Spatial Aggregation as a Means to Improve Data Quality

Min Sun[1, 3], David W. Wong[2, 3]

[1]NSF Spatiotemporal Innovation Center
George Mason University Fairfax, VA 22030 USA Telephone: (01)703-993-9612
Email: msun@gmu.edu

[2]Department of Geography University of Hong Kong Pokfulam, Hong Kong Telephone: (852) 9870-8773
Email: dwong2@hku.hk

[3]Department of Geography & GeoInformation Science
George Mason University Fairfax, VA 22030 USA Telephone: (01) 703-993-9260
Email: dwong2@gmu.edu

1. Introduction

The accuracy of spatial data may be broadly divided along two dimensions: accuracy of data representing the geometric characteristics of features and accuracy of attribute data (Goodchild and Gopal 1989). Research on the accuracy of attribute data often falls onto the laps of statisticians. Research in GIScience on this topic is mostly limited to representing and visualizing the quality of attribute data (e.g., Leitner and Buttenfield 2000), if error in mapped values is acknowledged. But in fact, most mapped values are in essence estimates based upon samples and sample size is likely the most influential factor in affecting attribute accuracy in most cases. Small sample sizes will likely produce unreliable estimates.

Using spatial data with unreliable attributes is undesirable, but often no alternative is available. For instance, the American Community Survey (ACS) is currently the only source providing detailed population and housing information for socioeconomic research in the U.S. Similarly, many health statistics, such as those provided by cancer registry, are not available elsewhere, regardless of how unreliable these data may be.

One possible approach to improve the reliability of attribute data (but with costs) is through aggregation (Salvo 2014). Data can be aggregated through the attribute space by reducing the number of variables or classes of a variable or through the geographical space by merging areal units. This paper reports an effort to develop a heuristic approach to aggregate areal units, both the space and the attribute through statistical method, in order to improve the reliability of the attribute data.

2. Data Quality and Spatial Aggregation

Reliability of an estimate (which is often a sample mean) can be reflected by the standard error of the estimate $\bar{x}$:

$$SE_{\bar{x}} = \frac{S}{\sqrt{n}}$$

where S is the standard deviation of sampled values and n is the sample size. Apparently, small sample size will likely have large SE, and increasing sample size can reduce SE, raising the estimate reliability. Merging areal units enlarges the sample size of each new units so as to reduce the standard error with some undesirable consequences, which is particularly due to the scale effect of the Modified Areal Unit Problem (MAUP).

Our general objective is to develop a procedure to aggregate areal units to reduce error, but minimize associated costs.

Methods to aggregate areal units to meet certain analysis or modelling objectives have been developed (e.g., Cockings and Martin 2005, Guo et al. 2001, Openshaw 1978). These methods determine how all units in the study region should be merged into larger but fewer zones by optimizing an objective function. In our current context, estimates of certain areal units have relatively high levels of uncertainty such that users may find uncomfortable of using the data. Our objective is to derive data for those areas with poor estimates such that the new data meet pre-defined reliability levels acceptable to the users. The general spatial aggregation approach based upon optimization can surely increase the sample sizes and therefore improves the reliability of estimates, but such an approach suffers from several major drawbacks.

The traditional aggregation approach removes the entire original geographical structure. Areal units with geographical meanings, such as communities or neighbourhoods may no longer be recognized after merging of areal units. A related problem is that the spatial resolution of the aggregated data is lower than that in the original data, making local-scale analysis more difficult and challenging. In most cases, some units may have estimates with acceptable reliability levels and therefore they should not be merged. These units with reasonable estimates are subject to the “risk” of aggregation, changing the geography that may not need to be changed. The black-box optimization approach of spatial aggregation may be difficult to incorporate analyst’s local knowledge of the study area, failing to recognize the boundaries of communities or neighbourhoods. Therefore, our proposed heuristic method assists analyst with local knowledge to create a “new” zoning system such that it resemble the original one as much as possible, but new data for areas with poor estimates are improved to an acceptable level.

