Nº 11-522-x au catalogue
ISSN : 1709822-X

Recueil du Symposium de 2021 de Statistique Canada
Adopter la science des données en statistique officielle pour répondre aux besoins émergents de la société

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

par Natalie Shlomo

Date de diffusion : le 22 octobre 2021

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2021 de Statistique Canada
Adopter la science des données en statistique officielle pour répondre aux besoins émergents de la société

Contrôle de la divulgation statistique et développements 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 formelles 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 les années 1990 à propos de ses travaux sur la diffusion de microdonnées d'échantillon du recensement au Royaume-Uni. De ces recherches sont nées diverses études où on a mesuré le risque de réidentification dans les microdonnées d'enquête au moyen de modèles probabilistes. Je m'attacherai 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. Nous allons suivre le canevas préparé par Chris en prévision de cette allocution qui me vient de son fils, Tom Skinner.

Mots clés : risque de réidentification, révolution des données, modèles de confidentialité, confidentialité différentielle

1. Introduction

Une séance commémorative spéciale a eu lieu en l'honneur de Chris Skinner au Symposium international de 2021 sur les questions de méthodologie de Statistique Canada avec de nombreux témoignages émouvants d'amis et de collègues venus marquer la vie et les réalisations du défunt. Chris a été lauréat du prix Waksberg en 2019 et projetait de présenter son discours à cette occasion dans le cadre du Symposium international de 2019 sur les questions de méthodologie. Malheureusement, sa maladie a empiré et il est décédé le 21 février 2020. J'ai eu l'immense privilège de présenter son travail sur le contrôle de la divulgation statistique (CDS) depuis le tout début et jusqu'à présent en soulignant les apports de Chris dans ce domaine. Les grandes lignes de l'exposé figurent dans une série de notes que Chris avait rédigées en prévision de son discours de 2019 et qui m'ont été remises par son fils, Tom Skinner.

Dans le présent document qui fera partie du recueil, je résume l'exposé que j'ai donné à cette séance commémorative. À la section 2, je traiterai de la première évolution du contrôle de la divulgation statistique et, à la section 3, de la situation aujourd'hui en fonction de la révolution des données. À la section 4, je décrirai les apports de Chris et ses travaux de pionnier sur le CDS. À la section 5, il sera question des recherches en cours sur ce même contrôle et sur la confidentialité des données, et d'une des dernières contributions de Chris visant à intégrer la confidentialité différentielle dans la trousse d'outils de CDS des organismes gouvernementaux. À la section 6, pour finir, je discuterai du rôle joué par les recherches de Chris dans les statistiques gouvernementales et sociales et la méthodologie d'enquête.

[1]Natalie Shlomo, Social Statistics Department, University of Manchester, UK, natalie.shlomo@manchester.ac.uk

2. Évolution et histoire initiales du CDS

Depuis les années 1960, le public est sensibilisé à la confidentialité et à la protection de la vie privée, ce qui l'a fait s'opposer à la collecte des données, plus particulièrement dans le cas des recensements en Europe. 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 travaux de la première époque de Barabba (1975), Cox (1976), Fellegi (1972) et Dalenius (1974). Fellegi (1972, p. 8) a écrit : [traduction] « Les instituts nationaux de statistique (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 réalisées en Suède, Dalenius (1977) est un des premiers à avoir défini et arrêté un cadre de contrôle de la divulgation statistique. Il a dit : [traduction] « Une partie non autorisée devrait être incapable par la diffusion de statistiques f(D) d'apprendre sur quelqu'un quelque chose qui ne peut s'apprendre sans un accès à f(D). »

Les travaux de Dalenius et d'autres ont établi le cadre voulu pour la recherche-développement sur le CDS au sein des organismes gouvernementaux, ainsi que pour la création en bonne et due forme de conseils de gouvernance de la diffusion des données statistiques. Dans les recherches réalisées aux États-Unis, on peut distinguer, par exemple, le Subcommittee on Disclosure-Avoidance Techniques (sous-comité des techniques de prévention de la divulgation) mis en place en 1976 par le Federal Committee on Statistical Methodology et parrainé par la division de la politique statistique de l'Office of Management and Budget (OMB), comme l'indiquent Jabine et coll. (1977) (voir aussi le rapport de 1978 et son annexe A sur les pratiques de prévention de la divulgation statistique au sein de certains organismes fédéraux). Cette annexe en cinq sections présente des recommandations sur le concept de CDS, la matière de la divulgation, les techniques d'évitement, les effets de la divulgation sur les sujets et les utilisateurs de l'information 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 dans la statistique régionale des données relatives à moins de 100 000 personnes.

De nouveaux travaux dans les années 1980 ont mis l'accent dans le CDS sur les produits 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'effet d'échantillonnage, du bruit additif et de la connaissance déjà acquise dans le cas présumé d'atteinte à l'intégrité de données. Dans son article, Paass (1988) 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 assumée 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. » Autre fait, Bethléem et coll. (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'uniques dans la population compte tenu des uniques dans l'échantillon, sur un ensemble de quasi-identificateurs recoupés. Pour en savoir plus sur cette méthode et les apports de Chris, voir la section 4.1.

Dans les années 1990, la demande de produits détaillés s'est faite beaucoup plus grande, plus particulièrement avec la disponibilité de meilleures solutions technologiques et l'avènement de l'ordinateur personnel. Et les utilisateurs des données s'inquiétaient de plus en plus de ce qu'ils aient à travailler avec des produits protégés ou perturbés. Cela a coïncidé avec le développement à grande échelle du domaine du contrôle de la divulgation statistique par une 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 sur la prévention de la divulgation statistique (voir à ce sujet

