NEW YORK STATE OF OPPORTUNITY. Department of Health

Geographic Aggregation Tool (GAT): A method for handling small numbers when calculating disease rates

Abigail Stamm, Center for Environmental Health, NYS Department of Health
August 2020

The geographic aggregation tool (or GAT) is a free software tool that was developed as part of a CDC funded Environmental Public Health Tracking (or EPHT) grant.

Current version of GAT as of 5/28/2020:
Abigail Stamm and Gwen Babcock (2020). gatpkg: Geographic Aggregation Tool (GAT). R package version 1.50.

Note for examples:
Talbot et al used R v1.33 (Talbot & Babcock, 2015)
Portal used R v1.49 (Stamm & Babcock, 2020)

Outline

- Need for subcounty data
- GAT
  - What it does
  - How it works
  - Application examples

NEW YORK STATE OF OPPORTUNITY. Department of Health

GAT was recently updated and is being used in the development of a new EPHT platform in NYS, which will show subcounty measures of health and exposure. This can help the DOH identify and better understand patterns and trends of disease at a higher resolution. In this presentation, I will cover why subcounty aggregation is important and how GAT addresses it.

Why display subcounty data?

Need: High risk areas

Issues:

- Smoothing/masking (county)
- Small numbers (tract, town)

Solution: Aggregation

NEW YORK STATE OF OPPORTUNITY. Department of Health

Counties want their residents' data at higher granularity than county-level, which can mask variation, especially in counties with a mix of urban and rural populations. However, showing data at town level won't work because many rural towns have very small populations. (Ex. Hamilton County has 4800 people spread across 9 towns.) Areas with small populations are likely to have unstable rates and few cases, which can put confidentiality of cases at risk.

We want to find a happy medium where numbers are large enough to protect confidentiality and provide stable rates and areas are small enough to show rate variations. This will allow local health departments and others to use rates to identify hot spots for targeted interventions.

One way is to aggregate small areas, such as towns or census tracts. We chose to aggregate census tracts so that we could combine smaller towns as well as create smaller areas within large cities to display urban variation.

GAT's objective

Aggregate small areas to:

1. Meet minimum counts
2. Standardize process

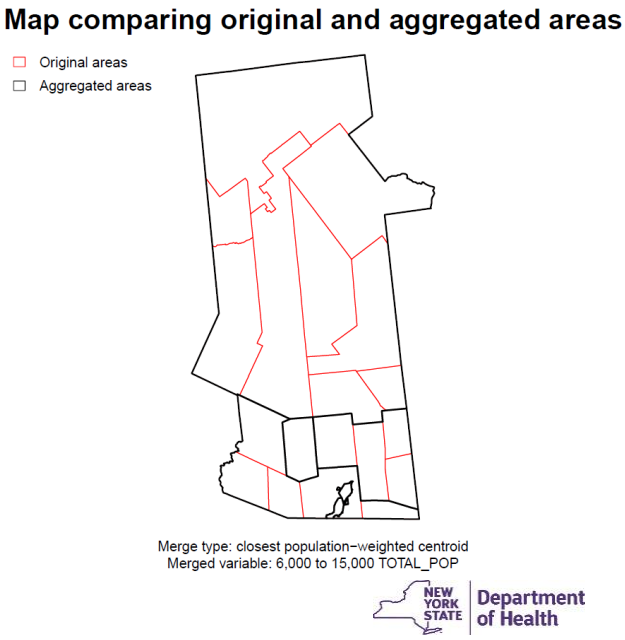
NEW YORK STATE OF OPPORTUNITY. Department of Health

We developed GAT to standardize and automate how we aggregate New York's 4900 census tracts. GAT reads in areas with small counts, such as population, deaths, or cases, and combines them until those counts meet the user's desired minimum value. At the end, GAT provides a log and other outputs to help the user record and repeat the settings they used. This allows users to map stable rates and explain to shareholders how they developed the areas they are displaying.

There are four aggregation methods included in GAT, so you can decide which method provides the most meaningful results for your data.

GAT's process

1. Request user inputs
2. Run aggregation
3. Output shapefiles and documentation



GAT is run through the free statistical software program R. In this version, we have converted it to a stand-alone package, or add-on, for R. To make it as user-friendly as possible, we developed a series of dialogs to help you select your settings.

