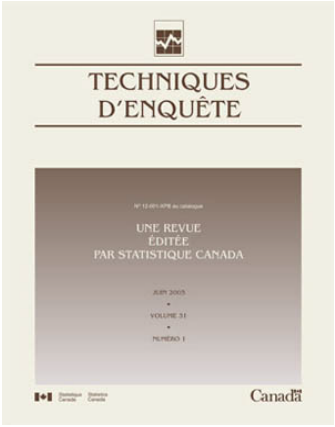
Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d’enquête

Contrôle de la divulgation statistique et avancées dans la protection officielle des renseignements : à la mémoire de Chris Skinner



par Natalie Shlomo

Date de diffusion : le 30 juin 2023

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-514-283-9350

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Techniques d'enquête, juin 2023
Vol. 49, nº 1, pp. 5-27
Statistique Canada, nº 12-001-X au catalogue

Contrôle de la divulgation statistique et avancées dans la protection officielle des renseignements : à la mémoire de Chris Skinner

Natalie Shlomo[1]

Résumé

Je donnerai un aperçu de l'évolution de la recherche sur le contrôle de la divulgation statistique (CDS) dans les dernières décennies et de son adaptation à la révolution des données à l'aide de définitions plus officielles de la confidentialité. Je soulignerai les nombreux apports de Chris Skinner aux domaines de recherche sur le CDS. Je passerai en revue ses recherches de pionnier en commençant par ses travaux des années 1990 sur la diffusion de microdonnées d'échantillon du recensement au Royaume-Uni. De ces recherches sont nées diverses études où l'on a mesuré le risque de réidentification dans les microdonnées d'enquête au moyen de modèles probabilistes. Je porterai principalement mon attention à traiter d'autres aspects des recherches en CDS de Chris. Chris Skinner a reçu le prix Waksberg en 2019 et n'a malheureusement jamais eu l'occasion de présenter son discours Waksberg au Symposium international sur les questions de méthodologie de Statistique Canada. Le présent article suivra le canevas préparé par Chris en prévision de cette allocution.

Mots-clés : Confidentialité différentielle; modèles de confidentialité; révolution des données; risque de réidentification.

1. Introduction

Une séance commémorative spéciale a eu lieu en l'honneur de Chris Skinner le 22 octobre 2021 au Symposium international de 2021 sur les questions de méthodologie de Statistique Canada. Au cours de celle-ci ont été entendus de nombreux témoignages émouvants d'amis et de collègues venus célébrer sa vie et ses réalisations. Chris a été lauréat du prix Waksberg en 2019 et projetait d'aller au Symposium international de 2019 sur les questions de méthodologie pour présenter son discours. Malheureusement, sa maladie a empiré et il est décédé le 21 février 2020. Durant la séance commémorative, tout comme dans le présent article, j'ai décrit l'évolution de la recherche sur le contrôle de la divulgation statistique (CDS), en mettant l'accent sur les contributions de Chris Skinner au domaine. Les grandes lignes de l'exposé et de l'article sont tirées d'une série de notes que Chris avait rédigées en prévision de son discours pour le prix Waksberg de 2019 qui m'ont été remises par son fils, Tom Skinner.

J'ai eu le grand privilège de travailler avec Chris, lorsqu'étudiante au doctorat, à l'élaboration de la théorie sur l'estimation du risque de réidentification présentée à la section 4.1, et plus tard, en tant que collègue à l'Université de Southampton, où nous avons continué à réaliser des progrès dans d'autres volets du CDS. Un excellent aperçu de la personne qu'était Chris et de sa contribution à la statistique sociale et à la méthodologie d'enquête est donné dans son entrevue publiée dans la Revue Internationale de Statistique (Haziza et Smith, 2019).

J'aborde les premiers progrès concernant le CDS à la section 2 et je passe aux défis auxquels nous faisons face aujourd'hui en raison de la révolution des données à la section 3. À la section 4, je décris les apports

1. Natalie Shlomo, Social Statistics Department, University of Manchester, Royaume-Uni. Courriel : natalie.shlomo@manchester.ac.uk.

de Chris et ses travaux de pionnier sur le CDS. À la section 5, je présente les recherches en cours et à venir sur le CDS et la protection des données, ainsi que l'une des dernières contributions de Chris, qui consistait à intégrer la définition de la confidentialité différentielle en informatique dans la trousse à outils de CSD des organismes gouvernementaux. À la section 6, je discute en conclusion du rôle joué par les recherches de Chris dans les statistiques gouvernementales et sociales et la méthodologie d'enquête.

2. Premières avancées et histoire du CDS

La sensibilisation du public à la confidentialité et à la vie privée est apparue dans les années 1960, entraînant le début d'une opposition publique à la collecte de données, en particulier pour les recensements en Europe après la Seconde Guerre mondiale. Il y a eu, par exemple, de nombreuses objections à la collecte de renseignements sur la population des Pays-Bas et le dernier recensement classique dans ce pays a eu lieu en 1971. Cette opposition a obligé les organismes gouvernementaux à répondre aux préoccupations populaires en matière de respect de la vie privée et de confidentialité (Dunn, 1967). Il en a aussi été question dans d'autres premiers travaux de Barabba (1975), de Cox (1976), de Fellegi (1972) et de Dalenius (1974). Fellegi (1972, page 8) a écrit : [traduction] « Les instituts nationaux de statistiques (INS) vivent de la bonne volonté et de la confiance du public et, par conséquent, le maintien de cette confiance est littéralement une question de vie ou de mort pour eux ». S'appuyant sur des recherches menées en Suède, Dalenius (1977) a été l'un des premiers à définir officiellement et à préciser un cadre pour le CDS de la façon suivante : [traduction] « Une partie non autorisée ne devrait pas pouvoir apprendre quoi que ce soit sur une personne par la diffusion d'une statistique calculée à partir de la base de données D, $f(D)$, qui ne peut être appris sans avoir accès à $f(D)$ ».

Les travaux de Dalenius et d'autres ont permis d'établir le cadre de recherche-développement sur le CDS au sein des organismes gouvernementaux, ainsi que de créer en bonne et due forme des conseils de gouvernance de la diffusion des données statistiques. La recherche effectuée aux États-Unis comprenait, par exemple, le Subcommittee on Disclosure-Avoidance Techniques, fondé en 1976 par le Federal Committee on Statistical Methodology, et appuyé par la division des politiques statistiques de l'Office of Management and Budget, comme l'ont indiqué Jabine, Michael et Mugge (1977) [voir aussi le rapport de 1978 et l'annexe A sur les pratiques d'évitement de la divulgation statistique dans certains organismes fédéraux]. Cette annexe en cinq sections présente des recommandations sur le concept de CDS, quoi diffuser, les techniques de prévention de la divulgation, les effets de la divulgation sur les sujets concernés et les utilisateurs, et les besoins en recherche-développement. Par ailleurs, des règles générales ont été mises en place en ce qui concerne, par exemple, l'interdiction de publier des données tirées de régions comptant moins de 100 000 personnes.

De nouveaux travaux dans les années 1980 ont mis l'accent sur le CDS en ce qui concerne les produits tirés de données d'enquête, car à l'origine on croyait à tort que l'échantillonnage protégeait contre les risques de divulgation (Dalenius, 1988). Paass (1988) est un des premiers à avoir estimé la fraction d'enregistrements identifiables dans des microdonnées d'enquête et à avoir tenu compte de l'échantillonnage, de la

méthode CDS du bruit additif et de la connaissance déjà acquise dans le cas présumé d'une attaque sur les données. Dans son article, Paass (1988) a écrit : [traduction] « Là où il y a déjà une vaste connaissance, l'obligation de protéger la vie privée et d'obtenir des données de grande qualité pourrait n'être respectée que si de bonnes restrictions organisationnelles et juridiques empêchent de faire le lien entre les fichiers de données en cause et cette vaste connaissance qui s'ajoute ». De plus, Bethlehem, Keller et Pannekoek (1990) ont été parmi les premiers à recourir à la modélisation probabiliste pour évaluer le risque de réidentification dans les microdonnées d'enquête en estimant le nombre d'enregistrements uniques de population compte tenu des enregistrements uniques d'échantillon dans un ensemble de quasi-identificateurs recoupés. De plus amples renseignements sur cette méthodologie et sur les contributions de Chris dans ce domaine sont présentés à la section 4.1.

Dans les années 1990, la demande de produits détaillés s'est faite beaucoup plus grande, plus particulièrement en raison de la disponibilité de meilleures solutions technologiques et de meilleurs ordinateurs personnels. Les utilisateurs des données s'inquiétaient de plus en plus du fait qu'ils devaient travailler avec des produits protégés, modifiés ou perturbés. Cela a coïncidé avec les avancées à grande échelle dans le domaine du contrôle de la divulgation statistique permises par l'évolution scientifique de la méthodologie et l'échange international des acquis théoriques et pratiques à l'occasion, par exemple, du symposium international tenu aux Pays-Bas en 1990, qui portait sur la prévention de la divulgation statistique (voir à ce sujet le numéro spécial de Statistica Neerlandica, 1992). Des collaborations croisées ont également eu lieu au sein de l'Union européenne dans le cadre du quatrième Projet de recherche-cadre sur le contrôle de la divulgation statistique (1996 à 1998) et de nombreux autres projets qui ont suivi, y compris la conception de logiciels de CDS : mu-ARGUS pour les microdonnées (Hundepool et coll., 2003) et tau-ARGUS pour les données tabulaires (en particulier, la suppression de cellules dans les tableaux quantitatifs contenant des statistiques sur les entreprises) (Hundepool et coll., 2011). Pour en savoir plus sur les projets de recherche en Europe ainsi que sur le livre de Hundepool et coll. (2012), consultez le site Web suivant : https://research.cbs.nl/casc/index.htm (en anglais seulement).

