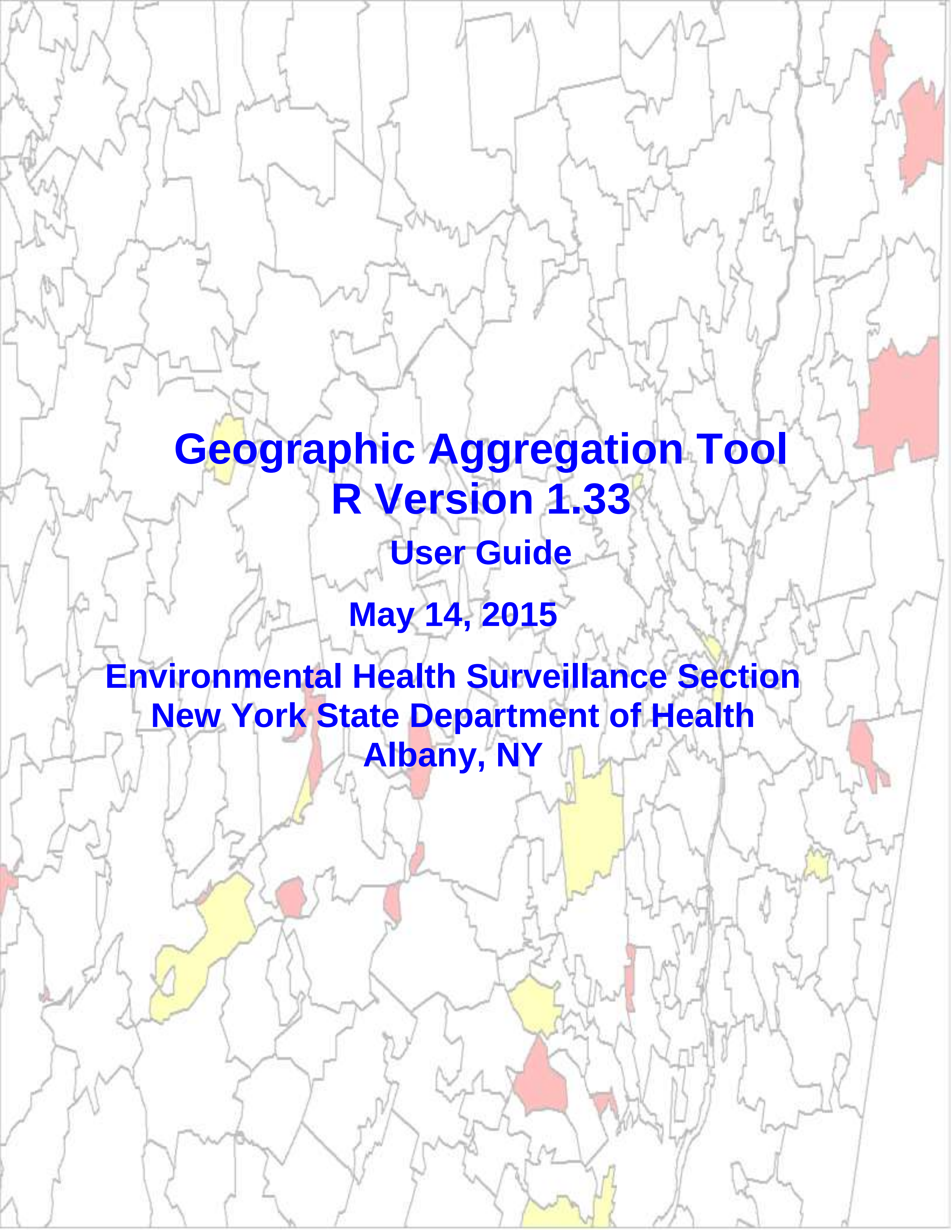


Geographic Aggregation Tool
R Version 1.33

User Guide

May 14, 2015

Environmental Health Surveillance Section
New York State Department of Health
Albany, NY

Contact Information:

Please send comments and suggestions to:

Thomas Talbot, Chief
Environmental Health Surveillance Section
New York State Department of Health
Albany, NY Email: thmstalbot@gmail.com

Introduction

Health outcome maps with fine geographic resolution can be misleading due to random fluctuations in disease rates caused by small numbers. In some cases these maps can also inadvertently disclose confidential data. To overcome these limitations the New York State Department of Health's Environmental Health Surveillance Section developed the Geographic Aggregation Tool to join neighboring geographic areas together until a user defined population and/or number of cases is reached. This tool can be used to produce maps for the public at the finest geographic resolution practicable.

The tool allows for restrictions so that areas are not unnecessarily merged across larger regions. For example, one may want to merge census tracts into larger areas but not have these larger areas cross county boundaries. One could also restrict the merging so that areas are not merged across physical features such as mountain ranges or large water bodies.

The tool has been tested in R 3.0.2 32 and 64 bit versions, under Windows 7. The user does not need R programming experience to run the tool. The user must have an input shapefile (*.shp). The shapefile must have the geographic boundaries of the areas the user would like to merge, a character variable that uniquely identifies these areas, and a numeric variable to merge by, such as the number of cases of disease or the population.

The program outputs a shapefile containing the aggregated regions with associated data, which can be used in GIS programs such as ArcGIS, MapInfo and QGIS. A KML file will also be produced, which can be displayed in Google Earth as well as other Internet-based mapping programs.

In addition to the tool we have also provided a sample map file to use with the tool. The file contains the simulated total number of low birth weight births, and the simulated number of births which occurred in each ZIP code area in New York State. The file also contains sociodemographic data derived from the US Census on race, poverty and educational attainment.

If the software is used we would appreciate a citation for our work.

Disclaimer: This is a version of the GAT and is provided as is. We would appreciate suggestions for improvement and reports of bugs.

Thomas O Talbot

Table of Contents

Requirements to run the tool:

The tool has been tested under Windows XP and Windows 7 and may work on other versions of Windows. We anticipate that the tool would require modification to run under Mac or Linux operating systems.

Files needed to run the tool:

Program files:

- Batch file: GAT vR13.bat
- R script file: GATinRv13.R
- Required Software (provided):
  - R 3.0.2 32 or 64 bit
- R packages: foreign, classInt, maptools, RColorBrewer, rgdal, rgeos, spdep, svDialogs, and their dependent packages

Data

You may use the provided sample data to learn how to use the program. This simulated birth outcome data “lbwbyzip” contains the total number of singleton births and the number of low birth weight births by ZIP Code over a 5 year period. These data also include information on educational attainment, race and the number of children living in poverty for each ZIP Code. A data dictionary is provided in the Technical Notes at the end of this document.

The GAT tool requires a cleaned shapefile, with no overlapping geographic areas. Having a large number of geographic areas relative to the amount of computer memory available to you will cause the program to fail. (The maximum number of areas is likely to be over 100,000, depending on the amount of computer memory you have). There should be one record for each geographic area to be merged, and each should have a unique ID. The ID should be a character variable. The data should also have one or two numeric variable(s) to merge by, such as the number of cases of disease and/or the population. The data may contain another boundary variable. If a boundary variable is used, GAT can first merge areas within the larger boundary areas before merging across the larger boundary regions. Your data should not contain any fields beginning with 'GAT' as the program uses variable names with this prefix. Your data should contain 250 fields or less, as the GAT tool will add fields, and the shapefile output can only accept a limited number of fields.

