Chapter 16
Alternative Strategies for Mapping ACS Estimates and Error of Estimation

Joe Francis, Nij Tontisirin, Sutee Anantsuksomsri, Jan Vink and Viktor Zhong

Introduction

The beloved "long form" is dead. Long live the American Community Survey! As Eathington (2011) recently exclaimed, "Beginning in 2011, regional scientists and other socio-economic data users must finally come to terms with major changes in U.S. Census Bureau methodologies for collecting and disseminating socioeconomic data." Eathington's proclamation holds all the more true for today.

The American Community Survey (ACS) is now the primary mechanism for measuring detailed characteristics of the population at the sub-state level and especially smaller geographies like townships, places, and tracts. It is the main vehicle for disseminating information about educational attainment, occupational status, income levels (including poverty), and much more. As Sun and Wong (2010) write, "Census data have been widely used to support a variety of planning and decision making activities."

Additionally, during the past decade, there has been increasing interest among demographers, economists, planners, and regional scientists in mapping census data including the ACS. The main reason is that a map can show the spatial distribution of demographic data better than any other medium. Maps add another tool to the demographer's analytic toolbox.

Compared to the past, mapping has become an easier and more straightforward task. The widespread availability of desktop GIS systems and trained GIS professionals assures that an increasing amount of decennial, ACS, Small Area Income and Poverty Estimates (SAIPE), and other survey data will become mapped. The Census Bureau itself now routinely publishes reference maps and hosts an automated, interactive mapping service that can be invoked as part of ACS data display via the American Fact Finder. TIGERmap has been launched, for example. At the same time, mapping sample survey data like the ACS and SAIPE present significant

J. Francis (✉) · N. Tontisirin · S. Anantsuksomsri · J. Vink · V. Zhong
Program on Applied Demographics, Cornell University, Ithaca, NY, USA
e-mail: joe.francis@cornell.edu

M. N. Hoque, L. B. Potter (eds.), *Emerging Techniques in Applied Demography,* Applied Demography Series 4, DOI 10.1007/978-94-017-8990-5_16,

cartographic challenges in portraying visually both the estimates and the error of estimation on maps.

As the ACS begins its second iteration, with updated Census 2010 based geographies and new vintages of 1, 3, and 5 year ACS now available, demographers as well as geographers face quandaries about how to present such information to the intelligent public. Given that the ACS is such as a relatively small sample, particularly in the sub-county, tract. and block group geographies, it has become ever more pressing to assure the user of our research that the information is reliable. But, particularly where there is high uncertainty, to so signal that low data quality.

With the dissemination of the American Community Survey, the U.S. Census Bureau began to report forthrightly the uncertainty of its sample estimates by including not only the estimates for the various TIGER geographies, but also the accompanying error of estimation. Specifically, the ACS estimates are published with associated margins of error (MOE) representing a 90% confidence level under an assumed Gaussian distribution. By contrast, the published 2000 Census data tables did not include this information with the estimates for data from the "long form." Today demographers increasingly recognize that this uncertainty needs to be expressed to a potential user along with the estimates. But how?

This quandary is particularly evident in mapping ACS data. Unfortunately, the most prevalent practice at present is to largely ignore the unreliability of ACS estimates when mapping these data. But this needs to change if users of our maps are to place confidence in our map making.

Because of the small sample sizes, particularly in sub-state geographies, differences between different units or percent change across time for the same unit may appear to be significant when they are not. Moreover, for GIS analysts, this becomes a problem as the apparent emergence of a spatial pattern based upon these differences of ACS estimates among areal units may not be real, but may be the result of sampling errors instead. As every spatial demographer knows, determining whether differences are significant is crucial in analyzing the spatial distribution of some characteristic before drawing any conclusion other than that of spatial randomness. On a more technical level, cartographers recognize that uncertainty, as revealed by measures of error, can also affect even the seemingly simple process of determining optimal class intervals or boundaries in a thematic classification. That is, the determination of class boundaries may be influenced by the errors of estimates.

This paper is about exploring ways to improve communication of the estimates and reliability of ACS estimates in map making and in our published map products, whether in static "printable" form (e.g. a pdf) or in web based interactive delivery format (html, JavaScript, KML). First we will summarize geo-visualization developments over the past two decades on ways to present uncertain data. Second, we will present selected works of others more directly related to the ACS situation—polygons with attributes derived from a continuous monthly sampling activity and updated periodically. Third we will present some of our own work at the Cornell Program on Applied Demographics exploring how to communicate simultaneously both the ACS estimates and margins of error for polygons at the county and subcounty levels of geography.

General Approaches to Identifying and Dealing with GIS Data Error

All GIS data have error to some degree. There are many reasons—measurement errors, interpretation errors, classification errors, interpolation errors, generalization errors—with consequences like uncertain propagations and poor decisions. Indeed one current view among GIS scientists is that both spatial (positional) and attribute information have an *inherent* associated uncertainty. Zhang and Goodchild (2002) classify uncertainty in GIS work into two broad categories—positional accuracy and attribute accuracy—and group data errors involved into three categories—error, randomness, and vagueness. Census geography has positional uncertainty issues in the TIGER files—notably boundary accuracy, missing and misaligned streets, address uncertainty for non-city style address—but considerable improvement has taken place during the last decade and work is underway to continue improvement of both TIGER and MAF accuracy, as witnessed by the Geographic Support System initiative under the direction of the Census Bureau's Geography Division (http://www.census.gov/geo/www/gss/index.html). On the other hand, the ACS, along with the SAIPE and similar surveys, has attribute uncertainty that affects the data quality we deal with as well. The latter is the main concern of this paper.

In one sense, dealing with spatial accuracy and attribute uncertainty is still a fairly new and evolving area of study in GIS and analytic cartography. A quick background search revealed that it was not until the early 1990s that GIS users as a whole begin to take notice of spatial and attribute uncertainty. One can only speculate as to the reason, but perhaps it was a special issue of *Cartography and Geographic Information Science* that served to focus and spawn a research program on the topic of 'Geo-visualization' (MacEachren and Kraak 2001). In that special issue Fairbairn et al. (2001) proposed that representation of uncertainty was a key research challenge, and even went so far as to say that attribute uncertainty could be characterized as a "new" component of data. Fairbairn et al. (2001, p. 20) for example stated, "…that a representation of uncertainty may supplement existing data or may be an item of display in its own right…" Similarly, Pang (2001, p. 12), after reviewing the visualization developments of the 1990s, expressed the viewpoint that "visualizing the uncertainty in geo-spatial data is as important as the data itself… There is a lot of opportunity to further improve the current suite of uncertainty visualization techniques to meet this challenge. Particularly, in creating new visualization techniques that treat uncertainty as an integral element with the data."

Among the early approaches taken toward mapping uncertainty was via manipulating graphic "primitives" like color, transparency, line width, and sharpness or focus. Examples that fell under this category included Yee et al. (1992) work on varying contour widths depending on certainty. Dutton (1992) explored mapping uncertainty parameters to different points in HSV (*hue*, *saturation*, and *value*) space. Monmonier (1990) experimented with using cross hatches to express the degree of unreliability. Beard et al. (1991) investigated the inclusion of "fog," where the amount of haziness corresponds to the amount of uncertainty or the decreasing

amount of "focus" is represented by the amount of blurring as uncertainty increases. Pang et al. (1994) used the degree of transparency to indicate confidence in an interpolated field. Cedilnik and Rheingans (2000) utilized "perturbing" and "blurring" overlaid grid lines. Interestingly, these same kinds of approaches are being explored still today as ways to deal with ACS uncertainty in data estimates.

Meanwhile, on a related front in GIS statistics, Chrisman (1995) in discussing S. S. Steven's widely influential work on the levels of measurement expressed the belief that Steven's nominal, ordinal, interval, and ratio scales were not adequate for geography. He gave several examples where Steven's scheme, based on an implied linear measurement epistemology, falls apart for GIS data. Most convincing of his examples is the use of circular measurement in GIS where the distance from 0° to 1° is the same as from 359° to 0°, an outcome not expected under linear measurement. A second example of the inherent limitation of Steven's measurement scale system is the ability to reduce two linear orthogonal measures (the X-axis and the Y-axis) to a single scale using a radian angle measurement. Perhaps the most damning critique comes from Chrisman's pertinent illustration involving the commonly used multidimensional measurements on some spatial object in GIS analysis. He writes (1995, p. 275) "Multidimensional measurements create interactions not imagined in the simple linear world of Stevens. Since GIS is inherently multidimensional, the linear model limits our understanding concerning the interactions of measurements." Researchers who conduct spatial analyses or use spatial statistic procedures are very familiar with outcomes being a function of multidimensional interactions of measures.

One convergence of these developments in geovisualization and measurement theory was that GIS researchers began exploring the idea of using fuzzy classification as techniques for expressing uncertainty in the estimation. Burrough et al. (1997) expressed the belief that "… there is still a need for GIS methods to visually explore results of fuzzy classification." MacEachren and Kraak (1997), as well as Burrough and McDonnell (1998), advanced the notion that new visualization techniques were needed to allow users to explore uncertainty in spatial data *visually* and to investigate the effects of different decisions in the classification process. Their concern was that the common practice of presenting discrete classes of phenomenon like soils using sets of colors in a choropleth map was too constraining. They referred to this as a "double crisp" approach wherein (1) the features were drawn using sharp boundaries to delineate soil bodies and (2) crisp classes were used to classify the different types of soils. Yet they state the reality of the distribution of soils types across a landscape is that they weren't really as crisp as the choropleth map may indicate. Zhang and Goodchild (2002) also discuss the use of fuzzy classes in preference to rigidly defined classes for GIS work. The question of sharp versus fuzzy classification of values in the context of sampling and measurement uncertainty remains an issue we face today in dealing with ACS estimates in the presence of error of estimation.

