Imputation de distributions dans les données administratives

Rong Huang, Dominique Ladiray

JMS – 15 mars 2005

Préambule

- Présentation dédiée à notre ami Alain Desrosières
  - *La politique des grands nombres – Histoire de la raison statistique* (La Découverte)
- Approche originale (?) : application de la CAH et de l'analyse discriminante non paramétrique.
  - Voir aussi présentation suivante
- Travail commun INSEE-StatCan

Plan

- ✓ Quelques mots sur les données fiscales
- ✓ Blocs, génériques et détails
- ✓ Procédure actuelle et homogénéité
- ✓ Une nouvelle procédure basée sur une classification automatique
- ✓ Comparer les méthodes d'imputation
- ✓ Quelques résultats

Données fiscales

- ✓ Fournies par l'Agence du Revenu du Canada
- ✓ Très utilisées à Statistique Canada
  - Pour alléger la charge de réponse des entreprises
  - Réduction de 50 à 60 % de la taille des échantillons des enquêtes annuelles d'entreprises
- ✓ "Exhaustives" mais doivent être "nettoyées" : valeurs manquantes, aberrantes etc.
- ✓ Les divisions clientes veulent des données complètes et "propres" au niveau de l'entreprise

Blocs, génériques et détails

- Blocs, génériques 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

- Imputer la distribution des détails si elle manque

Procédure actuelle (GDA)

- Sous groupes a priori: Bloc * Activité * Taille
- Distribution marginale calculée sur les entreprises ne reportant que des détails
- Détails « fréquents » : renseignés par au moins 10% des entreprises du sous-groupe
- Au moins 25 entreprises par sous-groupe
- Hypothèse :
  - La taille et l'activité sont des critères qui permettent de définir des sous-groupes « homogènes »

Mesurer l'homogénéité

- ✓ Dans un sous-groupe la distribution en détails ne dépend pas de l'entreprise
- ✓ Mesures basées sur le Chi2 : Pearson, Kendall etc.

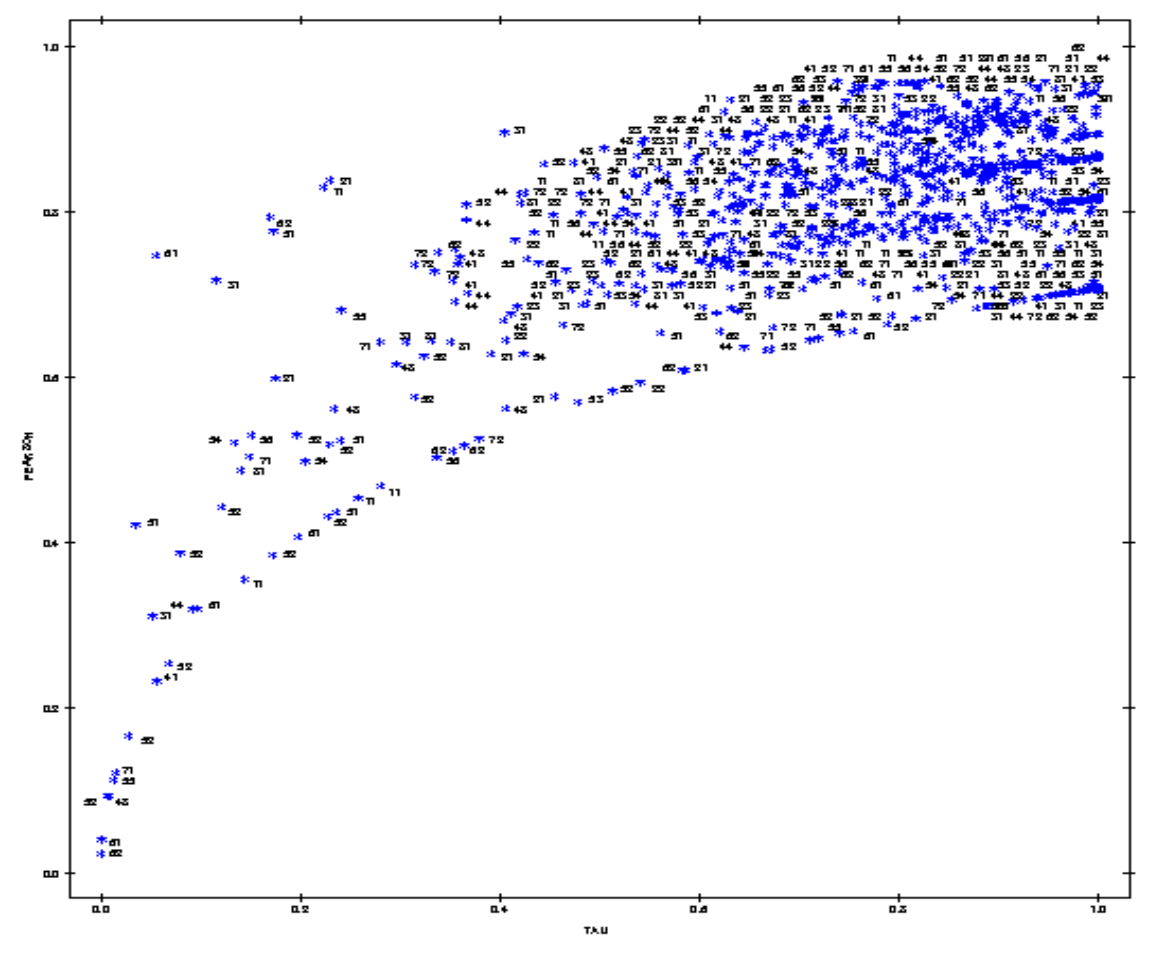
$$d^2 = \sum_i \sum_j \frac{\left(n_{ij} - \frac{n_{i.} n_{.j}}{n} \right)^2}{\frac{n_{i.} n_{.j}}{n}} = n \sum_i \sum_j \frac{\left(f_{ij} - f_{i.} f_{.j} \right)^2}{f_{i.} f_{.j}}$$

