Exploring the effects of spatial aggregation

Amelia McNamara*[1] and Aran Lunzer†[2]

[1] *Smith College*
[2] *Viewpoints Research Institute*

May 2016

1 Introduction

In this note, we present a tool that allows a user to manipulate the size and orientation of spatial aggregation units, in order to explore the effect of those parameters on the visual pattern presented. The tool is available at `https://tinlizzie.org/spatial/`.

2 Background

Aggregated data is notoriously hard for novices to grasp (Konold et al., 2014; Hancock et al., 1992). However, it is used in many situations, both symbolic (the mean) and visual (the histogram).

When we began showing our LivelyR (Lunzer et al., 2014) prototypes to statisticians and teachers, we were met with excitement about the ability to simply manipulate the bin width and bin offset of a histogram. Also popular was the 'histogram cloud' feature that superimposed a set of slightly different histograms of the same data (McNamara, 2015). Both interactions give a user the ability to easily manipulate parameters that they often would leave as defaults, and compare the results of different parameter choices.

Given the popularity of our work about the impact of parameter choices on 1-dimensional distributions, we began thinking about a 2-dimensional analogue. The most natural case was mapmaking and spatial data. One common way to depict spatial data is a choropleth map, which divide area into polygons, colored by the value of a parameter measured or estimated in that area. Figure 1 is a modern choropleth map, showing income disparities from average by county, over the United States (Aisch et al., 2015).

*Some work completed while the author was partially funded by the Communications Design Group, SAP Labs. Contact: amcnamara@smith.edu

†Contact: aran@acm.org

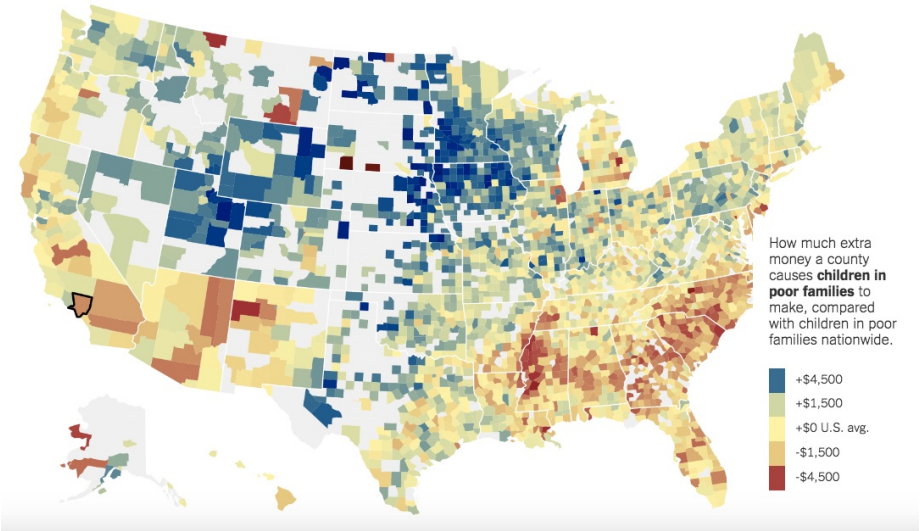

Figure 1: Map of income disparities from average per county, from Aisch et al. (2015).

Many choropleth maps use polygons in the shape of political boundaries, like states, counties, and zip codes. Others use more generic polygons, like the example from 1895 shown in Figure 2 and the more modern example in Figure 3.

Much like the case of histograms, the visual pattern in a choropleth map is imbued with meaning by readers who forget the arbitrariness of the chosen aggregation units.

Geographers call this phenomenon the Modifiable Areal Unit Problem (MAUP) (Ervin, 2015). Essentially, map-making and statistical analysis is highly sensitive to the area of aggregation (e.g. Census tracts, zip codes, neighborhoods) that is used. For data that comes as points, this problem can be skirted by using the raw point data to fit models, but much spatial data is distributed in already-aggregated forms. Again, the US Census is a prime example. Because of anonymity issues, the Census Bureau can only provide data at the level of Census blocks.

In maps where the spatial aggregation units vary in size and shape (for example, the US States), large, rural areas tend to have fewer incidences of things like homicide and cancer, which can make them appear to be "better" in some way than their small, densely-populated counterparts. Trying to correct for this flaw by using proportions based on population size or area adds another problem—the areas with smaller populations have higher variance, so they will appear to have higher or lower rates than average. Because of this, some statisticians argue that "all maps of parameter estimates are misleading" (Gelman and Price, 1999).

