Computers, Environment and Urban Systems 65 (2017) 15–27

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Spatial aggregation as a means to improve attribute reliability

Min Sun *, David W.S. Wong [1]

Department of Geography and Geoinformation Science, George Mason University, Fairfax, VA, USA

ARTICLE INFO

Article history:
Received 6 November 2016
Received in revised form 12 April 2017
Accepted 14 April 2017
Available online xxxx

Keywords:
Attribute reliability
Standard error
Spatial aggregation
Interactive-heuristic method

ABSTRACT

Attributes of areal units are often estimates derived from survey samples. Estimates of these attributes with large standard errors (*SEs*) discount the confidence and validity of spatial analytical results. Large *SE* for estimates of enumeration units are often the results of small sample sizes in areal units and imply unreliable attribute values. One way to suppress error in attributes is to merge areal units to raise sample size. Traditional regionalization methods serve this purpose, but may unnecessarily alter the geography of the study area. We propose an interactive-heuristic aggregation approach to assist analysts in selecting and merging only units with *SEs* larger than acceptable levels while preserving the original geography and data as much as possible. Results of this approach and a recent automated optimization method are comparable. Both methods successfully lower the *SEs* in attribute data, but the interactive approach flexibly adjusts the importance levels of different aggregation criteria across areal units, thus offering a high degree of transparency in the aggregation process. The interactive approach also incorporates subjective and local knowledge of neighborhoods in selecting areal units for aggregation.

1. Introduction

When spatially aggregated data are used in mapping, they are often assumed to be either accurate or errors in data are not substantial. These assumptions may be acceptable when using reasonably accurate data (e.g., decennial censuses). However, the number of spatial datasets, including many from public health and socioeconomic surveys, is growing. When sample data are gathered using relatively small sample sizes, resultant estimates serving as the attributes of areal units may have substantial sampling error. Sampling errors in these estimates for feature attributes are often ignored. When these areal data are used in spatial data analysis and visualization, results can be misleading and conclusion erroneous (MacEachren, Brewer, & Pickle, 1998, Heuvelink & Burrough, 1989). If data analysis fails to consider the magnitude of errors in data, especially sampling error, and the analysis results are used to support mission-critical decisions, the outcomes may be disastrous.

For more than two decades, many scholars have addressed issues related to spatial data accuracy (e.g., Beard, Buttenfield, & Clapham, 1991, Goodchild & Gopal, 1989). These may be broadly divided along two dimensions: 1) accuracy of data representing the geometric characteristics of features, including the positional accuracy, and 2) accuracy of attribute data (e.g., Caspary & Scheuring, 1993, Hunter & Goodchild, 1996, Griffith, Wong, & Chun, 2016). The error in attribute data is partly attributable to sampling. Most attribute data are estimates based upon samples of individual observations. While many factors may affect the accuracy of these estimates, sample size is often the most influential factor. Small sample sizes will likely produce estimates with large standard errors (*SEs*) reflecting the low reliability of the estimates. When the *SE* of an estimate is large, the estimate may deviate greatly from the true value, rendering them inaccurate and unreliable.

Our focus is to reduce the *SEs* of estimates attributable to sampling so that attribute values (i.e., estimates) for areal units in spatial analysis and mapping are more reliable. Herein, the term "error" refers to standard error or sampling error. Estimates in these data can be rates or counts in interval-ratio scale derived or aggregated from the original individual-level samples. But these individual samples are not available to data analysts - only estimates as attribute values and associated *SEs* or margin of errors (*MOEs*) for areal units are available.