Another of the new concepts advanced during this era was that of multiple membership maps, reflecting the multidimensional nature of geographic objects and how

to classify them. The notion was that multiple memberships were more complex than could be adequately handled by the traditional strict classification methods used in cartography (natural breaks, quantile, equal interval, defined interval, standard deviation, etc.). Instead multiple memberships could be better handled by means of different methods. Dovetailing with the work on fuzzy classification, the notion was that memberships based on multiple attributes should be derived using some continuous classification algorithm such as fuzzy k-means (DeGruijter and McBratney 1988). For example, Hengl et al. (2002), working with digital imagery, explored the use of pixel and color mixture as techniques to deal with visual fuzziness and uncertainty. Kardos et al. (2003) explored the value of hierarchical tree structures as a geovisualisation of attribute uncertainty technique. In their paper two such structures were compared: the region quadtree and the Hexagonal or Rhombus (HoR) quadtree, both variable resolution structures. The conclusion from these explorations was that an area where attribute data is uncertain will show less resolution through the data structure, whereas an area that is more certain will show greater resolution through the quadtree structures. While this work is a bit complex for the immediate concern of the present paper, their work is instructive on alternative approaches we might take to dealing with uncertainty of sample survey data like the ACS.

Another third set of developments came from working with digital imagery, where ideas centered on the notion that when we have multiple memberships for each pixel of a map, one could make conclusions about the *ambiguity,* i.e. *indistinctness,* of a specific class and overall *confusion* among all classes. Operationally, these researchers used what they called a confusion index to inspect confusion or fuzziness among multiple membership maps (Burrough and Frank 1997). Hengl's work along with Hootsmans (1996) was focused on color confusion as the means to create and detect fuzziness, but the idea of fuzzy boundaries isn't that far afield from Xiao's concerns with robustness in classification, which will be discussed later (Xiao et al. 2007)

In the early 2000s, analytic cartographers developed some of the new interactive data exploration techniques such as the use of slide bars, point and click events, blinking and animations as ways to give impressions of the amount of data uncertainty in estimates. The research on these techniques will be discussed in the next section of this paper because they relate more directly to handling uncertainty in ACS data.

While there has been considerable attention to issues surrounding attribute uncertainty over the past 20 years, a couple of general conclusions can be drawn from the review. One is that previous research provides a number of platforms on which to build in our efforts to portray estimate unreliability via maps of ACS data. Secondly, it is currently safe to say that uncertainty in spatial information is still an evolving field of GIS and analytic cartography. Evidence of this is the current dilemma we are facing in how to deal with errors of estimation in the ACS and related surveys.

Error of Estimation in the ACS

As mentioned, this paper is focused primarily on attribute accuracy or uncertainty in the ACS. Unfortunately, for the most part, communicating the data quality of ACS estimates have been either ignored or underplayed in maps of ACS data. Torrieri et al. (2011) present several examples of this pattern from the news media, governmental agencies, and academic research. One possible reason for this may be that, while guidelines for use of ACS data indicate the importance of indicating the measures of error along with the estimates, there is no consensus yet evolved as to a standard (or set of standards) for reporting the measures of error. Sun and Wong (2010, p. 287) note there have been various national committees that have deliberated on how to present this information but no standardization has emerged. Hence, this seems to signal the need for further exploration of alternative approaches, and this is the motivation for the discussion of various considerations that follow on how to handle error of estimation from survey data like the ACS and SAIPE. A number and a mixture of issues are involved.

While not an exhaustive classification, there seem to be at least nine major issues. One issue revolves around what to use as a measure of error, whether to use the traditional 90, 95 or 99 % statistical confidence interval under an assumed Gaussian theoretical distribution, or to employ a relative measure of error like the coefficient of variation. (A closely related issue should be whether a Gaussian distribution is always the most appropriate theoretical distribution to benchmark against, but given the generally large sample size and plethora of variables being estimated, this concern seems to have been conveniently either underplayed or ignored.) A second major issue surrounds the question of whether to present error of estimation in a separate map beside the map of estimates (the adjacency technique), or overlay them on the same map (the integrative technique) and use a "bivariate" legend to aid interpretation of patterns. A third major issue surrounds the rigidity of crisp classes for categorization of estimates in our choropleth maps given the errors of estimation and the likelihood that the "true" value of the variable placed arbitrarily in a given class may actually land in an adjacent class. Falling into the latter discussion is the more basic question of whether to present ACS data via classed or unclassed (unique values) thematic maps. A fourth major issue is the number of classes to employ for categorizing estimates in the face of uncertainty. A fifth concern is classification methodology. Symbolization of uncertainty is the sixth major concern. We will show various approaches to symbolization based on the principles espoused above plus use of cartograms. A seventh issue is whether it is better to present this combination of information via static maps (pdf's, jpeg's, eps, etc.) or through web based interactive maps (html, JavaScript, KML), which have more flexibility. Each of these issues is addressed below. Torrieri et al. (2011) identify an eighth issue specific to mapping ACS data for many geographic areas simultaneously. Ninth, there is the issue of map complexity and viewing audience.

Absolute vs. Relative Error

Regarding what to use as a measure of error as well as Li and Zhao (2005) argue for the using a relative error. Like other researchers, they recognize that absolute error measures are sensitive to the scale of the estimate. That is, the larger the estimate, X, the larger the value of an absolute error measure like the standard error (X). Wong and Sun's concern is that, because larger estimates have larger standard errors of estimates, researchers will draw inappropriate conclusions in comparing two or more estimates due to possible misinterpretations about the size of error–namely that attributes with large absolute error will draw the researcher/user to conclude, somewhat mindlessly, that the attribute has greater unreliability (and therefore shouldn't be used) rather than interpreting that error relative to the size of the estimate. On the other hand those who advocate for the uses of relative error measures assert that relative measures of error don't present this confusion and that the coefficient of variation is independent of the estimate scale. Li and Zhao (2005, pp. 2–3), using the definition that relative error is one that is simply relative to some referent, develop this idea further and propose three possibilities: the coefficient of variation, the estimate relative to measurement error, and a third class of relative errors where the estimation error of one estimator is relative to that of another estimator. We will concern ourselves in this paper only with the question of whether to use (1) an absolute measure of error like the traditional $\sigma(X)$ and a 90 % confidence interval, or (2) a relative error like the coefficient of variation (CV).

Our work at the Program on Applied Demographics leads us to conclude that the choice of whether to use a traditional confidence interval as the MOE or a relative one depends on the format of the variable being estimated. Specifically, while presenting information about counts like totals (e.g. number of housing units) and frequencies (occupied, vacant) or medians (e.g. median household income) and mean averages, then use of relative measures of error like the coefficient of variation (CV) seems more appropriate. But when representing information about proportions or percentages (e.g. percent Hispanic), or information about a ratio like the sex ratio, the standard confidence interval seems the more appropriate measure to employ. Likewise, when examining changes over time like the percent change in median housing costs, it may be better to use an "absolute" confidence interval rather than a relative measure of error. In short, one shouldn't mindlessly employ a relative error measure either.

Because these estimates are bounded by 0 to 1 in the case of proportions, or 0 to 100 in the case of percentages, the CV presents misinterpretation problems that are avoided when the traditional confidence interval is employed instead. To illustrate, consider a variable like percent foreign born. Let's say we estimate for a given geographic unit that 10 % of the population in a minor civil division is foreign born and is reported to have a MOE of ± 8 %. Here we have an estimate, p, with a certain standard error. On the other side of the dichotomy, we have an estimate of 1-p or 90 % native born with the same MOE of ± 8 %. These two facts are structurally

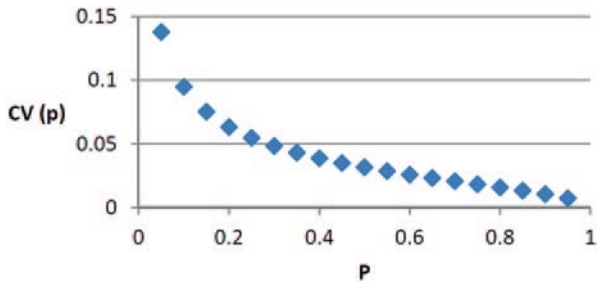

Fig. 16.1 Distribution of CV against percentage value (P)

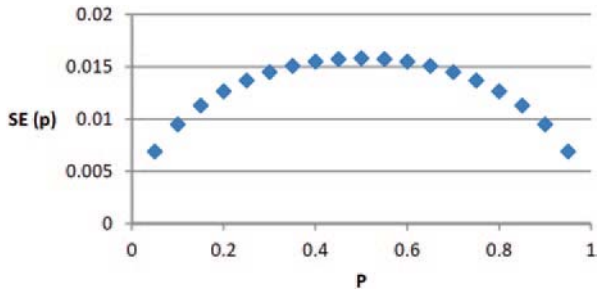

Fig. 16.2 Distribution of standard error of P against percentage value (P)

equivalent estimates, at least on the face of it, but when you calculate the CVs, the CV for the 10% foreign born is 48% (very unreliable), while the CV for the 90% native born is 5% (very reliable). Does this make sense?

