No 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2006 : Enjeux méthodologiques reliés à la mesure de la santé des populations

2006

Statistics Canada Statistique Canada

Canada

Recueil du Symposium 2006 de Statistique Canada
Enjeux méthodologiques reliés à la mesure de la santé des populations

Création de fichiers de microdonnées à grande diffusion pour la National Survey on Drug Use and Health (NSDUH)

Feng Yu, Lanting Dai, Moshe Feder et James R. Chromy[1]

Résumé

Le processus de création de fichiers de microdonnées à grande diffusion compte un certain nombre de composantes. L'un de ses éléments clés est la méthode novatrice MASSC mise au point par RTI International. Cependant, ce processus comporte d'autres composantes importantes, comme le traitement des variables d'identification non essentielles et des résultats extrêmes en guise de protection supplémentaire. Le contrôle de la divulgation statistique a pour but de contrer l'intrusion interne ainsi qu'externe. Les composantes du processus sont conçues en conséquence.

MOTS CLÉS : contrôle de la divulgation statistique, fichiers de microdonnées à grande diffusion, MASSC.

1. Introduction

1.1 La National Survey on Drug Use and Health

Le National Survey on Drug Use and Health (NSDUH) est une enquête nationale transversale annuelle sur la consommation de drogues et les problèmes connexes réalisée par RTI International pour le compte de la Substance Abuse and Mental Health Services Administration (SAMHSA). La NSDUH fournit des renseignements sur la consommation de drogues illicites, d'alcool et de tabac par les membres de la population civile américaine de 12 ans et plus ne résidant pas en établissement. Elle donne aussi des mesures relatives aux problèmes de santé mentale, y compris des données sur la dépression et sur la coexistence de la consommation de substances psychoactives et de problèmes de santé mentale, ainsi que des renseignements sur les problèmes de santé et les questions liées aux soins de santé.

1.2 Craintes concernant le respect de la confidentialité et la divulgation

Le risque de divulgation est une préoccupation importante étant donné le caractère confidentiel des renseignements (consommation de drogues et information sur la santé) recueillis dans le cadre de la NSDUH. Afin de protéger les renseignements personnels confidentiels fournis par les répondants, de même que pour respecter les règlements fédéraux et l'engagement à veiller à la confidentialité des données, un traitement de protection contre la divulgation doit être appliqué aux fichiers de données avant qu'ils soient diffusés au public.

1.3 Scénarios de divulgation

Une intrusion a lieu quand une personne non autorisée (un « intrus ») essaie de relier un enregistrement du fichier de microdonnées à un répondant identifiable. L'intrus peut ou non avoir une cible particulière, il peut ou non savoir que sa cible est présente dans l'échantillon, et il peut tenter de déclarer avec certitude que certains renseignements confidentiels concernant la cible sont vrais ou simplement de prouver la véracité de sa déclaration avec une probabilité élevée. La combinaison particulière de ces conditions définit le scénario de divulgation. Les variables utilisées pour identifier l'enregistrement correspondant à la cible sont appelées variables d'identification (VI). Ces variables peuvent inclure l'âge, le sexe, la race, etc.

[1]RTI International, P.O.Box 12194, Research Triangle Park, NC 27709-2194, États-Unis.

Dans la discussion qui suit, nous nous intéressons principalement à la distinction entre l'intrusion interne et l'intrusion externe. Une intrusion interne a lieu lorsque l'intrus connaît sa cible dans l'enquête et essaie de découvrir les renseignements confidentiels en combinant l'information identifiante présente dans le fichier de données et ses connaissances personnelles. Cette forme d'intrusion est préoccupante dans le cas de la NSDUH, parce qu'une personne pourrait savoir qu'un membre de sa famille fait partie de l'échantillon. Les chances de succès de l'intrus interne augmentent quand la cible est unique dans l'échantillon. Voici un exemple d'intrusion interne : un père sait que son fils participe à l'enquête et, en utilisant un ensemble de variables d'identification, constate que le fichier ne contient qu'un seul enregistrement dont le profil correspond aux caractéristiques démographiques de ce dernier; il est donc certain que cet enregistrement correspond à son fils. En poussant plus loin l'examen des renseignements confidentiels que contient l'ensemble de données, il pourrait arriver à déceler le comportement de consommation de drogues de son fils.

Une intrusion externe a lieu lorsque l'intrus ne sait pas si sa cible est présente dans l'échantillon et qu'il essaie d'identifier un enregistrement en l'appariant à une base de données externe. L'identification peut se faire par appariement des VI des répondants qui figurent également dans l'ensemble de données externe. Un intrus externe cible habituellement un répondant dont le profil (combinaison de VI) dans la population est unique (ou rare). Ces cas sont généralement dits uniques dans la population (ou rares dans la population).