Using spatial data with accurate attributes is preferred, but often highly accurate data are not available. For example, after the 2000 Census in the U.S., socioeconomic data for population and housing are only available through the American Community Survey (ACS). Unfortunately, ACS estimates for small areal units such as block groups or even census tracts are not highly reliable and their *MOEs* can be relatively large (Bazuin & Fraser, 2013, Citro & Kalton, 2007). In the U.S., no other publicly available nation-wide data can provide similar types of information to support socioeconomic research. Similarly, estimates of some public health datasets have substantial errors, partly attributable to the rare-event nature of certain health conditions or disease outcomes. An example of such datasets is the National Institute of Health's Surveillance,

* Corresponding author at: Department of Geography and Geoinformation Science, 4400 University Drive, MS 6C3, Fairfax, VA 22030-4444, USA.
E-mail addresses: msun@gmu.edu (M. Sun), dwong2@gmu.edu (D.W.S. Wong).
[1] Department of Geography and Geoinformation Science, 4400 University Drive, MS 6C3, Fairfax, VA 22030-4444, USA.

http://dx.doi.org/10.1016/j.compenvurbsys.2017.04.007
0198-9715/

Epidemiology, and End Results (SEER) Program.[2] These data programs and many surveys provide aggregated sample estimates with sampling error information (e.g., *SE* or *MOE*) for spatial units.[3] For those data for small areas or small population sizes, the sampling errors of estimates may be too large to be useful.

Developing methods to improve attribute accuracy and thus the reliability of survey estimates are much needed. One possible approach, with potentially significant cost, is aggregation. Data can be aggregated in the attribute space by collapsing the number of variables and/or reducing the number of classes of a variable (Salvo, 2014). Data can also be aggregated spatially by merging areal units and aggregating their respective attribute values. Aggregation may reduce error or suppress variance because when smaller enumeration units are merged to form larger units, new units have sample sizes larger than each of the original units. With larger sample sizes, the *SEs* (and *MOEs*) of new estimates, standard deviations divided by the square root of sample sizes, would likely be lower than the *SEs* of the original units (more explanations below).

This article presents an interactive heuristic approach and a tool to implement that approach. The tool helps select and merge areal units with large errors in their estimates with other units to improve the reliability of attribute data for mapping and spatial analysis. This approach is implemented in an interactive environment, offering high degree of flexibility to determine how units should be merged and ample information about the concerned areal units through various means of geovisual analytics. Due to the interactive nature of the approach, it is not intended to handle large datasets with thousands of records (areal units). In the next section, the connection between sampling error and aggregation is discussed and notes that research has been limited in recent years. In Section 3, the proposed approach and different components of the associated tool are presented. The tool is freely available to the public (details will be provided after the review of manuscript is completed). Using ACS data, we demonstrate the approach and associated tool in Section 4. While many other datasets can be used, ACS data are selected because they have been widely used without considering their reliability. We discuss some strengths and weaknesses of the proposed approach in the final section.

2. Data quality and spatial aggregation

Standard error is a common measure of the reliability of an estimate $\bar{x}$ (i.e., sample mean) and is defined as follows:

$$SE_{\bar{x}} = \frac{S}{\sqrt{n}} \tag{1}$$

where s is the standard deviation of sampled values and n is the sample size. Apparently, a small sample size n will produce a relatively large *SE*, and a larger sample size can reduce *SE*, and a more reliable estimate. When estimates have relatively large *SEs*, using these estimates will be risky.[4] Salvo (2014) suggested collapsing variables into fewer categories or classes such that the sample size in each category for each tabulation unit (e.g., a census enumeration unit) becomes larger, reducing the *SE*. Obviously, doing so will introduce some undesirable consequences, such as increasing the difficulties in differentiating observations among groups in the sample. A full discussion on the drawbacks of this aggregation approach is beyond the scope of this article.

