Recueil du Symposium 2014 de Statistique Canada
Au-delà des méthodes traditionnelles d'enquêtes : l'adaptation à un monde en évolution

Les différents contextes de l'utilisation de données administratives à des fins statistiques

Loredana Di Consiglio et Piero Demetrio Falorsi[1]

Résumé

Le projet Methodologies for an integrated use of administrative data (MIAD) du Réseau statistique a pour but d'élaborer des méthodologies en vue d'un usage intégré des données administratives (DA) dans le processus statistique. Le principal objectif du projet MIAD est de fournir des lignes directrices pour l'exploitation des DA à des fins statistiques. En particulier, les membres du projet ont élaboré un cadre de la qualité, ont fourni une représentation des utilisations possibles des DA et proposé un schéma des différents contextes informatifs. Le présent article est axé sur ce dernier aspect. En particulier, nous faisons la distinction entre les dimensions en rapport avec les caractéristiques de la source associées à l'accessibilité, d'une part, et les caractéristiques associées à la structure des DA et à leurs liens avec les concepts statistiques, d'autre part. Nous désignons la première catégorie de caractéristiques comme étant *le cadre de l'accès* et la deuxième catégorie de caractéristiques comme étant le *cadre des données*. Dans le présent article, nous nous concentrons principalement sur la deuxième catégorie de caractéristiques qui sont reliées spécifiquement au type d'information qui peut être obtenu à partir de la source secondaire. En particulier, ces caractéristiques ont trait à la population administrative cible et à la mesure de cette population ainsi que la façon dont elle est (ou pourrait être) liée à la population cible et aux concepts statistiques cibles.

Mots clés : cadre de la qualité; GSBPM; accessibilité; mesure.

1. Introduction

1.1 Description du projet Methodologies for an integrated use of administrative data

Le présent travail fait partie des activités planifiées dans le cadre du projet Methodologies for an integrated use of administrative data (MIAD; http://www1.unece.org/stat/platform/display/msis/Statistical+Network). Les membres du projet MIAD sont le National Institute of Statistics (ISTAT) d'Italie, Statistique Canada, l'Australian Bureau of Statistics, Statistics New Zealand et Statistics Sweden. Le but du projet est d'élaborer des stratégies cohérentes et bien fondées, permettant d'exploiter pleinement l'utilisation de sources de données administratives (DA) dans le processus statistique. Afin d'atteindre ce but, l'équipe du projet MIAD a d'abord visé à approfondir l'analyse des dimensions qu'il convient de prendre en considération pour évaluer l'usage d'une source externe dans le processus statistique. À cet égard, l'accent a d'abord été mis sur l'élaboration d'un cadre en vue d'évaluer la qualité des DA et la mesure dans laquelle elles sont utilisables à des fins statistiques — avant leur utilisation dans le processus statistique, afin de déterminer si une source de DA peut être utilisée à des fins statistiques et de quelle manière. En particulier, on a établi l'ensemble de caractéristiques souhaitées d'un cadre de la qualité et l'ensemble d'indicateurs de la qualité appliqués avant l'introduction des DA dans le processus statistique. Le projet MIAD prend aussi en considération le besoin de lignes directrices en vue de relier les résultats de cette évaluation préliminaire à l'usage effectif des DA dans le processus statistique, afin de suggérer des seuils pour établir de quelle façon il est approprié d'utiliser les DA, car celles-ci peuvent être utilisées, par exemple, pour des totalisations directes des données (après un prétraitement) ou comme variables auxiliaires dans le processus d'estimation.

[1]Loredana Di Consiglio, ISTAT, Via Cesare Balbo, 16 Italie, Rome, 00184 (diconsig@istat.it); Piero Demetrio Falorsi, ISTAT, Via Cesare Balbo, 16 Italie, Rome, 00184 (falorsi@istat.it).

Un autre champ d'activités connexe important cerné par l'équipe du projet MIAD a trait à la représentation de la façon dont les DA peuvent être et sont en fait utilisées dans le processus statistique. En particulier, la mise en concordance avec le Modèle de processus opérationnel statistique générique (GSBPM) sert de fondement. En outre, le GSBPM est réanalysé dans la perspective particulière de l'utilisation des DA.

Cette activité d'exploration des différents usages possibles des DA est reliée aux actions qui peuvent devoir être effectuées sur les DA et à l'aperçu des scénarios les plus fréquents en ce qui a trait aux types possibles de cadres d'information auxquels fait face l'INS pour produire des données statistiques en se servant de DA.

