Proceedings of Statistics Canada Symposium 2018
Combine to Conquer: Innovations in the Use of Multiple Sources of Data

Using multiple sources of data to create and refine geographic aggregations for sub-county surveillance

Angela K. Werner, and Heather Strosnider[1]

Abstract

The Center for Disease Control and Prevention's National Environmental Public Health Tracking Program's (Tracking Program) mission is to provide information from a nationwide network of integrated health and environmental data, driving actions to improve the health of communities. The Tracking Program plans on regularly disseminating data at a higher geographic resolution to improve environmental health surveillance and help drive more local-level changes. When displaying higher resolution data, several considerations, such as stability and suppression of those data, must be taken into account. This requires consideration of both temporal and spatial aggregation to minimize suppression and instability in displays while remaining classified as sub-county. The method to create these geographies must be standardized so the geographic units will be comparable across states, time, and datasets for use in a national surveillance system.

Using multiple sources of data, including census tract boundaries, health data, and population data, optimal aggregations were created for two aggregation schemes (i.e., a rare outcome aggregation scheme and a more common outcome aggregation scheme) for a set of pilot states. An initial review of the new aggregations and consultations with states revealed several issues such as cross-county merging, variations in merges, and geographic units with larger populations than needed. After establishing suitable population thresholds for the two aggregation schemes, an alternative method of merging using population-weighted centroids was explored. Future work includes further refinement of the aggregated geographies by addressing some of the challenges that were encountered and exploring the use of additional factors in the aggregation process.

Keywords: Aggregation; Environmental health; Small area; Sub-county; Surveillance; Tracking.

1. The Environmental Public Health Tracking Program

1.1 Introduction

The Centers for Disease Control and Prevention's (CDC) Environmental Public Health Tracking Program (Tracking Program) was started in 2002 in response to the January 2001 Pew Environmental Health Commission report calling for the development of a coordinated public health system to track and combat environmental health threats (McGeehin, Qualters, & Niskar, 2004). The Tracking Program works to bridge existing data gaps by combining health, hazard, and exposure data to impact public health. Since its inception, the Tracking Program improved communication and collaboration between health and environmental agencies, academia, and non-governmental organizations and also created uniform data standards and understandable case definitions (McGeehin et al., 2004). At present, the CDC funds health departments in 25 states plus New York City (recipients) and regularly disseminates data via the National Environmental Public Health Tracking Network (Tracking Network). The Tracking Network makes health, hazard, and environmental information readily available to a range of stakeholders and provides dynamic maps to view this information.

1.2 The Tracking Program's sub-county efforts

Currently, most health data that are displayed on the Tracking Network are at the state or county level. However, the Tracking Program has made a concerted effort to increase the availability and accessibility of sub-county data. The

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

Tracking Program conducted a sub-county pilot project in 2014 to better understand the challenges that need to be addressed when working with sub-county data. This pilot project resulted in a set of recommendations for the Tracking Program to consider when moving forward with this work, including the need for some type of standardized sub-county geographies (Werner, Strosnider, Kassinger, & Shin, 2018). It was suggested this be done for two aggregation schemes—one for the Tracking Program's more common outcomes and one for the Tracking Program's rarer outcomes.

There are several parts to the Tracking Program's current sub-county efforts, including:

- Collaboration with recipients to develop **geocoding guidelines** for transforming address-level health data to census tract.
- Evaluation of current **suppression rules** used by recipients and other data stewards for protecting sub-county data to develop new suppression rules for the Tracking Program.
- Evaluation of how different **population estimates** datasets impact rates.
- Creation of standardized sub-county geographies through the use of **geographic aggregation** to allow for data to be displayed at a finer scale compared to county-level data.

This paper will focus on the creation of the standardized sub-county geographies and will provide a brief overview of the methods used, challenges faced during the geographic aggregation process and refinement, and next steps.

2. Geographic aggregation

2.1 Methods

Several sources of data were used to produce the standardized geographies, including population data by sex and age from the U.S. Census Bureau 2010 decennial census (U.S. Census Bureau, 2017a), census tract shapefiles (U.S. Census Bureau, 2017b), 2010 county-level asthma emergency department (ED) visits, which recipients routinely submit to the Tracking Program, and 2010 lung and bronchus cancer data from the National Cancer Institute's Surveillance, Epidemiology, and End Results Program and CDC's National Program of Cancer Registries. Asthma ED visits was used for the more common outcome and lung cancer was used for the rarer outcome.

Due to current data use agreements, the Tracking Program only has county-level data from recipients, so expected census tract-level case counts were calculated to create the standardized geographies. This was done by combining case count data and population data to calculate county-level age- and sex-specific rates. Expected case counts were then calculated by multiplying the county rate by the population for each age and gender group.

Shapefiles containing expected census tract-level case counts and corresponding populations were used as the input for the Geographic Aggregation Tool (GAT). The GAT was created by the New York State Department of Health's Environmental Health Surveillance Section to join neighboring geographic areas based on user specifications (Talbot & LaSelva, 2010). Several population thresholds were tested for each aggregation scheme to decide which would maximize the number of geographic units while minimizing suppression and instability. Ultimately, a total population of 5,000 persons was selected as the threshold for the more common outcome aggregation scheme and a total population of 20,000 persons was selected for the rarer outcome aggregation scheme. A small group of recipients tested the boundaries for their states using real census tract-level health data.

