ERDC/CERL TR-18-32

US Army Corps of Engineers®
Engineer Research and Development Center

ERDC
ENGINEER RESEARCH & DEVELOPMENT CENTER

Military Facilities Engineering Technology

Conflating Survey Data into Sociocultural Indicator Maps

Charles R. Ehlschlaeger, Jeffrey A. Burkhalter, Natalie R. Myers, Carey L. Baxter, Matthew D. Hiett, Ellen R. Hartman, Scott A. Tweddale, James D. Westervelt, Robert C. Lozar, Yizhao Gao, Dandong Yin, Marina V. Drigo, and David A. Brown

October 2018

Construction Engineering Research Laboratory

The U.S. Army Engineer Research and Development Center (ERDC) solves the nation's toughest engineering and environmental challenges. ERDC develops innovative solutions in civil and military engineering, geospatial sciences, water resources, and environmental sciences for the Army, the Department of Defense, civilian agencies, and our nation's public good. Find out more at www.erdc.usace.army.mil.

To search for other technical reports published by ERDC, visit the ERDC online library at http://acwc.sdp.sirsi.net/client/default.

Conflating Survey Data into Sociocultural Indicator Maps

Charles R. Ehlschlaeger, Jeffrey A. Burkhalter, Natalie R. Myers, Carey L. Baxter, Matthew D. Hiett, Ellen R. Hartman, Scott A. Tweddale, James D. Westervelt, and Robert C. Lozar

Construction Engineering Research Laboratory
U.S. Army Engineer Research and Development Center
2902 Newmark Drive
Champaign, IL 61822

Yizhao Gao and Dandong Yin

Department of Geography and Geographic Information Science
University of Illinois at Urbana-Champaign
1301 W Green Street
Urbana, IL 61801

Marina V. Drigo

PERTAN Group Inc.
44 E Main St., #403
Champaign, IL 61820

David A. Brown

U.S. Pacific Command Joint Intelligence Operations Center
Camp Smith, HI 96701

Final report

Prepared for Assistant Secretary of the Army for Acquisition, Logistics, and Technology
103 Army Pentagon
Washington, DC 20314-1000

Under Project P2 458304, “Framework for Integrating the Complexity of Urban Systems (FICUS)”

Abstract

This report 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 used to combine any number of survey responses to stochastically weight survey cases and to simulate survey-case locations in a study area. Through these methods, known errors from each input source can be retained. By keeping the individual survey case 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 using data and output maps for Chittagong Division, 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 for each input source via the automatically generated summary statistics created at the application's geographic scale.

Contents

Figures and Tables

Figures

Tables

Preface

This study was conducted for the Office of the Assistant Secretary of the Army for Acquisition, Logistics, and Technology under Research, Development, Test, and Evaluation (RDT&E) Program Element 622784T41, "Military Facilities Engineering Technology"; Project P2 458304, "Framework for Integrating the Complexity of Urban Systems (FICUS)." The technical monitor was Ritchie L. Rodebaugh, CEERD-TZT.

The work was performed by the Land and Heritage Conservation Branch of the Installations Division (CEERD-CNC), U.S. Army Engineer Research and Development Center, Construction Engineering Research Laboratory (ERDC-CERL). At the time of publication, Dr. Michael L. Hargrave was Chief, CEERD-CNC; Michelle J. Hanson was Chief, CEERD-CN; and Ritchie L. Rodebaugh, CEERD-TZT was the Technical Director for Geospatial Research and Engineering. The Deputy Director of ERDC-CERL was Dr. Kirankumar Topudurti and the Director was Dr. Lance D. Hansen.

The Commander of ERDC was COL Ivan P. Beckman and the Director was Dr. David W. Pittman.

1 Introduction

1.1 Background

Military conflicts and other events involving the presence or attention of the U.S. Army increasingly occur in and around large urban environments located across most continents. Such missions require effective and ongoing urban characterization and an understanding of the people who live there. This effort requires data from many disparate sources such as population and housing census surveys. Spatial and attribute differences among various data sources compel users to resolve inconsistencies before using the data or being restricted to variables within a single survey instance. This problem can be addressed with *conflation tools*, which help users to reconcile data obtained from multiple sources and achieve the best possible data quality for analysis and mapping. Researchers at and working with the U.S. Army Engineer Research and Development Center, Construction Engineering Research Laboratory (ERDC-CERL) are developing an analytical methodology called the Framework for Integrating the Complexity of Urban Systems (FICUS), which is intended to provide data synthesis and integration capabilities to help military planners better understand megacities and other dense urban environments.

