Recueil du Symposium 2016 de Statistique Canada
Croissance de l'information statistique : défis et bénéfices

Trouver une aiguille dans une botte de foin : les fondements théoriques et empiriques de l'évaluation du risque de divulgation pour des microdonnées contextualisées

Kevin T. Leicht et Kristine Witkowski[1]

Résumé

Cette communication décrit divers facteurs qui posent un problème lorsque l'on évalue le risque de divulgation de microdonnées contextualisées, ainsi que certaines des étapes empiriques qui sont comprises dans leur évaluation. À partir d'ensembles synthétiques de répondants d'enquête, nous illustrons comment différents postulats modulent l'évolution du risque lorsque l'on tient compte : 1) des probabilités estimées que des régions géographiques non identifiées soient représentées dans une enquête; 2) du nombre de personnes dans la population qui partagent les mêmes identificateurs personnels et contextuels qu'un répondant; et 3) de l'ampleur prévue de l'erreur de couverture dans les chiffres de population du recensement et les fichiers existants qui fournissent des données d'identification (p. ex., le nom, l'adresse).

Mots-clés : confidentialité; diffusion; risque de divulgation

1. Introduction

De nombreux problèmes dans le domaine des sciences sociales contemporaines se prêtent à une analyse dans laquelle les personnes à l'étude sont placées en contexte, défini spatialement comme une rue, un îlot, une ville, un comté ou une autre unité spatiale. Les fournisseurs de données ont trouvé deux façons de présenter ces renseignements : identifier l'unité spatiale, afin que les utilisateurs des données puissent coupler les données contextuelles appropriées eux-mêmes, ou fusionner les données contextuelles, en ajoutant bien sûr les caractéristiques de l'unité spatiale où le sujet vit. Dans le deuxième cas, l'enregistrement pour chaque personne comprend les caractéristiques de cette personne (p. ex., l'âge du répondant) et les caractéristiques spatiales (p. ex., proportion de la population qui est pauvre dans le quartier du répondant).

Une raison de fournir les données contextuelles dès le départ, plutôt que d'identifier l'unité spatiale, est que cela rend plus difficile l'identification de l'unité spatiale où vit le répondant d'enquête (Armstrong, Rushton et Zimmerman, 1999). Toutefois, les données contextuelles proprement dites peuvent représenter suffisamment de renseignements pour être uniques au niveau géographique. Le fait de connaître le comté (ou encore le secteur de recensement ou le groupe d'îlots) ne signifie pas que l'on puisse identifier les personnes. Il s'agit uniquement d'un point de départ. Étant donné que les fichiers de microdonnées comprennent habituellement des mesures individuelles et contextuelles, une évaluation complète du risque nécessite une approche intégrée, qui tient compte à la fois des caractéristiques d'identification des répondants aux enquêtes et des lieux où ils habitent (lieux de résidence). La présente étude aide à jeter les premières bases d'une telle évaluation, grâce à son approche dite de « l'aiguille dans une botte de foin » en matière de divulgation et à un examen des préoccupations méthodologiques connexes.

Grâce à une approche analytique faisant le pont entre deux niveaux, notre étude éclaire la conception de fichiers de données à grande diffusion constitués d'enregistrements-personne comprenant des mesures contextuelles provenant

[1]Kevin T. Leicht, Département de sociologie, Université de l'Illinois à Urbana-Champaign, 3120 Lincoln Hall MC-454, 702 South Wright Street, Urbana, Illinois, 61801. (kleicht@illinois.edu); Kristine Witkowski, Inter-University Consortium for Political and Social Research (ICPSR), Institute for Social Research, Université du Michigan, C.P. 1248, Ann Arbor, Michigan 48106-1248. Courriel : kwitkow@umich.edu.

de comtés, de secteurs de recensement et de groupes d'îlots. Au moyen d'un fichier de données d'essai synthétiques, nous illustrons comment la ré-identification des personnes est affectée par l'agrégation des niveaux géographiques dans des ensembles de données d'aspect similaire. Nous évaluons en outre une dimension du risque associée aux microdonnées contextualisées, à savoir l'identification des répondants d'enquêtes dont les caractéristiques personnelles (p. ex., le sexe, l'âge et la race) se retrouvent rarement dans les populations qui partagent les mêmes contextes.