le numéro spécial de Statistica Neerlandica, 1993). Notons aussi les collaborations entre membres de l'Union européenne dans le 4e projet-cadre en recherche sur le CDS [1996 à 1998] et dans de nombreux autres projets européens qui ont eu lieu par la suite (mise au point, par exemple, des logiciels mu-ARGUS pour les microdonnées et tau-ARGUS pour les données en tableaux, avec plus particulièrement leur suppression de cellules dans les tableaux de données quantitatives en statistique des entreprises). Voir https://research.cbs.nl/casc/index.htm pour plus de détails sur les projets de recherche à l'échelle de l'Europe. Un numéro spécial du Journal of Official Statistics (vol. 14(4), 1998) sous le titre « Disclosure Limitation Methods for Protecting the Confidentiality of Statistical Data » (méthodes de limitation de divulgation pour la sauvegarde de la confidentialité des données statistiques) a eu un retentissement particulier et a mis en lumière les recherches à grande échelle consacrées au CDS. Un livre et un cours ont vu le jour (voir Willenborg et De Waal [1996] avec la collaboration de Chris Skinner et une réédition en 2001). Le travail se poursuivait parallèlement aux États-Unis et au Canada avec notamment le Federal Committee on Statistical Methodology (1994), le Committee on Maintaining Privacy and Security in Health Care Applications of the National Information Infrastructure (1997) et « Disclosure Control Issues at Statistics Canada » (Yeo et Robertson, 1995) incluant le progiciel CONFID avec ses opérations de suppression de cellules en tableau de données quantitatives.

Tout au long des années 1990, on s'est de plus en plus employé à élaborer des dispositions et des lois en matière d'accès et de gouvernance et à développer la notion d'accès aux données par palier. On a mis en place des archives de données et des centres de données de recherche avec des modes et des cadres de gouvernance de l'information pour une exploitation efficace des données statistiques. Un exemple en est le Cadre des cinq éléments de la sécurité au tableau 1 qui a été mis en application à l'Office for National Statistics en 2002 (Ritchie, 2009). Il a été suivi du cadre décisionnel de l'anonymisation (Elliot et coll., 2016 et https://ukanon.net/framework/).

Tableau 1
Cadre des cinq éléments de la sécurité

Projets sécuritaires	L'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'installation d'accès limite-t-elle l'utilisation non autorisée des données?
Données sécuritaires	Y a-t-il un risque de divulgation qui tient aux données mêmes?
Produits sécuritaires	Est-ce que les résultats statistiques ne se prêtent pas à la divulgation?

3. Révolution des données

Depuis la seconde moitié des années 2000, les données ont été plus ouvertes et accessibles dans le domaine public, qu'il s'agisse de sources ouvertes ou de mégadonnées, d'où un plus grand risque de non-respect de la vie privée et de la confidentialité, puisque ces sources d'information pourraient servir à porter atteinte aux données statistiques diffusées. Ajoutons que l'avancement des outils technologiques disponibles qui améliorent les recoupements et les manipulations de données augmente les chances de réidentification dans les données statistiques. Les organismes gouvernementaux ont mieux pris conscience que les méthodes CDS pourraient ne pas suffire à sauvegarder la confidentialité des unités statistiques et ont donc resserré les restrictions et les contrôles d'accès sur les données. Cela s'est par ailleurs manifesté par des modifications de la réglementation, plus particulièrement avec le Règlement général sur la protection des données (RGPD) de 2016 de l'Union européenne et ses dispositions et prescriptions en matière de traitement des données à caractère personnel des particuliers. On a également insisté davantage sur la protection de la vie privée dans les données sur la santé (El Emam et coll., 2011) et la génétique (Homer et coll., 2008;

Gymrek et coll., 2013). On a démontré dans le cas des données génétiques que le risque était grand et influait sur la diffusion des bases de données d'ADN. Dans le domaine commercial, les exemples d'atteintes à la vie privée abondent : mots clés de recherche AOL (Barbaro et coll., 2006), courses de taxi à New York (Douriez et coll., 2016), Cambridge Analytica et Facebook (Meredith, 2018), et ainsi de suite.

Avec l'avancement technologique et la possibilité de recouper les sources de données, on a songé à ce qu'on appelle des « tiers de confiance » pour faire les liens et sécuriser l'informatique multipartite. Ce dispositif est né des études en informatique et a été repiqué avec les études en statistique sur la façon d'exécuter une modélisation statistique avancée dans cette optique (Slavkovic et Nardi, 2007; Snoke et coll., 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 plus de détails). Parmi les auteurs s'étant exprimés sur les questions de confidentialité, Dwork et coll. (2017) ont écrit : [traduction] « Depuis le milieu des années 2000, la discipline de l'analyse statistique des données pour la protection de la vie privée a assisté à un afflux d'idées formées il y a quelque deux décennies déjà dans le milieu de la cryptographie. »

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

Le travail consacré officiellement par Chris au domaine du contrôle de la divulgation statistique (CDS) a débuté par des collaborations à l'Université de Manchester dans un plaidoyer pour la diffusion de microdonnées d'échantillon (échantillons d'enregistrements anonymisés) tirées du recensement du Royaume-Uni (Marsh et coll., 1991; Skinner et coll., 1994; Marsh et coll., 1994). Il s'est alors intéressé à la mesure du risque de réidentification dans les microdonnées d'enquête par la modélisation probabiliste, objet d'une première publication dans Skinner (1992) que nous allons décrire à la section 4.1. Il a aussi entrepris sa longue carrière à titre de conseiller auprès des comités de statistique gouvernementale et d'accès aux données : UK Census Design and Methodology Advisory Committee, Statistical Disclosure Control (SDC) Subgroup (2008-2010); Understanding Society Data Access Committee (2010-2013); Expert Advisory Group on Data Access avec Wellcome Trust, MRS, ESRC et Cancer Research UK (2012-2014).