Recommendations to run the tool:

For best performance, we recommend using a shapefile with relatively low resolution areas. This can be accomplished by reducing the number of boundary nodes of the geographic areas using mapping software. See

Thinning Geographic Boundaries for details.

The tool is best used in conjunction with mapping software such as ArcGIS, Google Earth, MapInfo, or QGIS.

Setting up the R Program and batch file

Installing the GAT Tool

The GAT tool is being distributed through the Internet and flash drives. If you received the tool on a USB drive, install the program on your computer or computer network by copying and pasting all the files and directories provided to the desired location on your computer.

If you received the GAT Tool through the Internet or file hosting service, the files are compressed into one .ZIP file. Unzip the files and then copy and paste all the files and directories provided to the desired location on your computer.

The files you received include **all** the necessary files to run the tool including R program packages. The program can be launched by opening the batch file *GAT vR13-64.bat* or *GAT vR13-32.bat* by following the path *drive:\ your_path\GAT\R\program\*. *GAT vR13-64.bat* should be used for 64-bit systems, *GAT vR13-32.bat* for 32-bit systems.

Use the Version of R provided

Because R and its packages are being constantly updated, it is recommended that you use the R version supplied, rather than downloading the R program and the necessary R packages. The R program is described on http://www.r-project.org/. Using versions of R and R packages other than those provided may cause the program to fail. We have tested the GAT tool with R-3.0.2. The program will not work with R 3.0.1, but we are unable to predict whether the program will work properly with other versions of R.

Batch file setup

The batch file calls the R program, which produces the aggregated areas. You can then bring the results of these programs into a GIS package. If you change the file structure or the location of the batch file, the batch file may need to be edited so that the locations of the R program and the R script are correct. To edit the batch file, right click on it and select “edit”. The batch file for 64-bit R contains the following text:

```
setlocal
set R_PROFILE_USER=GATinRv13.R
"..\..\R-3.0.2\bin\x64\Rgui.exe"
endlocal
```

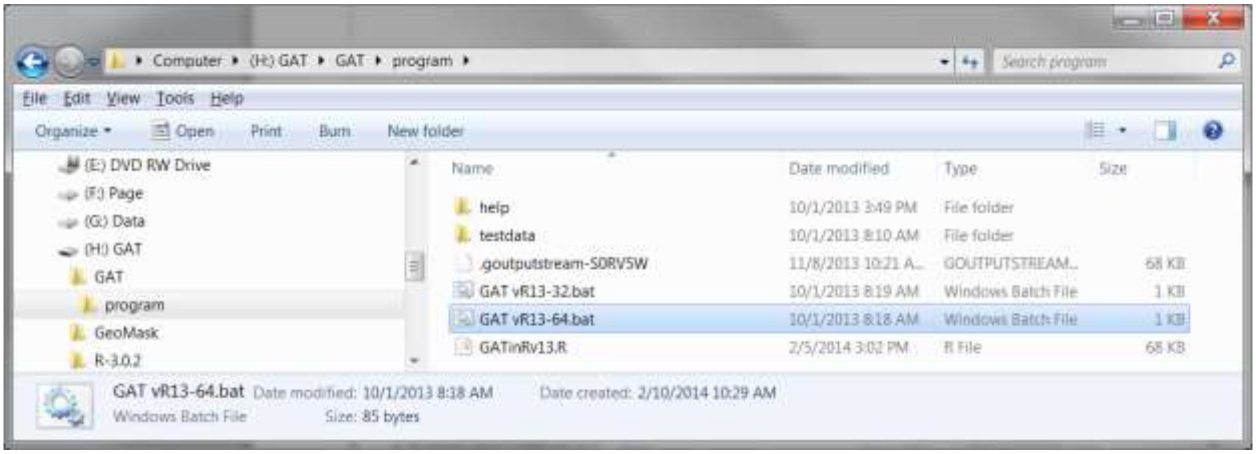
The second line contains the path and filename of the R script, and the third line contains the path and filename of the R program. If you change the location of the R program or the R script, you may need to change the path in your batch file to ensure that the program will work properly. The path in the drive(s) batch file must match the drive and directory(s) where the R program and R script are stored. The path in the batch file can be changed by opening the batch file in notepad and changing both paths in the batch file and then saving it. ***Please be careful to only change the path, even a deleted quotation mark may cause the batch file not to run. We recommend saving a copy of the original version of the batch file incase a problem does arise.*** Here is an example of an edited batch file:

```
setlocal
set R_PROFILE_USER=P:\Sections\EHS\Aggregation\GAT\R\GATinRv13.R
"S:\APPS\R-3.0.2\R-3.0.2\bin\x64\Rgui.exe"
```

endlocal

To run the tool:

Browse to the program folder. Double click on *GAT vR13-64.bat* (or *GAT vR13-32.bat* if you have a 32-bit operating system or don't know what kind of operating system you have).



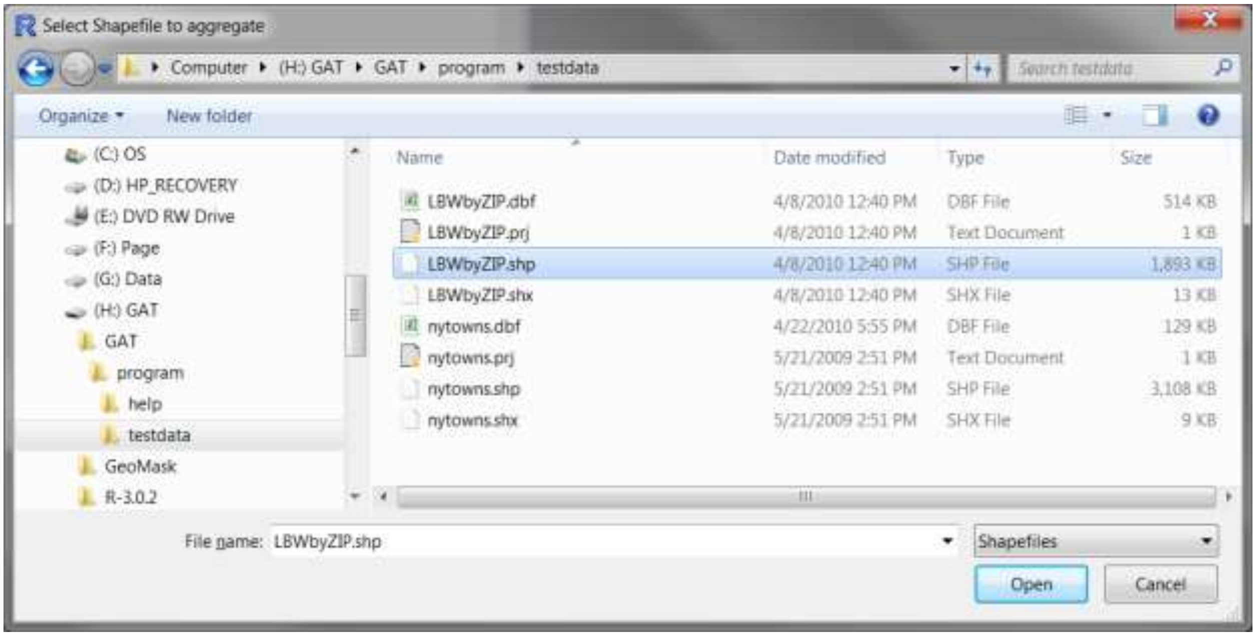
Name	Date modified	Type	Size
help	10/1/2013 3:49 PM	File folder	
testdata	10/1/2013 8:10 AM	File folder	
.goutputstream-S0RV5W	11/8/2013 10:21 A...	GOUTPUTSTREAM...	68 KB
GAT vR13-32.bat	10/1/2013 8:19 AM	Windows Batch File	1 KB
GAT vR13-64.bat	10/1/2013 8:18 AM	Windows Batch File	1 KB
GATinRv13.R	2/5/2014 3:02 PM	R File	68 KB

