Séminaire en ligne du groupe enquêtes
Société Française de Statistique
27 juin 2024

LES ASPECTS MÉTHODOLOGIQUES VUS DU CÔTÉ DU RESPONSABLE DE PRODUCTION STATISTIQUE

Philippe BRION

Introduction

- Objectif de la présentation : développer le point de vue du responsable de production statistique, qui est confronté à différentes contraintes et à différents choix

- L’illustrer à partir de l’exemple des statistiques d’entreprises

1. Introduction (2)

- Partir de deux papiers anciens :
  - Platek & Särndal (2001) : Can a statistician deliver ?
  - Lyberg (2012) : La qualité des enquêtes

- Du point de vue des utilisateurs : questions de délais, comparabilité, coûts (y compris ceux pesant sur les enquêtés) plus mises en avant que le sujet de l'erreur d'échantillonnage stricto sensu

- Sujets qui renvoient peu à des questions de formalisation mathématique

- Le paradigme de l'erreur totale : constat d'échec ?

1. Introduction (3)

- Le responsable de production :
  - Coincé entre ces différentes contraintes (avec en plus, de manière récente, la mise en place d'indicateurs de performance)
  - S'assure du bon fonctionnement du processus de production (qui comprend une part importante d'expertise manuelle menée par des équipes de gestionnaires)
  - Le respect des échéances fixées pour les statistiques nécessite une succession de « micro-décisions », en particulier du point de vue du travail d'expertise manuelle

1. Introduction (4)

- Suite de la présentation : illustration à partir du cas des statistiques d'entreprises (dont la population a une distribution très asymétrique)
  - Passage en revue de différents aspects méthodologiques
  - Retour sur la question de l'erreur totale
- Deux remarques préliminaires :
  - Point de vue personnel (même si marqué par mon passage à l'Insee)
  - La présentation n'aborde pas la question des unités statistiques (voir par exemple Haag, 2019)

Partie 2 : Aspects méthodologiques « classiques »

2. Aspects méthodologiques classiques : échantillonnage

- utilisation de plans de sondages stratifiés à un degré

- une question fondamentale pour la statistique d’entreprises : la répartition de la charge statistique –> mise au point de méthodes de coordination d’échantillons (voir par exemple Guggemos, Sautory, 2012)

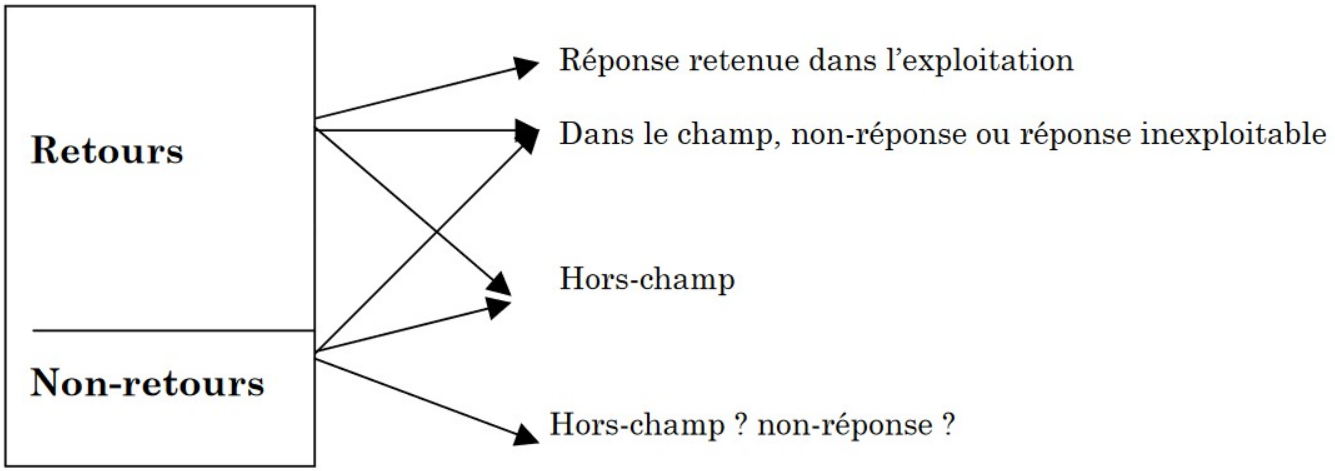
2. Aspects méthodologiques classiques : traitement des non-réponses (1)

- outils classiques de traitement de la non réponse, mais difficulté : distinguer, parmi les unités pour lesquelles il n'y a pas de retour, celles qui sont cessées et celles qui sont actives et non répondantes

- Pour affiner le traitement des non réponses, mobiliser des sources externes (par exemple déclarations TVA)

2. Aspects méthodologiques classiques : traitement des non-réponses (2)

Schéma tiré de Brion, Caron, Pietri-Bessy (2005)



