=1\ c=1$	44
$r=1\ c\neq 1$	6
$r\neq 1\ c=1$	60
$r\leq 25$ (%)	32.14

Statistique	Cramer	Pearson	Tau
Mean	0.838	0.799	0.786
Median	0.866	0.817	0.836
Mode	1.000	0.707	1.000
Standard Deviation	0.153	0.117	0.209
Variance	0.023	0.014	0.044
Range	0.977	0.935	0.999
Interquartile Range	0.193	0.127	0.262
100% Max	1.000	0.958	1.000
99%	1.000	0.955	1.000
95%	1.000	0.927	1.000
90%	1.000	0.911	1.000
75% Q3	0.958	0.867	0.951
50% Median	0.866	0.817	0.836
25% Q1	0.765	0.740	0.690
10%	0.653	0.702	0.519
5%	0.566	0.635	0.353
1%	0.262	0.319	0.057
0% Min	0.023	0.023	0.001

Figure 1 : Répartition des indicateurs de Pearson et de Kendall pour les sous-groupes 8040-Code2-taille.
(Les points sont représentés par leur code activité dans le graphique gauche et par leur taille dans le graphique droit).



Quelques commentaires :

- Ces indicateurs n'ont pas pu être calculés pour 110 sous-groupes (9.8%). Cela arrive quand le nombre d'entreprises ou le nombre de détails reportés (les dimensions de la table donc) sont égaux à 1. Un nombre de détails reporté égal à 1 est plutôt une « bonne chose » puisque cela signifie que, dans le sous-groupe, les entreprises ont le même comportement de réponse.
- Plus de 30% des sous-groupes contiennent moins de 25 entreprises. Dans la procédure d'imputation atuelle, ils seront donc agrégés à d'autres sous-groupes ce qui diminuera vraisemblablement l'homogénéité des groupes ainsi reconstitués. Les chiffres présentés dans le tableau donnent donc une vision sans doute « optimiste » de l'homogénéité des sous-groupes.
- La majorité des sous-groupes présentent des valeurs élevées des indicateurs : 75% d'entre eux ont des valeurs des indicateurs de Pearson, Cramer et Kendall supérieures à 0.70. Les distributions ne sont donc pas équivalentes au sein des sous-groupes. La **Figure 1** est particulièrement claire à cet égard puisque la grande majorité des points se trouve dans le coin supérieur droit du graphique. Ces graphiques ne montrent d'ailleurs pas de lien entre l'homogénéité d'un sous-groupe et la taille ou l'activité des entreprises qui le composent.

Les résultats obtenus sont similaires, quelque soit la classification a priori retenue. Le **Tableau 5** donne ainsi quelques statistiques sur la distribution du coefficient de contingence de Pearson pour les différentes classifications étudiées, suggérées par les divisions clientes de Statistique Canada.

1.4. Premières conclusions

1.4.1. Pourquoi cela ne marche-t-il pas bien ?

Comme nous l'avons vu, la procédure actuelle d'allocation repose implicitement sur une homogénéité des distributions de détails au sein d'un même sous-groupe, hypothèse qui ne paraît pas raisonnable dans la plupart des cas étudiés.

Le Tableau présente les données fournies par les entreprises du code activité 5111 ne reportant que des détails pour le bloc 8040.

Comme on peut le voir, dans ce sous-groupe les entreprises ne renseignent qu'un seul détail sur les 13 possibles : soit le détail 10, soit le détail 13. La distribution observée des détails est donc soit (100 ;0) soit (0 ;100) sans que la taille ni le code activité à un niveau fin ne permettent de déduire la distribution de l'entreprise.

Comme il n'y a que 2 entreprises « moyennes » et deux « grosses » entreprises, la procédure va agréger les sous-groupes pour ne retenir que le sous-groupe constitué de toutes les entreprises. La distribution marginale calculée sera alors (3.3 ; 96.7), une distribution assez proche de celle des 11 entreprises ne reportant que le détail 13. Une majorité d'entreprises, 25 sur 36 dans notre exemple, se verraient donc affecter une distribution « fausse ».

Par ailleurs, l'estimation de la distribution marginale est par nature très sensible à la présence de valeurs « atypiques ». Si par exemple on retire les deux valeurs les plus fortes des détails observés, la distribution marginale passera à (59.1 ; 40.9), une distribution très loin des distributions réellement observées.

Tableau 5 : Distribution du coefficient de contingence de Pearson selon la définition a priori des sous-groupes .