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

Bloc*Activité*Taille



Fabriquer des classes homogènes

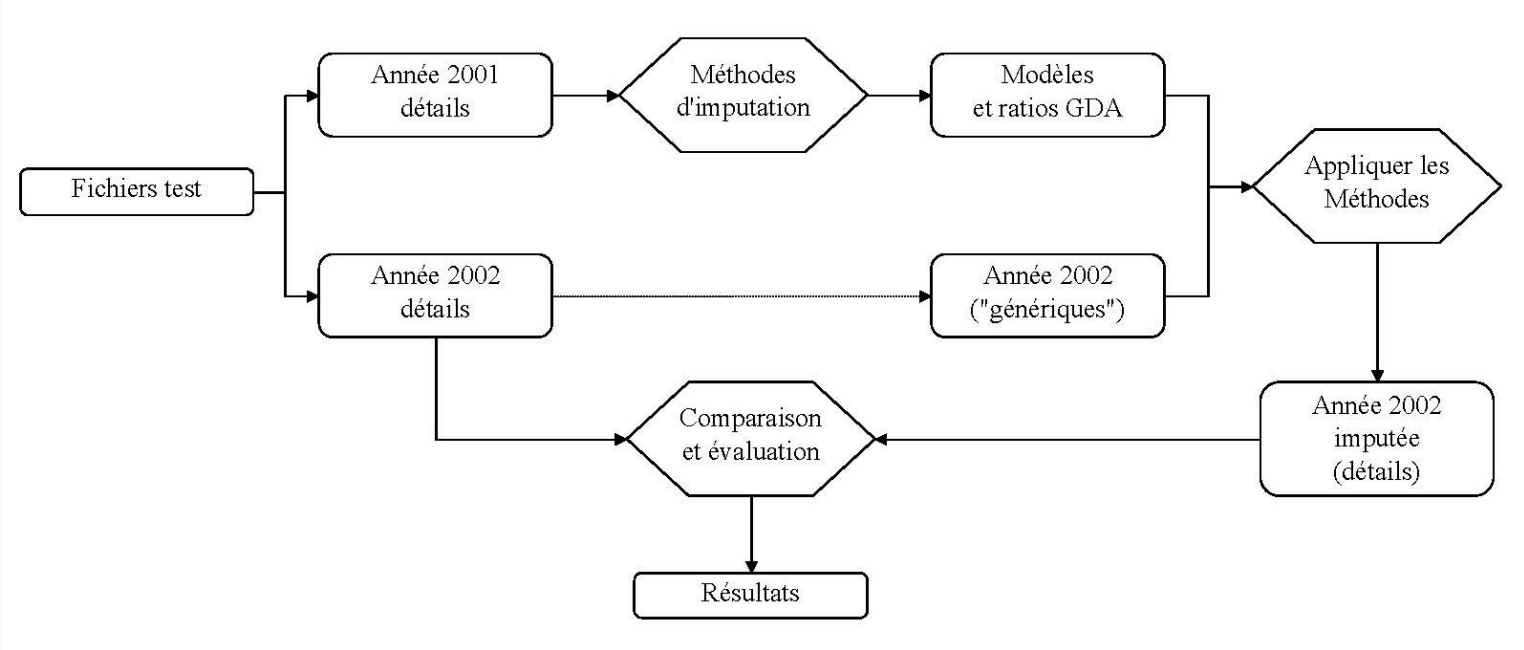
- Exemple du Bloc 8040 :
  - 5997 entreprises, 3 détails (C4113, C4122, C6)
- Définir des classes a priori. «Attracteur80»
  - détail >= 80% (3 classes), une classe « autres »
- Par classification automatique
- Coefficients de Pearson