Au cours d'une phase subséquente du projet, qui débutera en 2015, l'équipe du MIAD se penchera aussi sur les méthodes statistiques propres à l'exploitation et au traitement des DA, qui sont seulement survolées rapidement ici, et sur les mesures de la qualité des données de sortie.

Le présent travail porte sur l'activité d'exploration des cadres informatifs les plus fréquents auxquels fait face l'INS pour produire des données statistiques au moyen de DA et a principalement pour but d'établir un schéma des méthodes statistiques qui sont nécessaires pour intégrer les DA dans la production de statistiques. Ces considérations, conjuguées à une évaluation de la qualité des DA à l'entrée, serviront de fondement à l'élaboration de lignes directrices concernant le bon usage d'une source de DA.

1.2 Principales caractéristiques des scénarios des DA

Le présent travail a pour objet de donner un aperçu des contextes possibles de l'utilisation des DA et de commencer à relier les différentes actions possibles pour leur utilisation à des fins statistiques aux différents cadres des DA. En particulier, nous visons ici à décrire les dimensions contextuelles les plus pertinentes qui influent sur l'utilisation statistique des DA et les procédures nécessaires pour intégrer les DA dans le processus statistique, et qui déterminent les méthodes qu'il convient d'appliquer pour créer l'information statistique. Certaines de ces dimensions sont liées aux caractéristiques de la source associées à l'*accessibilité*, et de manière plus générale à l'environnement (p. ex. les relations entre l'INS et les fournisseurs externes de données ou les propriétaires des données), tandis que d'autres dimensions ont trait à la *structure* des *objets* dans les DA et à sa relation avec les concepts statistiques.

Nous appelons la première catégorie de caractéristiques *le cadre de l'accès* aux données et la deuxième catégorie de caractéristiques, la *structure intrinsèque des DA*. En fait, dans certaines circonstances, il semble que les DA doivent être exploitées pour un usage indirect seulement, tandis que dans d'autres, des données de sortie statistiques peuvent être obtenues directement à partir des DA, utilisées seules ou dans un système complexe de sources. En particulier, pour ce qui est du cadre de l'accès, nous faisons la distinction entre les cadres juridique et institutionnel. En ce qui concerne le premier, certains cadres juridiques nationaux accordent en fait plus de pouvoirs que d'autres en matière d'accès aux DA à des fins statistiques — en fixant les limites de cet accès et des utilisations des DA. Il existe souvent des restrictions spécifiant que les données peuvent seulement être utilisées à des fins statistiques particulières et que la confidentialité des dossiers individuels doit être protégée. Le cadre juridique influence la disponibilité des données à un niveau non agrégé (microdonnées) ou agrégé, qui a aussi une incidence sur la présence des identificateurs (pour des raisons de protection de la vie privée) et sur le genre de variables qui peuvent être partagées avec l'INS, ce qui à son tour a une incidence sur les usages possibles des DA et sur la qualité des données de sortie obtenues en les utilisant. Une loi permettant l'accès aux DA est en fait une condition préalable fondamentale à leur utilisation. En ce qui concerne la seconde dimension, c'est-à-dire le *cadre institutionnel*, nous entendons par là l'organisation établie pour acquérir les DA, en vue de les utiliser. Elle est principalement associée à des caractéristiques telles que l'actualité et la disponibilité des métadonnées, c'est-à-dire des dimensions qui ont une incidence sur la qualité, en particulier celle des données d'entrée, et elles sont habituellement incluses dans les cadres d'évaluation de la qualité.

En outre, un système fortement intégré présentant des liens efficaces entre l'INS et le ou les fournisseurs de données aide aussi le premier à s'assurer de pouvoir être informé à l'avance des changements prévus afin de planifier les mesures qui pourront être prises. Ces aspects ne sont pas décrits en détail ici; nous invitons le lecteur à consulter UNECE (2011) et Royce (2013) pour obtenir de plus amples renseignements sur les cadres existants que les INS ont établis pour l'utilisation des DA dans le processus statistique.

À la section suivante, nous nous concentrons sur le deuxième ensemble de caractéristiques que nous avons appelé structure des DA, en soulignant les aspects qui influent sur l'assurance de la qualité et les méthodes à mettre en œuvre pour obtenir des unités et des concepts statistiques et pour améliorer leur qualité.