2. Aspects méthodologiques classiques : estimateurs (1)

- Panoplie « standard » des estimateurs

- Mais une difficulté supplémentaire peut venir de la production de statistiques à partir de sources multiples (voir diapo suivante)

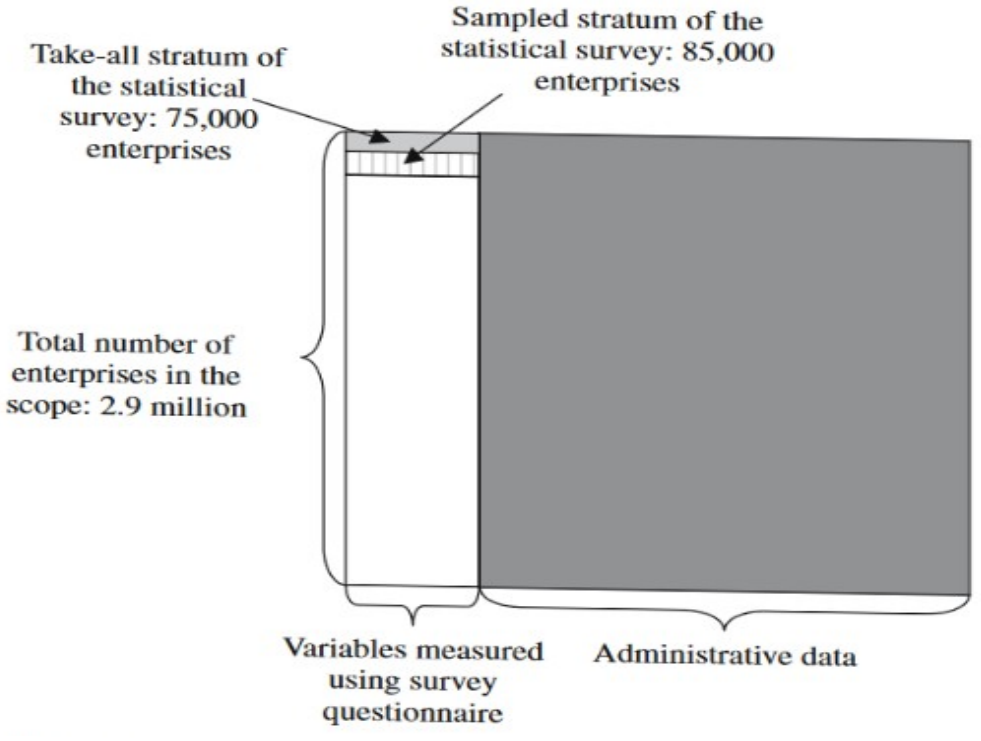
  - exemple : les statistiques qu'on cherche à produire résultent parfois d'un produit d'une variable qualitative et d'une variable quantitative ; exemple d'un chiffre d'affaires d'un secteur :

$$CA(secteur) = \sum 1(entreprise \in secteur) * chiffre\ d'affaires(entreprise)$$

2. Aspects méthodologiques classiques : estimateurs (2)

Un exemple d'estimation complexe, tenant compte de la diversité des sources : estimateurs composites (Brion, Gros, 2015)

- mise en place d'une procédure d'« accrochage » des données individuelles entre sources



2. Aspects méthodologiques classiques : estimateurs (3)

- Traitement des unités influentes (voir par exemple Favre-Martinoz, Deroyon (2017))

Partie 3 : Aspects liés au travail d’expertise manuelle

3. La relance des non répondants

- En raison de la distribution très asymétrique des entreprises, la non réponse n'a pas le même impact selon les unités

- Mettre en place des procédures de suivi des unités à relancer de manière individualisée s'appuyant sur des calculs d'agrégats en cours d'enquête (Buisson, 2009), les autres unités étant relancées de manière « industrielle »

