Nations Unies ECE/CES/GE.41/2022/3

Conseil économique et social

Distr. générale
23 juin 2022
Français
Original : anglais

Commission économique pour l'Europe

Conférence des statisticiens européens

Groupe d'experts des recensements de la population et des habitations

Vingt-quatrième réunion
Genève, 21-23 septembre 2022
Point 5 de l'ordre du jour provisoire
Transitions dans les méthodes appliquées pour les recensements : plans, expérience acquise et innovations

Méthodes perturbatrices pour les tableaux du recensement de 2021 dans le cadre européen

Note établie par Statistics Netherlands*

Résumé

Cela fait plusieurs décennies que les recensements constituent une partie essentielle du programme de travail des organismes nationaux de statistique. La perspective européenne est devenue une importante dimension supplémentaire pour la diffusion de tous les différents résultats de recensement. La combinaison des résultats des recensements nationaux nécessite évidemment un certain degré de coordination et d'harmonisation. Dans un premier temps, on a harmonisé la conception des tableaux de présentation des résultats du recensement au niveau européen. Cela facilite clairement le regroupement des tableaux de différents pays. Cependant, étant donné que les États membres n'ont pas utilisé les mêmes méthodes de contrôle de la divulgation d'informations aux fins de la protection de la vie privée de leurs habitants, les avantages d'une conception harmonisée des tableaux n'ont pas été aussi importants que prévu. Il fallait donc aussi une harmonisation des méthodes de contrôle de la divulgation de données statistiques. Deux projets européens ont été lancés en vue de proposer une telle approche harmonisée du contrôle de la divulgation des données dans les tableaux issus du recensement. Il n'est pas obligatoire pour les États membres d'utiliser les méthodes proposées, mais si de nombreux pays les utilisaient, cela améliorerait certainement la comparabilité des tableaux de recensement européens. Le présent document est en partie basé sur les résultats des deux projets européens qui visaient à élaborer et à appliquer une approche harmonisée. On y trouvera également mises en exergue certaines questions restant à prendre en compte lorsque les méthodes proposées seront utilisées pour le recensement européen de 2021.

* Document établi par Eric Schulte Nordholt et Peter-Paul de Wolf.
Note : Les appellations employées dans le présent document ne reflètent aucune prise de position du Secrétariat de l'Organisation des Nations Unies quant au statut juridique de pays, territoires, villes ou zones quelconques, ou de leurs autorités, ni quant au tracé de leurs frontières ou limites.

I. Introduction

1. L'année 2021 était une année de recensement pour l'Europe. En effet, tous les États membres de l'Union européenne (UE) devaient effectuer un recensement de la population et des habitations avec un jour de référence en 2021 (jour du recensement). C'est là un moyen important d'harmoniser les résultats des recensements européens. En outre, tous les pays de l'UE publieront un ensemble de tableaux harmonisés afin de rendre les comparaisons possibles. Cette série de tableaux liés, à multiples dimensions, qui donne une description précise des personnes vivant dans l'UE et de leur situation en matière de logement, s'appelle les « hypercubes » du recensement européen de 2021. En outre, pour la première fois, un ensemble de tableaux de données carroyées est obligatoire pour le recensement européen de 2021.

2. L'expérience du recensement de 2011 en Europe a montré qu'il fallait une harmonisation plus poussée pour rendre les données plus comparables au niveau international. Les différents pays européens ont appliqué des méthodes sensiblement différentes pour protéger leurs tableaux du recensement 2011, ce qui a beaucoup nui à la comparabilité des résultats entre les pays. S'il est juridiquement impossible de prescrire exactement comment les hypercubes du recensement européen de 2021 doivent être protégés, il est toutefois possible de faire un grand pas en avant en échangeant des informations sur les meilleures pratiques et en recommandant des méthodes de protection pour ces tableaux. De telles recommandations sont par conséquent présentées ici. Elles reposent sur les expériences de nombreux pays membres de l'UE ou extérieurs à celle-ci. D'importants progrès ont été réalisés dans le cadre de deux projets européens, dont les principaux résultats sont décrits dans le présent document.

3. La section II esquisse une perspective historique. Les méthodes de protection proposées sont décrites dans la section III, qui traite également de la combinaison de ces méthodes. La section IV décrit comment les méthodes ont été évaluées jusqu'à présent aux Pays-Bas. La section V est la conclusion, qui comprend quelques remarques sur les problèmes qui peuvent être rencontrés lors de l'utilisation concrète des méthodes proposées.

II. Perspective historique

4. Les recensements sont d'importantes opérations de collecte de données. Tous les États membres de l'Union européenne effectuent des recensements de la population et des habitations afin de fournir des données complètes sur leur population. Les États membres de l'Union européenne doivent fournir des données de recensement à Eurostat, l'office statistique de l'Union européenne. Eurostat compile les données de recensement au niveau européen sur la base des données fournies par les États membres. Dans la plupart des États membres, les données de recensement ne peuvent être publiées que si des mesures ont été prises pour empêcher la divulgation d'informations sur les répondants individuels. Par conséquent, le contrôle de la divulgation de données statistiques est une étape importante avant la publication des données de recensement. Dans Hundepool et al. (2012), on peut trouver plus d'informations sur le contrôle de la divulgation de données statistiques en général. Le présent document concerne la problématique particulière de la protection des données figurant dans des tableaux de recensement détaillés et liés.