À partir d'unités géographiques, et de leurs « contextes agrégés d'aspect similaire », comme unités d'analyse, du nombre de personnes dans la population comportant un ensemble distinct de caractéristiques personnelles comme résultat d'intérêt et d'indicateurs du risque sous-jacent, nous décrivons en détail la complexité des modèles de ré-identification, en évaluant la probabilité que de jeunes hommes adultes de race blanche et noire soient localisés avec précision dans des bottes de foin géographiques reconstituées, compte tenu : 1) de la taille de la population totale des contextes agrégés; 2) de la quantité d'erreurs dans les comptes de la population; et 3) des coûts (ou temps) de recherche différents découlant de contextes spatialement dispersés.

2. Définition du risque

Afin de contribuer à expliquer les fondements conceptuels des données anonymisées, nous aurons recours à l'expression française « chercher une aiguille dans une botte de foin », qui désigne une chose difficile à trouver parce qu'elle est cachée dans un ensemble plus grand d'objets (Cambridge University Press, 2003). Un répondant d'enquête (« l'aiguille ») est une personne déterminée dans un fichier de données à grande diffusion, qui affiche un ensemble particulier de caractéristiques individuelles qui peuvent facilement être identifiées par un intrus (sexe, âge, race, origine ethnique, état matrimonial ou niveau de scolarité). Ce répondant d'enquête est aussi membre d'un groupe de personnes (« la botte de foin ») dans la population qui partagent les mêmes caractéristiques individuelles et contextuelles d'identification. Comme c'est le cas pour la recherche d'une aiguille dans une botte de foin, la probabilité qu'un répondant soit ré-identifié correctement est considérée comme faible lorsqu'un intrus doit effectuer, dans un fichier d'identification existant, une recherche parmi un nombre suffisamment grand de personnes qui partagent les mêmes caractéristiques.

À l'étape initiale de notre expérience de ré-identification, nous présumons qu'un intrus, soit la personne voulant identifier un répondant en particulier à partir du jeu de données, dépendra de trois ensembles de variables de composition pour évaluer la difficulté de localiser avec précision des répondants. Le sexe, l'âge, la race et l'origine ethnique des répondants d'enquête définissent la composition des bottes de foin dans lesquelles ces « aiguilles » sont cachées. L'enrichissement des fichiers de microdonnées au moyen d'informations géographiques et de variables décrivant les contextes d'emplacements enquêtés inconnus (p. ex., pourcentage de personnes nées à l'étranger, personne vivant dans une région métropolitaine ou non), limite encore plus la localisation de ces répondants. Un troisième ensemble de variables de composition, qui prennent la forme de « scores de difficultés à dénombrer » (*hard-to-count scores*) et de taux de location du logement relié à la race, sert à évaluer l'exactitude des tailles des bottes de foin estimées et la quantité probable d'erreurs de couverture dans les fichiers existants. La corrélation entre ces trois ensembles de mesures contextuelles est considérable. C'est ce chevauchement dans les mesures qui déterminent les probabilités de toutes les ré-identifications possibles des répondants. Ce chevauchement reflètera à quel point le contexte agrégé est : isolé au niveau résidentiel, unique dans ses caractéristiques et difficile à recenser.

À partir d'une vaste gamme de questions sociales scientifiques, nous avons sélectionné cinq variables contextuelles devant être représentées dans notre ensemble de données d'essai : 1) % de personnes de race blanche non hispaniques; 2) % de personnes nées à l'étranger; 3) % de personnes vivant dans la pauvreté; 4) % d'unités de logement occupées par le propriétaire; et 5) % de gens sans emploi dans la population active. Par suite de nos travaux antérieurs (Witkowski, 2016), nous avons appliqué une technique de « masquage non perturbateur », avec comme objectif d'éclairer de façon large la conception d'ensembles de données comprenant des données contextuelles à diverses échelles spatiales. Après regroupement des valeurs extrêmes supérieures et inférieures de ces variables continues, afin de cacher les valeurs aberrantes, nous avons recodé les mesures contextuelles en dix catégories espacées de 10 % (p. ex., 0 à 9 %, 10 à 19 %, 20 à 29 %, 30 à 39 %, 40 à 49 %, 50 à 59 %, 60 à 69 %, 70 à 79 %, 80 à 89 %, 90 à 100 %). Les valeurs aberrantes ont été identifiées comme étant celles se situant parmi les 0,5 % valeurs les plus élevées et les 0,5 % valeurs les plus petites de chaque répartition de variables (Zayatz, 2005), selon des répartitions

géographiques définies par le statut métropolitain des régions géographiques. Les variables contextuelles sont recodées en catégories agrégées, selon leurs valeurs absolues (c.-à-d., recodage absolu).