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

Le scénario de risque de divulgation dans la diffusion de microdonnées d'échantillon contenant des enregistrements d'une enquête avec échantillon prélevé au hasard sur une population finie repose sur les hypothèses suivantes : (1) il y a un « intrus » (animé de l'intention malveillante de discréditer le bureau de la statistique) qui a accès aux microdonnées et à des renseignements auxiliaires sur la population qui lui permettent de recouper des sources de données de manière à reconnaître des individus 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 est un des premiers à avoir dressé un cadre de modélisation statistique pour l'estimation de la probabilité de réidentification en fonction des données diffusées et des hypothèses quant à 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 d'information, lesquelles sont généralement catégoriques (âge, sexe, lieu, groupe ethnique, etc.). Le recoupement de ces variables clés donne naissance à de grands tableaux de contingence des chiffres de l'échantillon; beaucoup de cellules de ces tableaux ont zéro ou l'unité comme valeur. À cet égard, nous nous intéressons tout

particulièrement au risque de divulgation avec les cellules de valeur un, qui sont les uniques de l'échantillon. Ce risque est fonction de la notion d'unique de population dans le tableau de contingence : s'il y a un unique d'échantillon observé dans la cellule d'un tableau de classification croisée des variables clés, quelle est la probabilité que cet unique soit aussi un unique de la population? Les mesures individuelles du risque par enregistrement sont estimées sous forme de probabilité de réidentification. On agrège ensuite ces mesures pour dégager un risque global à l'échelle du fichier, ce qui sert à prendre une décision éclairée sur le niveau d'accès à ce fichier.

La modélisation probabiliste conçue par Chris adopte une vue simplifiée qui restreint l'information connue des intrus (Skinner et Holmes, 1998; Elamir et Skinner, 2006). Désignons par F_k la taille de population dans la cellule k d'un tableau de variables clés comptant K cellules. Désignons par f_k la taille d'échantillon et posons $\sum_k F_k = N$ et $\sum_k f_k = n$. L'ensemble d'uniques d'échantillon se définit par $SU = \{k: f_k = 1\}$ correspondant aux enregistrements à haut risque possible comme uniques de population. Voici deux mesures globales de risque de divulgation (où I est la fonction indicatrice) :

1. nombre d'uniques d'échantillon qui sont des uniques de population :
   $\tau_1 = \sum_k I(f_k = 1, F_k = 1)$;
2. nombre attendu de justes appariements d'uniques d'échantillon dans l'hypothèse d'une répartition aléatoire dans la cellule k (la probabilité d'appariement) :
   $\tau_2 = \sum_k I(f_k = 1)\ 1/\ F_k$.

Nous supposons par ailleurs que les fréquences de population F_k sont inconnues et doivent être estimées à l'aide d'un modèle probabiliste où les mesures de risque de la forme suivante :

$$\hat{\tau}_1 = \sum_k I(f_k = 1)\hat{P}(\ F_k = 1|f_k = 1) \text{ et } \hat{\tau}_2 = \sum_k I(f_k = 1)\hat{E}(1/\ F_k\ |f_k = 1) \quad (1)$$

Chris a posé une distribution de Poisson et un modèle loglinéaire pour estimer les mesures de risque de divulgation dans (1). Dans ce modèle, lui et ses coauteurs supposent que $F_k \sim Pois(\lambda_k)$ pour chaque cellule k. Un échantillon de Poisson ou de Bernoulli est tiré avec une fraction d'échantillonnage π_k dans la cellule k : $f_k|F_k \sim Bin(F_k, \pi_k)$. Il s'ensuit que :

$$f_k \sim Pois(\pi_k\lambda_k) \text{ et } F_k|f_k \sim Pois(\lambda_k(1 - \pi_k)) \quad (2),$$

où, pour une cellule, le chiffre de population F_k est supposé indépendant du chiffre d'échantillon f_k.

Les paramètres λ_k font l'objet d'une estimation par modélisation loglinéaire. Les fréquences d'échantillon $f_k$sont issues de distribution indépendantes de Poisson avec une moyenne $\mu_k = \pi_k\lambda_k$. Un modèle loglinéaire pour les μ_k prend la forme $\log(\mu_k) = x_k{}'\boldsymbol{\beta}$, où x_k est un vecteur par plan qui désigne les principaux effets et éléments d'interaction du modèle pour les variables clés. On obtient l'estimateur de maximum de vraisemblance (EMV) $\widehat{\boldsymbol{\beta}}$ en résolvant les équations de valeurs résultantes :

$$\sum_k(\ f_k - \pi_k \exp(x_k'\boldsymbol{\beta}))x_k = 0 \quad (3).$$

Les valeurs ajustées sont alors calculées par $\hat{\mu}_k = \exp(x_k'\widehat{\boldsymbol{\beta}})$ et $\hat{\lambda}_k = \hat{\mu}_k/\pi_k$. Voici les mesures individuelles du risque de divulgation pour la cellule k :

$$P(\ F_k = 1|f_k = 1) = \exp\ (\lambda_k(1 - \pi_k))$$

$$E(1/\,F_k\,|f_k = 1) = (1 - \exp\,(\lambda_k(1-\pi_k)))/\,(\lambda_k(1-\pi_k)) \qquad (4).$$

Si on introduit $\hat{\lambda}_k$ pour λ_k dans (4), on obtient les estimations $\hat{P}(\,F_k = 1|f_k = 1)$ et $\hat{E}(1/\,F_k\,|f_k = 1)$, puis $\hat{\tau}_1$ et $\hat{\tau}_2$ dans (1).

Skinner et Shlomo (2008) conçoivent une méthode de sélection des principaux effets et éléments d'interaction pour le modèle loglinéaire par une estimation et une minimisation (approximative) du biais des estimations de risque $\hat{\tau}_1$ et $\hat{\tau}_2$. En définissant $h(\lambda_k) = P(\,F_k = 1|f_k = 1)$ pour τ_1 et $h(\lambda_k) = E(1/\,F_k\,|f_k = 1)$ pour τ_2, ils considèrent l'expression :

$$B = \sum_k E(I(f_k = 1))(\,h(\hat{\lambda}_k) - h(\lambda_k)).$$

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

$$B \approx \sum_k \pi_k\,\lambda_k \exp\,(-\lambda_k)(\,h'(\lambda_k)(\hat{\lambda}_k - \lambda_k) + h''(\lambda_k)(\hat{\lambda}_k - \lambda_k)^2/2).$$

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

$$\hat{B} \approx \sum_k\,\hat{\lambda}_k \exp\,(-\pi_k\hat{\lambda}_k)(\,-h'(\lambda_k)(f_k - \pi_k\hat{\lambda}_k) + h''(\lambda_k)((f_k - \pi_k\hat{\lambda}_k)^2 - f_k)/(2\pi_k)) \qquad (5).$$