Cluster	Cluster3	Cluster4	Attractor80	Attractor90	Attractor95	Attractor100
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

Stratégie

- ✓ Pour un bloc donné
- ✓ Préparer les données
  - brutes, discrétisées (combien de classes ?)
- ✓ Classification
  - Laquelle choisir ? Automatique ? A priori ?
- ✓ Retrouver la classe d'une entreprise ne reportant qu'un générique (modéliser)
  - Quelles variables (plusieurs centaines possibles)
  - Quelle méthode (DISCRIM, CATMOD)
  - Quels paramètres pour la méthode ?
- ✓ Pour répondre à ces questions: simuler

Comparer les méthodes d'imputation



- ✓ Critères statistiques de comparaison ?

Critères de comparaison

- ✓ Au niveau de l'entreprise (micro)
  - Coefficient de Pearson sur tableau « réel * imputé »
  - $Micro\_pseudo\_CV_j = \sqrt{\sum_i (x_{ij} - \hat{x}_{ij})^2} / \sum_i x_{ij}$
- ✓ Au niveau du bloc (macro)

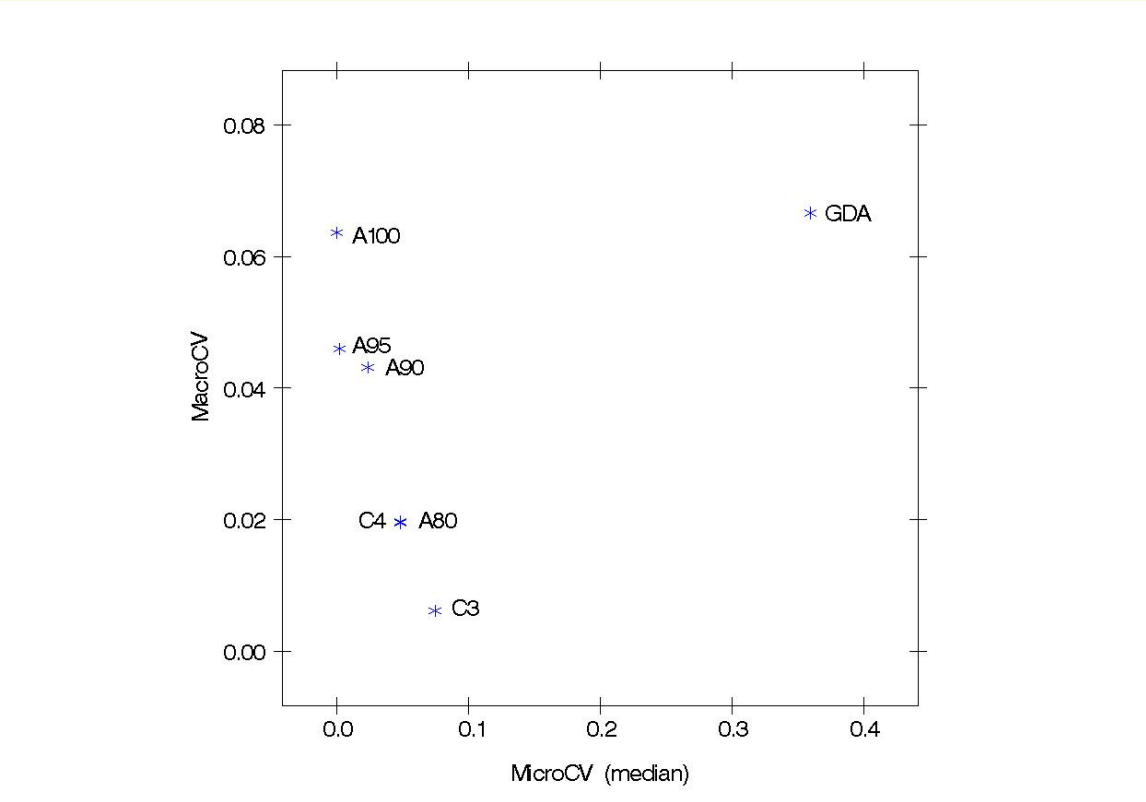
$$SSE = \sum_i (t_i - \hat{t}_i)^2$$

$$SSEP = \sum_i (\frac{t_i}{\hat{t}_i} - 1)^2$$

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

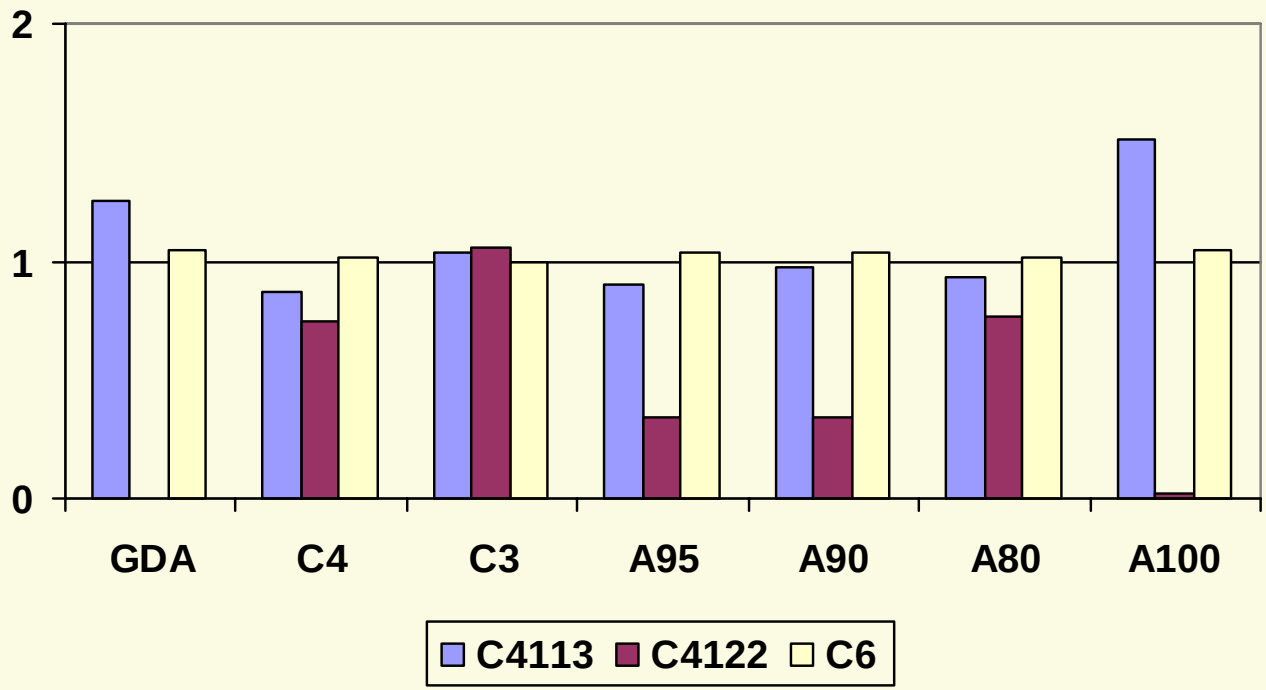
Résultats pour le Bloc 8040 (1)