Statistique	**Naics 2 taille**	**Ind80 taille Fréquent**	**Ind80 taille Important**	**Naics 4**	**Naics 4 Fréquent**	**Naics 4 Important**	**Naics 4 taille Fréquent**	**Naics 4 taille Important**	**Naics 6 Fréquent**	**Naics 6 Important**	**Naics 6 taille Fréquent**	**Naics 6 taille Important**
N	1010	605	560	1400	1192	1175	2074	1917	3074	3011	4360	4088
Missing values	110	163	208	158	366	383	1045	1202	1191	1254	2723	2995
$r=1\ c=1$	44	45	50	57	57	66	363	395	273	289	1091	1168
$r=1\ c\neq 1$	6	5	5	2	2	1	117	109	24	19	352	322
$r\neq 1\ c=1$	60	113	153	99	307	316	565	698	894	946	1280	1505
$r\leq 25$ (%)	32.14	45.05	45.31	26.19	26.57	26.51	58.06	58.19	38.15	38.45	61.13	61.34
Pearson's Coefficient												
Mean	0.799	0.708	0.730	0.811	0.720	0.745	0.689	0.714	0.710	0.732	0.683	0.709
Median	0.817	0.707	0.749	0.823	0.707	0.758	0.707	0.707	0.707	0.741	0.707	0.707
Mode	0.707	0.707	0.707	0.707	0.707	0.707	0.707	0.707	0.707	0.707	0.707	0.707
Standard Deviation	0.117	0.123	0.119	0.094	0.093	0.093	0.140	0.133	0.098	0.099	0.139	0.130
Variance	0.014	0.015	0.014	0.009	0.009	0.009	0.020	0.018	0.010	0.010	0.019	0.017
Range	0.935	0.864	0.868	0.948	0.888	0.844	0.894	0.935	0.894	0.882	0.894	0.935
Interquartile Range	0.127	0.114	0.115	0.118	0.110	0.110	0.125	0.124	0.117	0.113	0.129	0.125
100% Max	0.958	0.886	0.905	0.954	0.894	0.942	0.894	0.935	0.894	0.942	0.894	0.935
99%	0.955	0.866	0.892	0.946	0.866	0.897	0.861	0.889	0.864	0.890	0.863	0.884
95%	0.927	0.835	0.863	0.919	0.825	0.866	0.816	0.860	0.820	0.861	0.816	0.858
90%	0.911	0.816	0.851	0.907	0.815	0.852	0.812	0.842	0.813	0.844	0.810	0.836
75% Q3	0.867	0.790	0.809	0.875	0.792	0.811	0.781	0.802	0.783	0.804	0.774	0.797
50% Median	0.817	0.707	0.749	0.823	0.707	0.758	0.707	0.707	0.707	0.741	0.707	0.707
25% Q1	0.740	0.677	0.694	0.757	0.682	0.701	0.656	0.678	0.666	0.692	0.646	0.672
10%	0.702	0.589	0.617	0.707	0.617	0.640	0.558	0.586	0.599	0.616	0.537	0.578
5%	0.635	0.520	0.548	0.696	0.579	0.589	0.405	0.477	0.545	0.563	0.402	0.471
1%	0.319	0.112	0.191	0.542	0.405	0.462	0.083	0.146	0.372	0.403	0.089	0.147
0% Min	0.023	0.022	0.037	0.006	0.006	0.098	0.000	0.000	0.001	0.060	0.000	0.000

Tableau 6 : Bloc 8040, code activité 5111

Nombre de détails = 13

Nombre de détails reportés = 2

Nombre d'entreprises = 36

Coefficient de Pearson = 0.707

Code activité	Taille	Somme des détails	Détail 10	Détail 13
511120	1	751	0	751
511120	1	7194	0	7194
511110	1	12633	0	12633
511130	1	37123	0	37123
511110	1	93480	0	93480
511130	1	102917	0	102917
511120	1	143334	0	143334
511110	1	143513	0	143513
511110	1	289736	0	289736
511120	1	42	42	0
511120	1	44	44	0
511120	1	161	161	0
511120	1	256	256	0
511120	1	296	296	0
511120	1	327	327	0
511130	1	543	543	0
511130	1	765	765	0
511120	1	887	887	0
511130	1	1686	1686	0
511120	1	1902	1902	0
511130	1	2981	2981	0
511130	1	5349	5349	0
511130	1	11192	11192	0
511120	1	11302	11302	0
511130	1	12558	12558	0
511130	1	12992	12992	0
511130	1	14884	14884	0
511130	1	21483	21483	0
511130	1	55761	55761	0
511120	1	78360	78360	0
511130	1	211488	211488	0
511120	1	433875	433875	0
511130	2	857258	0	857258
511130	2	277076	277076	0
511120	3	33717045	0	33717045
511120	3	43832	43832	0
Distributions marginales				
Avec toutes les entreprises			3.3	96.7
Sans les 2 plus hautes valeurs			59.1	40.9

Notons enfin que dans ce cas, les deux détails reportés (10 et 13) sont pris en compte dans le calcul de la distribution marginale puisqu'ils sont reportés par plus de 10% des entreprises.

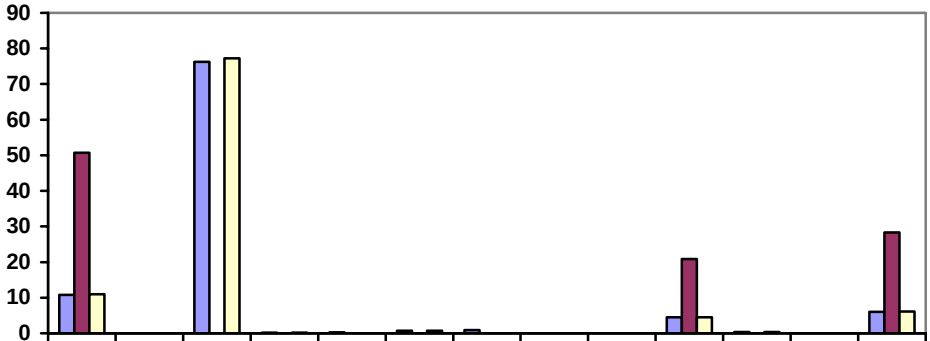
L'utilisation des détails les plus « fréquents » peut aussi considérablement affecter l'estimation de la distribution marginale dans la mesure où la détermination de ces détails ne fait pas intervenir les valeurs reportées.