Pour chaque unité géographique échantillonnée (1785 comtés, 8947 secteurs de recensement et 10 478 groupes d'îlots) et chaque unité d'aspect similaire dans un contexte agrégé (315 comtés, 2280 secteurs de recensement et 3090 groupes d'îlots), nous compilons trois ensembles de données concernant la population visée par l'enquête, ainsi que la taille et la composition de la population totale. Tout d'abord, nous dénombrons le nombre total de répondants dans chaque emplacement et chaque contexte. Puis nous corrigeons la taille de la population totale des régions individuelles et agrégées. Enfin, nous corrigeons le nombre total de personnes qui ont un ensemble sélectionné de caractéristiques personnelles, en fournissant des estimations de la taille de la botte de foin pour chaque emplacement et chaque contexte. Étant donné que les analyses pour une vaste gamme de bottes de foin dépassent la portée du présent document, nous simplifions l'étude en évaluant uniquement une botte de foin majoritaire et une botte de foin minoritaire comprenant les hommes âgés de 20 ans qui sont : 1) de race blanche non hispaniques seulement, ou 2) de race noire ou afro-américaine seulement. Grâce à cette approche, nous sommes en mesure de déterminer comment le statut de minorité influence la ré-identification des répondants, en maintenant constants leur sexe et leur âge.

3. Que faisons-nous maintenant?

Dans notre exercice d'analyse, nous 1) construisons un fichier de microdonnées constitué d'un échantillon synthétique unique de répondants d'enquête, en reliant l'information contextuelle aux enregistrements-personne; 2) déterminons les unités géographiques qui ressemblent au lieu de résidence d'un répondant, et ce, à partir des données contextuelles disponibles pour les comtés, les secteurs de recensement et les groupes d'îlots; 3) regroupons la population de base, la botte de foin, l'erreur de couverture et la dispersion spatiale pour les unités géographiques et les contextes agrégés, en joignant ces estimations de recherche aux enregistrements-personne; et 4) calculons des statistiques sommaires pour le fichier de microdonnées d'essai pour bien indiquer la répartition des répondants.

L'unité d'analyse la plus importante est la personne échantillonnée ou répondante. Même si les répondants d'enquêtes synthétiques reçoivent un ensemble de caractéristiques contextuelles qui peuvent permettre de retracer leur lieu de résidence, on ne leur attribue pas de caractéristiques personnelles. Ainsi, nous ne connaissons pas le sexe, l'âge ni la race d'un répondant en particulier. Toutefois, nous connaissons la taille des bottes de foin et de la population totale dans chaque unité géographique et contexte agrégé, ainsi que la proportion de répondants qui ont été tirés de ces régions. En regroupant ces données, nous pouvons établir la probabilité qu'un répondant donné ait un ensemble de caractéristiques personnelles définies par la botte de foin d'intérêt (dans ce cas, les hommes âgés de 20 ans de race blanche non hispaniques ou afro-américaine).

4. Résultats

Nous menons deux ensembles d'analyses qui évaluent le rôle possible de la taille de la botte de foin, de l'erreur de couverture et de la dispersion dans l'évaluation du risque de divulgation pour les microdonnées contextualisées. Pour le premier ensemble d'analyses, présentées dans le tableau 1, nous déterminons comment le processus de ré-identification est modifié lorsque nous joignons des données contextuelles plutôt que d'identifier directement les unités géographiques. Nous illustrons comment les caractéristiques de la botte de foin, précédemment définies, sont fonction d'unités géographiques peu ou densément peuplées et comment l'agrégation de régions géographiques d'aspect similaire réduit le risque de divulgation en augmentant l'incertitude et les coûts de recherche associés à l'identification des répondants. À des fins d'illustration, nous produisons des analyses distinctes pour les répondants d'enquêtes cachés dans les unités géographiques individuelles et les contextes agrégés qui sont très peuplés (« denses ») ou moins peuplés (« épars »), définis comme étant ceux dont la population est supérieure ou inférieure à 100 000 habitants respectivement.