GAT vR13-64.bat Date modified: 10/1/2013 8:18 AM Date created: 2/10/2014 10:29 AM
Windows Batch File Size: 85 bytes

You will briefly see a DOS/cmd.exe window; the RGUI window will appear; then, after a few seconds, you will see a popup window titled "Select Shapefile to aggregate".

Selecting the Data file

If you do not see the "Select Shapefile to aggregate "window, you may need to minimize all of your other windows to bring it to the front. Select your file and click "open".

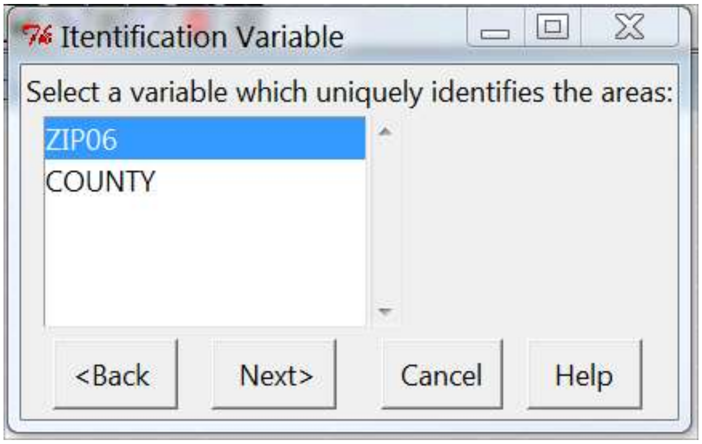


Name	Date modified	Type	Size
LBWbyZIP.dbf	4/8/2010 12:40 PM	DBF File	514 KB
LBWbyZIP.prj	4/8/2010 12:40 PM	Text Document	1 KB
LBWbyZIP.shp	4/8/2010 12:40 PM	SHP File	1,893 KB
LBWbyZIP.shx	4/8/2010 12:40 PM	SHX File	13 KB
nytowns.dbf	4/22/2010 5:55 PM	DBF File	129 KB
nytowns.prj	5/21/2009 2:51 PM	Text Document	1 KB
nytowns.shp	5/21/2009 2:51 PM	SHP File	3,108 KB
nytowns.shx	5/21/2009 2:51 PM	SHX File	9 KB

File name: LBWbyZIP.shp
Shapefiles

Selecting the Identification Variable

A window will pop with a list of character variables from the Shapefile you selected. Pick the variable which uniquely identifies the area (the ID variable) and click "Next". If the ID variable does not uniquely identify the areas, you will be asked to choose another variable if one is available. Examples of ID variables might be variables containing county names or zip codes



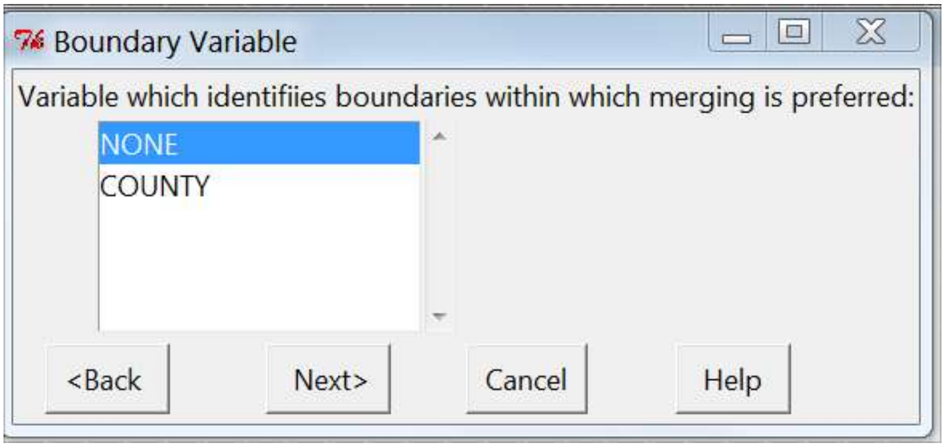
Itentification Variable

Select a variable which uniquely identifies the areas:

ZIP06
COUNTY

Selecting the Boundary Variable

Next, another window will pop up, asking you to enter the boundary variable. This variable identifies the areas within which merging is preferred. For example, if your map contains ZIP codes, you may prefer to have the ZIP codes aggregate within the same county if possible, and only if there are no ZIP codes in the same county, merge to the next county. Your boundary variable must be a character variable. Select "NONE" if you do not wish to use a boundary variable. Click "Next".



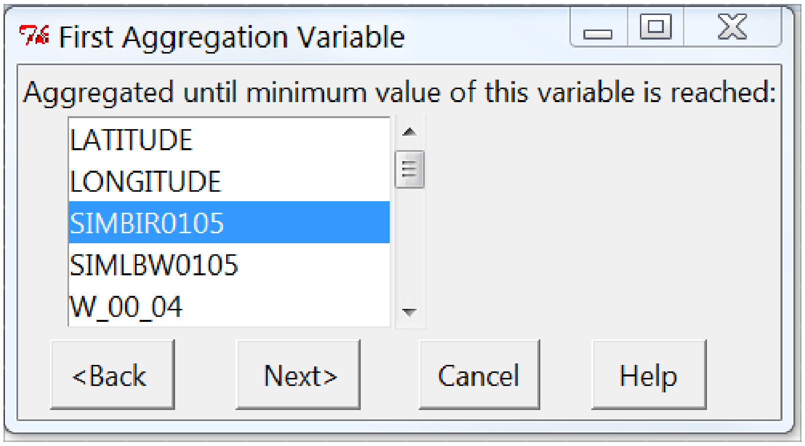
Boundary Variable

Variable which identifiies boundaries within which merging is preferred:

NONE
COUNTY

Selecting the Aggregation Variable

Next, another window will pop up asking for the aggregation variable. This is the first variable for which you wish to have some minimum value in all of the output regions. For example, if you wish all of the output regions to have a population of at least 20,000, select the variable which represents the population. Click "Next". If there is only one possible aggregation variable to choose, a different style window will appear.