La Figure 2 montre l'impact de la prise en compte des détails fréquents seulement. Cette figure montre aussi les distributions marginales calculées à partir des « détails importants ». Ces détails sont définis comme étant les détails les plus fréquents et dont la valeur cumulée est supérieure à 90% de la somme totale des détails.

Pour le bloc 8040 dans son ensemble, l'utilisation de la distribution marginale calculée à partir des « détails fréquents » sous-estime terriblement le détail 3 alors que l'utilisation des « détails importants » traduit assez bien la distribution réellement observée.

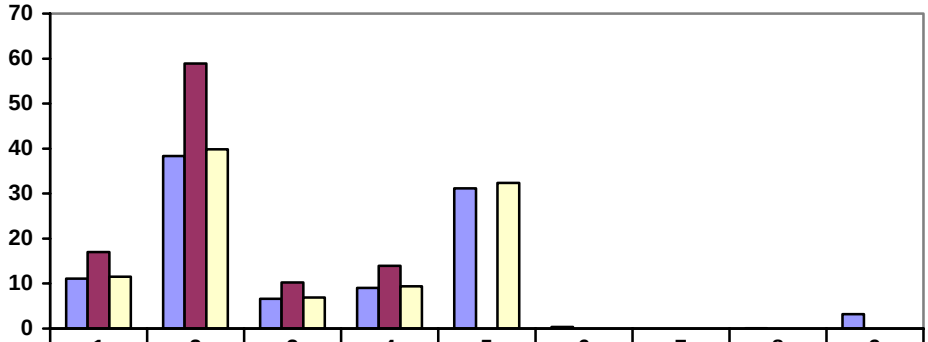
Figure 2 : Comparaison des distributions marginales calculées avec tous les détails, les détails fréquents et les détails importants selon le bloc.

Bloc 8040



	1	2	3	4	5	6	7	8	9	10	11	12	13
Total	10,82	0	76,23	0,17	0,29	0,7	0,92	0	0,01	4,46	0,33	0	6,05
Fréquents	50,74	0	0	0	0	0	0	0	0	20,9	0	0	28,36
Importants	10,96	0	77,19	0,17	0	0,71	0	0	0	4,51	0,33	0	6,13

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 situation est similaire : la distribution des « détails fréquents » ne prend pas en compte le détail 5 qui représente environ 30% du total des détails reportés par les entreprises tout en surestimant 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%.

1.4.2. Que peut on faire ?

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 des clefs de répartitions déduites d'enquêtes effectuées auprès des entreprises et qui s'intéresseraient justement aux détails. Malheureusement cette idée se heurte à deux obstacles importants :
   - L'un de principe. L'utilisation des données administratives répond à la volonté de diminuer la charge de réponse des entreprises. Il paraît donc peu souhaitable de continuer à faire des enquêtes pour corriger ces mêmes données administratives.
   - L'autre tient au fait que les concepts utilisés dans les déclarations fiscales sont généralement différents de ceux utilisés par les statisticiens dans leurs enquêtes. Pour résoudre ce problème, Statistique Canada a mis au point un plan comptable normalisé (COA, « Chart of Accounts ») pour les données sur la situation financière et les résultats financiers des entreprises. Cette nouvelle norme, mise en application en mai 2001, est conçue pour la déclaration statistique et sert également de modèle pour la diffusion des données sur la situation financière et les résultats financiers.

2. 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. Là encore, 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 Tableau , 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.

3. Le plan comptable normalisé (COA) 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.

4. 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.

2. Une méthode alternative basée sur une classification hiérarchique et une analyse discriminante.

2.1. Une première analyse du bloc 8040

Une première analyse descriptive de la façon dont les entreprises répondent aux variables du bloc 8040 permet de justifier une approche par classification mais aussi d'en apprécier les limites.

Le bloc 8040 tableau 7 est constitué de 13 détails libellés de 8041 à 8053, qui correspondent à trois variables COA, notées dans la suite C4113, C4122 et C6. Mais finalement, les entreprises ne reportent que peu d'entre eux. Ainsi (voir tableau 8) 84% de ces entreprises ne renseignent qu'un détail et 99% au plus trois.

Le Tableua 9 précise la configuration[2] de la réponse des entreprises. 73 configurations de réponse apparaissent dans le fichier de données mais une très grande majorité d'entreprises (4329 sur 5997 soit 72%) ne renseigne que les détails 1, 10 ou 13.

Notons au passage que quelques configurations (ombrées dans le tableau 9) montrent que certains détails sont présents dans les déclarations d'entreprises reportant à la fois un générique et des détails sans qu'ils soient présents dans aucune déclaration d'entreprise ne reportant que des détails. Cela montre les limites de imputation et suggère que les déclarations « mixtes » pourraient être traitées de façon différente des autres déclarations.