Un numéro spécial du *Journal of Official Statistics* (volume 14, numéro 4, 1998) intitulé « Disclosure Limitation Methods for Protecting the Confidentiality of Statistical Data » a eu des répercussions particulières à cette époque et a mis en évidence la recherche à grande échelle réalisée sur le CDS dans le milieu universitaire et au sein des organismes gouvernementaux. De plus, un livre a été écrit et un cours a été mis sur pied (voir Willenborg et De Waal (1996), avec contributions de la part de Chris Skinner, et plus tard deuxième édition dans Willenborg et De Waal (2001)). Le travail se poursuivait parallèlement aux États-Unis et au Canada notamment grâce au Federal Committee on Statistical Methodology (1994), au Committee on Maintaining Privacy and Security in Health Care Applications of the National Information Infrastructure (1997) et à la publication intitulée « Disclosure Control Issues at Statistics Canada » (Yeo et Robertson, 1995), sans oublier le progiciel CONFID, qui permettait de supprimer des cellules dans des tableaux de données quantitatives.

Tout au long des années 1990, on a mis de plus en plus l'accent sur l'élaboration de dispositions et de lois en matière d'accès et de gouvernance, ainsi que sur la notion d'accès par niveaux pour fournir des

données statistiques aux chercheurs. On a créé des entrepôts de données et des centres de données de recherche en même temps qu'on a créé des approches et des cadres de gouvernance des données permettant une utilisation efficace des données statistiques, par exemple le cadre des cinq éléments de la sécurité présenté au tableau 2.1, qui a été mis en application à l'Office for National Statistics du Royaume-Uni en 2002 (Ritchie, 2009) et, plus tard, le cadre décisionnel sur l'anonymisation (Elliot, Mackey, O'Hara et Tudor, (2016) accessible à l'adresse suivante : https://ukanon.net/framework/).

Tableau 2.1
Les cinq éléments de la sécurité

Projets sécuritaires	Cette utilisation des données est-elle appropriée ?
Personnes fiables	Peut-on faire confiance aux utilisateurs pour qu'ils emploient les données de façon appropriée ?
Lieux sécuritaires	L'emplacement où l'accès a lieu limite-t-il l'utilisation non autorisée des données ?
Données sécuritaires	Y a-t-il un risque de divulgation propre aux données ?
Produits sécuritaires	Les résultats statistiques sont-ils non-divulgateurs ?

3. La révolution des données

Depuis environ 2008, il y a une abondance de données accessibles dans le domaine public, qu'il s'agisse de données ouvertes ou de mégadonnées, d'où un plus grand risque de non-respect de la vie privée et de la confidentialité, puisque ces sources de données pourraient servir à porter atteinte aux données statistiques diffusées. Ajoutons que la disponibilité d'outils technologiques de pointe qui améliorent les appariements et les manipulations de données augmente les possibilités de réidentification dans les données statistiques. Les organismes gouvernementaux ont commencé à se rendre compte que les méthodes normalisées de CDS pouvaient ne pas suffire à protéger la confidentialité des unités statistiques et ont donc introduit des restrictions plus strictes et un accès plus contrôlé aux données pour accroître le contrôle. Cela s'est par ailleurs manifesté par des modifications de la réglementation, plus particulièrement le Règlement général sur la protection des données de l'Union européenne de 2016 et ses dispositions et prescriptions en matière de traitement des données à caractère personnel des particuliers. On s'est également davantage intéressé aux préoccupations relatives à la protection des renseignements personnels dans les données sur la santé (El Emam, Jonker, Arbuckle et Malin, 2011) et dans les données génétiques (Homer et coll., 2008; Gymrek et coll., 2013), et il a été démontré que ces dernières données présentent un fort risque, ce qui a eu des répercussions sur la diffusion et le partage des bases de données sur l'ADN. Dans le domaine commercial, plusieurs exemples d'atteintes à la vie privée ont été largement publicisés : mots clés de recherche sur AOL (Barbaro et Zeller, 2006), déplacements en taxi à New York (Douriez, Doraiswamy, Freire et Silva, 2016), données de Facebook obtenues par Cambridge Analytica (Meredith, 2018), notamment.

En raison des grands progrès technologiques et de la possibilité de coupler des sources de données, on a créé des tiers de confiance pour effectuer des couplages de données et le calcul multipartite sécurisé. Le calcul multipartite sécurisé a d'abord été développé dans les études informatiques, où il a été croisé avec les

études statistiques, particulièrement la façon d'effectuer la modélisation statistique avancée selon cette approche (Slavkovic, Nardi et Tibbits, 2007; Snoke, Brick, Slavkovic et Hunter, 2018). De plus, la collaboration entre informaticiens et statisticiens s'est resserrée, engendrant une importante évolution en matière de confidentialité des bases de données au sein des organismes gouvernementaux (voir la section 5 pour obtenir plus de précisions). Dans leur étude sur la protection des renseignements personnels, Dwork, Smith, Steinke et Ullman (2017) ont écrit : [traduction] « À compter du milieu des années 2000, le domaine de l'analyse statistique des données assurant la protection des renseignements personnels a vu un afflux d'idées élaborées quelque 20 ans plus tôt dans le milieu de la cryptographie. »

4. Apports de Chris Skinner à la recherche sur le contrôle de la divulgation statistique

La recherche officielle de Chris Skinner sur le CDS a commencé par ses collaborations à l'Université de Manchester, qui plaidaient en faveur de la diffusion des microdonnées d'échantillons (les échantillons d'enregistrements anonymisés) tirées du Recensement de 1991 du Royaume-Uni (Marsh et coll., 1991; Skinner, Marsh, Openshaw et Wymer, 1994; Marsh, Dale et Skinner, 1994). Il s'est alors intéressé à la mesure du risque de réidentification dans les microdonnées d'enquête au moyen de la modélisation probabiliste, qui a fait l'objet d'une première publication dans Skinner (1992) et que nous allons décrire à la section 4.1. Il a également commencé une longue carrière en tant que conseiller auprès des comités gouvernementaux de statistique et d'accès aux données, par exemple : le UK Census Design and Methodology Advisory Committee Statistical Disclosure Control Subgroup (2008 à 2010); le Understanding Society Data Access Committee (2010 à 2013); l'Expert Advisory Group on Data Access, Wellcome Trust, Medical Research Council, Economic and Social Research Council et Cancer Research UK (2012 à 2014).

4.1 Mesure du risque de réidentification dans les microdonnées et les prolongations d'enquête

Depuis les années 1990, de nombreux organismes gouvernementaux publient des microdonnées provenant d'enquêtes gouvernementales à grande échelle dans lesquelles l'échantillon est tiré aléatoirement d'une population générale finie et les fractions d'échantillonnage sont petites. À titre d'exemple, mentionnons les échantillons tirés de l'Enquête sur la population active et de l'Enquête sur les dépenses familiales. Normalement, le chercheur devait suivre un processus pour obtenir l'accès aux données, qu'elles aient été reçues sur une disquette ou au moyen d'un entrepôt de données. Néanmoins, on s'est rapidement aperçu que l'échantillonnage à lui seul ne pouvait pas assurer une protection suffisante dans les microdonnées, et une foule de recherches ont été entreprises sur la mesure du risque de réidentification dans les microdonnées d'échantillon et sur les techniques de prévention de la divulgation.

Le scénario de risque de divulgation lors de la diffusion de microdonnées d'échantillon tirées d'une population générale repose sur les hypothèses suivantes : 1) un « intrus » (une personne animée de l'intention malveillante de discréditer le bureau de la statistique) qui a accès aux microdonnées et aux autres renseignements auxiliaires sur la population, qui lui permettent de coupler des sources de données de manière à identifier des personnes dans les microdonnées; 2) il n'y a pas de « connaissance de la réponse », en ce sens que l'intrus ignore qui a été échantillonné aux fins de l'enquête. La définition fondamentale du risque de réidentification réside, par conséquent, dans la probabilité de faire le lien. Chris Skinner est l'un des premiers à avoir établi un cadre de modélisation statistique pour l'estimation de la probabilité de réidentification en fonction des données diffusées et des hypothèses sur la façon de produire les données (connaissance du processus d'échantillonnage). Le modèle vise des variables clés définies comme ensemble de quasi-identificateurs dans ces deux sources de données, lesquelles sont généralement catégoriques (âge, sexe, lieu, groupe ethnique, etc.). Le recoupement de ces variables clés donne naissance à de grands tableaux de contingence contenant des comptes de l'échantillon; beaucoup de cellules de ces tableaux ont une valeur nulle ou une valeur de un. À cet égard, nous nous intéressons tout particulièrement au risque de divulgation selon les cellules ayant une valeur de un, à savoir les enregistrements uniques d'échantillon. Le risque de réidentification repose sur la notion d'unicité de la population dans le tableau de contingence : s'il y a un enregistrement unique d'échantillon observé dans une cellule d'un tableau créé à partir du croisement des variables clés, quelle est la probabilité que la cellule soit aussi un enregistrement unique de population ? Les mesures individuelles du risque par enregistrement sont estimées sous forme de probabilités de réidentification. Ces mesures sont ensuite agrégées pour obtenir des mesures de risque globales pour l'ensemble du fichier, qui sont utiles pour prendre des décisions éclairées au sujet de la diffusion des données et du niveau d'accès fourni par l'entremise de conseils de gouvernance des données.

