Recueil du Symposium 2018 de Statistique Canada
Combiner pour vaincre : innovations dans l'utilisation de multiples sources de données

Utilisation de multiples sources de données pour créer et affiner des agrégations géographiques pour la surveillance des sous-comtés

Angela K. Werner et Heather Strosnider[1]

Résumé

La mission du National Environmental Public Health Tracking Program (Programme de suivi) du Center for Disease Control and Prevention est de fournir de l'information à partir d'un réseau national de données intégrées sur la santé et l'environnement, ce qui oriente les mesures visant à améliorer la santé des collectivités. Le programme de suivi prévoit la diffusion régulière de données à une résolution géographique plus élevée afin d'améliorer la surveillance de l'hygiène du milieu et de favoriser les changements à l'échelle locale. Lors de l'affichage des données à plus haute résolution, il faut tenir compte de plusieurs facteurs, comme la stabilité et la suppression de ces données. Pour ce faire, il faut examiner l'agrégation temporelle et spatiale afin de réduire au minimum la suppression et l'instabilité des affichages tout en veillant à ce qu'ils restent classés comme sous-comtés. La méthode de création de ces régions géographiques doit être normalisée afin que les unités géographiques soient comparables d'un État à l'autre, d'une heure à l'autre et d'une base de données à l'autre, et utilisées dans un système national de surveillance.

À l'aide de sources de données multiples, y compris les limites des secteurs de recensement, les données sur la santé et les données sur la population, des agrégations optimales ont été créées pour deux schémas d'agrégation (c.-à-d. un schéma d'agrégation des résultats rares et un schéma d'agrégation des résultats plus commun) pour un ensemble d'États pilotes. Un examen initial des nouvelles agrégations et des consultations avec les États a révélé plusieurs problèmes, comme la fusion entre les comtés, les variations dans les fusions et les unités géographiques avec des populations plus importantes que nécessaire. Une autre méthode de fusion utilisant des centroïdes pondérés en fonction de la population a été explorée après l'établissement de seuils de population appropriés pour les deux systèmes d'agrégation. Les travaux futurs comprennent l'amélioration des régions géographiques agrégées en abordant certains des défis rencontrés et en explorant l'utilisation de facteurs supplémentaires dans le processus d'agrégation.

Mots-clés : Agrégation; santé environnementale; petite région; sous-comté; surveillance; suivi.

1. Programme national de surveillance de la santé publique et de la salubrité de l'environnement

1.1 Introduction

Le Programme national de surveillance de la santé publique et de la salubrité de l'environnement (ci-après, « programme de surveillance ») des Centers for Disease Control and Prevention (Centres pour le contrôle et la prévention des maladies, CDC) a été lancé en 2002 en réponse au rapport de la Pew Environmental Health Commission (commission Pew de la santé environnementale) de janvier 2001 appelant à l'élaboration d'un système coordonné de santé publique pour surveiller et contrer les menaces à la santé environnementale ((McGeehin, Qualters et Niskar, 2004). Le programme de surveillance vise à combler les lacunes statistiques existantes en combinant les données sur

[1] Angela K. Werner, National Center for Environmental Health, Centers for Disease Control and Prevention, Atlanta, Georgie, États-Unis (awerner@cdc.gov); Heather Strosnider, National Center for Environmental Health, Centers for Disease Control and Prevention, Atlanta, Georgie, États-Unis (hstrosnider@cdc.gov)

la santé, le danger et l'exposition afin d'avoir une incidence sur la santé publique. Depuis sa création, le programme de surveillance a permis d'améliorer la communication et la collaboration entre les organismes sanitaires et environnementaux, le milieu universitaire et les organisations non gouvernementales, et a également permis de créer des normes uniformes relatives aux données et des définitions de cas compréhensibles (McGeehin et coll., 2004). À l'heure actuelle, le CDC finance les départements de la Santé de 25 États en plus de celui de la ville de New York (bénéficiaires) et diffuse régulièrement des données par l'intermédiaire du National Environmental Public Health Tracking Network (réseau national de surveillance de la santé publique et de la salubrité de l'environnement, ci-après « réseau de surveillance »). Le réseau de surveillance rend accessibles les renseignements sur la santé, les dangers et l'environnement à un éventail d'intervenants et fournit des cartes dynamiques pour visionner ces renseignements.