3. Le data editing (1)

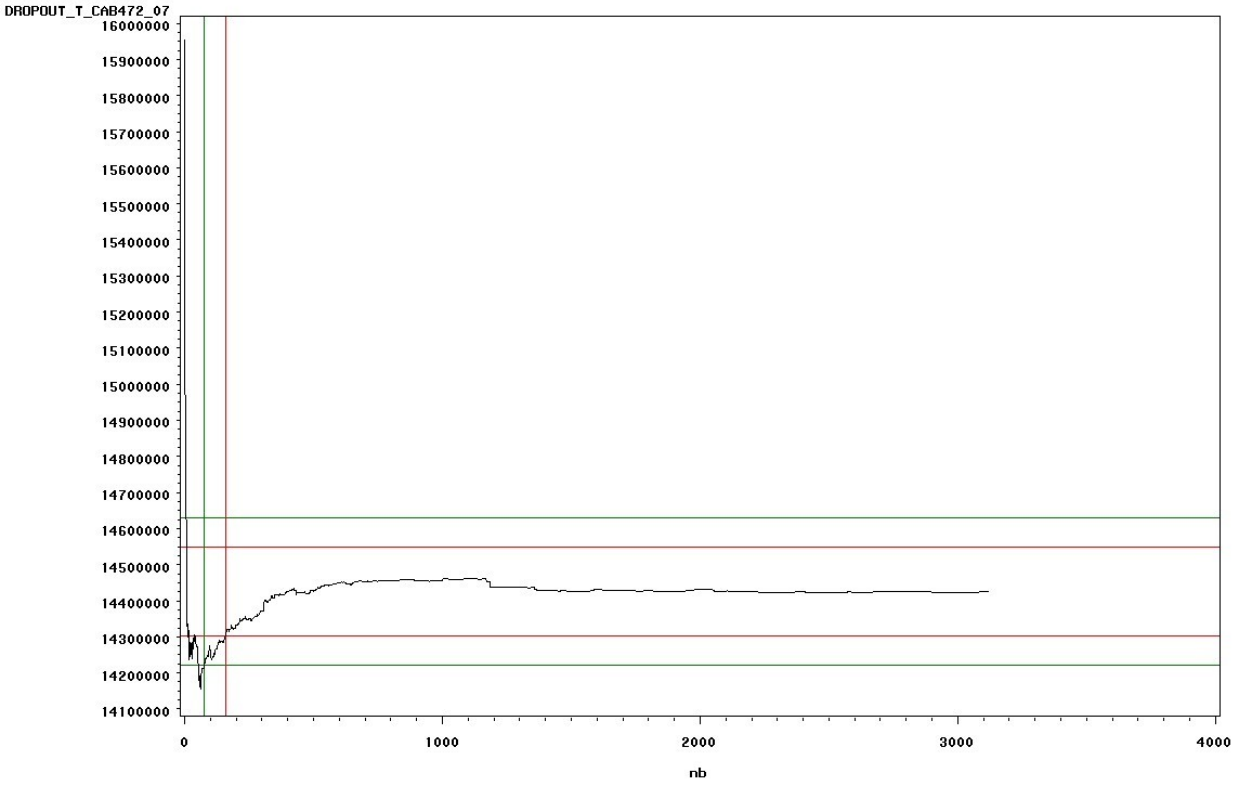
- Une grande partie des questionnaires « déclenchent » au moins un contrôle dans les chaînes mises en place

- Le travail de contrôle redressement des données consomme une part importante des ressources dédiées à la production des statistiques d'entreprises

- Une évolution dans la manière de mener le travail de vérification des données qui a conduit à se poser la question de « l'acharnement » et à proposer des méthodes adaptées (voir par exemple Granquist Kovar (1997) : *Editing of survey data : how much is enough ?,* ou De Waal, Pannekoek et Scholtus (2011*))*

3. Le data editing (2)

Illustration sur l'enquête annuelle d'entreprises (Insee, 2007) : estimation du chiffre d'affaires de la branche « commerce de détail alimentaire », en « intégrant » progressivement les corrections faites sur les données



3. Le data editing (3)

- L’expertise manuelle est donc rarement appliquée sur l'ensemble des données, et l'idée est alors de séparer les « questionnaires » en deux paquets :
  - un qui est examiné de manière manuelle par des gestionnaires
  - un pour lequel on utilise les données brutes fournies par les répondants, ou pour lequel on utilise une méthode de redressement automatique

3. Le data editing (4)

- La sélection des unités à vérifier manuellement se fait souvent via un score, calculé à partir de formules du type $w_i\left(x_i - \hat{x}_i\right)$

où : x_i *est la valeur « brute » de la variable X pour l'unité i*

et $\hat{x}_i$ *est un « prédicteur » de la vraie valeur X pour i*

- la formule ci-dessus donne un score « local » pour une variable ; un score global est ensuite calculé en agrégeant les scores locaux

3. Le data editing (5)

- Approche pragmatique, et difficultés pour les « réglages » (voir par exemple Gros (2012)):
  - Attention à stabilité des scores calculés (au fur et à mesure que les questionnaires rentrent)
  - Étude de l'ordre de grandeur de l'erreur d'échantillonnage pour « contenir » l'erreur d'observation par rapport à celle-ci
  - Qualité du prédicteur ?
  - Variables avec beaucoup de réponses à zéro
- Peu de papiers théoriques sur le sujet (voir par exemple Hesse (2005)) ; sujet peu ou pas enseigné dans les formations académiques

Partie 4 : Le paradigme de l’erreur totale

4. Retour sur la question posée sur l'erreur totale

- Creating a functional total survey error (TSE) model : an impossible dream ? (Platek, Särndal)