5. Pour le recensement européen de 2001, le fondement juridique faisait défaut. Les États membres s'étaient seulement engagés informellement à faire de leur mieux pour fournir des tableaux de recensement à Eurostat. Il ne s'agissait manifestement pas d'une base suffisamment solide pour que tous les tableaux dont on avait besoin soient produits par tous les États membres.

6. La situation s'était améliorée pour le recensement européen de 2011, grâce à l'introduction de la loi sur le recensement européen (Commission européenne, 2008). La livraison de microdonnées de recensement à Eurostat s'étant heurtée à des obstacles juridiques, le concept d'hypercubes de recensement a été introduit à cette époque. Les hypercubes sont des tableaux à multiples dimensions, dont on peut dériver de nombreux

tableaux plus simples à des fins de publication. Dans le même temps, le format dans lequel les tableaux devaient être fournis était passé d'Excel à SDMX.

7. Il est possible d'appliquer aux tableaux de recensement des méthodes classiques et non perturbatrices de contrôle de la divulgation de données statistiques. Ces méthodes comprennent la refonte des tableaux, le recodage global et les suppressions locales. La présentation des tableaux du recensement européen a été modifiée afin de faciliter les comparaisons entre les données des différents pays. Dans le recensement de 2011, la modification a été introduite avec le passage au nouveau format SDMX obligatoire. Toutefois, ces formats de tableau fixes signifiaient également qu'il n'était plus possible de recourir à la refonte des tableaux ni au recodage global pour protéger les informations individuelles contre la divulgation, alors que ces deux méthodes avaient été largement appliquées par plusieurs pays lors du recensement de 2001.

8. L'application de suppressions locales pour protéger un tableau à multiples dimensions peut sembler une solution réalisable. En effet, il est aussi souvent possible de protéger simultanément quelques tableaux liés. Cependant, l'ensemble des hypercubes du recensement de 2011 était beaucoup trop grand et complexe pour être protégé de manière optimale par des suppressions locales. Par optimale, nous entendons l'obtention d'un ensemble de suppressions tel qu'aucune des cellules primaires non sûres ne peut (même approximativement) être recalculée, tandis que les hypercubes conservent une quantité suffisante d'informations pour rester utiles aux utilisateurs.

9. Le problème de la protection adéquate d'un ensemble d'hypercubes de recensement a été reconnu et une équipe spéciale du contrôle de la divulgation des statistiques de recensement a été créée en 2008. Le travail de cette équipe s'est avéré compliqué car il n'existait pas encore de véritables hypercubes et les pays n'étaient pas juridiquement autorisés à partager leurs anciennes microdonnées de recensement. En outre, comme il incombe aux pays de protéger leurs propres tableaux de recensement, il n'est pas possible d'imposer une même manière de protéger les hypercubes de recensement. Par conséquent, chaque pays a protégé ses hypercubes du recensement de 2011 à sa manière. Une complication supplémentaire était le fait que les représentants des différents pays avaient des idées divergentes sur la question de savoir quelles informations étaient sensibles et devaient être protégées. Ce résultat était plutôt décevant.

10. L'harmonisation voulue par l'équipe spéciale aurait dû conduire à des résultats plus comparables entre les pays, mais dans la pratique, les hypercubes mis à disposition dans le Census Hub (voir https://ec.europa.eu/CensusHub2/) n'étaient pas toujours comparables d'un pays à l'autre en raison de la grande variété des méthodes de protection employées. Certains pays n'avaient pas du tout protégé leurs tableaux, de nombreux pays avaient introduit des valeurs manquantes dans les cellules sensibles et beaucoup d'autres cellules pour protéger les cellules sensibles et d'autres pays avaient ajouté du bruit de différentes manières pour protéger leurs hypercubes.

11. Deux variables sensibles comportant de nombreuses catégories ont été fortement sacrifiées dans les hypercubes du recensement de 2011 des Pays-Bas. Pour les variables « pays de naissance » et « pays de citoyenneté », seuls des agrégats ont été publiés et les informations sur les différents pays ont été supprimées lorsque les données ont été passées au format SDMX et publiées dans le Census Hub. Il est clair que cela a permis de protéger les informations sensibles, mais la perte d'informations a été énorme et a conduit à de nombreuses demandes de tableaux supplémentaires au cours des années suivantes. Ces demandes étaient normalement acceptées si aucun tableau à multiples dimensions n'était demandé, car ces tableaux conduiraient toujours à la divulgation d'informations individuelles. Statistics Netherlands ne souhaitait bien évidemment pas répéter cette approche pour les tableaux du recensement de 2021. De même, de nombreux autres pays n'étaient pas non plus satisfaits de l'approche qu'ils avaient suivie pour le contrôle de la divulgation de données statistiques dans les tableaux du recensement de 2011.

12. Il était clair que le processus de production des résultats du recensement européen devait être encore amélioré. Le Règlement européen concernant les recensements de la population et du logement (Commission européenne, 2008) est la base juridique des recensements européens de 2011 et de 2021. Pour le recensement de 2021, quatre règlements

d'application ont été adoptés pour préciser ce que les États membres doivent fournir (Commission européenne, 2017a, 2017b, 2017c et 2018). En outre, afin d'essayer d'éviter d'aboutir à une situation indésirable similaire concernant la protection des hypercubes pour le recensement européen de 2021, deux conventions de subvention ont été conclues sur la question ces dernières années au titre de l'accord de programme-cadre № 11112.2014.005-2014.533. Les résultats des travaux menés au titre de ces deux conventions sont examinés dans les deux sections suivantes.