À partir du tableau 1, les analyses révèlent que la divulgation de données contextuelles, plutôt que l'identification directe de lieux, comporte des ramifications importantes pour le risque de divulgation. Les probabilités qu'un répondant d'enquête se trouve dans une région densément peuplée augmentent de façon marquée lorsque les données contextuelles sont jointes aux enregistrements individuels. La plupart des répondants d'enquête (96 %) se trouvent

dans des contextes densément peuplés déterminés à partir de comtés, tandis qu'une majorité (59 à 67 %) vit dans des contextes très peuplés agrégés à partir d'unités géographiques à plus petite échelle (secteurs ou îlots). Pour toutes les échelles géographiques, la taille de la population totale (en moyenne) peut monter jusqu'à plus de 2,4 millions de personnes dans les contextes agrégés à forte densité, tandis que dans les contextes à faible densité, la population est d'au moins 29 000 personnes.

Pour toutes les unités géographiques, l'accumulation de populations de base dans des contextes agrégés donne lieu à des bottes de foin significativement plus grosses. Les contextes denses comptent au moins 11 000 hommes blancs non hispaniques âgés de 20 ans, tandis que les contextes épars comptent au moins 125 membres de cette sous-population majoritaire (en moyenne). Comme il fallait s'y attendre, la sous-population minoritaire de jeunes hommes noirs est associée généralement à des bottes de foin beaucoup plus petites que celles de leurs homologues majoritaires. Toutefois, l'agrégation des régions géographiques similaires augmente la taille des bottes de foin qui découlent des unités à petite échelle, au point où pas moins de 59 jumeaux se trouvent dans des contextes épars et jusqu'à 467 dans des contextes denses. Les bottes de foin minoritaires sont même plus importantes pour les contextes denses calculés à partir de comtés, avec une taille moyenne de 2641 membres cibles.

Lorsqu'on examine la modification des tendances de l'erreur de couverture découlant de l'agrégation, nous voyons bien la sélection de régions géographiques uniques dans des contextes comptant moins de 100 000 personnes. Avec deux fois plus de difficultés d'être bien dénombrées, de 47 à 71 % des régions géographiques d'aspect similaire dans des contextes épars sont probablement sous-dénombrées. La concentration des populations difficiles à dénombrer à l'intérieur de contextes relativement peu peuplés et uniques est démontrée davantage par les probabilités exceptionnellement faibles de sur-dénombrement (c.-à-d. 1 à 5 %). Avec des taux nationaux de sous-dénombrement allant de 1 à 2 %, le nombre disproportionné de secteurs de recensement difficiles à dénombrer pourrait se révéler un obstacle important à la ré-identification de répondants par un intrus.

Toutefois, la protection offerte par l'erreur de couverture et la transposition de scores de difficultés à dénombrer (*hard-to-count scores*) en taux réels de sous-dénombrement sont étroitement liées. Tout d'abord, on voit bien que le niveau de difficulté du dénombrement dont rendent compte les scores de difficultés à dénombrer ne reflète pas suffisamment la complexité des modèles d'isolement résidentielle entourant les groupes raciaux et les groupes ethniques sous-représentés. Peu importe l'échelle spatiale, la densité de la population et le détail géographique des contextes, les ménages de race blanche non hispaniques sont en fait plus faciles à dénombrer que les autres populations du même secteur de recensement, alors que les ménages de race afro-américaine ou de race noire sont en fait plus difficiles à dénombrer. Comparativement aux autres ménages du même secteur de recensement, parmi les ménages qui louent leur logement de 2 à 11 % de plus sont de race noire, alors que parmi les ménages qui sont propriétaires de leur logement de 13 à 17 % de plus sont blancs non hispaniques. Compte tenu de cette différence en matière d'accès à la propriété, l'erreur de couverture a tendance à rétrécir l'écart entre la minorité et la majorité par rapport au risque de divulgation. Toutefois, la probabilité qu'une petite botte de foin minoritaire ait été extrêmement bien cachée peut être réduite dans des contextes comportant de petites populations de base. Par exemple, la population totale d'un contexte épars (construit à partir de secteurs de recensement) est typiquement de 34 668 personnes. Même si 71 % de ces secteurs de recensement sont probablement sous-dénombrés, il serait peu probable que cette petite population comprenne un nombre exorbitant de membres cachés.

