Mapping Suicide Death Rates: Geographic Aggregation Tools and Spatial Smoothing with Hierarchical Bayesian Models

Lauren M. Rossen & Diba Khan

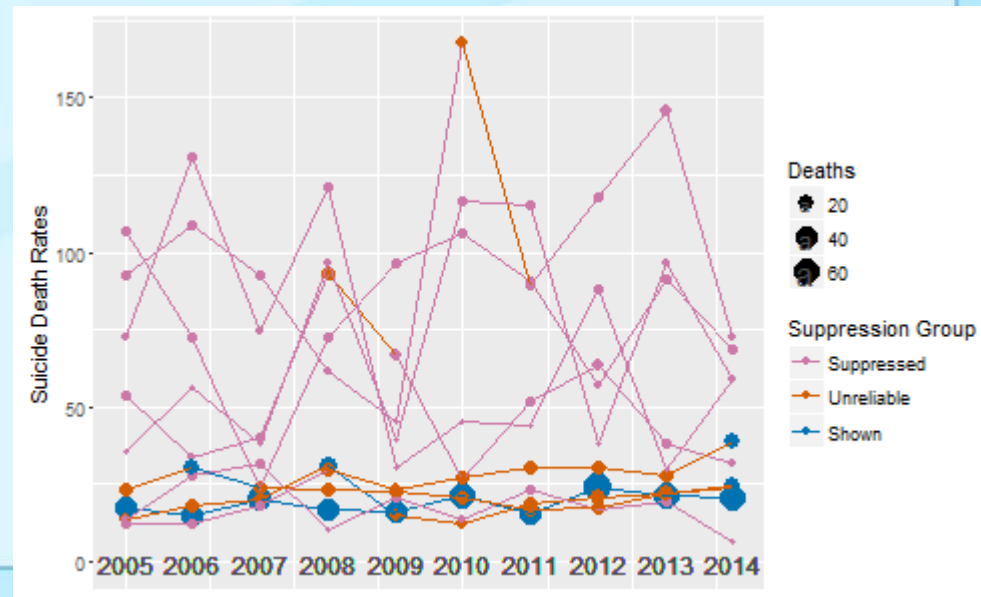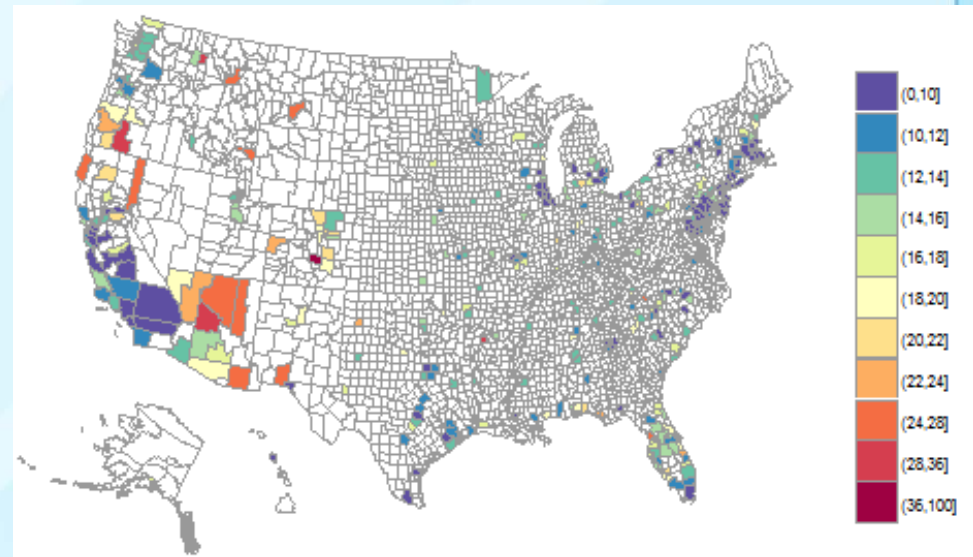National Center for Health Statistics

Geospatial Web Applications, Tools, and Data Workshop

November 18, 2016

# Objective: Map County-Level Suicide Death Rates

Two big problems:

- *Rare events* – **84% of counties had fewer than 20 suicides in 2014**
  - Largely blank map → → →

- **Rates are *unstable* in sparsely populated areas**
  - Suicide rates worse in rural areas? → → → → → → → →

Approaches

- **Option 1:**

  - Combine several years of data

    - 5 year aggregation (2010-2014):

      - 48% counties < 20 suicides

    - Combining more years may mask important temporal trends

Suicide Death Rates Increasing
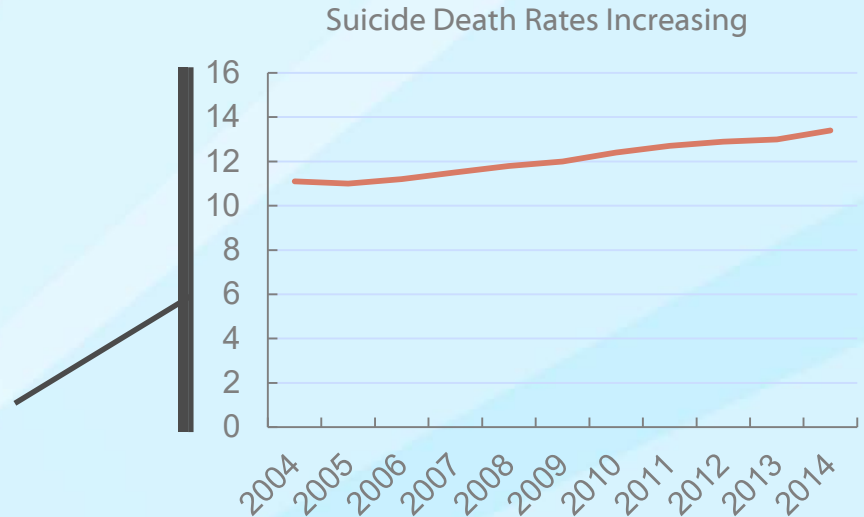
- **Option 2:**

  - Combine adjoining counties – Geographic Aggregation Tool

- **Option 3:**

  - Spatial smoothing with Hierarchical Bayesian Models

NOTE: Many, many other options not covered here.

