Nº 11-522-X au catalogue
ISSN 1709-822X

Recueil du Symposium de 2022 de Statistique Canada : Désagrégation des données : dresser un portrait de données plus représentatif de la société

Contrôle de la divulgation statistique et groupes présentant un intérêt particulier : une perspective européenne

par Peter-Paul de Wolf et Eric Schulte Nordholt

Date de diffusion : le 25 mars 2024

Statistique Canada Statistics Canada

Canada

Recueil du Symposium de 2022 de Statistique Canada
Désagrégation des données : Dresser un portrait de données plus représentatif de la société

Contrôle de la divulgation statistique et groupes présentant un intérêt particulier : une perspective européenne

Peter-Paul de Wolf et Eric Schulte Nordholt[1,2]

Résumé

Dans le contexte de la disponibilité de sources de données plus vastes et plus diverses, les instituts statistiques en Europe sont enclins à publier des statistiques sur des groupes plus petits qu'auparavant. En outre, des épisodes mondiaux à forte incidence, comme la crise de la COVID-19 et la situation en Ukraine, peuvent également nécessiter des statistiques sur des sous-groupes particuliers de personnes. La publication de données concernant de petits groupes ciblés soulève non seulement des questions sur la qualité statistique des chiffres, mais aussi sur le risque de divulgation statistique.

Le principe du contrôle de la divulgation statistique ne dépend pas de la taille des groupes sur lesquels les statistiques sont basées. Cependant, le risque de divulgation dépend de la taille du groupe : plus un groupe est petit, plus le risque est élevé. Les méthodes classiques de gestion du contrôle de la divulgation statistique lorsque la taille des groupes est réduite comprennent la suppression de données et le regroupement des catégories. Pour l'essentiel, ces méthodes consistent à augmenter la taille (moyenne) des groupes. Des approches plus récentes incluent des méthodes de perturbation des données visant à maintenir des groupes de petite taille pour préserver le plus d'information possible, tout en réduisant suffisamment le risque de divulgation.

Dans le présent article, nous mentionnerons quelques exemples européens de statistiques sur des groupes types présentant un intérêt particulier et évoquerons les implications sur le contrôle de la divulgation statistique. Nous aborderons, en outre, certains problèmes liés à l'utilisation de méthodes de perturbation des données, à savoir leur incidence sur le risque de divulgation et sur l'utilité, ainsi que les défis liés à une bonne communication à ce sujet.

Mots clés : divulgation; risque; groupes types; perturbation.

1. Introduction

1.1 Contexte

En Europe, les instituts nationaux de statistique (INS) sont liés au sein du système statistique européen (SSE). Au sein du SSE, Eurostat joue le rôle d'un institut européen de coordination qui combine les statistiques nationales aux statistiques européennes. Pour atteindre cet objectif, Eurostat essaie d'harmoniser les types de données qui doivent être recueillies ainsi que le niveau de diffusion des statistiques par les États membres de l'Union européenne. L'harmonisation du niveau concerne principalement les catégories de variables utilisées dans les publications tabulaires. Certaines statistiques produites par les INS sont obligatoires en vertu de la réglementation européenne.

Il va de soi que les INS appliquent des techniques de contrôle de la divulgation statistique (CDS) à leurs publications nationales afin d'empêcher la diffusion de renseignements susceptibles d'être liés à des unités individuelles identifiables. Les INS ont un deuxième moyen de diffuser leurs statistiques, l'envoi de leurs données à Eurostat, qui les combine ensuite en statistiques européennes. Les statistiques européennes nécessitent alors une deuxième série de techniques de CDS.

---

[1] Les opinions exprimées dans l'article sont celles des auteurs et ne reflètent pas nécessairement la politique du Bureau central de la statistique des Pays-Bas (CBS).

[2] Peter-Paul de Wolf, Bureau central de la statistique des Pays-Bas, Henri Faasdreef 312, La Haye, Pays-Bas, 2492JP, (pp.dewolf@cbs.nl); Eric Schulte Nordholt, Bureau central de la statistique des Pays-Bas, Henri Faasdreef 312, La Haye, Pays-Bas, 2492JP (e.schultenordholt@cbs.nl).

