Mise à jour de répertoires d'entreprises

JMS 2015





Philippe Brion
INSEE

INSEE
INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES

Plan de l’exposé

- 1 - Présentation du problème
- 2 - Exemple simplifié : problèmes de classement
- 3 - Approche économique du problème
- 4 - Conclusions

1. Présentation du problème (1)

- Le répertoire statistique d'entreprises : un matériau très riche pour être utilisé comme base de sondage
- Les variables généralement utilisées pour la définition du champ et les plans de sondage :
  - Code APE
  - Variables de taille (effectif, chiffre d'affaires)

1. Présentation du problème (2)

- Généralement on dispose de métadonnées :
  - Date de dernière mise à jour concernant chaque unité
  - Probabilité qu'une donnée soit inexacte, en fonction d'éléments recueillis par ailleurs
- Comment utiliser au mieux ces métadonnées si l'on dispose d'un certain budget pour mener des opérations d'amélioration du répertoire ?

2. Exemple simplifié : questions de classement sectoriel

- Un ensemble composé de deux parties : A et B (correspondant par exemple à deux codes APE)
- On cherche à estimer le total d'une variable Y (par exemple le chiffre d'affaires) sur le secteur A
- Le problème est que des unités sont mal classées : on note Pi la probabilité de chaque unité i d'être bien classée
- Dans l'approche proposée, l'enquête ne sert pas à réévaluer la valeur du code APE (on n'est donc pas dans le cadre utilisé pour Esane); on utilise donc les valeurs disponibles dans le répertoire pour l'estimateur

2. Exemple simplifié : questions de classement sectoriel, sans sondage (1)

- On se place ici dans le cas où on ne fait pas de sondage dans la partie A
- On se limite aux unités classées dans A, et on ne prend pas en compte celles classées en B
- Le classement du répertoire étant utilisé, l'estimateur est biaisé
- La modélisation proposée : c'est la grandeur qu'on cherche à estimer qui est aléatoire (elle dépend du classement « réel » de chaque unité)

2. Exemple simplifié : questions de classement sectoriel, sans sondage (2)

- Plus précisement, le paramètre à estimer vaut :

$$\theta = \sum_{A} 1(i \in A) Y_i + \sum_{B} (1 - 1(j \in B)) Y_j$$

- Et le risque associé à l'estimateur utilisé vaut :

$$\mathrm{E}(\text{estimat} - \theta)^2 = E\left(\sum_{A} (1 - 1(i \in A)) Y_i - \sum_{B} (1 - 1(j \in B)) Y_j\right)^2$$

- Ce qui vaut :

$$\left[\sum_{A} (1 - P_i) Y_i - \sum_{B} (1 - P_j) Y_j\right]^2 + \sum_{A+B} P_i (1 - P_i) Y_i^2$$

2. Exemple simplifié : questions de classement sectoriel, sans sondage (3)

- On met à jour des unités à la fois dans A et B, sur des sous populations A1 et B1
- Restent des sous-populations A2 et B2 sur lesquelles on reconduit la formule précédente
- Idée : maximiser le gain obtenu grâce à cette mise à jour

2. Exemple simplifié : questions de classement sectoriel, avec sondage

- On fait maintenant un sondage dans la partie A ; par exemple un SAS de taille n
- Deux niveaux aléatoire :
  - 1er niveau : classement dans A ou B
  - 2ème niveau : SAS de n unités dans A
- En décomposant :

$$\mathrm{E}(\text{estimat} - \theta)^2 = E_1 E_2 (estimat - E_2(estimat) + E_2(estimat) - \theta)^2$$

- Soit :

$$\mathrm{E}(\text{estimat} - \theta)^2 = E_1(\text{var } iancesond) + E_1(\sum_A Y_i - \theta)^2$$

- On peut penser qu'après mise à jour d'un certain nombre d'unités dans le répertoire, la valeur du premier terme ne sera pas très différente → agir sur le deuxième terme

3. Approche économique du problème

- Utiliser un modèle de coût :
  - C1 pour chaque mise à jour unitaire ; on procède à m mises à jour
  - C2 pour chaque unité enquêtée (n=taille échantillon)

$$C = mC1 + nC2$$

- ? Déterminer une fonction de gain du type :

$$gain = f(m) + \frac{V}{n}$$

- La difficulté est l'estimation de la fonction f(m)
- Si on arrive à modéliser f(m), utiliser la technique du multiplicateur de Lagrange pour déterminer le couple optimal m,n

4 - Conclusions

- Premiers éléments, mais travail à affiner :
  - Pousser un peu plus loin l'analyse (par exemple sur les types de sondage utilisés) …
  - Prendre en compte d'autres types de problèmes, en particulier les *stratum jumpers* (ce qui aura cette fois un impact sur la variance d'échantillonnage), ou des probabilités de cessation
  - Quantifier les différents phases décrites précédemment
- Encore plus difficile : la question des liens (pour définir des entreprises au sens de la LME)
- Ici, les questions de sous-couverture ne sont pas prises en compte

Questions ?

Contact
M. Philippe Brion
Courriel : philippe.brion@insee.fr

Insee
18 bd Adolphe-Pinard
75675 Paris Cedex 14
www.insee.fr
Informations statistiques :
www.insee.fr / Contacter l'Insee
09 72 72 4000
(coût d'un appel local)
du lundi au vendredi de 9h00 à 17h00