Le cadre de modélisation probabiliste conçu par Chris Skinner suit une approche simplifiée qui restreint les renseignements connus des intrus (Skinner et Holmes, 1998; Elamir et Skinner, 2006). Désignons par F_k la taille de la population dans la cellule k d'un tableau comprenant des variables clés ayant K cellules, désignons par f_k la taille de l'échantillon dans la cellule k, et supposons que $\sum_k F_k = N$ et que $\sum_k f_k = n$. L'ensemble d'enregistrements uniques d'échantillon est défini de la façon suivante : $\mathrm{UE} = \{k : f_k = 1\}$. Cet ensemble représente les enregistrements à risque élevé susceptibles d'être des enregistrements uniques de population. Voici deux mesures globales du risque de divulgation (où I est la fonction indicatrice) :

1. Nombre d'enregistrements uniques d'échantillon qui sont des enregistrements uniques de population : $\tau_1 = \sum_k I(f_k = 1, F_k = 1)$.
2. Nombre prévu de liens appropriés si nous devions faire correspondre les enregistrements uniques d'échantillon à la population (en supposant une assignation aléatoire de la population dans la cellule k). Par exemple, si un enregistrement unique d'échantillon correspond à trois personnes dans la population, la probabilité de correspondance de cet enregistrement unique d'échantillon

serait de 1/3. L'agrégation de toutes les probabilités de correspondance des enregistrements uniques d'échantillon nous mène à : $\tau_2 = \sum_k I\left(f_k = 1\right) 1/F_k$.

Si l'on connaît les fréquences de la population F_k, les mesures globales du risque de divulgation sont alors simples à calculer. Toutefois, on suppose généralement que les fréquences de la population F_k sont inconnues, et il faut utiliser une modélisation probabiliste pour estimer les mesures du risque de divulgation de la façon suivante :

$$\hat{\tau}_1 = \sum_k I\left(f_k = 1\right) \hat{P}\left(F_k = 1 \,\middle|\, f_k = 1\right) \quad \text{et} \quad \hat{\tau}_2 = \sum_k I\left(f_k = 1\right) \hat{E}\left(1/F_k \,\middle|\, f_k = 1\right). \tag{4.1}$$

Comme nous modélisons les comptes de population en fonction d'un tableau de contingence de fréquences d'échantillon définies par les variables clés, Chris Skinner a supposé une distribution de Poisson et un modèle log-linéaire pour estimer les mesures du risque de divulgation dans l'équation (4.1). Dans ce modèle, lui et ses coauteurs supposent que F_k sont des réalisations de variables aléatoires de Poisson indépendantes : $F_k \sim \text{Pois}\left(\lambda_k\right)$ pour chaque cellule k. Un échantillon est sélectionné par échantillonnage de Poisson ou de Bernoulli avec une fraction de sondage π_k dans la cellule $k : f_k \,\middle|\, F_k \sim \text{Bin}\left(F_k, \pi_k\right)$. Il s'ensuit que :

$$f_k \sim \text{Pois}\left(\pi_k \lambda_k\right) \quad \text{et} \quad F_k \,\middle|\, f_k \sim \text{Pois}\left(\lambda_k \left(1 - \pi_k\right)\right) \tag{4.2}$$

et les fréquences par cellule de population F_k, étant donné que les fréquences par cellule d'échantillonnage f_k sont également des réalisations de variables aléatoires de Poisson indépendantes.

Comme c'est le cas habituellement dans ce type de cadre, on estime les paramètres λ_k au moyen d'une modélisation log-linéaire. Les fréquences d'échantillon f_k suivent une distribution de Poisson indépendante selon une moyenne de $\mu_k = \pi_k \lambda_k$. Un modèle log-linéaire pour le μ_k est exprimé de la façon suivante : $\log\left(\mu_k\right) = \mathbf{x}_k' \boldsymbol{\beta}$, où $\mathbf{x}_k$ est un vecteur de plan qui désigne les principaux effets et éléments d'interaction du modèle pour les variables clés. On obtient l'estimateur du maximum de vraisemblance $\hat{\boldsymbol{\beta}}$ en résolvant les équations de valeurs résultantes :

$$\sum_k \left(f_k - \pi_k \exp\left(\mathbf{x}_k' \boldsymbol{\beta}\right)\right) \mathbf{x}_k = 0. \tag{4.3}$$

On calcule alors les valeurs prédites par : $\hat{\mu}_k = \exp\left(\mathbf{x}_k' \hat{\boldsymbol{\beta}}\right)$ et $\hat{\lambda}_k = \hat{\mu}_k / \pi_k$. Les mesures individuelles du risque de divulgation pour la cellule k, en supposant une distribution de Poisson, sont :

$$P\left(F_k = 1 \,\middle|\, f_k = 1\right) = \exp\left(\lambda_k \left(1 - \pi_k\right)\right)$$

et

$$E\left(1/F_k \,\middle|\, f_k = 1\right) = \left(1 - \exp\left(\lambda_k \left(1 - \pi_k\right)\right)\right) \Big/ \left(\lambda_k \left(1 - \pi_k\right)\right). \tag{4.4}$$

L'intégration de $\hat{\lambda}_k$ pour λ_k dans l'équation (4.4) conduit aux estimations des mesures du risque de divulgation au niveau de l'enregistrement $\hat{P}\left(F_k = 1 \,\middle|\, f_k = 1\right)$ et $\hat{E}\left(1/F_k \,\middle|\, f_k = 1\right)$, qui sont agrégées pour obtenir des estimations des mesures globales du risque de divulgation $\hat{\tau}_1$ et $\hat{\tau}_2$ pour l'équation (4.1).

Par exemple, supposons qu'un organisme statistique aimerait diffuser des microdonnées de l'Enquête sur la population active. L'organisme suppose les quasi-identificateurs (variables clés) suivants, qu'il aimerait diffuser dans les microdonnées et utiliser pour évaluer le risque de réidentification (nombre de catégories entre parenthèses) : sexe (2), groupe d'âge (10), état matrimonial (3), groupe ethnique (10), situation d'emploi (3) et groupe professionnel (10). Lorsqu'elles sont recoupées, ces variables clés mènent à des cellules k, $k = 1, \ldots, K$, où $K = 18\,000$ cellules. L'organisme détermine les cellules des variables clés recoupées contenant une seule unité d'échantillonnage : $f_k = 1$ (enregistrement unique d'échantillon). Pour chaque cellule qui est un enregistrement unique d'échantillon, nous estimons la mesure du risque de divulgation au niveau de l'enregistrement en fonction de la probabilité que l'enregistrement unique d'échantillon soit également un enregistrement unique de population $P\left(F_k = 1 \mid f_k = 1\right)$. Nous estimons également la probabilité d'appariement de l'enregistrement unique d'échantillon en fonction (hypothétiquement) de la correspondance de cet enregistrement à une population à l'aide des variables clés comme variables d'appariement pour obtenir $E\left[1 \,/\, F_k \mid f_k = 1\right]$. Ces mesures du risque au niveau de l'enregistrement sont ensuite agrégées pour obtenir des estimations des mesures globales du risque de divulgation τ_1 et τ_2, en supposant que, pour un K de grande taille, $\left(F_k, f_k\right)$ sont indépendantes.

Skinner et Shlomo (2008) ont élaboré une méthode de sélection des principaux effets et paramètres d'interaction pour le modèle log-linéaire qui permet de trouver le bon équilibre en tenant compte des valeurs nulles aléatoires et structurelles dans le tableau de contingence. La méthode repose sur l'estimation et la réduction au minimum (approximative) du biais des estimations de risque $\hat{\tau}_1$ et $\hat{\tau}_2$. En définissant $h\left(\lambda_k\right) = P\left(F_k = 1 \mid f_k = 1\right)$ pour τ_1 et $h\left(\lambda_k\right) = E\left(1 \,/ F_k \mid f_k = 1\right)$ pour τ_2, ils considèrent l'expression :

$$B = \sum\nolimits_k E\left(I\left(f_k = 1\right)\right)\left(h\left(\hat{\lambda}_k\right) - h\left(\lambda_k\right)\right).$$

Un développement en série de Taylor de h mène à l'approximation

$$B \approx \sum\nolimits_k \pi_k \lambda_k \exp\left(-\lambda_k\right)\left(h'\left(\lambda_k\right)\left(\hat{\lambda}_k - \lambda_k\right) + h''\left(\lambda_k\right)\left(\hat{\lambda}_k - \lambda_k\right)^2 \Big/ 2\right)$$

et les relations $E\left(f_k\right) = \pi_k \lambda_k$ et $E\left(\left(f_k - \pi_k \hat{\lambda}_k\right)^2 - f_k\right) = \pi_k^2 E\left(\hat{\lambda}_k - \lambda_k\right)^2$, selon l'hypothèse d'ajustement de la distribution de Poisson, conduisent à une nouvelle approximation de B sous la forme :

$$\hat{B} \approx \sum\nolimits_k \hat{\lambda}_k \exp\left(-\pi_k \hat{\lambda}_k\right)\left(-h'\left(\lambda_k\right)\left(f_k - \pi_k \hat{\lambda}_k\right) + h''\left(\lambda_k\right)\left(\left(f_k - \pi_k \hat{\lambda}_k\right)^2 - f_k\right) \Big/ \left(2\pi_k\right)\right). \quad (4.5)$$

Par exemple, pour τ_1 :

$$\hat{B}_1 \approx \sum\nolimits_k \hat{\lambda}_k \exp\left(-\hat{\lambda}_k\right)\left(1 - \pi_k\right)\left\{\left(f_k - \pi_k \hat{\lambda}_k\right) + \left(1 - \pi_k\right)\left[\left(f_k - \pi_k \hat{\lambda}_k\right)^2 - f_k\right] \Big/ \left(2\pi_k\right)\right\}. \quad (4.6)$$

Comme on peut le constater, les critères de qualité de l'ajustement $\hat{B}$ sont liés à la notion de mesure de la surdispersion et de la sous-dispersion élaborée dans les études économétriques (Cameron et Trivedi, 1998). La méthode permet de sélectionner le modèle à l'aide d'un algorithme de recherche avant (forward search) qui réduit au minimum l'estimation du biais standardisé $\hat{B}_i \,/\, \sqrt{\hat{v}_i}$ pour $\hat{\tau}_i$, $i = 1, 2$, où $\hat{v}_i$ sont des

estimations de la variance de $\hat{B}_i$. Les critères de qualité de l'ajustement $\hat{B}_i \,/\, \sqrt{\hat{v}_i}$ présentent, en valeur approchée, une distribution normale centrée réduite selon l'hypothèse que l'espérance de $\hat{B}_i$ est nulle.