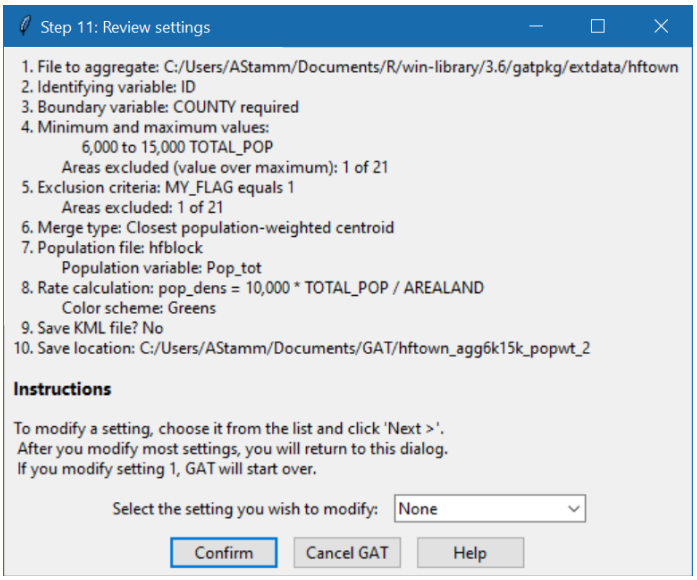
For the map shown, I ran New York State towns through GAT. This map was produced by GAT. It shows the following information.

Red borders show town boundaries and black borders show aggregated area boundaries. I chose a minimum population of 6,000 and a maximum population of 15,000. I aggregated each area with a population below 6,000 to its neighbor with the closest population-weighted centroid.

The new version of GAT produces two shapefiles, an aggregated file and a crosswalk. GAT also produces a PDF of maps and a log of the entire process, including your settings, any warnings, and a brief data dictionary. The PDF and log provide much more information than I am showing here. They are designed to help you evaluate and report your aggregation results and standardize your process.

User inputs

- Shapefile
- Minimum and maximum values
- Boundaries
- Exclusions
- Aggregation method



NEW YORK STATE Department of Health

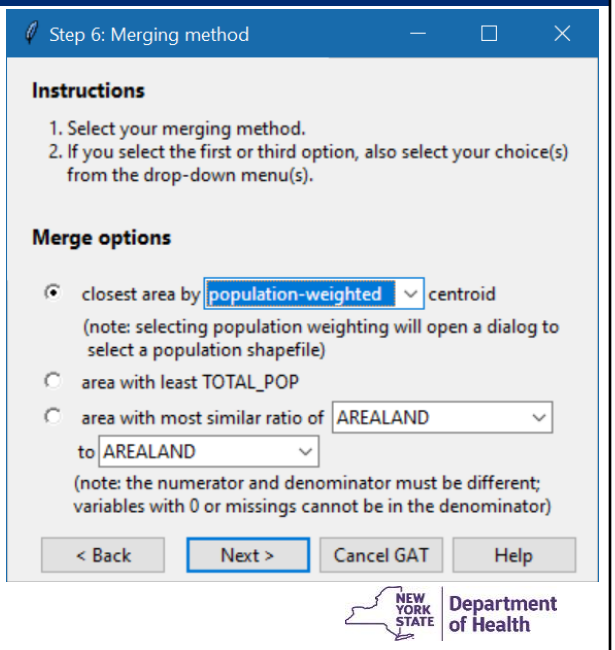
GAT requests user inputs through a series of dialogs, including menus, checkboxes, and text boxes, so no programming knowledge is necessary.

This screenshot is the final step of the user input section, where a user can verify that the information they entered is correct. The ability to return to previous steps is a new feature in GAT. This screenshot contains all settings that the user has entered, so someone else could recreate their results just from this image.

At this point, when the user clicks “confirm”, GAT proceeds to the aggregation, which can take between seconds and hours, depending on the shapefile size and settings. Two progress bars will be displayed so the user knows GAT has not stalled.

