An integrated classification scheme for mapping estimates and errors of estimation from the American Community Survey

Ran Wei[1]

Daoqin Tong[2]

Jeff M. Phillips[3]

[1]Department of Geography, University of Utah, Salt Lake City, UT, USA 84112;
Email: ran.wei@geog.utah.edu
[2]School of Geography and Development, University of Arizona, Tucson, AZ, USA 85721
[3]School of Computing, University of Utah, Salt Lake City, UT, USA 84112;

March 5, 2016

Thanks to supported by NSF CCF-1350888, IIS-1251019, ACI-1443046, and CNS-1514520.

An integrated classification scheme for mapping estimates and errors of estimation from the American Community Survey

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.

Keywords: classification; uncertainty; ACS

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 and 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 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 et al. 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 and Kalton 2007; Bazuin and Frazer 2013; Folch et al. 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 et al. 2003; Sun et al. 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 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.

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 and Pickle 2002; Armstrong et al. 2003). A variety of classification methods have been developed and detailed reviews can be found in Robinson et al. (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 et al. 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 and 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 et al. 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 et al. (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.

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.

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 and Caspall 1971; Cromley 1996; Murray and 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 Bhattacharya (1946) is a widely used similarity measure for probability distributions (Kailath 1967; Basseville 1989). Given two probability distributions $u(x)$ and $v(x)$, the Bhattacharyya distance between them is defined as:

$$B(u,v) = -\,ln \int \sqrt{u(x)v(x)}dx \qquad (1)$$

$0 \leq B(u,v) \leq \infty$. When $B(u,v) = 0$, the two distributions are identical. The larger the Bhattacharyya distance is, the more dissimilar the two distributions are. The Bhattacharyya distance can be used to measure similarity between any discrete or continuous probability distributions. Assuming that the two observed attribute values, i and j, conform to two normal distributions, $N_i(\mu_i, \sigma_i^2)$ and $N_j(\mu_j, \sigma_j^2)$, respectively, the Bhattacharyya distance can be calculated as (Nielson and Boltz 2011):

$$B(N_i, N_j) = \frac{1}{4}\frac{(\mu_i - \mu_j)^2}{\sigma_i^2 + \sigma_j^2} + \frac{1}{4}ln(\frac{1}{4}\left(\frac{\sigma_i^2}{\sigma_j^2} + \frac{\sigma_j^2}{\sigma_i^2} + 2\right)) \qquad (2)$$

where μ_i and μ_j are the means of the attribute at units i and j, and σ_i and σ_j are the corresponding standard deviation or the uncertainty level associated with an observation. The first term is associated with the mean difference, while the second term accounts for the variance difference. When the means are equal or observations at two units are similar, the first term will be zero or small, and more dissimilarity will be mainly associated with variances that differ significantly. The second term becomes zero if the variances are equal, and the dissimilarity will be introduced by the mean difference. A large mean difference along with small variances (less uncertainty) leads to large dissimilarity. Thus, the Bhattacharyya distance takes into account both the mean and variance differences and provides an overall measurement of the similarity between two normal distributions.

As the means and standard errors of ACS estimates are provided by the Census Bureau, and the sampling errors are commonly assumed to conform to normal distributions (see Sun et al. 2015), we can use the Bhattacharyya distance between normal distributions, equation (2), to measure the similarity between ACS estimates.

Minimizing within-class variation

Based on the Bhattacharyya distance similarity measure discussed previously, the overall within-class variation can be evaluated while considering the uncertainty associated with each estimate. In particular, we considered the sum of the Bhattacharyya distance of each member in the class from the class median. The class median, also referred to as medoid in Kaufman and Rousseuw (1990), is a representative observation whose average distance to all members of the class is minimal among all class members (Murray and Estivill-Castro 1998). Given this, an optimization model is structured that minimizes the total within-class variation of ACS estimates. Consider the following additional notation:

i, j = index of estimates

p = number of classes to be identified

$$X_j = \begin{cases} 1, & \text{if estimate } j \text{ is selected as class median;} \\ 0, & \text{otherwise;} \end{cases}$$

$$Z_{ij} = \begin{cases} 1, & \text{if estimate } i \text{ is assigned to class } j; \\ 0, & \text{otherwise;} \end{cases}$$

With this notation, a clustering or classification model that minimizes within-class variation under data uncertainty is formulated as follows:

Median Clustering Problem-Uncertainty (MCP-U)

$$Minimize \quad \sum_i \sum_j B(N_i, N_j) Z_{ij} \tag{3}$$

$$Subject\ to \quad \sum_j Z_{ij} = 1, \forall i \tag{4}$$

$$Z_{ij} \leq X_j, \forall i, j \tag{5}$$

$$\sum_j X_j = p \tag{6}$$

$$Z_{ij} = \{0,1\}, \forall i, j \tag{7}$$

$$X_j = \{0,1\}, \forall j$$

The objective, (3), is to minimize the total Bhattacharyya distance between estimates and corresponding class medians. Constraints (4) ensure that each estimate is assigned to one class. Constraints (5) require that an estimate can only be assigned to a class j if estimate j is selected as class median. Constraint (6) specifies that p classes will be identified. Constraints (7) impose binary restrictions on decision variables.