Alors qu'un intrus externe ne sait pas si les données contiennent un enregistrement représentant sa cible, il dispose habituellement de moyens plus complexes qu'un intrus interne et peut posséder une énorme quantité de données externes et des logiciels d'appariement. Il ne vise habituellement pas une personne précise, mais plutôt tout sujet qu'il peut identifier, ce qui pourrait éventuellement discréditer l'enquête.

1.4 Catégories de risque

Le risque de divulgation est plus élevé si le profil du sujet est rare. Donc, nous définissons la catégorie de risque d'un enregistrement en fonction de la combinaison de VI. Les enregistrements uniques sont ceux qui sont identifiés de façon unique par un ensemble donné de VI. Les enregistrements doubles sont les enregistrements dont le profil est en commun avec uniquement un autre enregistrement dans l'ensemble de données pour une combinaison donnée de VI; les enregistrements triples sont ceux dont le profil est en commun avec uniquement deux autres enregistrements dans l'ensemble de données pour une combinaison donnée de VI; les enregistrements autres sont ceux dont le profil est commun à au moins trois autres enregistrements de l'ensemble de données pour une combinaison donnée de VI. Afin d'assurer le respect de la confidentialité, chaque enregistrement de l'ensemble de données a la chance d'être traité en fonction de la catégorie de risque à laquelle il appartient.

1.5 Variables de nature délicates

Les fichiers de la NSDUH contiennent des renseignements confidentiels sur la consommation de substances psychotropes, la santé mentale et d'autres problèmes de santé. Certaines variables de consommation de drogues sont la consommation de cigarettes, d'alcool, de cocaïne ou de marijuana, le mois dernier, l'année dernière ou au cours de la vie. Les données sur les problèmes de santé ont trait, par exemple, au diabète l'année dernière/au cours de la vie, l'accident vasculaire cérébral l'année dernière/au cours de la vie ou l'obtention d'un traitement pour des problèmes de santé mentale. Les variables délicates (VD) sont des variables qui contiennent ce genre d'information.

2. Traitement contre la divulgation appliqué aux fichiers de microdonnées à grande diffusion de la NSDUH

2.1 Traitement contre la divulgation

Le but du traitement contre la divulgation est d'introduire une incertitude suffisante quant à la présence et à l'identité d'un répondant. Nous utilisons la méthode MASSC (Singh et coll., 2003, 2004) conjuguée à un traitement

supplémentaire des VI non essentielles et des variables à valeur extrême pour traiter les données de la NSDUH en vue de créer des fichiers à grande diffusion (FGD).

2.2 Méthode MASSC

La méthode MASSC est une méthode de contrôle de la divulgation statistique (CDS) qui combine la perturbation aléatoire et la suppression aléatoire en vue d'introduire une incertitude suffisante dans un cadre probabiliste. La perte d'information et le risque de divulgation après le traitement peuvent être contrôlés et mesurés simultanément. La méthode MASSC a été élaborée à RTI sous la direction d'Avinash Singh (Singh, 2002, 2006).

La méthode MASSC comprend les quatre grandes étapes suivantes : **M**icro**A**gglomération, **S**ubstitution, **S**ous-échantillonnage et **C**alage. À l'étape de la microagglomération, les variables posant un grand risque de divulgation sont catégorisées de façon à réduire les enregistrements uniques dans l'échantillon, puis des strates de risque sont créées en fonction de la combinaison de VI afin de contrôler le niveau de traitement aux dernières étapes. À l'étape de la substitution, les valeurs des VI d'un enregistrement sélectionné aléatoirement sont remplacées par celles d'un « enregistrement donneur de substitution ». À l'étape du sous-échantillonnage, certains enregistrements sont supprimés aléatoirement de la base de données diffusée. La substitution et le sous-échantillonnage sont exécutés afin de réduire au minimum les pertes associées à la divulgation : la substitution est destinée à introduire une incertitude au sujet de l'identité d'un répondant, tandis que le sous-échantillonnage a pour but d'introduire une incertitude au sujet de la présence d'un répondant dans l'enquête. Le biais dû à la substitution et la variance due au sous-échantillonnage sont limités par les contraintes de biais et les contraintes de variance, respectivement, de sorte que la perte d'information soit contrôlée. La dernière étape de la MASSC dans le traitement contre la divulgation est le calage des poids sur les variables sociales et démographiques afin de rajuster les poids dans le sous-échantillon aux poids totaux originaux dans le fichier analytique complet. Le calage minimise la perte de précision due aux traitements de sous-échantillonnage et de substitution.

