JOURNAL OF SPATIAL INFORMATION SCIENCE
Number 13 (2016), pp. 103–130
doi:10.5311/JOSIS.2016.13.268

RESEARCH ARTICLE

Mapping neighborhood scale survey responses with uncertainty metrics

Charles R. Ehlschlaeger[1,2,3], Yizhao Gao[2], James D. Westervelt[1], Robert C. Lozar[1], Marina V. Drigo[4], Jeffrey A. Burkhalter[1], Carey L. Baxter[1], Matthew D. Hiett[1], Natalie R. Myers[1], and Ellen R. Hartman[1]

[1]Engineer Research Development Center, Construction Engineering Research Lab, Illinois, USA
[2]Department of Geography, University of Illinois, Urbana, USA
[3]Geographic Information Systems Program, John Hopkins University, Baltimore, USA
[4]PERTAN Group Inc., Champaign, IL, USA

Received: November 11, 2015; returned: April 5, 2016; revised: July 6, 2016; accepted: August 24, 2016.

Abstract: This paper presents a methodology of mapping population-centric social, infrastructural, and environmental metrics at neighborhood scale. This methodology extends traditional survey analysis methods to create cartographic products useful in agent-based modeling and geographic information analysis. It utilizes and synthesizes survey microdata, sub-upazila attributes, land use information, and ground truth locations of attributes to create neighborhood scale multi-attribute maps. Monte Carlo methods are employed to combine any number of survey responses to stochastically weight survey cases and to simulate survey cases' locations in a study area. Through such Monte Carlo methods, known errors from each of the input sources can be retained. By keeping individual survey cases as the atomic unit of data representation, this methodology ensures that important covariates are retained and that ecological inference fallacy is eliminated. These techniques are demonstrated with a case study from the Chittagong Division in Bangladesh. The results provide a population-centric understanding of many social, infrastructural, and environmental metrics desired in humanitarian aid and disaster relief planning and operations wherever long term familiarity is lacking. Of critical importance is that the resulting products have easy to use explicit representation of the errors and uncertainties of each of the input sources via the automatically generated summary statistics created at the application's geographic scale.

Keywords: demography, data uncertainty, inequity analysis, spatial-temporal application credibility, planning analysis, survey analysis

1 Introduction

Traditional measures of survey quality have focused on sampling technique errors. For example, U.S. Congressional records [24–27] provide details on under-counting and over counting estimates from the 2000 census, providing good estimates of the uncertainty between reported census figures and the true population enumerations at the census block level. From the record, demographic modelers can determine the variability between official enumerations against sampled higher quality counts. Even though the data quality metrics available from the U.S. Census Bureau are as or more precise than most censuses available (and probably all surveys), these measures cannot be easily used to determine whether US Census data or American Community Survey information can be useful for a particular application. In order to quantify the utility of spatial data for a particular application, researchers often employ Monte Carlo simulation (MCS) [19], especially for complex applications. MCS requires dozens to thousands of realizations for each input data layer with the application repeatedly run for each version of the input data layers to properly identify the range or distribution of potential outcomes indicating the data's application utility [12]. Ideally, all knowledge of potential data errors is incorporated into the simulation model that generates the data layer realizations. The same is true of the simulation model or analytic process using the data: All errors and uncertainties identified in the model or analysis must be represented stochastically in the MCS for its results to fully represent the uncertainty. Using U.S. Census as an example, census block enumerations should be stochastically adjusted by $\pm 2\%$ in the accurate locations and much greater in census blocks with demographic conditions known to induce higher errors. In some locations, sampled census blocks were off by 40% [24–27]. Any unrepresented errors or uncertainty will underestimate the difference distribution between the simulation model results and the intended model goals.

This paper presents a methodology of mapping population-centric social, infrastructural, and environmental metrics at neighborhood scale. It utilizes and synthesizes survey microdata, Bangladesh administrative level four choropleth attributes, land use information, and ground truth locations of attributes to create neighborhood scale multi-attribute maps. This methodology is part of a multi-year ongoing research project attempting to quantify the uncertainty of survey microdata and make covariated responses in the survey available at geographic scales more precise than the original survey. The underlying goal of this research is to create a foundation of rich contextual metrics at neighborhood scale in order to form a more structured presentation of urban space. This research is inspired by "The Digital City," a series of conferences [5] representing complex urban environments. Recent efforts to represent demographic information include agent-based transportation modeling [16, 25], cluster detection analysis [7], and migration analysis [10]. These efforts used microdata and surveys to provide the inputs necessary to perform complex urban analysis. These efforts eliminated ecological fallacy by not aggregating multiple demographic attributes into arbitrary areas [13].

This research's conceptual approach locates potential household locations with a conflation model [4], which is also known as data fusion [28]. Conflation is the process by which multiple data layers are combined in order to generate a product containing the best aspects of each layer. In this research, five types of data are conflated: choropleth population enumeration estimates, survey responses or microdata, a household or population density map, maps of survey response constraints, and samples of ground truth information con-

taining specific survey responses. Each of the data inputs imparts specific types of precise information that other data types do not contain. This demographic model ensures that each type of precise information is represented in the data layer realizations at the level of accuracy known for each data layer. The model creates three types of demographic results:

1. Plausible realizations of every person and household within the study area with simulated locations for Monte-Carlo simulation (MCS) agent-based modeling.
2. Plausible realizations of survey response heat maps, suitable for geospatial analyses using MCS to explicitly represent uncertainty.
3. Maps that summarize the distribution of realization results as box plot variables [18] at all locations in the study area. The box plot variable maps are useful for cartographic analysis and communication in most common geographic information systems.

This article will articulate the proposed set of techniques, highlight the data sources developed and used, then proceed to illustrate the technique and its outputs using a case study in the Chittagong Division of Bangladesh. The research draws ideas from multiple disciplines, requiring knowledge of survey design and statistics, demographic modeling techniques, spatial statistics, habitat modeling, and spatial data uncertainty modeling to fully implement, making it difficult to cover all techniques in depth. The article concludes with some recommendations for future improvements.

2 Description of the proposed methodology

By necessity, the methodology for creating neighborhood scale survey response maps from censuses and surveys is both heuristic and complicated. Nearby land use characteristics and urban infrastructure can attract or distract people from living in an area. Such influences, however, are highly depending on cultural conditions, and thus vary between and even within countries. Each input theme, whether census, survey, land cover, land use, and physical or artificial characteristics has its own measures of accuracy which should be explicitly accounted for as a range of possible values. Otherwise, the constrained representation of those input layers will underestimate the survey response variability between the demographic model and reality.

This section presents the six step process for converting surveys into the three types of demographic results. Before stepping into the procedure, some important terms need to be defined. The households and people simulated are derived from a survey or microdata, defined as a set of questions asked using proper survey design techniques. The locations of simulated households are based on household density surface and census information. The household density map indicates the probability that a household is located at each place based on environmental and infrastructural information. The household density map can be derived via multiple techniques, and is discussed in section four. A survey case is the responses given by one person or household. More simply, a survey case is the set of answers given by a person to a survey's questions. One sample in census microdata has a more complex origin than a survey case. A census microdata sample represents a cluster of similar, but hardly ever identical, survey cases combined. A census microdata sample is used as a survey case in this methodology. Survey responses are the set of answers relevant to a particular operational need or analysis, which should be considered as a subset of a survey

case. Population enumerations are estimated counts of households, people, and their attributes within administrative areas defined at the finest scale possible, provided by census data and including an estimate of census error measures and adjusted to changes over time. Ground truth samples are point locations or binary maps where known attributes related to survey responses are located. This process locates known survey responses to specialize information not typically available to demographic models and will ensure those locations will have survey responses at those locations. One end product, survey response box plot variable maps, provide easy to understand maps of survey response covariates while providing a representation of uncertainty at every location with the information available in typical box plots. The box plot variable maps are summarized from many alternative realizations of every household or person in the study area, either as a regular lattice of kernel density estimates, or as a proportion within cells of a regularized grid. Figure 1 represents the process for converting the various inputs into a set of maps representing the variation of a survey response over a study area.

