Sharing Finer Resolution Protected Health Information

USING REGIONALIZATION TO CREATE HIPAA-COMPLIANT AGGREGATIONS

Abstract

Background: This project addresses the challenge of sharing finer-scale Protected Health Information (PHI) while maintaining patient privacy by exploring the use of regionalization to create higher resolution HIPAA-compliant geographical aggregations. **Methods:** Four regionalization methods 1) Max P Regions, 2) REDCAP, 3) MSOM (Max P + SOM), and 4) RSOM (REDCAP + SOM) were each used to create 30 configurations of regions that optimized intra-unit homogeneity and maximized the number of spatial units while meeting the minimum population threshold required for sharing PHI under HIPAA guidelines. The relative utility of each configuration was assessed according to: 1) model fit, 2) compactness, 3) homogeneity, 4) resolution, and 5) suppression. **Results:** Adding SOM to Max P significantly impacted all five assessment measures, while the addition of SOM to REDCAP resulted in much less of an impact on these measures. **Discussion:** Each method has its advantages and weaknesses. Deciding which regionalization approach works best will depend on its intended use.

Background

In order to protect patients' privacy the Health Insurance Portability and Accountability Act (HIPAA) limits the ways in which Protected Health Information (PHI) can be shared. Under the safe harbor method, patient address data must be aggregated to geocodes that contain populations no less than 20,000 people. ZIP code tabulation areas (ZCTAs) are the finest pre-existing units permitted to be shared under the HIPAA safe harbor rule, but not every ZCTA meets the required population threshold which means that some units are required to be suppressed from the dataset, leaving investigators with patchy coverage of their study area (Figure 1).

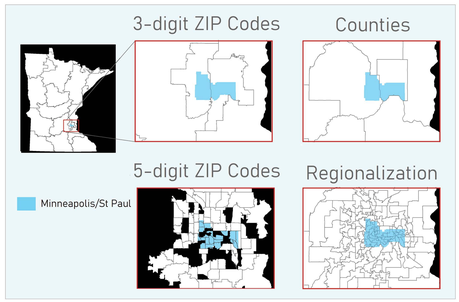


Figure 1. The greater Minneapolis/St Paul area shown in different map units with varying levels of resolution. Three-digit ZIP codes and counties offer very low spatial resolution. Five-digit ZIP codes require a great deal of suppression. Regionalization provides map units with much higher resolution and no suppression.

Methods

Four regionalization methods were each used to create 30 configurations of regions that optimized intra-unit homogeneity and maximized the number of spatial units while meeting the minimum population threshold.

1. Max P Regions
2. MSOM (Max P + Self Organizing Maps)
3. REDCAP
4. RSOM (REDCAP + Self Organizing Maps)

All four regionalization strategies used the same input: median household income, proportion with bachelor's degree or higher, and proportion people of color. These three sociodemographic variables were entered directly into the Max P and REDCAP methods and indirectly entered (via a u-matrix outputted from SOM) into the MSOM and RSOM methods. To account for random variation within the regionalization process, each of the four methods was repeated 30 times and assessed according to averages. 120 different configurations from all four regionalization strategies were assessed according to model-fit (Figure 3.), homogeneity (Figure 4.), compactness via the isoperimetric ratio (Figure 2.), resolution (Figure 5.), and suppression (Figure 6.).

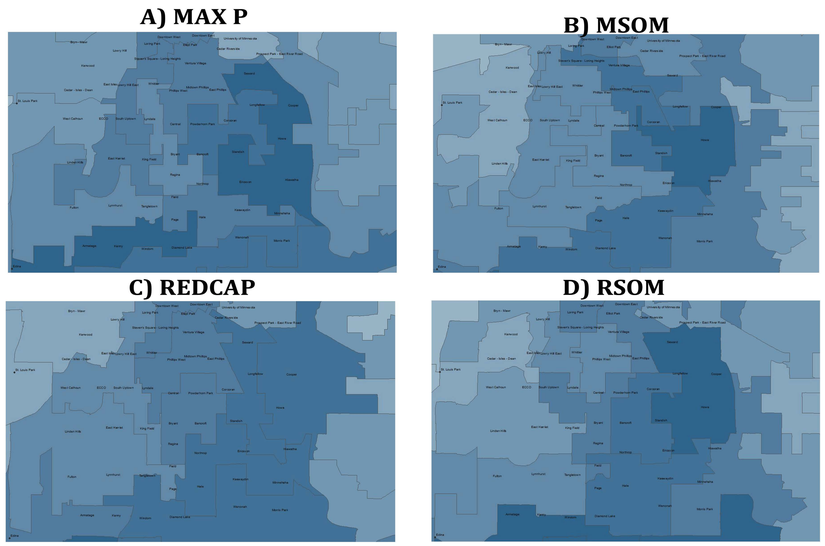


Figure 2. Mapping the same variable with all four regionalization approaches using the same map extent (zoomed in to show Minneapolis neighborhoods).