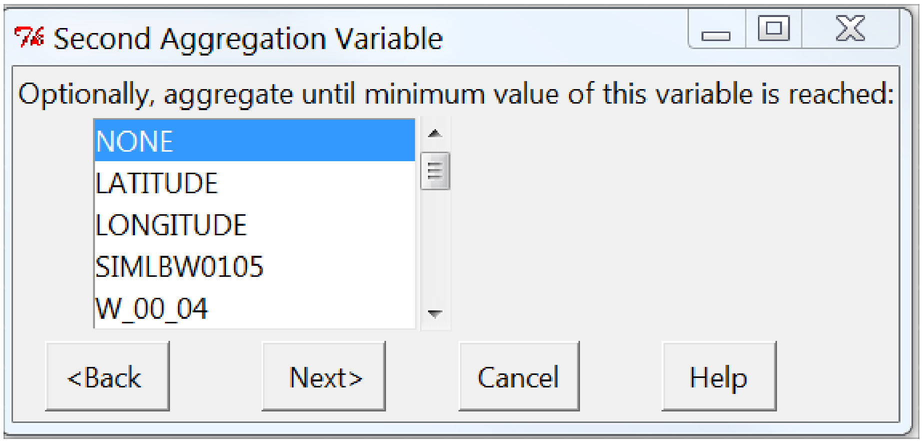


Aggregated until minimum value of this variable is reached:

LATITUDE
LONGITUDE
SIMBIR0105
SIMLBW0105
W_00_04

Selecting an Additional Aggregation Variable

Another window will appear which allows you to optionally select a second variable for aggregation. If you do not wish to select a second variable, select "NONE", otherwise, select the variable of interest. Click "Next".

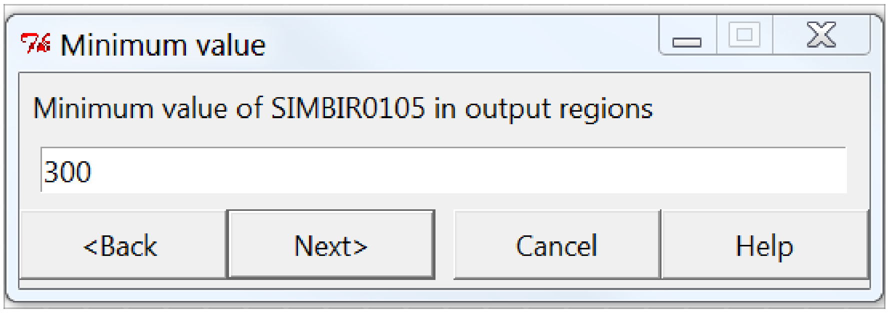


Optionally, aggregate until minimum value of this variable is reached:

NONE
LATITUDE
LONGITUDE
SIMLBW0105
W_00_04

Selecting the minimum Value(s) to aggregate to

Another window will appear, asking for the minimum value of the variable you just selected. The default value will be the largest value in your map.



Minimum value of SIMBIR0105 in output regions

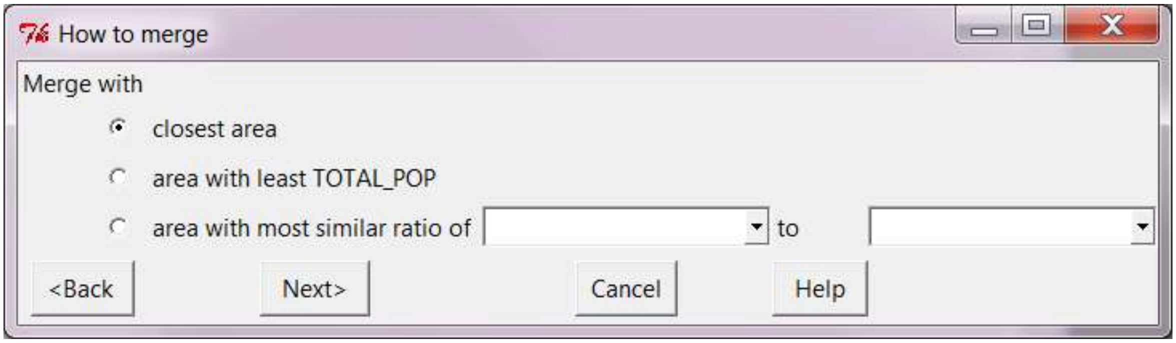
300

Enter the minimum value desired and select "Next".

If you have selected a second aggregation variable, you will be prompted to enter a second minimum value. Enter the minimum value desired and select "Next."

Selecting the merge type

Next, a window will appear, prompting you for the type of merge. Selecting the “closest area” option will result in fewer, more compact aggregated areas. Selecting the “area with least...” option will result in more, less compact aggregated areas. Selecting “area with most similar…” will favor merging areas with the smallest difference in the ratio of the variables selected. If you choose this option, the second variable must not contain any values of zero. See How the program merges for more information.



How to merge

Merge with

● closest area

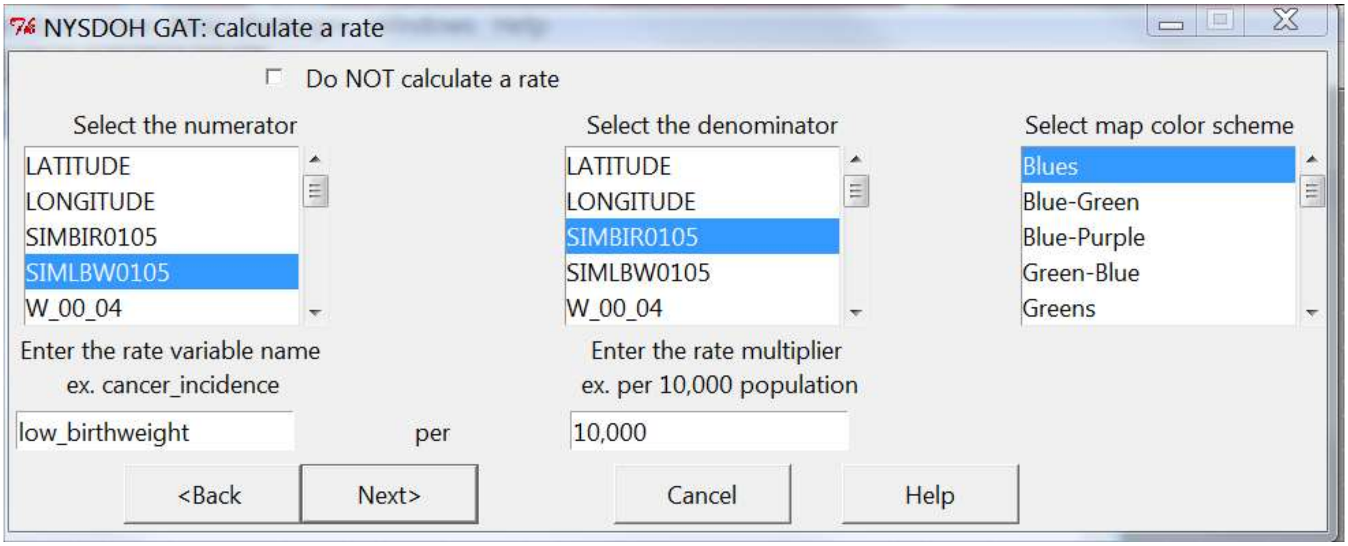
○ area with least TOTAL_POP

○ area with most similar ratio of ▾ to ▾

Select an option and click “Next”.