The MCP-U is an extension of p-median problem or standard median clustering problem (MCP) proposed by ReVelle and Swain (1970) as well as Murray and Estivill-Castro (1998). When the MCP is applied to choropleth classification as detailed in Cromley (1996) and Murray and Shyy

(2000), the objective is to minimize the total squared or absolute deviations between mean values, while the MCP-U aims to minimize the Bhattacharyya distances that account for the differences in both estimates and the associated variances (or the uncertainty). As the Bhattacharyya distances can be computed in advance and utilized as the input for the MCP-U, existing solution approaches for the standard MCP can also be employed to solve the MCP-U. Two solution approaches have been widely used to solve the MCP. One is an optimal or exact solution method using the linear-programming based branch-and-bound algorithm, and the other is an approximate or heuristic technique known as interchange algorithm or k-medoid algorithm (Kaufman and Rousseeuw 1990; Murray and Estivill-Castro 1998).

Due to the integration of data uncertainty, classes identified by optimally solving the MCP-U might overlap when mean values of estimates are used to generate the map legend and to display the results, which is also the typical method used in choropleth mapping. Such overlap theoretically makes sense as the classification is no longer a univariate classification after taking into account the estimate variances to address uncertainty, but overlapping class boundaries create ambiguities for map interpretation and also violate the traditional rules for choropleth mapping (Jenks and Coulson 1963). A new interchange heuristic is therefore developed to efficiently solve the MCP-U while maintaining the crisp class boundaries of the mean values of estimates (or the reported ACS estimates). The algorithm is as follows:

Interchange Heuristic for MCP-U

1. Sort n estimates based on the mean values of estimates, $a_1, a_2, \dots, a_n$, in ascending order.

2. Randomly select $p - 1$ class breaks, $b_1, b_2, \dots, b_{p-1}$, which are also in ascending order.

3. For each break b_j, swap b_j with a_i which are larger than b_{j-1} and smaller than b_{j+1}.

4. Identify the class median for each class and compute the objective (3).

5. If the objective (3) is improved, a_i replaces b_j as new class break for class j.

6. Repeat steps 3-5 until the objective (3) cannot be improved by interchanging.

7. A locally optimal solution is identified.

This heuristic starts by randomly selecting $p - 1$ class breaks and then improves the total within-class variation by interchanging the class breaks while maintaining crisp class boundaries. The crisp class boundaries are maintained by only allowing the interchange to take place between the selected break (b_j) and estimates with mean values that are larger than break b_{j-1} ($< b_j$) and smaller than larger break b_{j+1} ($> b_j$), as specified in step 3. Similar to the generic interchange algorithm as a heuristic approach (see Kaufman and Rousseeuw 1990), it is necessary to run the new interchange heuristic many times, each time starting with a different set of random class breaks, to avoid the bias associated with the initial class breaks. Previous studies suggested that a considerable number of random runs are likely to generate solutions of high quality for the MCP (Hansen and Mladenovic 1997; Church and Murray 2009).

While the objective (3) describes the total within-class variation under data uncertainty, it is difficult to use it as a direct measure for assessing classification quality as different data sets could result in significantly different total within-class variations. It is therefore necessary to develop a standardized measure, tabular accuracy index-uncertainty (TAI-U), to quantify the degree of class homogeneity under data uncertainty:

$$\text{TAI-U} = 1 - \frac{\sum_i \sum_j B(N_i, N_j) Z_{ij}}{\sum_i B(N_i, N_k)} \qquad (8)$$

where k is the median of all estimates. $\sum_i B(N_i, N_k)$ is the sum of Bhattacharyya distance between each individual estimate and the overall median, so the TAI-U is essentially a comparison of variation within classes to the total variation of all estimates. Given this, the TAI-U can be considered as an extension of the traditional measure for class homogeneity, the tabular accuracy index (TAI) proposed by Jenks and Caspall (1971). Similar to the TAI, the TAI-U range from 0 to 1. The higher the TAI-U is, the more homogeneous the classes are.

Application results

The proposed classification method is implemented in Python and executed on a Intel Core i7-4770 (3.40 GHz) computer running Windows with 16 GB RAM. Choropleth maps were generated using the proposed interchange heuristic as well as the existing schemes including class separability, natural breaks, quantile, and equal interval. The 2009-2013 ACS median household income estimates at three scales were used in the test, the Utah counties, census tracts in Salt Lake County, Utah, and the U.S. counties.

Table 1 presents some descriptive statistics of the three data sets. The MOE at the 90% confidence level is provided by the Census Bureau. In addition to the MOE, the Census Bureau commonly uses the coefficient of variation (CV) to assess the reliability of ACS estimates. The CV is computed as the ratio of the standard error to the estimate mean. The average CV of census tracts is 2.4 times larger than that of counties, indicating that estimates at the census tract level is less reliable than those at the county level. This is not a surprise given the much smaller sample size used for data collection in small geographic units. The average CV of U.S. counties is slightly larger than that of Utah counties, but much larger variation of CV can be observed for the U.S. counties given sampling variation across the U.S. The bivariate choropleth maps of the three data sets are generated using the natural break classification method and shown in Figure 1, where the variations of the median household income and the corresponding CVs are displayed.

Table 1: Descriptive statistics of the three data sets

Data sets	Number of units	Average mean	Average MOE	Average CV

		value of estimates ($)	($)	
Utah counties	29	53,591	3,111	3.77%
SLC tracts	210/212*	64,268	9,173	8.97%
U.S. counties	3109	45,756	2,888	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.

Each data set is classified using the interchange heuristic with the number of classes (p) varying from 3 to 10. The interchange heuristic was randomly initiated 1,000 times for each case. The corresponding TAI-Us are displayed in Figure 2. Clearly, the TAI-U decreases as p increases, which is common in classification as greater number of classes will lead to smaller amount of variation within classes. As a result, there is a trade-off between the number of classes and total within-class variation. Additionally, smaller data set is likely to have smaller within-class variation.