Voici un exemple pour τ_1 :

$$\hat{B}_1 \approx \sum_k\,\hat{\lambda}_k \exp(-\hat{\lambda}_k)\,(1-\pi_k)\{\,(f_k - \pi_k\hat{\lambda}_k) + (1-\pi_k)[(f_k - \pi_k\hat{\lambda}_k)^2 - f_k]/(2\pi_k)\} \qquad (6).$$

La méthode permet de sélectionner le modèle à l'aide d'un algorithme de recherche prospective qui minimise l'estimation du biais centré réduit $\hat{B}_i/\sqrt{\hat{v}_i}$ pour $\hat{\tau}_i$,, i= 1,2, laquelle sert de critère de qualité de l'ajustement où les $\hat{v}_i$ sont des estimations de la variance des $\hat{B}_i$. Les critères de qualité de l'ajustement $\hat{B}_i/\sqrt{\hat{v}_i}$ présentent en valeur approchée une distribution normale centrée réduite dans l'hypothèse d'une espérance nulle des $\hat{B}_i$.

Skinner et Shlomo (2008) traitent également de l'estimation des mesures de risque de divulgation dans des plans de sondage complexes avec stratification, mise en grappes et poids d'enquête. Si la méthode décrite pose comme hypothèse que tous les individus membres de la cellule k sont sélectionnés indépendamment par échantillonnage de Bernoulli, $(f_k = 1|F_k) = F_k\pi_k(1-\pi_k)^{F_k-1}$, cela pourrait ne pas être le cas dans un échantillonnage en grappes (ménages). Dans la pratique, les variables clés sont normalement l'âge, le sexe, la profession, etc. qui sont généralement en chevauchement sur les grappes. L'hypothèse qui précède se vérifie donc dans la pratique pour la plupart des enquêtes auprès des ménages et n'entache pas d'un biais l'estimation des mesures de risque. Par ailleurs, les probabilités d'inclusion sont susceptibles de varier selon les strates. La stratification la plus courante est géographique. Des indicateurs de stratification devraient toujours figurer dans les variables clés pour rendre compte des différences de probabilités d'inclusion dans le modèle loglinéaire. Dans un échantillonnage complexe, les λ_k peuvent être estimés avec cohérence dans un traitement de pseudo-maximum de vraisemblance (Rao et Thomas, 2003), où l'équation d'estimation en (3) se trouve ainsi modifiée :

$$\sum_k(\,\hat{F}_k - \exp(x_k'\boldsymbol{\beta}))x_k = 0 \qquad (7),$$

et où on obtient les $\hat{F}_k$ en sommant 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 en (4) et π_k est remplacé par l'estimation $\hat{\pi}_k = f_k / \hat{F}_k$. Le critère de qualité d'ajustement $\hat{B}$ est aussi adapté à la méthode du pseudo-maximum de vraisemblance. Voir Skinner et Shlomo (2008) pour une simulation et une application réelle démontrant la justesse de ce traitement tant pour un échantillon aléatoire simple que pour un plan de sondage complexe.

Avec la modélisation probabiliste présentée ici et dans d'autres études spécialisées apparentées, l'hypothèse est qu'il n'y a aucune erreur de mesure dans la façon dont les données sont enregistrées. Il faut savoir que, en dehors des erreurs habituelles de saisie des données, les variables clés peuvent être mal classées à dessein comme moyen de masquage de l'information, ce qui peut se faire, par exemple, par échange d'enregistrements ou par la méthode de postrandomisation (PRAM, Gouweleeuw et coll., 1998). Shlomo et Skinner (2010) adaptent l'estimation du risque de réidentification de manière à tenir compte des erreurs de mesure. Si nous désignons les variables clés recoupées dans la population et les microdonnées par *X* et supposons que les *X* des microdonnées sont entachés d'une certaine faille de classification ou d'une certaine erreur de perturbation dénotée par la valeur $\tilde{X}$ et déterminée indépendamment par une matrice de défaut de classification *M*,

où $M_{kj} = P(\tilde{X} = k | X = j)$ (8)

,

la mesure du risque de divulgation par appariement avec un unique d'échantillon en situation d'erreur de mesure devient :

$$\frac{M_{kk}(1-\pi_k M_{kk})}{\sum_j F_j M_{kj}/(1-\pi_k M_{kj})} \leq \frac{1}{F_k} \quad (9).$$

Sous l'hypothèse de petites fractions d'échantillonnage et de légères fautes de classification, la mesure de risque de divulgation peut être approchée par $M_{kk} / \sum_j F_j M_{kj}$ ou $M_{kk} / \tilde{F}_k$, où $\tilde{F}_k$ est le chiffre de population avec $\tilde{X} = k$. Après agrégation des mesures de risque de divulgation selon les enregistrements, la mesure globale de risque devient :

$$\tau_2 = \sum_k I(f_k = 1)\, M_{kk} / \tilde{F}_k \quad (10).$$

Il convient de noter que, dans le calcul de cette mesure, seule la diagonale de la matrice de défaut de classification doit être connue, soit les probabilités de ne pas être perturbé doivent être connues. Les chiffres de population sont généralement inconnus, de sorte que l'estimation en (10) peut s'obtenir par modélisation probabiliste sur l'échantillon en défaut de classification comme ci-dessus :

$$\hat{\tau}_2 = \sum_k I\left(\tilde{f}_k = 1\right) M_{kk}\, \hat{E}(1/\tilde{F}_k | \tilde{f}_k) \quad (11).$$

Dans des travaux plus récents avec Chris qui sont présentés pour la première fois dans Shlomo et Skinner (à paraître), une nouvelle direction est expérimentée pour la mesure du risque de réidentification dans le cas des sources de données non probabilistes. Plus précisément, il y a des registres du domaine public où l'appartenance est inconnue et revêt un caractère sensible. Comme exemples de tels registres, mentionnons le recensement des gens souffrant d'un état médical comme le cancer ou le VIH ou encore le dénombrement des adhérents à un programme de cartes de fidélisation. On peut étendre la chose au cas où des échantillons sont prélevés sur des registres et plus généralement aux échantillons non probabilistes issus, par exemple, d'enquêtes Web. Si nous élargissons le cadre qui précède, les microdonnées d'un échantillon aléatoire peuvent encore servir à estimer les paramètres de population avec un cadre

de modélisation probabiliste s'il s'agit de dégager le risque de réidentification, mais avec comme complication d'avoir à estimer en même temps la propension à l'appartenance chez les individus membres du registre.