Having large *SEs* in estimates is a commonly encountered issue in spatial epidemiology and mapping health statistics. These estimates, such as the rates of contracting a disease, fluctuate across units within a neighborhood (Bell, Hoskins, Pickle, & Wartenberg, 2006). These unstable rates usually have large errors, indicating that they are not reliable. They are often the results of events with low frequencies (rare-event statistics). Large errors can also be the results of small population sizes based on which rates are derived. With fluctuating estimates across neighboring units, it is difficult to detect regional trends. To reduce the rate fluctuation and improve the reliability to identify a spatial trend or determine a cluster, smoothing is a common solution (e.g., Aylin et al., 1999; Kadadar, 1997), and many smoothing methods have been adopted. For example, the "head-banging" algorithm uses the median to iteratively smooth the distribution (e.g., Mungiole, Pickle, & Simonson, 1999), and Rushton (2003) provided a detailed review of different smoothing methods. Spatial smoothing borrows information from neighboring units to improve the reliability of estimates. A drawback of using spatial smoothing is that original estimates of all units will be altered, regardless if they are reliable or not. In other words, originally reliable estimates may be changed unnecessarily. Also, because the smoothed estimates are derived from drawing partly from the original estimates of neighboring units, the reliability of the new estimates (e.g., *SE*) is unknown, even though such information is essential (U.S. Census Bureau, 2008).

Spatial aggregation is another method to suppress the fluctuation of estimates across areal units. In general, aggregating areal units is undesirable in the context of spatial analysis. This is known as the scale effect (discussed in the literature of the modifiable areal unit problem or the MAUP), which refers to the varying of analytical results when data of different spatial resolutions are used (Openshaw & Taylor, 1979; Armhein, 1995; Wong, 2009). Spatial aggregation may be regarded as a form of spatial smoothing by borrowing information from neighboring units. As spatial smoothing draws values from neighboring units based on fractional weights, spatial aggregation uses a binary weight (0, 1), determining if a neighboring unit will be merged or not with the original unit. After two enumeration units are merged, the resultant unit will have an *SE* lower than one or both of the original *SEs*. An obvious cost of spatial aggregation is that the data's spatial resolution will be lowered. In addition, bias is introduced to the new estimates because they are derived using spatially aggregated data rather than the original individual level data.

In this study spatial aggregation is used to improve the attribute reliability in spatial data. Aggregating areal units to achieve certain analytical or modeling objectives is not new (Openshaw, 1978). There are many ways to merge the geography and attributes of enumeration units. In general, enumeration units with smaller area or population size are more likely to be merged. Randomly aggregating units is usually not warranted except in the case of statistical testing (randomization test) (Openshaw & Rao, 1995). Nevertheless, using different aggregation schemes produce different datasets. Openshaw (1978) suggested evaluating the fitness of different aggregated schemes based on an objective function formulated according to the intention of analysis. An aggregated spatial configuration should be chosen to optimize the objective function. Such an aggregation approach has been applied in many studies to meet a variety of objectives, but none, except that of Spielman and Folch (2015), focused on suppressing error to improve data reliability (e.g., Cockings & Martin, 2005, Guo, Trinidad, & Smith, 2000, Haining, Wise, & Ma, 1998, Li, Goodchild, & Church, 2013, Martin, 1998, Martin, Nolan, & Tranmer, 2001, Openshaw & Rao, 1995). Spatial aggregation has two major drawbacks. The process removes the original zonal or geographical structure. During the process, areal units constituting communities or neighborhoods that are meaningful to residents may be merged, and they may no longer be

[2] https://seer.cancer.gov/.

[3] Numerous survey datasets are relevant to the discussion here. A large scale survey of demographic information is the Current Population Survey conducted by the U.S. Census Bureau and the U.S Bureau of Labor Statistics (http://www.census.gov/programs-surveys/cps/about.html). Two additional examples that the authors have dealt with are the Centers for Disease Control and Prevention (CDC) State Tobacco Activities Tracking and Evaluation (STATE) System (https://chronicdata.cdc.gov/Survey-Data/Graph-of-Cigarette-Use-Among-Adults-Behavior-Risk-/syfb-fzcd) and the State of Obesity: 2015 data collected by the Trust of America's Health and the R. W. Johnson Foundation (http://tfah.org/assets/files/TFAH-2015-ObesityReport-final.22.pdf) (Accessed on January 24, 2017).

[4] Therefore, many opinion polls try to get large numbers of respondents, besides deciding how to select respondents.