III. Méthodes proposées

A. Introduction

13. La première convention de subvention mentionnée dans la section II (№ 11112.2016.005-2016.367) concernait un projet qui a débuté en septembre 2016 et a duré un an. Les organismes de statistique de six pays européens (Allemagne, Finlande, France, Hongrie, Pays-Bas et Slovénie) ont participé au projet et Statistics Netherlands en était le chef de file. Dans le cadre de ce projet, intitulé « Harmonised protection of census data in the European Statistical System (ESS) » (Protection harmonisée des données de recensement dans le système statistique européen (SSE)), une enquête a été menée auprès des pays du SSE sur la protection de leurs tableaux de recensement. Ce questionnaire comprenait évidemment des questions sur les aspects juridiques, méthodologiques et techniques. L'objectif du projet était de fournir des recommandations pour la protection des tableaux de recensement de 2021. Celles-ci ne pouvant être formulées correctement que si les situations (juridiques) nationales étaient prises en compte, un certain nombre de questions ont été posées sur ces situations. En outre, des questions ont été posées sur l'évaluation par les pays de leurs méthodes de protection des hypercubes de recensement de 2011 et sur l'utilisation de mailles (carreaux) dans les hypercubes de recensement (nationaux). En définitive, 33 pays européens (27 des 28 États membres de l'époque et 6 des 7 pays en voie d'adhésion) ont répondu.

14. Il convient de noter que les recommandations issues de ce projet n'impliquent pas d'obligations légales pour les pays du SSE. L'objectif du projet était de fournir des conseils pour obtenir des tableaux de recensement bien protégés et faciles à comparer entre les pays.

B. Conclusions tirées du projet sur l'harmonisation de la protection des données de recensement

15. Les lois nationales qui s'appliquent à la publication des résultats des recensements sont souvent vagues quant à ce qu'il faut protéger et à la manière de le faire. Les hypercubes du recensement de 2021 sont un ensemble de tableaux à nombreuses dimensions liés entre eux. Cela implique que de nombreuses cellules des tableaux du recensement auront une valeur très faible ou seront même égales à 0. Il en ressort que les informations individuelles pourraient assez facilement être obtenues à partir des hypercubes du recensement de 2021. Cela montre clairement qu'une absence totale de mesures de contrôle de la divulgation serait illégale pour tous les pays du SSE. De même, bien que la perception de ce qui constitue une information sensible diffère selon les pays, le consensus est que les variables de recensement les plus problématiques semblent être le pays ou lieu de naissance et le pays de citoyenneté. En particulier, le niveau le plus détaillé de ces variables (pays particuliers) pourrait permettre la divulgation des informations individuelles dans les hypercubes dans lesquelles elles apparaissent.

16. Dans l'enquête, de nombreux pays ont mentionné que les méthodes post-tabulation n'étaient pas populaires. Cependant, à notre avis, sans méthodes post-tabulation, il sera pratiquement impossible de protéger correctement les hypercubes de recensement. Ce constat est lié à un fait dont la plupart des pays sont bien conscients : en protégeant les hypercubes de recensement, il faut également se pencher sur les tableaux de recensement nationaux. En effet, même si les hypercubes européens et les tableaux nationaux sont correctement protégés en eux-mêmes, la combinaison des deux sources n'est pas nécessairement sûre. Par conséquent, si par exemple les tableaux nationaux sont publiés en premier, les

hypercubes européens doivent être protégés en tenant compte des informations déjà publiées. Une situation similaire se présente lorsque l'on considère d'autres publications démographiques (nationales).

17. Dans le recensement 2021, de nouveaux types de tableaux ont été ajoutés aux hypercubes européens : des tableaux sur des carreaux de 1 km par 1 km. Ces tableaux ne sont pas détaillés dans leur contenu (pour chacun de ces tableaux, une seule caractéristique est incluse), mais détaillés dans leur structure (le nombre de carreaux est dans tous les pays beaucoup plus grand que le nombre de municipalités). De plus, les carreaux et les distributions régionales sont des variables non imbriquées. Cela implique que les pays doivent vérifier si des informations sur les individus peuvent être obtenues en croisant les données de ces carreaux avec celles des municipalités (LAU), le niveau le plus détaillé de la région dans les hypercubes européens. Les autres niveaux géographiques (pays, NUTS1, NUTS2 et NUTS3) dans les hypercubes sont des combinaisons de LAU. De plus, ces niveaux sont imbriqués, c'est-à-dire qu'ils suivent une structure hiérarchique. Les carreaux sont la seule variable géographique qui n'est pas imbriquée dans cette structure hiérarchique.

18. Les États membres produisent des hypercubes de recensement et les fournissent à Eurostat. Les variables de chaque hypercube et leurs catégories sont harmonisées entre les pays, ce qui permet de combiner les données des États membres au niveau européen. Cependant, les États membres peuvent appliquer les méthodes de leur choix pour le contrôle de la divulgation, et les différences de méthode d'un pays à l'autre pourraient avoir un effet négatif sur la qualité des données au niveau européen. Eurostat vise à harmoniser les méthodes de protection de la confidentialité dans les États membres afin d'améliorer la qualité des données. Plus il y aura d'États membres qui appliquent les méthodes recommandées pour le contrôle de la divulgation, plus les données pourront être harmonisées au niveau européen.