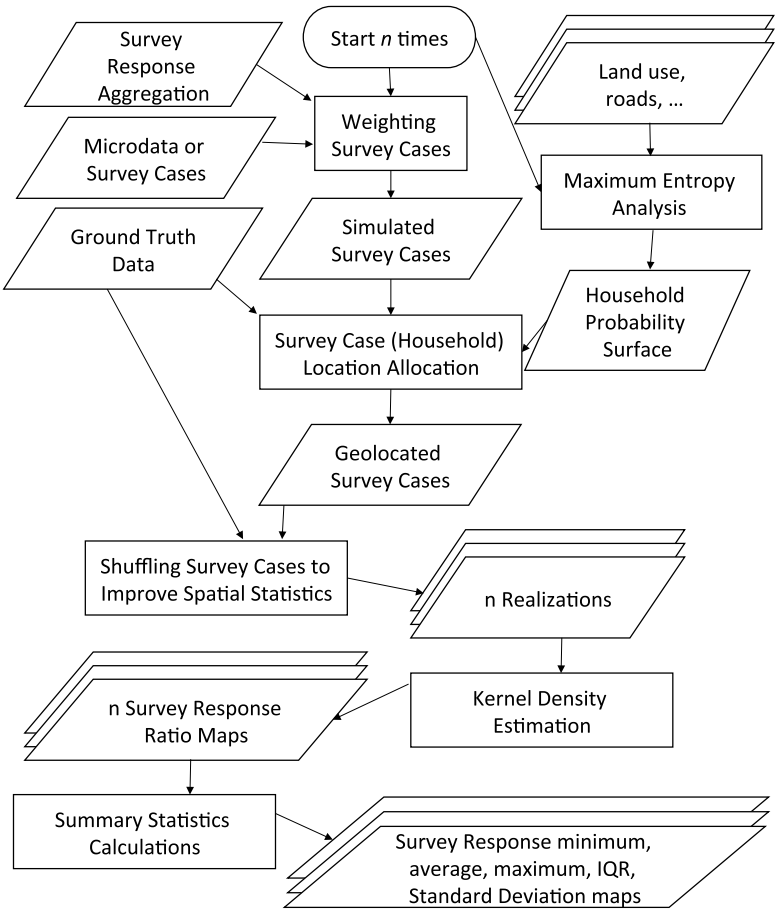


Figure 1: The survey response mapping process. Each step is performed n times for Monte Carlo simulation uncertainty analysis.

The process can be summarized into the following major steps:

1. Survey cases are replicated a number of times, **Weighting Survey Cases**, to match demographic characteristics in the overall estimated population enumerations. The replication process fits the results are weighted using a sum of least squares minimizing specific desirable survey responses.
2. A probability surface of household locations is created, **Maximum Entropy Analysis**, for each population realization.
3. Replicated geolocated survey case locations are realized, **Survey Case Location Analysis**, by optimizing a set of proportional for each population realization. Household locations are based on a household probability surface, ground truth data, and survey responses' proportional measures.
4. Survey case locations are shuffled, **Shuffling Survey Cases to Improve Spatial Statistics**, optimizing a set of proportional and spatial statistics for each population realization to create realistic clustering of survey responses.
5. For each desired combination of survey responses, proportion maps are generated, **Kernel Density Estimation**, on each population realization throughout the study area representing the percentages of simulated survey cases with such responses. This kernel density estimation process is done cell by cell across a regularized grid.
6. Throughout the study area, box plot summary statistic maps are compiled, **Summary Statistics Calculations**, on the minimum, maximum, median, medium, 1st quartile, 3rd quartile of realizations at all study area locations, as well as the standard deviation and interquartile range for these locations. Both the summary statistics and the kernel analysis for each realization provide error and uncertainty estimates.

Steps one through four are done repeatedly for dozens, hundreds, or even thousands of alternative realizations to create enough realizations to provide representative distributions for important survey answers at critical geographic locations. For example, a survey response that is answered seldom would require a larger number of realizations for its Poisson distribution to reflect the variation of reality while a survey response answered by about 50% of the households or people would take fewer realizations to define the resulting normal distribution.

If a more complex analysis of survey cases is desired, map algebra can be performed on population realizations of survey response maps as shown in Figure 2, the Stochastic Simulation Map Algebra Process. This process, which the authors named "quantum population geo-analytics" (QPG), begins with the n realizations of Geolocated Survey Cases from step four of the Survey Response Mapping Process, Figure 1. Its three step process is:

1. Construct a heat map of the proportion of survey cases fitting each of the survey response queries necessary for the complex application, **Kernel Analysis for each Map Algebra Theme**, at the neighborhood size desired for each realization.
2. Perform the map algebra analysis for each realization, **Map Algebra 1 to n Realizations**.
3. Throughout the study area, box plot summary statistic maps are compiled, **Summary Statistics Calculations**, on the minimum, maximum, median, medium, 1st quartile, 3rd quartile of realizations at all study area locations, as well as the standard deviation and interquartile range for these locations.

The case study presented in this paper shows one such complex analysis in Figure 8. That case study, being more complex than a query of survey responses, requires that map

algebra must be performed on a realization by realization basis with summary statistics calculated for each set of output maps in the procedure. Details about the application itself is in Section 5.

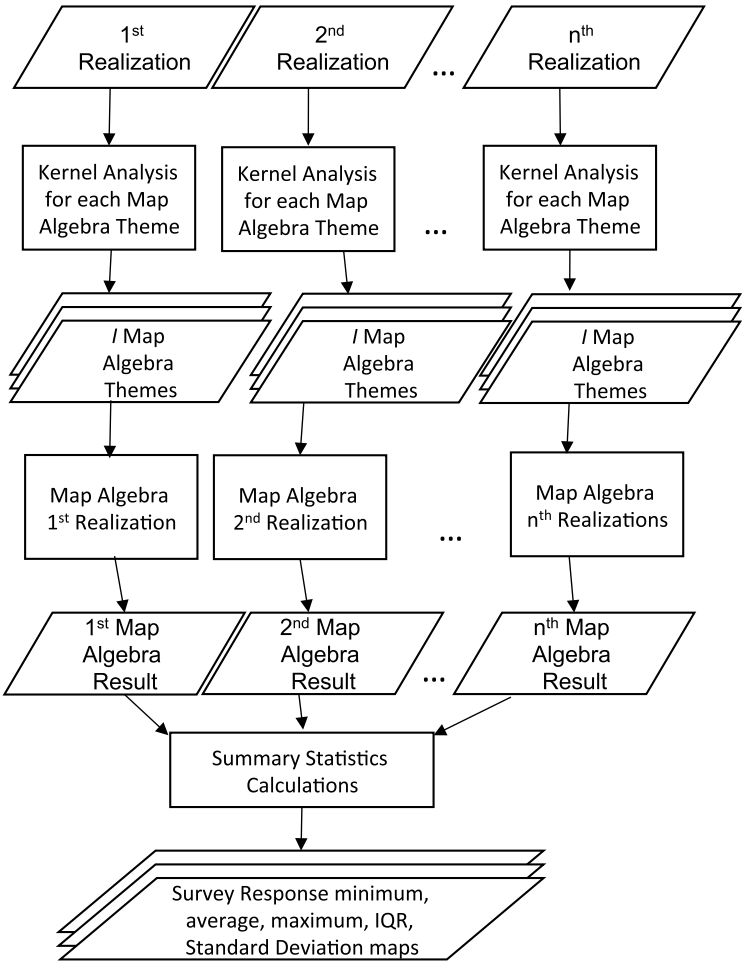


Figure 2: The stochastic simulation map algebra process.

2.1 Weighting survey cases to census data

Weighting survey cases to census data, **Weighting Survey Cases** in Figure 1, is the process of fitting survey cases to enumerated population estimates for administrate regions in the study area. In simpler terms, each survey case is replicated a number of times to match the overall estimated population enumerations. This process is known as "population weighting adjustment" [15] in survey literature and is appropriate for correcting surveys likely to be more biased than authoritative population enumerations. A subset of survey responses are that related to population enumerations are chosen to fit the survey. For the case study illustrated in this paper, IPUMS questions for member counts,

toilet facilities, building types, electricity availability, household water supply type, house ownership, and household religious preference were weighted against population enumerations provided by the country census enumerations at administrative level four. This step uses a "duplication of cases" approach [17], minimizing the sum of squared errors between survey case responses and authoritative population enumerations. This approach is convenient due to the complexity of the fitting criteria, survey nonresponses, and maintaining co-variance relationships inside survey responses. Kalton (1983) indicates that the multiple realizations generated by the overall technique almost completely eliminate the increased variances [16]. Survey cases are stochastically duplicated to minimize root mean squared errors of the questions' relationships.

2.2 Creating a probability surface of household locations

A probability surface of household locations, or household density map, provides a surface with population density values at different locations to guide the placement of simulated households. The spatial resolution of a household density map can be as fine as that of the most detailed land use map or remotely sensed imagery. Maximum entropy analysis [14] was used to determine which landscape characteristics were valuable for predicting where households would be located in the study area. (See **Maximum Entropy Analysis** in Figure 1 for its location in this process.) Maximum entropy analysis expresses the suitability of a location, household location in this case, as a function of the environmental variables at or near that location. The variables used were samples of density urban, suburban, and rural household locations fitted against distance to roads, hydrology information, topography, and various imagery layers. Alternative techniques to create density surfaces from land use information and census data include random forest modeling [22], and linear regression [7]. A stratified random sample of urban and rural household locations was used to calibrate and verify the model. For the case study in this paper, 23 map themes were tested for significance to samples of household locations. Those map themes were from 15 meter NaturalView Landsat Mosaic imagery layers including raw bands, unsupervised classifications pattern and texture measures; SRTM (shuttle radar topography mission) also provided topographic variables including hydrology; and open source augmented road information. A deterministic sequential-update algorithm [6] implemented in the Maxent open source software package [21] was used to determine which layers were significant. For a general discussion on the statistical properties of this process see Erith et al. (2011) [9]. While other imagery provided more accurate results, especially when pixel resolution became as fine as 2 m, the choice of imagery ensures near global availability of all urban areas.