This work does not try to address the underlying statistical issue of generat-

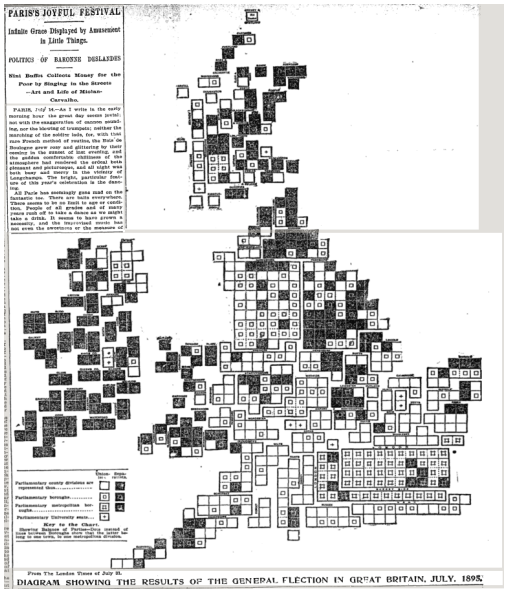

Figure 2: Geometric choropleth map from 1895, from Felton (2015).

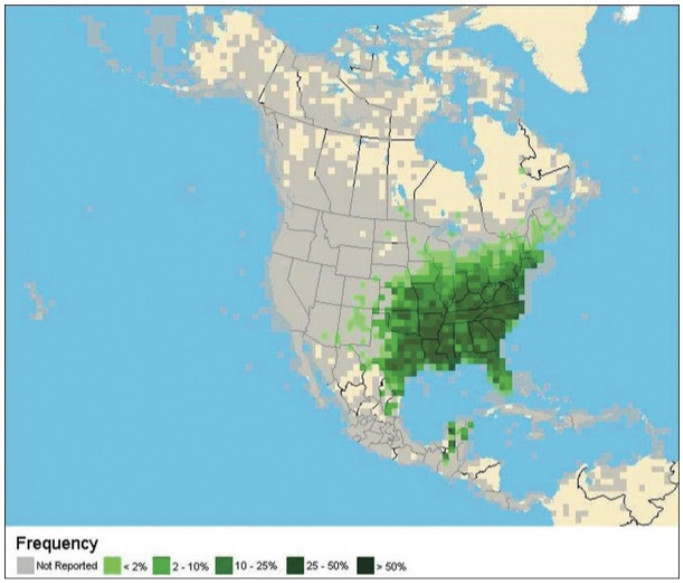

Figure 3: A geometric choropleth map from 2008, showing the incidence of Carolina Wrens, binned to 100-km blocks, from Sullivan et al. (2008).

ing more robust estimates from pre-aggregated data, but rather seeks to make visible the possible effects of slight variations in aggregation.

There are two varieties of MAUP that are commonly discussed: scale MAUP and zone MAUP. The scale problem is related to the scale of the areal unit you have chosen to use. Is it large (like a state) or small (like a Census block)? The zone problem has more to do with the shape of the units. Similarly sized, but differently shaped units (imagine squares versus hexagons) can provide very different patterns.

Notice that I am saying "imagine" and "like." This is because there are surprisingly few demos that actually show this phenomenon in action. Sometimes, information about MAUP will show two or three discrete possibilities of aggregations of the same data, but usually only with toy data. The figure that is used in most explanations is shown in Figure 4 (Penn State GEOG 486, 2014; Ervin, 2015).

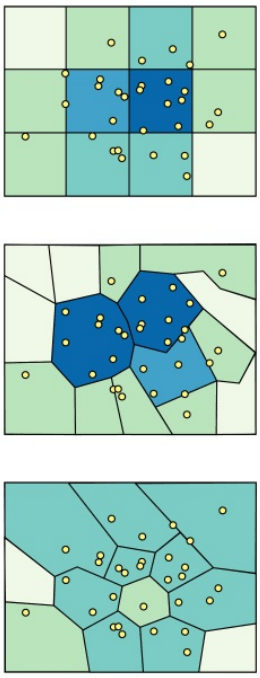

Figure 4: Typical figure used to illustrate the MAUP, taken from Ervin (2015).