19. Les méthodes classiques non perturbatrices comme le recodage global et la suppression de cellules ne sont pas, pour différentes raisons, une solution pour protéger les tableaux de recensement européens. Pour permettre les comparaisons entre pays, les formats des tableaux sont fixes et ne peuvent être modifiés, ce qui exclut le recodage global. Il est en outre pratiquement impossible d'appliquer de manière optimale la suppression de cellules à un ensemble aussi vaste de tableaux liés à multiples dimensions. Théoriquement, il serait possible d'appliquer la suppression de cellules avec beaucoup de sur-suppressions pour rendre sûr l'ensemble des tableaux, mais cela entraînerait une énorme perte d'informations, ce qui est inacceptable du point de vue de l'utilisateur. Un autre problème est la gestion du risque de divulgation par recoupement des différences entre les données au niveau des hypercubes et des carreaux, ce qui ajoute encore à la complexité des concepts de protection basés sur la suppression de cellules.

20. Une méthode harmonisée devrait offrir une certaine flexibilité afin que les pays puissent facilement l'adapter à leurs propres besoins et attentes concernant un niveau acceptable de risque résiduel de divulgation, d'une part, et un niveau acceptable de perte d'informations, d'autre part. La méthode doit pouvoir être adaptée en changeant les paramètres et doit être constituée de modules distincts pouvant être utilisés en combinaison. L'idée est alors née d'inclure des modules de perturbation non seulement post-tabulation mais aussi pré-tabulation.

21. Ainsi, l'équipe de projet a décidé de choisir la méthode pré-tabulation de la permutation ciblée d'enregistrements et la méthode post-tabulation des clés de cellule dans laquelle du bruit est ajouté aux cellules du tableau. Les deux sous-sections suivantes présentent brièvement les méthodes proposées. Les paramètres des deux méthodes ne sont pas fixes ; les États membres peuvent en décider. Ces deux méthodes ne conduisent pas à la suppression de données, par conséquent les données des États membres, si elles sont traitées par ces méthodes, peuvent être combinées en données de niveau européen.

22. Si de nombreux États membres utilisent la même méthode − même sous des formes différentes − cela aidera à rendre plus simple la préparation des données au niveau européen. Contrairement à la suppression de cellules, avec les méthodes perturbatrices proposées, les données seront disponibles pour toutes les cellules de l'hypercube. Cela constituera un grand

avantage pour tous les utilisateurs et améliorera considérablement la comparabilité des données entre les pays.

23. Afin de maintenir la cohérence entre les données publiées à l'échelon européen et national, les États membres sont encouragés à appliquer la même méthode de contrôle de la divulgation à tous les types de données qu'ils publient. Si, toutefois, une autre méthode est utilisée pour protéger les données nationales, les États membres devraient effectuer une vérification et éventuellement concevoir des variantes qui évitent les risques résiduels de divulgation qui pourraient survenir lorsque les utilisateurs comparent les données de l'hypercube européen aux données nationales.

C. Permutation ciblée d'enregistrements

24. La permutation d'enregistrements est une méthode pré-tabulation de contrôle de la divulgation, c'est-à-dire qu'elle est appliquée aux microdonnées avant la construction des hypercubes de recensement. L'idée générale de la permutation d'enregistrements est que des paires d'enregistrements sont sélectionnées et que les valeurs de certaines variables sont permutées entre enregistrements. La sélection des enregistrements est généralement effectuée de telle sorte que certaines propriétés analytiques particulières soient maintenues et que le biais introduit soit réduit au minimum.

25. La permutation ciblée d'enregistrements telle qu'elle est préconisée pour les hypercubes du recensement 2021 est basée sur une approche mise au point par l'Office for National Statistics (ONS) du Royaume-Uni. Quelques petits ajustements ont dû être apportés, parce que la façon dont l'ONS l'avait appliquée était adaptée à la situation du Royaume-Uni, alors que le projet couvert par la convention de subvention avait pour but d'arriver à une méthode générale applicable à tous les États membres.

26. La permutation ciblée est appliquée au niveau des ménages, en ce sens que seuls des ménages complets feront l'objet de permutations, et non des individus. C'est une façon d'éviter de trop modifier la distribution des caractéristiques des ménages. De plus, seules les variables géographiques seront permutées. De cette façon, les dépendances au sein des ménages ne seront pas trop affectées.

27. De manière générale, la permutation ciblée peut être décrite de la manière suivante (on suppose que les niveaux géographiques cités sont imbriqués) :

a) À chaque niveau géographique, on détermine les ménages dont le risque de divulgation dépasse un certain seuil ;

b) On commence alors au niveau le plus large des zones géographiques disponibles ;

c) On détermine les ménages « similaires » (c'est-à-dire les ménages qui ont les mêmes valeurs pour certaines caractéristiques spécifiées, mais pas nécessairement pour d'autres) dans d'autres régions au même niveau géographique pour obtenir un ensemble de ménages de départ ;

d) On sélectionne au hasard un des ménages de départ ;

e) On permute les variables géographiques du ménage à risque avec celles du ménage de départ sélectionné (c'est-à-dire qu'on permute ces variables géographiques pour tous les membres et enregistrements des ménages) ;