Tableau 7 : Détails du bloc 8040 et variables COA associées

SALES FROM RESOURCE PROPERTIES FLDS 8040 TO 8053

8040		Resource property sale amount	
Detail #	Detail		**COA**
1	8041	Petroleum and natural gas sale amount	4113
2	8042	Related party petroleum natural gas sale amount	4113
3	8043	Gas marketing income amount	4113
4	8044	Resource industry processing revenue amount	4113
5	8045	Pipeline income amount	4113
6	8046	Seismic sale amount	4113
7	8047	Mining income amount	4113
8	8048	Coal income amount	4113
9	8049	Oil sand income amount	4113
10	8050	Royalty income amount	4122
11	8051	Oil and gas partnership joint venture income amount	6
12	8052	Mining partnership joint venture income amount	6
13	8053	Other resource production income amount	4113

Le comportement de réponse des entreprises est donc fortement typé et il existe au sein de ce bloc des groupes d'entreprises présentant des distributions identiques.

Cela est d'autant plus vrai si on se restreint aux variables COA. Dans ce cas, seuls 8 comportements de réponse différents peuvent être observés (voir tableau10) et 60% des entreprises ne reportent que le détail C4113, 23% reportent le détail C4122 seulement, 4% le seul détail C6 et 12% un mélange des détails C4113 et C4122.

Tableau 8 : Répartition des entreprises du bloc 8040 selon leur type de réponse.

Nombre de détails	Générique seulement	Détails seulement		Générique & Détails	Total
0	1693	0	0.00	0	1693
1	0	5032	83.91	122	5154
2	0	729	12.16	9	738
3	0	198	3.30	5	203
4	0	32	0.53	1	33
5	0	5	0.08	0	5
6	0	1	0.02	0	1
Total	1693	5997	100.00	137	7827

[2] Cette configuration est représentée dans le tableau par une variable composée de 13 caractères '0' et de '1'. Le $i^{ème}$ caractère est '1' si l'entreprise a renseigné le détail i.

Pour essayer de retrouver ces sous-groupes, on fait une classification ascendante hiérarchique du tableau de contingence des distributions en valeur par détail COA pour chaque entreprise, en utilisant une métrique du χ^2 (adaptée à ce type de données) et la stratégie de Ward. L'arbre de classification de la figure 3 suggère un découpage en 4 classes dont les caractéristiques sont résumées dans le tableau 11.

Comme attendu, les classes sont assez homogènes et assez bien typées. La classe 1 contient essentiellement des entreprises ne reportant que le détail C4113, la classe 2 est liée au détail C4122 et la classe 4 au détail C6. La distribution marginale de la classe 3, peu nombreuse, est assez proche de la distribution d'ensemble.

Ces premiers résultats sont encourageants. Il est donc possible de dégager des groupes assez nombreux d'entreprises homogènes à partir d'une classification hiérarchique des entreprises ne reportant que des détails. Il reste maintenant à « prévoir » la classe à laquelle appartient une entreprise reportant une valeur générique seulement.

Tableau 9 : Répartition des entreprises du bloc 8040 selon leur comportement de réponse (détails originaux).

Configuration	Détails seulement	Générique & Détails
1000000000000	1666	10
0000000001000	1396	67
0000000000001	1267	29
1000000001000	444	1
0000000000100	201	3
0001000000000	135	7
1001000001000	114	1
0000001000000	98	0
0000100000000	72	0
1001000000000	70	0
0000010000000	64	4
0010000000000	57	1
0000000001001	55	2
1000000000001	53	0
0000000000010	36	0
1000000001001	35	0
1001000001001	20	0
0000000010000	16	0
0100000000000	15	0
1000000000100	15	0
0000000001100	13	1
0100000001000	13	0
1000000001100	13	0
1001000000001	12	0
0000000100000	9	1
1000010000000	8	0
1000010001000	8	0
0001000001000	3	4
1000100000000	6	0
1000100001000	6	0
1010000000000	6	0
1100000000000	6	0
0000001000001	5	0
0001000000001	5	0
1001000001100	4	0
1011000001000	4	0
0000000000101	3	0
0000010001000	3	0
0000100000001	3	0
0000000011000	2	0
0000000100010	2	0
0000001001000	2	0
0000110000000	2	0
0001000000100	2	0
0001000001001	0	2
0010000000001	2	0
1000001000000	2	0
1000001001000	2	0
1001000001101	2	0
0000000101000	1	0
0000010001001	1	0
0000100001000	1	0
0000100001001	0	1
0000100010000	1	0
0001001000000	1	0
0001010000100	1	0
0001100000000	0	1
0010000001101	0	1
0011100000000	0	1
0101100000000	1	0
1000100001001	1	0
1001010000000	1	0
1001010001000	1	0
1001010001001	1	0
1001100001000	1	0
1001100001101	1	0
1010000001001	1	0
1011000000000	1	0
1011000001001	1	0
1100000000100	1	0
1101000000000	1	0
1101100000001	1	0
1110000000000	1	0

Figure 3 : arbre de la classification selon les distributions en détails COA.