- ✓ Analyse discriminante non paramétrique sur variables discrétisées (30 groupes).



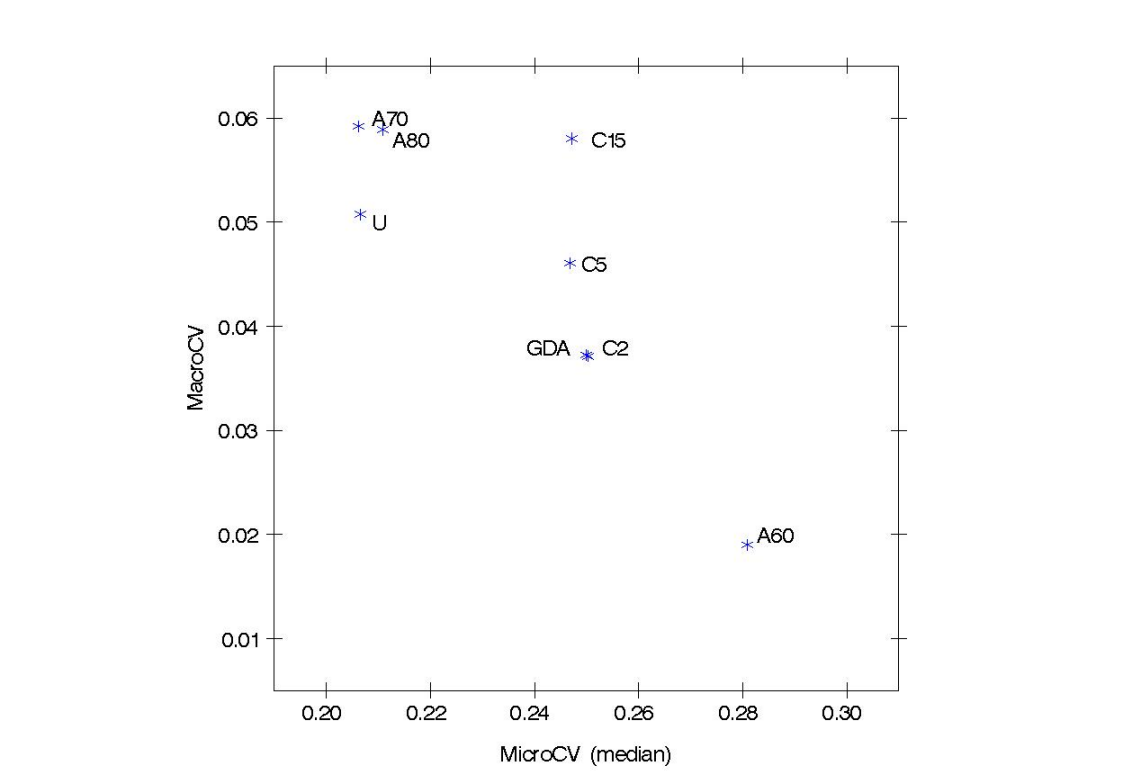
Résultats pour le Bloc 8040 (2)

- Préservation des distributions : C3



Résultats pour le Bloc 9760 (1)

- ✓ 21000 entreprises, 5 détails. Modèle linéaire sur données qualitatives.



Résultats pour le Bloc 9760 (2)

- Préservation des distributions : A60, A70 ou A80