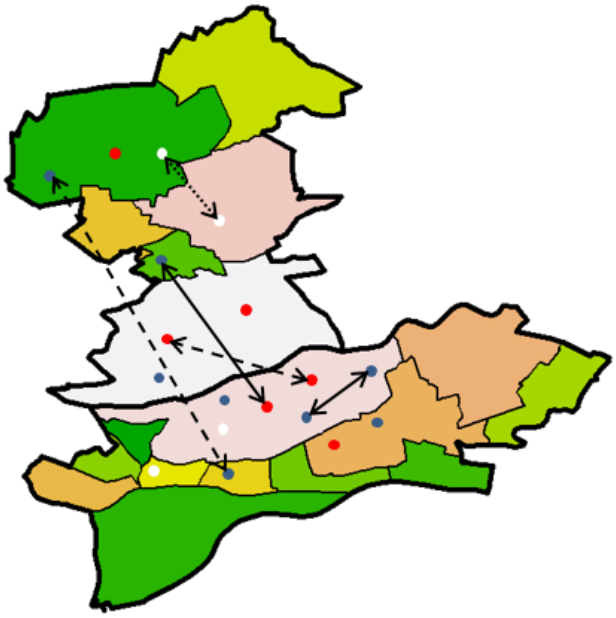
f) On passe ensuite au niveau suivant de détail géographique et on répète l'étape 3 jusqu'à ce que le niveau le plus détaillé soit atteint ;

g) Si le pourcentage d'enregistrements permutés est inférieur à un seuil prédéfini une fois que le niveau géographique le plus détaillé a été atteint, on permute des ménages supplémentaires de manière aléatoire au niveau géographique le plus détaillé jusqu'à ce que soit atteint le taux de permutation souhaité.

28. Une restriction supplémentaire qui est imposée au processus itératif susmentionné est qu'un ménage ne peut être permuté deux fois. En outre, il est clair que ce processus conduit à la permutation de tous les ménages qui sont à risque. Cela implique qu'il sera beaucoup plus difficile d'obtenir des informations individuelles après l'application de la permutation.

29. La figure I donne une représentation graphique de l'application de la permutation ciblée d'enregistrements. La figure présente deux niveaux géographiques, représentés par des frontières épaisses (niveau le plus large) et des zones colorées (niveau détaillé). Les points de couleur représentent les ménages, les ménages similaires ayant la même couleur. Les flèches avec un trait plein montrent les permutations qui ne sont pas autorisées : permutations de ménages non similaires, ou permutations de ménages dans la même zone au même niveau géographique. Les flèches avec des lignes en tirets montrent les permutations possibles au niveau le plus large, les flèches avec des lignes pointillées montrent une permutation possible au niveau le plus détaillé.

Figure I
Illustration de la permutation ciblée d'enregistrements



Note : Les flèches pleines indiquent les permutations qui ne sont pas autorisées, les flèches en tirets les permutations au niveau géographique le plus large, et les flèches en pointillés les permutations au niveau le plus détaillé.

30. Une fois appliquée la permutation ciblée d'enregistrements, les hypercubes de recensement peuvent être calculés. Il en résultera des hypercubes où les nombres de cellules peuvent différer des nombres de cellules d'origine en raison des permutations de ménages et de leurs membres.

D. Ajout de bruit à l'aide d'une méthode de clé de cellule

31. L'ajout de bruit à l'aide d'une méthode de clé de cellule est une méthode post-tabulation, qui s'applique à l'ensemble déjà construit de tableaux de recensement. Elle ne modifie par conséquent pas les microdonnées sous-jacentes et n'affecte que les tableaux de recensement. Cette méthode est basée sur celle introduite par le Bureau australien des statistiques (ABS) (voir par exemple Fraser et Wooton (2006)). La méthode de l'ABS s'appuie sur ce qu'on appelle des « clés de cellule » pour garantir que le bruit aléatoire ajouté à une cellule particulière sera toujours exactement le même, quel que soit l'hypercube de recensement particulier dans lequel elle apparaît. Nous avons légèrement adapté la méthode proposée par l'ABS, dans le sens où nous sommes un peu plus souples dans l'attribution des clés de cellule.

32. Pour garantir la cohérence du bruit ajouté entre différents hypercubes, le processus d'attribution des clés aux cellules doit être cohérent dès le départ. À cette fin, des « clés d'enregistrement » sont attribuées aux enregistrements dans les microdonnées qui sous-tendent tous les hypercubes de recensement. C'est-à-dire qu'un numéro aléatoire est attribué à chaque individu de la population. Chaque fois qu'une cellule d'un hypercube est construite, le nombre d'enregistrements qui relèvent de cette cellule est calculé, et les clés d'enregistrement de ces enregistrements produisent une clé de cellule qui sera utilisée pour sélectionner le bruit à ajouter. De cette façon, le caractère aléatoire des clés d'enregistrement détermine le caractère aléatoire du bruit, tandis que le caractère déterministe du calcul des clés de cellule assurera la cohérence entre les différents hypercubes.

33. D'une manière générale, la méthode des clés de cellule peut être décrite comme suit :

a) On attribue un nombre uniformément distribué [0,1) à chaque enregistrement des microdonnées du recensement ;

b) On établit un tableau de probabilités qui définit la distribution du bruit ;

c) Lors de l'agrégation des microdonnées en un hypercube de recensement, pour chaque cellule, on calcule en outre de manière déterministe la clé de la cellule en utilisant les clés des enregistrements qui relèvent de cette cellule ;

d) On utilise la clé de cellule ainsi que la valeur de la cellule pour déterminer à partir du tableau de probabilités le bruit à ajouter à la cellule ;

e) On ajoute le bruit à la valeur de la cellule.