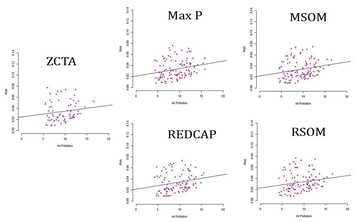


Figure 3. Model fit was assessed using the Akaike Information Criterion (AIC) from regression output of an unconditional model of depression & air quality.

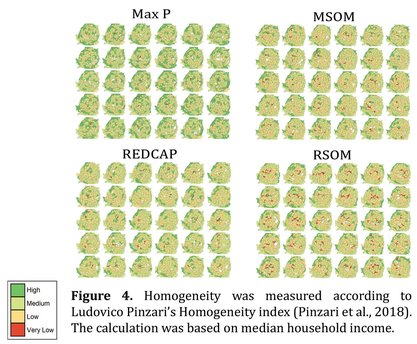


Figure 4. Homogeneity was measured according to Ludovico Pinzari's Homogeneity index (Pinzari et al., 2018). The calculation was based on median household income.

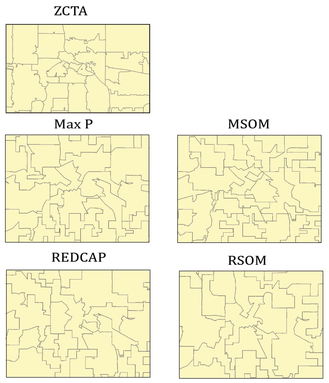


Figure 5. Vector resolution is often defined as the number of, or average area of, the regions.

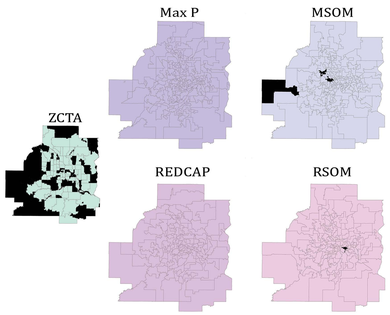
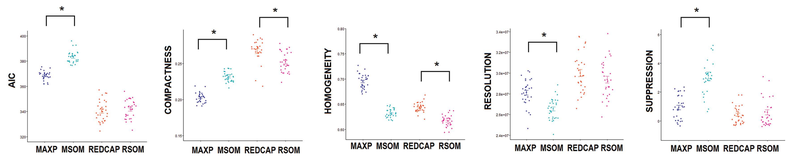


Figure 6. Suppression is simply defined as the number of regions that contain populations less than 20,001.

Results

Adding SOM to Max P significantly impacted all of the assessment measures. In specific, regionalization with MSOM created configurations that, on average, had higher measures of compactness, more desirable vector resolution (lower average areas), and better model-fit (higher negative AIC values) when compared to Max P. Adding SOM to REDCAP (RSOM) had much less of an impact. The only significant changes were a slight but significant decrease in homogeneity and compactness. These differences were not as striking as those observed between Max P to MSOM, which is largely due to the fact that Max P relies on a less constrained regionalization algorithm than REDCAP.



Discussion

Which configuration is best? It depends…
Max P provided the highest homogeneity on average, REDCAP created the most compact regions, and MSOM created configurations with the finest units (or highest resolution) and best model-fit.

It is important to consider how to prioritize the different region characteristics. Homogeneity, which although may seem like a very important factor, should not be prioritized over compactness if it means to degrade the integrity of the region shape in a way that mirrors gerrymandering (Figure 2a). Elongated or wildly regions may provide high homogeneity or great relative model-fit, but this is often an artifact of spatial p-hacking and therefore these configurations should not be used without a strong justification for doing so.

Since Max P was impractical in terms of compactness and RSOM was not significantly different from REDCAP in most respects, we should make a comparison between REDCAP and MSOM. When we compare these two, no clear winner emerges. REDCAP provided more compact regions with higher homogeneity, and less suppression, but MSOM configurations had the best model-fit in addition to providing the finest units possible for mapping under HIPAA guidelines.

Conclusions

By sharing data at finer resolutions that adhere to HIPAA guidelines, we can provide more accurate depictions of geographic variations in health without jeopardizing patient privacy. This project explored four different regionalization strategies, each having their own strengths and weaknesses in terms of the configurations they created. Investigators are encouraged to use the strategy that best suits the needs of the project at hand.

Acknowledgements

I'd like to thank Steven Manson, Mark Lindbergh, Michael Oakes, Eric Shook, Susan Everson Rose, and Brian Swies for the guidence they provided, and Constance Krzyzanowski for her photoshop assistance and graphic design input.

University of Minnesota

Brittany Krzyzanowski

Contact:
krzyz016@umn.edu