Skinner et Shlomo (2008) ont également traité de l'estimation des mesures de risque de divulgation dans des plans de sondage complexes au moyen de la stratification, de la mise en grappes et des poids d'enquête. Bien que la méthode décrite suppose que toutes les personnes dans la cellule k sont sélectionnées indépendamment à l'aide de l'échantillonnage de Bernoulli, c'est-à-dire $P\left(f_k = 1 \,\middle|\, F_k\right) = F_k\, \pi_k \left(1-\pi_k\right)^{F_k - 1}$, cela peut ne pas être le cas si l'échantillonnage est réuni en grappes (pour les ménages). Dans la pratique, les variables clés sont normalement l'âge, le sexe, la profession, et elles ont tendance à recouper les grappes. L'hypothèse qui précède se vérifie donc dans la pratique pour la plupart des enquêtes menées auprès des ménages et n'entraîne pas de biais lors de l'estimation des mesures de risque. Par ailleurs, les probabilités d'inclusion sont susceptibles de varier selon les strates, la stratification la plus courante se faisant selon la région géographique. Des indicateurs de strates devraient toujours figurer dans les variables clés pour rendre compte des probabilités différentielles d'inclusion dans le modèle log-linéaire. Dans un échantillonnage complexe, les λ_k peuvent être estimés avec cohérence au moyen du pseudomaximum de vraisemblance (Rao et Thomas, 2003), où l'équation d'estimation (4.3) se trouve ainsi modifiée :

$$\sum_k \left(\hat{F}_k - \exp\left(\mathbf{x}_k' \boldsymbol{\beta}\right)\right) \mathbf{x}_k = 0 \tag{4.7}$$

et $\hat{F}_k$ s'obtient en additionnant les poids d'enquête dans la cellule k : $\hat{F}_k = \sum_{i \in k} w_i$. Les estimations λ_k ainsi obtenues sont introduites dans les expressions de l'équation (4.4) et π_k est remplacé par l'estimation $\hat{\pi}_k = f_k \,/\, \hat{F}_k$. Les critères de qualité de l'ajustement $\hat{B}$ sont aussi adaptés à la méthode du pseudomaximum de vraisemblance. Une simulation et une application réelle démontrant cette approche pour une enquête ayant un plan à probabilité égale et une enquête ayant un plan complexe se trouvent dans Skinner et Shlomo (2008).

Pour la modélisation probabiliste présentée dans cet article et dans d'autres études spécialisées apparentées, la supposition est qu'il n'y a aucune erreur de mesure dans la façon dont les données sont enregistrées. En dehors des erreurs habituelles de saisie des données, les variables clés peuvent être mal classées à dessein afin de masquer les données, ce qui peut se faire, par exemple, par la permutation d'enregistrements ou par la méthode de la randomisation *a posteriori* (MRP) (Gouweleeuw, Kooiman, Willenborg et De Wolf, 1998). Shlomo et Skinner (2010) ont adapté l'estimation du risque de réidentification en fonction de τ_2, de manière à tenir compte des erreurs de mesure. Si l'on désigne les variables clés recoupées dans la population et les microdonnées par X et que l'on suppose que les X des microdonnées sont entachées d'une certaine erreur de classification ou de perturbation dénotée par la valeur $\tilde{X}$ et déterminée indépendamment par une matrice de classification erronée M :

$$M_{kj} = P\left(\tilde{X} = k \,\middle|\, X = j\right). \tag{4.8}$$

La mesure du risque de divulgation au niveau de l'enregistrement d'un appariement au moyen d'un enregistrement unique d'échantillon en situation d'erreur de mesure devient :

$$\frac{M_{kk}\left(1-\pi_k M_{kk}\right)}{\sum_j F_j M_{kj} \Big/ \left(1-\pi_k M_{kj}\right)} \leq \frac{1}{F_k}. \tag{4.9}$$

Dans une hypothèse de petites fractions d'échantillonnage et de légères erreurs de classification, on peut estimer la mesure de risque de divulgation par : $M_{kk} \Big/ \sum_j F_j M_{kj}$ ou $M_{kk} \Big/ \tilde{F}_k$, où $\tilde{F}_k$ est le compte de population et $\tilde{X} = k$. Après agrégation des mesures de risque de divulgation de l'enregistrement, la mesure globale de risque τ_2 devient :

$$\tau_2 = \sum_k I\left(f_k = 1\right) M_{kk} \Big/ \tilde{F}_k . \tag{4.10}$$

Il convient de prendre note que pour calculer la mesure, il faut seulement connaître la diagonale de la matrice d'erreur de classification, ce qui est la probabilité de ne pas être perturbé. Les comptes de population sont généralement inconnus, de sorte que l'estimation de l'équation (4.10) peut s'obtenir par modélisation probabiliste sur l'échantillon classé incorrectement comme ci-dessus :

$$\hat{\tau}_2 = \sum_k I\left(\tilde{f}_k = 1\right) M_{kk}\, \hat{E}\left(1 \Big/ \tilde{F}_k \,\middle|\, \tilde{f}_k\right). \tag{4.11}$$

Dans des travaux plus récents que j'ai réalisés avec Chris Skinner, et qui ont été présentés pour la première fois dans Shlomo et Skinner (2022), j'expérimente une nouvelle avenue pour la mesure du risque de réidentification dans le cas des sources de données non probabilistes. Plus précisément, il existe des registres du domaine public où l'appartenance est inconnue et revêt un caractère délicat. Comme exemples de tels registres, mentionnons le recensement des gens souffrant d'un problème médical comme le cancer ou le VIH, ou encore le dénombrement des adhérents à un programme de cartes de fidélité. On peut étendre l'approche au cas où des échantillons sont tirés des registres et, plus généralement, aux échantillons non probabilistes issus, par exemple, d'enquêtes en ligne. Si nous élargissons le cadre qui précède, les microdonnées d'un échantillon aléatoire peuvent encore servir à estimer les paramètres de population selon le cadre de modélisation probabiliste d'estimation du risque de réidentification mentionné précédemment. Cependant, il est compliqué d'estimer en même temps la propension à l'appartenance chez les personnes du registre.

Plus précisément, supposons que U et U_1 désignent la population et la population d'un registre, respectivement, et que $U_1 \subset U$. Supposons que R_i est la variable indicatrice du registre pour la personne i, et que $R_i = 1$ si $i \in U_1$, sinon $R_i = 0$. Comme on l'a mentionné, on suppose que l'appartenance R_i est une variable sensible dont la divulgation n'est pas souhaitable.

Nous désignons les fréquences de population du registre dans la cellule k par F_k^1. Les enregistrements les plus à risque sont ceux des cellules pour lesquelles $F_k^1 = 1$, et, comme pour le calcul présenté dans Skinner et Shlomo (2008), la mesure du risque s'obtient par

$$\tau_1^* = \sum_k P\left(F_k = 1 \,\middle|\, F_k^1 = 1\right) I\left(F_k^1 = 1\right). \tag{4.12}$$

Il n'y a aucun moyen d'estimer ces mesures avec cohérence seulement à partir des microdonnées du registre. Les microdonnées fournissent des renseignements sur la F_k^1, mais pas sur la F_k dans U, et la

distribution de X dans U_1 peut être très différente de celle dans U, de sorte que les microdonnées ne contiennent aucun renseignement direct sur la F_k. Nous utilisons donc le fichier de microdonnées d'échantillon aléatoire dans lequel les valeurs de X sont enregistrées pour un échantillon probabiliste s provenant de U. Supposons que f_k indique la fréquence dans la cellule k de s. Il faut prendre note que la f_k et la F_k^1 sont observées, mais pas la F_k. Si l'intrus a accès au fichier de microdonnées d'échantillon, on pourrait avoir intérêt à se concentrer sur les cellules pour lesquelles $f_k = 1$, ce qui permet d'obtenir la mesure suivante du risque :

$$\tau_1 = \sum_k P\left(F_k = 1 \,\middle|\, F_k^1 = 1, f_k = 1\right) I\left(F_k^1 = 1, f_k = 1\right). \tag{4.13}$$

Selon Skinner et Shlomo (2008), supposons que F_k suit une distribution de Poisson $F_k \sim \text{Pois}(\lambda_k)$, où le paramètre λ_k obéit au modèle log-linéaire :

$$\log(\lambda_k) = \mathbf{x}_k'\boldsymbol{\beta}. \tag{4.14}$$

Supposons que dans la cellule k, la variable d'appartenance inconnue R_i a pour valeur 1 selon une probabilité p_k, indépendamment pour chacune des F_k unités, afin que $F_k^1 \sim \text{Pois}(\phi_k)$, où $\phi_k = \lambda_k\, p_k$, et que les F_k^1 soient réparties de façon binomiale $F_k^1 \mid F_k \sim \text{Bin}(F_k, p_k)$ selon la F_k. Nous supposons en outre que p_k suit le modèle logistique :

$$\text{logit}(p_k) = \mathbf{x}_k'\,\boldsymbol{\xi}. \tag{4.15}$$

Comme le montrent Shlomo et Skinner (à paraître), la mesure du risque τ_1 est estimée par :

$$P\left(F_k = 1 \,\middle|\, F_k^1 = 1, f_k = 1\right) = \frac{\exp\left(-(1-\pi_k)(\lambda_k - \phi_k)\right)}{1 + (1-\pi_k)(\lambda_k - \phi_k)}$$

et pour évaluer τ_1^*, nous utilisons

$$P\left(F_k = 1 \,\middle|\, F_k^1 = 1\right) = P\left(F_k - F_k^1 = 0\right) = \exp\left(-(\lambda_k - \phi_k)\right) \tag{4.16}$$

puisque $F_k - F_k^1 \sim \text{Pois}(\lambda_k - \phi_k)$.

