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

ESTIMATION DU SOUS-DÉNOMBREMENT DES BLESSURES CAUSÉES PAR DES VÉHICULES EN NOUVELLE-ZÉLANDE : UNE APPROCHE FONDÉE SUR L'INTÉGRATION PROBABILISTE DES DONNÉES ET SUR LA CAPTURE-RECAPTURE

Ricardo Enrico C. Namay II[1]

RÉSUMÉ

Le programme de statistiques sur les blessures (*Injury Statistics Project*) de Statistics New Zealand relie par couplage trois bases de données afin de produire un tableau exhaustif des blessures qui surviennent en Nouvelle-Zélande. Ces bases de données sont fournies par les organismes suivants : Accident Compensation Commission (ACC), New Zealand Health Information Service (NZHIS) et Land Transport New Zealand (LTNZ).

Au moyen du couplage probabiliste des données, on obtient une base de données exhaustives sur les blessures. Pour évaluer la qualité des couplages réalisés, on calcule des taux de faux positifs et de faux négatifs. Toutefois, ces taux n'indiquent pas si les bases de données utilisées pour le couplage ont sous-dénombré les blessures (biais) et ne fournissent pas de marges d'erreur concernant l'incidence produite.

En nous inspirant de la notion de capture-recapture, nous envisageons une méthode d'estimation des sous-dénombrements et des marges d'erreur concernant l'incidence estimée des blessures causées par des véhicules.

MOTS CLÉS : Capture-recapture; taux de faux positifs; taux de faux négatifs.

1. INTRODUCTION

1.1 Méthodologie d'intégration des données

L'intégration des données consiste à coupler deux ou plusieurs sources de données, à raison de deux à la fois, pour repérer les enregistrements qui appartiennent à la même entité. On compare les enregistrements d'une base de données à ceux d'une autre en fonction de variables de comparaison sélectionnées, appelées variables d'appariement. On appelle *passage* chaque cycle de comparaison portant sur les variables d'appariement. Lorsqu'on compare des variables correspondantes et qu'on juge qu'elles concordent, on attribue à cette paire de variables un poids de concordance (valeur positive); autrement, on lui attribue un poids de discordance (valeur négative). On compare chaque paire correspondante des variables d'appariement sélectionnées. On calcule alors un poids composite pour la paire d'enregistrements comparée. Ce poids composite est simplement la somme de tous les poids de concordance et de discordance des variables d'appariement correspondantes de la paire d'enregistrements. Si l'on suppose que les variables d'appariement ne sont pas en corrélation très étroite, un poids composite positif indiquerait sans doute que la paire d'enregistrements appartient à la même entité (il s'agit d'un couplage; lorsque la paire d'enregistrements appartient **vraiment** à la même entité, on est en présence d'un appariement et non d'un couplage). Plus le poids composite est positif, plus il est probable que les enregistrements appartiennent à la même entité. Inversement, plus le poids composite est négatif pour une paire d'enregistrements, plus il est probable que les enregistrements de cette paire appartiennent à des entités différentes (il s'agit d'un non-couplage).

Dans la pratique, on ne tient pas pour acquis que le poids composite positif d'une paire d'enregistrements indique un couplage. On établit un point de coupure non négatif pour distinguer les couplages des non-couplages. Les paires d'enregistrements dont le poids composite se situe en deçà de ce point de coupure non négatif sont considérées

[1] Ricardo Enrico C. Namay II, Statistics New Zealand, Statistics House, Wellington, New Zealand, 6004 (ricardo.namay@stats.govt.nz)

comme des non-couplages. Il est donc possible qu'une paire d'enregistrements dont le poids composite est positif soit considérée comme un non-couplage si son poids composite se situe en deçà du point de coupure. Dans la pratique, une concordance partielle entre des variables (et, par conséquent, un poids de concordance partiel) est possible, si l'analyste en décide ainsi.

1.2 Méthodologie de capture-recapture

La capture-recapture est une méthode mise au point par des biologistes pour estimer une population faunique. Parmi les premières utilisations connues de cette méthode, on retient celle de Petersen, en 1896, pour estimer la population de poissons, et celle de Lincoln, dans les années 1930, pour estimer les populations d'oiseaux. En raison de ces travaux avant-gardistes, il n'est pas rare de voir, dans la documentation, l'estimateur de population désigné sous le nom d'indice de Petersen-Lincoln.

La capture-recapture vise à estimer la taille de population d'une espèce animale sauvage. À cette fin, on prélève un premier échantillon de l'espèce d'intérêt (capture) et on marque ces animaux, puis on les libère dans la nature. Après un certain temps, on prélève un deuxième échantillon de l'espèce (recapture) et, d'après la proportion d'animaux marqués par rapport aux animaux non marqués, on calcule une estimation de la population.