2. Différents scénarios des DA : éléments de la structure des données

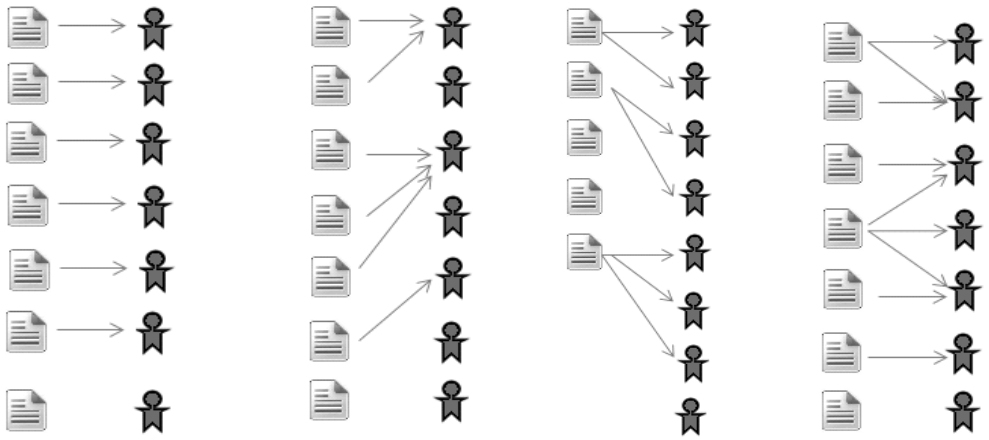
2.1 Le modèle à deux cycles de vie

Pour décrire la structure des DA, nous nous inspirons du modèle à deux cycles de vie de Zhang (2012). Ce modèle introduit explicitement le concept de réutilisation « secondaire » d'une source dans le modèle du cycle de vie. En premier lieu, Bakker (2010) a développé le modèle de Groves et coll. (2004), qui avait été introduit dans le contexte des enquêtes, pour traiter le cas des DA. Puis, Zhang (2012) a étendu le modèle de Bakker pour l'adapter à une situation de registre combiné. Ces modèles ont été introduits explicitement pour tenir compte des sources possibles d'erreurs dans le processus, et dans la réutilisation statistique d'une source, en mettant en relief le processus auquel doit être soumise la source externe. En particulier, pour notre projet, nous nous intéressons à la deuxième phase du modèle à deux cycles de vie de Zhang, qui décrit la relation entre les enregistrements de DA disponibles et les concepts cibles ainsi que la population à laquelle s'intéresse l'INS. Ce cadre est particulièrement utile pour décrire comment les différents scénarios (et les mesures connexes de la qualité des données d'entrée) influent sur les processus nécessaires pour obtenir l'information statistique souhaitée et le risque d'erreurs associé, et comment ils affectent la production et la qualité des données de sortie. Ici, nous décrivons les structures possibles des données dans différentes situations. Comme nous l'avons déjà mentionné en ce qui concerne la structure des données, nous entendons les types de caractéristiques et de mesures de la population administrative et la façon dont elles sont (ou pourraient être) reliées à la population statistique cible et aux concepts statistiques cibles. À l'instar de Zhang, nous faisons la distinction entre la représentation (objets ou unités) et la mesure — les variables et les diverses configurations possibles qui sous-tendent aussi la façon dont elles sont reliées.

2.2 Différents scénarios : des enregistrements administratifs aux unités statistiques

En nous inspirant du modèle de Zhang (2012), du *côté représentation* du cadre proposé, nous pouvons déterminer dans quelle mesure les enregistrements de DA concordent avec les unités statistiques cibles. Idéalement, chaque enregistrement de DA correspond à une unité de la population cible. En fait, en raison de la nature des DA saisies, les enregistrements dans les sources de DA correspondent souvent à des événements ou à des transactions, de sorte que l'information pour chaque unité statistique peut seulement être obtenue en combinant l'information de multiples enregistrements (cela est évidemment relié au *côté mesure* du modèle du cycle de vie) ou en tenant compte des complexités des liens entre les enregistrements et les unités. La figure 2.2-1 qui suit décrit les divers cas possibles de liens entre le contenu de la source de DA et les unités statistiques, représentant des cas d'appariement un à un, plusieurs à un, un à plusieurs ou, enfin, plusieurs à plusieurs. L'existence d'identificateurs (ou de variables clés) est un élément important de la configuration de ce que nous avons appelé le côté représentation de la structure des données, compte tenu du fait que leur absence se traduirait par une plus forte probabilité d'introduire des erreurs dans les procédures de couplage ou, en général, que la qualité de ce dernier dépend de la qualité des variables d'appariement. Notons que la disponibilité d'identificateurs, ou plus généralement de variables d'appariement, est souvent limitée par des questions de confidentialité, c'est-à-dire par le cadre juridique qui caractérise le scénario d'accessibilité.