Option 2: Geographic Aggregation Tool (GAT)

- **R & SAS code/macros available to combine adjacent counties**
  - Input shapefile with counts/data
    - Orders counties/units by # of events, combines adjacent counties until minimum threshold (e.g., 20 events) is reached
   - Some other options available
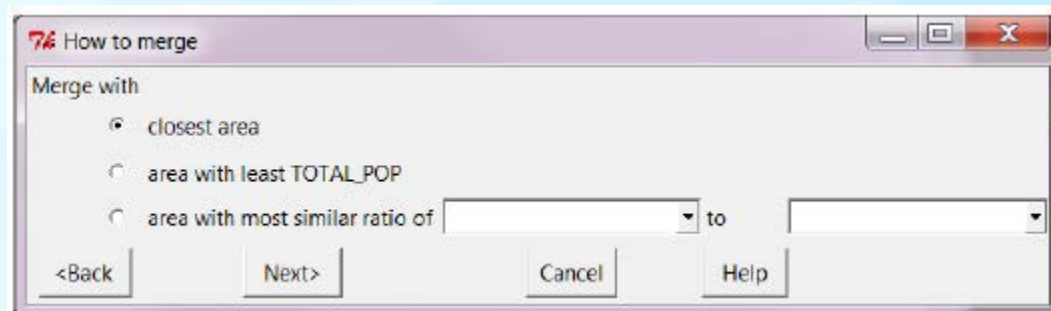
  - Output = set of new shapefiles with aggregated boundaries

More details on GAT available from: http://www.albany.edu/faculty/ttalbot/GAT/GAT_vR13_guide.pdf

Geographic Aggregation Tool (GAT)

- Started with **3140** counties → → → → →

- Aggregate until numerator (suicide deaths) and denominator (population) both >20
  - Choose closest area

- GAT produced a shapefile with **696** areas→ → → → → →
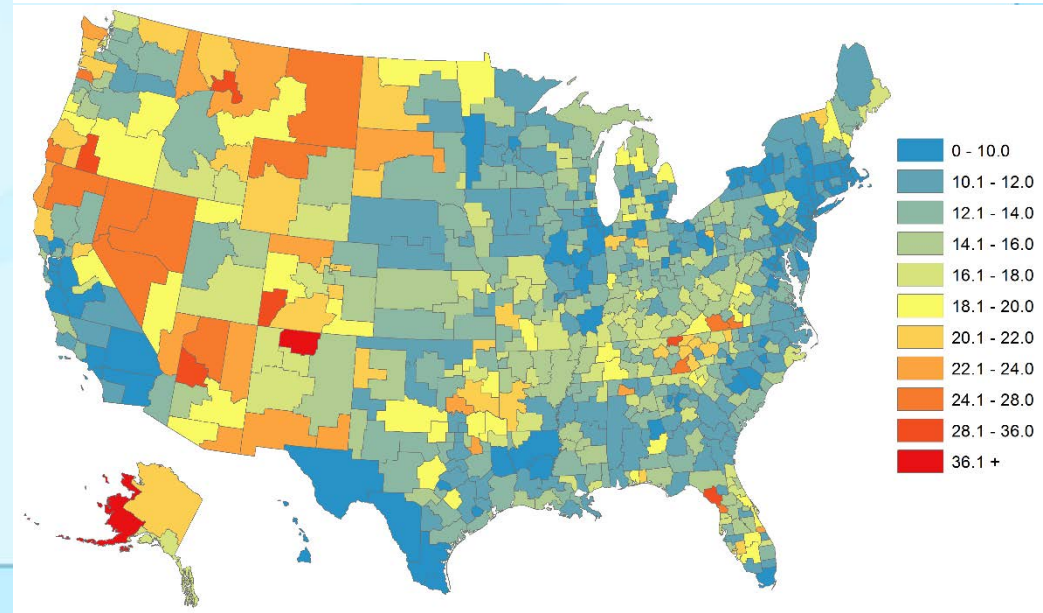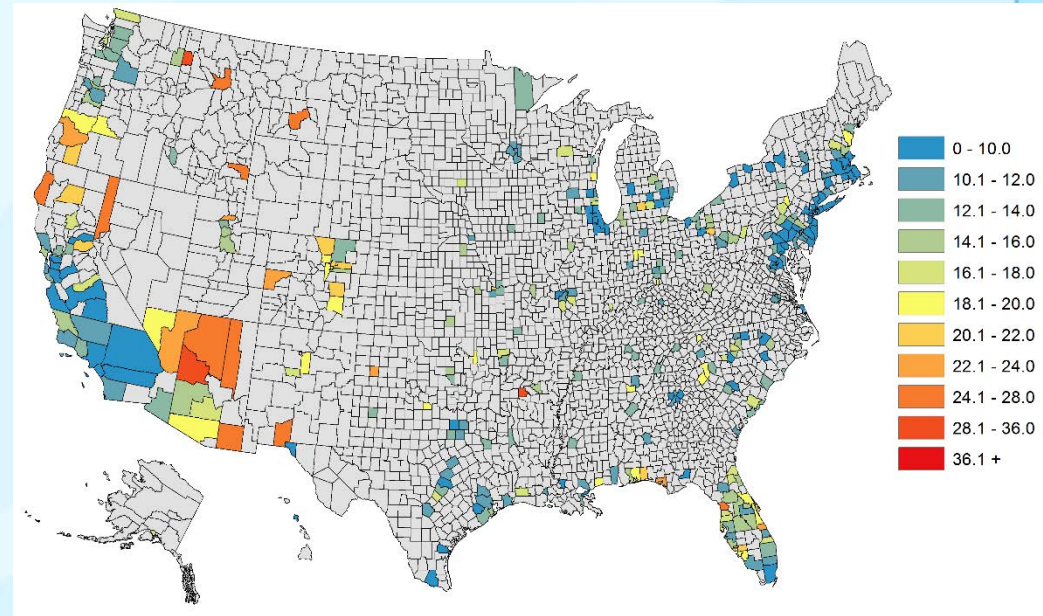
Option 3: Spatial Smoothing – Hierarchical Bayesian Models with R-INLA