The reason for this anomaly is the nature of the distribution of the CV as shown in the nearby plot (Fig. 16.1), which shows the declining, nonlinear relationship between the estimate and the CV. As the plot indicates, the smaller the estimated p value, the larger the coefficient of variation, while the larger the estimate, the smaller the coefficient of variation. Hence, even though the above two possible estimates of foreign born are equivalent structurally because they are two sides of a dichotomy, their certainty appears to be very different.

On the other hand, for variables like this, the confidence interval performs as one would expect. See Fig. 16.2. For both the estimate of $p = 10\%$ foreign born and $q = 90\%$ native born, the standard error of estimate is the same, approximately 0.01 when n = 1,000. This symmetry for placing a confidence bound on the estimate makes more sense both intuitively and statistically to us compared to a nonlinear relative error measure like the CV.

For our work, when presenting information like totals or frequencies, or summary statistics like medians and mean averages, we prefer to use relative measures of error like the coefficient of variation (CV). See Fig. 16.3, where the first map reflects Sun and Wong's cross-hatching approach with three data-driven categories, and the second is one showing our exploratory work at PAD on developing legends where category limits are set manually to values more meaningful to researchers and policy workers—0–15, 15.1–30, 30.1–60, and more than 60%.

However, when presenting information like proportions or percentages, we prefer to use absolute measures of error like the traditional standard error of estimate

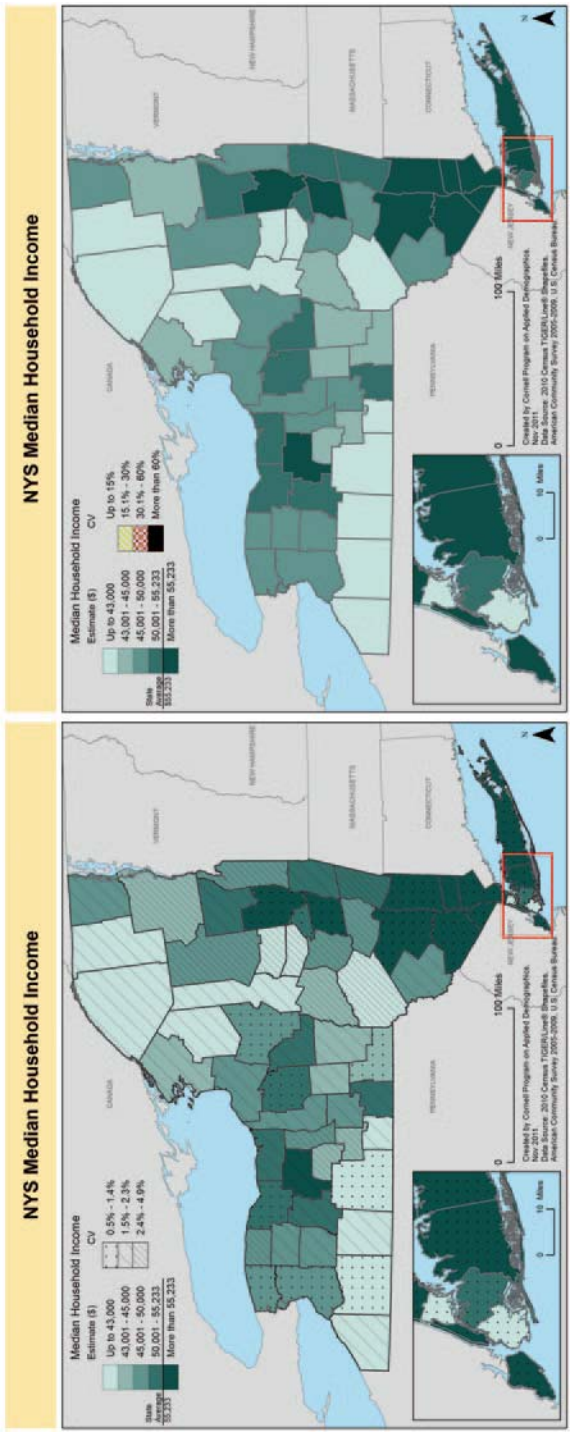

Fig. 16.3 Comparison of data driven category legends vs. researcher developed category legends

and confidence interval. See Fig. 16.4, where the first map represents the Sun and Wong crosshatching approach and the second shows some of our exploratory work with legends at the Program on Applied Demographics, here using cross hatching and color hue to differentiate

Side-by-Side Maps vs. Overlay Maps

A second major issue faced by spatial demographers and GIS analysts is whether to present error of estimation in a separate map beside the map of estimates (two map, side-by-side technique), or overlay them on the same map and use a "bivariate" legend to aid interpretation of patterns (single, integrated map technique). Compare the map layout in Fig. 16.5 for a side-by-side arrangement and Fig. 16.6 for an example of an overlay map.

MacEachren and colleagues have probably conducted the most extensive exploration of this question and have concluded that the single integrated map approach, wherein estimates are presented via color coding on a choropleth map and uncertainty of these estimates symbolized by cross-hatching, works better. Sun and Wong (2010, pp. 290–291) build on these results and present illustrations of both approaches, using New Jersey counties. In our own work, we found that both approaches were taxing to absorb for the general educated public viewer. However, for those more experienced with maps, the single integrated map was preferred.

We will address the issue of symbolization for these maps below in a later section, but in our work we find that viewers do not prefer (even dislike) the use of cross-hatching to portray uncertainty because they felt it obstructed viewing and understanding the classification of the estimate for the county or sub-county unit of geography. Compare for example the clarity of the maps in Fig. 16.7, which uses a less obstructive symbolization of error, with those in Fig. 16.6.

This judgment of cross-hatching producing an obstructed view was particularly true for maps with lots of geographic units being displayed simultaneously, like the 1,000-plus minor civil divisions (towns, cities, reservations) in New York.

Crisp vs. Modified Classes

A third major issue surrounds the structure of classes used to group estimates. For continuous attributes cartographers typically group values, for purposes of display on a choropleth map, into 4–6 classes using a classification methodology like natural breaks, equal interval, equal size (quantile), standard deviation, or some similar scheme. Doing so always raises the question of arbitrariness and rigidity of crisp class boundaries. Are these the optimal boundaries? How much alike are the spatial objects falling into a given class compared to those in an adjacent class? The latter question is of particularly high valence for spatial features "at the boundaries of the class." These questions take on an extraordinary relevance for categorization

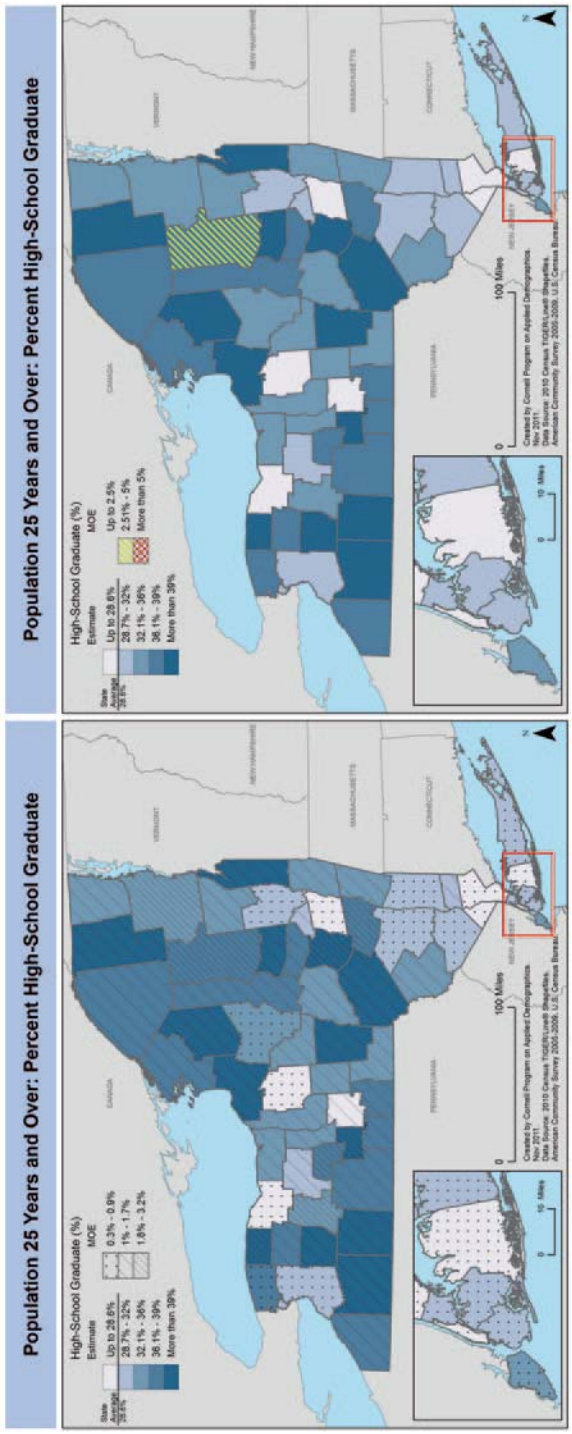

Fig. 16.4 Percent high-school graduates estimates with different MOE legend styles