2.3 Spatial allocation and shuffle households processes

There are three levels of model complexity when generating point patterns that mimic real world processes: 1) homogeneous Poisson processes, 2) heterogeneous Poisson processes, and 3) Cox processes [1]. Homogeneous Poisson processes assume that an object is equally likely to be located at any point in the study area ignoring population density, location attractors and detractors with this simple process. Heterogeneous Poisson processes recognize that different parts of the study area have different likelihoods of containing events. Applying only the **Survey Case Location Analysis** process, in Figure 1, would only simulate a heterogeneous Poisson process. Including the **Shuffling Survey Cases to Improve**

Spatial Statistics process provides a Cox process approach for realizing the locations of survey responses. The location of realized households are determined both by the density function defined by administrative level enumeration estimates, population density estimates, and the location of other realized households with similar survey responses. When survey responses cluster at scales finer than the available input data, central limit theorem will give variance values as if the population were randomly scattered across the enumeration region, underestimating variability. In an extreme example known to the authors, imagine requiring a demographic analysis of a census block composed of low density housing mostly composed of elderly couples and a large university dormitory. Without accounting for spatial clustering, the census block will average survey responses across that region with mean and variance rates of marital status, income, children, etc. at values inaccurate to either subpopulation. On the other hand, should the realizing process cluster elderly people and students separately, different realizations will have elderly people concentrated in the residential neighborhood or in the dormitory. (The dormitory, visible only as a large institutional residence, would then be perceived as a retirement community.) With a large enough number of Monte Carlo realizations, the survey response's average will remain the same. However, the variance will be larger, and more accurate, because individual realizations will have more extreme values than would be provided. Also, the bimodal distribution of demographic attributes will be shown across realizations, which will accurately be demonstrated in MCS. A Cox process ensures that similarly attributed households are more likely to be located near each other. In this case study, we chose infrastructural attributes to be clustered: households with piped sewer, electricity, or piped water were more likely to be clustered near households containing identical infrastructural attributes. The spatial allocation process and the shuffle households process locates household in the study area maintaining both first- and second-order attribute properties.

In the spatial allocation process, duplicated household survey cases are realized into plausible geographic locations. This process is performed stochastically using the household locations probability surface, enumerated population estimates, and samples of ground truth data. Ground truth data, in this context, is point or polygon locations were survey response characteristics are known to exist in the study area. This phase realizes survey case locations for the entire population of the study area. This point pattern process is discussed throughout the research literature [1, 20]. In the current spatial allocation process, each survey case in a realization is randomly given 10 eligible locations, based on survey response constraint maps and ground truth information. The survey case is placed in the one that will best "improve the fit" of important survey responses. Survey cases are initially placed across the study area fitting first-order properties of census enumerations. In the case study, the important variables are access to utilities, residential structures, household wealth metrics, and religion measures. This algorithm is sensitive to the number of survey cases already realized earlier in the process. The algorithm uses a least squared error approach for both attempting to realize survey cases and the exact enumerations of survey responses in the census data. Demographic modelers can determine a greater weight on census enumerations while virtually ignoring survey response counts or vice versa. If the application is for a year when the Census was done, users should place greater weight on the enumerations. However, if the application simulation year is far removed from an actual census and much closer to the date of the survey, it would be better to add greater weight to the survey case estimates. This step greatly diminishes the sampling methodology drawback to the surveys.

The shuffle households process represents survey response clustering by reducing each response's moment of inertia. The initial moment of inertia is determined by the stochastic placement of survey cases. Survey responses that that are more spatially autocorrelated than the information in available input maps must be given a "clustering parameter" setting a goal to reduce the moment of inertia at various lags of their variograms. Ideally, the demographic modelers would know the true variograms of the survey responses. However, survey collection techniques do not report second-order properties as part of their sampling methodology. Instead, modelers determine the proportional reduction of the moment of inertia at various lags to increase survey responses' spatial autocorrelation by analyzing expert knowledge of the social, environmental, and infrastructural characteristics of the locations against the patterns of household attributes of a heterogeneous Poisson process. The demographic modeler compares the size of neighborhoods with and without electricity to determine the maximum distance of autocorrelation. Demographic modelers also choose the amount of spatial dependence to increase over random placement at short distances. Demographic modelers would have to experiment with different spatial dependence parameter values, measured as moments of inertia in a semi-variogram, to fit the appropriate density of households without electricity in electricity prone areas and vice versa. Without prior knowledge of which remote villages have or will have biogas development,[1] this procedure will cluster positive rural electricity responses in the same villages, leaving other villages without electricity on a realization per realization basis. Since different realizations will have different villages with electricity, the summary statistics maps for across all realizations will retain overall proportions from the original census enumerations while providing a more accurate distribution of realistic survey responses at each location. Other researchers have attempted to eliminate the pattern of the uniform survey response density by using the pycnophylactic method [23]. However, that process and similar methods will underestimate the variability across realizations.

Figure 3 shows a subset of one realization's households in the greater Dhaka metropolitan area. Realizations can be converted into KML files for display in various mapping products, both point based and surface based, to aid the development of models and visually provide ways to understand the population's spatial and attribute distribution.

In experiments in rural environments, survey responses were almost always exactly matched when three or fewer responses were fitted. Fitting six or more survey responses in rural areas inevitably caused some responses to be less precisely fitted. This was expected as results occurred in other similar algorithms recreating multiple statistics [8]. These experiments in Chittagong Division Bangladesh, where the number of survey cases are much higher, required computational loads greater than available in typical scientific workstations, see Table 1.

2.4 Generating proportion maps

The spatial patterns of the responses to each survey question are represented as proportion maps. A proportion map indicates the ratio of people having a survey question's answer to all people answering that survey question:

$$r(x, y) = \frac{f_c(x, y)}{f_p(x, y)} \tag{1}$$

[1]Bangladesh's drive to provide all citizens with electricity is by both expanding the countries National Grid and encouraging remote rural villages to develop Bio Gas plants: http://bbdf.org.

Figure 3: A sampling of Bangladesh IPUMS survey cases' simulated locations from one realization as a KMZ file displayed in Google Earth (in Southern Dhaka). Process generates KMLs for cartographic communication, GeoTIFF for Map Algebra, and CSV files for agent-based model inputs.

Process	Computer	Computing Time
Household Density Analysis	PC Scientific Workstation	3 hours
Urban/Rural Analysis	PC Scientific Workstation	12 hours
Household Realization Process, 147 realizations	ROGER Cluster, 50 CPU cores	6 days
Kernel Analysis Maps (5 kernel bandwidths)	ROGER Cluster, 10 GPU nodes	15 days
Map Algebra (5 kernel bandwidths)	ROGER Cluster, 50 CPU cores	$6\frac{1}{2}$ days

Table 1: Computational requirements for this methodology.

where $r(x, y)$ is the proportion value at (x, y), and $f_c(x, y)$ and $f_p(x, y)$ are the spatial density of survey responses and population respectively.

In this research, the proportion maps are represented as raster maps with regular grid cells. A **Kernel Density Estimation** (KDE), in Figure 1, and **Kernel Analysis for each Map Algebra Theme**, in Figure 2, is used to calculate the spatial density of both the population and survey responses at each grid cell. KDE estimates the spatial density at a location by aggregating the contribution of its surrounding data points (i.e., people) through a distance-decay kernel function—a nearer point has a larger influence than a farther away point. A bandwidth h represents the radius of the surrounding area, which controls the degree of smoothness.

When the same kernel function $k(\,)$ and bandwidth h are used for both the population and survey responses, the final formula used in this research is:

$$r(x,y) = \frac{\sum_{i=1}^{n} k\left(\frac{x-x_i}{h}, \frac{y-y_i}{h}\right) N_i}{\sum_{i=1}^{n} k\left(\frac{x-x_i}{h}, \frac{y-y_i}{h}\right)} \tag{2}$$

where (x_i, y_i) are the locations of n persons, and N_i is the indicator of whether the ith person matches the survey response.

The radius of the scaled kernel should be appropriate for the operational analysis necessary. For example, understanding the population near potential healthcare facilities would require using kernel radii based on the expected distances people would travel to use those facilities. This may incorporate multiple transport modes, including traveling by foot, vehicle, or public transportation. In a situation where multiple distances would define the application's kernel radius, multiple proportion maps should be created with the resulting map being a weighted sum of the individual distances.

KDE provides the appropriate input maps to any map algebra analysis, **Map Algebra 1 to n Realizations**, in Figure 2. Map algebra is the subset of geographic information analysis processes that are performed location by location, originally known as map overlay. More complex forms of geographic information analysis would require algorithms performed on the household or population realizations created after **Shuffling Survey Cases**, in Figure 1. The authors typically use agent-based modeling systems for these more complex analyses because commercial geographic information systems are often overwhelmed by the amount of information (on the same computing hardware platforms).

2.5 Proportion map summary statistics

For important survey responses containing individual attributes, combinations of attributes, or analyses of multiple attributes, a proportion map is generated for each realization. Hence, the empirical distribution of the proportion values at each location (i.e., map cell) can be estimated. **Summary Statistics Calculations**, in Figures 1 and 2, which characterizes such distributions, are then calculated from these proportion maps on a cell-by-cell basis. These summary statistics include minimum, maximum, median, average, 1st quartile, 3rd quartile, standard deviation, and interquartile range for each map cell. When the goal is to provide a best presentation of real world conditions, it is expected that decision makers will generally use median or average statistics in their subsequent analyses. However, they will use minimum, maximum, standard deviation and inter-quartile range (IQR) as a measure of input data utility to help determine potential variability of results. Minimum, maximum, 1st quartile, or 3rd quartile estimates will also be useful when analysis requires ensuring that specific survey response ranges are met. For example, queries such as "where are locations with at least 80% female adults have at least 12 years of education" would result in maps with the probability of attaining that 80% level using minimum proportion maps.