Aggregation methods

1. Closest geographic centroid
2. Closest population-weighted centroid
3. Neighbor with the lowest count
4. Most similar neighbor

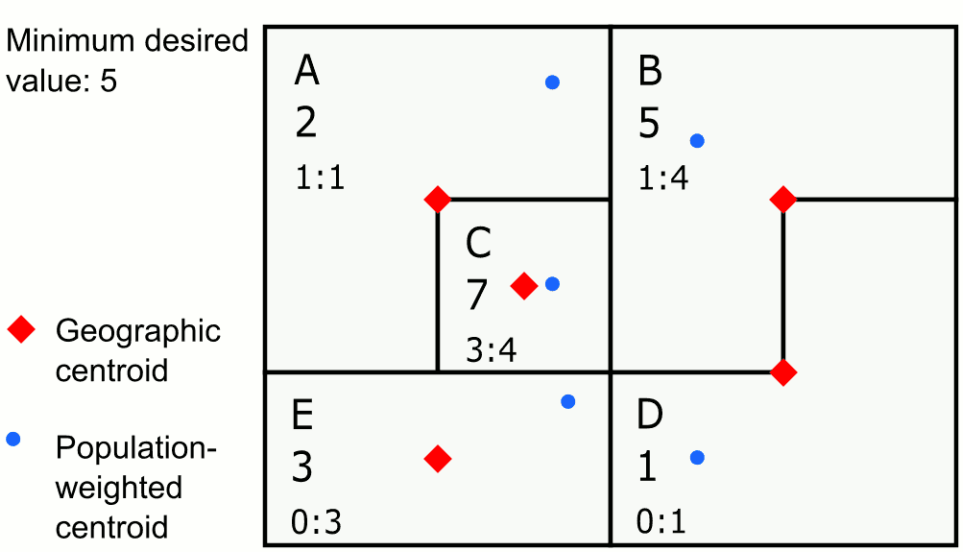


GAT contains four aggregation methods and several basic and advanced settings that allow you to refine these methods further. The screenshot shows the dialog to select your aggregation method. Using these options, you can decide which aggregation method would be most meaningful for the data you intend to show.

In my previous example, I used closest population-weighted centroid, which is a new feature in the R version of GAT. This might work well for disease rates, but what if I want to investigate something else?

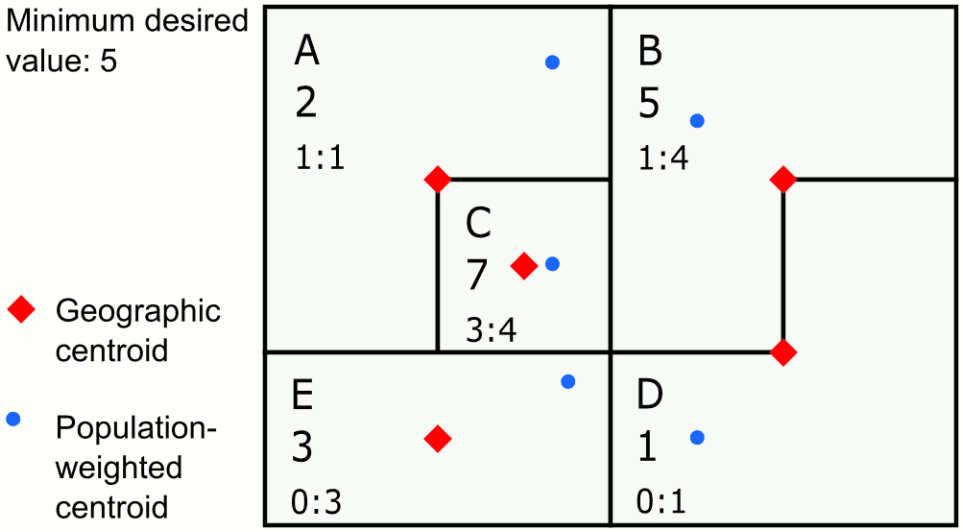
To study an environmental exposure like air pollution, you might choose geographic centroids. To maximize the number of areas, you might aggregate to the neighbor with the lowest count. To investigate social determinants of health, you might aggregate to the most similar neighbor based on a ratio of two values.