Other classification methods were also utilized to generate choropleth maps and the performance results for $p = 5$ and 6 are shown in Table 2. The interchange heuristic always produces the best TAI-U for the respective number of classes, p. As an example, when classifying the U.S. counties into five groups, the interchange heuristic has a TAI-U of 63%, which is greater than the next closest TAI-U found by the quantile approaches of 60%. While the class separability approach is designed for uncertain data classification, the within-class variations are quite large for medium and large data sets. Natural breaks generally produce good results in terms of the TAI-U, but equal interval and quantile might lead to classes with large within-class variations for some data sets.

Table 2: TAI-U of classification methods

Data sets	p	Interchange heuristic	Class separability	Natural breaks	Equal interval	Quantile
Utah counties	5	**88%**	88%	84%	80%	70%
	6	**91%**	89%	87%	81%	84%
SLC tracts	5	**78%**	18%	76%	70%	76%
	6	**83%**	55%	82%	70%	79%
U.S. counties	5	**63%**	4%	59%	42%	60%
	6	**68%**	4%	64%	47%	65%

The choropleth maps with five classes generated using various classification methods for Utah counties, Salt Lake county census tracts, and U.S. counties are displayed in Figures 3, 4, and 5, respectively. The maps produced by the interchange heuristic and class separability approaches for Utah counties are the same, but differ for those classified by natural breaks, equal interval and quantile. One significant difference is that the interchange heuristic groups almost all the southern counties together, while other approaches classify them into different groups. This is attributed to the high CV or uncertainty associated with the estimates in the southern counties. After taking into account such uncertainty, the variations between these estimates are no longer substantial and the new method grouped them into one class. This is also consistent with the local knowledge for Utah counties.

A similar pattern is also observed when comparing the interchange heuristic with the natural breaks approach for classifying the census tracts in Salt Lake County. The eastern census tracts are grouped together due to the integration of data uncertainty by the interchange heuristic. The U.S county maps produced by the interchange heuristic and natural breaks display an overall similar pattern, but local variations can also be found, such as the pacific coast of California, Montana and Wyoming. This shows that significant differences can be found in the spatial distribution even when the difference of TAI-U is not large.

When classifying the census tracts and U.S. counties using the class separability approach (see Figure 4b and Figure 5b), there is one dominant class containing over 95% (99% in the case of U.S. counties) of all the units, which explains the low TAI-U in Table 2. As the class separability approach classifies the data by ranking the separability between estimates, a dominant class will emerge in situations where there are a small number of outlier values, which is not uncommon in real data sets. More classes are probably needed to achieve better within-class homogeneity if the class separability approach is used to classify the estimates of medium or large data set.

Discussion and conclusions

The applications demonstrate that the interchange heuristic can consistently classify the ACS estimates into classes with the least within-class variations under data uncertainty. In addition to the better TAI-U criteria, the spatial distribution of resulting classes explicitly takes into account data uncertainty and could more accurately reflect the true spatial pattern of ACS estimates.

Worth further discussion is the restriction of non-overlapping class boundaries for displaying the mean values of ACS estimates. While the crisp class boundaries are maintained in the proposed interchange algorithm, it is possible to relax such constraints by allowing the interchange to take place among overlapping breaks. Optimally solving the MCP-U using exact solution approaches could result in overlapping classes as well. The overlapping classes make sense as the classification is no longer a univariate classification after taking into account data uncertainty. How to effectively visualize and comprehend overlapping classes remains a challenge and further research is needed before the adoption of overlapping classes.

While the MCP-U is proposed to classify the ACS estimates, it is also possible to use the model to identify clusters for the ACS estimates. Murray and Shyy (2000) have examined the relationship between clustering and classification approaches. When the MCP-U is used for clustering, there is no need to maintain crisp class boundaries. Relaxing such constraints will lead to clusters that have the least within-class variation.

In addition to the TAI-U, the average robustness measure proposed in Xiao et al. (2007) and average class separability measure proposed in Sun et al. (2015) have also been used in our study to evaluate the classification methods under uncertainty. The class separability approach generally achieves the highest average robustness and class separability for the tested data sets, similar to the results presented in Sun et al. (2015). This is associated with the existence of one dominant class (see Figure 4b and Figure 5b). The range of the dominant class is notably wider than the others, and the probability of an estimate falling into the class is high, resulting in high average robustness. As most units within the dominant class are quite far away from other classes due to the large value range of the dominant class, the average class separability is likely to be high as well. However, dominant classes should be avoided in choropleth mapping as they prevent the illustration of attribute variations. It is probably necessary to integrate the TAI-U when assessing a classification scheme under uncertainty to reflect the undesirability of dominant classes. Additionally, some trade-offs were observed between the three classification quality measures, including TAI-U, robustness and class separability. It is also worth noting that while the interchange heuristic did not explicitly take into account the between-class differences, the goal of minimizing within-class variations might naturally result in groups with large between-class differences as the natural breaks method performs. Therefore, given the specific classification purpose, it would be beneficial to explore such trade-offs using multi-objective optimization algorithms, but this remains for future research.

An important goal of map classification is to maximize class homogeneity so that similar values can be displayed using the same pattern/color and map users can visually explore spatial patterns. This article developed a new classification scheme that can maximize within-class homogeneity when the attribute values to be mapped are uncertain, such as ACS estimates. A standardized measure for within-class variation under uncertainty, the TAI-U, was also proposed to assess classification quality. Applications using three data sets of ACS estimates demonstrated the effectiveness of the developed classification approach and the TAI-U.

