gatpkg: Developing a geographic aggregation tool in R for non-programmers

Abigail Stamm
New York State Department of Health
July 2021

The NYS Department of Health developed the geographic aggregation tool (or GAT) as a way for anyone, irrespective of coding skills, to combine small geographic areas.

Citation (as of June 2021):
Abigail Stamm and Gwen Babcock (2021). gatpkg: Geographic Aggregation Tool (GAT). R package version 1.61.0.

Outline

- GAT: what and why
- Package development
  - User dialogs and options
  - Vignettes
  - Function examples

I'll provide a brief overview of what GAT is and why we created it. Then I will focus on strategies we used when developing GAT to reduce the need for users to know R. I'll speak from a public health perspective, but GAT can aggregate any geographic polygons by any numeric value, such as land area or species counts.

Why create GAT?

Need: Identify high risk areas

Issues:

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

Solution: Aggregation

NEW YORK STATE OF OPPORTUNITY. Department of Health

Why create GAT?
Regional data can mask variation, especially in regions with a mix of urban and rural populations. However, town-level data may be unstable due to very small populations. We aggregated census tracts for our project both to combine smaller towns and to create smaller areas within large cities.

GAT's objective

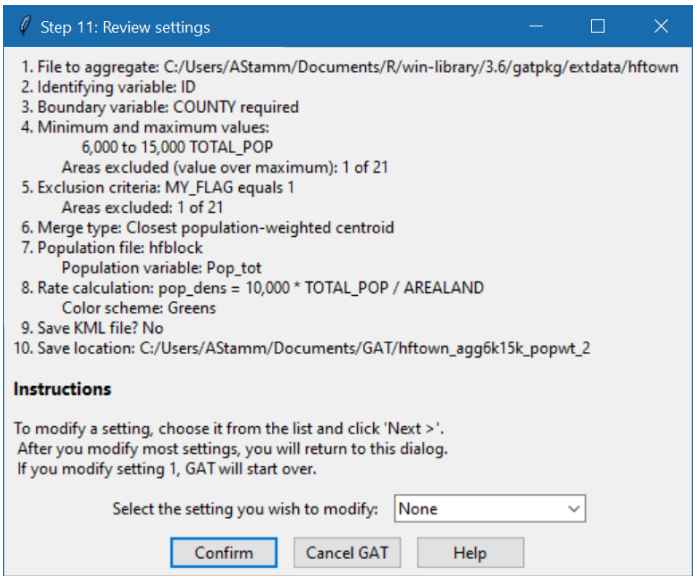
1. Aggregate small areas to:
   a. Meet minimum counts
   b. Standardize process
2. Be as user friendly as possible

NEW YORK STATE OF OPPORTUNITY. Department of Health

GAT's objective:
We developed GAT to standardize our aggregation. GAT combines areas with small counts until those counts meet the user's minimum desired value. Since many public health workers are not computer programmers, we developed GAT to be accessible to non-programmers. I'll cover a few of the strategies that we used.

User dialogs

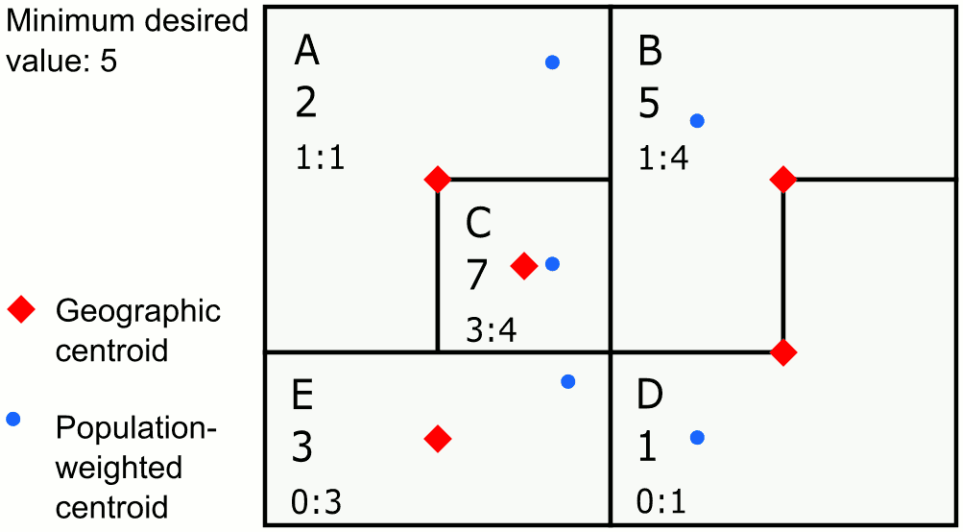
- Shapefile
- Minimum and maximum values
- Boundaries
- Exclusions
- Aggregation method



NEW YORK STATE Department of Health

User dialogs:
We developed a series of dialogs to walk the user through GAT's basic options using the packages tcltk and tcltk2. In the confirmation dialog shown, the user verifies all selections. The drop-down list at the bottom allows the user to return to a previous step, correct their selection, and jump seamlessly back to this dialog. Since this dialog shows all basic settings that the user selected, someone could replicate results just from this image and the user's function call.