2.6 Survey response mapping

The process described in this article generates three types of survey representations:

1. Multiple realizations of survey cases representing every household in the survey area, suitable for agent-based modeling of socio-cultural behavior. By applying each realization of survey responses to an agent-based model via a bootstrapping process, model designers and users can see the variability of model results inherent from the uncertainties of the demographic input data. See Figure 3 for the Bangladesh IPUMS households from one realization of the results. Analysis of this product will be covered in a different article.
2. Multiple realizations of kernel analysis for useful survey responses. These realizations provide a detailed visualization for response variability caused by input variables errors and uncertainties no matter the application or map. As mentioned earlier, these realizations will allow Monte Carlo simulation of geospatial applications to provide consumer error measures of demographic inputs' uncertainty. Figure 4 illustrates the variability of analysis at different kernel diameters. The smaller the kernel analysis performed, the greater the variability of kernel results.
3. Cell by cell summary statistics of survey responses are useful when those demographic variables are the desired indicators for an application or planning process. The box plot variable maps provide end users an easy to understand way to certify where the original data is useful for that application or planning process.

3 Data description

In this research, five types of data are conflated: choropleth population enumeration estimates, survey responses or microdata, a household or population density map derived from remote sensing as well as infrastructure data, maps of survey response constraints, and samples of ground truth information containing specific survey responses. Population enumerations and surveys or microdata are the two complex datasets used in this analysis. Both are composed of well-defined questions and answers with extensive rules and procedures to ensure accurate results. In the United States, the U.S. Census Bureau collects both types of data with oversight by the U.S. Congress [25]. Outside of the U.S., countries often get the assistance of the U.S. Census International Division or the United Nations Statistics Division. Bangladesh's 2011 census was a collaborative effort between Bangladesh's Bureau of Statistics, the United Nations Population Fund, the European Union, United States Census Bureau, and the United States Agency for International Development [2]. However, exact measures of survey accuracy were not easily available, at least to non-Bengali readers, via the internet. With one of this research's goals to provide near global coverage of neighborhood scale demographic characteristics, we increased the variance of uncertainty in equation 3, $\sigma^2(\mathbf{s})$, to reflect that unknown information.

3.1 Population enumeration

Population enumeration is the process of assigning population estimates in administrative areas at multiple levels. Since many countries collect census information to allocate government funds, censuses are usually the most accurate single source population estimates. However, census data for some countries are incomplete or less accurate for regions under conflict or ungoverned. A careful study of the country and time period of census is necessary to properly account for the uncertainty of this information. For this research's

case study location, Chittagong Division Bangladesh, 2011 census data was downloaded from the Bangladesh Bureau of Statistics website. This dataset contains the enumeration of various age groups, gender, and race for each upazila (analogous to county in the United States). These upazilas are further subdivided into urban and rural subpopulations, which this research representing those areas using imagery and maximum entropy analysis to represent the urbanicity in Bangladesh. We applied urban and rural population trends to the 2011 data in order to estimate the distribution of likely 2015 population variable enumerations using the following formula for each variable in each subdivided upazila.

$$P(\mathbf{s})_{t,v,i} = P(\mathbf{s})_v + \mu(\mathbf{s})_{t,v} + R(\mathbf{s})_{t,v,i} \times \sigma^2(\mathbf{s})_v \tag{3}$$

where:

i is a population realization,
$P(\mathbf{s})_{t,v,i}$ is the realized demographic variable v goal values for time t across the study area $\mathbf{s}$ for population realization i.
$P(\mathbf{s})_v$ is the stated population estimates in a census across the study area $\mathbf{s}$ and demographic variable v,
$\mu(\mathbf{s})_{t,v}$ is the estimated trend from when the census was collected to the time t of the demographic variable v goal values across the study area $\mathbf{s}$,
$R(\mathbf{s})_{t,v,i}$ is a random normal deviate for time t of the demographic variable v goal values across the study area $\mathbf{s}$ for population realization i, and
$\sigma^2(\mathbf{s})_v$ is the variance of the demographic variable v goal values across the study area $\mathbf{s}$.

This process is applied separately for each realization of population enumeration so that known uncertainties of population totals will be accurately reflected in the summary statistics. Each attribute's trending variables can be independent, whether age, gender, access to sanitary water, etc., and may or not be covariated with other attributes to account for known trends in the population. Those variables that are dependent have the equation applied to a single variable with cascading adjustments made to the other dependent variables.

3.2 Surveys or microdata

The data in U.S. PUMS or IPUMS Bangladesh microdata are presented in the form of "typical" households and "typical" population members. The tabular nature of the data is conducive to SQL queries. Any query made across a survey or microdata constitutes a survey response. The following query, for example, would find the subset of people that were Hindu, male, and over the age of 17:

```
select * from IPUMS where RELIGION=HINDU and AGE>17 and SEX=MALE
```

The recent source of U.S. available demographic data is from the American Community Survey (ACS) phone questionnaire [24]. ACS phone surveys are done every year with a data product similar to IPUMS. There are inherent advantages to annual surveys, especially in rapidly changing neighborhoods. As Goldstein et al. [11] and many others have discussed, uncertainty increases with the difference in time when a survey is completed and for when data is needed. In the slums of developing world cities and other rapidly changing environments, the frequency of surveys is more critical for accurate demographic

forecasts. By relying on four year old Bangladesh IPUMS data, the uncertainty represented by equation 3 is magnified significantly than if we were trying to represent Chittagong Division in 2011 or even 2013. Since the primary use of this analysis is to support future humanitarian aid and disaster relief (HA/DR) operations, it is important that map results forecast population trends into the future. Thus, if a plan improving HA/DR response will require three years to implement, population forecasts to time of implementation should be simulated.

When comparing Bangladesh census enumeration attributes against Bangladesh IPUMS, specific survey responses do not have the same proportions in both datasets: it is critical to perform population weighting adjustment [15] on microdata. Forecasting future Bangladesh population growth will also cause a divergence between the enumerations and the microdata. As the proportions of Bangladesh citizens gaining access to electricity, better educational outcomes, and other demographics, sampled households with lower quality attributes will need to weight adjusted lower. Survey results, which often have not been adjusted by statisticians, require weighting adjustments even more so. Measure DHS surveys from USAID, for example, contain many health related indicators not available in Bangladesh IPUMS.

3.3 Survey case density map

Population or household density maps, usually gridded raster layers, represent how many households exist within specific areas, usually as a measure of people per square km. (While most of the indicators used to create these maps are indicative of household density, making household density a more accurate phrase, the term population density is almost always used to describe them.) Household density is estimated using land use, topography, distance to transportation, and other factors that will attract or distract where people want to live. With the goal of neighborhood scale maps, $15\,\mathrm{m}$ resolution imagery, distance to all roads, topographic slope and aspect, and Bangladesh census enumeration values were used in a maximum entropy analysis to determine household locations. Table 2 shows the maximum entropy analysis results to create the density surface. The density surface is used to stochastically locate survey cases. When survey case locations are simulated within each enumeration zone, upazilas subdivided into urban and rural areas for the Bangladesh case study, each survey case is initially placed based on the relative weight of the household density map.

One of the largest potential uncertainty issues regarding household density maps is the speed at which urbanization changes. Especially in dynamic urban environments, newly developed subdivisions will be treated as vacant lands. Stochastically driven urban growth modeling provides forecasts of future land use patterns [11,29]. The Monte Carlo methods are well suited for different household density maps to be used with each realization of survey cases.

3.4 Survey response constraint maps

The spatial allocation and shuffle households Monte Carlo process allows for any number of constraints to be incorporated. These constraints can either be represented in the form of binary maps or point locations of specific survey responses. Maps associated with specific survey responses will ensure those locations are populated with survey cases with those

survey responses. For example, detailed infrastructure maps capturing publicly access to sanitary water will ensure that households with public water will only be located in those locations. Since many survey responses are collinear with each other, such as households having refrigerators is collinear with the households consuming electricity from a commercial grid, constrained survey responses will both accurately locate the geolocated response, electricity from a commercial grid in this example, as well as the collinear survey attributes (household refrigerator). Survey response constraint maps are only useful when the detail of the maps is finer than the census enumeration locations, upazilas for this case study.

3.5 Ground truth samples

Examples of ground truth information are also incorporated into process. Geolocated survey responses are matched to households with those characteristics into the spatial allocation process. During the household shuffling process, second-order properties will be fitted so that estimated clusters of responses will more likely be located in areas with examples of ground truth information. This process works well in rural low populated areas. Unfortunately, properly clustered survey responses in dense urban environments will require computational resources in excess of current personal computers. In order to the keep the fitting statistics accurate, the current technique has survey case swapping done sequentially, preventing this process from exploiting parallel processing techniques.