The most publicized examples of the MAUP tend to be related to political gerrymandering of electoral boundaries. Depending on how districts are drawn, election results can be swung wildly in favor of one candidate or another. However, even in the case of gerrymandering, theoretical language like "Suppose there's a state that [...]" is often used, rather than concrete interac-

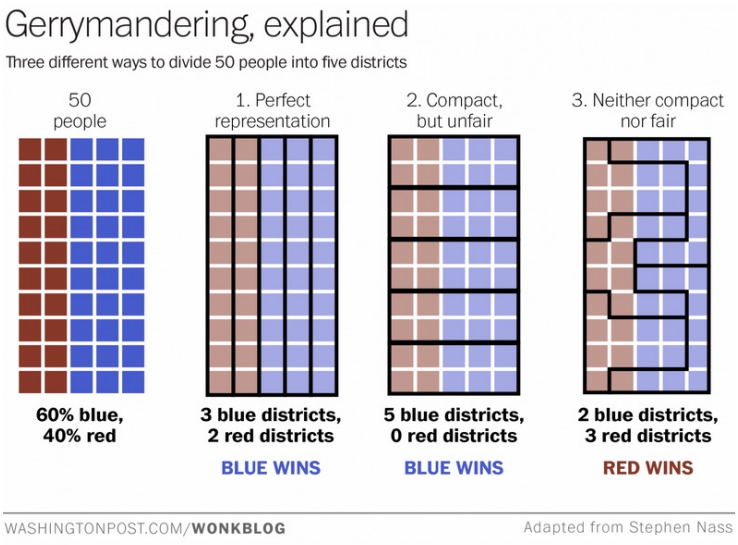

Figure 5: "The best explanation of gerrymandering you will ever see," from Ingraham (2015).

tives (Cohn, 2015). Even in pieces that use visuals, toy examples are king, as seen in Figure 5 (Ingraham, 2015).

Gerrymandering is a more complex version of the MAUP, because there is more at stake than simply the location of voters. Even people in the same political party can have very different ideas about how districts should be drawn—should they contain many like-minded people, or those with similar demographics? Or should they seek to be diverse in every sense of the word? Should they follow geographic boundaries or strive for geometric compactness?

Again, we are not trying to solve this problem. Rather, we aim to present a tool that allows users to see the effects of various spatial aggregation levels.

3 The tool

The tool, as it stands, is an HTML page using javascript (including d3.js and leaflet.js) to provide interaction. An OpenStreetMaps map is used as the base, and spatial point data is layered on top. Then, regular polygon aggregation units are provided, and colored according to the number of points they contain. A screenshot is shown in Figure 6.

It supports either squares or hexagons as the polygon aggregation units, and provides 8 possible sizes, which can be moved between smoothly.

The polygons can be scaled, moved, and rotated. The base map can also be zoomed and moved, although the interface does not support base map rotation.

Interaction controls, as they stand now:

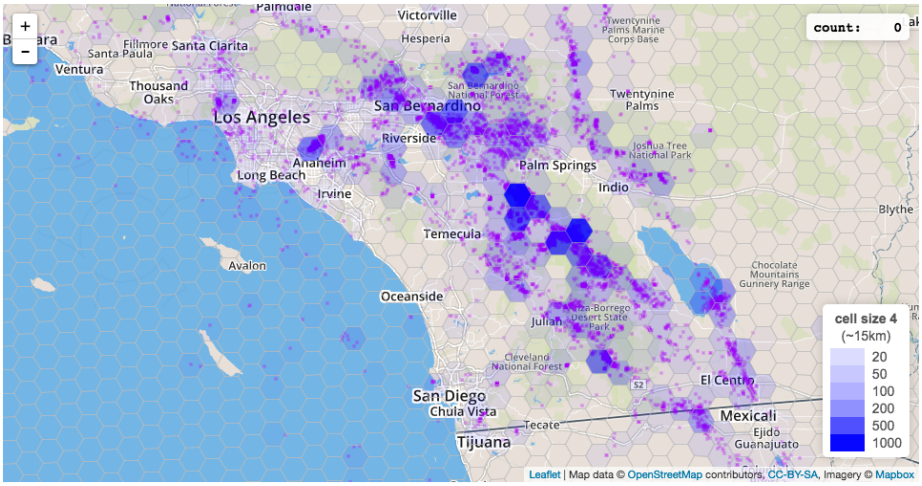

Figure 6: California earthquakes from 2014, binned using hexagons.

- **s** and **h** toggle the **s**quares and **h**exagons, respectively
- The **space bar** turns the data on and off, and fades out the base map
- **drag** moves the map, and the polygons along with it
- **Shift** + **drag** moves the polygons without moving the map
- **scroll wheel** or zoom buttons zoom the map, maintaining the size of polygon binning unless a threshold is breached. (The system only supports a certain number of polygons on a single screen, so if a user is zoomed in with small polygons and then zooms out, the interface will eventually change the relative size of the polygons)
- **Shift** + **scroll wheel** zooms the size of the polygons in or out
- **option** + **scroll** (side to side) rotates the polygons.