The potential contributions of FICUS can be illustrated by two figures shown on the next page. Figure 1 shows a map of literacy among adult Muslims obtained from the 2011 Bangladesh national census for each *upazila* (an administrative district analogous to county in the United States). However, additional data sources can be incorporated to better inform users about the location and prevalence of literacy among Muslims. Examples include the World Development Indicators reported by the World Bank and graduation reports from local educational institutions. By fusing multiple data sets together, the intent is to generate a more accurate and informative data layer, such as the one pictured in Figure 2. That figure represents a conflation of the Bangladesh national census (2011), the USAID Demographic and Health Survey (2012), and the DoD Vulnerable Population Survey (2013). Under the FICUS research effort, ERDC researchers sought to develop a methodology for representing the combination of sociocultural data layers in a way that end users will readily understand, and that will ultimately improve the utility of the information available for characterizing dense urban environments.

Figure 1. Literacy rate of Muslim adults (Bangladesh National Census, 2011).

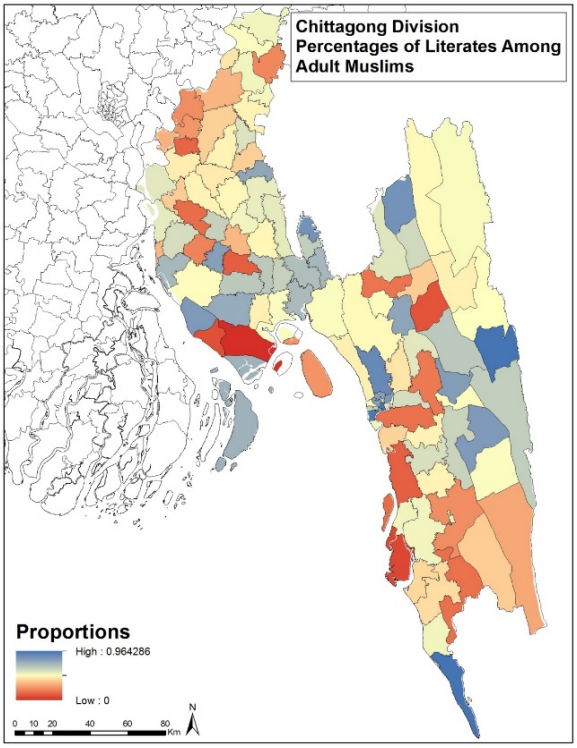
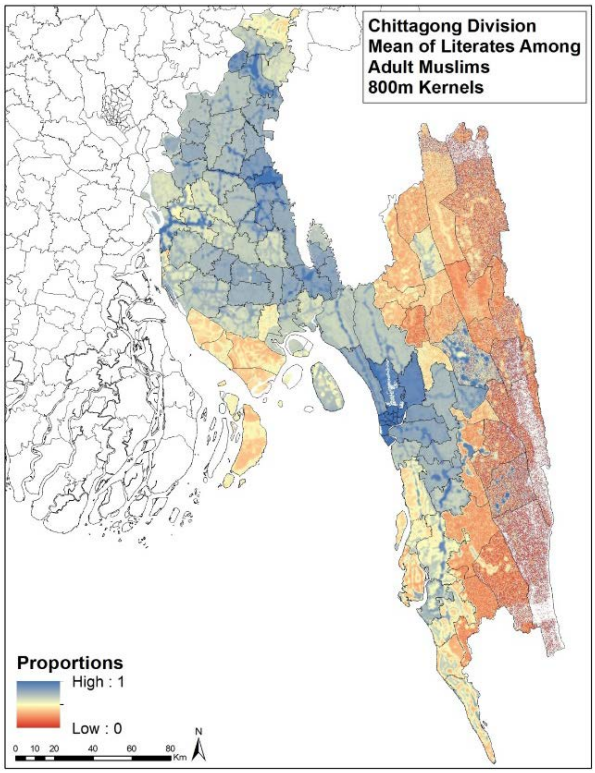


Figure 2. Literacy rate of Muslim adults obtained by conflating survey data.