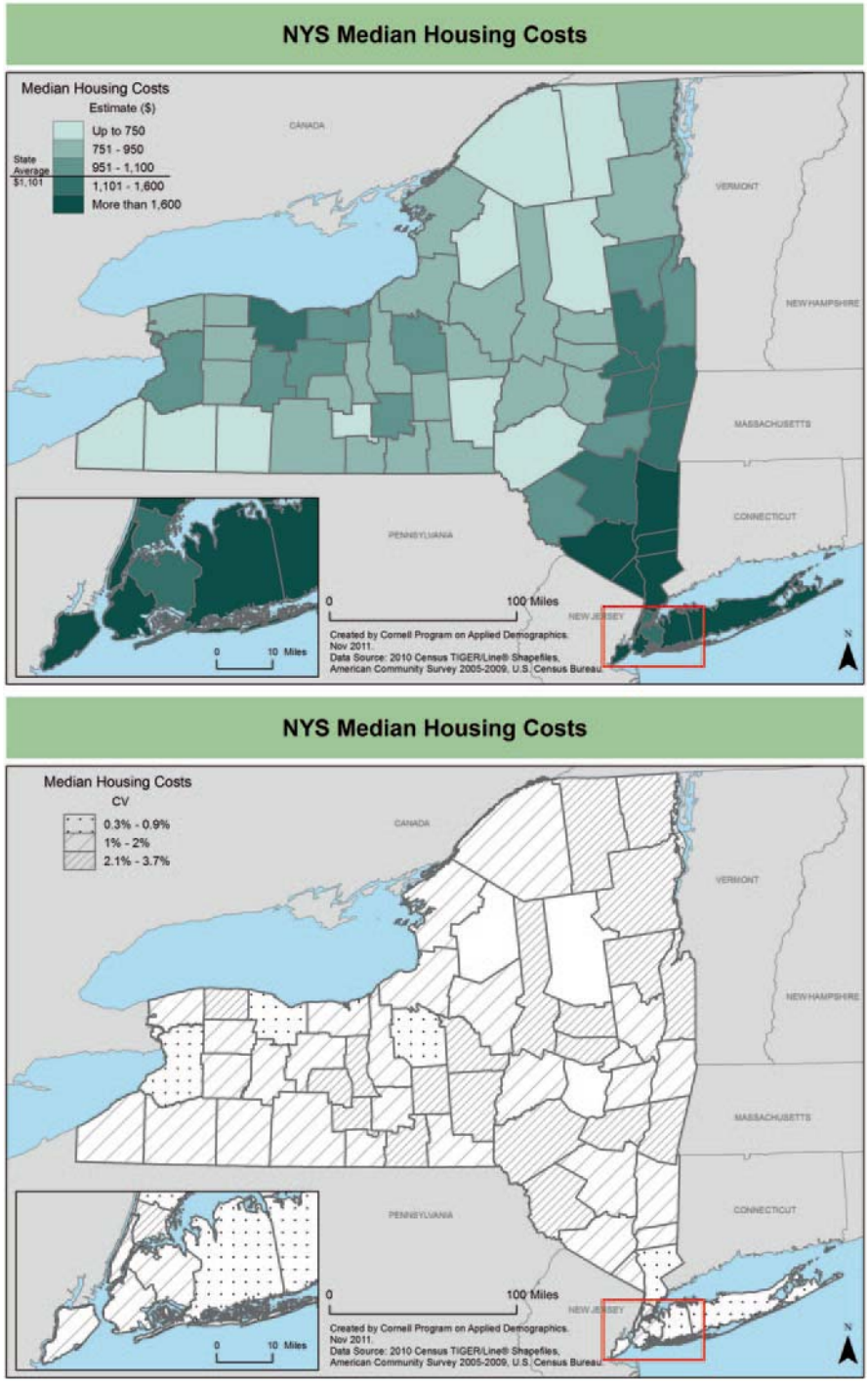

Fig. 16.5 Side by side maps, *left* showing estimates of median housing costs and *right* the MOEs

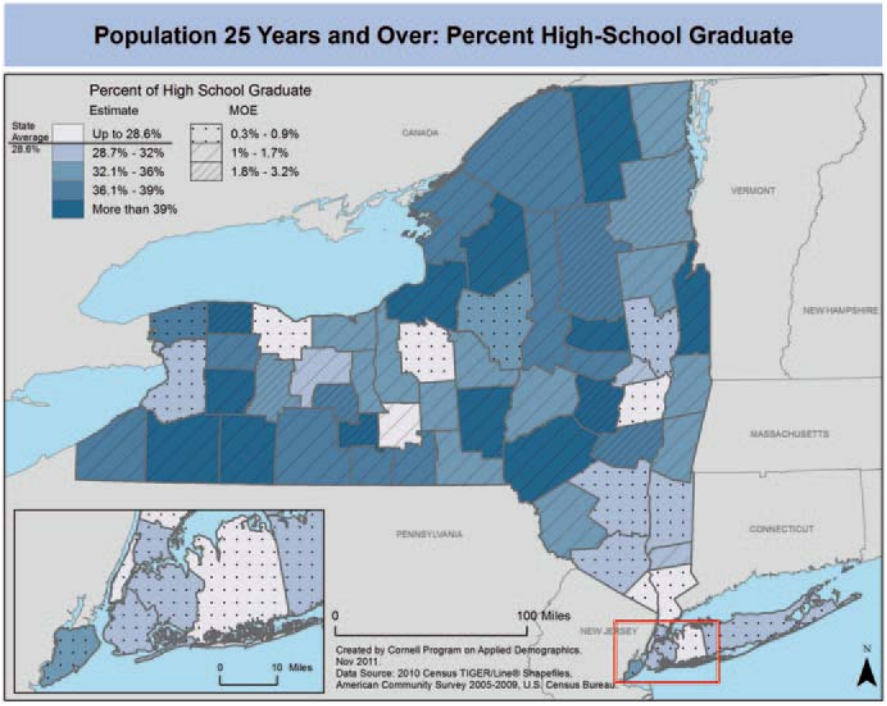

Fig. 16.6 Overlay map with one legend for estimates and the other for MOE

of estimates in our maps in the face of the errors of estimation and the likelihood that the "true" value of the variable placed arbitrarily in a given class may actually belong in an adjacent class. Xiao et al. (2007) investigated this issue and note, "the probability that an estimate is significantly different from values in other classes is a function of a number of factors, among them being the classification scheme and the size of the confidence interval."

Due to the uncertainty of estimation, settling on the number of classes, the class interval, and the class boundaries is more complex than when either (1) there is complete attribute certainty or (2) the error is small. As Xiao et al. (2007, p. 123) write, "When producing a choropleth map, it is important to realize that, owing to data uncertainty, each enumeration unit has a chance to fall into more than one class." In mapping ACS estimates, because the error is often large, margins of error need to be considered in setting the class boundaries and class interval such that when an estimate falls into a given class, the class boundaries are broad enough that they can include the confidence limits of the estimate as well.

Crisp classification of values in GIS follows the same principles as statistics—values assigned to a given class belong to that class and only that class. That is, there is no overlap of values of a given class into another class. However, because of the uncertainty of ACS estimates, unless error is taken into account in setting

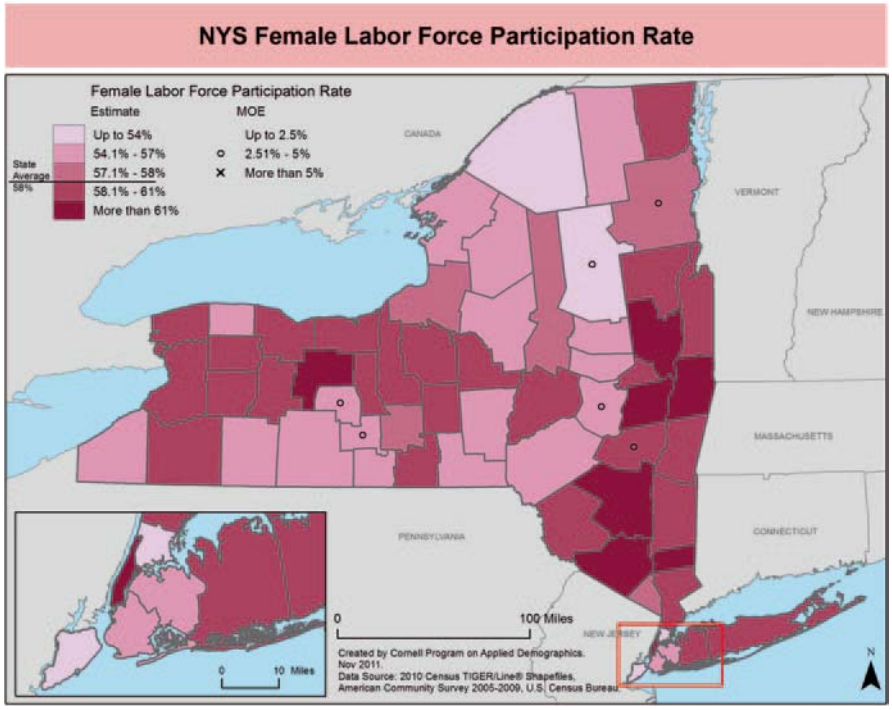

Fig. 16.7 Map of female labor force participation rate showing less obstructive MOE overlay

the class boundaries for a map legend, when the estimate is assigned to a class in the choropleth map, the confidence bounds may extend into other classes rather that coincide with the class boundaries specified (arbitrarily) for the legend. One consequence of uncertainty here is that the true value, represented by the estimate (we hope), may not be significantly different from values in those lower or upper classes in the map legend. In terms of symbology, the problem can be framed as one in which areas (e.g. townships) with different colors could have estimates that are not significantly different from each other, and areas with the same colors could have significantly different estimates. An alternative approach to dealing with this issue is to employ fuzzy classification as discussed earlier.

Sun and Wong (2010, p. 293) illustrate the issue of accommodative class interval width and boundary demarcations in the face of estimate unreliability with the following Figures (see Fig. 16.8, adapted from their Fig. 5), where the triangles represent the estimates, the class breaks (blue lines) are established at 20 and 30, and the confidence limits are represented by the error bars with round tips.