- **INLA = "Integrated Nested Laplace Approximations"**

- **R package to run Hierarchical Bayesian models**

  - Many different types of models available

  - Can include spatially structured random effects, as well as space-time interaction terms for spatiotemporal smoothing

  - Alternative to BUGS (OpenBUGS, GeoBUGS), but much, much faster

More details on R-INLA available from: http://www.r-inla.org/

See Rue, Martino & Chopin (2009)

Option 3: Spatial Smoothing – Hierarchical Bayesian Models with R-INLA
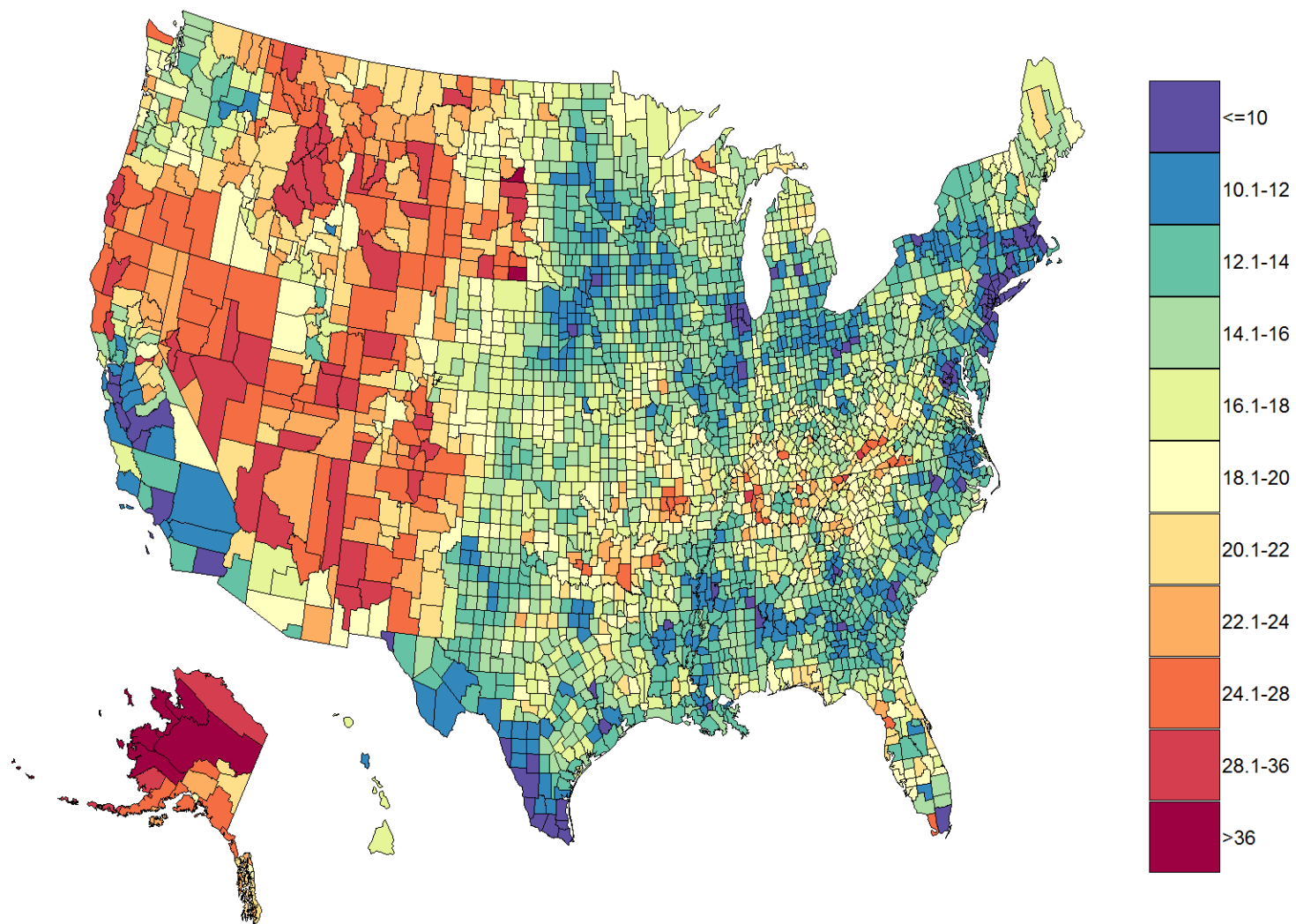
- **Suicide death rates (2005-2014) modeled as a function of:**

  - County-level non-spatial random effect (iid)

  - County-level spatially structured random effect

  - Year random effect (type 1 random walk)