Par conséquent, dans une première étape, l'estimation de ces mesures exige à la fois l'estimation de $\boldsymbol{\beta}$ à partir de $f_k \sim \text{Pois}(\pi_k\, \lambda_k)$ et, dans une deuxième étape, l'estimation de $\boldsymbol{\xi}$, en fixant λ_k à la valeur impliquée par l'équation (4.14). Nous utilisons alors l'équation (4.16) et le fait que $\phi_k = \lambda_k\, p_k$ pour écrire

$$\log\phi_k = \log\lambda_k + \mathbf{x}_k'\boldsymbol{\xi} - \log\left(1 + \exp\left(\mathbf{x}_k'\boldsymbol{\xi}\right)\right) \tag{4.17}$$

et estimer $\boldsymbol{\xi}$ du fait que $F_k^1 \sim \text{Pois}(\phi_k)$, en recourant à une estimation de maximum de vraisemblance et en traitant λ_k comme étant connu. Shlomo et Skinner (2022) ont proposé d'autres méthodes d'estimation. Cependant, un travail approfondi est nécessaire pour améliorer l'estimation simultanée des paramètres $\boldsymbol{\xi}$ et $\boldsymbol{\beta}$ du modèle.

Un autre type d'estimateur fondé sur le plan permettant de mesurer le risque de divulgation dans des microdonnées d'échantillon s'appelle la mesure de simulation de l'intrusion des données. Elle a été mise au

point dans Skinner et Elliot (2002) et étendue à des plans de sondage plus complexes dans Skinner et Carter (2003). Le risque de divulgation est alors fonction d'un scénario différent où l'intrus tire une unité au hasard dans la population, vérifie si elle se trouve dans l'échantillon et, si oui, estime la probabilité d'un juste appariement avec l'unité de l'échantillon (c'est ce qu'on appelle le « scénario à l'aveuglette »). Il faut prendre note que le scénario diffère grandement du scénario mentionné pour la modélisation probabiliste où l'intrus a accès à une unité dans les microdonnées diffusées et tente un appariement de l'unité avec la population. L'avantage du « scénario à l'aveuglette » est que la mesure peut facilement s'estimer sans qu'on ait à passer par une modélisation probabiliste. La mesure de simulation de l'intrusion des données se définit comme suit :

$$\theta = \sum_k I\left(f_k = 1\right) \Big/ \sum_k I\left(f_k = 1\right) F_k \tag{4.18}$$

et elle est estimée par :

$$\hat{\theta} = \pi n_1 \Big/ \left[\pi n_1 + 2\left(1-\pi\right) n_2\right] \tag{4.19}$$

où n_1 sont les enregistrements uniques d'échantillon, à savoir $\text{UE} = \left\{k : f_k = 1\right\}$, et n_2 sont les enregistrements en double d'échantillon, à savoir $\text{DE} = \left\{k : f_k = 2\right\}$. Skinner et Shlomo (2012) ont étendu cette approche pour permettre d'estimer les fréquences des fréquences de populations finies au-delà des enregistrements uniques d'échantillon.

4.2 Distinction entre le risque de divulgation et le préjudice

Chris a proposé dans Skinner (2012) un cadre conceptuel pour distinguer le risque de divulgation au préjudice causé par la divulgation, faisant ainsi le lien avec des études de Duncan et Lambert (1986) et de Lambert (1993). Ce cadre repose sur la théorie de la décision où les acteurs sont l'organisme, l'intrus et l'utilisateur dans une analyse faisant intervenir leurs actions et leurs fonctions de perte. Chris a insisté sur l'importance de distinguer ce qui peut se mesurer selon la théorie statistique (risque de divulgation possible) et de constater quels aspects de la décision nécessitent d'autres apports, comme le jugement de politique (préjudice de divulgation possible). Ce travail a également été motivé par le cadre du risque de divulgation par rapport à l'utilité des données de Duncan, Keller-McNulty et Stokes (2001) et les facteurs économiques de la protection des renseignements personnels de Abowd et Schmutte (2019).

Comme le révèlent ces exemples, Chris Skinner a accru la profondeur et l'étendue des recherches consacrées au CDS. Chris a eu également une influence considérable sur la recherche sur les associations entre la mesure des risques de divulgation dans le CDS et d'autres recherches apparentées, par exemple, des recherches sur le couplage d'enregistrements (Skinner, 2009) ou les sciences judiciaires (Skinner, 2007). Les prochaines sections portent sur les travaux plus récents de Chris sur le risque de divulgation et la protection des renseignements personnels.

5. Risque de divulgation et protection des renseignements personnels

Dans la littérature sur la protection des renseignements personnels en sciences informatiques, on définit officiellement la protection des renseignements personnels au moyen de modèles visant à assurer une protection contre une catégorie d'attaques. Les modèles de protection des renseignements personnels sont paramétrés selon un seuil de risque de divulgation déterminé *a priori*. Une fois le modèle défini, une technique de perturbation est élaborée pour garantir la protection contre les attaques, en fonction d'un certain seuil. Des exemples de modèles de protection des renseignements personnels comprennent : k-anonymat, la proximité t, la diversité l et la confidentialité différentielle. Comme on l'a mentionné précédemment, dans les études sur la protection des renseignements personnels, on recourt à des techniques perturbatrices, lesquelles entraînent une plus grande perte de renseignements que certaines des techniques le plus normalisées élaborées dans les études sur le CDS. La catégorie d'attaques dans les études sur la protection des renseignements personnels a normalement pour motif le traitement de la divulgation par déduction, qui comprend à la fois les risques de divulgation de l'identité et des attributs, même si le modèle k-anonymat (Sweeney, 2002) vise à prévenir des attaques par appariement de façon semblable à l'approche de CDS que nous décrivons à la section 4. Nous remarquons que même s'ils mettent de plus en plus l'accent sur la protection contre la divulgation d'attributs et la divulgation par déduction, les organismes gouvernementaux sont encore obligés de se protéger contre la divulgation de l'identité par d'éventuelles attaques par appariement, en raison de la loi et des codes de pratique sur la protection des entités statistiques contre la réidentification. Les études portant sur le CDS s'appuient sur l'appariement de variables quasi identificatoires, tandis que dans les études sur la protection des renseignements personnels, il n'y a aucune distinction entre les variables identificatoires et les variables sensibles.

Il convient toutefois d'indiquer que nombre de concepts présentés dans les études consacrées à la protection des renseignements personnels ne sont pas nouveaux en ce qui a trait au CDS. Par exemple, les attaques par reconstruction, mentionnées dans Garfinkel, Abowd et Martindale (2018), qui motivent l'utilisation de la confidentialité différentielle pour protéger les produits du Recensement de 2021 aux États-Unis, comportent les mêmes facteurs que celles qui ont motivé la conception de la suppression de cellules complémentaires. Cette approche a été pensée dans les années 1980 pour protéger les tableaux de données quantitatives de statistique des entreprises. Lors de la sélection de suppressions de cellules complémentaires, les bornes inférieure et supérieure des cellules supprimées sont calculées en fonction des renseignements provenant des valeurs marginales du tableau et des hypothèses de non-négativité. Bien sûr, les attaques par reconstruction ne concernent pas l'appariement, mais plutôt la divulgation d'attributs en raison du petit nombre d'unités, plus particulièrement dans les valeurs marginales. Comme on l'a mentionné, les études sur la protection des renseignements personnels ont pour principal objet la divulgation d'attributs et la divulgation par déduction, bien que les études sur le CDS aient également traité de ces sujets, par exemple le risque de divulgation prédictive mentionné dans Fuller (1993). Un autre modèle de protection des renseignements personnels dans les études informatiques est le dépistage des attaques permettant d'inférer qu'une personne se trouve dans un ensemble de données de nature délicate (par exemple Homer et coll., 2008), mais dans les études sur le CDS, on s'est demandé également si une personne concernée est visible dans un ensemble de données.