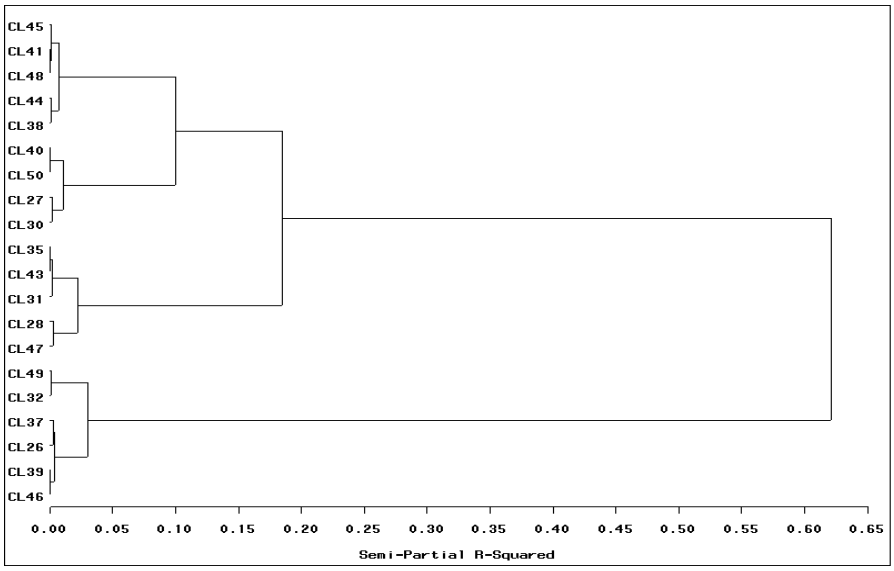


Tableau 10 : Répartition des entreprises du bloc 8040 selon les montants reportés pour les détails COA.

COA Configuration	C4113	C4122	C6	Nombre
001				
Détails seulement	$0	$0	$717,077,749	237
Générique & détails	$0	$0	$19,749	3
Total	$0	$0	$717,097,498	240
010				
Détails seulement	$0	$840,361,629	$0	1396
Générique & détails	$0	$110,469,698	$0	67
Total	$0	$950,831,327	$0	1463
011				
Détails seulement	$0	$1,026,185	$1,250,599	13
Générique & détails	$0	$15,243	$33,887	1
Total	$0	$1,041,428	$1,284,486	14
100				
Détails seulement	$50,901,515,822	$0	$0	3587
Générique & détails	$13,723,162	$0	$0	54
Total	$50,915,238,984	$0	$0	3641
101				
Détails seulement	$20,284,877,733	$0	$2,508,675,005	24
Générique & détails	$0	$0	$0	0
Total	$20,284,877,733	$0	$2,508,675,005	24
110				
Détails seulement	$4,510,964,932	$233,489,642	$0	720
Générique & détails	$15,643,828	$14,221,683	$0	11
Total	$4,526,608,760	$247,711,325	$0	731
111				
Détails seulement	$419,711,920	$13,193,176	$1,072,490,562	20
Générique & détails	$29,265	$44,597	$16,067	1
Total	$419,741,185	$13,237,773	$1,072,506,629	21
Total	$76,146,466,662	$1,212,821,853	$4,299,563,618	6134

Tableau 11 : caractéristiques des classes

Classe	Effectif	Pearson	Cramer	Tau	C4113	C4122	C6
1	4109	0.403	0.312	0.098	99.5	0.5	0.0
2	1527	0.593	0.521	0.279	1.3	98.6	0.1
3	93	0.346	0.261	0.038	61.3	38.5	0.2
4	268	0.651	0.607	0.458	2.9	0.6	96.5
ensemble	5997	0.811	0.981	0.954	69.1	26.5	4.4

2.2. La méthodologie proposée

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 » par nature ne peut reposer que sur des variables disponibles dans la réponse de l'entreprise c'est à dire essentiellement les génériques (ou somme des détails) ou les détails des autres blocs et certaines caractéristiques de l'entreprise (code 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 fixer la meilleure stratégie.

2.2.1. 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.

[3] faite avec la 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] avec la PROC FASTCLUS de SAS

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.

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 :

- Une analyse discriminante paramétrique
- Une analyse discriminante non paramétrique utilisant la méthode des k plus proches voisins (avec k=15 ou k=20)
- Un 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.

2.2.2. Quelques premiers résultats

Une simulation à grande échelle sur les données du bloc 8040 permet 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.

2.3. 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.