1.2 Objective

The specific objective of this work was to develop a functional data conflation model to convert massive demographic databases into map layers containing social, infrastructural, and environmental metrics that could be aligned specifically to operational use.

1.3 Approach

The research draws significantly on multiple disciplines, including survey design and statistics, demographic modeling techniques, spatial statistics, habitat modeling, and spatial data uncertainty modeling. Details of the methodology are presented in Chapter 2, and the types of input data sources for the process are specified in Chapter 3. In Chapter 4, the data conflation and map-generation processes are explained, and Chapter 5 presents and discusses example outputs of the process.

2 Methodology

2.1 Data conflation process

A conflation model (Chung et al. 1998) is the foundation for generating plausible locations of every person and household within a study area. The conflation process combines multiple data layers to generate a m containing the most useful aspects of each layer. In this work, five types of data are conflated:

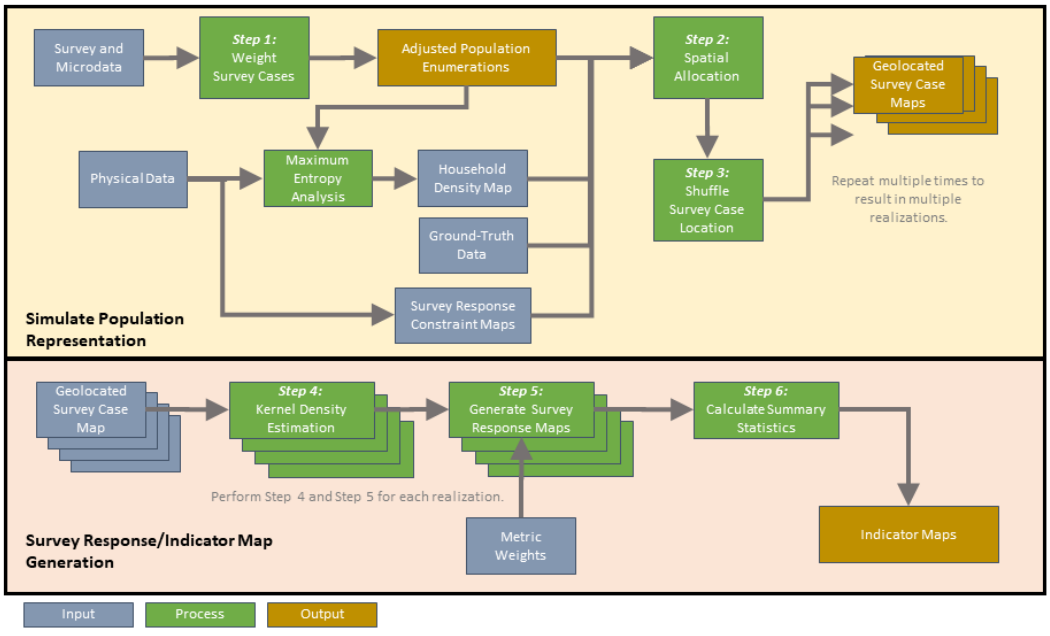
- Population enumeration estimates
- Survey responses or microdata
- Household or population density maps
- Maps of survey response constraints
- Samples of ground-truth information containing specific survey responses

Each data input imparts specific types of precise information that other data types do not contain. The goal is to ensure that each type of precise information is retained and represented in the model products.

2.2 Six-step survey response mapping process

Figure 3 represents the process for converting the various data inputs into a set of maps representing variations in a survey response over a study area.

Figure 3. Survey response mapping process.