Notice in Fig. 16.8 that only for scenario one are both the estimate and the confidence limits of the estimate within the class limits. In only that situation can we be assured that the estimate is significantly different from estimates assigned to the classes above and below it. For the other scenarios, at least one of the confidence limits reaches into another class, meaning that we cannot be assured that estimates

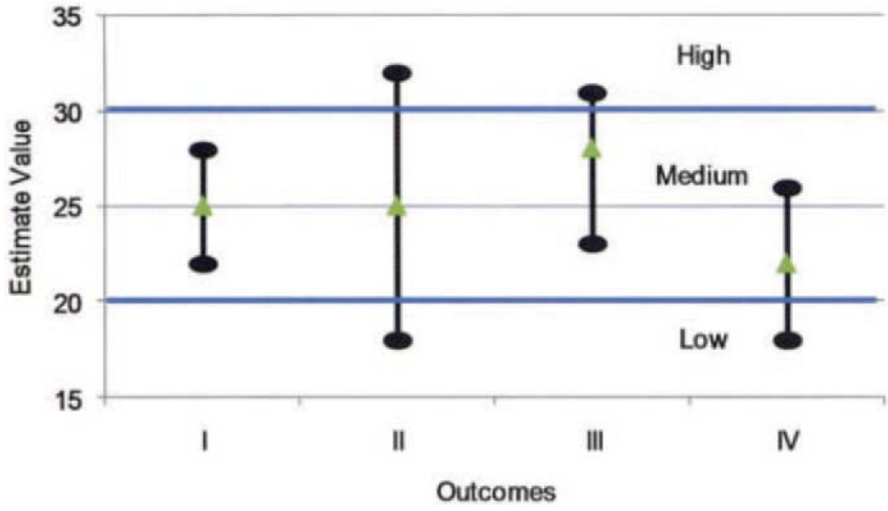

Fig. 16.8 Chart illustrating problem of establishing clear-cut category boundaries in face of measurement uncertainty

assigned to the focal class are significantly different from those assigned to the other classes.

Xiao et al. (2007) present the issue in a slightly different way. They use the term "robustness" to measure how well a classification works and define robustness of classification this way: "A classed choropleth map is robust if each enumeration unit has little chance of falling into a class other than the one to which it is assigned." To illustrate the application of their robustness concept, they ask the reader to consider two distinct enumeration units that have observation values *x* and *y*. See Fig. 16.9 (adapted from their Fig. 1).

The distributions of *x* and *y* are unknown, but with uncertainty we need to recognize that those x and y values could fall into any of the various classes depending on which of four classification schema are used. To illustrate the issue for at least two variables with uncertainty, Fig. 16.9 shows four possible classifications (A, B, C, or D), each using five classes, that the researcher may want to employ for classification and a map legend. In this illustration the class boundaries are represented by the vertical bars and the class interval represented by the width of the line between these bars.

Notice if classification scheme A is used, observations *x* and *y* will fall into classes 2 and 4, respectively. Xiao et al. (2007) indicate that under classification scheme A the classification of observation *x* is robust because all other possible values that could have occurred for ***x*** would have also fallen into class 2, since the interval of this class covers most of the distribution of ***x***. On the other hand scheme A is not robust for observation ***y*** since many of its other possible values would likely fall into class 3 or 5, instead of class 4. The reverse would be true for classification B, where the class limits for ***y*** are robust but are much too narrow for other plausible values

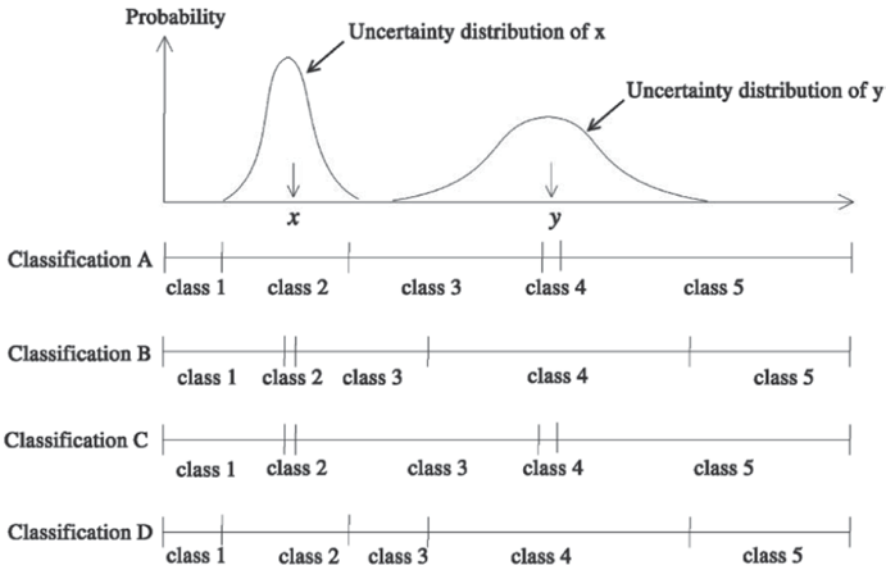

Fig. 16.9 Illustration of robustness of classification under estimate uncertainty

that ***x*** could assume. Classification scheme C is the least robust overall scheme as there is low probability that all or most of the possible values of ***x*** will fall into class 2 and likewise into class 4 for ***y***. By contrast classification D would best satisfy their criteria of overall robustness. The general point that Xiao et al. are making is that the uncertainty in attribute data makes it almost impossible to produce a perfectly robust choropleth map unless we somehow include robustness information as an element in the classification. But how do this?

Sun and Wong (2010, p. 294) suggest that one way around the problem is to first sort the estimates according to their value. Second, attach the corresponding confidence bounds to the estimates. Third, starting with some estimate, say the highest, compare it to nearby estimates to determine whether their confidence bounds overlay. If so, then group them into the same class. Fourth, proceeding onward, consecutive estimates whose confidence intervals do not overlap should be put into a different class. Their feeling is that this approach, which they call the "class comparison" approach, assures us that the estimates in a given class are significantly different from estimates in another class. On the other hand, a problem of this approach is that estimates within the same class may also be significantly different.

Xiao et al. (2007) have a much more elaborate approach that involves computing the probability of each observation x_i falling into each class j = 1…k, which they designate as p_{ij}. Letting p_{ij} be the probability that unit i belong in class j (1 < j < k), they calculate p_{ij} as follows:

$$\Pr(x_i \in I_j) = \int_{I_j} \pi_i(x)dx \tag{16.1}$$

where Ij is the interval of class j. Next they introduce the idea of a robustness measure q_α for the entire choropleth map such that when the robustness values for all enumeration units are obtained, one has a set $\{p_i | 1 \leq i \leq n\}$. Using q_α as their map robustness measure, where α is the tolerance level the researcher is willing to accept for complete classification accuracy, they require that $(1\text{-}q_\alpha)\%$ of its enumeration units have a p_i value greater than or equal to q_α.

With this conceptual development in place, Xiao et al. (2007, p. 125) state: "The key to obtaining p_i, and consequently q_α (which is the α^{th} quantile of {pi}), is to compute p_{ij} based on the uncertainty distribution of each unit. For many circumstances, the cumulative uncertainty distribution function can be analytically expressed (i.e. p_i can be written in closed form) so that for each class the exact p_{ij} values can be directly computed using equation (1)." Otherwise, they claim to be able to estimate the p_{ij} values via Monte Carlo approaches.

Xiao et al. (2007) also conducted a number of "experiments" of how well their robustness measures perform for various tolerances. They ran analyses on various combinations of five factors: type of data, type of uncertainty distribution, level of uncertainty, number of classes, and method of classification. Data consisted of four contrived datasets represented as polygons formatted into a regular lattice of size 33 by 33. The attributes for these polygons were invented continuous values that were scaled to range between 0 and 1, and then arranged to form four statistical surfaces: (1) uniformly distributed linear, (2) multimodal linear, (3) linear with a skewed data distribution, and (4) fractal. Regarding the level of uncertainty, they indicate that for a given polygon, two different uncertainty probability distributions were tested: a uniform and a Gaussian distribution. The number of classes used for classifying the data ranged from 5–100 with an increment of 5; hence 5, 10, 15…100. However, in reporting their results, they only focus on the "5 class" and "10 class" results. Equal-intervals, quantiles, and Jenks were the classifications schemes used. Their most general finding is that indeed the robustness of classification is a function of the level of uncertainty in the data. That is, as uncertainty increases, the probability of getting the polygon values into their most likely class decreases—regardless of the distribution of the values, the number of classes used, or the classification method employed. As Xiao et al. write (2007, p. 128), "This observation suggests that uncertainty effects must be c**o**nsidered as part of the classification process and, more importantly, that such effects should be revealed to map readers."

In our work at the Program on Applied Demographics we explored the idea of portraying the probability that the estimate belonged to the class to which we assign it. For static maps we tried the use of pie charts, where each slice of the pie represented the cumulative probabilities for a Gaussian distribution that the estimate belonged in the class to which it had been assigned by the Jenks method. See Fig. 16.10.

We also experimented with classifying and displaying the lower bound or the upper bound of the confidence intervals. This adds much complexity to the legend and interpretation of the map, but has the advantage that standard crisp classification methods can be applied. See Fig. 16.11.