  - Space-time interaction term (residual, iid)

```
>formula7<-numerator~1+f(countyid,model="iid")+
                    f(countyid2,model="besag",graph="suicides_map")+
                    f(year,model="rw1")+
                    f(resid,model="iid")

>result7<-inla(formula7,family="Binomial",Ntrials=denominator,data=data,
         control.compute=list(dic=TRUE,cpo=TRUE))
```
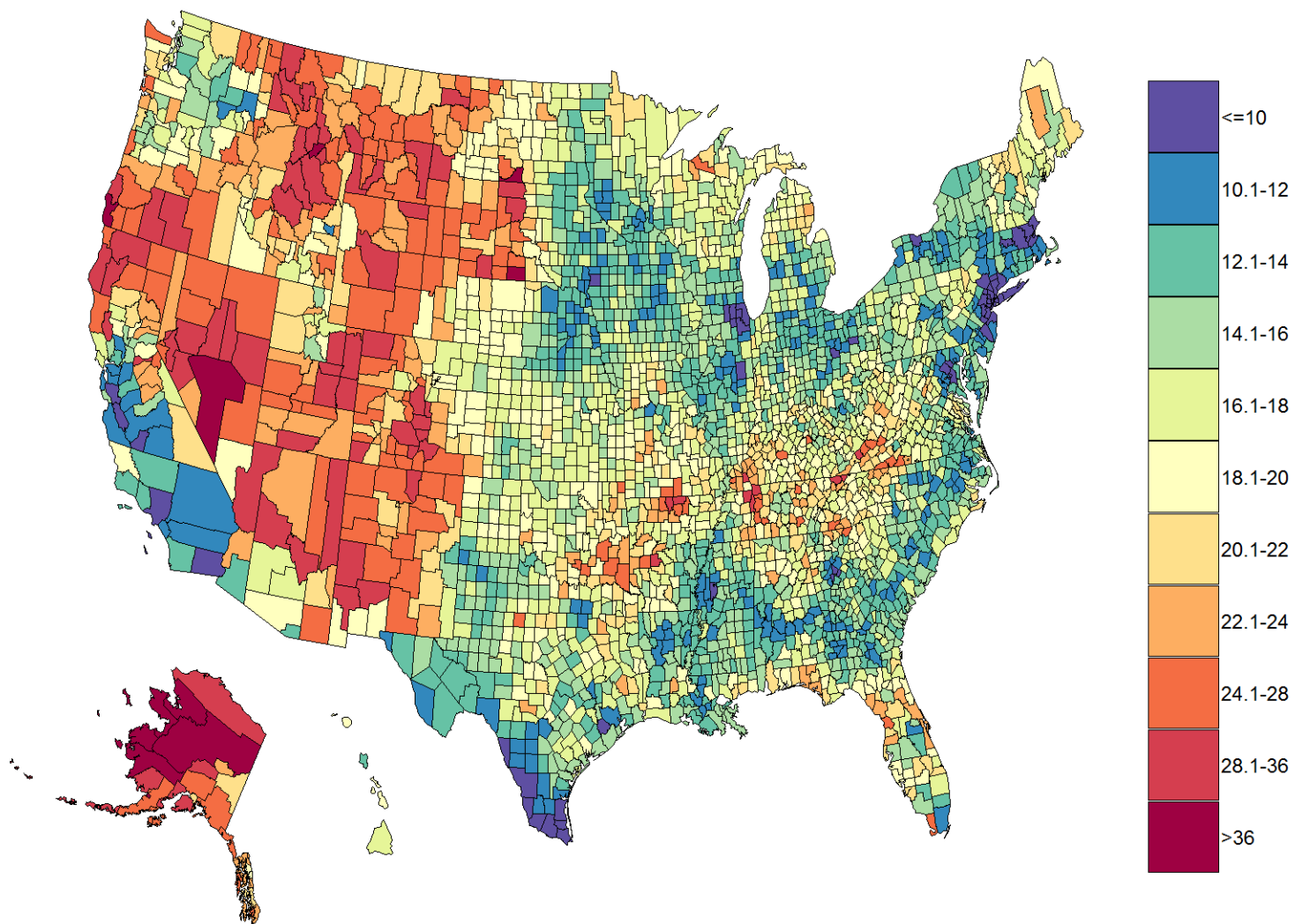
Results from INLA: Suicide Rates, 2005

Legend:
- <=10
- 10.1-12
- 12.1-14
- 14.1-16
- 16.1-18
- 18.1-20
- 20.1-22
- 22.1-24
- 24.1-28
- 28.1-36
- >36
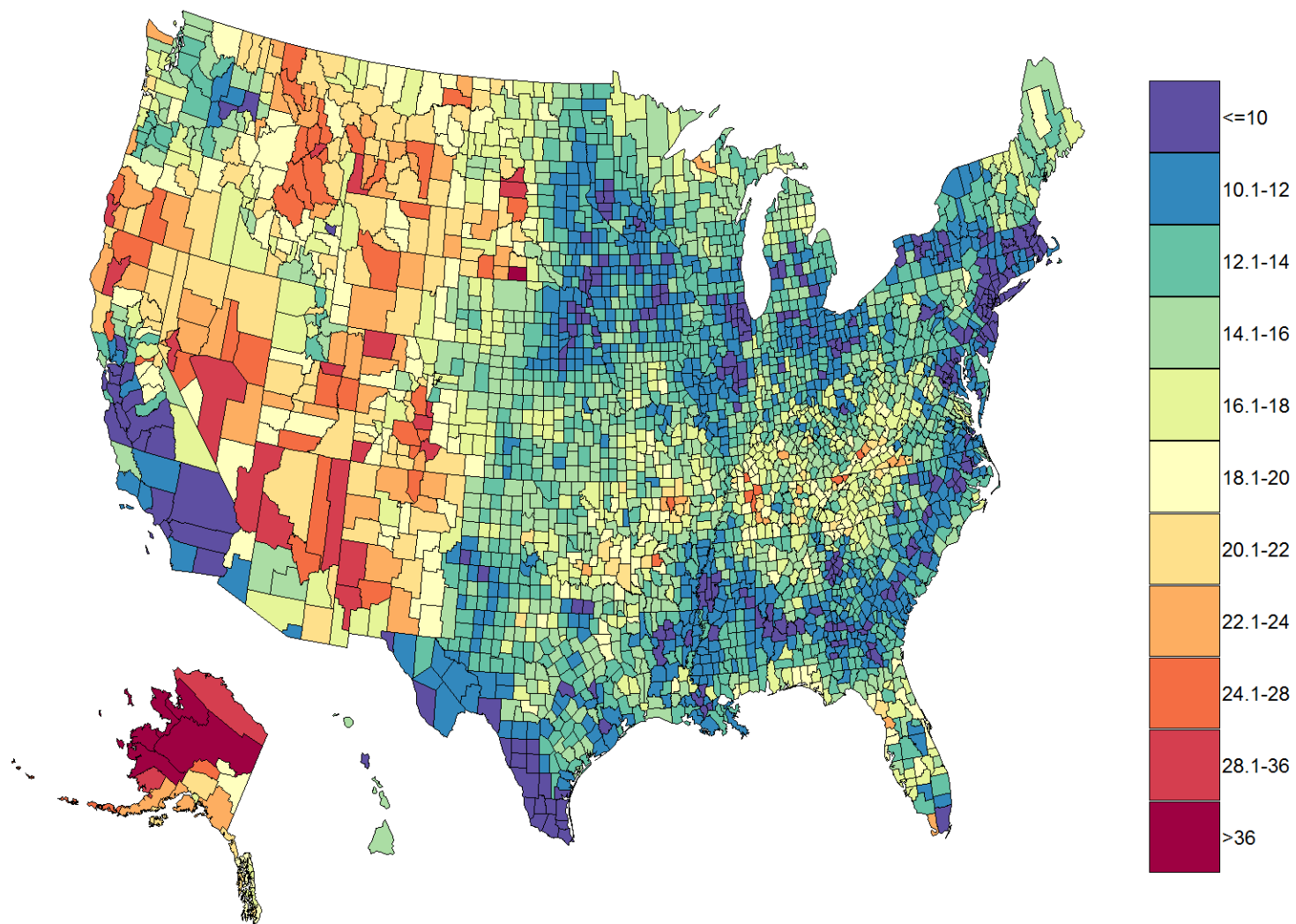
Results from INLA: Suicide Rates, 2010