4 Case study

This section presents the process of developing survey response maps to be used in a complex analysis for Chittagong Division Bangladesh. Each country has unique census techniques and attributes aligned to specific national goals. These specific data layers require three design decisions that needed to be made:

1. How to represent urbanicity and population density using maximum entropy analysis?
2. How many demographic variables to fit in the population representation process? and
3. Which demographic variables should be spatially autocorrelated in order for their second-order properties to be accurately represented in the realization process?

As discussed earlier, neglecting environmental factors that influence the attractiveness of a site for a household to locate will reduce the quality of the population density map, and more importantly reduce the accuracy of forecasting where urban and rural populations live. The second decision will determine which variables will be fully correlated between the census enumerations and the survey cases. Demographic variables not linked between enumerations and survey cases will still be somewhat correlated if those variables are correlated to variable chosen to be linked. The third decision has two effects on the results: to provide more realistic spatial patterns of the households' variables for agent based models, and to increase the variability of household variable proportions across realizations.

4.1 Weighting survey cases

For the case study illustrated in this paper, IPUMS questions for toilet facilities, residential building types, electricity availability, household water supply type, house ownership, urbanicity, and individual religious preference were weighted against population enumerations provided by the country census enumerations at administrative level four, upazilas. The upazilas were further subdivided into urban and rural areas. Since the census enumerations and IPUM survey responses included urban and rural designations, this information can then easily be clearly shown on the outputs.

4.2 Maximum entropy analysis

Maximum entropy analysis was used to determine household density required many choices to be made. Initially, potential factor map layers were standardized to 15 m cell resolution. We identified three types of housing in Chittagong Division: scattered homes in very rural settings, rural homes in small villages, and urban housing. Several hundred of each "species" of housing were geolocated to be used in the maximum entropy analysis training. We first performed maximum entropy analysis on 23 map layers, choosing eight of the map layers with the greatest permutation importance via jackknife tests. Maximum entropy analysis was rerun recalculating the contribution amounts, which was used for determining population density. Then, the maximum entropy analysis was calibrated to household density by raising the results to the 4th power to convert the relative suitability values into a more accurate distribution of household density. This process was chosen to provide higher precision population density than open source alternatives. Finally, grid cells within each upazila were declared to be urban or rural. The most densely populated cells were given urban status until the ratio of urban to rural density values equaled the upazila's ratio of urban households to rural households.

Map layer	% contribution to density	Permutation importance
Distance to Roads	84.8%	60.8%
Haralick Texture Band 2 Variance	5.4%	2.8%
TM Natural View Band 1	2.5%	22.7%
TM Natural View Band 2	0.3%	2.6%
TM Natural View Band 3	1.2%	7.4%
Haralick Texture Band 6 Different Entropy	5.6%	3.3%
Natural View Landsat Mosaic	0.2%	0.3%

Table 2: Seven maximum entropy analysis weighting factors chosen from 23 factors used to determine household density and the urban/rural divide within upazilas.

4.3 Survey case location analysis and shuffling survey cases

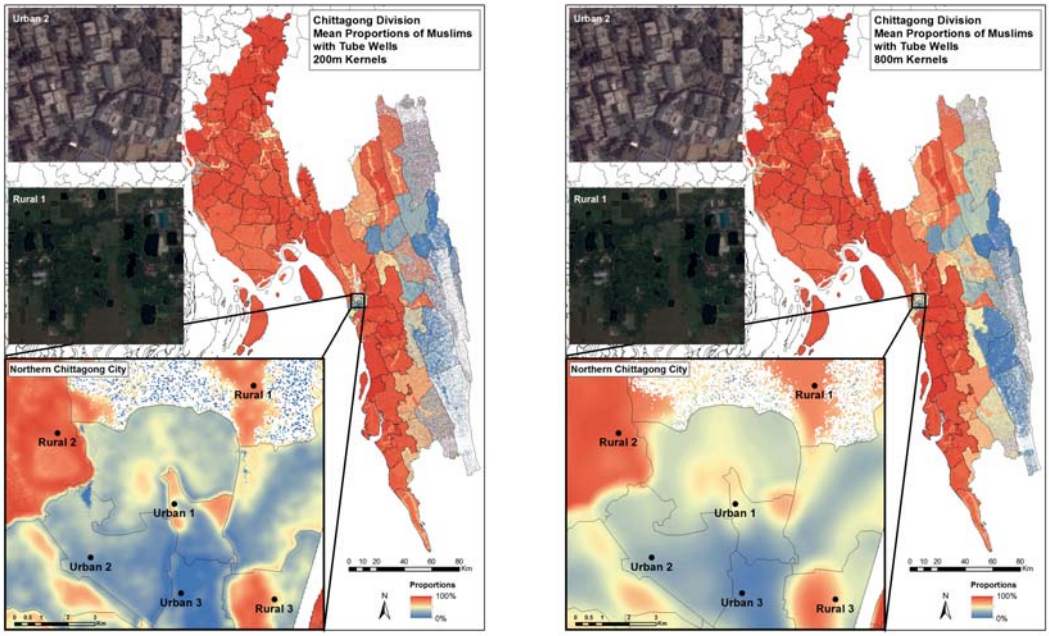
After household density across the study area is determined, household cases were initially located to match the same characteristics fitted when weighting those cases. For this

analysis, toilet facilities, residential building types, electricity availability, household water supply type, house ownership, urbanicity, and individual religious preference were fitted to the census enumerations of each upazila, split into urban and rural sections. We could have chosen to fit additional survey responses, however, increasing the criteria being fitted will reduce the overall quality of fit.

Once satisfactory location fitting values were observed, households were exchanged whenever spatial dependence of important survey responses were improved. In this case study, piped sewer, electricity, and piped water were more likely to be clustered near households containing identical infrastructural attributes.

4.4 Kernel density estimation

Figures 4a and 4b demonstrate the correlation between tube well water access and urbanicity with tube well water access proportions following rural areas. Tube wells are significant in HA/DR planning due to the increased likelihood of water contamination during flood events.



(a) (b)

Figure 4: Left (a) Bangladesh IPUMS survey probability density surface realizations for Muslim households getting their water via tube wells with kernel radii of 200 m, and right (b) 800 m.

4.5 Map algebra

This Bangladesh case study demonstrates the ability to understand neighborhood level information of complex IPUMS survey responses. There are 120 proportion map themes that are easily generated from the Bangladesh 2011 IPUMS individual responses alone. Figures 4a and 4b give an example of one of these map themes. For specific end users, individual survey response proportion maps and combinations of survey responses proportion maps must be created. One stakeholder for this research identified 47 social, environmental, and infrastructural indicators for monitoring the threat of potential humanitarian crises in Bangladesh. Of these 47 humanitarian crises indicators, we identified 14 indicators that are completely or mostly measurable from the IPUMS survey responses. If we include other surveys in Bangladesh, namely USAID's Demographic and Health Surveys as well as a proprietary survey, 26 of 47 indicators are modeled. When we combine all of our data sources and geographic analysis techniques, fully 44 of the 47 indicators are completely or partially represented. Our definition for "partially represented" is that the indicator's model represents a significant proportion of the desired indicator, but is not a 100% correlation. Thus, partially represented indicators can only be used as a proxy until either 1) better data becomes available, 2) that survey questions are adjusted to meet that indicator's needs, or 3) the Humanitarian Crisis Framework of the stakeholder is adjusted to account for available data. (Their framework was devised independent of knowing what information was available.) Some of the humanitarian crisis indicators include:

- refugee status,
- access to doctors,
- water shortfalls,
- energy deficits,
- lack of established civil authority,
- inadequate sanitation, and
- undernourished populations.

One of indicators that the stakeholder identified to monitor humanitarian crises is relative resource inequality between religious groups. To understand where resource deprivation or resource inequality exists, we mapped the relative difference in household resources (using household infrastructure access as the proxy) between the (minority) Hindu and (majority) Muslim population, Figure 5a. Based on the descriptions of house type, access to electricity, source of clean water, and sewage disposal, we weighted the quality of each survey response as shown in Table 3, summing the weights in each survey case.

We used a variation on a favorability function [3] to estimate relative resource deprivation:

$$I(\mathbf{s}) = F(\mathbf{s})_m / F(\mathbf{s})_h, \quad F(\mathbf{s})_i = \prod_{N=1}^{n} X_{N,i}(\mathbf{s}), \quad i = \{m, h\} \tag{4}$$

where:

$I(\mathbf{s})$ is the ratio of Muslim, m, household wealth at each location $\mathbf{s}$ in the study area to Hindu, h, household wealth,

$F(\mathbf{s})_i$ is the favorability of ethnicity i evaluated as a continuous value between 0.0–1.0 at each location $\mathbf{s}$ in the study area, and

$X_{N,i}$ is the value at $\mathbf{s}$ in the input map N coded to values between 0.0–1.0 with 1.0 being optimal and 0.0 being unable to sustain life.