References

Andrienko, G., N. Andrienko & A. Savinov. 2001. Choropleth Maps: classification revisited. In *Proceedings ICA*, 1109-1219.

Anselin, L. 1999. Interactive techniques and exploratory spatial data analysis. In *Geographical Information Systems: Principles, Techniques, Management and Applications* ed. M. G. P. Longley, D. Maguire, and D. Rhind. Cambridge: Geoinformation Int.

Armstrong, M. P., N. Xiao & D. A. Bennett (2003) Using genetic algorithms to create multicriteria class intervals for choropleth maps. *Annals of the Association of American Geographers,* 93**,** 595-623.

Basseville, M. (1989) Distance measures for signal processing and pattern recognition. *Signal Processing,* 18**,** 349-369.

Bazuin, J. T. & J. C. Fraser (2013) How the ACS gets it wrong: The story of the American Community Survey and a small, inner city neighborhood. *Applied Geography,* 45**,** 292-302.

Bhattacharyya, A. (1946) On a measure of divergence between two multinomial populations. *Sankhyā: The Indian Journal of Statistics***,** 401-406.

Brewer, C. A. & L. Pickle (2002) Evaluation of methods for classifying epidemiological data on choropleth maps in series. *Annals of the Association of American Geographers,* 92**,** 662-681.

Church, R. L. & A. T. Murray. 2009. *Business Site Selection, Location Analysis and GIS*. Hoboken, NJ: John Wiley & Sons, INC.

Citro, C. F. & G. Kalton. 2007. *Using the American Community Survey: benefits and challenges*. National Academies Press.

Cromley, R. G. (1996) A comparison of optimal classification strategies for choroplethic displays of spatially aggregated data. *International Journal of Geographical Information Systems,* 10**,** 405-424.

Cromley, R. G., S. Zhang & N. Vorotyntseva (2015) A concentration-based approach to data classification for choropleth mapping. *International Journal of Geographical Information Science***,** 1-19.

Folch, D. C., D. Arribas-Bel, J. Koschinsky & S. E. Spielman (2014) Uncertain uncertainty: Spatial variation in the quality of American Community Survey estimates. *Working paper*.

Garey, M. R. & D. S. Johnson (1979) Computers and intractability: a guide to the theory of NP-completeness. 1979. *San Francisco, LA: Freeman*.

Hansen, P. & N. Mladenović (1997) Variable neighborhood search for the p-median. *Location Science,* 5**,** 207-226.

Jenks, G. F. & F. C. Caspall (1971) Error on choroplethic maps: definition, measurement, reduction. *Annals of the Association of American Geographers,* 61**,** 217-244.

Jenks, G. F. & M. R. Coulson (1963) Class intervals for statistical maps. *International Yearbook of Cartography,* 3**,** 119-134.

Jiang, B. (2013) Head/tail breaks: A new classification scheme for data with a heavy-tailed distribution. *The Professional Geographer,* 65**,** 482-494.

Kailath, T. (1967) The divergence and Bhattacharyya distance measures in signal selection. *Communication Technology, IEEE Transactions on,* 15**,** 52-60.

Kaufman, L. R. & P. Rousseeuw. 1990. *Finding groups in data: An introduction to cluster analysis*. NJ: John Wiley & Sons Inc.

MacDonald, H. (2006) The American Community Survey: Warmer (more current), but fuzzier (less precise) than the decennial census. *Journal of the American Planning Association,* 72**,** 491-503.

Murray, A. T. & V. Estivill-Castro (1998) Cluster discovery techniques for exploratory spatial data analysis. *International Journal of Geographical Information Science,* 12**,** 431-443.

Murray, A. T. & T.-K. Shyy (2000) Integrating attribute and space characteristics in choropleth display and spatial data mining. *International Journal of Geographical Information Science,* 14**,** 649-667.

ReVelle, C. S. & R. W. Swain (1970) Central facilities location. *Geographical analysis,* 2**,** 30-42.

Robinson, A. H., J. Morrison, P. C. Muehrcke, A. Kimerling & S. Guptill. 1995. *Elements of cartography*. New York, USA: John Wiley & Sons

Spielman, S. E., D. Folch & N. Nagle (2014) Patterns and causes of uncertainty in the American Community Survey. *Applied Geography,* 46**,** 147-157.

Starsinic, M. 2005. American Community Survey: Improving reliability for small area estimates. In *Proceedings of the 2005 joint statistical meetings on CD-ROM*, 3592-3599.