Calculate a Rate

A dialog window will appear which allows you to select variable for a rate calculation. This is entirely optional. If you do not wish to calculate a rate, simply click in the “Do NOT calculate a rate” box and click “Next”. If you do wish to calculate a rate, select a variable for the numerator (this variable should be a count) and a variable for the denominator (should also be a count). Enter a name for the rate and a multiplier. Enter “1” in the multiplier box if no multiplier is needed. Optionally, change the color scheme for the rate map. To calculate the rate, the numerator will be multiplied by the multiplier and divided by the denominator.



NYSDOH GAT: calculate a rate

□ Do NOT calculate a rate

Select the numerator	Select the denominator	Select map color scheme
LATITUDE	LATITUDE	**Blues**
LONGITUDE	LONGITUDE	Blue-Green
SIMBIR0105	**SIMBIR0105**	Blue-Purple
SIMLBW0105	SIMLBW0105	Green-Blue
W_00_04	W_00_04	Greens

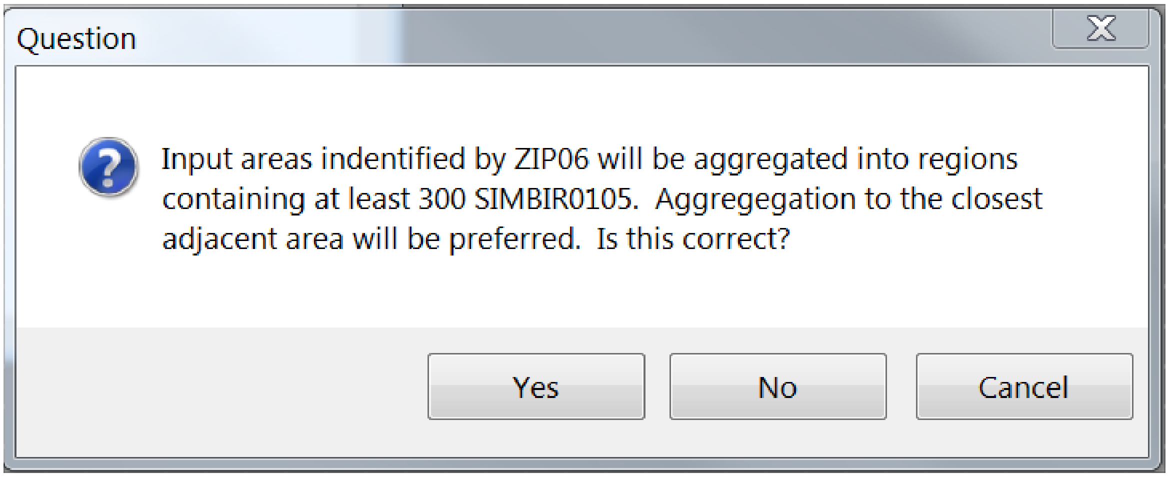
Enter the rate variable name
ex. cancer_incidence

Enter the rate multiplier
ex. per 10,000 population

low_birthweight per 10,000

Confirm your selections

A confirmation dialog will appear. Check that everything is correct, then click "Yes" if everything is fine, "No" if not.

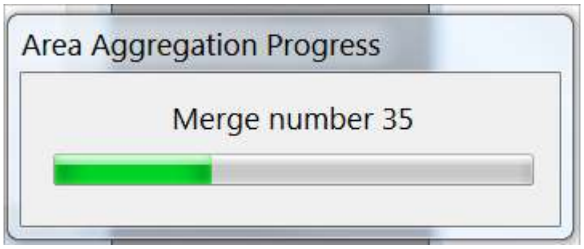


Input areas indentified by ZIP06 will be aggregated into regions containing at least 300 SIMBIR0105. Aggregegation to the closest adjacent area will be preferred. Is this correct?

Wait while the program runs

The program will run. A map or maps of the areas before aggregation will appear.

The aggregation may take a few minutes. The progress of the aggregation can be followed by watching the status bar or progress bar. The progress bar will appear when aggregation has started; it will count the merges and show them by 100's.

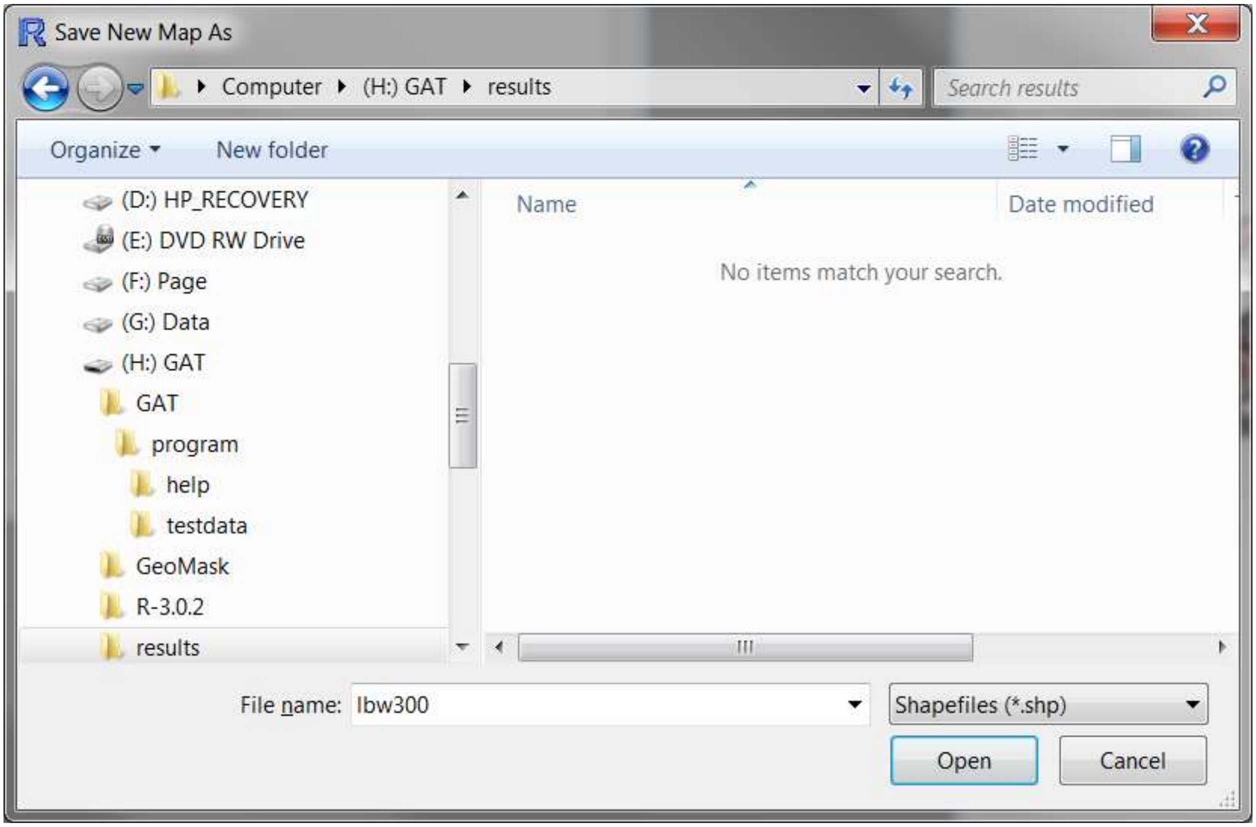


Merge number 35

Messages displayed in the progress bar and at the bottom of the "RGui: NYSDOH Geographic Aggregation Tool (GAT)" window indicate the number of merges made. When all the merges have been completed, the program will display maps showing the aggregated areas:

Save your output files

You will then be prompted to save the output files. Enter a filename and click “open”.