2.3.1. 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

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

fréquente dans d'autres domaines : tests d'algorithmes statistiques (voir [9]), tests de méthodes de prévisions en séries temporelles (voir [7]) etc.
La procédure de test repose ici sur les données des années 2001 et 2002 :

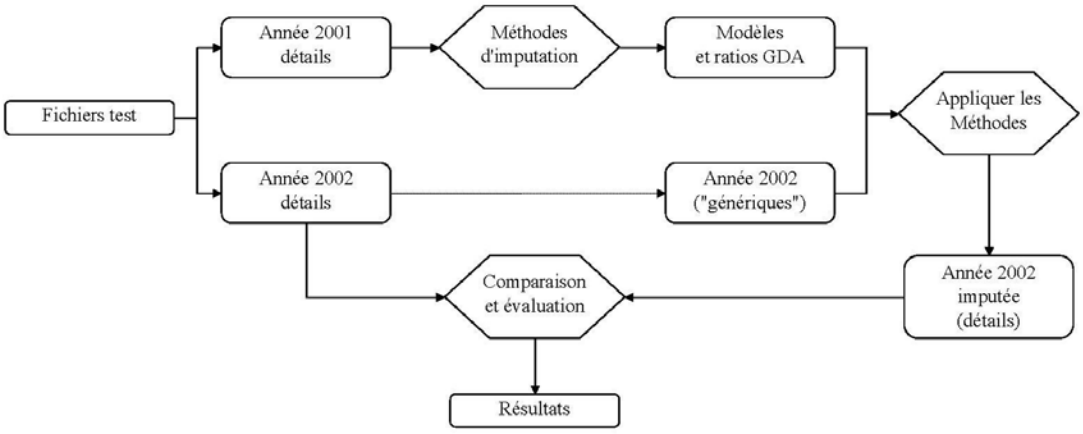
> Les données de l'année 2001 sont utilisées pour calculer les distributions qui seront utilisées pour imputer les données de l'année 2002.
>
> Pour chaque bloc étudié, on sélectionne les entreprises qui en 2002 n'ont reporté que des détails. Ces valeurs sont agrégées pour définir un générique « fictif » 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 3 résume le processus d'évaluation mis en oeuvre.

2.3.2. 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).

Figure 4 : Diagramme du processus d'évaluation des méthodes d'imputation



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 :

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

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.

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

2.3.3. Quelques éléments sur les blocs étudiés

L'évaluation s'est faite sur deux blocs : le bloc 8040 (variables de revenu) et le bloc 9760 (variables de dépense)

Le bloc 8040

Pour l'année 2001 (resp. 2002), 6145 (resp. 6226) entreprises ont reporté des détails seulement. Ce bloc correspond à 3 détails COA.

Les 6145 entreprises ont été classées en utilisant des classifications automatiques à 3 (Cluster3) et 4 (Cluster4) classes, et différents « attracteurs » : Attractor80, Attractor90, Attractor95. Une nouvelle classification ad hoc (Attractor100) a été définie en formant une classe supplémentaire en retirant des classes définies par Cluster4, les entreprises reportant à la fois les détails C4113 et C4122.

Le 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.

2.4. Principaux résultats

2.4.1. Bloc 8040

Homogénéité

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

Tableau 12 : Homogénéité des groupes obtenus par les différentes classifications (bloc 8040, coefficient de Pearson)

Cluster	Cluster3	Cluster4	Attractor 80	Attractor 90	Attractor 95	Attractor 100
1	0.536	0.333	0.432	0.314	0.174	0
2	0.566	0.566	0.180	0.176	0.171	0
3	0.315	0.315	0.331	0.325	0.271	0
4		0.493	0.607	0.617	0.590	0.445
5						0.501

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

Précision

Pour chacune des 7 variables de classification retenues, on a sélectionné les 10 meilleurs modèles de prédiction de la classe en fonction du taux d'erreur de classement.

Les figures 5 et 6 représentent les meilleurs modèles des 7 types de variables de classe en fonction des critères de précision précédemment définis : SSE et Macro CV, coefficients de Pearson et Micro CV. Les 2 graphiques donnent le même message : pour ce bloc, la nouvelle stratégie est meilleure que la stratégie actuellement employée.

En particulier, les modèles basés sur la variable Attractor100 sont les meilleurs en terme de coefficients de Pearson et de MicroCV ; alors que ceux basés sur la variable C3 sont les meilleurs en termes macro (SSE et Macro CV).

Pour faire un choix définitif, on peut regarder quels modèles respectent au mieux la répartition en termes de pourcentages. La figure 7 représente les performances des meilleurs modèles (au sens du SSE) de chaque variable de classe.

Figure 5 : Représentation des meilleurs modèles pour chaque variable de classe en fonction du SSE et de la médiane du coefficient de Pearson (Bloc 8040**)**

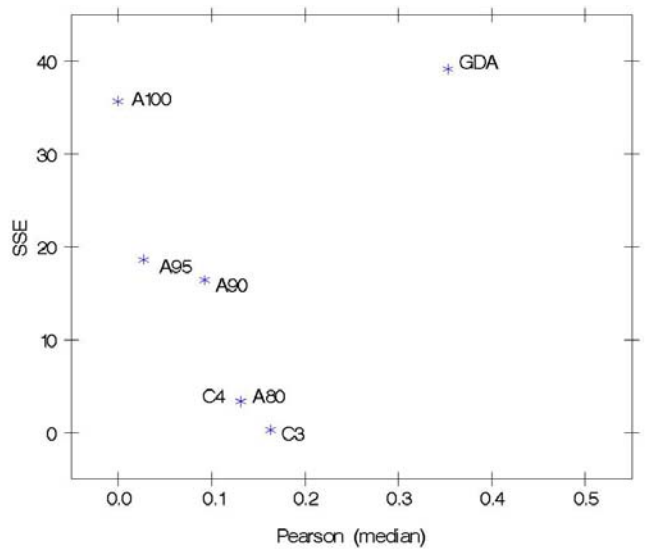


Figure 6 : Représentation des meilleurs modèles pour chaque variable de classe en fonction du Macro_Pseudo_CV et de la médiane du Micro_Pseudo_CV (Bloc 8040)

