HACKING SPATIAL DATA: AN EXAMPLE OF AGGREGATION PROBLEMS

Ningchuan Xiao

Department of Geography, The Ohio State University, Columbus, OH, United States – xiao.37@osu.edu

Commission IV, WG IV/4

KEY WORDS: Spatial aggregation, Spatial data science, Hacking spatial data, Spatial autocorrelation

ABSTRACT:

Many applications using spatially aggregated data tend to treat the spatial units as given. For example, in the United States, analyses using the social and economic data often rely on the existing and fixed spatial units of census blocks or tracts. However, these spatial units are often aggregated arbitrarily. It is therefore important to ask this question: what if the spatial units are aggregated differently? Will the results obtained using the existing units still hold? This paper addresses questions like these. We first develop a search algorithm that can be used to find alternative aggregations with relatively equal total populations among the aggregated units. Then a number of experiments are conducted to test the algorithm and to demonstrate how alternative aggregations will affect the analysis. These experiments clearly suggest the significant effects of spatial aggregation on the analysis results.

1. INTRODUCTION

In recent years, data science has quickly emerged as a field that brings together researchers from a wide range of disciplines such as computational science, statistics, social sciences, and geography (Hey et al. 2009). At the heart of this movement is the ever growing need of understanding the data. Noticeably much of such data has a strong spatial context. However, it is also reasonable to argue that the spatial units under these data are typically taken as given. For example, spatially aggregated data at different levels (e.g., states, counties, census blocks, or wards) are commonly used in many geographic and social science applications. But what if different kind of spatial units are used? In other words, what if we aggregate the census blocks in different ways but still at the same level of the tracts? Will the result from the existing units still hold? Is it possible to *hack* spatial data so that we can examine it from a different angle? Spatial data has become increasingly more accessible, which makes it more important to fully understand the impacts of using such data.

It is difficult to find alternative aggregations of spatial units because such a process often takes exceedingly amount of computation and there are many equally good alternatives to be considered (Xiao 2008; Kim and Xiao 2017). Methods developed in the literature (e.g., Openshaw and Rao 1995; Martin 1998; Cockings and Martin 2005) are often designed to search a single alternative instead of multiple.

The purpose of this paper is to demonstrate new ways of exploring spatial aggregation. Specifically, we develop a new computational method that can be used to find multiple aggregations as alternatives to the existing one. We then test this method using the census data for a county in the United States. We compare and contrast the difference between the official census data and the alternatives. We conclude that the use of alternatively aggregated data have significant impacts on spatial analysis.

2. METHODOLOGY

Given a set of n spatial units, we assume each unit is associated with a population (or other types of weights or attributes). The goal of an aggregation problem is to group these units into a number of contiguous regions so that the overall difference in the population of the aggregated regions is minimized.

To search for alternative aggregations, we first randomly generate a pool of valid aggregations and then an efficient algorithm called give-and-take (Kim 2011) is used to improve each aggregation by repeatedly swapping units from one region to its neighboring regions. At the end of these improvements is a pool of relatively high quality aggregations. Then we randomly select pairs of aggregations from the pool and recombine them, and the recombined aggregation will be inserted into pool to replace the worst aggregation if the former is a better one. This recombination process is repeated many times as specified by the user.

3. DATA AND RESULTS

We first test the search algorithm using is a small data set (Figure 1) where each spatial unit (cell) is randomly assigned a population value and these units are grouped into 3 regions. The algorithm uses a pool of 20 and all the 20 alternatives found at the end have 3 regions with exactly the same total population. Figure 2 shows 3 examples of these aggregations.

34	38	21	45	19	40	28	41	32	47
49	17	12	45	41	26	23	26	20	39
22	35	24	10	19	28	46	25	28	20
30	39	50	31	30	49	29	48	18	20
15	45	20	36	37	36	26	22	31	33
34	15	30	19	42	47	17	18	15	48
15	23	46	14	30	41	12	27	20	25
26	50	30	29	24	25	28	18	35	13
47	49	14	10	43	32	14	32	26	31
24	15	23	23	47	49	41	16	34	21

Figure 1. A random test data set on a 10x10 grid.

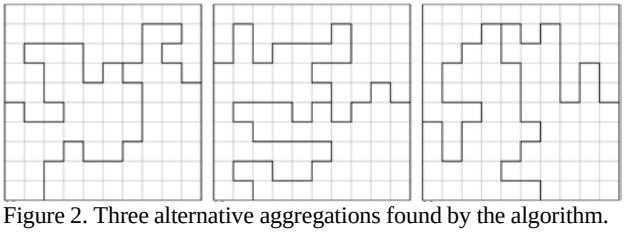

Figure 2. Three alternative aggregations found by the algorithm.

