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

LES PARADONNÉES DE LA CONCEPTION À LA RÉALISATION

Fritz Scheuren[1]

Le texte qui suit est une version annotée de la présentation en PowerPoint de la communication que j'ai faite au Symposium de méthodologie 2005. Même annotées, les diapositives ne donnent à plusieurs égards qu'un compte rendu incomplet de l'expérience. Font notamment défaut les questions et réponses qui ont suivi l'exposé. Certaines de celles-ci, comme les commentaires de David Binder, ont été clarifiantes et des corrections ont, naturellement, été apportées en vue d'éclaircir certains points. Les nombreuses autres remarques intéressantes qui ont été faites n'ont pu être traitées de cette manière. Je me contenterai de dire que la présentation doit être complétée par les expériences de ceux qui sont aujourd'hui profondément engagés dans le processus, parmi lesquels plusieurs ont pris la parole plus tard durant la conférence ou à la conférence complémentaire du comité fédéral américain sur la méthodologie statistique (www.FCSM.gov) tenue à Washington. Aucun logiciel n'a été recommandé durant l'exposé, mais plusieurs autres conférenciers avaient utilisé le système Nesstar. Ce dernier pourrait être un outil utile pour les débutants qui souhaiteraient appliquer les idées exposées ici à leurs propres enquêtes. Une référence plus complète à Nesstar est incluse à la fin de l'article.

Le mot « paradonnées » est peut-être encore peu familier pour certains d'entre vous, mais les idées sous-jacentes devraient être bien connues de tous. Ce mot a été inventé par Mick Couper il y a quelques années (p. ex., Lyberg et Couper, 2005).

Les paradonnées sont des données au sujet du processus d'enquête. Il peut s'agir de simples données sommaires, comme

les taux de refus;
les taux de non-prise de contact;
les taux de non-réponse partielle;

voire même de la fraction des données qui a dû être obtenue en recourant à l'interview par procuration.

Les paradonnées peuvent être employées au niveau des macrodonnées, ou niveau sommaire, comme ci-dessus, ou au niveau des microdonnées, ou niveau individuel. Les éléments de microdonnées qui pourraient figurer dans chaque enregistrement de données du fichier d'exploitation pourraient être, par exemple,

la langue de l'interview;
les personnes présentes durant l'interview;
le nombre de tentatives d'appel avant d'établir un contact;
le refus initial ou non de participer à l'interview.

Aujourd'hui, on enregistre généralement la durée d'une interview sur place. Dans le cas de l'IPAO ou de l'ITAO, il est même possible de consigner le temps nécessaire pour répondre à chaque question et d'avoir accès à ces renseignements, au moins dans un fichier interne de l'enquête.

Bref, les paradonnées font partie des métadonnées informatisées qui entourent de nombreuses enquêtes à grande échelle gouvernementales et autres, comme l'Enquête sur la population active réalisée mensuellement au Canada.

Pourquoi parler des paradonnées maintenant, surtout si elles sont déjà si répandues? Et bien, peut-être, précisément pour cette raison. Il est certain que, même si elles sont très répandues et de grande portée, les paradonnées sont

[1] Fritz Scheuren, NORC Université de Chicago

rarement utilisées à leur plein potentiel, que ce soit durant les opérations d'enquête ou, plus fréquemment, à l'étape de l'inférence. Pour certaines raisons particulières de rappeler à tout le monde l'existence des paradonnées, voir la figure 1.

Figure 1
Pourquoi parler des paradonnées maintenant?

- Accroissement de l'informatisation.
- Effet sur l'accès aux données.
- Évolution vers une concentration plus grande sur les clients.
- Besoin croissant de quantifier tous les effets de l'enquête sur l'inférence.
- Abandon croissant du clivage échantillonnage–non-échantillonnage (Scheuren 2005b).

Les changements imputables en partie au coût décroissant et à la vitesse croissante du traitement de l'information ont transformé le paradigme de l'exécution d'une enquête qui est devenue un art nettement plus partagé. Les clients effectuent aujourd'hui une part beaucoup plus importante du travail qui était par le passé le domaine exclusif des producteurs de données.

Les paradonnées sont utilisées, ou pourraient l'être, à presque toutes les étapes d'une enquête. La figure 2 en énumère quelques-unes.

Figure 2
Utilisations des paradonnées à diverses étapes d'une enquête

- Gestion de l'enquête.
- Planification des améliorations.
- Dans les profils de qualité.
- En tant qu'éléments des métadonnées.
- De plus en plus, dans l'inférence.

Nous allons maintenant examiner chacune de ces idées l'une après l'autre. Il s'agit essentiellement d'un exposé historique assorti d'exemples tirés principalement de ma propre pratique.

Figure 3
Gestion et planification d'une enquête

- Traditionnellement, au sein du département.
- Principalement, planification fonctionnelle.
- De plus en plus aujourd'hui, exercice global.
- Quantification plus fréquente des coûts de la qualité.
- Partage des compromis entre temps et coût.

La figure 3 qui précède décrit la façon dont les travaux sur les paradonnées ont évolué pour passer d'une activité effectuée simplement au sein d'un département ou d'une fonction d'enquête à une activité qui est de plus en plus accomplie de façon unifiée.

Cette évolution est due en partie au fait que nombre d'enquêtes sont devenues plus coûteuses, à cause de la baisse des taux de réponse et de la hausse du coût de la main-d'œuvre. Cette situation a obligé les producteurs de données à analyser leurs méthodes d'un œil plus critique.

L'un des nouveaux outils les plus efficaces élaboré ces dernières décennies est ce que l'on appelle le « profil de qualité ». Dans le présent exposé, un profil de la qualité s'entend d'un rapport qui fait l'historique d'une enquête du début à la fin, principalement à l'aide de macroparadonnées sommaires.

La figure 4 énumère les raisons pour lesquelles les profils de la qualité sont partagés avec les clients et d'autres producteurs de données, aussi bien pour diverses enquêtes que pour une même enquête au cours du temps. Un assez grand nombre de ces profils ont été établis, à commencer par ledit « profil d'erreur » de la Current Population Survey (CPS). Ce dernier a été créé dans le cadre des travaux du comité fédéral américain sur la méthodologie statistique (www.FCSM,gov). Les principaux auteurs étaient Barbara Bailar et Camilla Brooks.

Figure 4
Partage des profils de qualité

- A débuté sous forme de « profil d'erreur ».
- Résumait les objectifs de l'enquête.
- Était fondé sur des paradonnées agrégées.
- Plus grande responsabilisation du producteur.
- Avec les clients et d'autres producteurs.
- Particulièrement au cours du temps (méta-analyse).

Puisque j'ai participé à cet effort, je peux dire avec une certaine assurance, quoique la mémoire joue parfois des tours, que nous n'avions pas envisagé le profil de la qualité comme un outil jouant un rôle dans les méta-analyses. Nous percevions sa valeur en tant qu'instrument d'une plus grande responsabilisation du producteur et de la croissance du rôle joué par les clients dans l'analyse des données d'enquête.

À l'heure actuelle, nous n'avons encore effectué aucune tentative complète de méta-analyse des données d'enquête axée sur le processus. Par contre, les efforts partiels abondent. Nous ne pensions pas non plus, à l'époque, à construire systématiquement un fichier informatique combiné de toutes les paradonnées d'une enquête. Pourtant, nous sommes capables de faire ces deux choses aujourd'hui. Et nous devrions!

À mon avis, nous ne profitons pas assez rapidement des enseignements que nous pouvons tirer de notre pratique. La croissance de l'informatisation et des idées statistiques mathématiques supplante de loin l'emploi des nouveaux outils susmentionnés. Cette situation était peut-être inévitable, mais un meilleur usage, plus systématique, des paradonnées pourrait nous aider à cet égard; et je le recommande vivement.

Néanmoins, nous avons réalisé certains progrès, comme le savent la plupart d'entre vous. La chronologie de la création de métadonnées informatisées résumée à la figure 5 nous permettra de le constater.

Figure 5
Intégration dans les métadonnées
Chronologie

- Production de clichés d'enregistrement uniquement.
- Fourniture des tables de codage.
- Tables de fréquence des questions.
- Codes (p. ex. items manquants).
- Profils de qualité agrégés.
- Percée de la pensée systémique.

Quand ma carrière de statisticien a débuté, les producteurs de données ne conservaient généralement que des copies électroniques de fichiers, appelées topogrammes de bande. À l'époque, le transfert de ces topogrammes d'un support papier dans le fichier de données proprement dit était considéré comme un progrès (et cela l'était). Un grand nombre

d'autres éléments de documentation fonctionnels ont bien sûr été créés, mais ils n'étaient généralement pas regroupés systématiquement. Il fallait presque faire partie de la culture du producteur de données pour vraiment apprécier leur profondeur (Scheuren 2005a).

Durant les années 60 et 70, les tables de codage informatisées et les fréquences question par question (univariées) étaient encore des nouveautés. La création de fichiers à grande diffusion (p. ex., Mulrow et Scheuren, 2000) a facilité ce changement, mais nous n'avons toujours pas, en tant que profession, regroupé typiquement toutes les étapes d'une enquête en une seule base de données.

Deming nous a prescrit de recourir à la « pensée systémique ». Mais nous n'avons pas suivi son conseil. Et cela, alors que nous sommes capables de le faire depuis longtemps déjà.

L'un de ses 14 points ou principes de la qualité (Deming, 1986) consiste à abattre les barrières entre les départements. Pourtant, dans le cas des grandes enquêtes, il est rare que nous construisions une base de données combinée unique. Ainsi, l'information sur la base de sondage n'est pas toujours présentée. Les renseignements sur les intervieweurs, même simplement les codes (sans identificateur) qui relient les intervieweurs à leurs tâches, ne sont pas transmis et inclus dans le fichier final, etc.

Pourquoi agissons-nous ainsi en pratique? Je ne le sais pas vraiment, mais permettez-moi de conjecturer et d'offrir trois raisons :

> Les clients ne demandent pas grand-chose de ce que nous pourrions fournir.
>
> Les coûts sont toujours un facteur déterminant et la valeur de la création de ce genre de fichier n'a pas été prouvée.
>
> Ni les producteurs de données ni les clients n'ont conceptualisé entièrement les utilisations des paradonnées dans l'inférence.

Il existe des exceptions dignes d'être mentionnées, comme les codes pour les données imputées, mais de façon générale, nous nous trouvons dans ce que nous pourrions appeler une « impasse ».

Figure 6
Utilisations possibles des paradonnées durant l'inférence

- Descriptivement pour une meilleure interprétation qualitative.
- Comme dans les profils de qualité.
- Révolution cognitive (Groves).
- Prise en compte de l'effet du biais et de la variance.
- Utilisation quantitative au moment de l'inférence, même par les clients.

Nous créons rarement des fichiers combinés et, donc, n'avons ni appris comment le faire efficacement ni exploré complètement leur valeur à l'étape de l'inférence. La figure 6 qui précède énumère quelques-unes des possibilités qui s'offriraient si nous modifions nos pratiques.

Chronologiquement, quel type de données les clients ont-ils reçu? En général, comme l'illustre la figure 7, avant 1960 environ, principalement des agrégats.

Figure 7
Inférences types avant 1960 environ

$$\int f(Y,M,R,C,D)dMdRdCdD = g_1(Y)$$

Classiquement, cette fonction $g_1(Y)$ était simplement un vecteur ou une matrice de totaux, disons, sous forme tabulaire.

Le producteur de données corrigeait ou « désincorporait » des données (Y) ou « fixait » toutes les erreurs de réponse (R) qu'il pouvait, et « réglait » aussi les problèmes de données manquantes (M), ainsi que les questions de couverture (C). À partir des variables du plan d'échantillonnage (D), il obtenait des poids de sondage qu'il appliquait à l'élaboration de tables d'agrégats $g_1(Y)$ qui étaient alors publiées.

Or, l'implication évidente de cette approche classique était que les producteurs de données savaient mieux que personne ce qu'il fallait faire et que les clients ne pouvaient pas ajouter grand-chose de valable. Pratiquement dès le début, cette approche a été trop contraignante. Ainsi, des outils tels que la régression étaient nécessaires pour la résolution de certains problèmes et la nature itérative de ce genre d'application a obligé les producteurs de données à créer une certaine forme de produits de microdonnées.

La figure 8 représente cette étape de développement — un progrès important. En fait, le produit de microdonnées $g_2(Y,D)$ a donné le jour à une littérature abondante sur la façon d'analyser les données d'enquête complexe, un grand pas en avant pour notre profession.

Figure 8
Inférences types avant 1960 environ

$$\int f(Y,M,R,C,D)dMdRdC = g_2(Y,D)$$

La fonction $g_2(Y,D)$ est un produit de microdonnées beaucoup plus riche qui commence à déplacer le fardeau de l'inférence.

Naturellement, cette formulation du problème d'inférence continue de laisser une grande partie du travail réalisé au moyen des paradonnées dans les mains du producteur de données, comme l'illustre la figure 9.

Soulignons que, comme cela est habituellement le cas, pour pouvoir faire l'objet d'une grande diffusion, les microdonnées doivent être sous forme désidentifiée. Autrement dit, il pourrait être nécessaire de sacrifier la fourniture de renseignements complets sur le plan de sondage (D), ce qui oblige à un compromis en ce qui concerne l'estimation directe des variances d'échantillonnage.

Figure 9
Séparation initiale par rapport à l'inférence

- Hypothèse implicite que la vérification des données et la résolution d'autres erreurs de mesure sont effectuées par le producteur.
- Les données manquantes, après correction, sont ignorables.
- Seuls les poids de sondage sont nécessaires pour produire des estimations sans biais.
- Les variances peuvent être calculées en utilisant uniquement les variables « D » du plan de sondage.
- Souvent les variables « D » sont réduites à des cas sans identification.

Les paradonnées n'avaient que peu de place, voire aucune, à cet étape précoce de production de microdonnées à grande diffusion. Toutefois, de plus en plus de clients souhaitaient obtenir des données sur les mesures qui avaient été prises pour procéder à l'imputation pour les données manquant à cause de la non-réponse partielle.

Ici, nous commençons à observer un début d'utilisation des paradonnées. L'ajout de « drapeaux » aux microdonnées pour signaler les données manquant à une question peut, comme je l'ai caractérisé à la figure 10, être considérée comme une révolution. Les travaux fondamentaux de Rubin (1987) sur l'imputation multiple représentaient l'une, mais une seule, des utilisations de cette approche.

Figure 10
Inférences types après la révolution concernant les données manquantes

$$\int f(Y,M,R,C,D)dRdC = g_3(Y,M,D)$$

Le produit $g_3(Y,M,D)$ établit le lien avec l'introduction de la notion de données manquantes dans l'inférence à partir de données d'enquête.

Comme l'illustre la figure 10, il ne peut y avoir, au plus, que deux autres étapes dans le transfert ou le partage de la tâche d'inférence entre le producteur de données et le client. L'une de celles-ci, décrite à la figure 11 qui suit, correspond au déplacement des préoccupations concernant les réponses (R) existantes, ou du moins une partie de celles-ci, dans le premier membre de l'équation. Voir Scheuren (à paraître). Bob Groves compte parmi ceux qui soutiennent que ce changement est nécessaire.

Jusqu'à présent, les questions de couverture et la non-réponse totale continuent d'être traitées par les producteurs de données, mais aucune raison ne les empêche de partager cette tâche avec les clients également.

Nous ne sommes probablement pas encore tout à fait prêts à faire ce dernier pas, et il n'est pas nécessaire que nous le fassions maintenant. Nous avons assez de chats à fouetter ailleurs. Néanmoins, nous devons commencer à produire une meilleure documentation sur nos enquêtes afin de pouvoir procéder plus tard à des méta-analyses.

Figure 11
Nouvelles inférences avec utilisation complète de paradonnées

$$\int f(Y,M,R,C,D)dC = g_4(Y,M,R,D)$$

$g_4(Y,M,D,R)$ reflète le début de l'utilisation complète de paradonnées pour l'inférence, comme il est préconisé dans le présent article.

Trois exemples précoces, inspirés de ma propre pratique, méritent d'être mentionnés. Ils ont une portée de plusieurs années et présentent de nombreux éléments communs dont je discuterai à la fin.

En premier lieu, il y a eu les fichiers de la Survey of Economic Opportunity de 1966-1967 (achevés en 1971). Ils ont été produits au moment où je terminais mes études supérieures. Ils se trouvent maintenant dans les archives nationales des États-Unis. Ils sont aussi vieux que cela. Puis, il y a eu les fichiers d'appariement exact des données des CPS-IRS-SSA de 1973. Bien que la production de ces fichiers se soit fondamentalement terminée en 1980, ils sont mis à jour périodiquement et sont encore utilisés par l'administration de la sécurité sociale des États-Unis. Enfin, il y a eu les fichiers de 1997 et de 1999 de la National Survey of America's Families (achevés en 2000), qui sont disponibles à l'Urban Institute.

Certains attributs communs de ces fichiers sont énumérés à la figure 12 qui suit. Les fichiers CPS-IRS-SSA de 1973, contrairement aux autres, contenaient des données administratives en plus des résultats de l'enquête.

Figure 12
Certains attributs communs

- Tous contiennent des enregistrements pour les cas d'interviews et de non-interviews.
- Tous possèdent un code permettant de relier, mais non d'identifier, les interviews réalisées par un même intervieweur.
- Tous contiennent un codage permettant de repérer les groupes naturels connus d'après la base de sondage, comme la grappe ultime (mais non les UPE identifiables).
- Tous contiennent un codage étendu pour les données manquantes et un code pour signaler les imputations.

Néanmoins, certains éléments que l'on aurait souhaité inclure ne l'ont pas été. Des sondages auprès des intervieweurs ont été faits, mais les résultats n'ont jamais été ajoutés aux fichiers de microdonnées des enquêtes. L'ajout de ce genre de renseignements est controversé, mais a peut-être été effectué ailleurs. Je serais très heureux d'en prendre connaissance, le cas échéant.

L'utilisation des variables de paradonnées à l'étape de l'inférence n'était que modeste et presque inexistante chez les clients. La théorie ayant trait à l'utilisation de ces variables a été laissée à d'autres, mais commence à poindre maintenant (p. ex., Beaumont 2005).

Pour conclure, permettez-moi de me hasarder à décrire quelques-unes des prochaines étapes que nous devrions envisager en tant que profession :

> Opérations. Selon moi, un beaucoup plus grand nombre de choses pourraient être faites durant les opérations d'enquête si nous faisions simplement progresser le fichier d'enquête d'une étape à l'autre du processus, en l'étoffant au fur et à mesure, et en éliminant par conséquent partiellement les barrières entre les départements.
>
> Inférence. La remarque de Deming concernant la « pensée systémique » doit de toute évidence être étendue aux clients également. Nombre d'organismes statistiques fédéraux américains offrent des ateliers aux utilisateurs de leurs données. Favoriser l'utilisation de paradonnées par les clients pourrait contribuer puissamment à l'accélération de notre croissance en tant que profession et à l'amélioration de la qualité de notre pratique.
>
> Apprentissage. J'espère que les personnes présentes ici aujourd'hui utiliseront beaucoup plus fréquemment les paradonnées dans l'avenir. En fait, je sais qu'un grand nombre d'entre vous se penchent sur cette question à l'heure actuelle. Naturellement, d'autres statisticiens que ceux de Statistique Canada peuvent s'amuser aussi. J'aimerais vraiment jouer un rôle, mais, quoiqu'il en soit, je vous présente mes meilleurs vœux à tous.

POSTFACE

Dans la documentation remise à la conférence, j'ai aussi fournit un pastiche de la qualité (disponible séparément dans le numéro d'AMSTAT NEWS de septembre 2005) qui jouait sur le mot « paradonnées ». Je demande ici à Mick Couper d'être indulgent, mais ce pastiche a peut-être fait ressortir certains points intéressants qui sont couverts dans le présent article plus traditionnel.

RÉFÉRENCES

Biemer, P. et L. Lyberg (2003), *Introduction to Survey Quality*. New York: Wiley. Pour un exemple complètement opérationnel d'une qualité d'enquête supérieure, voir les travaux d'Arthur Kennickell et de ses collègues concernant la Survey of Consumer Finances, qui peuvent être consultés à http://www.frb.gov/.

Beaumont, Jean-François (2005), "L'utilisation de renseignements sur le processus de collecte de données pour traiter la non-réponse totale au moyen de l'ajustement de poids", *Techniques d'enquête*, 31, pp.249-254.

Couper, M. et L. Lyberg (2005), "The Use of Paradata in Survey Research", International Statistical Institute Meetings, Avril 2005, Sydney.

Deming, W. E. (1986), *Out of the Crisis*. Cambridge, MA: Massachusetts Institute of Technology, Center for Advanced Engineering Study.

Dippo, C. et Sundgren (2000), "The role of metadata in statistics". Voir aussi Colledge, M. and Boyko, E. (2000), "Collection and classification of statistical metadata; the real world of implementation". Articles présentés à la deuxième conférence internationale sur les enquêtes auprès des établissements, Buffalo, États-Unis.

Ishikawa, K. (1990), *Introduction to Quality Control*. Tokyo: 3A-Corporation.

Jabine, T. (1994), *Quality Profile for SASS: Aspects of the Quality of Data in the Schools and Staffing Surveys (SASS).* National Center for Education Statistics.

Juran, J. M. (1988), *Juran on planning for quality*. New York: Free Press.

Kalton, G., Winglee, M., Krawchuk, S., et Levine, D. (2000), *Quality Profile for SASS Rounds 1-3: 1987-1995, Aspects of the Quality of Data in the Schools and Staffing Surveys (SASS).* National Center for Education Statistics.

Mulrow, J. et Scheuren, F. (2000), "A Confidentiality Fable", *JSM Proceedings, 2000*.

Nesstar Publisher, un outil de traitement des métadonnées, qui peut être utilisé pour publier des données et la documentation qui les accompagne dans un catalogue sur un serveur Nesstar. De là, ces ressources peuvent être mises à la disposition de la collectivité plus générale par la voie de Nesstar WebView.

Rubin, D. (1987), *Multiple Imputation*, Wiley, New York.

Scheuren, F. (2005a), "Reminisce on Multiple Imputation", *The American Statistician.*

Scheuren, F. (2005b), "Seven Rules of Thumb for Nonsampling Error in Surveys", National Institute of Statistical Science (NISS) Total Survey Error Conference, Washington, Mars 2005.

Scheuren, F., "Macro and micro paradata for quality assessment in surveys", rapport non publié, *Journal of Official Statistics.*