Results from INLA: Suicide Rates, 2014

Geographic Aggregation Tool (GAT)

PROS

- **Easy to use, very fast (e.g., 10 minutes)**

- **Produces direct estimates (no model)**

- **Works in R and SAS**

- **Can specify how areas should be aggregated**

CONS

- **Different outcomes or years might produce different aggregation schemes**

- **Might 'over-smooth' very rural areas with high rates**

- **Interface and output somewhat clunky**

Hierarchical Bayesian Models with R-INLA

PROS

- **Very flexible, can handle large data sets**

- **Much faster than MCMC/BUGS**

- **Produces estimates for all (small) geographies, can implement geographically weighted regression and model point processes**

- **Can get estimates of uncertainty, include covariates**

CONS

- **Not as easy to implement as GAT, though easier than MCMC/BUGS**

- **Good luck explaining to a lay audience**

- **Can take a few hours/days to run compared to GAT**

- **Not as flexible as BUGS with respect to missing data**

Questions?

Contact Information:

Lauren Rossen

LRossen@cdc.gov

Diba Khan

ild1@cdc.gov

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the National Center for Health Statistics or the Centers for Disease Control and Prevention.

National Center for Health Statistics
Division of Vital Statistics

EXTRA SLIDES AND RESOURCES

Screen shots and additional tools to explore

OTHER TOOLS

- GeoDa
  - http://geodacenter.github.io/index.html
  - Free, open-source spatial data analysis program

- SaTScan
  - http://www.satscan.org/
  - Free program for space-time scan statistics, space-time cluster detection, etc.

- WinBUGS/OpenBUGS/GeoBUGS
  - http://www.mrc-bsu.cam.ac.uk/software/bugs/thebugs-project-geobugs/
  - Free program(s) to fit spatial and spatio-temporal models using MCMC
  - Can integrate with R through R2OpenBUGS package

OTHER TOOLS

- Many other R packages for spatial data analysis:
  - Bivand, Pebesma & Gomez-Rubio. Applied Spatial Data Analysis with R. Available online:
    - http://gis.humboldt.edu/OLM/r/Spatial%20Analysis%20With%20R.pdf

  - Another good overview of select R spatial packages:
    - http://www.spatialanalysisonline.com/HTML/index.html?r-project_spatial_statistics_s.htm

RUNNING GAT

1. Save the 'GATinRv13.R' file
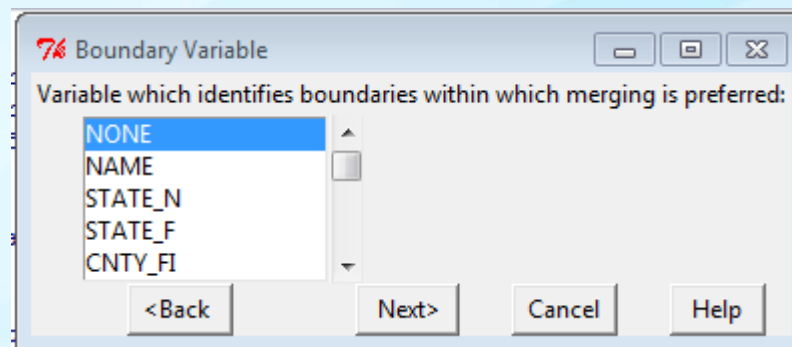2. Run their batch file, after pasting in the proper path names:
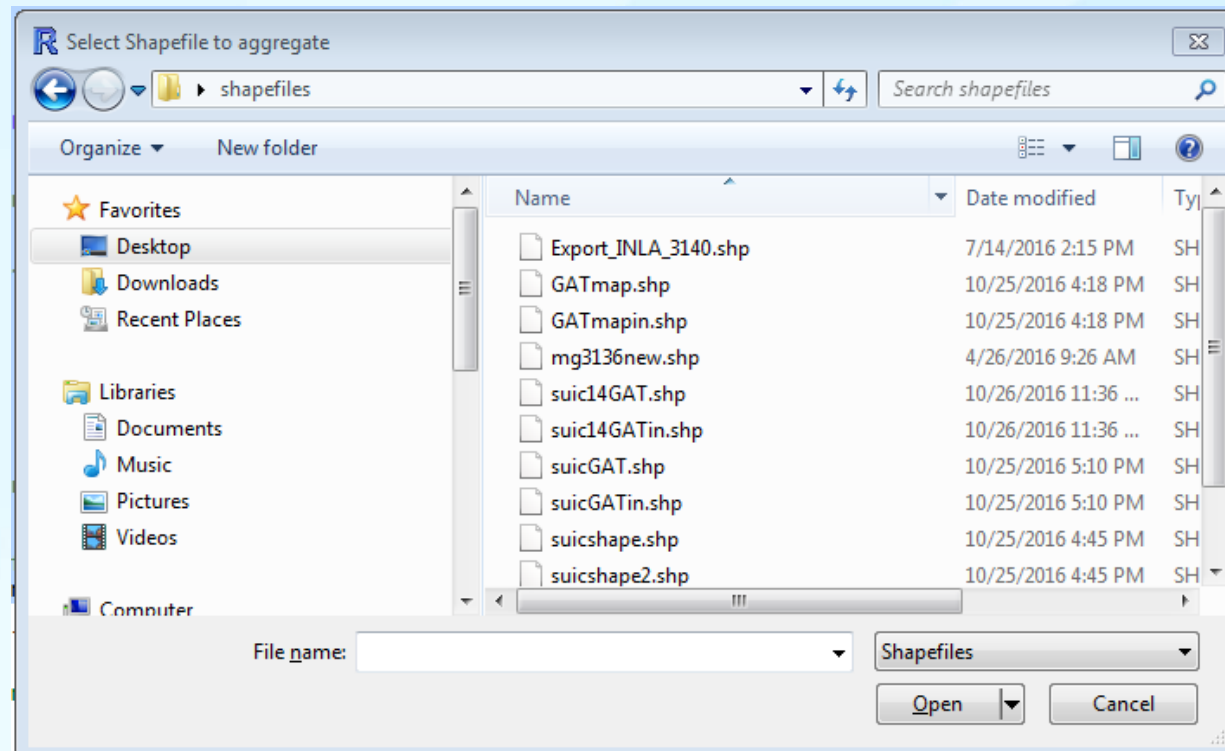
3. Select your shapefile