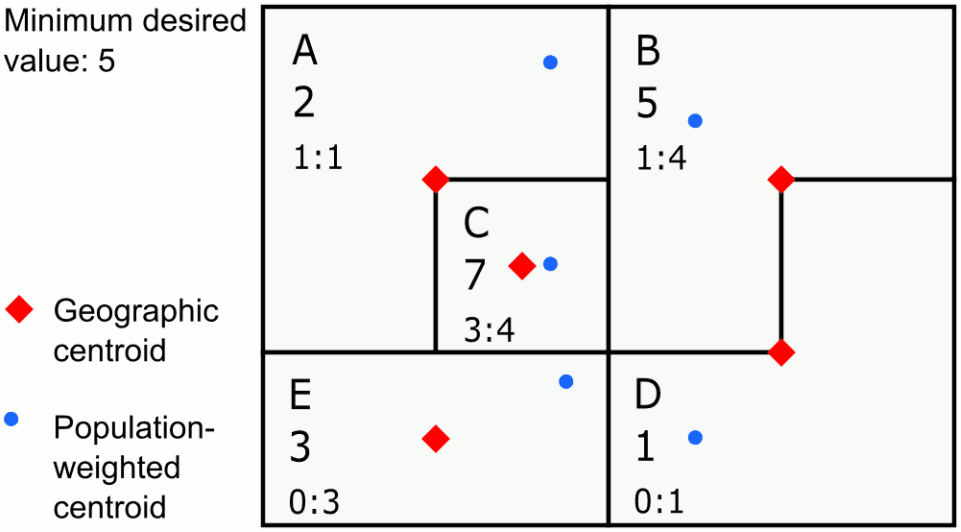
Preparing for aggregation



NEW YORK STATE Department of Health

Preparing for aggregation:
GAT begins with the area with the highest count below the minimum desired value and selects the area to merge first based on the aggregation method chosen. For this illustration, we will aggregate to the closest geographic centroid.

How GAT aggregates

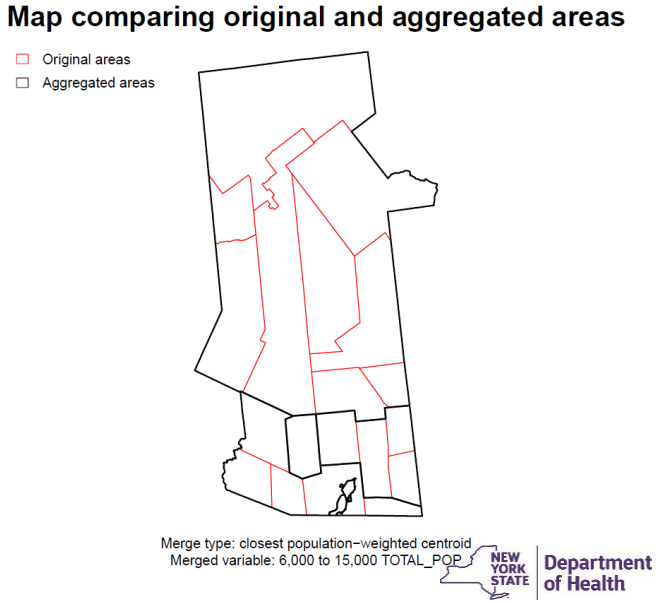


NEW YORK STATE Department of Health

How GAT aggregates:
Here, the minimum desired value is 5. The area with the highest value below 5 is E. GAT evaluates E's neighbors and determines the closest area is C, so GAT joins E to C. Aggregation continues until all areas contain values of at least 5.

GAT's process

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



GAT's process:
GAT generates several maps, including the one shown, which it writes to a PDF. GAT also writes a log of the entire process, a settings file, and on request, a KML file. The PDF, settings file, and log are designed to help the user evaluate and report aggregation results. Last, GAT saves two shapefiles, a shapefile of aggregated areas and a crosswalk shapefile that adds aggregated area IDs as a new variable to the original shapefile.