4 Use of the tool

The tool was demoed at OpenVisConf in McNamara's talk, "Do you know Nothing when you see it?" (McNamara, 2016).

5 Further work

The tool as it stands is already a fun thing to play with. As you move bins around, you can watch the spatial pattern shift (or not). It provides a clear improvement on static maps using similar spatial aggregation (Sullivan et al.,

2008; Felton, 2015). We have observed that it is just as interesting to watch someone else manipulate the interface, because you can separate yourself from the immediacy of the interaction.

However, we are already thinking of many extensions. Most important will be a way to compare some number of maps of the same data, with different aggregations. One thought is just to display a trailing history of the last 12 views, but there could be a way to "lock" one view into the list for comparisons.

Past that, a few particularly salient examples of spatial patterns that show large variation in spatial pattern will help motivate the usefulness of the tool. We encountered a catch-22 when building this tool. It was hard to think of an example that would show large variation without the tool to visualize it, and it was hard to build the tool without a motivating example. Once some illustrative cases have been identified, we hope to publish an interactive essay for a broad audience.

In the context of gerrymandering, it would be great to have a more specialized interface to allow users to manipulate voting districts to see the predicted effect on election outcomes. We have thought through several approaches to this problem, and the method that currently seems most reasonable would be to display district boundaries with the Census blocks of which they are composed, and allow users to indicate blocks they want to switch in and out of districts. A more automated approach would have the interface automatically displace districts by some number of blocks in each direction, and display the most different patterns that could be produced.

From a more statistical perspective, this platform could provide a nice way to visualize the effects of up-, down-, and side-scaling (Atkinson, 2013; Kyriakidis, 2004). It could also be combined with data augmentation to help add more detail to pre-aggregated data, as with the tool Disser (Martin-Anderson, 2014).

References

Gregor Aisch, Eric Bluth, Matthew Bloch, Amanda Cox, and Kevin Quealy. The best and worst places to grow up: How your area compares. *The New York Times*, May 2015.

Peter M Atkinson. Downscaling in remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 22:106–114, 2013.

Nate Cohn. Voting case has potential to put house further out of reach for democrats. *The New York Times*, June 2015.

Daniel Ervin. Maup: An introduction to the modifiable areal unit problem. `http://gispopsci.org/maup/`, 2015.

Nicholas Felton. Geometric choropleths 1895 vs 1978. `http://feltron.tumblr.com/post/126340096801/geometric-choropleths-1895-vs-1978`, 2015.

Andrew Gelman and Phillip N Price. All maps of parameter estimates are misleading. *Statistics in Medicine*, 18:3221–3234, 1999.

Chris Hancock, James J Kaput, and Lynn T Goldsmith. Authentic inquiry with data: Critical barriers to classroom implementation. *Educational Psychologist*, 27(3):337–364, 1992.

Christopher Ingraham. This is the best explanation of gerrymandering you will ever see. *The Washington Post*, March 2015.

Clifford Konold, Traci Higgins, Susan Jo Russell, and Khalimahtul Khalil. Data seen through different lenses. *Educational Studies in Mathematics*, 2014.

Phaedon C Kyriakidis. A geostatistical framework for area-to-point spatial interpolation. *Geographical Analysis*, 36(3), 2004.

Aran Lunzer, Amelia McNamara, and Robert Krahn. LivelyR: Making R charts livelier. In *useR! Conference*, 2014.

Brandon Martin-Anderson. Building precise maps with disser. `http://conveyal.com/blog/2014/04/08/aggregate-disser/`, April 2014.

Amelia McNamara. *Bridging the Gap Between Tools for Learning and for Doing Statistics*. PhD thesis, University of California, Los Angeles, June 2015.

Amelia McNamara. Do you know Nothing when you see it? In OpenVisConf, `https://www.youtube.com/watch?v=hps9r7JZQP8`, April 2016.

Penn State GEOG 486. Choropleth maps. `https://www.e-education.psu.edu/geog486/node/1864`, 2014.

Brian Sullivan, Steven T Kelling, Christopher Wood, Marshall Iliff, Daniel Fink, Mark Herzog, Doug Moody, and Grant Ballard. Data exploration through visualization tools. *Proceedings of the fourth international partners in flight conference: Tundra to tropics*, pages 415–418, 2008.