Il fut un temps où les INS (et Eurostat) diffusaient leurs statistiques sous forme de tableaux, c'est-à-dire sous forme de données agrégées. De nos jours, ces données tabulaires sont parfois présentées au moyen de visualisations. Dans ce cas, les visualisations sont souvent fondées sur des données tabulaires auxquelles des techniques de CDS ont déjà été appliquées. La tendance récente dans la diffusion est la demande de plus en plus grande de détails. Chercheurs, décideurs politiques, instituts gouvernementaux, tous sont à la recherche d'informations détaillées. Pour répondre à cette quête, les renseignements statistiques doivent concerner de petits groupes d'unités. Des événements mondiaux à incidence élevée, comme la récente crise de la COVID-19 et la situation en Ukraine, peuvent eux aussi nécessiter des statistiques sur des sous-groupes de la population présentant un intérêt particulier.

La publication sur de petits groupes ciblés d'unités soulève non seulement des questions sur des aspects de la qualité statistique, comme l'exactitude, mais aussi sur le contrôle de la divulgation statistique. Au fil des années, des méthodes ont été élaborées pour répondre au besoin de statistiques plus détaillées en tenant compte du CDS. À titre d'exemple, les chercheurs peuvent maintenant accéder directement aux microdonnées (de façon sécuritaire) et ainsi appliquer leurs modèles aux données. Citons les exemples des Scientific Use Files ou SUF (fichiers scientifiques) et des Secure Use Files ou ScUF (fichiers à accès sécurisé) qui ont été mis à la disposition des chercheurs autorisés. Pour éviter la divulgation de renseignements personnels, les SUF ou fichiers d'utilisation scientifique ne contenaient pas toujours toutes les variables de l'enquête, et certaines variables seulement dans un format moins détaillé. Le contenu des ScUF ou fichiers à accès sécurisé est habituellement plus riche que celui des fichiers d'utilisation scientifique, mais on peut y accéder uniquement dans un centre de données de recherche (sur place) ou dans un environnement d'accès à distance sécurisé. Les SUF et les ScUF sont encore utilisés par les chercheurs autorisés.

Pour permettre à d'autres personnes que les chercheurs autorisés d'accéder à des renseignements plus détaillés, des données tabulaires (beaucoup) plus détaillées sont fournies, comme les données du recensement au niveau géographique de mailles(cellules de grille) de 1 km × 1 km. Les données à ce niveau géographique sont par exemple mentionnées dans un règlement d'application du recensement européen de 2021 (Commission européenne, 2018), qui oblige les États membres à fournir des statistiques au niveau de ces cellules de grille à Eurostat.

2. Contrôle de la divulgation statistique et groupes présentant un intérêt particulier

2.1 Définitions

Dans l'article, nous désignons par le terme « groupes présentant un intérêt particulier » soit de petits groupes d'unités (p. ex. ménages sur des cellules de grille de 1 km × 1 km dans des régions rurales) soit des groupes particuliers d'unités (p. ex. patients atteints de la COVID-19 ou réfugiés ukrainiens). Par ailleurs, nous entendons par contrôle de la divulgation statistique (CDS) le domaine de recherche qui traite, d'une part, de l'évaluation du risque de divulgation de renseignements sur des unités identifiables à partir des sorties statistiques et, d'autre part, de l'application de techniques pour réduire ce risque de divulgation. Un synonyme de contrôle de la divulgation statistique est limitation de la divulgation statistique (LDS). Pour un aperçu des mesures des risques et de l'utilité et des méthodes de CDS, voir p. ex. Hundepool et coll. (2012).