2.3 Traitement supplémentaire après la procédure MASSC

Le traitement par la méthode MASSC ne s'applique qu'aux variables d'identification clés. Or, de nombreuses autres variables non soumises à la procédure MASSC peuvent encore poser un risque de divulgation. Par conséquent, les variables qui contiennent des renseignements d'identification personnels ou celles d'après lesquelles des renseignements d'identification personnels peuvent être inférés requièrent un traitement supplémentaire après la procédure MASSC. Nous procédons au regroupement des valeurs extrêmes supérieures (top coding) et inférieures (bottom coding) des variables présentant des valeurs extrêmes, à la fusion des niveaux pour les variables présentant des réponses rares et à la suppression des variables qui posent un risque élevé de divulgation, mais qui ont peu de valeur analytique. En outre, de nombreuses variables de la base de données sont corrélées aux VI clés utilisées dans la méthode MASSC. Afin de maintenir la cohérence interne, si les valeurs des VI clés sont substituées dans un enregistrement, les valeurs des variables connexes sont également remplacées par les valeurs correspondantes provenant du même enregistrement donneur.

3. Évaluation après le traitement

L'évaluation après le traitement comprend la mesure du risque de divulgation et de la perte d'information dans la base de données traitée. Nous mesurons le risque de divulgation en calculant la valeur delta (définie à la section 3.1) pour les enregistrements appartenant aux diverses catégories de risque et la perte d'information, en procédant à une comparaison des estimations et de leurs erreurs-types avant et après le traitement.

3.1 Risque de divulgation après le traitement : la mesure de delta

La procédure MASSC permet de quantifier le risque de divulgation dans la base de données traitée. Ce risque est mesuré grâce au calcul de δ, qui est défini comme étant la probabilité qu'un enregistrement qui figure en tant

qu'enregistrement unique ou non unique dans la base de données traitée pose un risque de divulgation. Les valeurs de δ sont définies d'après un ensemble de probabilités associées à la substitution et au sous-échantillonnage.

3.2 Évaluation de l'utilité analytique

Pour évaluer l'utilité analytique du fichier traité, nous procédons à une évaluation approfondie de la qualité analytique. Cette évaluation comporte les éléments suivants : comparaison avant et après le traitement des estimations et de leurs erreurs types, des contrastes et de leurs erreurs types, et des coefficients de régression univariés et de leurs erreurs types. Les estimations et les contrastes sont calculés pour une variable de consommation de drogues ou pour deux variables de consommations de drogues dans divers domaines démographiques, par exemple, la consommation de cocaïne l'année précédente, la consommation de marijuana le mois précédent, et la combinaison de ces deux variables de consommation de drogues. Nous calculons les régressions des variables de consommation de drogues sur certaines variables démographiques. Nous utilisons le ratio des estimations de la prévalence, des contrastes ou des coefficients de régression et de leurs erreurs-types provenant du fichier à grande diffusion à celles calculées sur échantillon complet pour les variables de consommation de drogues, afin de quantifier l'écart des estimations fondées sur le fichier à grande diffusion par rapport à celles calculées d'après le fichier d'accès restreint :

Ratio = (estimation sur sous-échantillon du FGD)/(estimation sur échantillon d'accès restreint)

Nous examinons la distribution des ratios et calculons leurs statistiques. En principe, la moyenne et la médiane des ratios devraient s'approcher de un. Si elles s'en écartent beaucoup, la procédure MASCC est réexécutée en utilisant des contraintes de biais et de variance plus rigoureuses.

4. Résultats sommaires du traitement contre la divulgation des données de la NSDUH de 2003

Comme pour les autres données de la NSDUH, le traitement contre la divulgation des données de 2003 a été mis en œuvre en appliquant la procédure MASSC, ainsi que le traitement supplémentaire des variables d'identification (VI) non essentielles et des variables présentant des valeurs extrêmes pour accroître la protection. Le FGD résultant (http://webapp.icpsr.umich.edu/cocoon/ICPSR-STUDY/04138.xml) contenait en tout 55 230 enregistrements. Le risque de divulgation est maintenu au niveau minimal après le traitement. Afin de déterminer l'effet de ces procédures sur le biais et la précision des estimations comparativement au fichier complet, nous avons étudié la précision de 140 estimations et de 190 contrastes fondés sur les combinaisons de 14 domaines et de 10 variables réponses, ainsi que pour 70 coefficients de régression provenant de 10 variables réponses dont la régression a été calculée sur quatre variables indépendantes démographiques. Nous avons calculé les ratios des estimations et de leurs erreurs types produites d'après le fichier complet et d'après le fichier à grande diffusion pour chaque combinaison domaine-variable réponse. Les statistiques sommaires de ces ratios pour les estimations et leurs erreurs-types sont présentées au tableau 1.