Closest geographic centroid



The graphic shows GAT's aggregation process. For all methods, GAT begins with the area with the highest count below the minimum desired value, which is E here. GAT evaluates E's neighbors, A, C, and D, to choose the most appropriate candidate based on the settings you select.

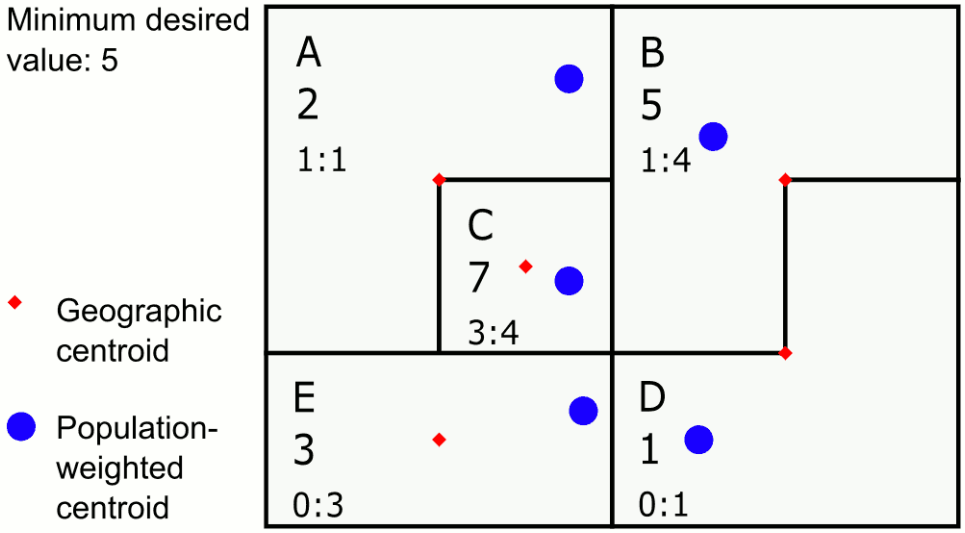
Closest geographic centroid



Here, we are aggregating by closest geographic centroid (the red diamonds). The closest neighbor to E is C, so GAT assigns E to merge with C.

With this method, each time two areas are aggregated, the new centroid becomes the mean of the previous two. This speeds up GAT considerably, but can result in centroids that do not match the area's center. This method provides the most compact areas and can work well for environmental exposures like air pollution and temperature.

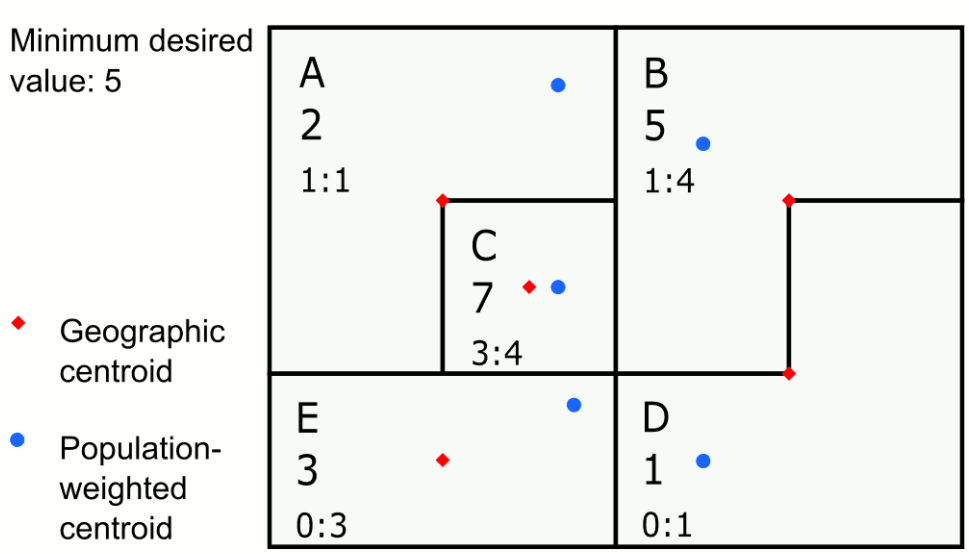
Closest population-weighted centroid



Here, GAT evaluates population-weighted centroids (the blue circles). Once again, GAT begins with area E. GAT determines that the closest neighbor is D, so GAT assigns E to merge with D.