* Corresponding author

We then use the algorithm to aggregate the 887 census block groups in Franklin County, Ohio. In the official census data for Franklin County, there are 284 census tracts, and our goal is to find alternative aggregations with 284 regions. We again set the pool size to 20.

Figure 3 shows the histograms of population for the official census tracts (left) and one of the alternatives found by the algorithms. It is clear that the official 284 census tracts have a wider range of population (from as small as 8 to as high as 14,652), while the alternative 284 regions have a much narrower range of the population centered around 4,000.

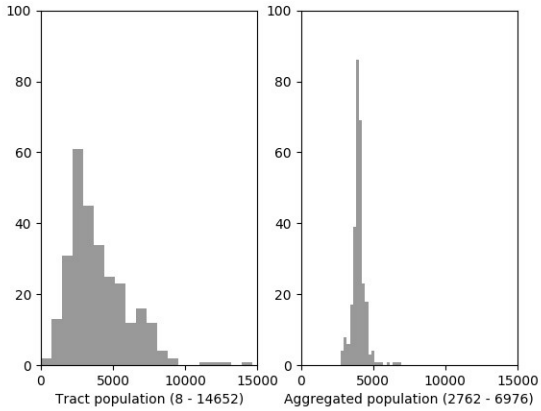

Figure 3. Histograms of population for the official census tracts (left) and aggregated tract level regions (right).

Maps in figures 4 and 5 compare the total population of the official census tracts and one of the aggregations yielded by the algorithm (maps use the quantile classification method). It is clear that the official tracts have a wider range of data and the units with the same color tend to be close to each other as compared with the pattern from the aggregated units.

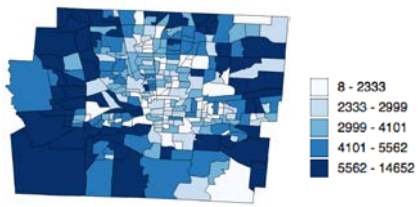

Figure 4. Total population of the official census tracts.

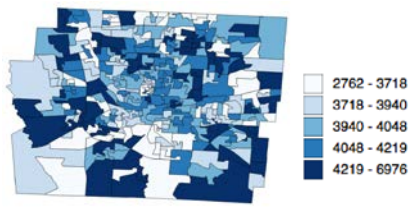

Figure 5. Total population of the 284 aggregated regions.

We further examine the spatial autocorrelation of the total population and minority (non-white) rates. The total population of the official census tracts shows a significant spatial autocorrelation with a Moran's *I* value of 0.349, while the 20 alternative aggregations do not show such spatial autocorrelation with their Moran's *I* values ranging from -0.063 to 0.094.

It is clear that by aggregating spatial units into regions with relatively equal population, we are able to reduce or even eliminate the effect of spatial autocorrelation in statistical analysis. For example, our preliminary results shows that *R*-square value of a regression model between minority rate and median household income at the block group level is 0.231. When the official tracts data is used, a higher *R*-square value at 0.304 is yielded, but our aggregated data have their *R*-square values ranging from 0.242 to 0.259, which is closer to what is found at the original block groups level.

4. CONCLUSIONS

The experiments discussed in the previous section suggest that different ways of aggregating spatial units will lead to different results (as in the cases of spatial autocorrelation and correlation between different socioeconomic variables). It is therefore important to explore alternatives when aggregated data are used to address social science questions.

The method and results presented here are still work in progress. However, our experiments show that the method developed is effective in handling certain sizes of spatial data. The future plan is to utilize new high performance computing technologies to handle larger data sets for larger areas.

5. REFERENCES

Cockings, S. and D. Martin 2005. Zone design for environment and health studies using pre-aggregated data. *Social science & Medicine* 60(12), 2729–2742.

Hey, T., Tansley, S. and Tolle, K.M., 2009. *The fourth paradigm: data-intensive scientific discovery*. Redmond, WA: Microsoft research.

Kim, M. J. 2011 . *Optimization Approaches to Political Redistricting Problems*. Ph.D. Thesis, The Ohio State University, Columbus, OH.

Kim, M. J. and N. Xiao. 2017. Contiguity-based optimization models for political redistricting problems. *International Journal of Applied Geospatial Research*, 8(4): 1-18.

Martin, D. 1998. Optimizing census geography: the separation of collection and output geographies. *International Journal of Geographical Information Systems* 12(7), 673–685.

Openshaw, S. and L. Rao 1995. Algorithms for reengineering 1991 census geography. *Environment and Planning A* 27, 425–446.

Xiao, N. 2008. A unified conceptual framework for geographical optimization using evolutionary algorithms. *Annals of the Association of American Geographers* 98(4), 795–817.