1.2 Les efforts du programme de surveillance dans les sous-comtés

À l'heure actuelle, la majorité des données sur la santé affichées dans le réseau de surveillance le sont à l'échelle de l'État ou du comté. Or, le programme de surveillance a déployé des efforts concertés dans le but d'accroître la disponibilité et l'accessibilité des données relatives aux sous-comtés. Le programme de surveillance a effectué un projet pilote de sous-comté en 2014 pour mieux comprendre les défis à relever lorsque l'on travaille avec des données de sous-comtés. Ce projet pilote a donné lieu à un ensemble de recommandations aux fins d'examen pour le programme de surveillance dans la poursuite de ce travail, y compris la nécessité d'un certain type de géographies de sous-comté normalisées (Werner, Strosnider, Kassinger et Shin, 2018). Il a été proposé que cela soit effectué pour deux schémas d'agrégation – un pour les résultats plus courants du programme de surveillance et l'autre pour les résultats plus rares du programme de surveillance.

Les efforts actuels du programme de surveillance quant aux sous-comtés sont composés de plusieurs parties :

- la collaboration avec les bénéficiaires pour élaborer des **lignes directrices sur le géocodage** pour transformer les données sur la santé à l'échelle des adresses à un secteur de recensement;
- l'évaluation des **règles de suppression** actuelles utilisées par les bénéficiaires et les autres coordonnateurs de la gérance des données pour protéger les données des sous-comtés afin d'élaborer de nouvelles règles de suppression pour le programme de surveillance;
- l'évaluation de la manière dont les différents ensembles de données sur les **estimations de la population** influencent les taux;
- la création de géographies de sous-comté normalisées au moyen de l'**agrégation géographique** pour permettre aux données de s'afficher à une échelle plus fine que celle des données à l'échelle des comtés.

Le présent article sera centré sur la création de géographies de sous-comté normalisées et fournira un aperçu des méthodes utilisées, des difficultés auxquelles on est confronté dans les processus d'agrégation géographique et d'amélioration, et les étapes suivantes.

2. Agrégation géographique

2.1 Méthodologie

Plusieurs sources de données ont été utilisées afin de produire les géographies normalisées, y compris les données sur la population en fonction du sexe et de l'âge provenant du recensement décennal de 2010 du Census Bureau des États-Unis (U.S. Census Bureau, 2017a), les fichiers de formes des limites des secteurs de recensement (U.S. Census Bureau, 2017b), les visites en 2010 à l'échelle des comtés au département des urgences pour l'asthme, que les bénéficiaires présentent régulièrement au programme de surveillance, et les données de 2010 sur le cancer du poumon et des bronches du Programme Surveillance Epidemiology and End Results du National Cancer Institute et le National Program of Cancer Registries (programme national des registres du cancer) du CDC. Les visites au département des urgences pour l'asthme ont été utilisées pour le résultat le plus courant et les visites pour le cancer du poumon ont été utilisées pour le résultat plus rare.

En raison des ententes courantes en matière d'utilisation de données, le programme de surveillance ne dispose que des données des bénéficiaires à l'échelle des sous-comtés, donc les nombres de cas prévus à l'échelle des secteurs de recensement ont été calculés pour créer les géographies normalisées. Cela a été effectué en combinant les données des nombres de cas et les données de la population pour calculer les taux en fonction de l'âge et du sexe à l'échelle des comtés. Les nombres de cas prévus étaient ensuite calculés en multipliant le taux des comtés par la population pour chaque groupe d'âge et de genre.

Les fichiers de formes comportant les nombres de cas prévus à l'échelle des secteurs de recensement et les populations correspondantes ont été utilisés en tant qu'entrée pour le Geographic Aggregation Tool (outil d'agrégation géographique, GAT). Le GAT a été créé par la Health Surveillance Section (section de la surveillance sanitaire) du Département de la Santé de l'État de New York pour joindre les régions géographiques avoisinantes en fonction des spécifications des utilisateurs (Talbot et LaSelva, 2010). Différents seuils de population ont été mis à l'essai pour chaque schéma d'agrégation afin de décider de celui qui permettrait de maximiser le nombre d'unités géographiques tout en réduisant au minimum la suppression et l'instabilité. Ultimement, une population totale de 5 000 personnes a été sélectionnée comme seuil pour le schéma d'agrégation du résultat le plus courant, et une population totale de 20 000 personnes a été sélectionnée pour le schéma d'agrégation du résultat plus rare. Un petit groupe de bénéficiaires a mis à l'essai les limites de leurs États au moyen de données réelles sur la santé à l'échelle des secteurs de recensement.