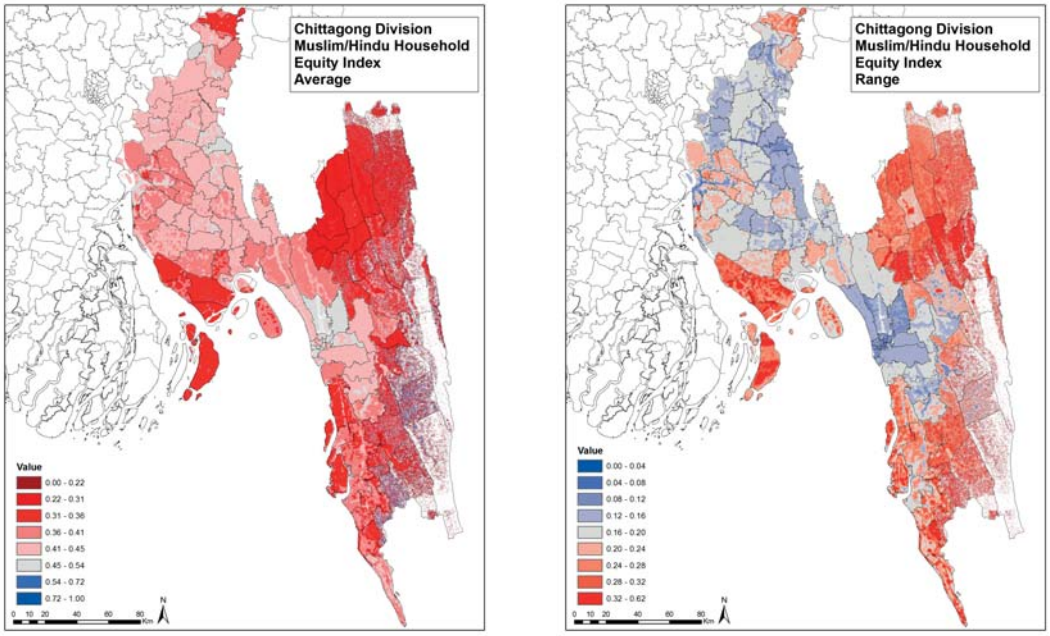
Survey Question and Responses	Range of Values
Source of drinking water	
Tap	1.0
Tube-well	0.8–0.6
Other	0.4–0.2
Electricity Connection	
Yes	1.0
No	0.6–0.2
Toilet Facilities	
Sanitary (with water seal)	1.0
Sanitary (no water seal)	0.9–0.7
Non-sanitary	0.6–0.3
None	0.2–0.1
Home Ownership	
Owned	1.0
Rented	0.8–0.5
Rent-free	0.4–0.2
Type of House	
Pucka (permanent, brick and concrete)	1.0
Semi-pucka (semi-solid, mostly wood)	0.9–0.8
Kutcha (mud/bamboo)	0.6–0.4
Jhupri (makeshift)	0.3–0.1

Table 3: Weights of survey responses indicating quality based on IPUMS responses.

4.6 Summary statistics calculations

Kernel analysis was performed on all realizations at both 200 m and 800 m radius, summarizing the results. Summary statistics is performed across realizations to determine the applicability of the IPUMS survey to the case study.

The blue areas in Figure 5a are where Muslims are substantially wealthier as compared to Hindus, and red is the opposite extreme. The gray is where wealth is more or less equal. It can be clearly seen where zones in which one religious group is wealthier than its counterpart in that 1600 m diameter area. Each of the grid cells on this map is 50 m although precise results could be performed at resolutions as fine as 15 m. There are two sources of error for this application: from the data and from the understanding of the application. Table 3 represents the uncertainty of application's parameters in whether the modeler's knowledge of the population's relative value placed on each attribute relative to the highest quality attribute. (Since the application modelers have never lived in Bangladesh and do not have intimate knowledge of the quality of each of the survey responses, they gave a wide range of relative values for each of the responses.) Figure 5b shows the range of application results using the maximum variation of parameter values. Blue areas, which are both urban and very densely populated, usually have high quality infrastructure and little variation between minimum and maximum parameter estimates. Rural and lightly populated upazilas have greater variation. Unlike in Dhaka Division, where different regions favored wealth inequity for Muslims, usually in areas with government housing,



(a) (b)

Figure 5: Left (a) Muslim/Hindu household equity index average with 800 m kernel radius, right (b). The range of the index is the spread of index values across all realizations, providing a measure of the application uncertainty at every location in the study area.

and wealth inequity favoring Hindus, usually in productive agricultural areas, Chittagong Division has no regions favoring Muslims, when accounting for the range of uncertainty from application error.

Figure 6 shows the variation of survey responses a different spatial scales to help understand uncertainty from the data. By analyzing these bar charts, one can determine the smallest neighborhoods that can be accurately analyzed. The bar charts reflect the variation of application results across all data realizations. Areas with larger ranges reflect zones where you would want to find better information to build the demographic model or perform the analysis with a longer kernel radius. From a decision makers' point of view, most neighborhoods in Chittagong have enough demographic information from the IPUMS responses to determine the level of inequity if they considered that Hindus would perceive their neighborhood to be those homes within 800 m of their dwelling. However, application results are so varied with neighborhoods smaller than 800 m in both urban and rural areas of Chittagong Division to prevent decision makers from being confident their analysis results are correct. These results surprised the demographic modelers who assumed that the more densely urban areas would have smaller ranges. (When a similar analysis was done in Dhaka Division, a small sampling of urban and rural locations indicated greater variation in rural areas.) If the decision maker needed higher quality data, they could include constraint maps of utilities or other survey specific household attributes and rerun the process to generate higher quality maps. The software developed as part of this

research automates the creation of the proportion maps that measure these indicators. As input data layers are improved, updated, or added, all maps can be rebuilt automatically.

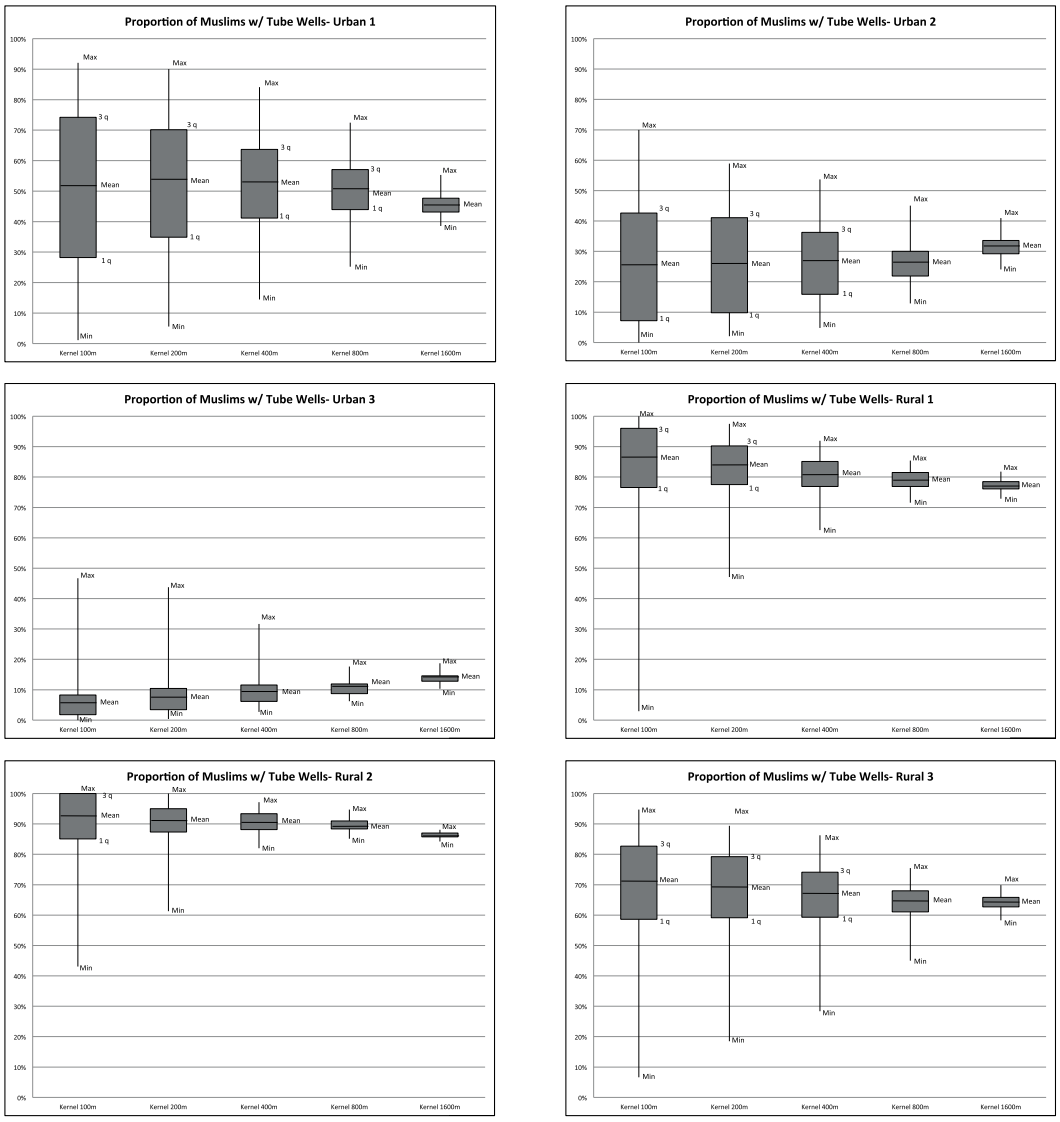
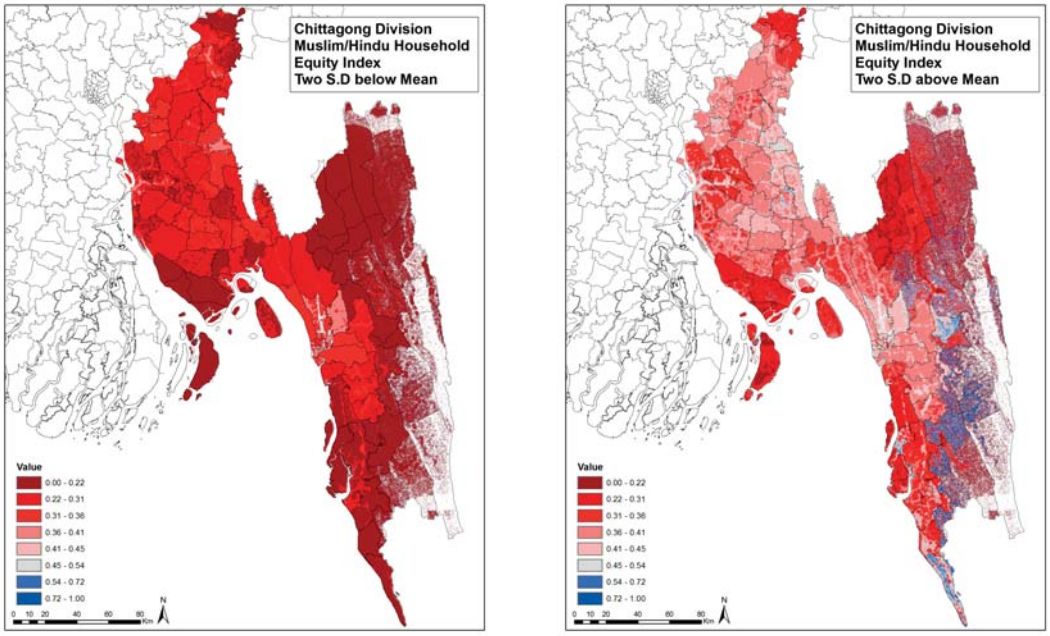