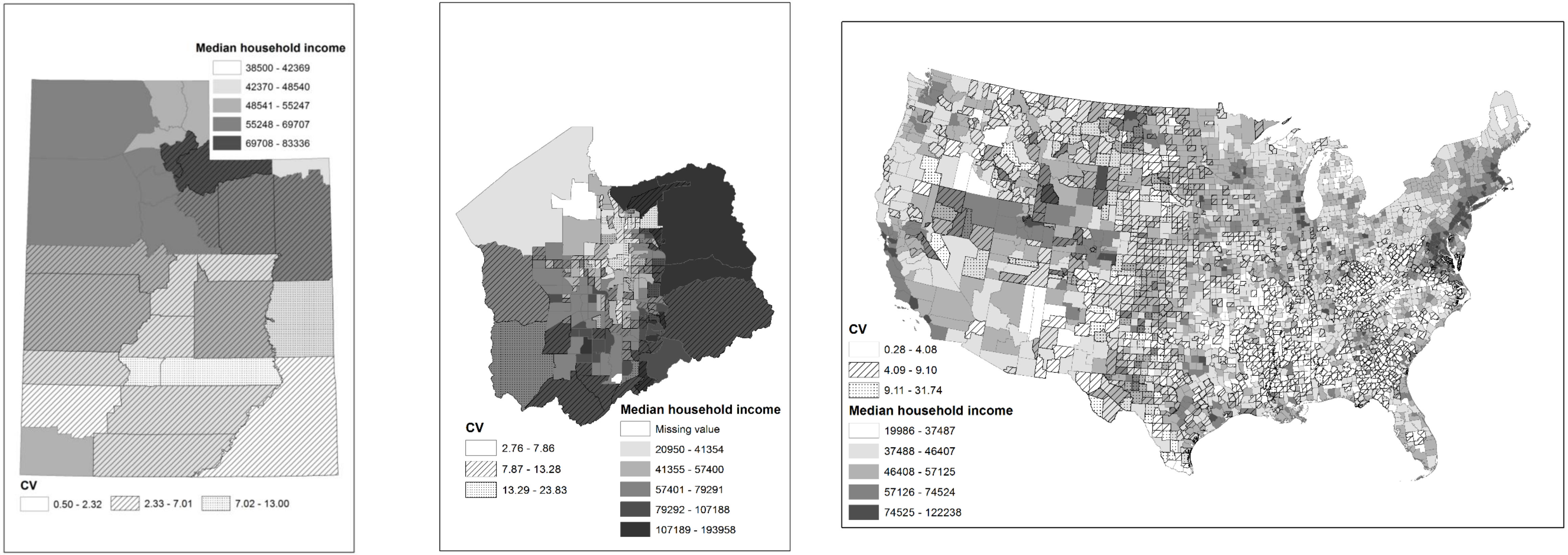
Sun, M. & D. W. Wong (2010) Incorporating data quality information in mapping American community survey data. *Cartography and Geographic Information Science,* 37**,** 285-299.

Sun, M., D. W. Wong & B. J. Kronenfeld (2015) A Classification Method for Choropleth Maps Incorporating Data Reliability Information. *The Professional Geographer,* 67**,** 72-83.

Traun, C. & M. Loidl (2012) Autocorrelation-Based Regioclassification–a self-calibrating classification approach for choropleth maps explicitly considering spatial autocorrelation. *International Journal of Geographical Information Science,* 26**,** 923-939.

Xiao, N., C. A. Calder & M. P. Armstrong (2007) Assessing the effect of attribute uncertainty on the robustness of choropleth map classification. *International Journal of Geographical Information Science,* 21**,** 121-144.

Figure



(a) (b) (c)