2.2 Le concept de divulgation

La nécessité d'évaluer le risque de divulgation des publications statistiques est en partie attribuable à des obligations légales et réglementaires. En Europe, par exemple, le règlement général de protection des données (RGPD, Commission européenne, 2016) concerne la protection des données personnelles, la définition d'une donnée à caractère personnel étant *toute information se rapportant à une personne physique identifiée ou identifiable (« personne concernée »)*. La plupart des États membres de l'UE ont aussi des lois relatives aux statistiques qui comprennent l'obligation pour les instituts nationaux de statistique (INS) de protéger le respect de la vie privée de leurs répondants. À titre d'exemple, les articles 37 à 42 de la loi néerlandaise sur la statistique (une traduction en anglais se trouve à l'adresse https://www.cbs.nl/en-gb/about-us/organisation) régissent la façon dont le Bureau central de la statistique des Pays-Bas est autorisé à traiter et à publier des renseignements statistiques. L'article 37, par exemple, stipule que les données reçues par le Bureau central de la statistique des Pays-Bas *doivent être publiées de manière à ce qu'aucune donnée reconnaissable au sujet d'une personne, d'un ménage, d'une entreprise ou d'une institution, ne puisse en être dérivée, à moins qu'il y ait, dans le cas de données relatives à une entreprise ou à une*

institution, de bonnes raisons de supposer que la société ou l'institution concernée n'aurait pas d'objection à la publication.

Le libellé concernant la façon dont les INS doivent traiter les données qui leur sont fournies et sur lesquelles ils sont autorisés à publier des statistiques ne mentionne pas l'origine ni la taille du groupe d'unités visé par la statistique. Le concept de divulgation statistique pour les groupes présentant un intérêt particulier ne diffère donc pas de celui des statistiques « ordinaires ». De toute évidence, les petits groupes peuvent être plus vulnérables à la divulgation, c'est-à-dire que les petits groupes et les groupes présentant un intérêt particulier sont plus susceptibles d'enfreindre les restrictions du RGPD et des lois nationales sur la statistique que les statistiques ordinaires. Le *concept* reste toutefois exactement identique.

2.3 Intérêt public

Dans le RGPD, l'alinéa 6 (1)(e) prévoit que les responsables du traitement peuvent traiter des données à caractère personnel si *un traitement est nécessaire à l'exécution d'une mission d'intérêt public*. En outre, en ce qui concerne les catégories particulières de données à caractère personnel, l'alinéa 9 (2)(g) prévoit que le traitement est permis s'il est *nécessaire pour des motifs d'intérêt public important, sur la base du droit de l'Union ou du droit d'un État membre qui doit être proportionné à l'objectif poursuivi, respecter l'essence du droit à la protection des données et prévoir des mesures appropriées et spécifiques pour la sauvegarde des droits fondamentaux et des intérêts de la personne concernée* et l'alinéa 9 (2)(i) prévoit que le traitement est permis s'il est *nécessaire pour des motifs d'intérêt public dans le domaine de la santé publique*. Ceci montre que dans des circonstances particulières, au nom de l'intérêt public, le traitement de données à caractère personnel (particulières) peut être autorisé en vertu du RGPD.

Le risque de divulgation est une combinaison de la probabilité qu'une divulgation se produise et de l'incidence possible de la divulgation sur les unités individuelles reliées. Il est presque impossible de publier des renseignements utiles sans risque de divulgation. En effet, la loi fondamentale du contrôle de la divulgation statistique (formulée par Dwork et Naor [2010] comme une impossibilité rigoureuse de l'objectif de Dalenius en matière de respect de la vie privée) énonce essentiellement que toute information statistique *utile* comporte *nécessairement* un *risque* résiduel non nul de divulgation de renseignements sur certaines personnes. Par conséquent, les INS doivent mettre en place une politique sur la quantité de risque résiduel acceptable. Les choix faits pour une telle politique peuvent dépendre de la sensibilité des variables en question, de l'importance des résultats pour le grand public et de nombreux autres facteurs.

Même si cela peut être considéré comme une question d'éthique, dans le cas de groupes présentant un intérêt particulier, l'importance pour le bien collectif peut l'emporter sur l'incidence sur les individus de ces groupes. Par conséquent, il se peut que le *concept* de contrôle de la divulgation statistique pour les groupes types (présentant un intérêt particulier) ne soit pas différent de la situation ordinaire, mais que le *risque résiduel acceptable* choisi le soit. Cela peut notamment se traduire dans la façon dont les paramètres des mesures de risque sont choisis.