The survey response mapping process consists of following six steps:

1. **Weight survey cases.** Survey cases are replicated a number of times to match demographic characteristics in the overall estimated population enumerations. The replication process fits the results and is weighted using a sum of least squares, minimizing specific desirable survey responses.
2. **Perform spatial allocation.** Household survey cases are realized into plausible geographic locations. Ultimate household location maps are based on household density maps, ground-truth data, and survey shuffling for optimization. The process of maximum entropy analysis generates household density maps.
3. **Shuffle survey case location.** Survey case locations are shuffled to improve spatial statistics. This step optimizes a set of proportional and spatial statistics for each population realization to create realistic clustering of survey responses.
4. **Estimate kernel density.** For each desired combination of survey responses, proportion maps are generated on each population realization throughout the study area, representing the percentages of simulated survey cases with such responses. This kernel density estimation process is applied cell by cell across a regularized grid.
5. **Generate survey response maps.** Map algebra analysis generates survey response maps 1 to n realizations.
6. **Calculate summary statistics.** Throughout the study area, box plot summary statistic maps are compiled on the minimum, maximum, median, medium, 1st quartile, and 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.

The first three steps focus on modeling the population within the landscape. This includes accurately representing population densities and fitting population demographics within that representation. A key aspect of to this representation is understanding the environmental factors that influence the attractiveness of a site for a household to locate. The last three steps of the process generate indicator maps of the simulated households. These steps focus on retaining the errors and uncertainties of the input data in a way that helps end users understand the impacts on their application and ultimately improve the utility of the information for decision makers.

The following chapters organize the description of this process by inputs (Chapter 3), processes (Chapter 4), and outputs (Chapter 5).

2.3 Definitions

Before stepping into the procedure, some important terms need to be defined. The households and people modeled are derived from a survey or *microdata*, the latter term being 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. A *survey case* comprises the responses given by one person or household. One sample in *census microdata* has a more complex origin than a survey case. A census microdata sample represents a cluster of similar, but almost never identical, survey cases combined. In this methodology, a census microdata sample is used as a survey case. *Survey responses* each represent the set of answers relevant to a particular operational need or analysis, and these 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. These are provided by census data; they include an estimate of census error measures and are 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 specialized information not typically available to demographic models, and it will ensure that those locations will provide survey responses for those locations. One end product category, 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.

3 Input Data

This chapter describes the five types of data conflated by the demographic model. Data for Chittagong Division, Bangladesh, are used for illustrative purposes.

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 inform the allocation of government funds, censuses are usually the most accurate single source for population estimates. However, census data for some countries are incomplete or less accurate for regions that are 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. Using Chittagong Division, Bangladesh, to illustrate, 2011 census data were downloaded from the Bangladesh Bureau of Statistics website. Bangladesh's 2011 census was a collaborative effort between the Bangladesh Bureau of Statistics, the United Nations Population Fund, the European Union, United States Census Bureau, and the United States Agency for International Development (Bangladesh Bureau of Statistics (2012). This data set contains the enumeration of various age groups, gender, and race for each upazila. These areas are further subdivided into urban and rural subpopulations, with this research representing those areas by means of imagery and maximum entropy analysis to represent 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:

$$P(\mathbf{s})_{t,v,i} = P(\mathbf{s})_v + \mu(\mathbf{s})_{t,v} + R(\mathbf{s})_{t,v,i} \times \sigma^2(\mathbf{s})_v \quad (1)$$

where

i = a population realization
$P(\mathbf{s})_{t,v,i}$ = the realized demographic variable v goal values for time t across the study area $\mathbf{s}$ for population realization i
$P(\mathbf{s})_v$ = 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 $\mathbf{s}$

$R(\mathbf{s})_{t,v,i}$ = a random normal deviate for time t of the demographic variable v goal values across the study area $\mathbf{s}$ for population realization i

$\sigma^2(\mathbf{s})_v$ = 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.

3.2 Surveys or microdata

The data in the U.S. Integrated Public Use Microdata Sample (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 all 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) telephone questionnaire (U.S. Congress 13 June 2001). ACS telephone 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, Candau, and Clarke (2004) have discussed, in addition to many other authors, uncertainty increases with the difference in time when a survey is completed and for when data are needed. In the slums of developing world cities and other rapidly changing environments, the frequency of surveys is especially critical for accurate demographic forecasts. By relying on four-year-old Bangladesh IPUMS data, the uncertainty represented by Equation 1 is

* SQL: Structured Query Language.

magnified more significantly than if we were trying to represent Chittagong Division using current data. It is important to understand the intended use of the analysis so that map results match those expectations. Forecasting population trends will often be necessary.