Plus précisément, soit U et U_1 désignant la population tout court et la population d'un registre respectivement avec $U_1 \subset U$. Soit R la variable indicatrice du registre pour l'individu i avec R_i =1 si $i \in U_1$ et avec R_i = 0 sinon. Comme nous l'avons mentionné, nous supposons que l'appartenance R_i est une variable sensible dont la divulgation n'est pas souhaitable.

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

$$\tau_1^* = \sum_k P(F_k = 1|F_k^1 = 1)I(F_k^1 = 1) \quad (12).$$

Il n'y a aucun moyen possible d'estimer ces mesures avec cohérence à partir des seules microdonnées du registre. Les microdonnées nous renseignent sur les F_k^1 , mais non sur les F_k de U. La distribution de X dans U_1 peut être fort différente de celle de U, si bien que les microdonnées ne donneront aucun renseignement direct sur les F_k. Nous nous reportons donc au fichier de microdonnées d'échantillon aléatoire où les valeurs de X sont connues pour un échantillon probabiliste s de U. Soit f_k la fréquence de la cellule k dans s. À noter que les f_k et F_k^1 sont observés au contraire des F_k. Si l'intrus a accès au fichier de microdonnées d'échantillon, on pourrait avoir intérêt à concentrer son attention sur les cellules où $f_k = 1$, ce qui donne la mesure suivante du risque :

$$\tau_1 = \sum_k P(F_k = 1|F_k^1 = 1, f_k = 1)I(F_k^1 = 1, f_k = 1) \quad (13).$$

Comme Skinner et Shlomo (2008), supposons que F_k épouse une distribution de Poisson, $F_k \sim Pois(\lambda_k)$, où le paramètre λ_k suit le modèle loglinéaire :

$$\log(\lambda_k) = x_k'\boldsymbol{\beta} \quad (14).$$

Supposons que, dans la cellule k, la variable d'appartenance inconnue R_i prend la valeur 1 avec la probabilité p_k indépendamment de chacune des unités F_k, d'où $F_k^1 \sim Pois(\phi_k)$ avec $\phi_k = \lambda_k p_k$ et où les F_k^1 sont en distribution binomiale $F_k^1 | F_k \sim Bin(F_k, p_k)$ conditionnellement sur les F_k. Nous supposons en outre que p_k suit le modèle logistique :

$$logit(p_k) = x_k'\boldsymbol{\xi} \quad (15).$$

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

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

Pour évaluer τ_1^*, nous prenons

$$P(F_k = 1|F_k^1 = 1) = P(F_k - F_k^1 = 0) = \exp(-(\lambda_k - \phi_k)) \quad (16),$$

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

Aussi l'estimation de ces mesures exige-t-elle à la fois l'estimation de $\boldsymbol{\beta}$ à partir de $f_k \sim Pois(\pi_k \lambda_k)$ et, dans une seconde étape, l'estimation $\boldsymbol{\xi}$ fixant λ_k à la valeur impliquée par (14). Nous prenons alors (16) et $\phi_k = \lambda_k p_k$ pour écrire

$$log\phi_k = log\lambda_k + x_k'\boldsymbol{\xi} - log\,(1 + \exp(x_k'\boldsymbol{\xi})) \qquad (17)$$

et estimer $\boldsymbol{\xi}$ du fait que $F_k^1 \sim Pois(\phi_k)$ en recourant à une estimation de maximum de vraisemblance et en traitant λ_k comme connu. D'autres modes d'estimation sont proposés dans Shlomo et Skinner (étude à paraître).

Un autre type d'estimateur par plan de mesure du risque de divulgation dans des microdonnées d'échantillon est ce qu'on appelle la mesure DIS mise au point dans Skinner et Elliot (2002) et étendue dans Skinner et Carter (2003) à des plans de sondage plus complexes. 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 retrouve 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 de pêche »). À noter qu'il 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 avec une unité de la population. L'avantage avec le « scénario de pêche » est que la mesure peut facilement s'estimer sans qu'on ait à passer par une modélisation probabiliste. Voici comment se définit la mesure DIS :

$$\theta = \sum_k I(f_k = 1) / \sum_k I(f_k = 1) F_k \qquad (18).$$

Elle est estimée par :

$$\hat{\theta} = \pi n_1 / [\pi n_1 + 2(1 - \pi) n_2] \qquad (19),$$

où les n_1 sont les uniques et les n_2 sont les doubles. Skinner et Shlomo (2012) étendent cette approche à l'estimation de la fréquence des fréquences de populations finies au-delà des uniques de l'échantillon.

4.2 Distinction entre le risque et le préjudice de divulgation

Chris a proposé dans Skinner (2012) un cadre conceptuel pour distinguer le risque du préjudice de 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 insiste sur l'importance de distinguer ce qui peut se mesurer par la théorie statistique (risque de divulgation possible) et de voir quels aspects de la décision exigent d'autres apports, comme le jugement de politique (préjudice de divulgation possible). Ce travail a comme motivation le cadre d'utilité des données de risque de divulgation dans Duncan et coll. (2001) et les aspects économiques de la confidentialité dans Abowd et Schmutte (2009).

Comme le font voir ces exemples, Chris a accru la profondeur et l'étendue des recherches consacrées au CDS. D'autres recherches considérablement marquées par l'influence de Chris concernent les associations entre la mesure des risques de divulgation CDS et d'autres recherches apparentées portant, par exemple, sur le couplage d'enregistrements (Skinner, 2009) et les sciences judiciaires (Skinner, 2007). Les travaux plus récents de Chris sur le risque de divulgation et la confidentialité seront l'objet des prochaines sections.