RUNNING GAT

4. Select your geography ID variable (e.g., FIPS)
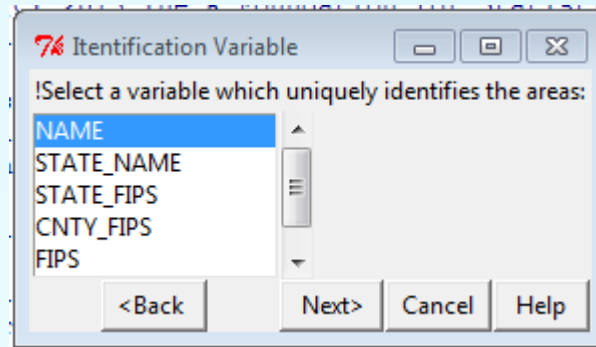
5. Select boundaries within which merging is preferred (e.g., only merge counties from the same state)
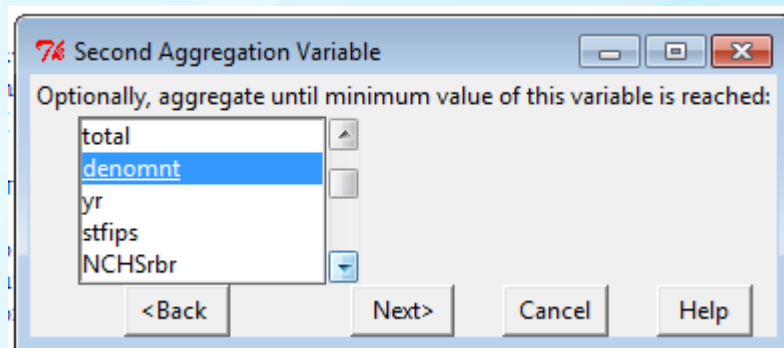
RUNNING GAT

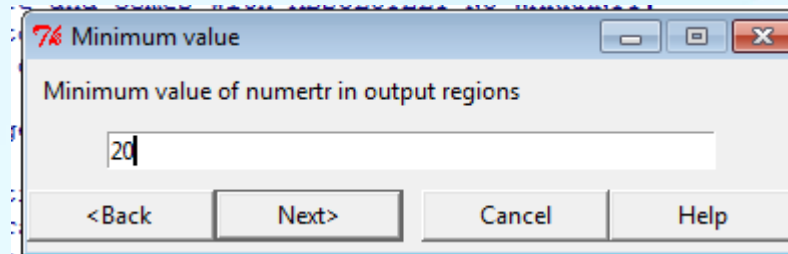6. Select the variable on which you want to aggregate (e.g., numerator)
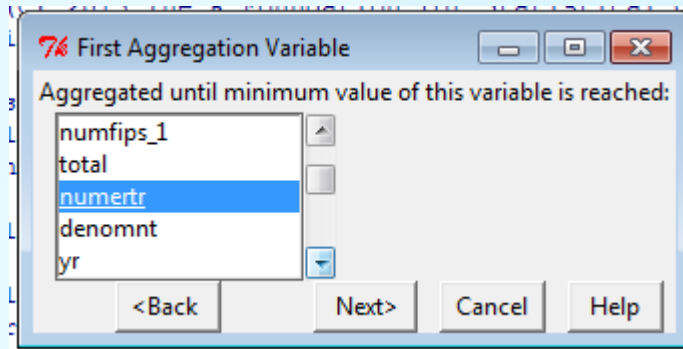
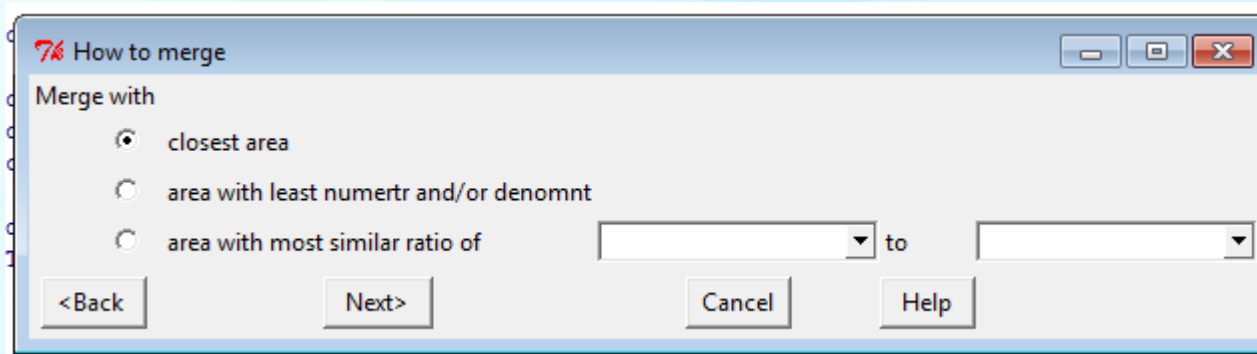7. Optionally, aggregate on a second variable (e.g., denominator)