Figure 6: Bangladesh IPUMS survey variability across realizations for Muslims with access to water from tube wells with kernel radii of 100 m, 200 m, 400 m, 800 m, and 1600 m for three urban and three rural locations. Generally speaking, kernel radii of 800 m or more gave reasonably accurate results in denser urban and rural areas, but not in low household density environments.



(a) (b)

Figure 7: Left (a) Estimated data error rates with 800 m kernel radius, shows two deviations below the mean of Muslim/Hindu household inequity, right (b) shows two deviations above the mean. Red areas have maximum Hindu household wealth inequity, blue areas has maximum Muslim household wealth inequity, gray areas have equal equity (0.5 is equal).

5 Conclusion and discussion

The ability to convert any survey into a neighborhood scale survey response maps has benefits for survey designers, socio-cultural analysts, and decision makers. This research is still evolving towards the goal of all socio-cultural analyses automatically retaining the errors and uncertainties of input data. The next two sections explain some of these benefits and challenges.

5.1 Benefits of mapping survey responses

The need for mapping socio-cultural indicators is a critical component for allowing non experts to understand the complexity of demographic mapping. Each of the separate input data layers contains different types of errors and uncertainties. Traditional measures of error of the various input layers were developed for the producers of the data, and are mostly useless for end users. A technique to represent the combination of all errors in a way that end users will understand the impacts on their application will improve the utility of the information for decision makers. Figure 7 presents the spread of application results across a range of four standard deviations, quickly communicating where in Chittagong

Division that not enough data was collected to provide useful results at neighborhood sizes of 1600 m in diameter.

The biggest benefit to this technique is that any combination of survey responses can be represented as a surface map of any scale without end users worrying whether ecological inference fallacy is being committed. Figures 8 and 9 compare the results of collecting survey responses using traditional SQL queries in IPUMS data versus this MCS technique. Figure 8 was prepared by querying IPUMS data on an upazila by upazila basis to generate a survey response because the census enumerations did not have literacy information. The large differences in literacy rates, especially in mostly rural low populated upazilas are typical from this type of analysis and reflects the lack of population weighting adjustment to obtain more accurate results. Figure 9 analysis of Muslim literacy reflects the automatic population weighting adjustment in both rural and areas, providing better estimates of actual rates in more precise areas. Of course, in very low populated areas, an 800 m kernel is too small to encompass enough people to ensure the standard deviation is low enough to be useful for decision makers, Figure 9b. In that case, the analysis should be done at much higher kernel radii until the margin of error is adequate for the analysis.

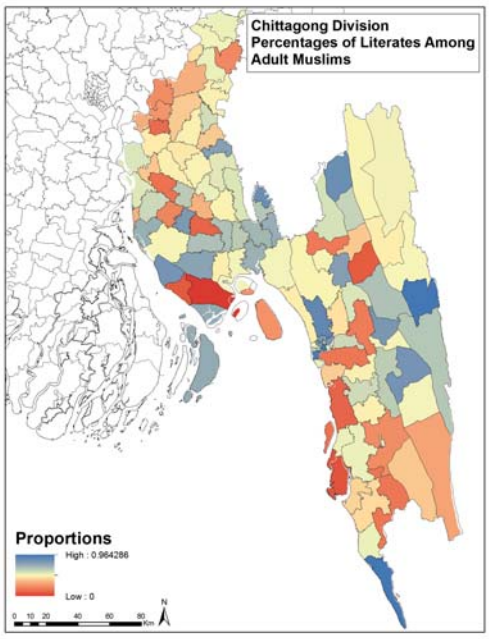
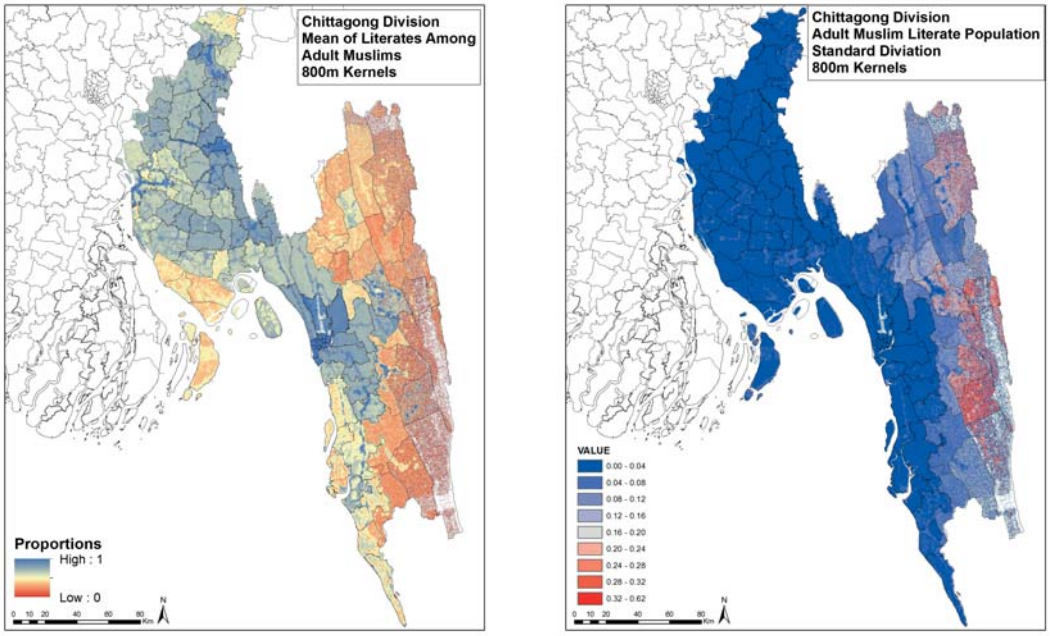


Figure 8: Choropleth map for literacy rate of Muslim adults.

The software developed for this technique generates both GeoTIFF grids and KML files for each survey response desired. The software is designed so that the entire process is automated once the demographic model is developed. That way, whenever changes to the input data layers are made, a single script will recreate all of the output maps. For example, creating the survey responses for the Bangladesh IPUMS survey generates 120 single response maps, with any number of multiple survey responses possible. The case study described earlier required a short Linux shell script.



(a) (b)

Figure 9: Left (a) Adult Muslim literacy rate map accounting for sub-upazila information, right (b) accuracy map.

Another potential benefit to this process is the ability for survey developers to form a more accurate understanding of how many survey cases need to be collected for their surveys to be useful for known users. The maps created provide precise measures of where information is most useful and less useful, with error bars at all locations. Of course, as more samples are collected, error bars will get smaller. Also, survey collectors could better proportion urban and rural subpopulations to minimize the larger error bars of the most important survey responses.

5.2 Areas for further research

While this set of techniques provides significant improvements in representing population information at the neighborhood scale, opportunities exist to further improve the results. One area of potential improvement is in modeling the second-order properties of the population density and demographic indicators. Population density becomes increasingly important when representing finer resolution details of demographic indicators in dense urban environments. The population density's second-order properties are modeled by semi-variogram, and the variogram lags are normally differentiated by distance (on a ratio level of measurement [22]). Instead of defining the lags by set distances, if lags were sets of households defined by proximity—the closest 10 households form the 1st lag, the 11th–20th closest households compose the 2nd lag, etc.—then the dense urban areas'

second-order properties would be prevented from dominating the second-order properties in rural areas.