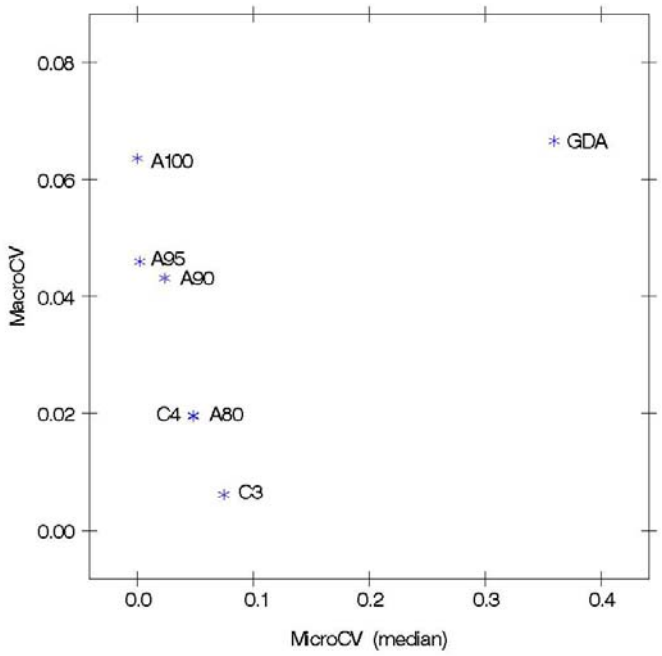
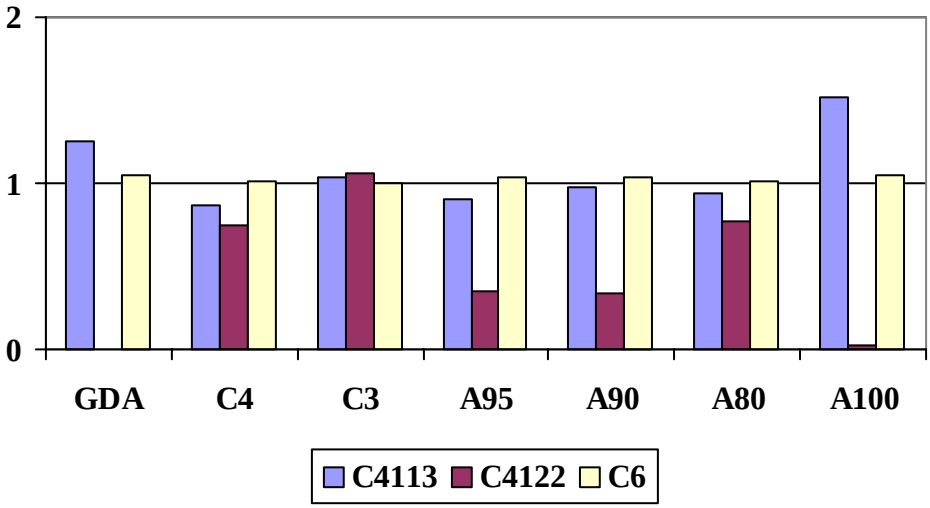


Figure 7 : Respect des poids de chaque détail (SSEP) pour les meilleurs modèles de chaque variable de classe.



La procédure actuelle d'imputation (GDA) sous estime énormément le détail C4122, comme le font d'ailleurs les autres modèles sélectionnés à l'exception du modèle relatif à la variable de classification C3.

Ce modèle est de plus particulièrement simple puisqu'il ne fait intervenir que 4 variables (Revenue L8436 ratio et N21).

En résumé, pour le bloc 8040, la meilleure stratégie d'imputation des détails est basée sur une classification automatique en 3 classes, la classe étant ensuite prédite par une analyse discriminante sur un modèle à 4 variables explicatives (Revenu L8436 ratio et N21) discrétisées en 30 groupes, et en utilisant une méthode d'estimation non paramétrique.

2.4.2. Bloc 9760

Homogénéité

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

Tableau 13 : 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

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). Comme pour le bloc 8040, ces modèles ont été choisis en fonction le leur taux d'erreur de classement.

Figure 8 : Représentation des meilleurs modèles pour chaque variable de classe en fonction du SSE et de la médiane du coefficient de Pearson (Bloc 9760)