2.4 L'utilité et ses aspects

Comme il s'agit d'un phénomène double relativement à la loi fondamentale sur le contrôle de la divulgation statistique mentionnée plus haut, l'application de toute méthode de CDS réduit l'utilité de l'information statistique publiée. Conformément au *concept* de la divulgation, les petits groupes exigent souvent une application rigoureuse des méthodes de CDS, qui a nécessairement de fortes conséquences sur toute publication de données concernant de petits groupes.

L'application de ces méthodes à de petits groupes affecte non seulement l'utilité des statistiques sur ces groupes eux-mêmes, mais parfois aussi l'utilité d'autres statistiques. Cela est particulièrement vrai quand les statistiques sur de petits groupes sont demandées à la suite d'événements inattendus, c'est-à-dire aux fins de publications non planifiées, hors du programme de publication ordinaire. Ces phénomènes spécialisés sont néanmoins souvent liés à d'autres statistiques régulières. L'application des méthodes de CDS pour protéger les statistiques concernant de petits groupes peut limiter sans risque les possibilités de publication de ces mêmes statistiques sur des groupes connexes plus importants. En effet, la publication de statistiques sur les grands groupes doit non seulement être sûre en soi, mais elle ne doit pas non plus entraver la protection de la publication sur les petits groupes. Cette interrelation entre la protection de statistiques sur de petits groupes et d'autres statistiques implique non seulement des statistiques liées en matière de

sujet, mais aussi des statistiques liées en matière de temps. Dans tous les cas, les publications doivent être protégées de façon uniforme ou *conditionnellement à* la protection des statistiques connexes ou publiées antérieurement.
Si l'on considère le contrôle de la divulgation statistique (CDS) comme l'art de publier autant que possible sans divulguer de renseignements individuels, les INS tenteront de réduire au minimum la perte de renseignements étant donné un risque de divulgation résiduel maximal. Comme cela a été dit plus haut, souvent, les groupes présentant un intérêt particulier nécessitent intrinsèquement une application assez sévère des méthodes de CDS, ce qui peut avoir une incidence considérable sur l'utilité de ces statistiques. Cela peut se produire au moyen de méthodes « classiques » de CDS, comme le regroupement de catégories et la suppression de cellules de tableau, mais aussi au moyen de méthodes plus récentes de CDS, comme la perturbation et l'ajout de bruit.

Une complexité supplémentaire découle de l'observation selon laquelle l'utilité n'est pas un terme absolu : les utilisateurs n'ont pas tous des préférences identiques. De plus, il est souvent impossible de définir l'utilité pour une seule publication : les utilisateurs souhaitent souvent comparer les données dans le temps ou entre différentes populations. Les aspects relatifs à l'utilité à long terme pourraient alors limiter les aspects de l'utilité à court terme. Cette observation vaut, par exemple, pour les données de recensement de la population et du logement qui doivent être comparables dans le temps et entre pays (de l'Union européenne).

3. Quelques exemples européens

3.1 Groupes présentant un intérêt particulier

Dans l'Union européenne, les statistiques sur les réfugiés ukrainiens et celles sur les effets de la crise de la COVID-19 sont de récents exemples de statistiques concernant des groupes présentant un intérêt particulier. Ces deux exemples sont présentés dans les sous-sections ci-dessous.

3.1.1 Crise de la COVID-19

La crise de la COVID-19 a bouleversé la vie de nombreux citoyens et citoyennes de l'Union européenne (UE). Cela a été particulièrement vrai au printemps 2020, quand la plupart des gouvernements de l'UE ont imposé des mesures de confinement à la population pour empêcher la propagation du virus. Pendant cette période, la vie professionnelle s'est déroulée en ligne ou a été interrompue momentanément. Afin de contrer l'effet de ces mesures de confinement sur la population active, la plupart des gouvernements de l'UE ont également adopté des mesures d'aide temporaires. Quand la situation s'est de nouveau détériorée à l'automne 2020, la plupart des gouvernements ont décidé de poursuivre les mesures d'aide. Voici des exemples de mesures d'aide gouvernementales :

- travail de courte durée et régimes assimilés, où les employés travaillent un nombre d'heures inférieur et sont (en partie) rémunérés pour les heures non travaillées;
- mises à pied temporaires, où les employés restent rattachés à leur emploi dans le cadre d'un contrat de travail, mais ne travaillent pas temporairement.