Figure 1: ACS estimates and coefficients of variations

(a) Utah counties

(b) Census tracts in Salt Lake County, Utah

(c) U.S. counties

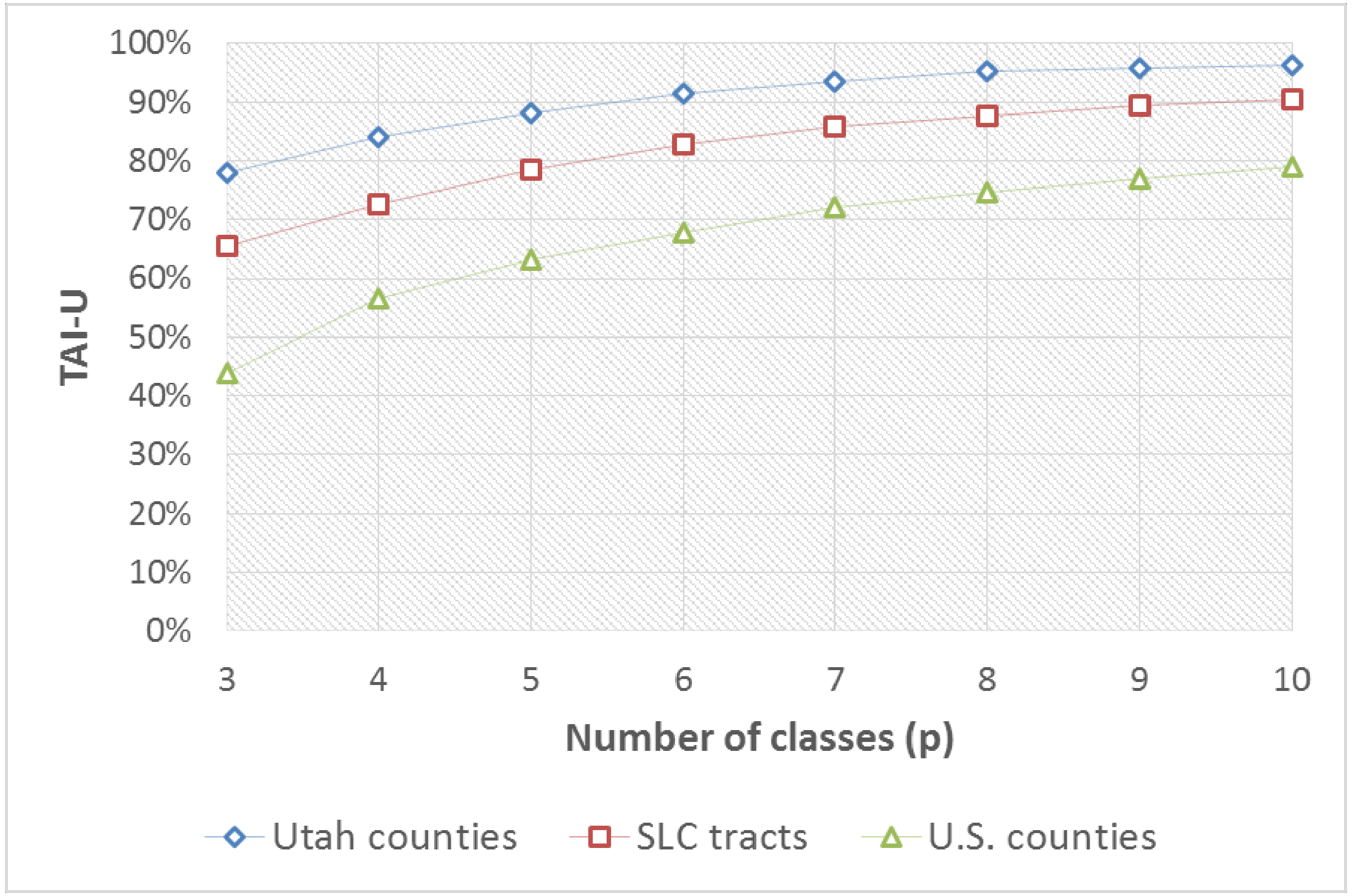
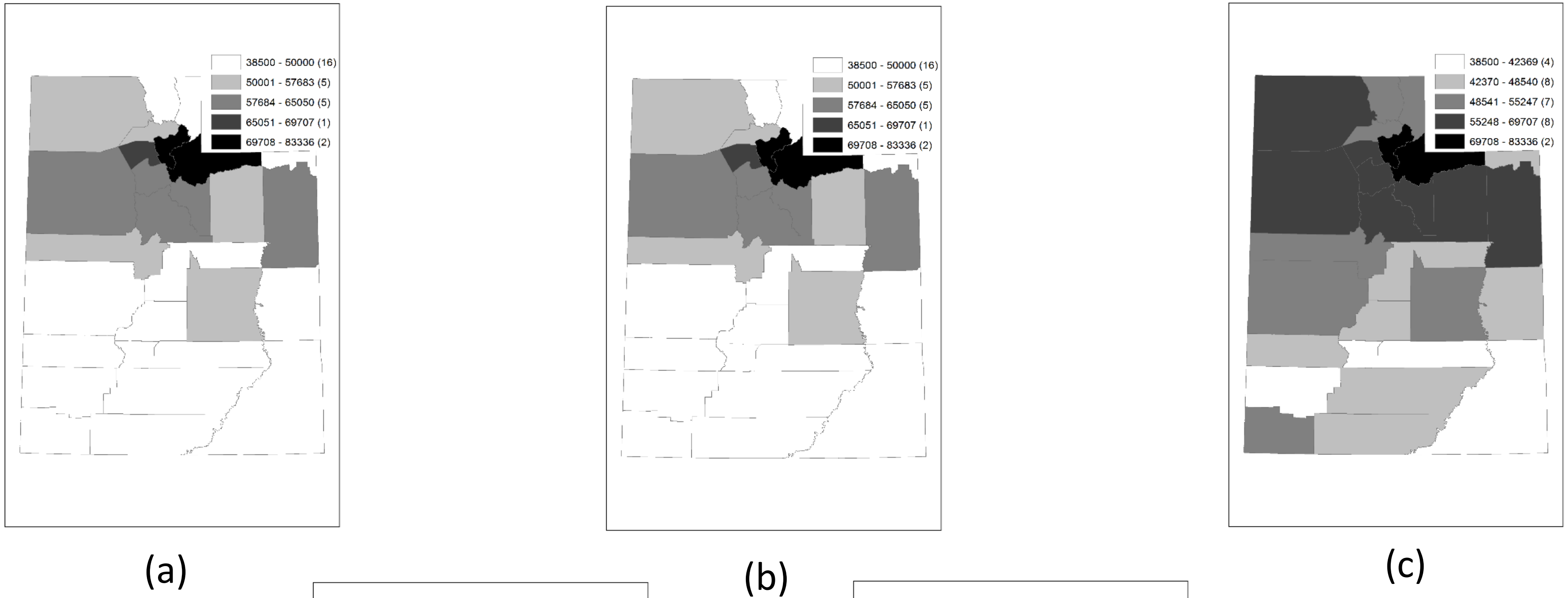