RUNNING GAT
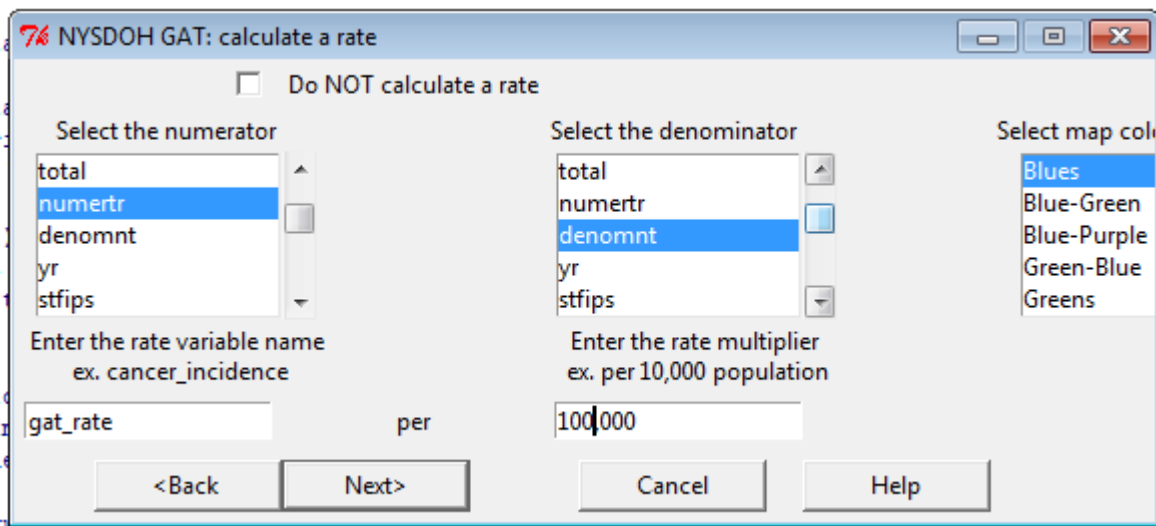
6. Specify minimum count/threshold

7. Choose priority for merging areas – proximity, smallest # of events, or areas with the 'most similar ratio of variable1 to variable 2' if you have certain characteristics like poverty rates, etc.

RUNNING GAT

8. Specify whether the tool should calculate the rate

9. Proceed through the confirmation, and GAT will begin to aggregate (it takes a few moments), displaying a progress bar
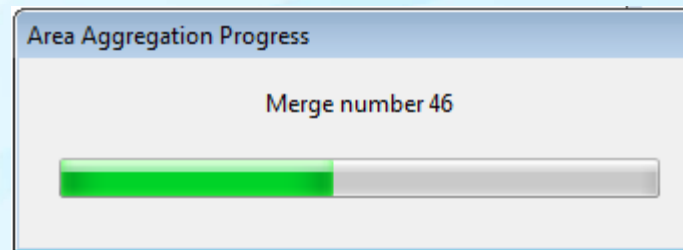
RUNNING GAT

10. Some maps will pop up of the variables being aggregated to show the distribution before aggregating
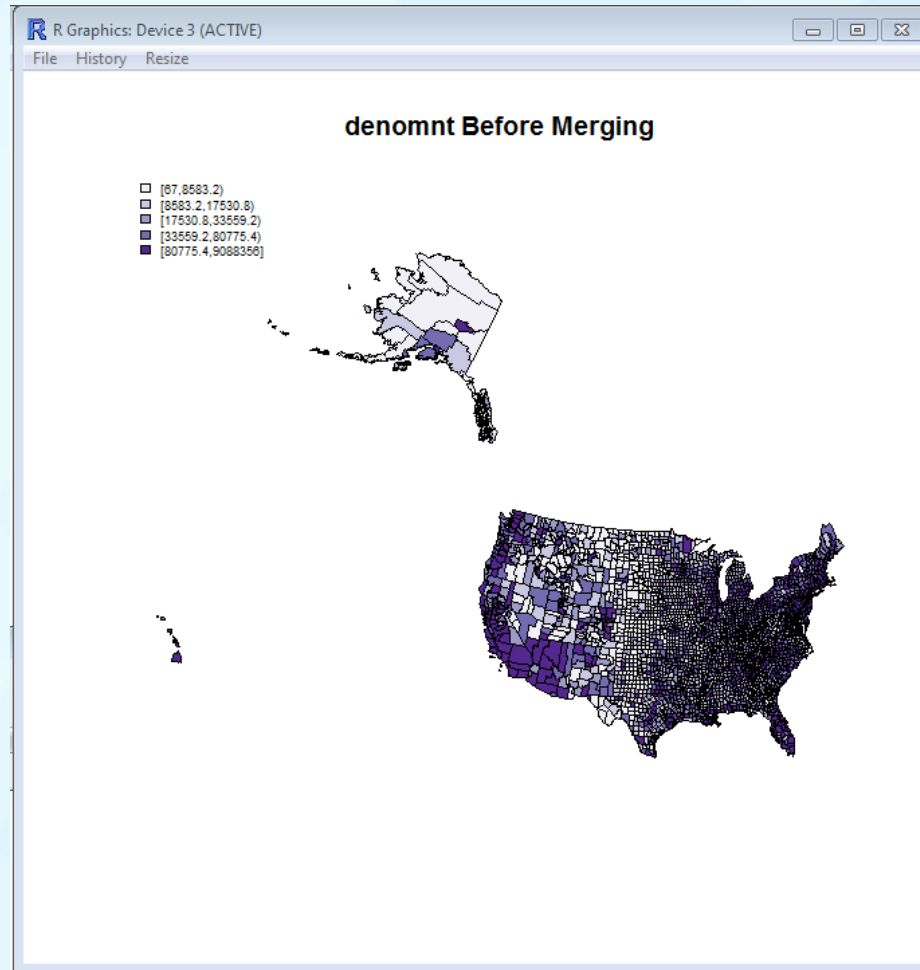
RUNNING GAT