Même si le nombre possible des bottes de foin cachées dans des contextes épars n'est peut-être pas suffisamment important, les mêmes taux élevés d'erreur pourraient indiquer l'absence de couverture dans les fichiers d'identification existants. L'intrus devra probablement effectuer des activités de recherche supplémentaires pour identifier les répondants dans ces régions. Même s'il est impossible d'associer un coût à ces activités, nous savons que la recherche concernant les répondants dans des contextes agrégés comptant moins de 100 000 personnes couvrirait en moyenne : 2 comtés, 7503 milles carrés et 2 hectares; 9 secteurs de recensement, 581 milles carrés et 5 états; ou 23 groupes d'îlots, moins de 1 mille carré et 10 états. Si un intrus souhaitait prendre un risque calculé, il pourrait réduire son temps de recherche en laissant de côté les unités géographiques qui sont peu susceptibles de compter un répondant d'enquête. Les économies pourraient être prononcées, particulièrement pour les populations minoritaires résidant dans des contextes agrégés épars, où 58 à 78 % des régions géographiques d'aspect similaire ne comptent aucun homme de race noire âgé de 20 ans.

Tableau 1. Agrégation d'unités géographiques échantillonnées dans des contextes d'aspect similaire, pondérée pour rendre compte de la répartition spatiale des répondants d'enquête (N = 1785; 8947; et 10 478 des comtés, secteurs de recensement et groupes d'îlots échantillonnés respectivement; N = 315; 2280; et 3090 des contextes agrégés échantillonnés fondés sur les comtés, les secteurs de recensement et les groupes d'îlots respectivement; N = 11 562 répondants d'enquêtes synthétiques)

	Répondants d'enquête résidant dans un											
	Comté comme base contextuelle				Secteur de recensement comme base contextuelle				Groupe d'îlots comme base contextuelle			
	Unité géographique		Contexte agrégé		Unité géographique		Contexte agrégé		Unité géographique		Contexte agrégé	
	<100K	100K+	<100K	100K+	<100K	100K+	<100K	100K+	<100K	100K+	<100K	100K+
Proportion de répondants dans le contexte	0,34	0,66	0,04	0,96	1,00	0,00	0,33	0,67	1,00	0,00	0,41	0,59
Taille de la population totale à l'intérieur du contexte échantillonné	41 756	955 163	51 233	3 469 881	4 898	---	34 668	2 484 499	1 613	---	29 243	2 489 206
Taille moyenne de la sous-population des bottes de foin												
Personnes de race blanche non hispaniques seulement, hommes âgés de 20 ans	275	3 128	264	17 715	25	---	155	11 808	8	---	126	11 155
Personnes de race afro-américaine ou de race noire seulement, hommes âgés de 20 ans	33	1 222	130	2 641	5	---	76	467	2	---	59	318
Proportion de secteurs de recensement dans les unités géographiques échantillonnées et les contextes agrégés qui sont probablement sous-dénombrés	0,17	0,28	0,47	0,23	0,29	---	0,71	0,08	0,29	---	0,59	0,07
Proportion des secteurs de recensement dans les régions géographiques échantillonnées et les contextes agrégés qui sont probablement sur-dénombrés	0,18	0,28	0,05	0,25	0,19	---	0,01	0,28	0,19	---	0,03	0,32
Différence dans la proportion de locataires												
Personnes de race blanche non hispaniques seulement (moins les autres)	-0,16	-0,15	-0,13	-0,16	-0,16	---	-0,15	-0,17	-0,16	---	-0,16	-0,16
Personnes de race afro-américaine ou de race noire seulement (moins les autres)	0,02	0,11	0,04	0,09	0,10	---	0,09	0,10	0,10	---	0,11	0,11
Nombre moyen d'unités géographiques ressemblant aux comtés échantillonnés	117	14	---	---	381	---	---	---	1 098	---	---	---
Nombre moyen d'unités géographiques dans le contexte agrégé[1]	---	---	2	51	---	---	9	566	---	---	23	1 838
Nombre moyen de milles carrés de superficie d'unités géographiques échantillonnées et de contextes agrégés	1 848	1 272	7 503	44 691	113	---	581	39 759	0,03	---	0,44	32,52
Nombre moyen d'états dans le contexte agrégé	---	---	2	12	---	---	5	31	---	---	10	37
Proportion d'unités géographiques comptant une sous-population (en contexte)												
Personnes de race blanche non hispaniques seulement, hommes âgés de 20 ans	1,00	1,00	0,99	1,00	0,95	---	0,98	1,00	0,88	---	0,96	1,00
Personnes de race afro-américaine ou de race noire seulement, hommes âgés de 20 ans	0,79	1,00	0,79	1,00	0,59	---	0,94	1,00	0,38	---	0,92	1,00