Figure 2: TAI-U of interchange heuristic



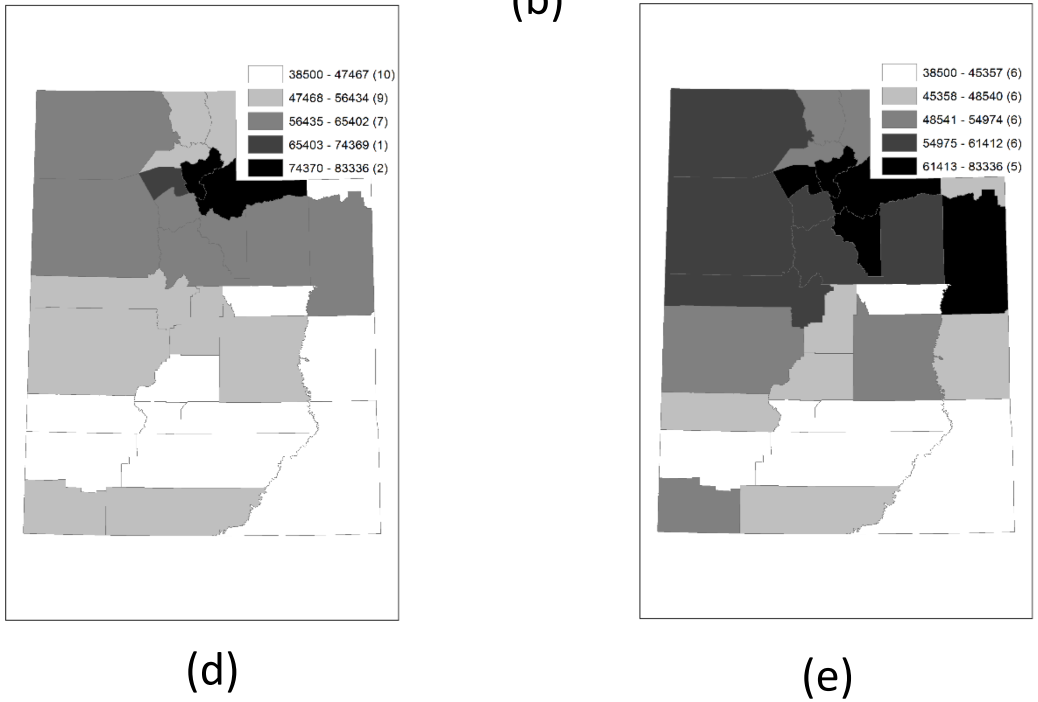
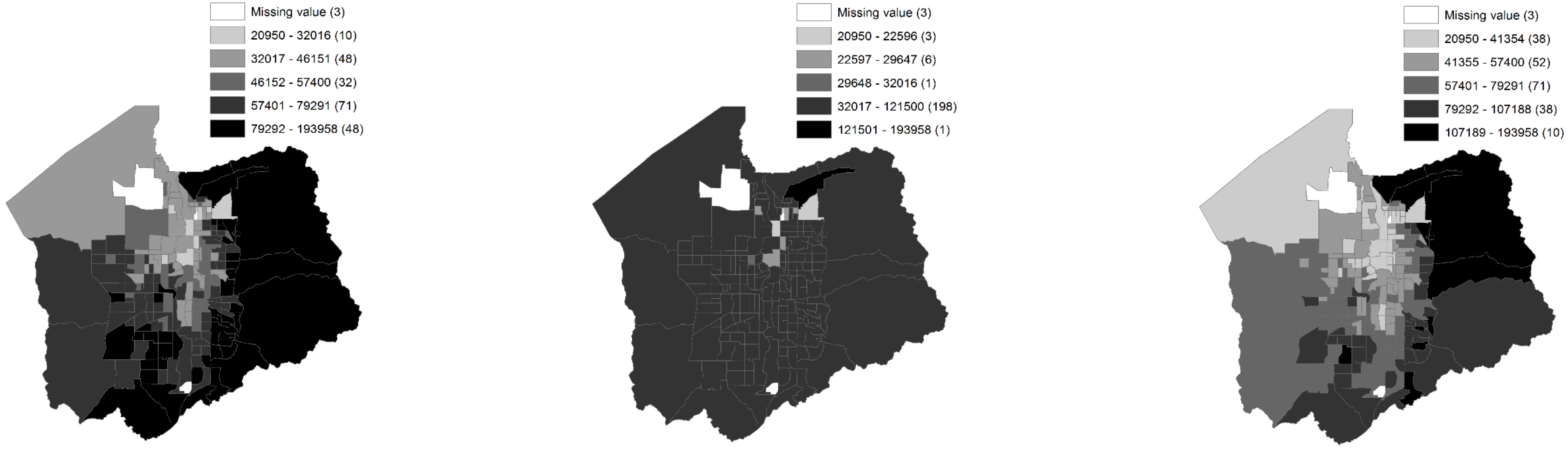
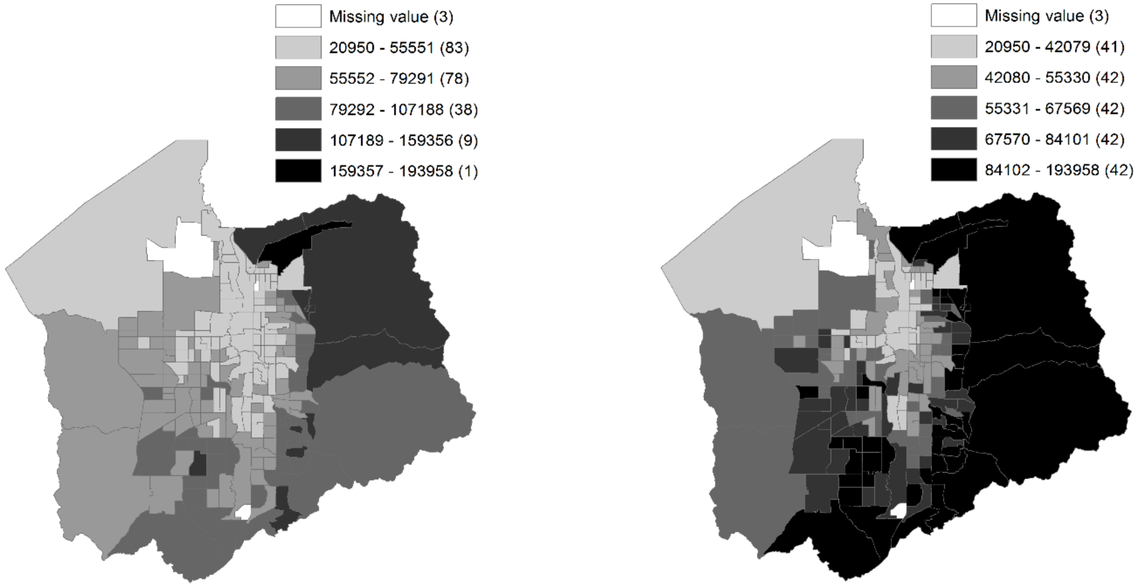