Reproducibility: Log

```
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
```

NEW YORK STATE | Department of Health

Reproducibility: Log

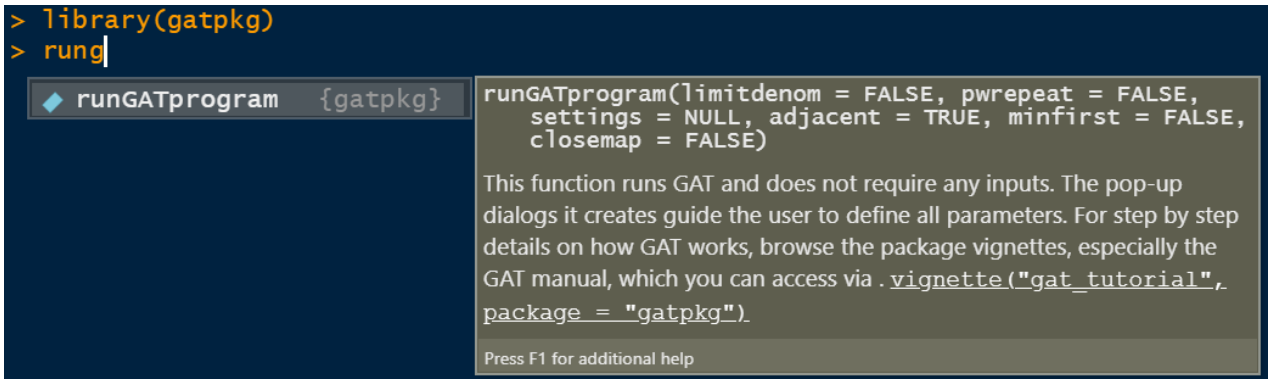
The log contains information about GAT's run, the input shapefile, merge settings, and aggregation variables. The log helps users to:

(1) remember what settings they used six months later

(2) identify areas that may not have aggregated correctly

(3) read the output shapefile into another GIS program

Advanced options



NEW YORK STATE | Department of Health

Advanced options:
The function to run GAT is aptly named runGATprogram. These options were placed here instead of the dialogs for two reasons.
(1) Defaults for these options are what we most commonly use.
(2) These options are complicated to explain.
We felt that we could cover them more clearly in the technical notes and other documentation than in the brief dialog instructions.

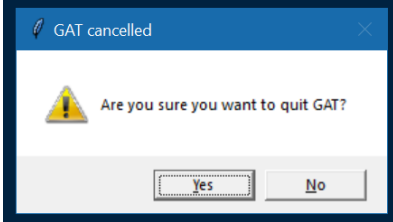
Accessibility: function help

confirmGATquit {gatpkg} R Documentation

Confirm Quitting GAT

Description

This function opens a dialog window for the user to select whether to quit or continue GAT.



Click on one of the following buttons.

- Click `Yes` if you would like to quit GAT.
- Click `No` if you do not want to quit GAT.

NEW YORK STATE | Department of Health

Accessibility: function help
This example shows the simplest dialog in GAT. All dialogs and maps are drawn using functions that standardize their appearance. Because GAT relies heavily on dialogs, we included images of those dialogs in both the help functions and the vignettes. Help buttons in most dialogs can send users to relevant help files in case they do not know how to access help via the console.

Accessibility: vignettes

- Readable
- Relevant
- Easy to navigate
- Self-contained

Accessibility: vignettes

We have developed eight vignettes in three general categories: package information, technical notes, and using GAT. The vignettes on using GAT are designed for users with little to no prior knowledge of R. They include:

(1) How to structure the shapefile to be read into GAT based on what the user plans to do
(2) A step-by-step tutorial of using GAT with an embedded shapefile
(3) How to identify areas that may not have merged correctly so the user can evaluate them separately

The tutorial includes links to other vignettes and help documentation for the relevant functions, so an advanced user who wants to customize GAT will know where to start.

Takeaways

Ways to increase accessibility

- Dialog boxes
- Scaled documentation
- Detailed examples

NEW YORK STATE OF OPPORTUNITY. Department of Health

Takeaways:
At all stages of development, we asked how to tailor an R tool for people who do not regularly use R. We wrote dialog boxes to nearly eliminate the need to write code, documentation that catered to people with a range of R knowledge, and examples that detailed the necessary function inputs and could be run using only the resources bundled in GAT and the packages it requires.

Acknowledgements

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

GAT online: https://github.com/ajstamm/gatpkg
or email me at abigail.stamm@health.ny.gov

I acknowledge GAT coauthor Gwen LaSelva, the CDC, and the many people who have tested and provided feedback during GAT's development. If you are interested in using GAT, please visit the GitHub link provided or email me.