5. Risque de divulgation et confidentialité

Dans les études en informatique sur la confidentialité, on trouve des définitions plus formelles de la confidentialité avec des modèles qui protègent contre une catégorie d'atteintes et où les modèles sont paramétrés par des valeurs de seuil de risque de divulgation établies a priori selon un budget de protection des renseignements personnels. En général, les auteurs spécialisés dans les questions de confidentialité réclament des techniques plus perturbatrices et une plus grande perte d'information en fonction des seuils de la modélisation de la confidentialité. La catégorie d'atteintes a normalement pour racine le traitement de la divulgation par déduction qui comprend à la fois les risques de divulgation par identité et par attributs, bien que le modèle de traitement dit du k-anonymat (Sweeney, 2002) vise à prévenir des atteintes par recoupement comme dans les études du CDS que nous décrivons à la section 4. À souligner l'importance sans cesse renouvelée de mesurer la divulgation par identité selon le risque de réidentification compte tenu de la réglementation et des atteintes possibles par recoupement, bien que, dans les études spécialisées sur la confidentialité, on ne fasse pas la distinction entre variables d'identification et variables de sensibilité.

Il convient toutefois de noter que nombre de concepts présents dans les études consacrées à la confidentialité n'ont rien de nouveau du point de vue du contrôle de la divulgation statistique. Ainsi, les atteintes par reconstruction que mentionnent Garfinkel et coll. (2018) dans leur plaidoyer pour une protection CDS plus stricte dans le recensement américain de 2021 font appel aux mêmes considérations que pour la suppression complémentaire de cellules imaginée dans les années 1980 dans la protection des tableaux de données quantitatives des entreprises, et notamment le calcul des bornes supérieures et inférieures sur les cellules en suppression. Bien sûr, l'atteinte par reconstruction ne concerne pas le recoupement, mais plutôt la divulgation par attributs à cause de petits chiffres de cellule, plus particulièrement de valeurs marginales. Comme il a été mentionné, les études de la confidentialité ont pour objet premier la divulgation par attributs et la divulgation par déduction, bien qu'on sache que les études du CDS ont traité d'autres thèmes, du risque de divulgation prédictive mentionné dans Fuller (1993), par exemple. Un autre modèle de confidentialité vise la divulgation par repérage où on peut établir par inférence si un individu figure dans un ensemble de données sensibles (voir Homer et coll. (2008), par exemple), mais les auteurs spécialisés en CDS se sont aussi demandé en priorité si un sujet de l'information était visible dans l'ensemble de données. La différence entre le traitement spécialisé de la confidentialité et le contexte CDS réside dans ce que l'unité statistique consent à fournir ses données à des fins statistiques et que l'organisme gouvernemental ait donc l'obligation en droit et en éthique d'assurer une protection contre les problèmes de divulgation.

Depuis 2005, il y a eu quatre réunions de concertation entre le milieu du CDS et le milieu en informatique de la confidentialité, ce qui a fait largement comprendre les approches des uns et des autres pour la sauvegarde de la confidentialité et le maintien d'une utilité suffisante des données. Dans Nissim et coll. (2017, p. 5) par exemple, il est dit : [traduction] « La confidentialité est une propriété d'une relation d'information entre entrée et sortie, et non de la seule sortie. » Cela a amené à relâcher quelque peu les strictes garanties de sauvegarde dans les études de la confidentialité. Pour sa part, le milieu du CDS a reconnu le besoin de se doter de garanties plus formelles en matière de confidentialité, plus particulièrement si on considère l'exigence d'un accès aux données statistiques par le canal des applications de diffusion Web. Les collaborations entre le milieu du CDS et le milieu de la confidentialité en informatique ont fait naître une revue lancée en 2005 sous le titre *Journal of Privacy and Confidentiality* (https://journalprivacyconfidentiality.org), dont Chris a été un des premiers corédacteurs (Abowd et coll., 2009).

4.3 Confidentialité différentielle

Dwork et Naor (2010) démontrent que la définition par Dalenius (1977) d'atteinte à la confidentialité à la section 2 est inévitable en proposant non pas de comparer l'information avec et sans f(D), mais plutôt f(D) et f(D'), où D' est la base de données D sans une unité particulière. C'est ce qu'on appelle le modèle de confidentialité différentielle (Dwork et coll., 2006).

Dans un tel modèle, on permet un « scénario du pire » où l'éventuel intrus dispose d'une information complète sur toutes les unités de la base de données sauf l'unité d'intérêt. La définition d'un mécanisme de perturbation M répond à la définition de confidentialité différentielle « ε » si, pour toutes les recherches dans des bases de données voisines $D, D' \in A$ différant par un individu et pour tous les résultats possibles définis comme sous-ensembles $S \epsilon Range(M)$, nous avons :

$$p(M(D) \epsilon S) \leq e^{\varepsilon} p(M(D') \epsilon S).$$

Un assouplissement s'offre par la définition de la confidentialité différentielle (Ɛ,δ) :

$$p(M(D) \epsilon S) \leq e^{\varepsilon} p(M(D') \epsilon S) + \delta.$$

Cela signifie que l'observation d'une sortie perturbée S n'apprend à peu près rien (jusqu'à un degré de e^{ε}) et que l'intrus est incapable de déterminer si la sortie vient de la base de données D ou D'. En d'autres termes, le rapport $p(M(D) \epsilon S) / \ p(M(D') \epsilon S)$ est borné et la probabilité au dénominateur ne peut être nulle. Dans ce cas, la confidentialité différentielle constitue une borne formelle pour un risque accru de divulgation par participation à la base de données. Dans le cadre de la confidentialité différentielle (Ɛ,δ), un léger glissement est permis pour cette contrainte.
La solution garantissant une confidentialité différentielle chez les auteurs spécialisés en informatique consiste à ajouter du bruit ou de la perturbation aux sorties de la recherche sous un paramétrage bien précis. Chez les auteurs en confidentialité, le bruit est produit par la distribution de Laplace (pour les données quantitatives, une distribution de Laplace discontinue peut être employée).