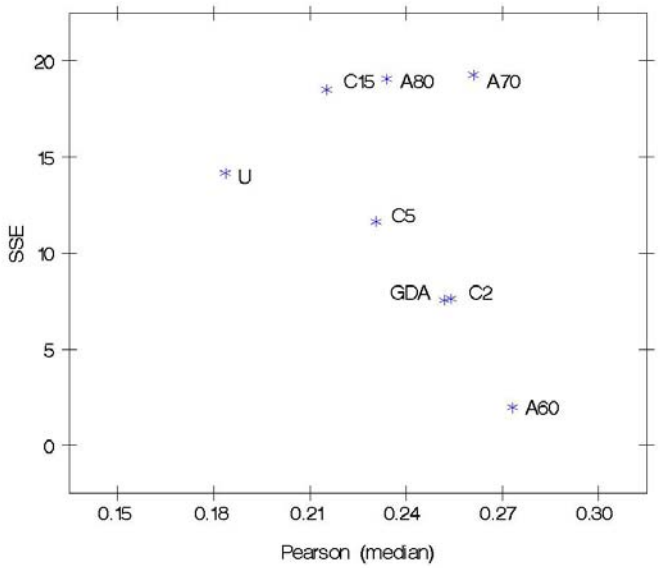
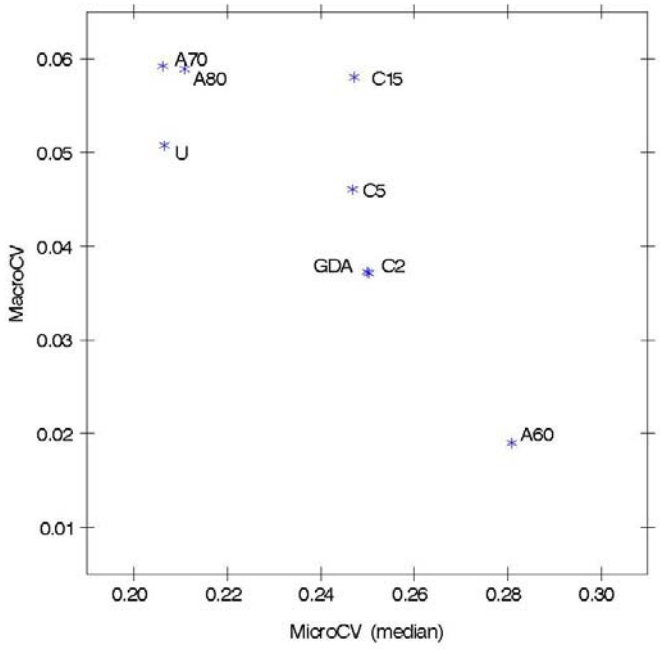


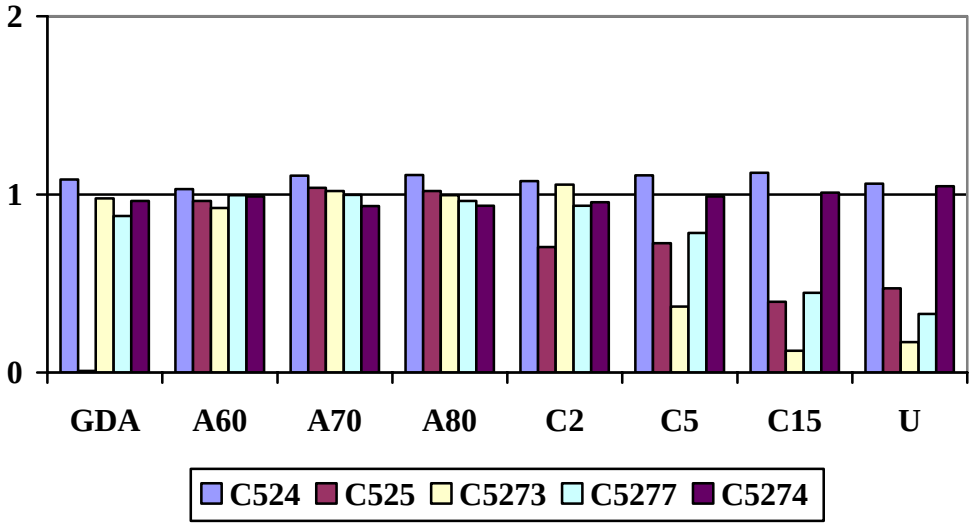
Figure 9 : Représentation des meilleurs modèles pour chaque variable de classe en fonction du Macro_Pseudo_CV et de la médiane du Micro_Pseudo_CV (Bloc 9760)



Les figures 8 et 9 représentent les meilleurs modèles des 8 types de variables de classe en fonction des critères de précision précédemment définis : SSE et Macro CV, coefficients de Pearson et Micro CV. Les 2 graphiques donnent un message similaire. 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.

Là encore, c'est essentiellement la performance des modèles sur la restitution des pourcentages des différents détails qui fera la différence. La figure 10 présente les performances des meilleurs modèles (au sens du SSE) de chaque variable de classe.

Figure 10 : Respect des poids de chaque détail (SSEP) pour les meilleurs modèles 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.

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.

3. 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.

Bibliographie

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

[2] Helmer, P., Lafontaine-Sorgo, D., Lalande, D., Thibault, D. (2001), Generic to Detail Allocation: GIFI 1999 Documentation, Statistics Canada.

[3] Rondeau, C. (2003), Specifications of the GDA ratios calculation, Tax data Division, Statistics Canada.

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

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

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

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

[8] NIST, (19ti, A. (1996), *An Introduction to Categorical Data Analysis*, New York: John Wiley & Sons, Inc.

[9] 98), "StRD: Statistical Reference Datasets for Assessing the Numerical Accuracy of Statistical Softwares", http://www.nist.gov/itl/div898/strd.