Computers, Environment and Urban Systems xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

An integrated classification scheme for mapping estimates and errors of estimation from the american community survey

Ran Wei [a,*], Daoqin Tong [b], Jeff M. Phillips [c]

[a] Department of Geography, University of Utah, Salt Lake City, UT 84112, USA
[b] School of Geography and Development, University of Arizona, Tucson, AZ 85721, USA
[c] School of Computing, University of Utah, Salt Lake City, UT 84112, USA

ARTICLE INFO

Article history:
Received 14 September 2015
Received in revised form 5 March 2016
Accepted 6 April 2016
Available online xxxx

Keywords:
Classification
Uncertainty
ACS

ABSTRACT

Demographic and socio-economic information provided by the American Community Survey (ACS) have been increasingly relied upon in many planning and decision making contexts due to its timely and current estimates. However, ACS estimates are well known to be subject to larger sampling errors with a much smaller sample size compared with the decennial census data. To support the assessment of the reliability of ACS estimates, the US Census Bureau publishes a margin of error at the 90% confidence level alongside each estimate. While data error or uncertainty in ACS estimates has been widely acknowledged, little has been done to devise methods accounting for such error or uncertainty. This article focuses on addressing ACS data uncertainty issues in choropleth mapping, one of the most widely used methods to visually explore spatial distributions of demographic and socio-economic data. A new classification method is developed to explicitly integrate errors of estimation in the assessment of within-class variation and the associated groupings. The proposed method is applied to mapping the 2009–2013 ACS estimates of median household income at various scales. Results are compared with those generated using existing classification methods to demonstrate the effectiveness of the new classification scheme.

1. Introduction

The U.S. Census Bureau initiated the operational testing of the American Community Survey (ACS) in 1995 to provide continuous and timely demographic and socio-economic data that were collected by the long form questionnaire of the decennial census. With the first full implementation of ACS in 2005, the nationwide data were first made available in 2006 and since then they have been increasingly relied upon in many planning and decision making contexts due to its timely and current estimates (MacDonald, 2006; Sun & Wong, 2010). Starting from 2010, the ACS completely replaced the decennial long form and became the primary data source for detailed characteristics of the U.S. population. A considerable increase in the use of ACS data can be expected in the future.

On the other hand, the timely, annual ACS estimates present great challenges to data sampling and inferences to be made based on the data. Given the limited budget and time constraints, the ACS utilizes a much smaller sample than the decennial long form, resulting in a potentially much larger sampling error (MacDonald, 2006; Spielman, Folch, & Nagle, 2014). As an example, Starsinic (2005) found that sampling errors of ACS estimates are generally 75% larger than those of the decennial long form at the census tract level. To support the assessment of the reliability of ACS estimates, the U.S. Census Bureau publishes a margin of error (MOE) at the 90% confidence level alongside each estimate. A growing number of literature have also recognized the issue and discussed the uncertainty involved, highlighting the necessity for ACS data users to understand the data quality and its implications in future analysis and inferences (MacDonald, 2006; Citro & Kalton, 2007; Bazuin & Fraser, 2013; Folch, Arribas-Bel, Koschinsky, & Spielman, 2014; Spielman et al., 2014).

This article focuses on addressing ACS data uncertainty issues in choropleth mapping, one of the most widely used methods to visualize and explore spatial distributions of demographic and socio-economic data (Armstrong, Xiao, & Bennett, 2003; Sun, Wong, & Kronenfeld, 2015). The Census Bureau, as an example, now hosts a Web Service allowing users to interactively map various census data including ACS estimates. While the use of choropleth maps to present the ACS estimates is extensive, the quality of estimates is mostly disregarded in the process. Sun and Wong (2010) have demonstrated that overlooking the ACS data uncertainty can result in biased or erroneous map patterns.

In order to more accurately present spatial distributions of the ACS estimates, a new map classification method is developed by explicitly integrating estimation errors in determining the best groupings for

* Corresponding author at: University of Utah, Department of Geography, 332 S 1400 E, Rm. 217 Salt Lake City, UT 84112, USA.
E-mail address: ran.wei@geog.utah.edu (R. Wei).

http://dx.doi.org/10.1016/j.compenvurbsys.2016.04.003
0198-9715/

the estimates. The next section reviews existing map classification approaches. A similarity measure of ACS estimates under data uncertainty is then formally structured. Following this, an optimization model that minimizes within-class variation is presented along with a heuristic algorithm to solve the problem. The proposed method is applied to mapping the 2009–2013 ACS estimates of median household income in Utah at the census tract level and county level. Results are compared with those generated using existing classification methods to highlight the effectiveness and efficiency of the new classification scheme.

2. Background

Choropleth mapping is an important exploratory spatial data analysis (ESDA) technique and has been extensively used to visually explore the spatial pattern of attribute distributions across a region (Anselin, 1999). As an essential procedure in choropleth mapping, determining class intervals to suitably group spatial units has attracted significant attention from the cartography and GIS community (Brewer & Pickle, 2002; Armstrong et al., 2003). A variety of classification methods have been developed and detailed reviews can be found in Robinson, Morrison, Muehrcke, Kimerling, and Guptill (1995); Murray and Shyy (2000) and Brewer and Pickle (2002).