11. When the aggregation completes, a prompt will ask where to save the output shapefiles, and then the tool will try to automatically quite R

12. But only after displaying some maps:

Map of original areas (red) and aggregations (black)

RUNNING GAT

Errors and issues:

- The program is pretty picky about things (e.g., the shapefile must be on a letter drive, not a network drive)

- The .dbf component of the shapefiles have to have the extension written out as .dbf or the program will say your shapefile doesn't have any numeric data

- The program will fail if you use the most recent version of R, so they provide an older version that will run using the batch file they provide

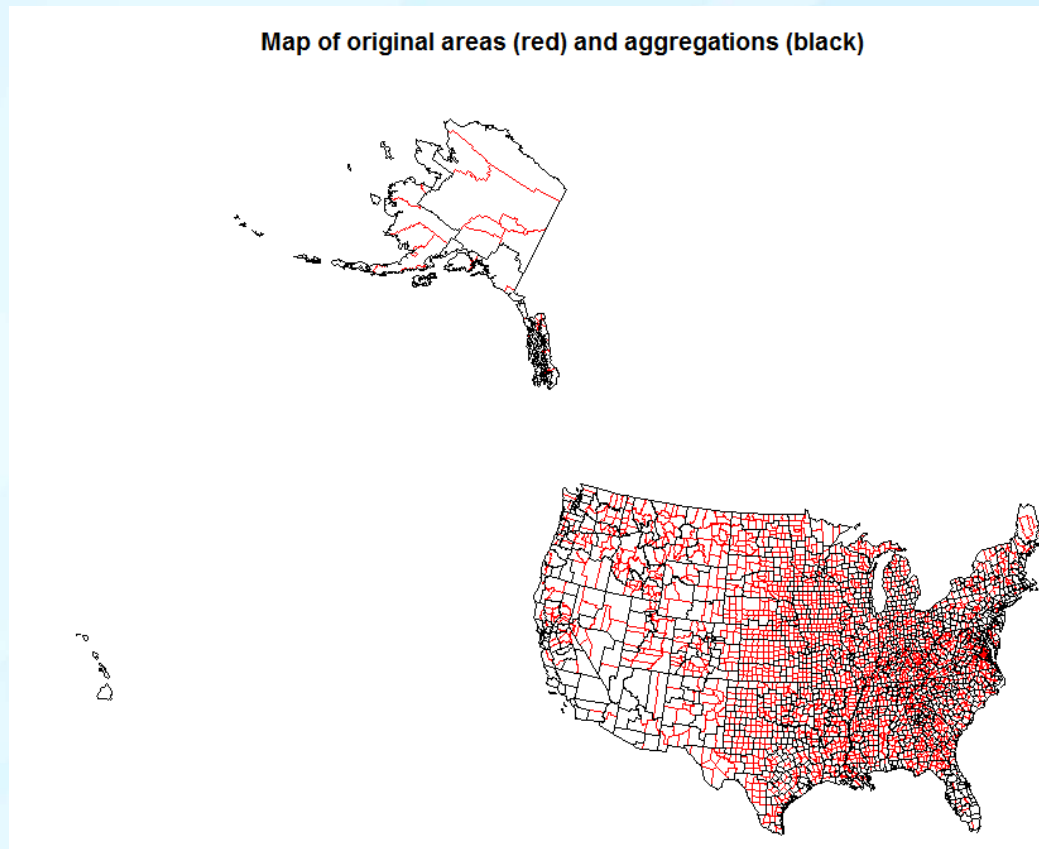- Sometimes there are no neighbors to merge within a given region/state, so check results carefully

```
[1] "No physically adjacent neighbors found in within same boundary"
[1] "Found not physically adjacent but in same boundary"
[1] "merge number 49"
[1] "merge number 50"
[1] "merge number 51"
[1] "merge number 52"
[1] "merge number 53"
[1] "merge number 54"
[1] "merge number 55"
[1] "merge number 56"
[1] "merge number 57"
[1] "merge number 58"
[1] "merge number 59"
```

RUNNING INLA

```r
#read in shapefile and data
county.map = readShapePoly('//path to shapefile here/shapefilename.shp',IDvar="NUMFIPS")

suic<-read.csv("//path to data here/datafilename.csv",header=TRUE)

#create spatial data frame
polys<-SpatialPolygonsDataFrame(county.map,data=as.data.frame(county.map),match.ID=TRUE)

#obtain lat long coordinates
coords<-coordinates(polys)
polys$x<-coords[,1]
polys$y<-coords[,2]

#create adjacency matrix, neighbors list - here using Delaunay Triangulation
triang<-tri2nb(coords, row.names=NULL)

neib<-nb2WB(triang)

#calculate sum of number of neighbors
neib$sumnb<-sum(neib$num)

#how many neighbors for each county?
summary(neib$num)

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      3.00    5.00    6.00    5.99    7.00   12.00
```

RUNNING INLA

```r
#set seed if you want to replicate results
set.seed(1234)

#create a file with the required info about what counties/units are neighbors
inla.geobugs2inla(neib$adj, neib$num, graph.file="suicides_map")

#create the model - here a binomial model for suicide deaths/population, including a
# random effect for county (iid), a spatially structured county-level
#random effect (besag), a random effect for time (type 1 random walk),
# and a county-year specific iid residual term
countyid<-rep(1:3140,each=10)            #number of counties
countyid2<- countyid                     #number of counties for second random effect
resid<-rep(1:31400)                      #number of county-year observations
year<-rep(1:10,len=31400)                #year variable
numerator<-suic$numerator
denominator<-suic$denominator

data<-data.frame(numerator, denominator, countyid, countyid2, resid, year)

formula7<-numerator~1+f(countyid,model="iid")+
             f(countyid2,model="besag",graph="suicides_map")+f(year,model="rw1")+f(resid,model="iid")

result7<-inla(formula7,family="Binomial",Ntrials=denominator,data=data, control.compute=list(dic=TRUE,cpo=TRUE))

#get fit statistics
result7$dic$dic;result7$dic$p.eff
## [1] 134594.2
## [1] 2641.959
```

posterior predictions (in result7 above) can then be saved/exported for mapping/analysis