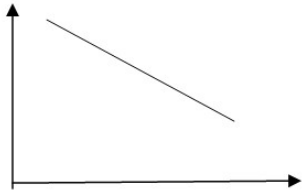
- De premiers éléments de réponse :
  - Un papier comme celui de Linacre et Trewin (1992) donne des éléments sur la prise en compte de différents types d'erreur :
    - Il compare différentes combinaisons d'options (couverture, type de collecte, suivi des non réponses) en mettant en regard des éléments de coût
  - Quand on procède au réglage des seuils utilisés pour la sélection des unités à vérifier de façon manuelle, on s'inscrit d'une certain façon dans une démarche TSE

4. Modéliser l’erreur totale, ou au moins une partie ?

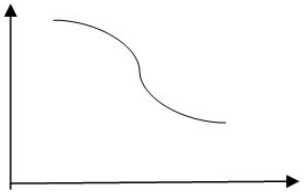
- Partir d’un cadre formalisé (voir par exemple Biemer, 2010), et écrire l’erreur totale d’une statistique produite sous forme de fonction de différents paramètres de coût ?
    - Nombre de questionnaires
    - Type de collecte utilisé
    - Mise à jour de la base de sondage
    - Temps passé à relancer les non répondants, et à contrôler les données
- Chercher un optimum concernant l’erreur totale sous contrainte de coût ? Attention à la robustesse des résultats obtenus relativement aux choix concernant les coûts unitaires …
- Le plus difficile : essayer de formaliser le comportement de l’erreur d’observation

4. Retour sur la question des délais de publication

- sujet sur lequel la pression est de plus en plus forte : être capable d'exhiber une courbe donnant en abscisse le temps, et en ordonnée l'erreur totale ? :



ou



???

4. La mise à jour des répertoires

- Les questions liées à l'imperfection des bases de sondage (Sautory, 2015)

- Leur mise à jour nécessite des travaux qui, pour certains, peuvent être réalisés de manière « industrielle », et qui, pour d'autres, demandent une expertise manuelle

- L'importance de certaines variables catégorielles

- Inscrire ces travaux dans une approche d'économie globale d'un dispositif de production ?

Conclusion

- Un ensemble de sujets méthodologiques sur lesquels des avancées sont possibles … et dont certains souffrent d'un manque de visibilité chez les méthodologues ?

- Ne pas oublier la composante humaine du dispositif de production statistique, avec un ensemble de tâches réalisées de façon manuelle

Bibliographie

- Biemer, P. (2010), *Total survey error – design, implementation and evaluation*, Public Opinion Quaterly, Vol 74, n°5
- Brion, Ph., Caron, N., Pietri-Bessy, P. (2005), *Redresser la non-réponse totale dans les enquêtes entreprises : les pièges à éviter – illustration avec l'enquête innovation*, Journées de méthodologie statistique, Insee
- Brion, Ph., Gros, E. (2105), *Statistical Estimators Using Jointly Administrative and Survey Data to Produce French Structural Business Statistics*, Journal of Official Statistics Vol 31, n°4
- Buisson, B. (2009), *Estimateurs en cours d'enquête et priorités de relance*, Journées de méthodologie statistique, Insee
- De Waal, T., Pannekoek, J., Scholtus, S. (2011). *Handbook of statistical data editing and imputation*, J. Wiley

Bibliographie

- Favre-Martinoz, C., Deroyon, T. (2017), *Traitement des valeurs influentes dans les enquêtes*, Fiche méthodologique, site web Insee
- Granquist, L., Kovar, J. (1997), *Editing of survey data : how much is enough ?* in Survey Measurement and Process Quality, John Wiley
- Gros, E. (2012), *Assessment and improvement of the selective editing process in Esane (French SBS)*, UNECE work session on statistical data editing, Oslo
- Guggemos, F., Sautory, O. (2012), *Sampling coordination of business surveys conducted by Insee*, communication à la conférence ICESIV, Montréal
- Haag, O. (2019), *Le profilage à l'Insee – une identification plus pertinente des acteurs économiques*, Courrier des Statistiques n°2, Insee

Bibliographie

- Hesse, C. (2005), Vérification sélective des données quantitatives, document de travail E2005/04, Insee
- Linacre, S., Trewin, D. (1993), *Total survey design : application to a collection of the construction industry*, Journal of Official Statistics, vol 9, n°3
- Lyberg, L. (2012), *La qualité des enquêtes*, Techniques d'enquête, vol 38, n°2
- Platek, R., Särndal, C.-E. (2001), *Can a statistician deliver ?*, Journal of Official Statistics, vol 17, n°1
- Sautory, O. (2015), *Les enjeux méthodologiques liés à l'usage de bases de sondage imparfaites*, Journées de méthodologie statistique, Insee