Among various classification methods, there are three most widely used categories, including equal intervals, equal frequencies and statistically optimal classification (Robinson et al., 1995; Andrienko, Andrienko, & Savinov, 2001). Equal interval methods divide the overall range of attribute values into multiple equally sized intervals, while equal frequency methods place the same number of spatial units into each class (Robinson et al., 1995). Statistically optimal classification methods determine class breaks by optimizing one or more statistical properties. The Jenks natural breaks method is one of most highly regarded approaches; it identifies groups by minimizing the overall absolute deviation of each attribute value from the corresponding group mean (Jenks & Caspall, 1971). Traun and Loidl (2012) extended the Jenks natural breaks classification method to take into account spatial autocorrelation. Cromley (1996) proposed several additional statistical formulas, such as minimizing the sum of squared deviations between attribute values and the group mean or the group median, minimizing the maximum attribute value deviation, and minimizing the boundary error associated with the attribute value deviations of adjacent units. While previous classification studies focused on optimizing a single property, Murray and Shyy (2000) and Armstrong et al. (2003) examined the optimization of multiple criteria simultaneously. The idea of all these statistically optimal classification methods is to identify class breaks that give the highest within-class homogeneity so that spatial patterns can be best highlighted in choropleth mapping, though the criteria used to define homogeneity may vary in different studies. New classification schemes have also been developed for data with specific characteristics, such as head/tail breaks method for data with a heavy-tailed distribution (Jiang, 2013) and concentration-based classification scheme for rate data (Cromley, Zhang, & Vorotyntseva, 2015).

While considerable research efforts have been made to develop classification schemes, only a few studies explicitly account for data uncertainty in map classification. Xiao, Calder, and Armstrong (2007) developed a statistical measure to examine the impacts of data uncertainty on the robustness of classification schemes. In the study, uncertainty existed an observed attribute value and it was assumed to follow a certain probability distribution. The probability of the actual value falling into the range of assigned class is computed for each spatial unit and then used to measure the robustness of the classification of each unit. This probability measure is statistically meaningful and can be used to evaluate the reliability of any classification scheme. In the article the robustness of equal interval, quantile and natural breaks methods were assessed but how to directly incorporate data uncertainty into the computation of a classification scheme remains unsolved.

A further step was taken by Sun and Wong (2010) who proposed a data-driven classification method that explicitly takes into account data uncertainty. This method was further enhanced by Sun et al. (2015) and referred to as the class separability classification method. The class separability approach incorporated the uncertainty associated with each observation and introduced a separability metric based on a statistical assessment of the difference between two units. They defined the separability between two classes as the minimum separability among all combinations of two individual units with each coming from a different class. A heuristic method was also developed to identify class breaks that maximize the separability between adjacent classes. This algorithm started by sorting the observed attribute values in the ascending order. Then for each potential break, the units falling into the left side of the break were grouped into one class and those into the right side of the break were placed into another. The separability between these two classes is computed for each potential break, and the $p-1$ breaks that lead to highest separability are selected as the final class breaks assuming p classes are needed.

This approach is intriguing as it explicitly integrates data uncertainty into the determination of class breaks. However, it does not take into account within-class homogeneity, one of the most important criteria for map classification as described above. Without the consideration of within-class homogeneity, units whose attribute values are significantly different might be grouped into one class, weakening the capability of choropleth mapping in presenting meaningful groups. Thus a new classification method is needed to account for within-class homogeneity to address map classification under data uncertainty.

Table 1
Descriptive statistics of the three data sets.

Data sets	Number of units	Average mean value of estimates ($)	Average MOE ($)	Average CV
Utah counties	29	53,591	3111	3.77%
SLC tracts	210/212*	64,268	9173	8.97%
U.S. counties	3109	45,756	2888	4.10%

* There are 212 census tracts in Salt Lake county, Utah, but no median household income estimates were reported for two tracts due to zero population.

3. Methodology

In order to explicitly integrate data uncertainty into map classification, we first proposed a measure to assess the similarity between ACS estimates under uncertainty, and then structured an optimization model that can minimize the total within-class variation. A heuristic algorithm is also developed to solve the model.

3.1. Similarity measure

Similarity measure that quantifies to what extent two observed attribute values are similar or different is essential for classification approaches. When the observed attribute values to be grouped are accurate with no uncertainty, the absolute or squared deviation between attribute values are commonly utilized as similarity measure (see Jenks & Caspall, 1971; Cromley, 1996; Murray & Shyy, 2000). However, when the observed attribute values to be mapped are highly uncertain, such as ACS estimates, the absolute or squared deviation of observed values is no longer valid as the true value might be different from the observed one. In this case, due to the uncertainty, our observation can be assumed to be a random variable following a certain probability distribution. The similarity between two uncertain attribute values can be considered as a similarity between two probability distributions instead of similarity between two mean values.

The Bhattacharyya distance developed by Bhattacharyya (1946) is a widely used similarity measure for probability distributions (Kailath,

Please cite this article as: Wei, R., et al., An integrated classification scheme for mapping estimates and errors of estimation from the american community survey, *Computers, Environment and Urban Systems* (2016), http://dx.doi.org/10.1016/j.compenvurbsys.2016.04.003