Figure 3: Choropleth maps of Utah counties (p=5)
(a) Interchange heuristic
(b) Class separability
(c) Natural breaks
(d) Equal intervals
(e) Quantile



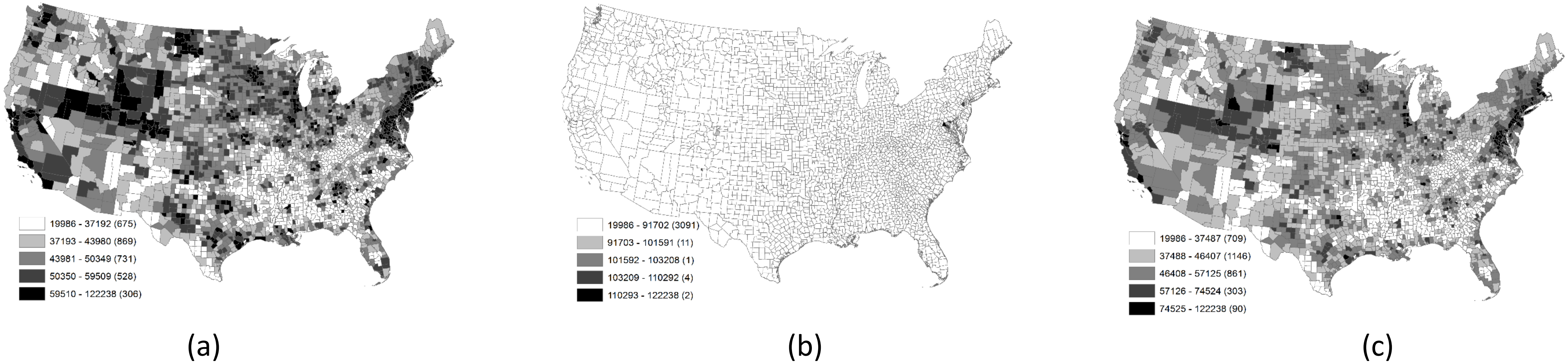
(a) (b) (c)



(d) (e)

Figure 4: Choropleth maps of census tracts in Salt Lake County (p=5)

(a) Interchange heuristic
(b) Class separability
(c) Natural breaks
(d) Equal intervals
(e) Quantile



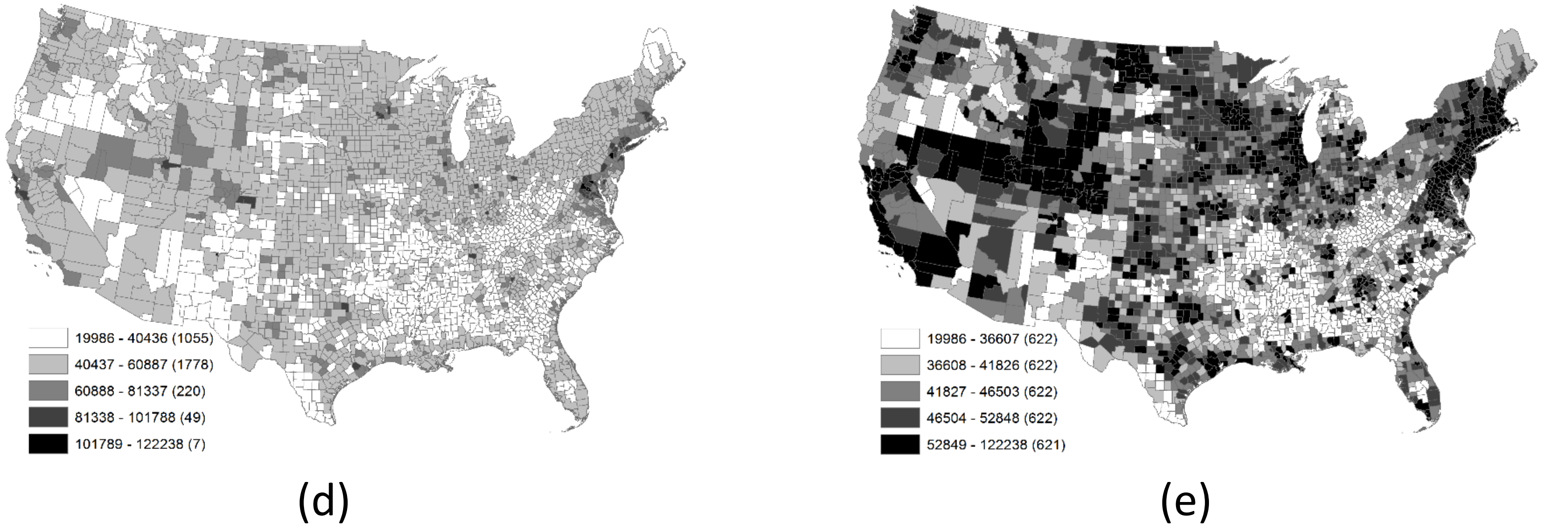


Figure 5: Choropleth maps of U.S. counties (p=5)
(a) Interchange heuristic
(b) Class separability
(c) Natural breaks
(d) Equal intervals
(e) Quantile