Depuis 2005, il y a eu quatre réunions de concertation entre le milieu du CDS et la communauté informatique de la protection des renseignements personnels, lesquelles ont largement permis de comprendre les différentes approches en matière de maintien de la confidentialité et de maintien de l'utilité suffisante des données. Par exemple, dans Nissim et coll. (2018, page 5), il est mentionné ce qui suit : [traduction] « La protection des renseignements personnels est une propriété d'une relation informationnelle relative à une donnée d'entrée et à une donnée de sortie, et non une propriété relative à une donnée de sortie seulement », ce qui a entraîné certains assouplissements des garanties strictes en matière de protection des renseignements personnels dans les études informatiques. Un exemple d'assouplissement peut se trouver ci-dessous dans la formule (5.2) à la section 5.1. Par ailleurs, le milieu du CDS a reconnu la nécessité d'avoir des garanties plus officielles en matière de protection des renseignements personnels, particulièrement en raison des demandes croissantes pour que les organismes gouvernementaux permettent l'accès aux données statistiques au moyen d'applications de diffusion Web (par exemple des générateurs de tableaux flexibles, l'accès à distance, l'analyse à distance). La diffusion au moyen d'applications ouvertes signifie que les organismes renoncent à exercer un certain contrôle strict sur les données qui peuvent être diffusées. Les collaborations entre le milieu du CDS et la communauté informatique de la protection des renseignements personnels ont donné naissance en 2005 à une revue intitulée *Journal of Privacy and Confidentiality* (https://journalprivacyconfidentiality.org [en anglais seulement]), dont Chris Skinner a été un des premiers corédacteurs (Abowd, Nissim et Skinner, 2009).

Dans la prochaine section, nous nous concentrons sur le modèle de protection des renseignements personnels appelé confidentialité différentielle, puisque Chris Skinner a participé directement à l'élaboration de cette approche en tant que méthode possible à inclure dans la trousse à outils de CDS des organismes statistiques.

5.1 Confidentialité différentielle

Un modèle de protection des renseignements personnels qui a suscité beaucoup d'intérêt dans le milieu du CDS est celui de la confidentialité différentielle (Dwork, 2006). Dwork et Naor (2010) ont démontré que l'atteinte à la vie privée définie par Dalenius (1977), et présentée à la section 2, est inévitable et ont proposé qu'au lieu de comparer les renseignements avec et sans la $f(D)$ statistique, on compare plutôt la $f(D)$ et la $f(D')$, où D' est la base de données D avec une unité en moins.

En confidentialité différentielle, on considère un « scénario de la pire éventualité » où l'éventuel intrus dispose de renseignements complets sur toutes les unités de la base de données sauf une unité d'intérêt. La définition d'un mécanisme de perturbation M satisfait aux critères de ε -confidentialité différentielle si, pour toutes les requêtes dans les bases de données avoisinantes $D, D' \in A$ où A est le domaine des bases de données et D, D' différent d'une personne, et pour tous les résultats possibles définis comme sous-ensembles $S \in \text{Étendue}(M)$, nous obtenons :

$$p\left(M\left(D\right) \in S\right) \leq e^{\varepsilon} p\left(M\left(D'\right) \in S\right). \tag{5.1}$$

Un assouplissement est offert par la définition de la (ε, δ)-confidentialité différentielle :

$$p\left(M\left(D\right)\in S\right)\leq e^{\varepsilon}\,p\left(M\left(D'\right)\in S\right)+\delta. \tag{5.2}$$

Cela signifie que l'observation d'un produit perturbé S ne nous apprend à peu près rien (jusqu'à un degré de e^{ε}) et que l'intrus est incapable de déterminer si le produit provient de la base de données D ou D'. Autrement dit, le ratio $p\left(M\left(D\right)\in S\right)\big/\,p\left(M\left(D'\right)\in S\right)$ est borné et la probabilité du dénominateur ne peut pas être nulle. Ainsi, la confidentialité différentielle limite formellement l'accroissement du risque de divulgation pour une personne en raison de la présence de ses données dans la base de données D, risque auquel elle n'aurait pas été confrontée si ses données ne s'étaient pas trouvées dans la base de données D (Dwork et Roth, 2014). Dans le cadre de la (ε,δ)-confidentialité différentielle, un léger glissement est permis pour cette contrainte.

La solution pour garantir la confidentialité différentielle dans les études informatiques consiste à ajouter du bruit ou de la perturbation aux données de sortie des requêtes selon des paramétrages précis, par exemple en générant du bruit supplémentaire à partir de la distribution de Laplace (ou d'une distribution discrète de Laplace pour ajouter du bruit aux données de fréquence).

Shlomo et Skinner (2012) se sont d'abord demandé si les méthodes normalisées de CDS sont des mécanismes de confidentialité différentielle selon la définition fournie à l'équation (5.1). Ils ont constaté que l'échantillonnage, ainsi que d'autres méthodes de CDS non perturbatrices, comme la diminution du niveau de détail des variables, n'est pas différentiellement confidentiel. Dans ce contexte, il y a deux définitions possibles de la base de données : la base de données de population $\mathbf{x}_U=\left(x_1,\ldots,x_N\right)$ et la base de données d'échantillon $\mathbf{x}_s=\left(x_1,\ldots,x_n\right)$, où N correspond à la taille de la population U et n, à la taille de l'échantillon s. Supposons qu'un vecteur de fréquences de taille K est diffusé à partir de l'échantillon $\mathbf{f}=\left(f_1,\ldots,f_K\right)$, où $f_k=\sum_{i\in s} I\left(x_i=k\right)$. Soit $P\left(\mathbf{f}\,\middle|\,\mathbf{x}_U\right)$, qui indique la probabilité de $\mathbf{f}$ par rapport à l'échantillonnage et $\mathbf{x}_U$, qui est traité étant comme fixe. Selon cette configuration, la ε-confidentialité différentielle est maintenue si :

$$\max\left|\ln\left(\frac{P\left(\mathbf{f}\,\middle|\,\mathbf{x}_U\right)}{P\left(\mathbf{f}\,\middle|\,\mathbf{x}'_U\right)}\right)\right|\;\leq\;\varepsilon \tag{5.3}$$

pour certains $\varepsilon>0$, où le maximum est supérieur à toutes les paires $\left(\mathbf{x}_U,\mathbf{x}'_U\right)$ qui diffèrent par un seul élément et selon toutes les valeurs possibles de $\mathbf{f}$. Dans les stratégies d'échantillonnage aléatoire, il y a généralement une probabilité positive qu'un enregistrement unique d'échantillon pour une cellule k puisse être unique à une population : $f_k=F_k=1$. Pour une $\mathbf{f}$ donnée et un plan d'échantillonnage où f_k peut être égale à F_k en ayant une probabilité positive, il existe une base de données $\mathbf{x}_U$ pour laquelle $f_k=F_k\geq 1$ pour certaines k et $P\left(\mathbf{f}\,\middle|\,\mathbf{x}_U\right)\neq\mathbf{0}$. Si nous changeons un élément de $\mathbf{x}_U$ qui prend la valeur k pour générer $\mathbf{x}'_U$, nous obtenons alors $F'_k=F_k-1<f_k$ et $P\left(\mathbf{f}\,\middle|\,\mathbf{x}'_U\right)=\mathbf{0}$.

Par ailleurs, on peut rendre des méthodes perturbatrices différentiellement confidentielles dans la trousse à outils de CDS si les mécanismes en cause ne présentent pas de probabilité nulle de perturbation. À titre d'exemple, les méthodes de CDS ne perturbent habituellement pas les cellules à fréquence nulle des tableaux

de recensement contenant des fréquences de l'ensemble de la population, mais introduisent plutôt stochastiquement plus de valeurs nulles par la perturbation au moyen d'une méthode telle que l'arrondissement aléatoire. Si le but est toutefois de rendre la méthode de perturbation différentiellement confidentielle, les valeurs nulles (aléatoires) du tableau doivent aussi être perturbées.

5.2 Générateurs de tableaux flexibles en ligne

Les organismes gouvernementaux ont manifesté un grand intérêt pour la conception de générateurs de tableaux flexibles en ligne au moyen de plateformes Web personnalisées. Les utilisateurs génèrent et téléchargent leurs propres tableaux de recensement à partir d'un ensemble de variables et de catégories prédéfinies sélectionnées au moyen de listes déroulantes. De légères vérifications de la divulgation sont effectuées pour chaque tableau généré afin de déterminer si le tableau peut être diffusé ou non et, si c'est le cas, des méthodes de contrôle de la divulgation sont appliquées au tableau avant sa diffusion. Une application pour générer des tableaux flexibles a été conçue par l'Australian Bureau of Statistics, ou ABS (voir https://www.abs.gov.au/statistics/microdata-tablebuilder/tablebuilder). L'application s'appuie sur un vecteur de perturbation pour changer les valeurs quantitatives des cellules en fonction de leur valeur initiale. Le mécanisme de perturbation a comme propriété d'être borné et exempt de biais et de posséder une entropie maximale. Il permet seulement d'entraîner des perturbations non négatives sans perturber les cellules comportant des valeurs nulles. Shlomo et Young (2008) ont proposé une méthode de transformation des vecteurs de perturbation qui fait en sorte que les fréquences marginales sont préservées en espérance par l'introduction de la propriété d'invariance dans le mécanisme de perturbation.

Dans le générateur de tableaux flexibles en ligne de l'ABS, un petit chiffre aléatoire est attribué à chaque personne dans les microdonnées de recensement. Lorsque quelqu'un demande un tableau et que les personnes sont agrégées dans les cellules du tableau, il y aura aussi agrégation des chiffres aléatoires de ces personnes dans chaque cellule. La valeur aléatoire agrégée sert ensuite de « graine » pour déterminer la perturbation à effectuer (Fraser et Wooton, 2005). Cela signifie que chaque fois qu'une même cellule apparaît dans un quelconque tableau de recensement demandé, elle comporte toujours la même perturbation. Il n'y a donc aucun risque de pouvoir « déconstruire » une valeur réelle de cellule en faisant la moyenne des perturbations indépendantes selon des demandes multiples d'un même tableau. Avec ce traitement, on s'assure en outre que le mécanisme de perturbation est « non interactif », puisque, pour l'essentiel, tous les résultats provenant de la perturbation dans les tableaux de recensement demandés seront connus d'avance dans le générateur de tableaux flexibles en ligne.