Le financement de ces mesures pouvait se faire directement entre l'employé et l'État, ou indirectement quand l'employeur est rémunéré par l'État et qu'il attribue le financement à ses employés.

L'Union européenne demandait des statistiques sur ces mesures, par exemple le nombre d'emplois qui en bénéficiaient. Des renseignements ont été recueillis sur le nombre total d'emplois pour lesquels un soutien avait été demandé, pour lesquels un soutien avait été autorisé et ceux qui avaient réellement bénéficié des mesures. En général, les statistiques devaient être publiées par secteur de la Nomenclature statistique des activités économiques dans la Communauté européenne (NACE). De plus, des statistiques étaient demandées sur le nombre d'unités locales utilisant la prestation et le nombre d'heures non travaillées. Dans la mesure du possible, ces statistiques étaient normalisées par le nombre total d'emplois, le nombre total d'unités locales et le nombre total d'heures travaillées, afin de faciliter les comparaisons entre pays. Les données ont été recueillies mensuellement à partir de janvier 2020, les dates de référence étant le dernier jour de chaque mois.

Voici quelques-uns des facteurs qui compliquent la production de ces statistiques au niveau européen.

- Les données ont été recueillies par des pouvoirs publics nationaux différents. Les données concernant le nombre total d'emplois provenaient de l'INS ou de l'enquête sur les forces de travail de l'UE. Les données concernant les unités locales provenaient soit de sources nationales soit de l'enquête sur la structure des entreprises de l'UE.
- Rien ne garantissait que les données étaient harmonisées. Il y avait d'importantes différences entre les règles déterminant si le financement était direct ou indirect.
- La participation variait selon l'indicateur et dans le temps. Le taux de participation le plus élevé était observé dans le nombre total d'emplois soutenus par des mesures gouvernementales, que le soutien ait été réellement utilisé ou approuvé. Au début de la pandémie (avril-juin 2020), 23 États membres au total y ont participé. La participation la plus faible concernait le nombre d'heures non travaillées, pour lequel seuls sept États membres ont transmis des données à Eurostat.

Cette collecte de données européenne a pris fin en 2021, en raison de l'influence décroissante de la pandémie sur l'économie, mais aussi du fait que l'accès aux données requises était parfois problématique, car les INS n'étaient généralement pas les producteurs et les propriétaires des données.

3.1.2 Réfugiés ukrainiens

Une directive relative à la protection temporaire a été activée en mars 2022 (voir Commission européenne, 2022). Cette directive régit la collecte de données sur les statuts de protection temporaire accordés au cours du mois de référence aux personnes fuyant l'Ukraine ainsi que sur le stock de bénéficiaires des statuts de protection temporaire à la fin du mois de référence. Les données doivent être transmises à Eurostat dans un délai d'un mois à compter de la fin du mois de référence.

Les données du premier trimestre de 2022 ont été transmises à Eurostat fin mai 2022 et publiées début juin 2022. Les données annuelles pour 2022 seront transmises à Eurostat d'ici la fin de mars 2023 et publiées au début d'avril 2023.

Bien que ce cadre juridique ait permis de commencer la collecte de données européennes sur les personnes fuyant l'Ukraine, dans la pratique, il a fallu un certain temps avant que les données des États membres n'arrivent au bureau d'Eurostat. De plus, la rapidité de mise en œuvre a entraîné des problèmes de comparabilité des données entre les différents pays. Cela a néanmoins permis la publication de données pertinentes dans un délai relativement court. Ces données sont moins détaillées que les produits du recensement détaillés dont la publication est moins actuelle. Cela signifie que différents utilisateurs reçoivent des données présentant différents degrés de détail et une actualité différente. Enfin, il est clair que les données de recensement plus détaillées doivent être davantage protégées et que la façon dont ces données plus détaillées sont protégées dépend des données publiées auparavant.

3.2 Petits groupes

Les statistiques sur des groupes de petits domaines dans les recensements européens sont un exemple récent de statistiques concernant de petits groupes. La sous-section suivante présente les statistiques nouvellement introduites au niveau des zones de 1 km × 1 km du recensement européen de 2021.