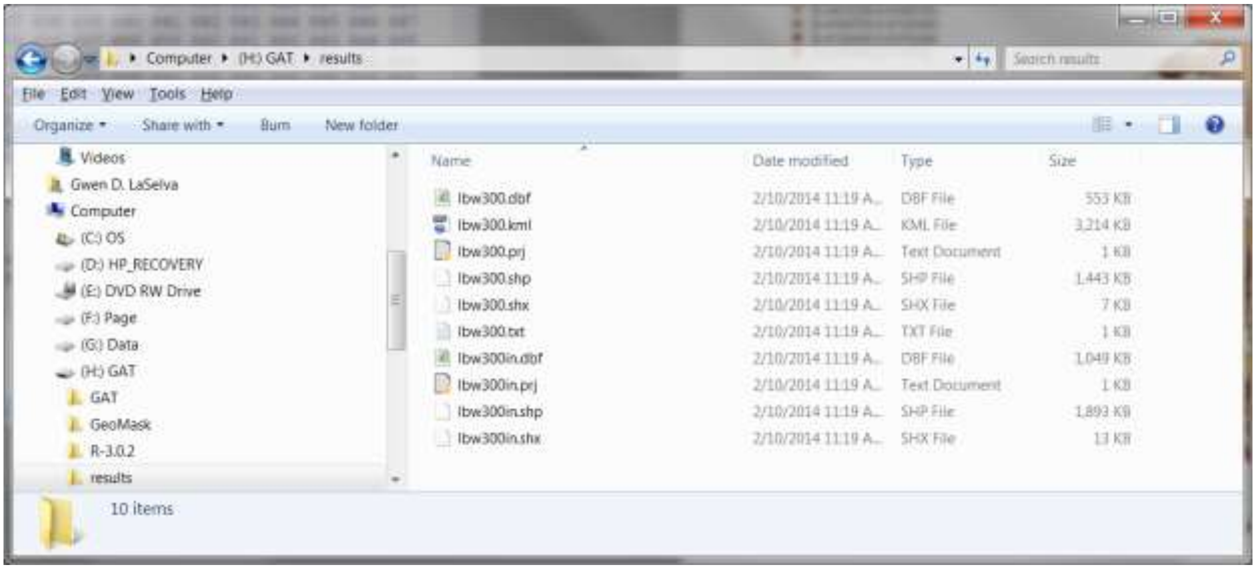


File name: lbw300

Shapefiles (*.shp)

There are 10 files which will be saved to the same location. First, a shapefile (.shp) file set (four files) is created which includes the original boundaries and data along with a new column (“gatid”) indicating which areas are merged together. This file has the suffix “in” before the file extension. Second, a shapefile file set (four files) is produced with the new aggregated boundaries. In this file, numeric fields will be summed over all the merged areas, except for fields named “Latitude”, “Longitude”, “x”, “y”, “Lat”, “lon”, or “long”, which will be averaged over all the merged areas. Character variables (except for the variable which identifies the areas) will be assigned the value of the area last merged into the region. The identification variable will remain the same in the new shapefile if the area was not merged, but aggregated areas will be assigned a unique ID beginning with “GAT”. The variables “GATx”, “GATy”, and “cratio” will be added to the dataset. “GATx” and “GATy” represent the geographic centroids as calculated by GAT. “cratio” is a measure of compactness based on the ratio of the area to the area of the minimum enclosing circle area; it ranges from 0 to 1; 1 is the most compact and 0 is the least compact. In addition, a variable representing the number of areas per region will be added; the variable will be named “num_”+first 5 characters of the identification variable+”s”. If you requested that a rate be calculated, the rate variable will also be added. Third, a text file (.txt) of the same name will be created which contains information about the settings used to produce the output. Fourth, a KML file (.kml) of the same name will also be created.

Here is an example of the output files:



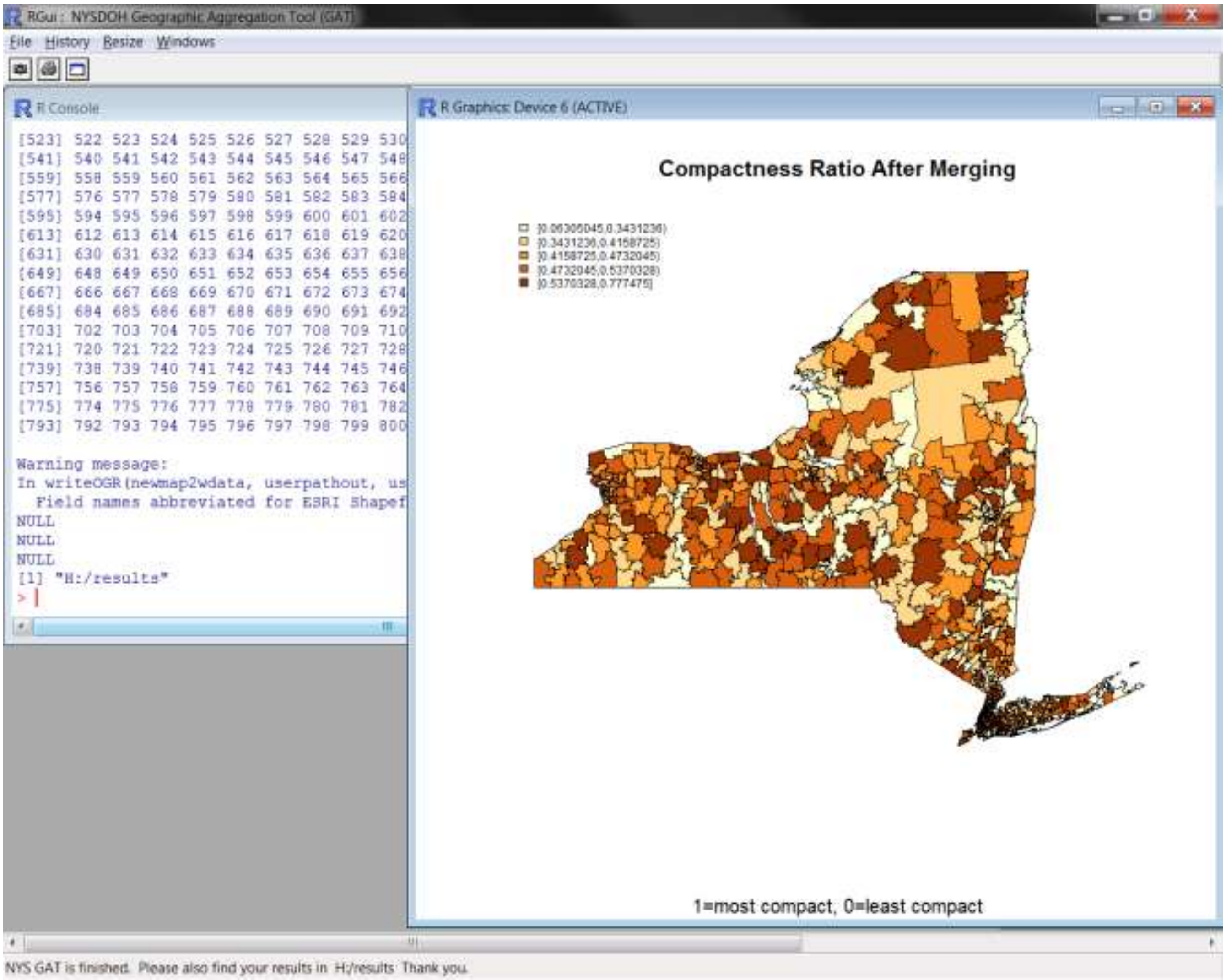
Name	Date modified	Type	Size
lbw300.dbf	2/10/2014 11:19 A...	DBF File	553 KB
lbw300.kml	2/10/2014 11:19 A...	KML File	3,214 KB
lbw300.prj	2/10/2014 11:19 A...	Text Document	1 KB
lbw300.shp	2/10/2014 11:19 A...	SHP File	1,443 KB
lbw300.shx	2/10/2014 11:19 A...	SHX File	7 KB
lbw300.txt	2/10/2014 11:19 A...	TXT File	1 KB
lbw300in.dbf	2/10/2014 11:19 A...	DBF File	1,049 KB
lbw300in.prj	2/10/2014 11:19 A...	Text Document	1 KB
lbw300in.shp	2/10/2014 11:19 A...	SHP File	1,893 KB
lbw300in.shx	2/10/2014 11:19 A...	SHX File	13 KB