For dynamic, internet mapping one can provide this information about the probability of the estimate belonging to the class to which it was assigned as a feedback when the user clicks or enters a polygon on the screen. We will illustrate this later in

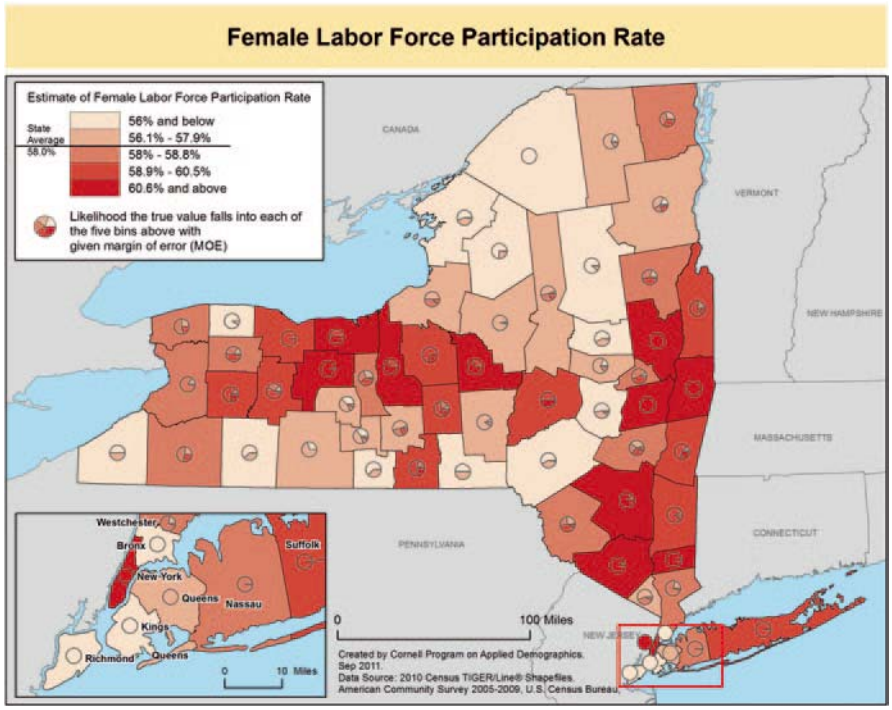

Fig. 16.10 Use of pie charts to symbolize cumulative probabilities of the estimate being what is shown for the geography

the section of this paper under the topic of static versus dynamic mapping. We have not explored fuzzy classification. This is a possible area for further work.

Number of Classes

One of the concerns we have in mapping ACS data with uncertainty is how many classes to use. Should one use more classes to achieve more certainty in classification or are fewer classes with wider class intervals better? The work of Xiao et al. is instructive here. They found that a small number of classes should be used if map robustness is a significant concern. As they state (2007, p. 131), "In general, an increase in the number of classes will induce a corresponding increase in the number of enumeration units with uncertainty distributions that overlap multiple intervals. Consequently, map robustness is reduced." Hence, when the data have high uncertainty, one can only create a robust classification map by keeping the number of classes low.

In our work, we have mostly kept the number of classes to five, sometimes using four. However, we haven't really explored the interaction of uncertainty and modification of class boundaries resulting in fewer classes with wider class intervals. This seems a useful area for additional research.

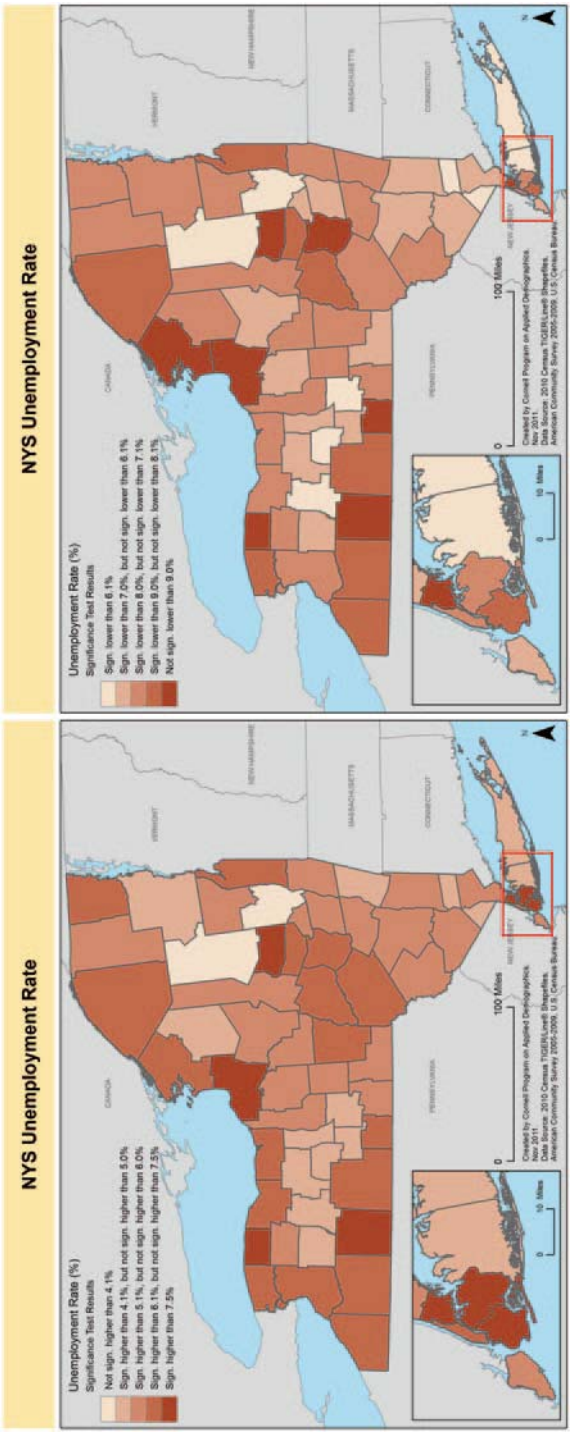

Fig. 16.11 Two approaches to estimating unemployment rate being significantly higher or lower than category boundaries

Method of Classification

One might expect that the Jenks method would provide the best classification of the attribute values under most circumstances, as it tries to minimize within-class variance while maximizing between-class variance to the extent possible. However, Xiao et al. (2007) found that among the four types of distributions they examined—linear, multimodal, skewed, and fractal—the difference between the three classification methods (equal-intervals, quantiles, and Jenks) was small under conditions of high uncertainty (low robustness). On the other hand, the Jenks method did outperform the others under conditions of (1) low uncertainty (high map robustness) and (2) use of only a few classes (i.e. less than 6). In general Xiao and colleagues found that for a dataset with a distribution similar to that of the multimodal or fractal data used in their paper, the Jenks optimal classification method appeared to be a superior choice, especially when a small number of classes was used.

In our own work we employed the Jenks classification method for "binning" the estimates and found that it seems to work well. We also experimented with the use of equal intervals for proportions, which allows one to express the corresponding MOE's in terms of the interval width.

Symbolizing Uncertainty

One aspect of research on spatially referenced attribute uncertainty during the past decade is in the area of geovisualization. As Kardos et al. (2003, p. 2) write, "Most research in attribute uncertainty has been focused on generating an uncertainty measure and then using visualization techniques to show uncertain areas." They further note that different attribute data uncertainty models are used for different spatial data. In GIS and analytic cartography, some models are ideal when using soil data, like fuzzy set theory, to express vagueness in soil type boundaries (Goodchild 1994). Other models like Monte Carlo can be used sequentially to express propagation of error and are good when dealing with random inaccuracies (Longley et al. 2001). Sun and Wong (2010), building on the work of MacEachren and Kraak (1997, 2001) for symbolizing errors of estimation for polygons, use metaphors (models) involving color and cross-hatching. But this isn't the complete extent of the work that has been done in the area and it is instructive to review some of the work of others at this point.

Lots of models or metaphors have been tried to improve the user's viewing of spatial information. Dent (1993) discusses using metaphors in representation through geometric shapes such as circles, squares, and triangles. In our own work on ACS data, we explored the use of circles, triangles, and squares as an abstraction to represent attribute uncertainty (Fig. 16.12).

MacEachren (1992) demonstrates the use of visual metaphors that included fog cover to hide the uncertain map parts and the blurring of uncertain areas. Kardos et al. (2003, p. 815) provide a very informative summary of things that have been

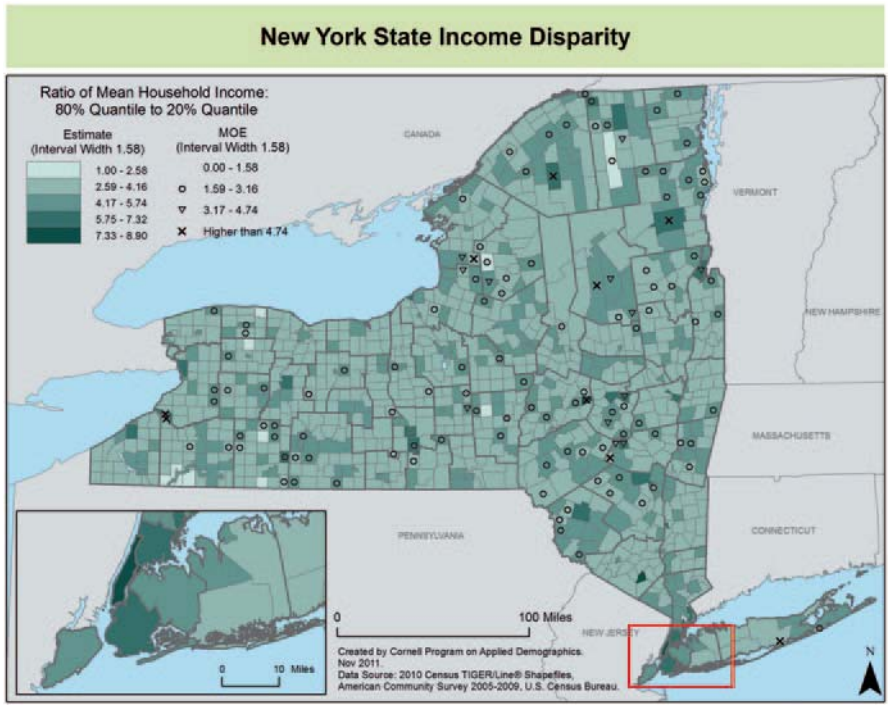

Fig. 16.12 Illustration of use of hollow geometric shapes to symbolize MOE overlaying estimates of income disparity