2.2 Défi

De nombreux points de décision sont survenus pendant le processus de création des géographies de sous-comté normalisées et de l'amélioration de ces géographies. Tout d'abord, le principal défi consistait à élaborer une méthode systématique pour créer les géographies normalisées qui seraient utilisées à l'échelle nationale plutôt que de travailler à l'échelle de chaque État. Bien qu'un processus propre à chaque État permettrait de créer la géographie normalisée optimale pour cet État, un processus à l'échelle de chaque État nécessiterait un grand nombre de ressources et pourrait donner lieu à des géographies qui ne sont pas comparables d'un État à l'autre. Deuxièmement, une décision devait être prise pour déterminer les données servant de dénominateurs à utiliser comme fondement des géographies et pour le calcul des taux. Ce projet a utilisé les données du recensement décennal de 2010, mais a soulevé des questions relatives à l'effet des données servant de dénominateurs utilisées, surtout alors que les données s'éloignent des années où un recensement décennal a eu lieu. D'autres questions ont été soulevées sur la manière de traiter les logements de groupes – la question de savoir si ces populations sont habituellement incluses dans le numérateur pour les résultats en santé sur lesquels le programme de surveillance recueille des données, et si les secteurs de recensement comportant un certain pourcentage de logements des groupes examinés devraient être inclus ou retirés. Il y avait également une préoccupation au sujet de la manière de traiter les régions comportant des chiffres de population plus grands qu'il n'était nécessaire. Il y avait également des questions au sujet de la façon d'évaluer les limites et ce qui est jugé « préférable » pour chaque option présentée. Il s'agissait d'un élément qui était plutôt évasif, il était donc difficile pour une personne qui évalue les limites d'établir des paramètres clairs pour définir ce qui est « préférable ». D'autres enjeux et les décisions finales à leur égard sont résumés dans le tableau 2.2-1, ci-après.

Tableau 2.2-1
Enjeux qu'il fallait aborder en cours de création et d'amélioration des géographies de sous-comté normalisées du programme de surveillance et le dernier point de décision

Enjeux	Points dont il faut tenir compte	Décision finale
Schémas d'agrégation	Il est possible que le comté soit trop grand, mais que le secteur de recensement nécessite trop de suppressions. Il pourrait s'avérer préférable pour le programme de surveillance de se doter de deux schémas d'agrégation – un pour les résultats plus courants, l'autre pour les résultats plus rares.	Utiliser des schémas d'agrégation pour les résultats plus courants et pour les résultats plus rares.
Secteurs de recensement comme fondement	On peut choisir parmi plusieurs géographies de sous-comté, y compris le secteur de recensement et le code postal. Chaque option disponible comporte des avantages et des inconvénients. Alors que les codes postaux sont mieux connus du public et que les données moyennes à l'échelle des adresses n'ont pas à être géocodées, ils ont été créés pour le service des postes et donnent lieu à de nombreuses interprétations au moment de leur création. Les délimitations des codes postaux changent fréquemment et ne sont pas relativement homogènes comme les secteurs de recensement.	Utiliser les secteurs de recensement comme fondement des géographies normalisées.
Seuils de population	Il n'est pas évident de déterminer les seuils de population à utiliser pour chaque schéma d'agrégation. Quelle population ou sous-population minimale fournira les taux stables et réduira au minimum la suppression? Les taux d'agrégation ayant fait l'objet de mises à l'essai pour une population totale de 5 000, de 10 000, de 15 000 et de 20 000 personnes et les sous-populations (deux personnes âgées de 65 ans et plus pour chaque personne âgée de zéro à quatre ans) de 1 000, de 2 500 et de 5 000 personnes. La prévalence, les intervalles de confiance, et l'erreur type relative ont été calculés pour déterminer le niveau d'agrégation comportant le meilleur rendement lorsqu'il a été déterminé que le taux acceptable de stabilité et de suppression est de 30 %.	Un total de 5 000 personnes pour le niveau d'agrégation relatif au schéma d'agrégation des résultats les plus courants et un total de 20 000 personnes pour le niveau d'agrégation relatif au schéma d'agrégation des résultats les plus rares
Nombres médians de cas	Il y a incertitude quant à la détermination des résultats en santé qui s'inscrivent dans les deux schémas d'agrégation. Des travaux supplémentaires ont permis d'examiner le nombre médian de cas pour des résultats donnés en santé à l'échelle des secteurs de recensement. Des cas (ou fractions de cas) ont été ajoutés ou soustraits pour améliorer le nombre médian de cas nécessaires à la production de taux stables à suppression minimale.	Les nombres médians de cas (annuel, secteurs de recensement) qui ont été recommandés : • secteur de recensement : ≥ 17,0 cas; • schéma d'agrégation des résultats les plus courants (5 000 personnes) : de 7,3 à 16,9 cas; • schéma d'agrégation des résultats les plus rares (20 000 personnes) : de 1,9 à 7,2 cas.
Géographies hiérarchiques	À l'origine, le niveau d'agrégation de 5 000 ne s'emboîtait pas dans le niveau d'agrégation de 20 000. Il faudrait disposer de géographies hiérarchiques afin qu'elles s'emboîtent les unes dans les autres comme les géographies du recensement.	Utiliser des géographies emboîtées.