When comparing Bangladesh census enumeration attributes against Bangladesh IPUMS, specific survey responses do not have the same proportions in both data sets. Therefore, it is critical to perform a population weighting adjustment (Kalton 1968) 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 socioeconomic amenities, the sampled households with less access to such amenities must be weight-adjusted lower. Survey results, which often have not been adjusted by statisticians, require weighting adjustments even more.

3.3 Survey case density map

Population or household density maps, usually gridded raster layers, represent how many households are located within specific areas, usually as a measure of people per square kilometer (km^2). (Most of the indicators used to create these maps are indicative of household density, which makes household density a more accurate term than population density. However, the term population density is almost always used to encompass these indicators.) Household density is a generated input map. It is estimated using land use, topography, distance to transportation, and other physical factors that will attract or repel populations. With the goal of neighborhood-scale maps, we used 15 m resolution imagery, distance to all roads, topographic slope and aspect values, and Bangladesh census enumeration values in a maximum entropy analysis to determine household locations for Chittagong Division. (Maximum entropy analysis is described in Chapter 4). When survey-case locations are simulated within each enumeration zone, which are upazilas subdivided into urban and rural areas for the Bangladesh case study,* each survey case is initially placed in the landscape based on the relative weight of the household density map.

One of the largest potential uncertainty issues about household density maps is the speed at which urbanization changes. Especially in dynamic

* An ERDC/CERL technical report on this case study is currently in preparation.

urban environments, newly developed subdivisions are treated as vacant lands. Stochastically driven urban growth modeling provides forecasts of future land use patterns (Westervelt, Bendor, and Sexton 2011); Goldstein, Candau, and Clark 2004). 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 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 public access to sanitary water will ensure that households with public water will only be located in those locations. Since many survey responses have collinearity with each other, such as households having refrigerators is collinear with 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.

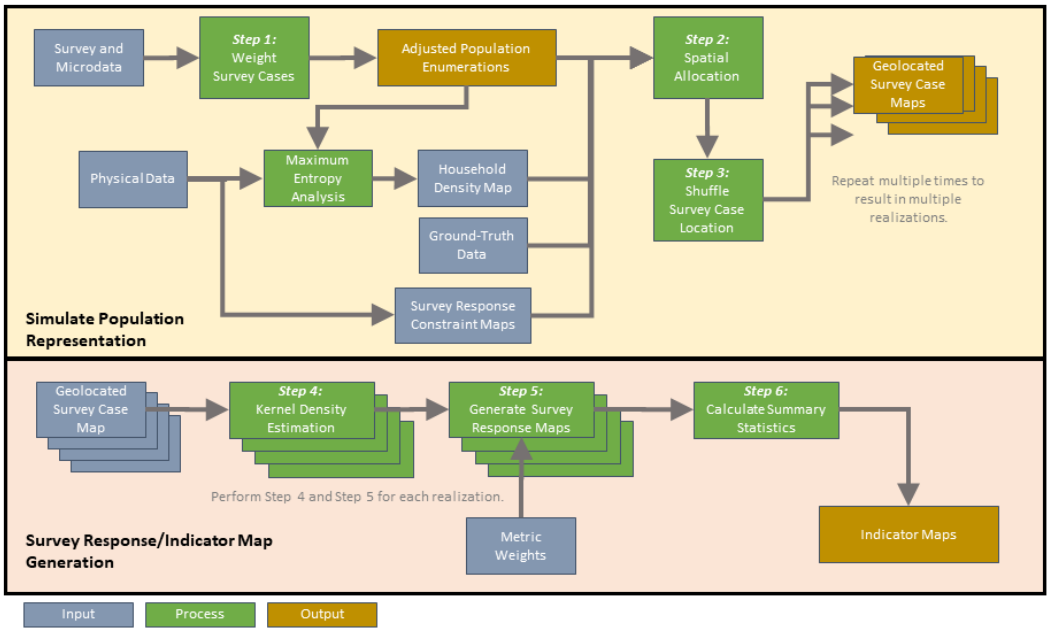
3.5 Ground truth samples

Ground truth data are point or polygon locations where survey response characteristics are known to exist in the study area. Geolocated survey responses are matched to households with those characteristics in 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 low-populated rural 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, which prevents this process from exploiting parallel processing techniques.

4 Processes

This chapter describes the process boxes depicted in section 2.2, Figure 3. That figure is repeated immediately below for the reader's convenience.