With this method, GAT reads in a second shapefile, for example, a census block file, to calculate initial population-weighted centroids. GAT then takes a weighted average of the centroids each time two areas are aggregated. This speeds up GAT considerably, but can result in centroids that do not match the population center. For example, in a test run of NYS census tracts, using weighted centroid averages took about 40 minutes, but recalculating the centroid for each area took 3.5 hours. This method provides areas that may sprawl a bit, but are representative of population centers and can work well for disease rates.

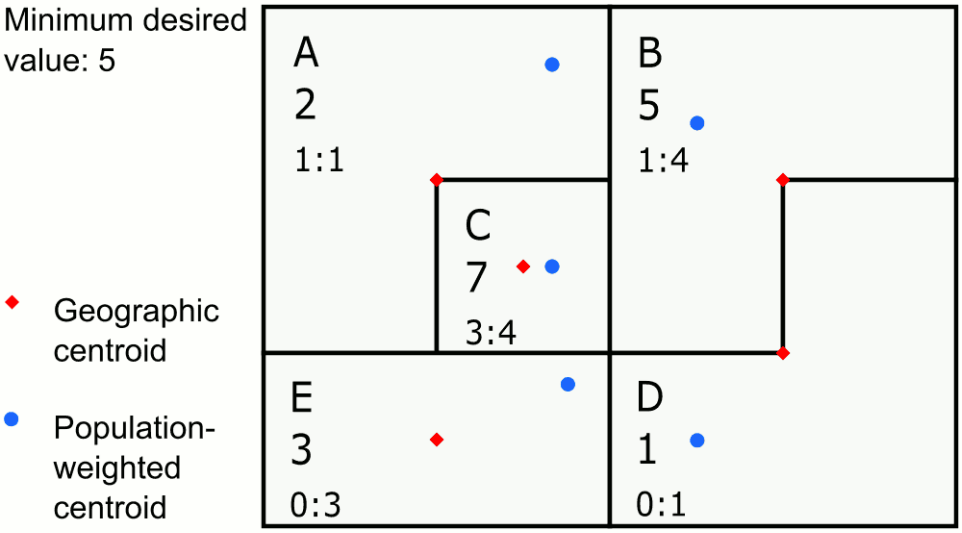
Neighbor with the lowest count



Again, GAT begins with area E. Here, GAT evaluates neighbors' populations and determines that the neighbor with the lowest population is D, so again GAT assigns E to merge with D.

With this method, each area is aggregated to its neighbor with the lowest value. This method provides the largest number of areas and therefore the greatest granularity. However, it can result in weird snaky shapes and possibly donuts, where a rural area entirely surrounds an urban one.

Most similar neighbor



Here, GAT evaluates ratios to determine the most similar neighbor. Once more, GAT begins with area E. E's ratio is zero, which is more similar to D (also zero) than to A (which is 1) or C (which is three-fourths). Again GAT assigns E to merge with D.

With this method, each area is aggregated to its neighbor with the most similar ratio of two values, for example Hispanic and non-Hispanic. This method works well if you want to create areas of similar populations to investigate social determinants of health.

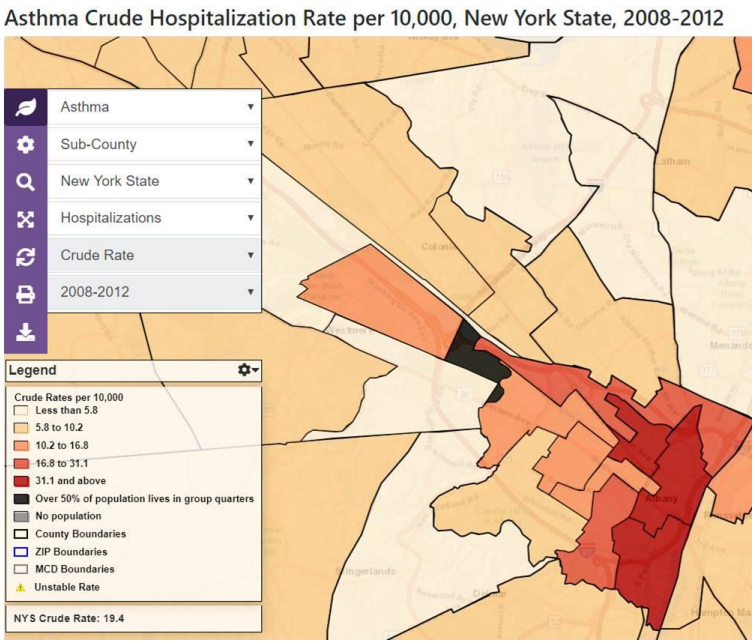
Differences between GAT 2015 and GAT 2020