Shlomo et Skinner (2012) se sont d'abord demandé si les méthodes CDS normalisées sont des mécanismes au caractère différentiellement privé selon la définition qui précède. Ils ont constaté que l'échantillonnage comme méthode CDS n'a pas ce caractère, puisque quelqu'un peut être observé dans l'échantillon (protégé), mais que s'il est retiré de la base d'information D pour l'obtention de D', il s'ensuit une situation impossible où le chiffre d'échantillon est supérieur au chiffre de population. En fait, toute méthode de CDS non perturbatrice, par le recours au grossissement de variables par exemple, n'est pas différentiellement privée, car il est toujours possible de relever un cas où le dénominateur du rapport sera nul en raison du caractère déterministe de la protection des données. On peut rendre des méthodes perturbatrices différentiellement privées dans la trousse d'outils CDS si les mécanismes en cause n'ont pas de probabilités nulles de perturbation. Pour prendre un exemple, les méthodes CDS ne vont pas habituellement perturber les cellules zéro dans les tableaux de recensement présentant des chiffres de population entière, mais induisent stochastiquement plus de zéros par perturbation, s'ils procèdent par arrondissement aléatoire, par exemple. Si le but est toutefois de rendre le mode de perturbation différentiellement privé, les zéros (aléatoires) du tableau doivent aussi être perturbés.

5.2 Générateurs de tableaux flexibles en ligne

Les organismes gouvernementaux ont manifesté beaucoup d'intérêt pour la génération de tableaux flexibles en ligne en vue de l'obtention de tableaux de recensement permettant à l'utilisateur de définir et télécharger ses propres totalisations censitaires, d'ordinaire à l'aide d'un site Web dédié avec un ensemble préétabli de variables et leurs catégories sélectionnées sur listes déroulantes. De légers contrôles de divulgation portent sur les tableaux ainsi produits avant leur diffusion. Une de ces applications a été conçue par l'Australian Bureau of Statistics ou ABS (voir : https://www.abs.gov.au/statistics/microdata-tablebuilder/tablebuilder). Cette application utilise un vecteur de perturbation pour changer les valeurs quantitatives des cellules en fonction de leur valeur initiale. Le mécanisme de perturbation a comme propriétés d'être borné, exempt de biais et à entropie maximale et permet seulement une perturbation non négative sans toucher aux cellules portant des zéros. Shlomo et Young (2008) ont proposé une approche de transformation telle des vecteurs de perturbation que les chiffres marginaux soient préservés en espérance par l'introduction d'une 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 individu dans les microdonnées de recensement. Si un tableau est demandé et que les individus sont agrégés dans ses cellules, il y aura aussi agrégation des chiffres aléatoires de ces individus dans chaque cellule. La valeur aléatoire agrégée sera ensuite la « graine » (seed en anglais) de la perturbation à effectuer (Fraser et Wooton, 2005). Dans ce cas, chaque fois qu'une cellule réapparaît dans tout tableau de recensement demandé, elle portera toujours la même perturbation. Il n'y a donc aucun risque de pouvoir « déconstruire » une valeur réelle de cellule en faisant la moyenne de perturbations indépendantes avec des demandes multiples d'un même tableau. On s'assure en outre avec ce traitement que la perturbation relève d'un mécanisme « non interactif », puisque, pour l'essentiel, tous les résultats de perturbation dans les tableaux censitaires demandés seront connus d'avance avec le générateur de flexibilité en ligne.

Une des dernières initiatives de Chris avant sa maladie avait été de prendre l'initiative de créer 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 avec succès un programme de croisement et d'anonymisation des données (appuyé par la subvention EP/K032208/1 de l'Engineering and Physical Sciences Research Council (EPSRC) 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 un générateur de tableaux flexibles en ligne dans la production de tableaux de recensement. Le résultat en est l'étude produite par Rinott, O'Keefe, Shlomo et Skinner (2018). La grande différence avec l'approche initiale de l'ABS a été d'utiliser un mécanisme de perturbation différentiellement privé (appelé mécanisme exponentiel et consistant pour l'essentiel en une distribution discontinue de Laplace) et de perturber les cellules à zéros (aléatoires). Toute perturbation négative obtenue était alors ramenée à zéro dans les tableaux du recensement. Dans une hypothèse de perturbations indépendantes, le mécanisme exponentiel se définit ainsi : pour une valeur quantitative donnée de cellule a, on choisit $b \epsilon B$ (où B est la plage des b) avec une probabilité proportionnelle à $\exp\left(\frac{\left(\frac{\varepsilon}{2}\right)u}{\Delta u}\right)$ où u est la perturbation et où Δu est la différence maximale de chiffre de cellule entre les bases de données D et D', ce qui correspond à la valeur de un dans le cas d'un tableau censitaire formé de cellules internes. Si on tient compte des valeurs marginales dans les tableaux, on élève la complexité du vecteur de perturbation (voir Rinott et coll. (2018) pour plus de détails sur les valeurs marginales). Pour garantir l'utilité, les valeurs de perturbation ont été limitées à ± 7 et, par conséquent, le mécanisme respectait le critère de confidentialité différentielle (ε, δ). Voici un exemple de vecteur de perturbation avec $\varepsilon = 1{,}5$ et $\delta = 0{,}00002$ et une borne de ± 7 :

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

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

On trouvera des exemples et des applications dans Rinott et coll. (2018). On y verra aussi comment adapter les analyses statistiques lorsqu'elles portent sur des données perturbées, si nous considérons que le mécanisme de perturbation est connu et non secret dans un traitement de confidentialité différentielle.

Jointe à des garanties plus formelles par plan contre les risques de divulgation par attributs et de divulgation par déduction, la confidentialité différentielle peut fournir des solutions de protection des données statistiques diffusées en source ouverte et par des applications Web; elle peut trouver sa place dans la trousse d'outils CDS des organismes gouvernementaux. Le US Census Bureau appliquera un cadre de confidentialité différentielle à ses produits de recensement en 2021 (Abowd, 2018). Il faudra pousser la recherche pour voir en quoi les budgets de confidentialité se trouvent influencés lorsqu'on combine ce cadre à d'autres cadres CDS comme ceux du grossissement, de l'échantillonnage et de la suppression de variables. Chez les auteurs en confidentialité, la recherche se poursuit sur les gains d'utilité dans les mécanismes de perturbation différentiellement privés avec, par exemple, la confidentialité différentielle bornée que proposent Kifer et Machanavajjhala (2014).