Figure 2.2-1
Liens possibles entre les objets et les unités statistiques



À titre d'exemple d'une situation d'appariement complexe plusieurs à plusieurs, nous mentionnons le cas des statistiques sur l'agriculture, où des DA sur les fermes peuvent être disponibles (un registre des fermes ou un registre des demandes de subvention), tandis que les unités statistiques cibles sont les ménages ruraux en rapport avec ces unités administratives. D'autres exemples typiques sont la transcription d'événements se rapportant à la population d'étudiants ou l'information sur les emplois; dans ces cas, nous pourrions habituellement être dans la situation d'un appariement de type plusieurs à un. En outre, ces exemples montrent que de nombreuses populations statistiques cibles différentes peuvent être étudiées à partir de l'ensemble d'enregistrements de DA, c'est-à-dire les fermes proprement dites ou inversement les ménages ruraux; les emplois ou les employés ou les employeurs, et le scénario des DA sera manifestement d'une complexité informative différente pour différents objectifs statistiques. Zhang (2012) envisage aussi des situations plus complexes, où des unités statistiques complexes doivent être dérivées d'objets de DA, par exemple des « entités juridiques » dans le domaine des entreprises ou des « ménages » dans le domaine social. Zhang (2011) propose une théorie de l'erreur au niveau de l'unité fondée sur une représentation matricielle des appariements entre unités simples et complexes. Un aspect important de la configuration de la source est celui des problèmes de couverture. Premièrement, le problème de couverture peut être dû au fait que la cible administrative pourrait différer de la population, étant par exemple un sous-ensemble de cette dernière, et de surcroît, une couverture insuffisante ou excessive par rapport à sa propre cible peut avoir lieu en raison des retards d'enregistrement des événements. Les erreurs de couverture peuvent aussi être causées par des erreurs d'appariement des unités dans les différentes sources. Enfin, comme nous l'illustrerons au paragraphe suivant, différents ensembles d'unités de la population peuvent être couverts par une quantité différente d'information administrative.

2.3 Différents scénarios : des mesures aux concepts statistiques

Considérons maintenant les mesures enregistrées dans la source externe. Avant tout, nous pouvons considérer le cas où les mesures administratives peuvent être utilisées comme une variable auxiliaire ayant un lien prononcé avec le concept cible, mais où elles sont utilisées pour remplacer les variables cibles proprement dites. Ces conditions représentent la situation habituelle et impliquent éventuellement un usage indirect de la source. Inversement, certaines des variables enregistrées dans la source de DA peuvent être considérées comme les variables d'intérêt. Cependant, même dans ce dernier cas, elles peuvent ne pas refléter exactement les concepts statistiques. D'abord, parce qu'il peut y avoir une différence due à des définitions dissemblables ou à des différences entre les

classifications appliquées; par conséquent, une mise en correspondance ou une dérivation de nouvelles variables est nécessaire.

La nécessité de dériver de nouvelles variables statistiques est également liée au genre de relations qui existent entre les objets et les unités statistiques que nous avons décrites à la section précédente. En fait, la contiguïté serait nécessaire pour déterminer les valeurs de la ou des variables cibles sur les nouvelles unités complexes dérivées (voir Zhang, 2012, pour des renseignements plus détaillés). En outre, les variables administratives pourraient différer de la mesure statistique idéale requise parce que la méthode de collecte proprement dite produit une *erreur de mesure*. De surcroît, lorsque les mêmes concepts sont mesurés dans des sources (de DA) différentes, il faut décider comment combiner les mesures éventuellement dissemblables. Dans un système intégré de sources de DA différentes, la figure 2.3-1 représente un schéma possible d'information.

Figure 2.3-1
Schéma possible d'information

	Mesures						
	DA1	**DA2**			**DA3**		
		Y1	Y2	Y3	Y1	Y2	Y4	Y5
Unités								

Ce schéma illustre deux problèmes types que l'on observe dans un système de sources. Premièrement, comme il est mentionné à la section précédente, la couverture en ce qui concerne la population statistique et, deuxièmement, des mesures de rechange de la même variable.