Nota : Sont exclues des analyses les unités géographiques n'ayant pas de population, ce qui donne lieu à 3141 comtés, 5 174 secteurs de recensement et 208 125 groupes d'îlots considérés à partir de la population des unités géographiques.

Nota : L'ensemble de données au niveau du comté, du secteur de recensement et du groupe d'îlots comprend cinq mesures contextuelles de : 1) % de personnes de race blanche non hispaniques; 2) % de personnes nées à l'étranger; 3) % de personnes vivant dans la pauvreté; 4) % d'unités de logement occupées par le propriétaire et 5) % de gens sans emploi dans la population active, recodées en catégories de 10 % (c.-à-d. 0 à 9 %, 10 à 19 %, 20 à 29 %, 30 à 39 %, 40 à 49 %, 50 à 59 %, 60 à 69 %, 70 à 79 %, 80 à 89 %, 90 à 100 %). Cet ensemble de données permet aussi d'identifier directement la situation de RSM des unités géographiques : 1) RSM de 1 million ou plus, 2) RSM de moins de 1 million et 3) non-RSM.

Nota : Pondérées pour rendre compte de la répartition spatiale des répondants d'enquête, les valeurs moyennes sont tirées d'ensembles d'unités géographiques et de contextes agrégés qui comptent une population totale dont la taille est de moins de 100 000 personnes ou de plus de 100 000 personnes, montrant : 1) le nombre d'unités géographiques ressemblant à l'unité géographique échantillonnée; 2) le nombre d'unités géographiques dans le contexte agrégé; 3) la taille de la population totale dans une unité géographique ou un contexte agrégé (c.-à-d. répartie entre les unités géographiques d'aspect similaire); 4) la proportion de secteurs de recensement dans des unités géographiques individuelles ou des unités géographiques dans un contexte agrégé qui comportent un score de difficultés à dénombrer « élevé » ou « faible »; 5) le nombre d'états dans un contexte agrégé; et 6) le nombre de milles carrés de superficie dans une unité géographique ou un contexte agrégé (c.-à-d. répartis entre toutes les unités géographiques d'aspect similaire).

Nota : Des détails concernant la construction des scores de difficultés à dénombrer sont fournis par Bruce et Robinson (2003). Les scores au niveau du secteur de recensement sont attribués à des groupes d'îlots intégrés et sont agrégés en estimations au niveau du comté et du contexte. Des niveaux « élevés » et « faibles » d'erreur dans les données existantes sont reflétés dans les quartiles supérieurs et inférieurs des scores de difficultés à dénombrer, à partir des répartitions des secteurs de recensement.

Nota : « 0,00 » indique une proportion de répondants dans des contextes supérieurs à 0, mais inférieurs à 0,05, alors que « --- » indique une valeur absolue de zéro. « --- » indique aussi qu'une catégorie particulière de taille de population ne s'appliquait pas à un ensemble de régions géographiques et, par conséquent, que les statistiques connexes n'ont pas été calculées.