3.2.1 Groupes de petits domaines dans les recensements européens

L'année 2021 a été une année de recensement européen. Cela signifie que tous les États membres de l'Union européenne (UE) ont dû effectuer un recensement de la population et des logements avec un jour de référence en 2021 (jour du recensement). Il s'agit d'un moyen important d'harmoniser les résultats des recensements européens. De plus, tous les pays de l'UE doivent publier au moins un ensemble de tableaux harmonisés pour faciliter grandement les comparaisons. Cet ensemble de tableaux à haute dimensionnalité donne une description précise des personnes vivant dans l'UE et de leur situation de logement. Cet ensemble est appelé les hypercubes du recensement européen de 2021.

De plus, dans le recensement européen de 2021, pour la première fois, un ensemble de tableaux de mailles de 1 km × 1 km était obligatoire. Ces tableaux ne sont pas détaillés par leur contenu (pour chacun de ces tableaux, une seule caractéristique est incluse), mais ils sont détaillés par leur structure (dans tous les pays, le nombre de mailles est

beaucoup plus élevé que le nombre de municipalités). En outre, les mailles et les distributions régionales sont des variables non emboîtées. Cela implique que les pays doivent vérifier si les informations sur les individus peuvent être divulguées par croisement de ces mailles avec les municipalités (unités d'administration locale), le niveau de région le plus détaillé dans les hypercubes européens. Les autres niveaux de région (pays, NUTS1, NUTS2 et NUTS3) dans les hypercubes sont des combinaisons d'unités d'administration locales. Ces niveaux régionaux sont emboîtés, c'est-à-dire qu'ils présentent une structure hiérarchique. Cependant, les mailles sont une variable régionale non emboîtée dans cette structure hiérarchique.

Les méthodes classiques non perturbatrices telles que le recodage global et la suppression de cellules ne constituent pas, pour différentes raisons, une solution de protection des tableaux de recensement européens. Pour permettre les comparaisons entre pays, les formats des tableaux sont fixes et non modifiables. C'est pourquoi le recodage global n'est pas possible. Il est pratiquement impossible d'appliquer de façon optimale la suppression de cellules à un aussi grand ensemble de tableaux couplés à haute dimensionnalité. En théorie, il serait possible d'appliquer la suppression de cellules avec une sursuppression pour sécuriser l'ensemble de tableaux. Cela entraînerait toutefois une perte d'information considérable, qui serait inacceptable du point de vue de l'utilisateur. Ajoutons que la gestion du risque de différenciation entre les données d'hypercube et les données au niveau de la grille serait aussi un problème qui compliquerait davantage encore les concepts de protection fondés sur la suppression de cellules.

À partir de l'expérience de nombreux pays à l'intérieur et à l'extérieur de l'UE, deux méthodes ont été recommandées : l'échange ciblé d'enregistrements (TRS pour *Targeted Record Swapping*) pré-tabulaire et la méthode de la clé de cellule post-tabulaire (CKM pour *Cell Key Method*). Essentiellement, les deux méthodes consistent à ajouter du bruit aux cellules des tableaux. Bien qu'il soit recommandé d'employer les deux méthodes pour protéger les hypercubes du recensement de façon harmonisée, les différents États membres de l'UE ont une certaine liberté pour décider de leur mode d'utilisation de ces méthodes. Ils peuvent non seulement choisir des valeurs de paramètres différentes, mais aussi décider d'utiliser une seule des méthodes ou une combinaison des deux méthodes. En effet, compte tenu des différentes règles de confidentialité en vigueur dans les différents pays européens ainsi que de la différence de taille de ces pays, il a été conseillé de ne pas recommander une seule méthode. Néanmoins, le fait de limiter le nombre de méthodes recommandées facilite la comparaison des statistiques de recensement protégées entre pays, qu'elle soit réalisée par Eurostat ou d'autres utilisateurs. Puisque les deux méthodes ne conduisent pas à la suppression de données, les données des États membres, si elles sont traitées par ces méthodes, peuvent être combinées en données au niveau européen.