En résumé, dans une situation idéale, les DA contiennent exactement les objets et les mesures nécessaires pour obtenir l'information pour la production de statistiques. Dans ce cas, la qualité des données de sortie est égale à la qualité des données d'entrée. En pratique, nous n'observerons jamais cette situation idéale et un processus de traitement doit être appliqué aux données originales, de la même façon que cela se fait pour les données recueillies dans une enquête statistique. De nouvelles erreurs peuvent être incluses durant le traitement nécessaire pour obtenir les données statistiques finales intégrées. Il est très important de prendre ces aspects en considération lorsqu'on évalue l'utilité de la ou des sources. Cette évaluation doit être effectuée en tenant compte du système complet de sources administratives.

3. Différents scénarios d'utilisation des DA : un premier aperçu des méthodes

Wallgren et Wallgren (2013) donnent un ample aperçu des méthodes et des opérations nécessaires dans un cadre fondé sur des registres. Ici, nous mentionnons brièvement les principales méthodes statistiques. Avant tout, en ce qui concerne les opérations nécessaires pour déterminer les unités, il faut procéder à un appariement et un couplage : tant pour augmenter l'information d'un registre de base, que pour vérifier s'il existe des enregistrements en double dans une source unique. Dans ce cas, le scénario des DA aura habituellement une influence sur le genre de méthode de couplage que l'on peut réellement appliquer. En fait, on procède ordinairement au couplage déterministe d'enregistrements lorsqu'un identificateur est disponible; sinon, des méthodes de couplage d'enregistrements probabiliste (Fellegi et Sunter, 1969) permettent de contourner son absence. Dans ce dernier cas, l'estimation doit tenir compte de ce processus dans les analyses subséquentes. Voir Chambers (2009) pour une estimation sans biais

quand un couplage d'enregistrements est appliqué, et Di Consiglio et Tuoto (2014) pour l'analyse de sensibilité de l'effet des erreurs de couplage d'enregistrements sur les analyses de régression linéaire et logistique.

Le cadre complexe dans un contexte de sources multiples de DA peut aussi être exploité pour concevoir un sondage. Falorsi et Righi (2012) ont proposé des stratégies de sondage optimales sous une approche assistée par modèle dans le cas d'une utilisation conjointe de données d'enquête et de DA, où les secondes ne couvrent que des sous-ensembles de la population. Pour ce qui est des mesures que l'on peut obtenir dans un cadre à sources multiples, comme nous l'avons déjà mentionné, lorsque l'on a affaire à un système complexe de sources différentes présentant divers degrés de couverture de la population cible statistique, certains problèmes particuliers se posent. D'abord, lorsqu'on intègre plusieurs sources de données, la cohérence des données devient un aspect essentiel, parce que l'intégration accroît les conflits possibles entre les données dans l'information disponible. La détermination de variables statistiques quand des mesures sont présentes dans plus d'une source peut se faire par hiérarchisation des différentes sources, en se fondant sur une évaluation de la qualité de la variable dans la source choisie. Cependant, plus récemment, la littérature a porté principalement sur l'étude des erreurs de mesure de toutes les sources concernées, sans établir *a priori* une source (ou enquête) particulière comme source de référence. Dans le contexte de la conception de questionnaires, l'utilisation de modélisations par équations structurelles (MES) linéaires pour évaluer la qualité de la mesure des variables étudiées est une approche bien établie. Elle a été étendue au contexte administratif dans Bakker (2012) et dans Scholtus et Bakker (2013). Pavlopoulos et Vermunt (2013) ont appliqué un modèle à classes latentes complexe (une chaîne de Markov cachée) pour évaluer la mesure d'une variable catégorique. Dans le dernier cas, le modèle sert aussi à obtenir une mesure, au lieu d'être utilisé uniquement pour évaluer la validité des variables observées dans la source pour le concept cible. Meijer et coll. (2013) ont fourni des prédicteurs de rechange dans un contexte similaire pour une variable cible continue, en supposant que la mesure provenant de registres administratifs peut être entachée d'erreurs dues au mauvais appariement de la source de DA et de l'enquête par sondage. La figure 2.3-1 illustre les situations d'information incomplète dans la source de DA, auquel cas le lien avec le modèle peut être utilisé : d'une part, on peut appliquer une imputation massive; sinon, on peut appliquer la pondération ou la repondération (sur d'autres registres disponibles) pour combiner les registres (avec des enquêtes), voir Renssen et coll. (2001). Quand les DA sont utilisées comme source de variables auxiliaires, les méthodes statistiques classiques peuvent être appliquées dans les différentes situations : des estimateurs fondés sur un modèle ont été explorés dans de nombreuses applications (voir ESSnet, 2011, pour l'application de DA dans des enquêtes entreprises). Kim et Rao (2012) poursuivent une approche assistée par modèle où un estimateur par projection est proposé pour combiner différentes sources. Voir aussi Luzi et coll. (2014) pour une application de l'estimation selon le projet à un cas réel.