Technique **Metaphor of:**	**Fog** – Detail	**Blur** – Focus	**Blinking Pixels** – Stability	**Colour Mix** – Clarity	**Pixel Mix** – Fuzziness
Certain Data **Metaphor of:**	Clear – Revealing detail	Sharp Focus – Focused	No Blinking – No Movement, therefore more stable	High Saturation – High clarity, less recessive	Single Hue – Low fuzziness
Uncertain Data **Metaphor of:**	Foggy – Hiding Detail	Blurry – Unfocused and merging	Blinking over areas – Less stable, more unsettling	Low Saturation – Low clarity, more recessive	Multiple Hues – High fuzziness

Fig. 16.13 Suggested visual metaphors to signal estimate uncertainty

tried, particularly the effects being sought through the symbolization metaphors that various researchers have either proposed or used. See Fig. 16.13 (adapted from their Table 1).

Kardos, et al. (2003) have summarized also the details of research that GIS scientists and analytic cartographers have either proposed or used in the symbolization metaphors over the past decade. See their Table 1).

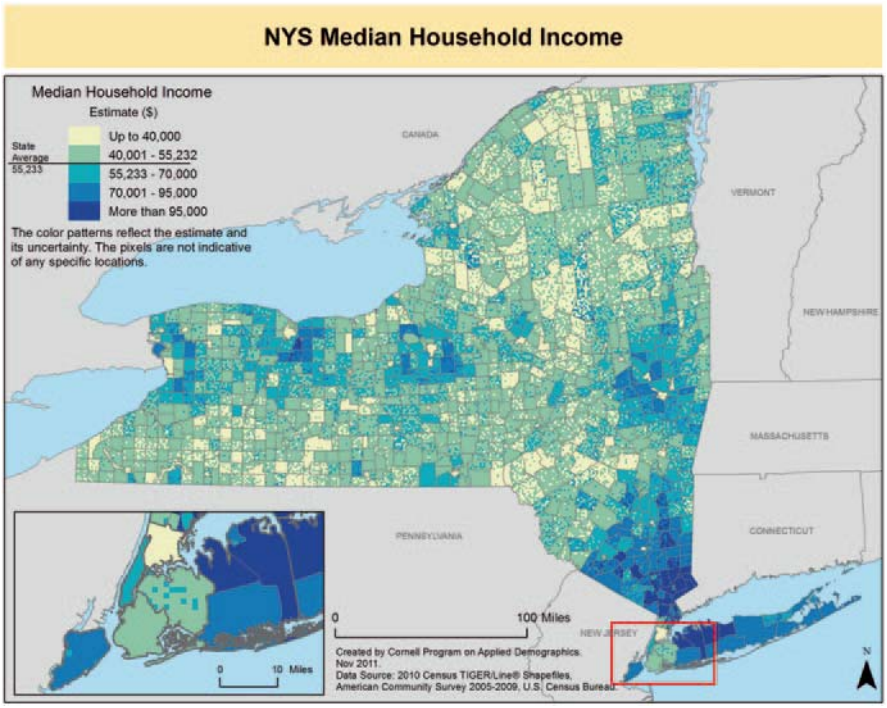

Fig. 16.14 Illustration of the use of pixel mixture to symbolize amount of estimate uncertainty overlying the estimate

Since examples of adjacent and overlay maps have already been presented, of interest here might be examples of a pixel mixture (pixelated) map (Fig. 16.14) and a saturated color map (Fig. 16.15).

In this pixel mixture map, pixels values were assigned proportional to their value under the Gaussian distribution. Hence, the greater the number of pixels having the same category value as an estimate value (low mixture), the greater the certainty of the estimate. On the other hand, when the polygon contains a lot of "spottiness" (high mixture), the lower is the certainty of the estimate.

Another alternative approach to expressing uncertainty along with an estimate value on the same map mentioned by Kardos et al. (2003) above is a map using color saturation. Figure 16.15 gives an illustration of one using a combination of saturation and intensity to symbolize both the value of the estimate and the uncertainty of the estimate. Here darker color symbolizes polygons with estimates of higher median household income while simultaneously intensity (dark to light) symbolizes the degree of uncertainty of the estimate (from 0–100 %). So subcounty units just north of New York City, as well as parts of Long Island, have high median household income and the coefficient of variation is low. By contrast, areas further north up the Hudson River have low estimated median household income and the

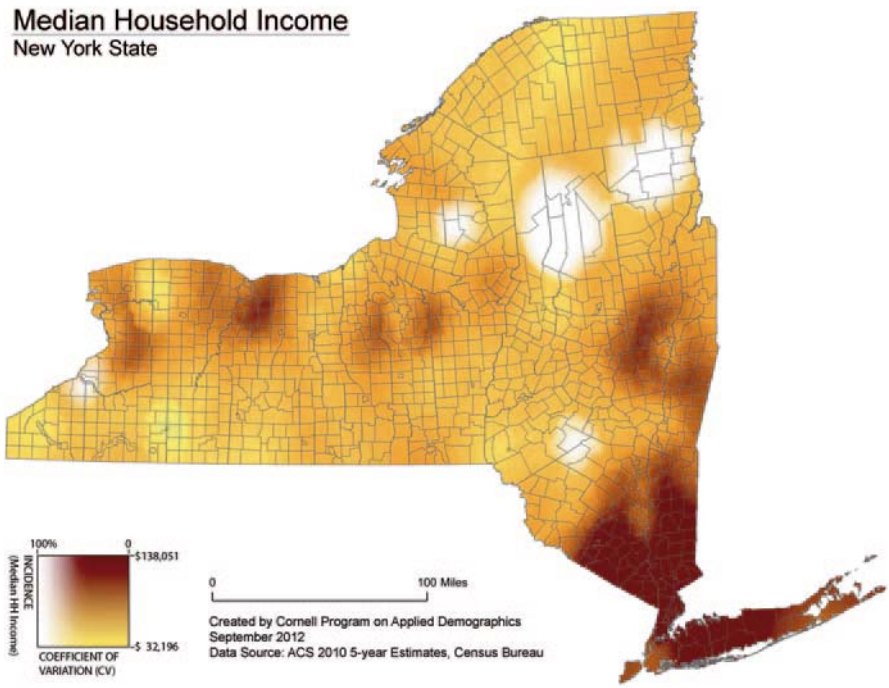

Fig. 16.15 Illustration of the use of color saturation and intensity to symbolize estimate uncertainty in combination with the estimate

coefficient of variation is large, reflecting the uncertainty that comes with smaller sized samples for those areas.

Kardos et al. (2003) also conducted a bit of research on how users perceived and utilized various symbolization (geovisualization) techniques designed to communicate data uncertainty. Their research is especially pertinent to the ACS estimate and error of estimation context. The data used was from the New Zealand 2001 census. They derived an uncertainty value for each feature using data from the 2001 post enumeration survey. Their survey was conducted via the internet and was designed to provide answers to three aspects of various symbolizations: (1) the visual appeal, (2) speed of comprehension of the information, and (3) overall effectiveness of the symbolizations. The respondents were all experienced GIS users.

The maps presented to the respondents to evaluate in terms of visual appeal, speed of comprehension, and symbology effectiveness consisted of the following "treatments." Nine different techniques were assessed and rated—adjacent maps, overlays, blurring, fog, pixel mix, saturation of color, sound, blinking pixels, and animation–using a five-point assessment scale (excellent, good, moderate, limited, and ineffective), along with the option of stating that the metaphor was "not useful." After examining the performance of their nine techniques on the usefulness, visual appeal, and speed of comprehension criteria, Kardos et al. (2003) drew the conclusion that the blinking of areas metaphor/technique outperformed the other

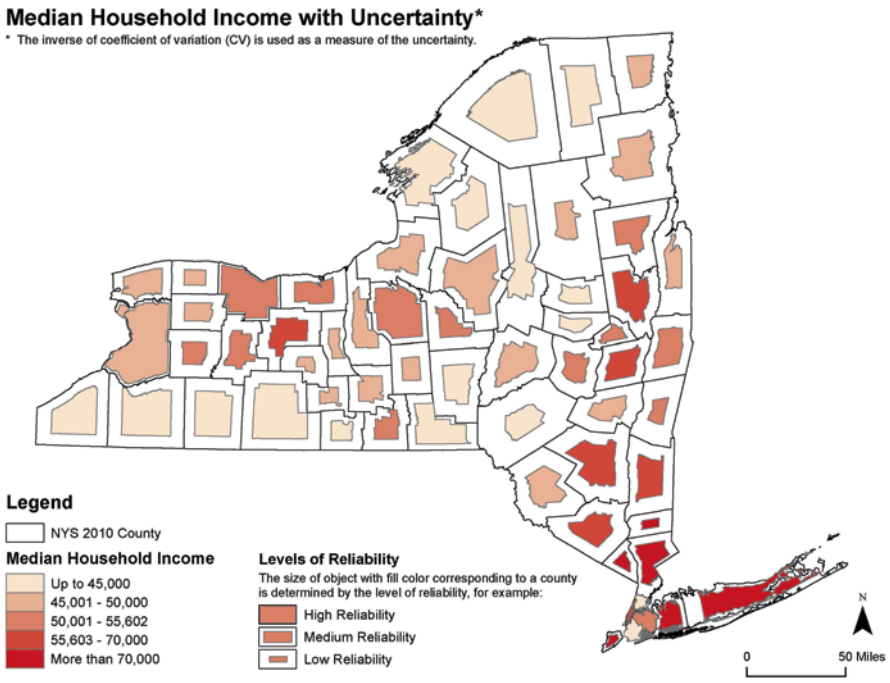

Fig. 16.16 Illustration of use of cartograms to symbolize uncertainty in combination with estimates of median household income

techniques. Overlay was found to be useful by 84 % of the respondents, while 78 % found adjacent maps (one for the estimate and one for uncertainty) were useful, with "fogging" and "blurring" next most useful.