2.2 Challenges

Numerous decision points arose during the process of creating the standardized sub-county geographies and refining those geographies. First and foremost, the primary challenge was developing a systematic method to create standardized geographies for use at the national level rather than working on a state-by-state basis. While a state-specific process would create the optimal standardized geography for that state, a state-by-state process would be resource intensive and potentially lead to geographies that are not comparable across states. Second, a decision had to be made as to which denominator data to use as the basis of the geographies and for calculating rates. This project

used 2010 decennial census data, but it raised questions of the impact of denominator data source used, particularly as the data move away from decennial census years. Other questions arose on how to handle group quarters—whether these populations are typically included in the numerator for the health outcomes the Tracking Program collects data on and if census tracts with a certain percentage considered group quarters should be included or removed. There was also a concern about how to address areas with larger than necessary population counts. There were also questions about how to evaluate the boundaries and what is considered the 'best' with each option presented. This was somewhat flexible and it was challenging to set clear parameters to define this for someone evaluating the boundaries. Additional issues and their final decisions made are summarized in Table 2.2-1 below.

Table 2.2-1
Issues that needed to be addressed during the creation and refinement of the Tracking Program's standardized sub-county geographies and the final decision point.

Issue	Points to consider	Final decision
Aggregation schemes	County may be too big but census tract may require too much suppression. May work best for the Tracking Program to have two aggregation schemes—one for more common outcomes and one for rarer outcomes.	Use more common and rarer outcome aggregation schemes.
Census tracts as the foundation	Several sub-county geographies to choose from including census tract and ZIP code. Pros and cons with each available option. While ZIP codes are more familiar to the public and mean address-level data do not need to be geocoded, they were created for the postal service and are open to numerous interpretations when they are created. ZIP code boundaries change frequently and are not relatively homogenous like census tracts.	Use census tracts as the foundation for the standardized geographies.
Population thresholds	Unclear on which population thresholds to use for each aggregation scheme. Which minimum population or sub-population will provide stable rates and minimize suppression. Tested aggregation levels for total population of 5,000, 10,000, 15,000, and 20,000 persons and sub-population (2:1 split for 65+ and 0-4 year olds) of 1,000, 2,500, and 5,000 persons. Calculated prevalence, confidence intervals, and relative standard error to determine which aggregation level performed best using 30% as acceptable for stability and suppression.	Total person 5,000 aggregation level for the more common outcome aggregation scheme and total person 20,000 aggregation level for the rare outcome aggregation scheme.
Median case count ranges	Unclear on which health outcomes fit into the two aggregation schemes. Additional work examined the median case count for given health outcomes at the census tract level. Cases (or fractions of cases) added or subtracted to refine the median case count range needed to produce stable rates with minimal suppression.	Median case count (annual, census tract) ranges recommended for three geographies: • Census tract: ≥ 17.0 cases • Common outcome aggregation scheme (5k persons): 7.3 to 16.9 cases • Rare outcome aggregation scheme (20k persons): 1.9 to 7.2 cases
Hierarchical geographies	Originally, the 5k aggregation level was not nesting within the 20k aggregation level. Should have hierarchical geographies so they nest within one another like census geographies.	Use nested geographies.
Remove zero population census tracts	Consider removing prior to aggregation as these areas are typically airports or large parks. If these are included in the aggregation, they can artificially increase the size of an area, misrepresenting data on a choropleth map.	Remove zero population tracts prior to aggregation.
Use of population-weighted vs. geometric centroid	Consider the rationale for using one method versus another for aggregating. The population-weighted centroid approach makes more sense to use due to basing aggregations on population rather than the physical centroid. Tested both options to examine differences.	Use the population-weighted centroid approach.

3. Next steps

The Tracking Program's sub-county efforts are ongoing and involve multiple components. Additional refinements will be explored for the geographic aggregation work to increase the utility of the standardized sub-county geographies that will be presented on the Tracking Program's public portal.

The first refinement is to remove any counties that do not meet either the 5,000 or 20,000 persons population threshold prior to aggregation. This should prevent county boundaries from crossing, particularly in more rural areas, and will keep the hierarchical structure so all of the geographic units nest within county boundaries. The second refinement is to consider group quarters and whether tracts with a certain percentage classified as group quarters should or should not be removed prior to aggregation.

The third refinement is addressing the issue of geographic units having higher population counts than necessary, and a possible option to explore this is to remove tracts that already meet the 5,000 or 20,000 persons threshold prior to aggregation. The fourth refinement is to explore restricting within city boundaries and/or using urban and rural classification to allow tracts of a certain type to aggregate with one another before aggregating with tracts that may be more dissimilar. Finally, another option to explore in the future is combining the population thresholds with sociodemographic variables, although this may be done more on a case-by-case basis as this raises questions of how to define community in a systematic way and which sociodemographic variables would be appropriate to factor into the aggregations that would be applicable across all of the Tracking Program's health outcomes.

Finally, as mentioned, these standardized sub-county geographies will be presented on the Tracking Program's public portal. This means that an aggregation of census tracts will have to translate in a way that is understandable to end users. Several options to increase the understandability of the aggregations were discussed, including the use of a geolocator so a person can place themselves on the map or having opaque layers, such as county boundaries, so a person can have context and better understand where they are located within an aggregation. Ultimately, the use of the standardized sub-county geographies will allow the Tracking Program to disseminate finer resolution data while addressing stability issues and maintaining confidentiality.

References

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

Talbot, T. O., and 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, and 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), pp. E20-E27.