	GAT 2015	GAT 2020
Format	SAS and R scripts	R package
Log	Minimal	Comprehensive
Maps	Simple, not saved	Detailed, saved to PDF
Change settings dialog	No	Yes
Population weighting	SAS yes, R no	Yes
Exclusion criteria	No	Yes
Maximum values	No	Yes

If you have used GAT before, this slide shows some of the changes we have made. If you have questions about these changes, please ask at the end.

Applying GAT: disease

- aggregation by population
- closest population-weighted centroid



(Update this image later with the revised app and larger font size)

In this example, the NYS EPHT program is developing an online portal to display various disease rates at subcounty level.

For our project, we ran two aggregations. For both aggregations, we excluded census tracts with zero population and census tracts in which at least 50% of the population lives in group quarters. (That black shape in the middle is the University at Albany dormitories.)

For the first aggregation, we merged areas within MCD boundaries to a minimum of 7500 population. For the second aggregation, we merged MCDs below 7500 population to other MCDs within the same county. In our evaluations, we determined that requiring a population of at least 7500 results in stable rates for most areas while providing meaningful detail within counties and larger cities.

Applying GAT: mortality

- aggregation by number of deaths
- closest geographic centroid

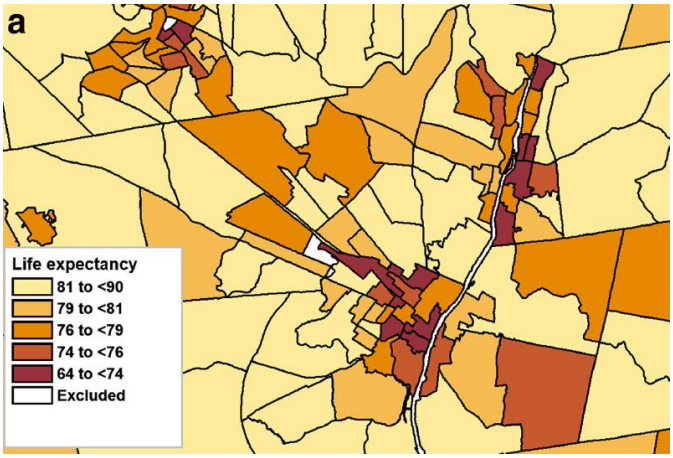


Fig. 6 Thematic Maps of the New York State Capital District after aggregation. a by life expectancy (image from Talbot et al. Population Health Metrics (2018) 16:1)

NEW YORK STATE Department of Health

In this example, published in Population Health Metrics, the researchers used a previous version of GAT to aggregate census tracts by number of deaths to at least 60 deaths per area. Their goal was to calculate life expectancies with low standard errors.

Takeaways

How GAT can help you

- Small areas with stable rates
- Standardization and documentation
- Customization

NEW YORK STATE OF OPPORTUNITY. Department of Health

GAT provides many options that you can customize to create small areas that are meaningful to you and your stakeholders, provide stable rates for whatever you are measuring, and protect patient confidentiality. GAT automatically produces documentation to help you keep track of and reproduce your process, share your methodology with others, and evaluate which settings best meet your needs.

Acknowledgements

CDC for funding
Gwen LaSelva for code and testing
NYS DOH EPHT team for testing and feedback

Email me at abigail.stamm@health.ny.gov

NEW YORK STATE OF OPPORTUNITY. Department of Health

We hope to make GAT available online soon. In the meantime, if you think GAT can help you in your work, please email me for a copy of the package and instructions to get you started.