Figure 3. Survey response mapping process (reproduction from section 2.2).



4.1 Weighting survey cases to census data

Weighting survey cases to census data 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" in survey literature (Kalton 1968) and is appropriate for correcting surveys likely to be more biased than authoritative population enumerations. A subset of survey responses related to population enumerations is chosen to fit the survey. This step uses a *duplication of cases* approach (Kish 1990), 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 covariance relationships inside survey responses. Kalton (1983) indicates that the multiple realizations generated by the overall technique almost

completely eliminate the increased variances. Survey cases are stochastically duplicated to minimize root mean square errors of the questions' relationships.

4.2 Spatial allocation and maximum entropy

The *spatial allocation* process duplicates household survey cases into plausible geographic locations. This process is performed stochastically using the enumerated population estimates, household locations density map, samples of ground-truth data, and survey response constraint maps. This phase ultimately geolocated survey case locations for the entire population of the study area. This point pattern process is discussed throughout the research literature (Bailey and Gatrell 1995; O'Sullivan and Unwin 2003). In the spatial allocation process, each survey case in a realization is randomly given 10 eligible locations. Survey cases are initially placed across the study area fitting first-order properties of census enumerations. The remaining nine eligible locations are based on the input data. The survey case is placed in the one that will best improve the fit of important survey responses. In the example outputs presented in Chapter 5, case study, the important variables are utility access, 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.

A 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 expresses the suitability of a location—household location in this case—as a function of the environmental variables at or near that location (Jaynes 1957). Maximum entropy analysis was used to determine household density. 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 category 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 layers with the greatest permutation importance using jackknife tests. Table 1 lists these eight map layers. 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 fourth power in order 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.

Table 1. Maximum entropy analysis weighting factors chosen to determine household density and the urban/rural divide within upazilas.

Map Layer	Percent 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%

The variables used were samples of 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* (Stevens et al. 2015) and *linear regression* (Ehlschlaeger 2004). A stratified random sample of urban and rural household locations was used to calibrate and verify the model. For the example outputs in Chapter 5 of this report, 23 map themes were tested

for significance to samples of household locations. Those map themes were from 15 m NaturalView Landsat Mosaic imagery layers, including raw bands and unsupervised classifications; and pattern and texture measures. Shuttle Radar Topography Mission (SRTM) also provided topographic variables, including hydrology and open-source augmented road information. A deterministic sequential-update algorithm (Dudik, Philips, and Schapire 2004) implemented in the Maxent open-source software package (Philips, Anderson, and Schapire 2006) was used to determine which layers were significant. A more detailed description of the Maxent analysis completed for the Bangladesh case study in this paper can be found in Lozar et al. 2018. See Elith et al. (2011) for a general discussion on the statistical properties of this process. 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.

4.3 The shuffle households processes

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 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 analysing 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 areas served by the grid and vice versa. Without prior

knowledge of which remote villages currently have or will have biogas development,* 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 a *pycnophylactic method* (Tobler et al. 1997). However, that process and similar methods will underestimate the variability across realizations.

4.4 Kernel density estimation and generating response maps

4.4.1 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:

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

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.

The proportion maps are represented as raster maps with regular grid cells. A *kernel density estimation* (KDE) 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.

* Bangladesh's drive to provide all citizens with electricity involves both expanding the national grid and encouraging remote rural villages to develop biogas plants (http://bbdf.org/).

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

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

where (x_i, y_i) are the locations of n persons, and N_i is the indictor of whether the i^{th} 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 value may incorporate multiple transport modes, including traveling walking, private 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.