34. Les clés de cellule sont calculées comme suit : on additionne les clés de tous les enregistrements qui se trouvent dans cette cellule particulière, puis on prend la partie fractionnaire du résultat comme clé de cellule. Ainsi, les clés de cellule sont également des valeurs uniformément distribuées [0,1). Ces valeurs peuvent ensuite être utilisées pour sélectionner parmi les distributions dans le tableau des probabilités. Les clés de cellule sont essentiellement prises comme arguments de la distribution inverse pour obtenir une réalisation du bruit.

35. Les tableaux de probabilité qu'il est préconisé d'utiliser pour les tableaux de comptage des fréquences (c'est-à-dire pour les hypercubes de recensement) ont certains paramètres particuliers qui peuvent être définis :

a) La variance du bruit ajouté, désignée par V ;

b) La valeur maximale du bruit ajouté, désignée par D, à partir de laquelle le bruit peut être distribué $\{-D, -D+1, \dots, -1, 0, 1, \dots, D-1, D\}$;

c) La valeur positive minimale autorisée pour la cellule après l'ajout du bruit, désignée par $j_s + 1$.

36. De plus, la distribution dans un tableau de probabilités doit être telle qu'il est impossible d'avoir des nombres de cellules négatifs alors que l'espérance du bruit ajouté est nulle pour chaque cellule. Par conséquent, les cellules zéro (cellules dont le nombre est égal à zéro) ne peuvent pas être perturbées en nombres positifs. De plus, comme il arrive parfois que les cellules zéro n'aient aucun contributeur, modifier les valeurs de ces cellules ne contribuerait pas à la protection des tableaux.

37. Il convient de noter que j_s pourrait par exemple être utilisé pour empêcher l'apparition des valeurs 1 et 2 dans les tableaux de comptage de fréquences, comme l'exigent certaines législations nationales. Toutefois, ce paramètre permet encore de perturber un nombre de cellules positif jusqu'à la valeur zéro. Si pour j_s la valeur 2 est choisie, alors l'ensemble des valeurs possibles des cellules après application de la méthode des clés de cellule sera $\{0, 3, 4, \dots\}$.

E. Combinaison des deux méthodes

38. Même s'il est recommandé d'utiliser les deux méthodes présentées pour protéger les hypercubes de recensement de manière harmonisée, les États membres de l'UE disposent encore d'une certaine latitude quant à la manière de les appliquer. Ils peuvent non seulement choisir différentes valeurs de paramètres, mais également décider d'utiliser une seule des méthodes ou une combinaison des deux. En effet, compte tenu des différentes règles de confidentialité applicables d'un pays à l'autre ainsi que des différences de taille0 de ces pays, il convenait de ne pas recommander une seule méthode. Cependant, en limitant le nombre de méthodes recommandées, il sera plus facile pour Eurostat ainsi que pour les autres utilisateurs de comparer entre les pays les statistiques de recensement protégées.

39. Un avantage de la combinaison des deux méthodes serait que les paramètres utilisés pour chaque méthode peuvent être fixés de manière moins stricte par rapport à une situation où une seule des méthodes est utilisée. En outre, la méthode des clés de cellule vise spécifiquement à protéger contre le recoupement des différences, alors que la perturbation ciblée d'enregistrements introduit une incertitude en général, mais essentiellement au niveau des enregistrements.

VI. Évaluation des méthodes

40. Au cours du projet financé par la deuxième convention de subvention (№ 2018.0108), des versions préliminaires d'outils permettant d'appliquer les méthodes ont été mises à la disposition du public et les États membres ont été invités à les tester et à fournir des observations. Malheureusement, seul un nombre limité d'États membres a effectivement fourni un retour. Leurs observations concernaient principalement des problèmes d'installation et des questions conceptuelles, par exemple comment choisir les paramètres. D'autres travaux (de recherche) sur le choix de valeurs de paramètres adéquates sont actuellement menés dans plusieurs pays européens.

41. Nous avons connaissance de plusieurs évaluations des méthodes proposées. Nous décrivons ici brièvement celle de Statistics Netherlands (SN).

42. Étant donné que SN réalise son recensement à partir de données administratives, il a pu produire à blanc un ensemble de données de recensement plus récent que le recensement de 2011. L'ensemble de données utilisé pour évaluer les deux méthodes appliquées était basé sur les données de population de 2017. L'intention de SN était d'utiliser une combinaison de permutations ciblées d'enregistrements et de clés de cellules. L'idée était que la principale protection contre le recoupement des différences proviendrait de la méthode des clés de cellule. Toutefois, pour que cette méthode offre une protection suffisante à elle seule, il faudrait probablement adopter des paramètres relativement stricts, ce qui nuirait gravement à l'utilité des données.

43. Les permutations ciblées d'enregistrements et les clés de cellules ont chacune leur propre impact sur l'utilité et le risque de divulgation. En combinant les deux méthodes, la charge sur les paramètres pourrait être répartie sur les deux méthodes. Cela permet de définir pour chaque méthode des paramètres qui ne sont pas trop stricts. En outre, les effets sur l'utilité pourraient également être répartis sur les deux méthodes. Sur la base de ces considérations, SN n'a évalué que l'application d'une combinaison de perturbation ciblée d'enregistrements et de méthode des clés de cellule.