Performance measures: the compactness ratio

After saving the files, the program will display a map of the compactness ratios of the newly created regions. Darker regions are the most compact.

If the program fails to run there may be warning messages in the R Console window. These messages, however, are often difficult to interpret without being an R programmer. Some warnings are normal and the program will successfully produce the desired output.

We recommend closing the RGUI after running the program and inspecting the maps. When prompted to save the workspace image, say “no”.



```
[523] 522 523 524 525 526 527 528 529 530
[541] 540 541 542 543 544 545 546 547 548
[559] 558 559 560 561 562 563 564 565 566
[577] 576 577 578 579 580 581 582 583 584
[595] 594 595 596 597 598 599 600 601 602
[613] 612 613 614 615 616 617 618 619 620
[631] 630 631 632 633 634 635 636 637 638
[649] 648 649 650 651 652 653 654 655 656
[667] 666 667 668 669 670 671 672 673 674
[685] 684 685 686 687 688 689 690 691 692
[703] 702 703 704 705 706 707 708 709 710
[721] 720 721 722 723 724 725 726 727 728
[739] 738 739 740 741 742 743 744 745 746
[757] 756 757 758 759 760 761 762 763 764
[775] 774 775 776 777 778 779 780 781 782
[793] 792 793 794 795 796 797 798 799 800

Warning message:
In writeOGR(newmap2wdata, userpathout, us
  Field names abbreviated for ESRI Shapef
NULL
NULL
NULL
[1] "H:/results"
> 
```

NYS GAT is finished. Please also find your results in H:/results Thank you.

Troubleshooting

Problem	Solution
When I import my shapefile into my GIS program, a blank map appears	When importing a shapefile, be sure that the same projection is selected in the GIS program as in the original map file
No KML file is produced	Do not end the program before the "NYS Geographic Aggregation Tool is finished" message appears in the taskbar and the "Compactness Ratio After Merging" map appears.
I'm not asked where to save my output files	The dialog box may be hidden beneath other windows. Close extraneous windows or check the taskbar for the dialog box.
When I click on a region in Google Earth, the data does not appear in the description callout.	Use Google Earth 5.0 or greater. The data may not appear in GE 4.x. Try clicking on the area's link in the left sidebar to see the data.
Output files are not saved and this message appears in the R console: "Error in writeOGR(newmapwdata, userpathout, userfileout, driver = "ESRI Shapefile", : GDAL Error 4: Failed to open shapefile G:/aggtest/output/testin\.shp. It may be corrupt or read-only file accessed in update mode."	Be sure to enter a valid file name in the save dialog. Remove all files with extensions only (such as ".shp", ".dbf", ".prj") from the directory you are trying to save to
Shapefile not saved and a message similar to this appears in the R console: "Error in writeOGR(newmapwdata, userpathout, userfileout, driver = "ESRI Shapefile", : GDAL Error 1: Invalid index : -1 In addition: Warning message: In writeOGR(newmapwdata, userpathout, userfileout, driver = "ESRI Shapefile", : Non-fatal GDAL Error 6: Normalized/laundered field name: 'num_GEOID10s' to 'num_GEOID1'"	Be sure you are using the most recent version of the GAT tool. This is a bug in older versions. Shorten the name of your identification variable to five characters or less. Reduce the number of fields in your data.
This error appears in the R console: "Error in if (tobemerged[lowpop[1], aggvar]/minval >= tobemerged[lowpop2[1], : missing value where TRUE/FALSE needed"	Be sure you are using the most recent version of the GAT tool. This is a bug in older versions. Ensure that you name the rate variable; do not accept the default "gat_rate".

Problem	**Solution**
The program stops and this error appears in the R console: “Error in structure(.External(.C_dotTclObjv, objv), class = "tclObj") : [tcl] window ".2.5" was deleted before its visibility changed.”	Reduce the number of fields in your Shapefile
Program is very slow	Try reducing the number of geographical boundary nodes (by thinning or simplifying polygons: see technical note)
After merges begin, an error message appears: “Error in `row.names<-.data.frame`(`*tmp*`, value = c("36001000300", "36001000401", : duplicate 'row.names' are not allowed In addition: Warning message: non-unique values when setting 'row.names': ‘GATid1’, ‘GATid10’, ‘GATid101’, ‘GATid102’, ‘GATid103’, ‘GATid104’, ‘GATid108’, ‘GATid11’, ‘GATid110’, ‘GATid111’, ‘GATid114’, ‘GATid116’, ‘GATid12’, ‘GATid121’, ‘GATid123’, ‘GATid124’, ‘GATid126’, ‘GATid127’, ‘GATid128’, ‘GATid129’, ‘GATid13’, ‘GATid130’, ‘GATid133’, ‘GATid134’, ‘GATid135’, ‘GATid136’, ‘GATid138’, ‘GATid139’, ‘GATid140’, ‘GATid141’, ‘GATid143’, ‘GATid146’, ‘GATid147’, ‘GATid148’, ‘GATid150’, ‘GATid153’, ‘GATid154’, ‘GATid156’, ‘GATid157’, ‘GATid158’, ‘GATid16’, ‘GATid162’, ‘GATid163’, ‘GATid167’, ‘GATid17’, ‘GATid175’, ‘GATid179’, ‘GATid180’, ‘GATid181’, ‘GATid182’, ‘GATid183’, ‘GATid184’, ‘GATid185’, ‘GATid19’, ‘GATid20’, ‘GATid21’, ‘GATid28’, ‘GATid29’, ‘GATid3’, ‘GATid32’, ‘GATid34’, ‘GATid37’, ‘GATid39’, ‘GATid40’, ‘GATid41’, ‘GATid42’, ‘GATid43’, ‘GATid44’, ‘GATid45’, ‘GATid46’, ‘GATid47’, ‘GATid48’, ‘GATid49’, ‘GATid5’, ‘GATid50’, ‘GATid51’, ‘GATid52’, ‘GATid53’, ‘GATid55’, ‘GATid57’, ‘GATid58’, ‘GATid59’, ‘GATid61’, [... truncated] “	If you are aggregating data multiple times, change the ID variables created by the GAT tool from the first aggregation so they do not begin with the prefix ‘GATid’. For example, change ‘GATid1’ to ‘NewID1’, ‘GATid10’ to ‘NewID10’, etc.