Tableau 1. Statistiques sommaires des ratios des estimations et des erreurs-types entre le sous-échantillon du FGD et l'échantillon complet

Distribution	Ratio après/avant (Est.)			Ratio après/avant (E.-T.))		
	Estimations	Contrastes	Coeff. régr.	Estimations	Contrastes	Coeff. régr.
100 % Max.	1,1164	10,5482	3,5817	1,3417	1,3533	1,4097
75 % Q3	1,0169	1,0358	0,9867	1,1495	1,1630	1,1750
50 % Médiane	1,0051	0,9930	0,9774	1,0823	1,0684	1,0865
25 % Q1	0,9920	0,9343	0,9372	1,0106	1,0043	1,0124
0 % Min.	0,9352	-3,4553	-3,9541	0,8292	0,7692	0,8345
Moyenne	1,0063	1,0511	0,8730	1,0831	1,0839	1,0925
N	140	190	70	140	190	70

Les données qui précèdent montrent que le ratio moyen des prévalences pour les dix variables de consommation de drogues sur les 14 domaines était de 1,01 et que l'accroissement moyen de l'erreur type des estimations était de 8 %. Les ratios médians pour les 190 contrastes et les 70 coefficients de régression étaient de 0,99 et 0,98, respectivement. Enfin, l'accroissement médian de l'erreur type des contrastes et des coefficients de régression était de 7 % et de 9 %, respectivement. Ici, nous avons utilisé les valeurs médianes pour évaluer l'effet du traitement sur la qualité analytique des données pour les contrastes ainsi que les coefficients de régression, parce que les deux grandeurs peuvent être proches de zéro, de sorte que les ratios risquent de présenter des variations extrêmes, même si le changement est faible.

Afin d'évaluer l'effet du traitement contre la divulgation sur l'inférence statistique à partir de la base de données traitée, nous avons également effectué un test *t* pour déterminer la variation de la signification au seuil de 5 % pour les contrastes ainsi que les coefficients de régression. Le tableau 2 montre que, sur les 190 contrastes, la signification au seuil de 5 % a changé pour neuf d'entre eux, parmi lesquels six sont passés d'une valeur significative à une valeur non significative, et trois sont passés d'une valeur non significative à une valeur significative; sur les 70 coefficients de régression, la signification au seuil de 5 % a changé pour trois d'entre eux, pour passer d'une valeur significative à une valeur non significative dans chaque cas.

Tableau 2. Changement de signification au seuil de 5 % pour les contrastes et les coefficients de régression

Analyse	Contrastes	Coeff. régr.
	190	70
N^{bre} total de changements	9 (4,74 %)	3 (4,29 %)
N^{bre} de changements de significatif à non significatif	6 (3,16 %)	3 (4,29 %)
N^{bre} de changements de non significatif à significatif	3 (1,58 %)	0 (0 %)

5. Conclusion

La combinaison de la méthode MASSC et d'un traitement supplémentaire offre une bonne base de contrôle de la divulgation pour les fichiers à grande diffusion de la NSDUH. L'analyse des données après l'application de ces procédures montre que le risque de divulgation dans la base de données traitée a été minimisé et que l'effet sur les niveaux et l'exactitude des estimations est minime.

6. Remerciement et avis de non-responsabilité

Le présent projet est financé par la Substance Abuse and Mental Health Services Administration, Office of Applied Studies, sous le numéro de contrat 283-2004-00022 et le numéro de projet 0209009. La communication sur laquelle est fondé le présent article a été parrainée par la Division de la statistique et de l'épidémiologie de RTI International.

Les opinions exprimées dans le présent article ne reflètent pas forcément les politiques officielles du Department of Health and Human Services; en outre, la mention de noms de marque, de pratiques commerciales ou d'organismes ne sous-entend pas que le gouvernement des États-Unis les cautionne.

Références

Singh, A. C. (2002, 2006), "Method for Statistical Disclosure Limitation", *US Patent Application Pub. No. US 2004/0049517A1*: Patent granted June 2006. Patent no. US7058638B2.

Singh, A.C., Yu, F., et Dunteman, G.H. (2003), "MASSC: A new data mask for limiting statistical information loss and disclosure", *Proceedings of the Joint UNECE/EUROSTAT Work Session on Statistical Data Confidentiality*, Luxembourg, pp. 373-394. (www.unece.org)

Singh, A., F. Yu, et D.H. Wilson. (2004), "Measuring Disclosure Risk and Information Loss for MASSC-treated Micro-data.", *Proceedings of the American Statistical Association*, Toronto, Canada, pp. 4374-4381.