Une des dernières initiatives de Chris Skinner avant sa maladie a été de diriger la mise sur pied d'un programme de collaboration entre statisticiens, informaticiens, praticiens et spécialistes des sciences sociales à l'Institut Isaac Newton de l'Université de Cambridge. Avec le professeur David Hand, il a mené à bien le programme d'appariement et d'anonymisation des données (grâce à la subvention EP/K032208/1 de l'Engineering and Physical Sciences Research Council du Royaume-Uni) de juillet à décembre 2016. C'est dans le cadre de ce programme qu'un groupe de statisticiens a voulu déterminer si la confidentialité

différentielle pouvait offrir une solution viable pour générer des tableaux flexibles de recensement en ligne, ce qui explique l'étude produite par Rinott, O'Keefe, Shlomo et Skinner (2018). La grande différence par rapport à l'approche initiale de l'ABS a été d'utiliser un mécanisme de perturbation différentiellement confidentiel (appelé mécanisme exponentiel et consistant pour l'essentiel en une distribution discrète de Laplace) et de perturber les cellules à fréquence nulle (aléatoires). Toute perturbation négative obtenue était alors ramenée à zéro dans les tableaux du recensement. En supposant des perturbations indépendantes, le mécanisme exponentiel se définit de la façon suivante : pour une valeur de fréquence de cellule a donnée, sélectionnons $b \in B$ (où B est la fourchette de b) et une probabilité proportionnelle à $\exp\left(\left(\varepsilon/2\right)u \,/\, \Delta u\right)$, où u est la perturbation et Δu, la différence maximale d'une fréquence par cellule dans la base de données D par rapport à D', ce qui, dans le cas des tableaux du recensement comportant des cellules internes, a une valeur de 1. La prise en compte des valeurs marginales dans les tableaux du recensement accroît la complexité de Δu et du vecteur de perturbation (voir Rinott et coll., 2018, pour en savoir plus sur les valeurs marginales). Afin de garantir l'utilité, les perturbations sont plafonnées à ± 7, le mécanisme satisfait donc aux critères de $\left(\varepsilon, \delta\right)$-confidentialité différentielle et nous permettons le glissement d'un ratio non borné, à la condition que δ soit très faible.

Dans ce cas-ci, nous mettons également en œuvre la méthodologie intitulée « même cellule, même perturbation » de Fraser et Wooton (2005), ce qui en fait essentiellement un mécanisme de confidentialité différentielle non interactif et, par conséquent, un budget de protection des renseignements personnels est nécessaire pour protéger les requêtes de tableaux dans le générateur de tableaux flexibles en ligne. Toute demande visant le même tableau entraînera toujours la même perturbation et aucun autre budget de protection des renseignements personnels ne sera consacré à d'autres perturbations au-delà de la perturbation initiale.

Le tableau 5.1 présente un exemple de vecteur de perturbation pour $\varepsilon = 1{,}5$ et pour $\delta = 0{,}00002$ et une perturbation limitée à ± 7, dans lequel la probabilité de perturbation est fondée sur une distribution discrète de Laplace : $p\left(u\right) = \frac{1}{C}\exp\left(-\varepsilon\left|u\right|\right)$, où C est une constante de normalisation faisant en sorte que le vecteur de perturbation s'additionne à 1. Comme il a été mentionné précédemment, toute valeur négative qui en résulte est ramenée à zéro, ce qui ne viole pas les critères de confidentialité différentielle. On trouvera des exemples et des applications dans Rinott et coll. (2018). Les auteurs montrent également la façon d'ajuster les inférences statistiques lorsqu'elles sont effectuées dans des tableaux de recensement perturbés, puisque le mécanisme de perturbation relatif à la confidentialité différentielle est connu. Cela contraste fortement avec l'approche de CDS dans laquelle on garde en général les paramètres secrets et l'on ne les communique pas aux chercheurs. Par exemple, lorsque l'on ajoute du bruit à des variables continues, la variance de la distribution du bruit n'est pas diffusée.

Tableau 5.1
Vecteur de perturbation pour le mécanisme de confidentialité différentielle $\varepsilon = 1{,}5$; $\delta = 0{,}00002$ et limite à ± 7

u	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
p(u)	0,00002	0,00008	0,00035	0,00157	0,00706	0,03162	0,14172	0,63516	0,14172	0,03162	0,00706	0,00157	0,00035	0,00008	0,00002

La confidentialité différentielle offre des garanties plus officielles de protection des renseignements personnels et protège contre les risques de divulgation d'attributs et de divulgation inférentielle. Par conséquent, elle peut constituer une meilleure solution pour la protection des données statistiques diffusées sous forme de données ouvertes ou au moyen d'applications ouvertes sur le Web, pour lesquelles il y a moins de contrôle et d'intervention de la part des administrateurs de données des organismes statistiques. Pour cette raison, la confidentialité différentielle devrait être incluse comme méthode supplémentaire dans la trousse à outils de CDS. Le US Census Bureau appliquera la confidentialité différentielle à ses produits du recensement de 2021 (Abowd, 2018). Il faudra approfondir la recherche pour comprendre en quoi les budgets en matière de confidentialité se trouvent influencés lorsqu'on combine cette méthode à d'autres méthodes de CDS, comme la diminution du niveau de détail, l'échantillonnage et la suppression de variables. Dans le domaine de la protection de la vie privée, la recherche se poursuit sur l'amélioration des mécanismes de perturbation de la confidentialité différentielle. Notons par exemple la confidentialité différentielle bornée que proposent Kifer et Machanavajjhala (2014).

6. Mot de la fin

En résumé, un des grands traits de la façon dont Chris Skinner a abordé la recherche sur le contrôle de la divulgation statistique et d'autres domaines de recherche a été de tenter de trouver des solutions pratiques à des problèmes statistiques réels. Ses recherches ont été influentes parce qu'il a pu mettre la théorie en pratique et résoudre des problèmes réels, faisant ainsi progresser les connaissances scientifiques dans les sciences sociales, les statistiques gouvernementales et sociales et la méthodologie d'enquête. Chris Skinner a eu une influence considérable sur des domaines de recherche autres que le CDS. Il a révisé des livres influents et rédigé de nombreux articles de revues spécialisées dans divers champs de recherche en statistique d'enquête, y compris les données manquantes, les erreurs de mesure, l'intégration des données, l'analyse de plans de sondage complexes et l'estimation à bases multiples. Chris était la voix définitive d'une génération dans la recherche et le développement des statistiques sociales et il nous manquera.

Bibliographie

Abowd, J.M. (2018). The U.S. Census Bureau adopts differential privacy. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2867. https://doi.org/10.1145/3219819.3226070.

Abowd, J.M., et Schmutte, I.M. (2019). An economic analysis of privacy protection and statistical accuracy as social choices. *American Economic Review*, 109(1), 171-202.

Abowd, J.M., Nissim, K. et Skinner, C.J. (2009). First issue editorial. *Journal of Privacy and Confidentiality*, 1(1), 1-6.

Barabba, V.P. (1975). The right to privacy and the need to know. Dans *US Bureau of the Census: A Numerator and Denominator for Measuring Change*. Document technique, 37, 23-29.

Barbaro, M., et Zeller, T. Jr. (2006). A face is exposed for AOL searcher, no. 4417749. *The New York Times*.

Bethlehem, J., Keller, W. et Pannekoek, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association*, 85(409), 38-45.

Cameron, A.C., et Trivedi, P.K. (1998). *Regression Analysis of Count Data*. Cambridge, Royaume-Uni: Cambridge University Press.

Cox, L.H. (1976). *Statistical Disclosure in Publication Hierarchies*. Rapport nº 14 du projet de recherche Confidentiality in Surveys, Dept. of Statistics, University of Stockholm, disponible à l'adresse https://hdl.handle.net/1813/111306.

Dalenius, T. (1974). The invasion of privacy problems and statistics production-an overview. *Statistisk Tidskrift*, 3, 213-225.

Dalenius, T. (1977). Towards a methodology for statistical disclosure control. *Statistik Tidskrift*, 15, 429-444.

Dalenius, T. (1988). *Controlling Invasion of Privacy in Surveys*. Dept. of Development and Research, Statistical Research Unit, Statistics Sweden.

Douriez, M., Doraiswamy, H. Freire, J. et Silva, C.T. (2016). Anonymizing NYC Taxi Data: Does it matter? *IEEE International Conference on Data Science and Advanced Analytics (DSAA2016)*, 140-148.

Duncan, G., et Lambert, D. (1986). Disclosure-limited data dissemination (avec discussion). *Journal of the American Statistical Association*, 81(393), 10-28.

Duncan, G., Keller-McNulty, S. et Stokes, S. (2001). Disclosure Risk vs. Data Utility: The R-U Confidentiality Map. Rapport technique LA-UR-01-6428. Statistical Sciences Group, Los Alamos, N.M.: Los Alamos National Laboratory.

Dunn, E.S. (1967). The idea of a national data centre and the issue of personal privacy. *American Statistician*, 21, 21-27.

Dwork, C. (2006). Differential privacy. Dans *ICALP*, (Éds., M. Bugliesi, B. Preneel, V. Sassone et I. Wegener). Heidelberg: Springer, note de cours en informatique 4052, 1-12.

Dwork, C., et Naor, M. (2010). On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. *Journal of Privacy and Confidentiality*, 2(1), 93-107.

Dwork, C., et Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4), 211-407.

Dwork, C., Smith, A., Steinke, T. et Ullman, J. (2017). Exposed! A survey of attacks on private data. *Annual Review of Statistics and its Application. Annual Review of Statistics and its Application*, 4, 61-84.