Bibliographie

BAKKER, Bart F. M. 2010. « Micro-integration: State of the Art ». Document du *Joint UNECE/Eurostat Expert Group Meeting on Register-Based Censuses*, La Haye, Pays-Bas.

BAKKER, Bart F. M. 2012. « Estimating the validity of administrative variables », *Statistica Neerlandica*, vol. 66, nº 1, p. 8 à 17.

FELLEGI, Ivan P., et Alan B. SUNTER. 1969. « A Theory for record linkage », *Journal of the American Statistical Association*, vol. 64, nº 328, p. 1183 à 1210.

CHAMBERS, Ray. 2009. « Regression analysis of probability-linked data », *Official Statistics Research Series*, vol. 4.

Di CONSIGLIO, Loredana, et Tiziana TUOTO. 2014. « When adjusting for bias due to linkage errors: a sensitivity analysis », *Proceedings of the European Conference on Quality in Official Statistics (Q2014)*, Vienne, du 3 au 5 juin.

ESSnet on use of administrative and account data in business statistics. 2011. Livrables de WP3 et WP4, http://www.cros-portal.eu/content/admindata-sga-3.

FALORSI, Pietro Demetrio, Stefano FALORSI et Paolo RIGHI. 2012. « Optimal Survey Strategies in the Multivariate Multidomain Context WithMultiple Sources of Administrative Information Covering Different Population Subsets », dans : Electronic proceedings, of the Seminar on New Frontiers for Statistical Data Collection, Commission économique des Nations Unies pour l'Europe, http://www.unece.org/stats/documents/2012.10.coll.html.

GROVES, Robert M., Floyd J. FOWLER Jr., Mick P. COUPER, James M. LEPKOWSKI, Eleanor SINGER et Roger TOURANGEAU. 2004. *Survey methodology*, New York: Wiley.

KIM J. K., et J. N. K. RAO. 2012. « Combining data from two independent surveys: a model assisted approach », *Biometrika*, vol. 99, nº 1, p. 85 à 100.

LUZI, O., U. GUARNERA et Paolo RIGHI. 2014. « The new multiple-source system for Italian Structural Business Statistics based on administrative and survey data », *Proceedings of the European Conference on Quality in Official Statistics (Q2014)*, Vienne, du 3 au 5 juin.

MEIJER, Erik, Susann ROHWEDDER et Tom WANSBEEK. 2012. « Measurement error in earnings data: Using a mixture model approach to combine survey and register data », *Journal of Business & Economic Statistics*, vol. 30, nº 2, p. 191 à 201.

PAVLOPOULOS, Dimitris, et Jeroen K. VERMUNT. 2013. « Mesure de l'emploi temporaire. Les données d'enquête ou de registre disent-elles la vérité? », *Techniques d'enquête*, sous presse.

RENSSEN, Robbert H., A.H. KROESE et Ad WILLEBOORDSE. 2001. « Aligning estimates by repeated weighting », document de recherche, BPA-nº 491-01-TMO, Statistics Netherlands, Heerlen.

ROYCE, Don. 2013. « A Survey of International Frameworks for the Statistical Use of Administrative Data », Secrétariat des données administratives, Statistique Canada.

SCHOLTUS, Sander, et Bart F. M. BAKKER. 2013. « Estimating the Validity of Administrative and SurveyVariables by Means of Structural Equation Models », *Actes du NTTS 2013 disponibles en ligne.*

Commission économique des Nations Unies pour l'Europe. 2011. « Using Administrative and Secondary Sources for Official Statistics A Handbook of Principles and Practices ».

WALLGREN, Anders, et Britt WALLGREN. 2014. *Register based statistics: Administrative Data for Statistical Purposes*, New York: Wiley.

ZHANG, Li-Chun. 2011. « A Unit-Error Theory for Register-Based Household Statistics », *Journal of Official Statistics*, vol. 27, nº 3, p. 415 à 432.

ZHANG, Li-Chun. 2012. « Topics of statistical theory for register-based statistics and data integration », *Statistica Neerlandica*, vol. 66, nº 1, p. 41 à 63.