Parallel processing techniques, while used in various components of these techniques, need to be used in all components of this process to ensure that recreating the population realizations and survey response proportion maps will be done quickly as new ground truth information is made available. Since this research generates explicit uncertainty information tied to geographic locations, collecting survey responses in areas with higher uncertainty ranges for important indicators will occur to improve map utility. Due to the massive amounts of computations required to perform this analysis, Table 1, this technique can only be performed on a dedicated supercomputer. Operational use of these techniques will require dozens or hundreds of thematic maps, which will need updating whenever input maps or additional ground truth information is collected. Ongoing research in high performance computing environments is necessary before this technique are easily implemented in production settings.

If there is no conditional ground truth demographic information used in the analysis, the fidelity of output results will be no finer than intersection of the input data layers. For example, a map including locations where specific types of utilities were available as one of the inputs would ensure that a greater proportion of the households with that level of utility access would only be located in those areas. Thus, the fidelity of survey responses would be accurate to both upazila boundaries as well as utility boundaries. Since this analysis separates urban and rural households within each upazila, that fidelity is also included in the results. Without explicit cartographic or geographic information analysis techniques, the first- and second-order effects of survey responses to all specified geographic boundaries are accounted for in the final products.

Based on the needs of the application, any data uncertainty model may be quite simple or very complex. Scientists usually try to determine the simplest model to explain a phenomenon, while data conflation techniques are by definition complex. The complex set of techniques used in this approach is necessary to model uncertainty may generate realizations that have little probability of being an actual representation of reality at very small geographic distances. The stochastic process of simulating household locations likely has many clumps of households too close to each other to be physically possible. These clumps require that agent-based modelers do not design their computational models requiring accurate micro-density to perform. The process of generating kernel density maps at multiple magnitudes of area are not affected by this problem. A simple fix in the future would be to ensure that households are not placed within calculated distance of other households during that phase.

Acknowledgments

The authors thank the many people who made this research possible. We especially appreciate the continuing efforts of the Minnesota Population Center at the University of Minnesota for making Bangladesh's household data, as well as many other countries' data, easy to download and use. Also, Dr. Shaowen Wang at the CyberGIS Center at the University of Illinois gave the authors great advice and access to the ROGER supercomputer, without which we would not have been able to complete this research. The ROGER supercomputer is supported by NSF grant number: 1429699.

References

[1] BAILEY, T. C., AND GATRELL, A. C. *Interactive spatial data analysis*, vol. 413. Longman Scientific & Technical Essex, 1995. doi:10.2307/2265559.

[2] BANGLADESH BUREAU OF STATISTICS. Population and housing census 2011, socio-economic and demographic report. Bangladesh National Series, Volume 4, 2012.

[3] BONHAM-CARTER, G. F. *Geographic information systems for geoscientists: modelling with GIS*, vol. 13. Elsevier, 2014. doi:10.1016/B978-0-08-041867-4.50001-1.

[4] COBB, M. A., CHUNG, M. J., FOLEY III, H., PETRY, F. E., SHAW, K. B., AND MILLER, H. V. A rule-based approach for the conflation of attributed vector data. *GeoInformatica 2*, 1 (1998), 7–35. doi:10.1023/A:1009788905049.

[5] CRAGLIA, M. Cogito ergo sum or non-cogito ergo digito? The digital city revisited. *Environment and Planning B: Planning and Design 31*, 1 (2004), 3–4. doi:10.1068/b3101ed2.

[6] DUDIK, M., PHILLIPS, S. J., AND SCHAPIRE, R. E. Performance guarantees for regularized maximum entropy density estimation. In *International Conference on Computational Learning Theory*, Springer, pp. 472–486. doi:10.1007/978-3-540-27819-1_33.

[7] EHLSCHLAEGER, C. Incorporating second-order properties for cluster detection analysis and agent based modeling. In *GeoComputation* (2005). http://www.geocomputation.org/2005/Ehlschlaeger.pdf.

[8] EHLSCHLAEGER, C. R. Representing multiple spatial statistics in generalized elevation uncertainty models: Moving beyond the variogram. *International Journal of Geographical Information Science 16*, 3 (2002), 259–285. doi:10.1080/13658810110099116.

[9] ELITH, J., PHILLIPS, S. J., HASTIE, T., DUDÍK, M., CHEE, Y. E., AND YATES, C. J. A statistical explanation of MaxEnt for ecologists. *Diversity and distributions 17*, 1 (2011), 43–57. doi:10.1111/j.1472-4642.2010.00725.x.

[10] GARCIA, A. J., PINDOLIA, D. K., LOPIANO, K. K., AND TATEM, A. J. Modeling internal migration flows in sub-Saharan Africa using census microdata. *Migration Studies* (2014), mnu036. doi:10.1093/migration/mnu036.

[11] GOLDSTEIN, N. C., CANDAU, J., AND CLARKE, K. C. Approaches to simulating the "march of bricks and mortar". *Computers, Environment and Urban Systems 28*, 1 (2004), 125–147. doi:10.1016/s0198-9715(02)00046-7.

[12] HEUVELINK, G. B. *Error propagation in environmental modelling with GIS*. CRC Press, 1998.

[13] HOLT, D., STEEL, D., TRANMER, M., AND WRIGLEY, N. Aggregation and ecological effects in geographically based data. *Geographical analysis 28*, 3 (1996), 244–261. doi:10.1111/j.1538-4632.1996.tb00933.x.

[14] JAYNES, E. T. Information theory and statistical mechanics. *Physical review 106*, 4 (1957), 620. doi:10.1103/physrev.106.620.

[15] KALTON, G. Standardization: A technique to control for extraneous variables. *Applied Statistics* (1968), 118–136. doi:10.2307/2985676.

[16] KALTON, G. Compensating for missing survey data. Tech. rep., Insitute for Social Research, University of Michigan, 1983. http://www.popline.org/node/412157.

[17] KISH, L. Weighting: Why, when, and how? In *Proc. Survey Research Methods Section* (1990), pp. 121–130.

[18] MCGILL, R., TUKEY, J. W., AND LARSEN, W. A. Variations of box plots. *The American Statistician 32*, 1 (1978), 12–16. doi:10.1080/00031305.1978.10479236.

[19] METROPOLIS, N., AND ULAM, S. The Monte Carlo method. *Journal of the American Statistical Association 44*, 247 (1949), 335–341. doi:10.2307/2280232.

[20] O'SULLIVAN, D., AND UNWIN, DAVID, J. *Geographic Information Analysis*. Willy Interscience, 2003. doi:10.1002/9780470549094.

[21] PHILLIPS, S. J., ANDERSON, R. P., AND SCHAPIRE, R. E. Maximum entropy modeling of species geographic distributions. *Ecological Modelling 190*, 3–4 (2006), 231–259. doi:10.1016/j.ecolmodel.2005.03.026.

[22] STEVENS, F. R., GAUGHAN, A. E., LINARD, C., AND TATEM, A. J. Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PloS one 10*, 2 (2015), e0107042. doi:10.1371/journal.pone.0107042.

[23] TOBLER, W., DEICHMANN, U., GOTTSEGEN, J., AND MALOY, K. World population in a grid of spherical quadrilaterals. *International Journal of Population Geography 3*, 3 (1997), 203–225.

[24] UNITED STATES. CONGRESS. HOUSE. COMMITTEE ON GOVERNMENT REFORM. SUBCOMMITTEE ON THE CENSUS. The Census Bureau's proposed American community survey (ACS). Hearing before the Subcommittee on the Census of the Committee on Government Reform, Serial no. 107-9, June 13 2000.

[25] UNITED STATES. CONGRESS. HOUSE. COMMITTEE ON GOVERNMENT REFORM. SUBCOMMITTEE ON THE CENSUS. Oversight of the 2000 census: Examining the status of key census 2000 operations. Hearing before the Subcommittee on the Census of the Committee on Government Reform, Serial no. 106-139, February 8 2000.

[26] UNITED STATES. CONGRESS. HOUSE. COMMITTEE ON GOVERNMENT REFORM. SUBCOMMITTEE ON THE CENSUS. Oversight of the 2000 census: Status of non-response follow-up and closeout. Hearing before the Subcommittee on the Census of the Committee on Government Reform, Serial no. 106-225, June 22 2001.

[27] UNITED STATES. CONGRESS. HOUSE. COMMITTEE ON GOVERNMENT REFORM. SUBCOMMITTEE ON THE CENSUS. The success of the 2000 census. Hearing before the Subcommittee on the Census of the Committee on Government Reform, Serial no. 107-7, February 14 2001.

[28] WALD, L. Some terms of reference in data fusion. *IEEE Transactions on Geoscience and Remote Sensing 37*, 3 (1999), 1190–1193. doi:10.1109/36.763269.

[29] WESTERVELT, J., BENDOR, T., AND SEXTON, J. A technique for rapidly forecasting regional urban growth. *Environment and Planning B: Planning and Design 38*, 1 (2011), 61–81. doi:10.1068/b36029.