6. Mot de la fin

Bref, un des grands traits de la façon dont Chris 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 eu de l'influence, parce qu'il a su passer de la théorie à la pratique et résoudre des problèmes concrets pour l'avancement des sciences sociales, des statistiques gouvernementales et sociales et de la méthodologie d'enquête. Il a définitivement été le porte-parole d'une génération avec ses décennies d'étude du CDS et d'autres thèmes de recherche en statistique des enquêtes, qu'il s'agisse des données manquantes et des erreurs de mesure, de l'intégration des données, de l'analyse de plans de sondage complexes ou de l'estimation de bases multiples, pour ne citer que ces exemples.

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).

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.

Barbaro, M., et Zeller Jr, T. (9 août 2006), A face is exposed for AOL searcher no. 4417749. New York Times.

Bethlehem, J., Keller, W., et Pannekoek, J. (1990), Disclosure Control of Microdata. JASA 85, 38-45.

Cox, L.H. (1976), Statistical disclosure in publication hierarchies. Rapport n° 14 du projet de recherche « Confidentiality in Surveys », Département de statistique, Université de Stockholm.

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. Statistique Suède.

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 (with discussion). JASA, 81, 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.

Dwork et Naor (2010), On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. Journal of Privacy and Confidentiality.

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. Université Harvard.

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

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

Fellegi, I.P. (1972), On the question of statistical confidentiality. JASA, 7-18.

Fraser, B., et Wooton, J. (2005), A Proposed Method for Confidentialising Tabular Output to Protect Against Differencing. Séance de travail entre la CEE-ONU et Eurostat sur la confidentialité des données statistiques, Genève, 9-11 novembre.

Fuller, W.A. (1993), Masking Procedures for Micro-data Disclosure Limitation. JOS 9, 383-406.

Garfinkel, S., Abowd, J.M., et Martindale, C. (2018), Understanding Database Reconstruction Attacks on Public Data. ACM QUEUE 16 (5)

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., et coll. (2013), Identifying personal genomes by surname inference. Science, 339 (6117) 321-324.

Homer, N., Szelinger, M., Redman, D., et coll. (2008), Resolving individuals contributing trace amounts of DNA to highly complex mixures using high-density SNP genotyping microarrays. PLoS Genet, 4(8).

Hundepool, A., et coll. (1998), mu-ARGUS user's manual and tau-ARGUS User's Manual. Department of Statistical Methods, Statistique Pays-Bas, Pays-Bas.

Jabine, T.B., Michael, J.A., et Mugge, R.H. (1977), Federal Agency Practices for Avoiding Stastical Disclosure: Findings and Recommendations. Asasrms.org.

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

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., et 6 coauteurs (1991), A case for samples of anonymised records from the 1991 Census. Journal of the Royal Statistical Society, A, 154, 305-340.

Meredith, S. (10 avril 2018), Facebook-Cambridge Analytica: A timeline of the data hijacking scandal". CNBC.

Nissim, K., Steinke, T. Wood, A., Altman, M., et 5 coauteurs (2017), Differential Privacy: A Primer for a Non-technical Audience. Disponible à : 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 (dir. R.L. Chambers et C.J. Skinner), p. 85–108. Chichester, Royaume-Uni : Wiley.

Rinott, Y., O'Keefe, C., Shlomo, N., et Skinner, C. (2018), Confidentiality and Differential Privacy in the Dissemination of Frequency Tables. Statistical Sciences, vol. 33, nº 3, 358-385.

Ritchie, F. (2009), Designing a national model for data access. Comparative Analysis of Enterprise (Micro) Data 2009.

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

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, vol. 4, numéro 1.

Shlomo, N., et Skinner, C.J. (à paraître), Measuring Risk of Re-identification in Microdata: State-of-the Art and New Directions. À paraître dans Journal of the Royal Statistical Society, série A.

Shlomo, N., et Young, C. (2008), Invariant Post-tabular Protection of Census Frequency Counts, dans PSD'2008 Privacy in Statistical Databases, (dir. J. Domingo-Ferrer et Y. Saygin), Springer LNCS 5261, 77-89.

Skinner, C.J. (2007), The probability of identification: applying ideas from forensic statistics to disclosure risk assessment. Journal of the Royal Statistical Society, série 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, 29-30 mai 2008, Publications méthodologiques et documents de travail statistiques, Luxembourg, Communautés européennes, 11-23.

Skinner, C.J., Marsh, C., Openshaw, S., et Wymer, C. (1994), Disclosure Control for Census Microdata. Journal of Official Statistics 10, 31-51.

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 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 Carter, R.G. (2003), Estimation d'une mesure du risque de divulgation pour les microdonnées d'enquête dans le cadre d'un échantillonnage probabiliste inégal. Techniques d'enquête 29, 177-180.

Skinner, C.J., et Shlomo, N. (2008), Assessing identification risk in survey microdata using Log Linear models. JASA 103 (483), 989-1001.

Skinner, C.J. (2012), Statistical disclosure risk: separating potential and harm. Revue internationale de statistique, 80, 349-368, avec analyse et réplique 379-391.

Skinner, C.J., et Shlomo, N. (2012), Estimating Frequencies of Frequencies in Finite Populations. Statistics and Probability Letters, vol. 82, 2206-2212.

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 (1993). Numéro spécial : Proceedings of the International Symposium on Statistical Disclosure Avoidance, vol. 46, nº 1.

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, Lecture Notes in Statistics, 111. New York: Springer Verlag.

Willenborg, L., et De Waal, T. (2001), Elements of Statistical Disclosure Control in Practice. Lecture Notes in Statistics, 155. New York: Springer Verlag.

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-fra.pdf.