Technical Notes:

How the program merges

The overall goal of the merging is to create a relatively large number of compact regions which meet the specified criteria. To accomplish this, the areas in the input shapefile are aggregated pairwise until all regions have the minimum values specified. First, all the areas which need to be merged are sorted from the highest values of the aggregation variable(s) to the lowest. If there is only one aggregation variable, the area with the highest value is merged first. If there is more than one aggregation variable, the aggregation variable is divided by the minimum value specified for it, and the area with the highest proportion is selected to be aggregated first. For example, suppose we wish to merge areas until they have a population of 1000 and at least 100 births. An area with 90 births and 200 population will be merged ahead of an area with 50 births and 400 population, because the maximum of 90/100=0.9 and 200/1000=0.2 is 0.9, but the maximum of 50/100=0.5 and 400/100=0.4 is 0.5, which is less.

Next, a list of the neighbors of the selected area is obtained. Areas are considered neighbors if they share at least two points with the selected area. If you specified a boundary variable, and neighbors exist within a boundary, they are the first candidates for merging. If there are no neighbors within the boundary, the closest area within the boundary will be merged. Only after all areas within a boundary are merged will merging occur across boundaries. If you did not specify a boundary variable or if you specified a boundary variable and areas do not exist within the boundary, then all neighbors are considered as candidates for merging. If the “closest” option was selected, then the closest area is merged with the previously selected area. Closeness is calculated as the distance between the areas' centroids. If the “least” option is selected, the area with the lowest proportion of aggregation variable to minimum value specified is selected. If the “most similar” option is selected, then the area with the least absolute difference between the ratios of the variables chosen for similarity comparison is merged with the selected area. For example, suppose the variables chosen for similarity comparison are counts of persons under poverty and total population. If the area to be merged has a percent under poverty of 10%, and it has neighbors have 9%, 13%, and 15% under poverty, then then it would be merged with the area with 9% under poverty. If an area has no neighbors, then it is merged with the closest area to form a new region.

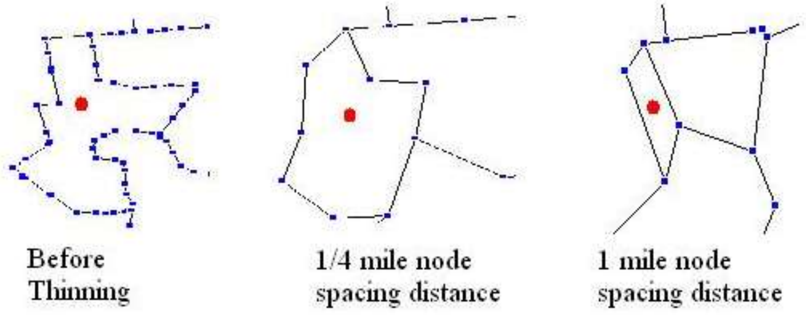
The new region then becomes a candidate for additional merging, and merging continues until all regions contain the minimum values of the aggregation variables specified.

Citation for the Geographic Aggregation Tool

Thomas O. Talbot and Gwen D. LaSelva. *Geographic Aggregation Tool, Version 1.31*, New York State Health Department, Troy NY, July 2010

Thinning Geographic Boundaries

The program may merge very slowly if the geographic boundaries are unnecessarily complex. Boundaries can be simplified by a process known as “thinning” (or “simplification” or “integrating” or “generalizing”). Thinning is removing nodes based on how far apart they are and/or their collinearity to decrease the complexity of the maps (see figure below). GIS systems like MapInfo and ArcGIS can be used to thin maps. After thinning, areas should merge much more quickly in the GAT, but the map may appear crude. If high resolution maps are needed for display, the thinned map used for processing data can be linked back to the high resolution boundaries using unique area identifiers.



Additional notes:

For more information on how the GAT was used to calculate life expectancy for small areas see:

Thomas O. Talbot, Douglas H. Done and Gwen D. Babcock.
Calculating census tract-based life expectancy in New York state: a generalizable approach.
Population Health Metrics 2018 16:1

Description of the simulated low birth weight data set for New York State

Birth outcome fields:

We used the distribution of New York State birth data at the ZIP code level for years 2001-2005 obtained from the NYS Department of Health and the NYC Department of Health and Mental Hygiene as the basis for a simulation. Simulated data was developed to protect the confidentiality of persons represented by the data. We simulated the number of total singleton births and number of low birth weight singleton births using a binomial distribution for each ZIP Code.

Socio-demographic fields:

Education and poverty data were added to block points based on 2000 US Census data for block groups. Proportions of high school graduates and children <5 living under poverty from block groups were multiplied by the population of children <5 assigned to each block point. Race and ethnicity data were assigned to each block point from the US Census 2000 SF1 files. We spatially joined the simulated ZIP code LBW data with census block points to add SF1 data fields for race and ethnicity. Block points were assigned ZIP codes based on whether or not the point was contained within the ZIP code polygon. We used census 2000 population data for children less than 5 years of age to provide socio-demographic information for each of the ZIP Code areas.

The ZIP code level dataset contains 20 Fields:

Field	*Description*
ZIP06	2006 ZIP Codes
county	County containing ZIP
Latitude	Latitude for population-weighted ZIP Centroid
Longitude	Longitude for population-weighted ZIP Centroid
simbir0105	Simulated N total births for ZIP 2001-2005
simlbw0105	Simulated N low birth weight babies born for ZIP 2001-2005
W_00_04[1]	Single race white population <5 years in ZIP
B_00_04[1]	Single race black population <5 years in ZIP
AI_00_04[1]	Single race American Indian population <5 years in ZIP
A_00_04[1]	Single race Asian population <5 years in ZIP
PI_00_04[1]	Single race Pacific Islander population <5 years in ZIP
O_00_04[1]	Single race Other population <5 years in ZIP
M_00_04[1]	Multiple race population <5 years in ZIP
H_00_04[1]	Hispanic ethnicity population <5 years in ZIP
NH_00_04[1]	Non-Hispanic ethnicity population <5 years in ZIP
total0_4[1]	Total population <5 years in ZIP
Upov0_4[1]	number of persons <5 years living under poverty level
Opov0_4[1]	number of persons <5 years living at or above poverty level
lths[1]	number of persons >25 years with less than high school education
hsg[1]	number of persons >25 years with high school or more education

[1] *Data Sources: 2000 Census SF1 File, 2000 Census SF3 File*