Retrait des secteurs de recensement à population nulle	Examiner leur retrait avant de procéder à l'agrégation, car ces secteurs sont habituellement des aéroports ou de grands parcs. Si ces derniers sont inclus dans l'agrégation, ils peuvent accroître artificiellement la taille d'une région et ne pas représenter adéquatement les données sur une carte choroplèthe.	Retirer les secteurs à population nulle avant l'agrégation.
Utilisation d'un centroïde pondéré d'après la population ou d'un centroïde géométrique	Examiner les raisons d'utiliser une méthode plutôt qu'une autre pour l'agrégation. Utiliser la méthode du centroïde pondéré d'après la population est plus judicieux, car on fonde les agrégations sur la population plutôt que le centroïde physique. Les deux options ont fait l'objet d'essais afin d'en examiner les différences.	Utiliser la méthode du centroïde pondéré d'après la population.

3. Prochaines étapes

Les efforts du programme de surveillance sont en cours en ce qui concerne les sous-comtés et font intervenir différentes composantes. Des améliorations supplémentaires seront examinées pour les travaux d'agrégation géographique afin d'accroître l'utilité des géographies de sous-comté normalisées qui seront présentées dans le portail public du programme de surveillance.

La première amélioration consiste à retirer les comtés qui ne satisfont pas au seuil des populations de 5 000 ou de 20 000 personnes avant l'agrégation. Cette pratique devrait empêcher les limites de comtés de se chevaucher (particulièrement dans les régions rurales) et permettra de conserver la structure hiérarchique, de sorte que toutes les unités géographiques s'emboîtent dans les limites des comtés. La deuxième amélioration consiste à examiner les logements de groupe et à déterminer si les secteurs comportant un certain pourcentage classé comme logement de groupe devraient être retirés avant l'agrégation.

La troisième amélioration consiste à aborder la question des unités géographiques comportant des chiffres de population plus grands qu'il ne le faut, et une option éventuelle pour s'y pencher est de retirer les secteurs qui satisfont déjà au seuil des 5 000 ou 20 000 personnes avant l'agrégation. La quatrième amélioration consiste à examiner la restriction dans les limites des villes ou à utiliser une classification urbaine et rurale afin de permettre aux secteurs d'un certain type de s'agréger les uns avec les autres avant de s'agréger avec des secteurs qui sont moins semblables. Enfin, une autre option est de se pencher, à l'avenir, sur la combinaison des seuils de population avec les variables sociodémographiques, bien que cela puisse être effectué au cas par cas, étant donné que cette pratique soulève des questions sur la manière de définir la collectivité de façon systématique et de déterminer les variables sociodémographiques appropriées dont il faut tenir compte dans les agrégations qui seraient applicables dans l'ensemble des résultats en santé du programme de surveillance.

Enfin, tel qu'il a été mentionné, ces géographies de sous-comté normalisées seront présentées dans le portail public du programme de surveillance. Par conséquent, l'agrégation des secteurs de recensement devra s'appliquer d'une façon compréhensible pour les utilisateurs finaux. Plusieurs options pour l'accroissement de la compréhensibilité des agrégations ont fait l'objet de discussions, y compris l'utilisation d'un géolocalisateur afin qu'une personne puisse se situer sur une carte ou visionner des couches opaques, comme les limites d'un comté, afin d'avoir un contexte et de mieux comprendre sa situation dans une agrégation. En fin de compte, l'utilisation de géographies de sous-comté normalisées permettra au programme de surveillance de diffuser des données à résolution plus fine, et ce, tout en réglant les problèmes de stabilité et en préservant la confidentialité.

Bibliographie

McGeehin, M. A., J. R. Qualters, et A. S. Niskar (2004), « National Environmental Public Health Tracking Program: Bridging the Information Gap », *Environmental Health Perspectives,* 112(14), p. 1409-1413.

Talbot, T. O.,et G. D. LaSelva (2010), *Geographic aggregation tool, version 1.31*.

U.S. Census Bureau (2017a), *American FactFinder*.

U.S. Census Bureau (2017b), *Cartographic boundary shapefiles - census tracts*.

Werner, A. K., H. Strosnider, C. Kassinger, et M. Shin (2018), « Lessons Learned From the Environmental Public Health Tracking Sub-County Data Pilot Project », *Journal of Public Health Management and Practice,* 24(5), p. E20-E27.