[1] Le faible nombre d'unités géographiques dans les contextes agrégés comptant moins de 100 000 personnes rend compte de la sélection d'un ensemble limité d'unités « d'aspect similaire » relativement peu peuplées dans ces contextes.

5. Conclusion

Notre étude montre que les microdonnées contextualisées peuvent constituer une méthode valable pour diffuser de façon sécuritaire des données géographiques riches en information. Cette constatation est particulièrement pertinente dans le cas des données contextuelles au niveau du comté, où seulement 4 % des répondants d'enquête se trouvent dans des contextes agrégés comptant moins de 100 000 personnes. Par contre, d'autres travaux devraient être effectués pour bien comprendre les répercussions des hypothèses qui sous-tendent les seuils de taille de population. Même si nos résultats montrent le rôle potentiellement important de l'erreur de couverture pour assurer l'anonymat des répondants, d'autres études seraient nécessaires pour créer et analyser des données qui rendent mieux compte de la variation spatiale du sous-dénombrement de différentes sous-populations. De plus, cette étude était limitée à un seul ensemble de données contextuelles et à deux bottes de foin émergentes. Une évaluation comprenant un ensemble complet de bottes de foin, ainsi qu'un ensemble élargi de données contextuelles, qui varierait au chapitre de la composition et des détails de la mesure, serait nécessaire et plus exhaustive.

Bibliographie

Armstrong, Marc P., Gerard Rushton, and Dale L. Zimmerman. 1999. "Geographically Masking Health Data to Preserve Confidentiality." *Statistics in Medicine* 18: 497-525.

Cambridge University Press. (2003). *Cambridge Dictionary of American Idioms*. Cambridge University Press: Cambridge, UK.

Witkowski, Kristine M. (2016). "Disclosure Risk of Contextual Data: The Role of Spatial Scale, Identified Geography, and Measurement Detail in Public-Use Files." Submitted to *Journal of Official Statistics*.

Zayatz, Laura. (2005). Disclosure Avoidance Practices and Research at the U.S. Census Bureau: An Update. Revised August 31, 2005: Research Report Series (Statistics #2005-06). Washington, DC: Statistical Research Division, U.S. Census Bureau.

U.S. Census Bureau, Population Division. 2002. Census 2000 PHC-T-3. Ranking Tables for Metropolitan Areas: 1990 and 2000 (Table 3: Metropolitan Areas Ranked by Population). Last Revised: July 31, 2002
< Web Page: http://www.census.gov/population/www/cen2000/phc-t3.html>
< Direct Link: http://www.census.gov/population/cen2000/phc-t3/tab03.xls>

U.S. Census Bureau, Geography Division, Cartographic Products Management Branch. 2005. *Cartographic Boundary Files*. Last Revised: August 24, 2005.
<http://www.census.gov/geo/www/cob/index.html>

U.S. Census Bureau, Population Division. 2006a. *Geographic Relationship Files: 1999 MA to 2003 CBSA* (Excel file). Last Modified: August 18, 2006.
<http://www.census.gov/population/www/estimates/metroarea.html>
<Direct Link: http://www.census.gov/population/www/estimates/CBSA03_MSA99.xls>

U.S. Census Bureau. 2006b. 2000 *Census of Population and Housing, Summary File 1 (Matrices P1)* generated by Kristine Witkowski; using American FactFinder; <http://factfinder.census.gov>; (6 November 2006).

U.S. Department of Commerce, Bureau of the Census. CENSUS OF POPULATION AND HOUSING, 2000a [UNITED STATES]: SUMMARY FILE 1 SUPPLEMENT, STATES [Computer file]. ICPSR release. Washington, DC: U.S. Dept. of Commerce, Bureau of the Census [producer], 2003. Ann Arbor, MI: Inter-university Consortium for Political and Social Research, [distributor], 2003.

U.S. Department of Commerce, Bureau of the Census, and Inter-university Consortium for Political and Social Research. CENSUS OF POPULATION AND HOUSING, 2000b [UNITED STATES]: SELECTED SUBSETS FROM SUMMARY FILE 3 [Computer file]. 2nd ICPSR ed. Washington, DC: U.S. Dept. of Commerce,

Bureau of the Census, and Ann Arbor, MI: Inter-university Consortium for Political and Social Research [producers], 2004. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor],