Nations Unies ECE/CES/2013/29

Conseil économique et social

Distr. générale
16 avril 2013
Français
Original: anglais

Commission économique pour l'Europe

Conférence des statisticiens européens

Soixante et unième réunion plénière
Genève, 10-12 juin 2013
Point 4 b) de l'ordre du jour provisoire
Comment les bureaux de statistique nationaux devraient-ils réagir en passant du souci d'éviter les risques à la gestion des risques?

Modalités novatrices d'accès aux microdonnées − confidentialité instantanée

Note du Bureau australien de statistique (ABS)*

Résumé

Le présent document porte sur les approches novatrices de l'analyse des microdonnées. Il donne un aperçu général des déterminants qui ont amené l'ABS à s'efforcer d'utiliser ces techniques nouvelles. Il décrit la série de serveurs d'analyse à distance, ainsi que la méthode instantanée adoptée par le Bureau. Il fait un bref tour d'horizon des approches internationales permettant de donner accès aux microdonnées et s'achève par un exposé des orientations futures de la recherche.

L'ABS a reconnu qu'il convenait de «repenser» l'accès aux microdonnées afin de répondre à la demande des utilisateurs qui souhaitent disposer de données plus détaillées sur les unités statistiques pour un plus large éventail d'ensembles de données. Des démarches plus souples s'imposent au-delà de la simple confidentialisation des données avant leur publication. Par conséquent, l'ABS a mis en place un serveur d'accès à distance appelé Remote Execution Environment for Microdata (REEM) comprenant un générateur de tableaux (*table builder*) de données d'enquêtes et un serveur d'analyse. Par rapport aux méthodes classiques de confidentialisation, cette nouvelle technique réduit au minimum les effets sur la qualité des données en n'appliquant les modifications nécessaires qu'au produit final et non pas aux données de base. Tous les produits sont anonymisés «instantanément» de manière à protéger la confidentialité des données tout en préservant le degré de détail et la qualité des données nécessaires.

* Le présent document a été soumis avec retard en raison d'une transmission tardive des apports provenant d'autres sources y relatives.

I. Introduction

1. Les organismes de statistique recueillent de très grandes quantités de microdonnées provenant de recensements, d'enquêtes et de sources administratives. Ces microdonnées peuvent être utilisées pour élaborer et évaluer des politiques bénéfiques ou utiles à la société. Par conséquent, la demande d'accès à des microdonnées de cette nature a continué d'augmenter depuis l'examen de la question de la confidentialité et de l'accès aux microdonnées par la Conférence des statisticiens européens à sa réunion plénière de 2003.

2. Bureau australien de statistique (ABS) a pour mission de faciliter et d'encourager la prise de décisions éclairées, la recherche et le débat au sein de l'administration et des collectivités, en tant que service national de statistique de haute qualité, objectif et réactif. L'ABS a dû, comme bon nombre d'autres organismes nationaux de statistique (ONS), «repenser» l'accès aux microdonnées, face à la difficulté de concilier l'obligation juridique de prévenir le risque de divulgation d'informations concernant une personne ou une organisation particulière, et la nécessité de publier des microdonnées plus détaillées pour la prise de décisions éclairées, la recherche et le débat dans l'intérêt de la société.

3. La gestion du risque de divulgation est communément appelée «contrôle de la divulgation de données statistiques». Même après avoir éliminé dans les microdonnées les informations permettant d'identifier les personnes, telles que le nom et l'adresse, le risque de divulgation subsiste (voir par exemple Willenborg and de Waal, 2001). Eu égard à la croissance du volume de données qui peuvent être obtenues auprès de sources administratives et apparentées, favorisée par le progrès technologique, il y a lieu de croire que le risque de divulgation des microdonnées s'accentue sans cesse.

4. Il existe plusieurs logiciels de protection de la confidentialité des données. La majorité d'entre eux sont conçus pour être utilisés par les ONS afin de rendre les données confidentielles avant leur divulgation. Il peut s'agir de logiciels applicables aux données mises en tableaux (par exemple, Tau-Argus[1] et sdcTable[2]) ou aux microdonnées (par exemple, le Special Uniques Detection Algorithm[3] et SDCMicro[4]). Par ailleurs, bon nombre d'ONS ont mis au point leurs propres processus et logiciels sur mesure adaptés aux prescriptions de la loi.

5. L'ABS a progressivement élaboré toute une gamme de méthodes, de processus et d'applications en vue d'offrir aux utilisateurs une série de produits ouvrant accès à ses données statistiques, notamment via la publication de tableaux statistiques sur son site Web, la diffusion de tableaux personnalisés par un service d'information et de référence et l'analyse de fichiers de microdonnées préconfidentialisés, appelés «Fichiers d'enregistrements unitaires confidentialisés» (CURF). Les CURF sont disponibles sous deux formes: CURF de base que les analystes peuvent utiliser dans leur propre cadre de recherche ou CURF élargi accessible en soumettant des demandes à distance au laboratoire de données avec accès à distance (RADL) ou en visitant sur place l'un des

[1] Tau-Argus est un logiciel libre destiné à protéger les tableaux statistiques, téléchargeable à partir de l'adresse suivante: http://neon.vb.cbs.nl/casc/tau.htm.

[2] sdcTable est un progiciel libre visant à contrôler la divulgation des données tabulaires, téléchargeable à partir de l'adresse ci-après: http://cran.r-project.org.

[3] Special Uniques Detection Algorithm est un système de détection et de gradation de données enregistrées à caractèrte unique. On s'en sert pour confidentialiser les ensembles de données en repérant d'abord toutes les données présentant un caractère unique puis en les déguisant ou en les éliminant.

[4] SDCMicro est un logiciel libre permettant de produire des microdonnées protégées à l'usage des chercheurs et du grand public. Il peut être téléchargé à l'adresse suivante: http://cran.r-project.org/web/packages/sdcMicro/index.html.

laboratoires de données de l'ABS (ABSDL). En général, les utilisateurs autorisés présentent leurs demandes dans les langages statistiques SAS, STATA ou SPSS[5]. Dans le cas du RADL, les réponses aux demandes sont vérifiées automatiquement par le système et les données de sortie acceptées sont mises à la disposition des utilisateurs sur leur poste de travail. En outre, un échantillon de demandes est suivi de manière permanente à des fins d'inspection manuelle. Dans le cas de l'ABSDL, la diffusion de tous les produits issus du laboratoire est autorisée manuellement. Le degré de confidentialisation préalable appliqué à chaque fichier CURF dépend du mode de diffusion, le CURF de base étant fortement sécurisé tandis que le CURF élargi est plus détaillé. Des volumes considérables de ressources en personnel et en temps sont nécessaires pour produire l'un ou l'autre type de CURF.

6. Confronté à un certain nombre de déterminants du changement, notamment à la complexité accrue des ensembles de données risquant de mettre en échec les approches traditionnelles, l'ABS a récemment investi dans la mise en place d'une série d'applications qui permettent aux analystes autorisés de soumettre par Internet des demandes auxquelles une réponse est apportée en temps réel (instantanément) au moyen des microdonnées disponibles, avec retour du produit confidentialisé à l'analyste. Au niveau international, les applications de cette nature sont communément appelées «Serveurs d'analyse à distance» (RASs). Ils offrent aux utilisateurs un moyen de contrôler les produits particuliers qu'ils souhaitent extraire d'un ensemble de données. La difficulté pour les organismes de statistique consiste à assurer le contrôle de la divulgation des données pour les différents produits accessibles.

7. La section II du présent document contient un tour d'horizon des principaux déterminants qui ont conduit l'ABS à rechercher des approches novatrices de l'analyse des microdonnées. La section III présente la série de serveurs d'analyse à distance de l'ABS et l'approche instantanée du contrôle de la divulgation de statistiques. La section IV donne un bref aperçu de différentes approches internationales de l'accès aux microdonnées. En conclusion, on trouve dans la section V une vue d'ensemble des orientations futures de la recherche et des possibilités de collaboration internationale.

II. Déterminants du changement

8. De nombreux ONS, notamment le Bureau australien de statistique, sont en train de mettre à jour leur mode de diffusion des microdonnées aux chercheurs. Parmi les déterminants majeurs de ce changement, citons:

a) Les demandes croissantes appelant à un accès renforcé, plus souple et plus rapide aux microdonnées détaillées;

b) Le développement de l'expérience des utilisateurs grâce à la mise en place d'interfaces conviviales fonctionnant par menus qui n'exigent pas de compétences en matière de programmation statistique;

c) La réduction des coûts résultant des méthodes existantes de diffusion des microdonnées qui s'utilisent manuellement et qui nécessitent une grande quantité de ressources, comme le processus de création de fichiers d'enregistrements unitaires confidentialisés (CURF);

[5] Ces solutions logicielles statistiques à usages multiples sont employées pour la production et l'analyse de statistiques: SAS (Statistical Analysis System), STATA et SPSS (Statistical Package for the Social Sciences).

d) La volonté de rendre les microdonnées accessibles dans de meilleurs délais (actuellement, il faut attendre jusqu'à six mois après la diffusion des tableaux statistiques pour que les CURF soient disponibles);

e) La lutte contre le risque croissant de divulgation lié à la hausse de la puissance de calcul (tant au niveau du matériel que des logiciels), l'augmentation du volume de produits diffusés et l'accessibilité accrue de larges ensembles de données;

f) Le souci de faciliter l'analyse de sources nouvelles de données (y compris les données transactionnelles, administratives et intégrées) pour lesquelles les méthodes traditionnelles de protection contre la divulgation sont insuffisantes pour pouvoir limiter le risque accru d'identification;

g) La prise de conscience croissante que toutes les ressources statistiques essentielles ne sont pas détenues par les ONS, d'où le besoin de concevoir des méthodes et infrastructures utilisables par d'autres organismes; et

h) L'évolution du modèle de l'analyste de données classique qui pousse les organismes à rechercher une meilleure accessibilité des produits grâce à un système d'interrogation de machine à machine plus efficace, comme l'utilisation de services SDMX (échange de données et de métadonnées statistiques) en ligne.

9. Ces déterminants du changement ont conduit de nombreuses ONS, notamment l'ABS, à entreprendre la mise en place de serveurs d'analyse à distance perfectionnés.

III. Serveurs d'analyse à distance de l'ABS – confidentialité instantanée

10. La série de serveurs d'analyse à distance de l'ABS permet à celui-ci de rendre disponible le détail d'ensembles de données tout en minimisant la perte d'utilité grâce à l'application d'une protection contre la divulgation des statistiques adaptée à chaque produit spécifique. Quant aux utilisateurs, ils contrôlent les produits particuliers qu'ils désirent extraire d'un ensemble de données. Il s'agit d'une réorientation fondamentale du processus. Le paradigme traditionnel voulait que l'ABS décide en totalité de la diffusion des produits. Désormais, les utilisateurs peuvent préciser ce qui leur faut, en fonction de l'évolution de leurs besoins au moment opportun.

11. Il convient de lutter contre un risque de divulgation réel s'agissant de la diffusion des données de tableaux et de résultats d'analyses. Plusieurs documents ont été publiés à ce sujet, notamment des propositions de stratégies en vue de gérer le risque de divulgation. En ce qui concerne les analyses, consulter Gomatam *et al.* (2005), Bleninger *et al.* (2011) ainsi que Sparks *et al.* (2008) et pour ce qui est des tableaux, voir Shlomo (2007). Ces ouvrages visent à lutter contre les atteintes aux données, autrement dit à empêcher qu'un analyste, en utilisant le produit issu d'un serveur d'analyse, y compris des graphiques et des évaluations des modèles, reconstitue les attributs correspondant à un ou plusieurs enregistrements car, s'il y parvient, une tentative d'identification est réalisable. L'enjeu pour l'ABS consiste à assurer une protection contre la divulgation statistique pour chaque produit disponible à la demande.

12. Pour un serveur d'analyse à distance, un modèle très simpliste se présente comme suit:

a) L'organisme de statistique met un fichier de microdonnées à la disposition des chercheurs. Ce fichier est détenu en toute sécurité par l'organisme. L'analyste n'est généralement pas en mesure de voir les microdonnées à caractère sensible;

b) Un analyste soumet une demande par Internet au serveur d'analyse des organismes. Cette demande est comparée aux microdonnées à caractère sensible;

c) Les produits statistiques (par exemple des coefficients de régression ou une tabulation) répondant à la demande sont modifiés grâce à une méthode de sécurisation spécialement adaptée à l'analyse en question afin d'assurer la protection contre la divulgation statistique;

d) Le serveur d'analyse envoie le produit à l'analyste via Internet. Certains produits peuvent faire l'objet de restrictions au motif qu'ils pourraient permettre à un analyste de reconstituer les attributs d'un enregistrement arbitraire.

13. Le serveur d'analyse de la prochaine génération de l'ABS est composé de trois applications: Census TableBuilder, ABS TableBuilder et ABS DataAnalyser. Une description de chacun de ces outils figure aux paragraphes suivants. L'analyste ne voit jamais les microdonnées sous-jacentes.

14. À l'occasion du recensement de la population et du logement réalisé en 2006, l'ABS et Space-Time Research Pty Ltd ont élaboré conjointement le générateur de tableaux de données du recensement **Census TableBuilder**, un produit en ligne mis à la disposition des analystes non rattachés à l'ABS. Le Census TableBuilder comporte une méthode de perturbation (Fraser and Wooton (2005)) qui permet de protéger automatiquement les tableaux de données chiffrées du recensement. Cette méthode a été conçue pour limiter les risques de divulgation résultant de demandes de tableaux similaires, de demandes répétées de tableaux identiques ainsi que de demandes répétées de la même cellule figurant dans différents tableaux. Elle vise à garantir un accès plus large aux données relatives aux sous-populations et à favoriser la mise en place de systèmes en ligne permettant aux utilisateurs de définir leurs propres tableaux. La même méthode est employée pour protéger tous les tableaux du recensement de 2006, notamment les tableaux établis pour les publications par le personnel de l'ABS. Dès lors, il existe des systèmes internes d'application de cette méthode et du Census TableBuilder en ligne destinés aux analystes.

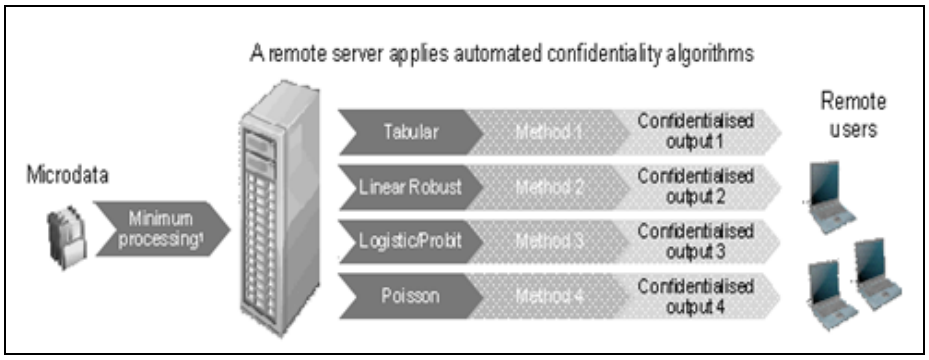
15. Sur cette base, l'ABS a continué de mettre au point une série d'algorithmes et de protections sur mesure assurant la confidentialité qui fonctionnent instantanément et minimisent la perte d'utilité, tout en rendant improbable l'identification d'une personne interrogée à partir des produits obtenus. Chaque méthode de confidentialité est adaptée à un produit spécifique, ce qui réduit le degré de confidentialisation nécessaire. À titre de comparaison, les systèmes existants de fichiers d'enregistrements unitaires confidentialisés (CURF) exigent une plus grande confidentialité pour protéger les données de la totalité des produits potentiellement réalisables.

16. Le générateur de tableaux de données **ABS TableBuilder**, qui a succédé au Census TableBuilder et a également été élaboré en collaboration avec Space Time Research Pty Ltd, inclut des mécanismes dynamiques de confidentialité pour les données d'enquête pondérées qui vont maintenant au-delà de simples dénombrements d'une population statistique et s'appliquent aussi aux statistiques récapitulatives clefs fondées sur des données quantitatives (telles que les intervalles personnalisés, les valeurs totales, moyennes et médianes et les quantiles). Outre la perturbation, plusieurs protections et restrictions ont été intégrées, par exemple pour restreindre les combinaisons d'éléments de données dans un même tableau, limiter la production de tableaux peu fournis qui comprennent un grand nombre de petites cellules et exiger une taille minimale de population pour le calcul de médianes et de quantiles. L'ABS TableBuilder simplifie la demande de microdonnées de machine à machine par l'intermédiaire d'un service SDMX en ligne.

17. L'analyseur de données **ABS DataAnalyser** a été élaboré pour faciliter l'analyse exploratoire de données et l'établissement de modèles de régression. Il s'agit d'un système sûr fondé sur des menus qui permet de réaliser des analyses statistiques via une interface

utilisateur à distance. Grâce à ce système, les utilisateurs peuvent estimer à distance les paramètres des modèles statistiques adaptés aux données de l'ABS tout en préservant la confidentialité des fournisseurs. Tous les produits statistiques pouvant être vus par l'utilisateur sont automatiquement rendus confidentiels grâce à diverses méthodes de contrôle de la divulgation, y compris la perturbation de l'équation d'estimation. Cette perturbation à elle seule ne suffisant pas, un ensemble de restrictions et de protections contre des atteintes spécifiques ont aussi été intégrées au système, notamment des affichages graphiques rendus confidentiels qui aident les utilisateurs à évaluer la justesse du modèle. Un résumé de la stratégie de perturbation et une description des protections supplémentaires sont disponibles dans le document rédigé par Chipperfield, Gare et Yu (2011). La version initiale de l'ABS DataAnalyser permet aux utilisateurs de transformer et de manipuler les données, de créer des tableaux, d'effectuer des analyses exploratoires de données ainsi que de procéder à des modélisations linéaires, robustes, logistiques, de Probit, de Poisson ou multinomiales. Le système est actuellement testé auprès d'utilisateurs et une édition complète est prévue pour le troisième trimestre 2013. L'ABS DataAnalyser est illustré par le diagramme 1.

Diagramme 1
Analyseur de données (DataAnalyser) de l'ABS comportant une confidentialité instantanée



18. La stratégie de l'ABS relative aux serveurs d'analyse à distance présente, entre autres, les avantages suivants:

a) L'analyse porte sur les microdonnées réelles en conservant les relations complexes dans les données;

b) Le produit statistique est modifié dans une mesure spécialement adaptée au type d'analyse entrepris et au niveau permettant de minimiser la perte d'informations;

c) Une fois le serveur mis en place, il peut traiter de multiples analyses en temps réel;

d) Tous les programmes soumis peuvent être enregistrés et vérifiés, et si une tentative d'identification est repérée, l'analyste peut voir son autorisation d'accès au serveur révoquée; et

e) Grâce au menu de type «pointer-cliquer», l'utilisateur n'a pas besoin de suivre une longue formation ni d'apprendre un nouveau langage de programmation.

19. La stratégie de l'ABS présente, entre autres, les inconvénients suivants:

a) L'analyste peut uniquement utiliser les techniques d'analyse et procéder aux transformations et manipulations de données prises en charge par le serveur;

b) Une analyse peut durer plus longtemps lorsqu'elle passe par des serveurs à distance que lorsque les microdonnées sont disponibles sur l'ordinateur personnel de l'analyste; et

c) Un investissement important en termes de temps et d'argent doit être consenti afin d'élaborer des mécanismes logiciels de confidentialisation pour chaque nouvelle fonctionnalité analytique.

IV. Méthodes internationales pour fournir un accès aux microdonnées

20. En ce qui concerne l'accès aux microdonnées, les organismes nationaux de statistique (ONS) ont adopté des stratégies très différentes, ce qui s'explique en grande partie par la diversité des exigences législatives. Plusieurs ONS publient des fichiers à usage public. Ces fichiers sont rendus hautement confidentiels en vue d'un usage général.

21. Les ONS font aussi largement appel aux centres de données de recherche, semblables au laboratoire de données de l'ABS, pour l'analyse de microdonnées détaillées. L'inconvénient de ces méthodes est que les produits retirés de ces centres doivent généralement être contrôlés manuellement. Ce processus mobilisant une grande quantité de ressources, le nombre de chercheurs qui peuvent obtenir un accès doit être restreint. L'ABS s'efforce de trouver une solution pour assurer un accès le plus large possible.

22. Certains centres de recherche sont situés sur place tandis que d'autres tirent parti des progrès technologiques pour créer un centre de recherche virtuel auquel il est possible d'accéder depuis des terminaux non intelligents installés dans d'autres organismes. Le degré de détail des données pouvant être consultées varie aussi fortement en fonction de la législation de l'ONS. Dans certains cas, des chercheurs qui ont reçu une autorisation ou bénéficient de la confiance de l'organisme disposent du même accès aux microdonnées détaillées que le personnel de l'ONS. En Australie, ceci n'est pas permis par la législation applicable à l'ABS.

23. Par comparaison, les systèmes en ligne de l'ABS sont conçus pour permettre aux utilisateurs externes d'accéder aux données à distance. Ceci ne requiert pas une confidentialisation de grande ampleur des microdonnées avant l'analyse mais passe par des méthodes de confidentialité appliquées en temps réel. Conscientes des avantages offerts par les serveurs d'analyse à distance qui intègrent la confidentialité instantanée, plusieurs ONS ont lancé des programmes de recherche-développement. Il est intéressant de citer deux projets relativement avancés, à savoir Morpheus (Höninger (2011)), élaboré par l'Institut de statistique de Berlin-Brandebourg, et le système d'analyse de microdonnées (Lucero *et al.* (2011)), en cours d'élaboration au Census Bureau des États-Unis.

V. Orientations futures et possibilités de collaboration internationale

24. Les serveurs d'analyse à distance de l'ABS offrent l'avantage de fournir des produits de grande qualité dérivés de fichiers de microdonnées ainsi qu'une facilité d'accès aux utilisateurs sans compromettre la confidentialité des données. Il reste encore de nombreux défis à relever.

25. Les travaux de recherche actuels portent principalement sur l'évaluation des méthodes existantes et leur efficacité quant aux ensembles de données liés, qui présentent un risque de divulgation supérieur. À l'avenir, la recherche sera axée sur l'établissement de méthodes de confidentialité instantanée pour la diffusion d'ensembles de données sur les entreprises et données longitudinales par l'intermédiaire de serveurs d'analyse à distance et la mise en place de mesures permettant aux chercheurs de limiter la perte d'utilité (Marley and Leaver, 2011).

26. L'ABS accorde un grand intérêt à l'examen du potentiel des microdonnées synthétiques, ne serait-ce que pour permettre aux analystes de tester et de vérifier leurs modèles avant de les appliquer aux données réelles stockées en toute sécurité dans l'ABS DataAnalyser.

27. L'ABS souhaiterait travailler avec la communauté internationale des statisticiens afin d'effectuer des recherches visant à relever les défis que posent ces ensembles de données. Il serait également prêt à fournir les méthodes statistiques et algorithmes connexes aux membres de cette communauté qui désireraient intégrer ces méthodes dans leurs applications.

VI. Bibliographie

Bleninger, P., Drechsler, J., et Ronning, G., 2011, «Remote Data Access and the Risk of Disclosure from Linear Regression: An Empirical Study», *Privacy in Statistical Databases*, Springer. Voir http://www.idescat.cat/sort/sortspecial2011/Data Privacy. 1.bleninger-etal.pdf.

Chipperfield, J., Gare, M., et Yu, F., 2011, «Providing access to microdata for statistical purposes − experiences of the Australian Bureau of Statistics with Remote Analysis Servers», document présenté en 2011 lors du colloque sur la méthodologie de Statistique Canada, Ottawa, Canada, 1-4 novembre.

Fraser, B., et Wooton, J., 2005, «A proposed method for confidentialising tabular output to protect against differencing», document présenté lors de la réunion de travail commune CEE/Eurostat sur la confidentialité des données statistiques, Genève, Suisse, 9-11 novembre.

Gomatam, S., Karr, A. F., Reiter, J. P., et Sanil, A. P., 2005, «Data Dissemination and Disclosure Limitation in a World Without Microdata: A Risk –Utility Framework for Remote Access Analysis Servers», *Statistical Science*, 20, p. 163 à 177.

Höninger, J., 2011, «An Innovative Approach to Remote Data Access», cinquante-huitième Congrès mondial de la statistique de l'Institut international de statistique, Dublin, Irlande, 21-26 août 2011.

Lucero, J., Zayatz, L., Singh, L., You, J., DePersio, M., et Freiman, M., 2011, «The Current Stage of the Microdata Analysis System at the U.S. Census Bureau», cinquante-huitième Congrès mondial de la statistique de l'Institut international de statistique, Dublin, Irlande, 21-26 août 2011.

Marley, J., et Leaver, V., 2011, «A method for confidentialising user-defined tables: statistical properties and a risk-utility analysis», document présenté à la session de l'Institut international de statistique, Dublin, Irlande, 22-26 août.

Shlomo, N., 2007, «Statistical disclosure control methods for census frequency tables», International Statistical Review, 75 (2), 199-217.

Sparks, R., Carter, C., Donnelly, J., O'Keefe, C. M., Duncan, J., Keighley, T., et McAullay, D., (2008), «Remote Access Methods for Exploratory Data Analysis and Statistical Modelling: Privacy-Preserving AnalyticsTM», *Computer Methods and Programs in Biomedicine* 91, p. 208 à 222.

Willenborg, L., et de Waal, T., 2001, «Elements of Disclosure Control», *Lecture Notes in Statistics*, vol. 155, ISBN 978-0-387-95121-8, Springer.