Comme nous l'avons dit, la protection des groupes de petits domaines influe sur la façon dont les autres publications (les hypercubes) doivent être protégées. En outre, la publication des cellules de grille et des hypercubes généraux a lieu à des moments différents : les cellules de grille doivent être transmises à Eurostat en décembre 2022, tandis que les hypercubes du recensement doivent parvenir à Eurostat au plus tard en mars 2024.

De plus, il existe un lien entre les publications du recensement et les statistiques (nationales) régulières sur la population et la démographie. L'application de méthodes récentes de CDS, comme l'échange ciblé d'enregistrements ou la méthode de la clé de cellule (CMK), aux publications du recensement n'est pas facilement transférable à toutes les statistiques démographiques connexes. Encore une fois, il s'agit d'un exemple de la façon dont les méthodes de CDS appliquées aux groupes types ou à de petits groupes d'unités influent sur les publications d'autres statistiques.

4. Résumé, conclusions et perspectives

Dans l'article, nous avons analysé le cas de statistiques portant sur des groupes présentant un intérêt (particulier) ou de petits groupes d'unités. Nous avons fait valoir que le *concept* de contrôle de la divulgation statistique (CDS) n'est pas différent de celui appliqué aux statistiques « ordinaires ». Seule la *norm*e que les instituts nationaux de la statistique (INS) choisissent dans leur politique sur le risque de divulgation résiduelle acceptable peut différer entre statistiques sur les groupes présentant un intérêt (particulier) et statistiques « ordinaires ». Cette différence peut se fonder, par exemple, sur la pertinence et l'importance pour l'intérêt public de l'information spécialisée et actuelle concernant des groupes particuliers de personnes. Nous avons illustré notre propos par des exemples européens de statistiques sur des groupes présentant un intérêt particulier (réfugiés ukrainiens) et de petits groupes (tableaux avec mailles de 1 km × 1 km dans le recensement européen de 2021).

L'utilisation du concept (normalisé) de CDS sur de petits groupes d'unités individuelles montre qu'intrinsèquement, ces statistiques exigent souvent une application rigoureuse des méthodes de CDS pour rendre possibles des publications au risque de divulgation limité. Les travaux récents vont dans le sens de l'utilisation de méthodes perturbatrices, comme l'ajout de bruit, plutôt que des méthodes classiques de CDS, comme le regroupement de catégories et la suppression de cellules de tableau. Dans le cadre des projets du recensement européen de 2021, l'échange ciblé d'enregistrements (TRS) et l'ajout de bruit au moyen de la méthode de la clé de cellule (CKM) sont proposés dans le but d'harmoniser davantage les techniques de CDS employées par les États membres de l'Union européenne.

Selon nous, l'application des méthodes de CDS aux publications sur des groupes présentant un intérêt particulier et sur de petits groupes d'unités individuelles a une incidence sur la protection d'autres publications qui leur sont liées. À titre d'exemple, les hypercubes du recensement et les tableaux de cellules de grille peuvent encore avoir des marginales communes, même si les variables régionales respectives ne sont pas emboîtées. Au minimum, les totaux nationaux se trouvent dans les deux classifications régionales. De plus, la publication des tableaux de cellules de grille doit être faite bien avant la publication des hypercubes. Ajoutons que les INS publient d'autres statistiques connexes concernant la population et les données démographiques.

Dans un avenir proche, les statistiques démographiques et les statistiques de recensement relèveront d'un seul cadre juridique européen et devront donc être protégées de manière uniforme. Les leçons tirées des expériences actuelles en matière de protection des groupes présentant un intérêt particulier et des groupes de petits domaines dans le cadre du recensement et les répercussions sur d'autres statistiques connexes pourraient et devraient être intégrées à cette méthode de protection plus unifiée. Dans une perspective encore plus lointaine, le succès de cette approche unifiée pourrait donner lieu à une tentative de définition de méthodes unifiées dans d'autres domaines de la statistique européenne.

Malgré la façon dont le CDS est appliqué actuellement à des groupes présentant un intérêt particulier et à de petits groupes d'unités individuelles, il reste des questions à régler.