3. Spatial Aggregation Procedure

The proposed heuristic spatial aggregation procedure involves three phases. The first phase identifies areal units to be the seeds of aggregation. These are units with relatively large error levels. Based upon relevant constraints, the second phase identifies candidates to be merged with the seeds. Results of different aggregation scheme and corresponding consequences are computed and evaluated. Eventually, analyst selects the most desirable choices for the aggregation.

The phases of the proposed aggregation procedure are facilitated by a set of visual-analytic tools and involve the active participation of the analyst in evaluating options. During phase one, error levels of attributes are shown graphically. Analyst can experiment different levels of error as the cut-off to select areal units to be the seeds. During phase two, spatial computational tools running in the background will identify all aggregation candidates corresponding to each seed. These candidates are evaluated according to the selected criteria and constraints. Some constraints and criteria include:

1) A candidate should be contiguous to the seed.

2) Error of the merged unit should be lower than the acceptable error thresholds determined in phase one.

3) If no single candidate meets criterion #2, then multiple candidates should be considered together in the aggregation process.

4) The bias of the new estimate for the merged unit (new compares with the original estimates) should be minimized.

5) Compactness and some other characteristics of the merged unit should be considered (Datta et al. 2012, Li et al. 2013)

Graphical displays coupling with computational tools in the background will show the trade-off relationship between selected criteria across all candidates meeting the constraints.

Analyst will evaluate and experiment with different options (candidates) based upon computed statistics for selected criteria and constraints. The trade-off relationship between different criteria will be revealed gradually through experimentation of different aggregation schemes. Analyst has to determine the preferred aggregation schemes. The visual-analytic environment includes visualization tools, real-time computational capabilities and interactive user operations that facilitate such heuristic processes.

4. An Example and Conclusion

We selected two variables with relatively large errors from ACS to demonstrate the usefulness of the proposed approach. Seeds for aggregation were selected based upon two threshold values of margin of error. Given the criteria to evaluate candidates, multiple zoning systems were possible. Summary statistics of their error levels are reported to show that aggregation can lower over error levels, but the new estimates create moderate levels of bias. The spatial patterns of the new zonal systems were also evaluated with spatial autocorrelation statistics and were compared with the original system.

5. Conclusion

The proposed heuristic spatial aggregation approach has several advantages over previous black-box optimization methods: less “intrusive” as it will preserve the current spatial configuration as much as possible; allow rooms to take into account of local information in the aggregation process. The proposed method has several limitations. The process does not guarantee finding the optimal solution. Involving a large number of attributes may overload the analyst with information beyond comprehension.

6. Acknowledgements

Research reported in this publication was supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number R01HD076020. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

7. References

Cockings, S. and D. Martin. 2005. Zone design for environment and health studies using pre-aggregated data. Social Science & Medicine, 60: 2729-2742.

Datta, D., J. Malczewski and J.R. Figueira. 2012. Spatial aggregation and compactness of census areas with a multiobjective genetic algorithm: a case study in Canada. Environment and Planning B: Planning and Design, 39(2) 376 – 392.

Goodchild MF and Gopal S. 1989. Accuracy of Spatial Databases. Taylor and Francis, London, UK.

Guo, J.Y., G. Trinidad and N. Smith. 2001. MOZART: a Multi-objective Zoning and AggRegation Tool. Paper presented at the TRB 80th Annual Meeting, Transportation Research Board, Washington, DC.

Leitner M and Buttenfield BP, 2000, Guidelines for the display of attribute certainty. Cartography and Geographic Information Science 27(1): 3-14.

Li, W., M.F. Goodchild and R. Church. 2013. An efficient measure of compactness for two-dimensional shapes and its application in regionalization problems. International Journal of Geographical Information Science, 27 (6): 1227–1250.

Openshaw, S. 1978. An optimal zoning approach to the study of spatially aggregated data. In I. Masser and P.J.B. Brown (eds.), Spatial Representation and Spatial Interaction, Martinus Nijhoff, Leiden: 93-113.

Salvo J. 2014. Using small-area data from the ACS: Issues, challenges, solutions, presentation in Workshop sponsored by the PAA Committee on Population Statistics (http://www.prb.org/pdf14/section2-salvo.pdf, access on January 10, 2015)