Supposons que le nombre d'animaux capturés dans le premier échantillon est n_A et que le nombre total d'animaux recapturés est n_B, dont n_{AB} sont des animaux marqués. L'estimateur naturel, $\hat{N}$, pour le total inconnu de population *N*, s'obtient facilement par le calcul suivant :

$$\hat{N} = \frac{n_A n_B}{n_{AB}} \tag{1}$$

où $\frac{n_{AB}}{n_B}$ est la probabilité de capture. Eberhardt (2003) observe que la distribution pertinente est hypergéométrique et en propose une explication intuitive.

Toutefois, comme il est possible qu'aucun animal marqué ne se trouve dans l'échantillon recapturé, l'estimateur a alors un dénominateur nul; il s'agit d'un « biais infini ».

Eberhardt cite l'ajustement proposé par Chapman (1951, 1954) pour contourner ce problème en proposant certains termes de correction. L'estimateur de Chapman ainsi obtenu (appelé aussi estimateur de Petersen-Lincoln-Chapman) est le suivant :

$$\hat{N} = \frac{(n_A + 1)(n_B + 1)}{n_{AB} + 1} - 1 \tag{2}$$

et la variance est donnée par

$$Var(\hat{N}) = \frac{(n_A + 1)(n_B + 1)(n_A - n_{AB})(n_B - n_{AB})}{(n_{AB} + 1)^2 (n_{AB} + 2)}. \tag{3}$$

En outre, si la taille inconnue de la population *N* est importante ($N > 1000$) et que la probabilité de recapture $\frac{n_{AB}}{n_B}$ est supérieure à 0,05, on peut alors utiliser la distribution normale en tant qu'approximation de la distribution hypergéométrique. Ce résultat est particulièrement utile pour construire un intervalle de confiance autour d'une estimation. Ainsi, pour un *N* important et une probabilité de recapture supérieure à 0,05, l'intervalle de confiance de 95% pour l'estimation $\hat{N}$ est donné par

$$CI = \hat{N} \pm 1{,}96 * \sqrt{Var(\hat{N})}. \tag{4}$$

2. SOURCES DES DONNÉES

2.1 Accident Compensation Corporation (ACC)

La Accident Compensation Corporation (ACC) est une société d'État qui administre le programme d'indemnisation des accidents de la Nouvelle-Zélande. Ce programme offre une assurance-accident aux citoyens, aux résidents et aux visiteurs temporaires de la Nouvelle-Zélande.

Dans le fichier source de l'ACC, l'unité déclarante est une demande d'indemnisation. Toutefois, en mettant au point les passages d'intégration de données, les responsables du projet ont constaté que les données de l'ACC contenaient des paires de demandes d'indemnisation en double, c'est-à-dire deux demandes d'indemnisation déposées pour la même blessure. L'objet de l'intégration consistait non seulement à compter le nombre total de blessures, mais aussi à ajouter, pour une blessure, de l'information provenant de chaque source administrative afin de dresser un portrait plus complet. Les doubles ont une incidence sur ces deux objectifs, car ils entraînent un léger surdénombrement des blessures. En outre, l'information concernant la même blessure risque d'être jointe à deux demandes d'indemnisation. Par exemple, les coûts d'une blessure seraient répartis entre deux demandes déposées en double. On s'est penché sur cette question et on a mis au point une méthodologie pour repérer les doubles et les retirer des données.

Les demandes d'indemnisation pour blessure qui sont refusées par l'ACC sont comprises dans la base de données à titre de blessures, car elles correspondent à la définition d'une blessure énoncée dans le document « *Injury Statistics Project Pilot: Definitions of Injury* » [10].

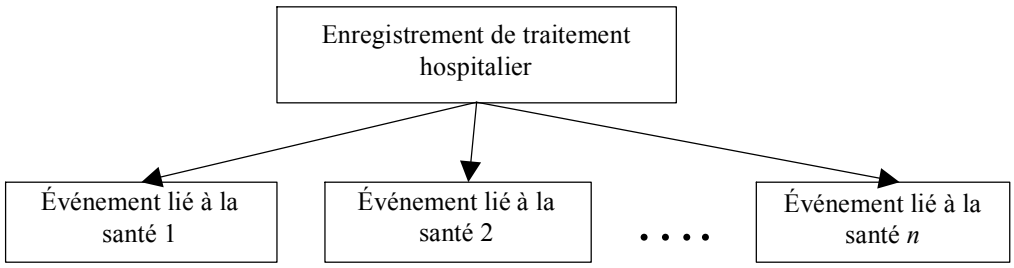
2.2 New Zealand Health Information Service (NZHIS)

Au sein du ministère de la Santé, le New Zealand Health Information Service (NZHIS) est un groupe chargé de recueillir et de diffuser les données relatives à la santé.