Projects that have cited GAT

Sherman RL, Henry KA, Tannenbaum SL, Feaster DJ, Kobetz E, Lee DJ. Prev Chronic Dis 2014;11:130264. DOI: http://dx.doi.org/10.5888/pcd11.130264 (referenced R v1.2)

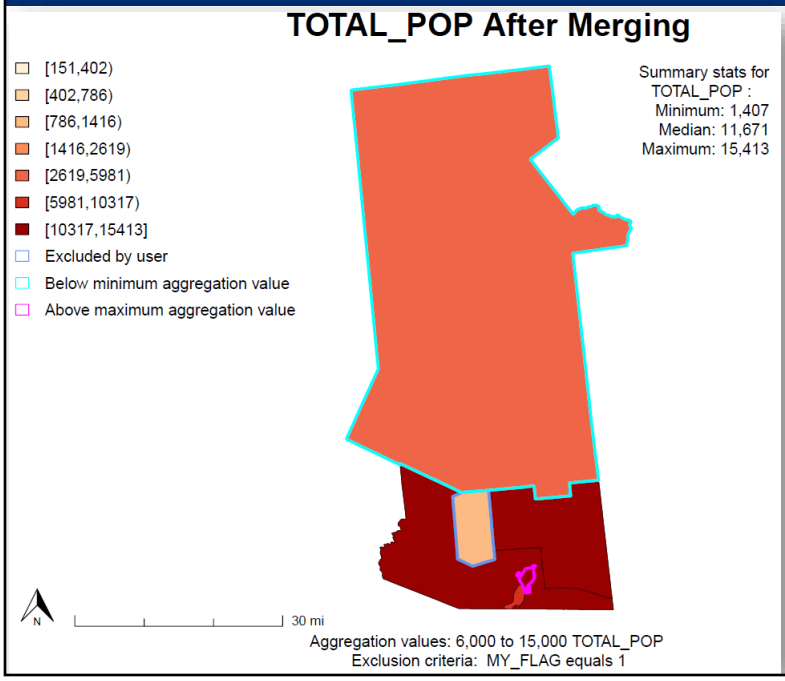
Werner AK, Strosnider HM. Spatial and Spatio-temporal Epidemiology 2020;33. DOI: https://doi.org/10.1016/j.sste.2020.100339 ((used SAS v1.31)

Werner AK, Strosnider H, Kassinger C, Shin M. *J Public Health Manag Pract*. 2018;24(5):E20-E27. doi:10.1097/PHH.0000000000000686 ((used SAS v1.31)

Boscoe FP, Talbot TO, Kulldorff M. *Geospat Health*. 2016;11(1):304. Published 2016 Apr 18. doi:10.4081/gh.2016.304 (used SAS v1.31)

Boothe VL, Fierro LA, Laurent A, Shih M. Global Diaspora News. Published 3/28/2020. https://www.globaldiasporanews.com/a-tool-to-improve-community-health-and-advance-health-equity/ (used R v1.33)

NEW YORK STATE OF OPPORTUNITY. Department of Health



Example assessment map

NEW YORK STATE OF OPPORTUNITY. Department of Health

Example log excerpt

```
NYSDOH Geographic Aggregation Tool (GAT) Log
  Version & date: 1.52 2020-07-14
  Date run: 2020-07-22
  Time GAT took to run: 5.73 minutes

Input file:           C:/Users/AStamm/Documents/R/
  Projection:         +proj=longlat +datum=NAD27 +
  Field names:        TOWN, ID, COUNTY, AREALAND,
  Identifier:         ID
  Boundary variable:  COUNTY
    You chose to require the aggregation to respec

Output file: C:/Users/AStamm/Documents/GAT/hftown_
  Number of input areas:     21
  Number of output areas:    6
  Number of aggregations:    15
  Number of excluded areas:  1

Merge type: closest population-weighted centroid
  Population file: C:/Users/AStamm/Documents/R/win
  Population variable: Pop_tot

Exclusion criteria:
  1.  MY_FLAG equals 1

First aggregation variable: TOTAL_POP
  Minimum value: 6,000
  Maximum value: 15,000
```

NEW YORK STATE OF OPPORTUNITY. Department of Health