The reason the respondents stated for preferring the blinking technique over others is that it didn't obstruct their viewing of the original information values. While they found the overlay technique useful, the felt it interfered with their understanding of the values symbolized by color. These are very useful finding.

In our own work, we found the same problem of confusion when presenting both the estimate and uncertainty information overlay. Side-by-side maps were just too awkward. So, we explored a modification of a "blinking" technique. For our static (pdf) maps, we first present the estimate for the geographic areas of interest) and then, with one mouse click, the viewer overlays the error of estimation information.

A final example of a map overlaying both ACS estimate and error of estimation is the Cartogram. This approach was suggested by Jerzy Wieczorek (2012, http://civilstat.com/?p=50), who provided an illustration that spurred the development of the cartogram approach in Fig. 16.16.

In this cartogram approach, the polygons (counties of NY) were kept in accurate size but the internal polygons were distorted in size to reflect the degree of uncertainty in three classes—high, medium, and low. Hence, the more the internal polygons are shrunk relative to the county boundaries, the lower the reliability of the estimate.

Static vs. Dynamic Interactive Maps

Dynamic interactive maps permit much more flexibility in presenting information compared to static maps. As Torrieri et al. (2011, p. 11) state, "Digital maps that include options for concealing or displaying information relating to the quality of the data displayed offer greater flexibility to the map designer." For one thing, you can program the application that serves out the information to "blink" those areas with high uncertainty as well as having MOEs displayed when the user passes a mouse cursor over the area. With the increasing availability of APIs for implementing these techniques, this is an area that should be explored by the Census Bureau and others. Jan Vink has done this for ACS data and New York geographies). Nicholas Nagel has built a version for SAIPE data and Tennessee geographies). For these interactive, internet maps the user has only to move the mouse over a geographic unit of interest and the error of estimation is displayed.

A slightly different approach to an internet-based map containing a layer of estimates and a layer of measures of error of estimation is one in which each of the layers of information can be toggled on or off, as can various boundaries.

While we think these approaches will improve the effectiveness of presenting both the estimates and their error, there are limitations to these techniques as well. For one, they require the user to make that mouse movement. Secondly, digital display techniques are generally outside the training of most spatial demographers, so this portends slow adoption of this approach.

ESRI has experimented with presenting maps of ACS estimates and error of estimation in the ESRI Business Analyst Online. It is served out via the internet but is only available via subscription. Moreover, unlike the Vink and the Nagel dynamic maps, the Business Analysts Online maps are served one at a time and are not overlain. Hence, they are really akin to the side-by-side maps discussed earlier.

Research is needed comparing these internet served interactive maps with static maps to see what users find more useful, understandable, and appealing.

Number of Geographic Units on Map

Lastly, Torrieri et al. (2011, p. 10) have noted that the overlay approach to communicating error of estimation via an integrated map has limitations when the number of geographic units in the map display is numerous. They illustrate this by asking the reader to imagine presenting both kinds of information (estimate and MOE) on a map for all 3,143 counties in the United States. For this situation they suggest that the map maker present only selected regions of the entire geographic coverage at time.

In our work at the Program on Applied Demographics, we explored the idea of using unfilled symbols of different shapes overlaid on a choropleth map of approximately 1,000 sub-county geographies (towns, cities, and reservations) in New York State.

References

Beard, M. K., & Buttenfield, B. P. (1991). *NCGIA Research Initiative 7: Visualization of Spatial Data Quality*. NCGIA, Technical Paper, pp. 91–26.

Burrough, P. A., & Frank, A. U. (1997). Geographic objects with indeterminate boundaries. *Transactions in GIS,* 2(4), 377–378.

Burrough, P. A., & McDonald, R. A. (1998). Principles of geographical information systems. Oxford: Oxford University Press.

Cedilnik, A., & Rheingans, P. (2000). *Procedural Annotation of Uncertain Information*. Proceedings of visualization 00, pp. 77–84. IEEE Computer Society Press, 2000.

Chrisman, N. (1995). *Beyond Stevens: A revised approach to measurement for geographic information*. http://mapcontext.com/autocarto/proceedings/auto…/pages281–290.

Dent, B. D. (1993). *Cartography: Thematic map design*. Dubuque: William C. Brown Publishers

Dutton, G. (1992). *Handling Positional Uncertainty in Spatial Databases*. Proceedings 5th international symposium on spatial data handling, pp. 460–469. University of South Carolina, August 1992.

De Gruijter, J. J., & McBratney, A. B. (1988). A modified fuzzy k-means method for predictive classification. In H. H. Bock (Ed.), *Classification and related methods of data analysis* (pp. 97–104). Amsterdam: Elsevier.

Eathington, L. (2011). *Holy MOE! Don't Marginalize the Error in the American Community Survey*. Paper was prepared for presentation at the 2011 annual meeting of the Southern Regional Science Association in New Orleans, LA.

Fairbairn, D., Andrienko, G., Andrienko, N., Buziek, G., & Dykes, J. (2001). Representation and its relationship with cartographic visualization. *Cartography and Geographic Information Science, 28*(1), 13–28.

Goodchild, M. F. (1994). *Spatial Accuracy*. Proceedings, 22nd annual conference, Australasian Urban and Regional Information Systems Association, Sydney, addendum. [216].

Hengl, T., Walvoort, D. J. J., & Brown, A. (2002). *Pixel and Colour Mixture: GIS Techniques for Visualisation of Fuzzyness and Uncertainty of Natural Resource Inventories*. In G. Hunter (ed), Proceedings: International symposium on spatial accuracy in natural resources and environmental science10–12 July 2002, Melbourne, Australia.

Hootsmans, R. M. (1996). *Fuzzy sets and series analysis for visual decision support in spatial data exploration*. PhD Thesis, University of Utrecht, Utrecht, 167 pp.

Kardos, J. D., Moore, J. A., & Benwell, G. L. (2003). *The Visualization of Uncertainty in Spatially-Referenced Attribute Data Using Trustworthy Data Structures*. Proceeding of 15th annual colloquium of the Spatial Information Research Centre (SIRC 2003: Land, Place and Space).

Li, X. R., & Zhao, Z. (2005). *Relative Error Measures for Evaluation of Estimation Algorithms*. Paper presented at IEEE conference, 2005.

Longley, P. A., Goodchild, M. F., Maguire, D. J., & Rhind, D. W. (2001). *Geographic information systems and science*. New York: Wiley.

MacEachren, A. M. (1992). Visualizing uncertain information. *Cartographic Perspective, 13*(Fall), 10–19.

MacEachren, A. M., & Kraak, M. J. (1997). Exploratory cartographic visualization: Advancing the agenda. *Computers & Geosciences, 23*(4), 335–343.

MacEachren, A. M., & Kraak, M. J. (2001). Visualization in modern Cartography. *Cartography and Geographic Information Science, 28*(1), 3–12.

Monmonier, M. (1990). *Strategies for the Interactive Exploration of Geographic Correlation*. Proceedings of the 4th international symposium on spatial data handling, Vol. 1, pp. 512–521. IGU, July 1990.

Pang, A. (2001). *Visualizing Uncertainty in Geo-Spatial Data*. Paper prepared for a committee of the computer science and telecommunications board.

Pang, A., Furman, J., & Nuss, W. (February 1994). Data quality issues in visualization. In J. Robert, I. I. Moorhead, D. E. Silver, & S. P. Uselton (Eds.), *SPIE Vol. 2178 Visual data exploration and analysis* (pp. 12–23). Bellingham: SPIE.

Sun, M., & Wong, D. S. (2010). Incorporating data quality information in mapping American community survey data. *Cartography and Geographic Information Science, 37*(4), 285–300.

Torrieri, N., Wong, D., & Ratcliffe, M. (2011). *Mapping American Community Survey Data*. Paper dated September 2011.

Wieczorek, J. (2012). (http://civilstat.com/?p=50) in his March 9, 2012 posting.

Xiao, N., Calder, C. A., & Armstrong, M. P. (2007). Assessing the effect of attribute uncertainty on the robustness of choropleth map classification. *International Journal of Geographical Information Science, 21*(2), 121–144.

Yee, L., Goodchild, M., & Lin, C.-C. (1992). *Visualization of Fuzzy Scenes and Probability Fields*. Proceedings 5th international symposium on spatial data handling, pp. 480–490. University of South Carolina, August 1992.

Zhang, J. X., & Goodchild, M. F. (2002). *Uncertainty in geographic information*. New York: Taylor and Francis.