Les hôpitaux doivent fournir un enregistrement au NZHIS chaque fois qu'un malade ayant reçu des soins de santé financés par l'État sort de l'hôpital. Appelés « événements liés à la santé », ces enregistrements de sortie sont les unités déclarantes du fichier source du NZHIS.

Plusieurs événements liés à la santé peuvent être nécessaires au traitement d'une blessure donnée. Avant de procéder au couplage d'enregistrements entre les données du NZHIS et de l'ACC, on a mis au point un processus permettant de repérer les événements liés à la santé qui concernent la même blessure, de les regrouper et de créer un seul enregistrement pour cette blessure. On appelle « enregistrement de traitement hospitalier » un groupe d'événements liés à la santé concernant la même blessure.

Figure 1. Relation des enregistrements de traitement hospitalier avec les événements liés à la santé.



Bien qu'on ait consacré beaucoup de travail à l'élaboration des règles de groupement pour obtenir des résultats aussi exacts que possible, la qualité de certaines données sur les événements liés à la santé risque d'entraîner des erreurs dans les enregistrements de traitement hospitalier. Pour les besoins du présent exposé, et pour simplifier, on suppose qu'il n'y a pas d'erreurs dans les groupements d'événements liés à la santé.

2.3 Land Transport New Zealand (LTNZ)

Appelé Land Transport Safety Authority (LTSA) au début du projet, Land Transport New Zealand (LTNZ) est un organisme relevant du ministère des Transports qui recueille des données sur les accidents causés par des véhicules et les enregistre dans le système d'analyse des accidents (*crash analysis system* ou CAS).

Sur le lieu d'un accident causé par un véhicule, un policier remplit le rapport d'accident. Il inscrit sur le formulaire les noms des automobilistes ou des passagers blessés. Parfois, lors d'un accident grave, l'ambulance peut avoir quitté les lieux avec les victimes de l'accident avant l'arrivée du policier, ou encore, d'autres victimes de l'accident peuvent s'être dispersées; aussi le policier remplit-il le rapport à l'aide des renseignements fournis par des témoins. Si les personnes blessées ont quitté les lieux avant l'arrivée du policier, ce dernier est censé assurer un suivi pour recueillir les renseignements supplémentaires nécessaires à l'établissement du rapport.

Le LTNZ établit des identificateurs uniques de personne et d'accident qui lui permettent de mettre à jour ses enregistrements lorsqu'il obtient de nouveaux renseignements concernant l'accident.

3. LA SÉQUENCE D'INTÉGRATION DES DONNÉES ET LA MÉTHODOLOGIE DE CAPTURE-RECAPTURE

On a d'abord couplé les données de l'ACC et du NZHIS en effectuant diverses combinaisons de champs communs. On a utilisé des identificateurs uniques (numéros de demande d'indemnisation, numéros de formulaire, etc.) et des identificateurs personnels (nom, âge, etc.) pour cerner et coupler les enregistrements de ces bases de données. Après le couplage ACC-NZHIS, on a estimé des taux de faux positifs et de faux négatifs en suivant des règles administratives formulées pour définir ces taux.

L'ensemble d'entités saisies se compose d'enregistrements de blessures causées par des véhicules provenant de l'ACC et du NZHIS. Donc, les enregistrements d'intérêt pour le premier échantillon (capture) peuvent être un enregistrement ACC-NZHIS couplé, un enregistrement ACC non couplé ou un enregistrement NZHIS non couplé, pourvu que ces enregistrements concernent une blessure causée par un véhicule. Les enregistrements ACC et NZHIS comportent un indicateur de blessure causée par un véhicule, ce qui permet de filtrer les enregistrements pertinents à l'étude.

Dans le présent document, comme les taux de faux positifs et de faux négatifs pour les enregistrements ACC-NZHIS couplés ont été estimés sans égard à la valeur de l'indicateur de blessure causée par un véhicule (indiquant si l'accident a été causé ou non par un véhicule), l'estimation des sous-dénombrements dans la base de données du LTNZ suppose que ces taux sont homogènes entre le sous-échantillon des blessures causées par des véhicules et celui des blessures qui ne le sont pas. Le dénombrement, *M*, de blessures causées par des véhicules dans le premier échantillon (capture) est donc la somme :

$$n_A = ACC_{uv} + (1 - fn_{ACC-NZHIS}) * NZHIS_{uv} + (1 + fp_{ACC-NZHIS}) * (ACC \leftrightarrow NZHIS)_{lv} \quad (5)$$

où

ACC_{uv} = nombre d'enregistrements ACC non couplés de blessures causées par des véhicules