44. Comme la publication d'hypercubes basés sur des mailles était nouvelle, SN a concentré son évaluation sur ce type de tableaux. L'organisme a envisagé non seulement le risque de recoupement des différences entre tableaux, mais aussi entre diverses variables géographiques. Par exemple, il a pris en compte le recoupement de différences entre les mailles et les régions LAU.

45. Plusieurs variantes de la combinaison de la permutation ciblée d'enregistrements et de l'utilisation de clés de cellules ont été envisagées en appliquant différentes valeurs pour les paramètres. Au moment de la rédaction du présent article, plusieurs pays européens avaient décidé d'utiliser la permutation ciblée d'enregistrements, l'utilisation de clés de

cellules ou une combinaison des deux pour protéger leurs hypercubes du recensement 2021, mais aucune conclusion définitive n'avait été tirée de l'évaluation réalisée par SN. Cependant, les recherches menées jusqu'à présent à SN nous ont appris que, pour les tableaux de recensement, un faible taux de permutation dans la méthode de la permutation ciblée d'enregistrements (par exemple, 1 %), où au moins tous les enregistrements à risque sont permutés, contribuera de manière significative à la protection, dans le sens où il est beaucoup plus difficile de juger si une divulgation trouvée est réelle ou non. Des taux de permutation plus élevés entraîneraient de graves pertes d'information. Pour la méthode des clés de cellules, il a été constaté que même avec des variances assez larges (par exemple, 2 ou 3), la perte d'information n'était pas nécessairement si élevée. Ces leçons ont été utilisées pour trouver des valeurs de paramètres appropriées en utilisant une combinaison de la permutation ciblée d'enregistrements et de l'utilisation de clés de cellules.

V. Conclusions

46. L'harmonisation de la présentation des tableaux de recensement est au programme d'Eurostat depuis longtemps. Les progrès récemment obtenus dans le domaine du contrôle de la divulgation des données dans les hypercubes du recensement européen semblent prometteurs. Plusieurs pays ont l'intention d'utiliser la permutation ciblée d'enregistrements ou la méthode des clés de cellule comme stratégie de contrôle de la divulgation pour les hypercubes du recensement de 2021. Les évaluations des méthodes faites par les États membres suggèrent que la permutation ciblée d'enregistrements ne devrait pas être utilisée seule car le risque de divulgation restant est encore trop élevé. La méthode des clés de cellule pourrait être utilisée seule, mais en la combinant avec la permutation ciblée d'enregistrements il semble que l'on parvienne à atténuer considérablement la perte d'utilité qui en découle.

47. Même si nous sommes convaincus que de nombreux États membres de l'UE auront recours à une approche beaucoup plus harmonisée pour le recensement de 2021, il reste encore quelques problèmes à surveiller lors de l'utilisation de méthodes perturbatrices.

48. Tout d'abord, le lien avec d'autres publications de données. Dans de nombreux pays, les hypercubes du recensement européen ne seront pas publiés seuls, mais seront accompagnés de produits standards tels que des tableaux de recensement nationaux et d'autres tableaux démographiques. Ces produits supplémentaires sont évidemment liés aux hypercubes du recensement européen s'ils ont le même jour de référence, comme c'est le cas dans un certain nombre de pays. Lorsque les tableaux nationaux et les autres tableaux démographiques sont protégés par des méthodes différentes (c'est-à-dire qu'ils ne sont pas protégés par l'application de la permutation ciblée d'enregistrements ou la méthode des clés de cellule), les différentes publications pourraient conduire à une situation indésirable dans laquelle des tableaux, qui sont correctement protégés en eux-mêmes, pourraient être combinés pour en extraire des informations individuelles. Une idée simple pour tenter de contourner ce problème serait d'appliquer les mêmes méthodes aux autres publications. Plus précisément, lors de l'application de la permutation ciblée d'enregistrements les mêmes microdonnées perturbées devraient être utilisées (ou, au moins, les mêmes permutations devraient être présentes), et lors de l'application de la méthode des clés de cellule il faudrait utiliser les mêmes clés d'enregistrement. Cela pourrait être possible tant que la publication supplémentaire est purement basée sur les mêmes microdonnées. Toutefois, pour certaines publications supplémentaires, les données du recensement sont combinées avec des données ne relevant pas du recensement. Il serait alors difficile d'appliquer correctement les perturbations originales.

49. Aujourd'hui, il est de plus en plus courant que des chercheurs (accrédités) mènent des analyses sur des ensembles de microdonnées mis à disposition par les organismes nationaux de statistique. Pour eux, il serait difficile de protéger leur production avec la permutation ciblée d'enregistrements et la méthode des clés de cellule, car ces méthodes sont destinées à protéger des tableaux de comptage de fréquences. Or les chercheurs ne produisent pas nécessairement ce type de résultats : ils peuvent envisager des estimations fondées sur des modèles plus complexes ou combiner les données de recensement avec d'autres données pour produire des tableaux de magnitude.