El Emam, K., Jonker, E., Arbuckle, L. et Malin, B. (2011). A systematic review of re-identification attacks on health data. *PLoS ONE*, 6(12), 1-12.

Elamir, E., et Skinner, C.J. (2006). Record-level measures of disclosure risk for survey microdata. *Journal of Official Statistics*, 22, 525-539.

Elliot, M., Mackey, E., O'Hara, K. et Tudor, C. (2016). *The Anonymisation Decision Making Framework, Manchester: UKAN*. Disponible à l'adresse https://ukanon.net/framework/.

Fellegi, I.P. (1972). On the question of statistical confidentiality. *Journal of the American Statistical Association*, 67(337), 7-18.

Fraser, B., et Wooton, J. (2005). *A Proposed Method for Confidentialising Tabular Output to Protect Against Differencing*. Joint UNECE/Eurostat work session on statistical data confidentiality, Genève, 9 au 11 novembre.

Fuller, W.A. (1993). Masking procedures for micro-data disclosure limitation. *Journal of Official Statistics*, 9, 383-406.

Garfinkel, S., Abowd, J.M. et Martindale, C. (2018). Understanding database reconstruction attacks on public data. *ACM QUEUE*, 16(5), 1-26.

Gouweleeuw, J., Kooiman, P., Willenborg, L.C.R.J. et De Wolf, P.P. (1998). Post randomisation for statistical disclosure control: Theory and implementation. *Journal of Official Statistics*, 14, 463-478.

Gymrek, M., McGuire, A.L., Golan, D., Halperin, E. et Erlich, Y. (2013). Identifying personal genomes by surname inference. *Science*, 339(6117), 321-324.

Haziza, D., et Smith, P.A. (2019). An interview with Chris Skinner. *Revue Internationale de Statistique*, 87(3), 451-470.

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J.V, Stephan, D.A., Nelson, S.F. et Craig, D.W. (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet*, 4(8), 1-9.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte-Nordholt, E., Spicer, K. et de Wolf, P.P. (2012). *Statistical Disclosure Control*. New York: John Wiley & Sons, Inc.

Hundepool, A., van de Wetering, A., Ramaswamy, R., de Wolf, P.P., Giessing, S., Fischetti, M. Salazar, J.J., Castro, J. et Lowthian, P. (2011). tau-ARGUS User's Manual (Version 3.5). Numéro BPA : 769-02-TMO, Statistics Netherlands Projet : Essnet-project, Statistics Netherlands, Pays-Bas. Disponible à l'adresse https://research.cbs.nl/casc/Software/TauManualV3.5_rev.pdf.

Hundepool, A., van de Wetering, A., Ramaswamy, R., Franconi, L., Capobianchi, A., de Wolf, P.P., Domingo, J., Torra, V., Brand, R. et Giessing, S. (2003). mu-ARGUS user's manual (Version 3.2). Numéro BPA : 768-02-TMO, Statistics Netherlands Projet : CASC-project, Statistics Netherlands, Pays-Bas. Disponible à l'adresse https://research.cbs.nl/casc/deliv/manual3.2.pdf.

Jabine, T.B., Michael, J.A. et Mugge, R.H. (1977). *Federal Agency Practices for Avoiding Stastical Disclosure: Findings and Recommendations*. Disponible à l'adresse http://www.asasrms.org/ Proceedings/y1977/Federal%20Agency%20Practices%20For%20Avolding%20Statistical%20Disclosure%20-%20Findings %20And%20Reconwnendatlons.pdf.

Kifer, D., et Machanavajjhala, A. (2014). Pufferfish: A framework for mathematical privacy definitions. *ACM Transactions on Database Systems*, 39(1), 1-36.

Lambert, D. (1993). Measures of disclosure risk and harm. *Journal of Official Statistics*, 9, 313-331.

Marsh, C., Dale, A. et Skinner, C.J. (1994). Safe data versus safe setting: Access to microdata from the British Census. *Revue Internationale de Statistique*, 62, 35-53.

Marsh, C., Skinner, C.J., Arber, S., Penhale, B., Openshaw, S., Hobcraft, J., Lievesley, D. et Walford, N. (1991). A case for samples of anonymised records from the 1991 Census. *Journal of the Royal Statistical Society A*, 154, 305-340.

Meredith, S. (2018). Facebook-Cambridge Analytica: A Timeline of the Data Hijacking Scandal. CNBC.

Nissim, K., Steinke, T. Wood, A., Altman, M., Bembenek, A., Bun, M., Gaboardi, M., O'Brien, D.R. et Vadhan, S. (2018). *Differential Privacy: A Primer for a Non-technical Audience*. University of Harvard, Disponible à l'adresse https://privacytools.seas.harvard.edu/files/privacytools/files/pedagogical-document-dp_0.pdf.

Paass, G. (1988). Disclosure risk and disclosure avoidance for microdata. *Journal of Business and Economic Statistics*, 6(4), 487-500.

Rao, J.N.K., et Thomas, D.R. (2003). Analysis of categorical response data from complex surveys: An appraisal and update. Dans *Analysis of Survey Data* (Éds., R.L. Chambers et C.J. Skinner), Wiley, Chichester, Royaume-Uni, 85-108.

Rinott, Y., O'Keefe, C., Shlomo, N. et Skinner, C. (2018). Confidentiality and differential privacy in the dissemination of frequency tables. *Statistical Sciences*, 33(3), 358-385.

Ritchie, F. (2009). Designing a national model for data access. *Proceedings of the Comparative Analysis of Enterprise Data Conference*, Tokyo, Japon. Disponible à l'adresse https://gcoe.ier.hit-u.ac.jp/CAED/ papers/ id213_Ritchie.pdf.

Shlomo, N., et Skinner, C.J. (2010). Assessing the protection provided by misclassification-based disclosure limitation methods for survey microdata. *Annals of Applied Statistics*, 4(3), 1291-1310.

Shlomo, N., et Skinner, C.J. (2012). Privacy protection from sampling and perturbation in survey microdata. *Journal of Privacy and Confidentiality*, 4(1), 155-169.

Shlomo, N., et Skinner, C.J. (2022). Measuring risk of re-identification in microdata: State-of-the art and new directions. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 185, 4, 1644-1662. Disponible à l'adresse http://dx.doi.org/10.1111/rssa.12902.

Shlomo, N., et Young, C. (2008). Invariant post-tabular protection of census frequency counts. Dans *PSD' 2008 Privacy in Statistical Databases*, (Éds., J. Domingo-Ferrer et Y. Saygin), Springer LNCS 5261, 77-89.

Skinner, C.J. (1992). On identification disclosure and prediction disclosure for microdata. *Statistica Neerlandica*, 46, 21-32.

Skinner, C.J. (2007). The probability of identification: Applying ideas from forensic statistics to disclosure risk assessment. *Journal of the Royal Statistical Society, Series A*, 170, 195-212.

Skinner, C.J. (2009). Record linkage, correct match probabilities and disclosure risk assessment. Dans *Insights on Data Integration Methodologies: ESSnet-ISAD workshop*, Vienne, 29 au 30 mai 2008, Eurostat Methodologies and Working papers, Luxembourg, European Communities, 11-23.

Skinner, C.J. (2012). Statistical disclosure risk: Separating potential and harm. *Revue Internationale de Statistique*, 80, 349-368, avec discussion et réponse, 379-391.

Skinner, C.J., et Carter, R.G. (2003). Estimation d'une mesure du risque de divulgation pour les microdonnées d'enquête sous échantillonnage avec probabilités inégales. *Techniques d'enquête*, 29, 2, 197-201. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2003002/ article/6784-fra.pdf.

Skinner, C.J., et Elliot, M.J. (2002). A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society B*, 64, 855-867.

Skinner, C.J., et Holmes, D. (1998). Estimating the re-identification risk per record in microdata. *Journal of Official Statistics*, 14, 361-372.

Skinner, C.J., et Shlomo, N. (2008). Assessing identification risk in survey microdata using Log Linear models. *Journal of the American Statistical Association*, 103(483), 989-1001.

Skinner, C.J., et Shlomo, N. (2012). Estimating frequencies of frequencies in finite populations. *Statistics and Probability Letters*, 82, 2206-2212.

Skinner, C.J., Marsh, C., Openshaw, S. et Wymer, C. (1994). Disclosure control for census microdata. *Journal of Official Statistics*, 10, 31-51.

Slavkovic, A.B., Nardi, Y. et Tibbits, M.M. (2007). Secure logistic regression of horizontally and vertically partitioned distributed databases. Seventh IEEE International Conference on Data Mining Workshops, (ICDMW2007), 723-728.

Snoke, J., Brick, T. Slavkovic, A. et Hunter, M.D. (2018). Providing accurate models across private partitioned data: Secure maximum likelihood estimation. *Annals of Applied Statistics*, 12(2), 877-914.

Statistica Neerlandica (1992). Volume 46(1), disponible à l'adresse https://onlinelibrary.wiley.com/toc/14679574/1992/46/1 [dernière consultation le 14 mars 2023].

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5), 557-570.

Willenborg, L., et De Waal, T. (1996). *Statistical Disclosure Control in Practice.* Notes de course en statistique, New York: Springer Verlag, 111.

Willenborg, L., et De Waal, T. (2001). *Elements of Statistical Disclosure Control in Practice*. Notes de course en statistique, New York: Springer-Verlag, 155.

Yeo, D., et Robertson, D. (1995). *Disclosure Control Issues at Statistics Canada*. https://publications.gc.ca/collections/collection_2017/statcan/11-613/CS11-617-96-5-eng.pdf.