$NZHIS_{uv}$ = nombre d'enregistrements NZHIS non couplés de blessures causées par des véhicules

$(ACC \leftrightarrow NZHIS)_{lv}$ = nombre d'enregistrements ACC-NZHIS couplés de blessures causées par des véhicules

$fp_{ACC-NZHIS}$ = taux de faux positifs d'après le couplage ACC-NZHIS
= pourcentage d'enregistrements NZHIS *couplés* de traitement hospitalier incorrectement couplés à un enregistrement ACC

$fn_{ACC-NZHIS}$ = taux de faux négatifs d'après le couplage ACC-NZHIS
= pourcentage d'enregistrements NZHIS *non couplés* de traitement hospitalier qui auraient dû être couplés à un enregistrement ACC

Après le couplage ACC-NZHIS, on a couplé les enregistrements LTNZ aux enregistrements ACC et NZHIS intégrés. Encore une fois, on a estimé des taux de faux négatifs et de faux positifs. La taille des enregistrements LTNZ qui sont couplés aux enregistrements ACC et NZHIS de blessures causées par des véhicules correspondrait donc à un ensemble d'animaux marqués recapturés de taille n_{AB}, rajusté en fonction des taux de faux positifs et de faux négatifs d'après le couplage de l'ensemble LTNZ avec l'ensemble ACC/NZHIS. Si, pour simplifier, on suppose un taux constant de faux positifs pour les différents types d'enregistrement LTNZ couplés, alors

$$n_{AB} = [1 + fp_{ACC/NZHIS-LTNZ}] * [(ACC \leftrightarrow LTNZ) + (NZHIS \leftrightarrow LTNZ) + (ACC \leftrightarrow NZHIS \leftrightarrow LTNZ)] \qquad (6)$$

où

$(ACC \leftrightarrow LTNZ)$ = nombre d'enregistrements ACC et LTNZ couplés

$(NZHIS \leftrightarrow LTNZ)$ = nombre d'enregistrements NZHIS et LTNZ couplés

$(ACC \leftrightarrow NZHIS \leftrightarrow LTNZ)$ = nombre d'enregistrements ACC, NZHIS et LTNZ couplés

$fp_{ACC/NZHIS-LTNZ}$ = taux de faux positifs d'après le couplage ACC/NZHIS avec LTNZ
= pourcentage d'enregistrements LTNZ *couplés* incorrectement couplés à un enregistrement du fichier ACC/NZHIS

Le dénombrement des blessures causées par des véhicules dans le deuxième échantillon (recapture) est donc le nombre d'éléments dans le fichier LTNZ et est donné par

$$n_B = n_{AB} + [1 - fn_{ACC/NZHIS-LTNZ}] * LTNZ_u \,. \qquad (7)$$

Si l'on soupçonne que les taux de faux positifs ou de faux négatifs ne sont pas homogènes, la stratification à l'égard des divers types d'enregistrement peut améliorer les estimations.

4. CONCLUSION

Nous avons présenté ici un cadre de travail simple pour améliorer les estimations de populations d'intérêt en faisant appel à la notion de capture-recapture et au couplage probabiliste d'enregistrements. De plus, par le biais de la capture-recapture, nous avons présenté une façon simple de construire des intervalles de confiance pour des estimations provenant de données issues d'un couplage probabiliste.

RÉFÉRENCES

Chapman, D.G. (1951), "Some properties of the hypergeometric distribution with application to zoological censuses", *Proceedings of the second Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley: University of California Press, pp. 131-160.

Chapman, D.G. (1954), "The estimation of biological populations", *Annals of Mathematical Statistics*; 25, pp. 1–15.

Eberhardt, L.A. (2003), *Course in Quantitative Ecology*, National Marine Mammal Laboratory, Alaska Fisheries Science Center.

Gill, G.V., Ismail A.A. and Beeching, N.J. (2001), "The Use of Capture-recapture Techniques in Determining the Prevalence of Type 2 Diabetes", *Q J Med*; 94, pp. 341-346.

Gill, L. (2001), *Methods for Automatic Record Matching and Linking and their Use in National Statistics*, London: Office for National Statistics.

Nanan, D.J. and White, F. (1997). "Capture-recapture: Reconnaissance of a Demographic Technique in Epidemiology", *Chronic Diseases in Canada*, 18, pp. 144-148.

Rohatgi, V.K. (1976) *An Introduction to Probability Theory and Mathematical Statistics*, New York: Wiley.