Parallèlement à l'échange ciblé d'enregistrements et la méthode de la clé de cellule, recommandés dans le recensement européen, la notion de protection différentielle de la vie privée est utilisée dans le recensement des États-Unis. Essentiellement, cette notion consiste aussi à ajouter du bruit aux tableaux du recensement. Il faudrait comparer les deux méthodes et leurs avantages et inconvénients éventuels.

L'ajout de bruit pour protéger les statistiques contre la divulgation de renseignements personnels dans les publications générales est relativement nouveau, non seulement pour les organismes de statistique, mais aussi pour les utilisateurs des données publiées et les fournisseurs de données individuelles (répondants). Cela contraint les INS à informer clairement et correctement les utilisateurs et les répondants. Il n'est pas facile d'expliquer de quelle façon l'ajout de bruit protège vraiment contre la divulgation de renseignements personnels ni en quoi les statistiques perturbées restent utiles. Souvent, les utilisateurs considèrent que les « statistiques officielles » sont des chiffres officiels et « donc » de toute évidence des chiffres exacts, alors qu'en réalité les « statistiques officielles » sont des estimations, qui comportent une incertitude par définition. En ce qui concerne la protection de l'aspect de communication, il faut noter que les chiffres perturbés peuvent donner lieu à des opinions différentes. Un répondant pourrait conclure que l'organisme de statistique ne présente pas sa réponse correctement, surtout si un répondant ou une répondante affirme qu'il ou elle s'identifie dans la publication. Il faudrait alors pouvoir indiquer clairement qu'en général, quiconque voit cette information ne peut pas être certain de la divulgation individuelle en raison du bruit (éventuellement) ajouté : seul le répondant lui-même peut être certain de la bonne information.

Nous devons aussi aborder un deuxième aspect concernant l'utilisation de méthodes perturbatrices et l'ajout de bruit : le mode de traitement des données par les chercheurs qui ont accès aux microdonnées à des fins de recherche statistique ou scientifique dans les locaux d'un institut national de statistique ou par une connexion à distance sécurisée. Doivent-ils également ajouter du bruit à leurs résultats, de façon cohérente avec les publications officiellement publiées? Si des méthodes prétabulaires sont utilisées, le chercheur doit -il utiliser les microdonnées perturbées? Dans tous les cas, leurs publications doivent être protégées par elles-mêmes et par rapport aux statistiques déjà publiées.

Bien que les méthodes perturbatrices comme l'ajout de bruit semblent de plus en plus courantes en cas de groupes présentant un intérêt particulier et de petits groupes d'unités individuelles, il ne doit pas y avoir de place pour la

complaisance. La popularité croissante de la recherche sur les microdonnées et le besoin de statistiques spécialisées et actuelles sur des groupes particuliers dans des circonstances particulières exigent que les INS et Eurostat vérifient toujours et mettent à jour continuellement leurs mesures de protection de la vie privée. L'objectif restera d'améliorer les possibilités de recherche scientifique et de servir l'intérêt public par des statistiques officielles sûres et utiles.

Bibliographie

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

Commission européenne. (2016), Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (règlement général sur la protection des données), *Journal officiel de l'Union européenne*, L119, p. 1-88.

Commission européenne. (2018), Règlement d'exécution (UE) 2018/1799 de la Commission du 21 novembre 2018 relatif à l'établissement d'une action statistique directe temporaire pour la diffusion de thèmes sélectionnés du recensement de la population et du logement de 2021 géocodés selon une grille de 1 km^2, *Journal officiel de l'Union européenne*, L296, p. 19-27.

Commission européenne. (2022), Décision d'exécution (UE) 2022/382 du Conseil du 4 mars 2022 constatant l'existence d'un afflux massif de personnes déplacées en provenance d'Ukraine, au sens de l'article 5 de la directive 2001/55/CE, et ayant pour effet d'introduire une protection temporaire, *Journal officiel de l'Union européenne*, L71, p. 1.

Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, E. Schulte Nordholt, K. Spicer et P.P. de Wolf. (2012), *Statistical Disclosure Control*, Wiley series dans Survey Methodology, John Wiley & Sons, Ltd, ISBN: 978-1-119-97815-2.