50. Une deuxième question concerne la communication des résultats perturbés. Pour les utilisateurs de données, il doit être clair que les tableaux publiés restent des tableaux « valables ». Pour l'utilisateur général, la publication de tableaux non additifs doit être expliquée. Cela pourrait être fait de la même manière que l'on explique que les tableaux comprenant des chiffres arrondis sont parfois non additifs. Pour les utilisateurs plus expérimentés des données publiées, il convient de quantifier l'incertitude supplémentaire due à la permutation ciblée d'enregistrements et à la méthode des clés de cellule. Toutefois, comme l'indiquent Enderle et al. (2020), la connaissance de la perturbation maximale dans la méthode des clés de cellule pourrait entraîner un risque accru de divulgation.

51. Il convient non seulement d'expliquer l'utilité des tableaux perturbés, mais aussi de faire comprendre au grand public que ceux-ci protègent effectivement sa vie privée. Les organismes nationaux de statistique eux-mêmes ne savent pas encore comment choisir les paramètres pour équilibrer de manière optimale le risque et l'utilité. Cela montre qu'il sera encore plus difficile pour l'utilisateur non spécialisé de saisir l'idée que les tableaux perturbés restent néanmoins utiles mais que, dans le même temps, la publication de ces tableaux ne porte pas atteinte à sa vie privée.

52. Une troisième question concerne le choix des paramètres de la méthode. Comme indiqué à la section IV, certains organismes nationaux de statistique ont récemment tenté d'évaluer cette question. Eurostat a également contribué à cette discussion (Bach 2021). Selon nous, le choix exact dépendra de la situation locale des organismes : taille de la population de l'État membre, différences culturelles, etc. Pour faciliter les évaluations futures, nous appelons l'attention sur deux publications récentes. La première est intitulée « How to select noise parameters for the Cell Key Method? » (Comment sélectionner les paramètres de bruit pour la méthode de la clé de cellule ?) par Giessing et al. (2021). Ce document a précisément pour objet de quantifier le risque de divulgation (restant). Un deuxième article, de Ricciato et al. (2021), traite d'un outil open source permettant d'expérimenter des schémas de perturbation basés sur le bruit, et qui vise à quantifier l'utilité des résultats.

53. Malgré les problèmes mentionnés ci-dessus, nous pensons toujours que l'harmonisation de la protection des hypercubes du recensement européen de 2021 proposée est un grand pas en avant. C'est la meilleure solution disponible lorsqu'il n'est pas possible de prescrire l'utilisation obligatoire de certaines méthodes de protection.

Références

Bach, F. (2021), « Differential Privacy and Noisy Confidentiality Concepts for European Population Statistics », *in* Journal of Survey Statistics and Methodology, 2021 (smab044, https://doi.org/10.1093/jssam/smab044).

Enderle, T., Giessing, S. et Tent, R. (2020), « Calculation of risk probabilities for the cell key method », *in* J. Domingo-Ferrer et Muralidhar, K. (Eds.), *Privacy in Statistical Databases*, pp. 151-165. New York : Springer-Verlag, LNCS, volume 12276.

Commission européenne (2008), Règlement (CE) nº 763/2008 du Parlement européen et du Conseil du 9 juillet 2008 concernant les recensements de la population et du logement. *Journal officiel de l'Union européenne*, L218, p. 14-20.

Commission européenne (2017a), Règlement d'exécution (UE) 2017/543 de la Commission du 22 mars 2017 établissant les règles pour l'application du règlement (CE) nº 763/2008 du Parlement européen et du Conseil concernant les recensements de la population et du logement en ce qui concerne les spécifications techniques des thèmes et de leurs subdivisions Journal officiel de l'Union européenne, L78, p. 13-58.

Commission européenne (2017b), Règlement (UE) 2017/712 de la Commission du 20 avril 2017 établissant l'année de référence et le programme des données et des métadonnées statistiques concernant les recensements de la population et du logement prévu par le règlement (CE) nº 763/2008 du Parlement européen et du Conseil. *Journal officiel de l'Union européenne*, L105, p. 1-11.

Commission européenne (2017c), Règlement d'exécution (UE) 2017/881 de la Commission du 23 mai 2017 portant mise en œuvre du règlement (CE) nº 763/2008 du Parlement européen et du Conseil concernant les recensements de la population et du logement, en ce qui concerne les modalités et la structure des rapports de qualité ainsi que le format technique pour la transmission des données, et modifiant le règlement (UE) nº 1151/2010. *Journal officiel de l'Union européenne*, L135, p. 6-14.

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.

Fraser, B. et J. Wooton (2006), « A proposed method for confidentialising tabular output to protect against differencing », in *Monographs of Official Statistics*. Réunion de travail sur la confidentialité en matière de statistique, Eurostat/Office des publications officielles de l'Union européenne, Luxembourg, pp. 299-302.

Giessing, S., Enderle, T. et Tent, R. (2021), « How to Select Noise Parameters for the Cell Key Method? », présenté à NTTS 2021, 9-11 mars 2021, résumé étendu à l'adresse https://coms.events/NTTS2021/data/x_abstracts/x_abstract_10.docx.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K., de Wolf, P.P. (2012), Contrôle de la divulgation de données statistiques. Wiley series in Survey Methodology, John Wiley & Sons, Ltd, ISBN : 978-1-119-97815-2.

Ricciato, F., Stocchi, M., Bach, F., Bujnowska, A. et Kloek, W. (2021), *An open source tool for experimenting with noise-based perturbation schemes*, présenté à NTTS 2021, 9-11 mars 2021, résumé étendu à l'adresse https://coms.events/NTTS2021/data/x_abstracts/x_abstract_105.pdf.