4.4.2 Combining proportion maps into indicator maps

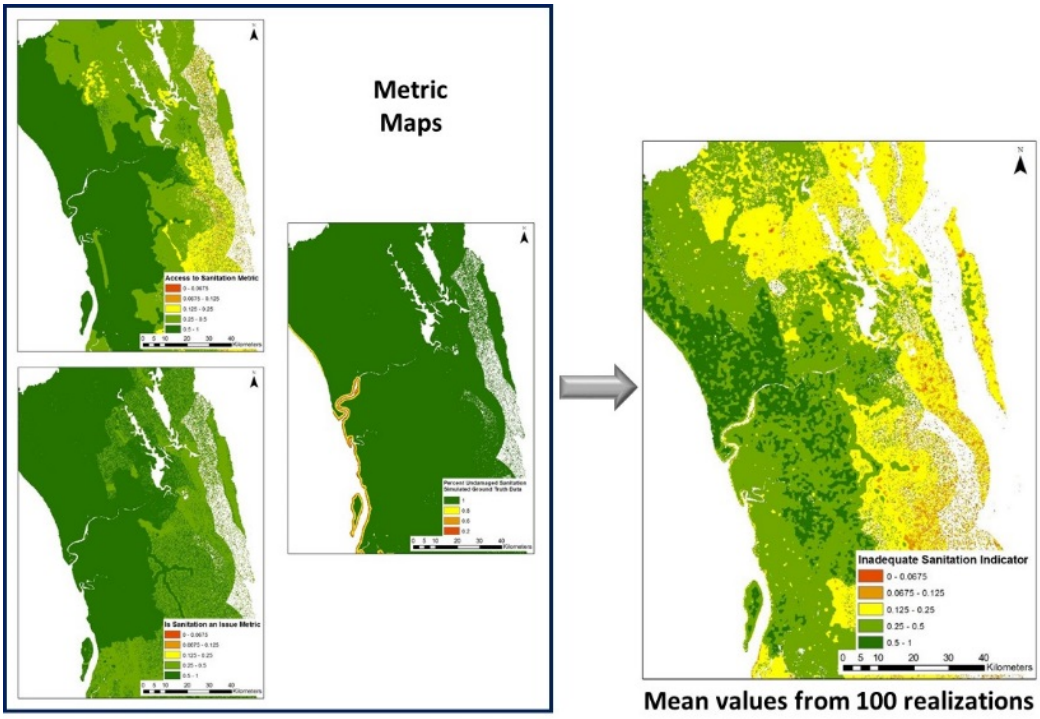
For specific end users, individual survey response proportion maps and combinations of survey response proportion maps are useful. This has proved particularly true in FICUS, where an objective is to represent the combination of sociocultural data layers for operational use. An example of this use is sustainability or risk frameworks. Table 2 shows an example of such a framework. Specifically, it defines the risk of service failure.

Table 2. Indicators of conditions. An HC framework (report in development) is organized by conditions. Factors and indicators are used to evaluate each conditional performance. Multiple conditions make up the complete HC framework, which serves to assess the combination of the probability of a disaster and its negative consequences.

Condition	Factors	Indicators
Services Failure	Unknown Service Failure Factor	
	Law Enforcement & Policing Deficiencies	Policing/Patrol Deficits
		Inadequate Investigations and Prosecution
		Prisons and Jails (lack of capacity)
		Inadequate Facilities/Property Protection
		Unknown Law Enforcement Deficiencies Indicator
	Health & Medical Service Insufficiencies	Doctor (Health Care Professionals) and Access to Primary Care
		Hospitals/Clinics and Secondary Care (Medical Specialists)
		Availability of Pharmaceuticals (Antibiotics)
		Delayed or Deficient Mortuary Affairs
		Unknown Medical Service Insufficiencies Indicator
	Utilities Disruption	Inadequate Sanitation
		Water Shortfalls
		Lack of Communications Availability
		Energy Deficits
		Unknown Utilities Disruption Indicator

Every condition, factor, and indicator map will consist of values between 0.0 (extreme risk) and 1.0 (low risk). Framework components in each column are informed by the rows across from its rightward column. For example, the factor map "Utilities Disruption" is a function of the indicator maps "Inadequate Sanitation," "Water Shortfalls," "Lack of Communications Availability," "Energy Deficits," and "Unknown Utilities Disruption Indicator." (The final map represents the information that is not available as well as the incomplete understanding of what "Utilities Disruption" should be.) Figure 4 represents the combination of sociocultural data layers.

Figure 4. Illustration of how combining metrics (survey responses) can be processed to form indicators.



Like framework components, metric maps represent the level of risk. The metric maps are composed of the information from available surveys, simulated across the entire population in the study area. They illustrate the proportion of people or households with the subject condition. For example, the "Sanitary Toilet Facility" metric map contains the proportion of households that have sanitary toilets within a radius of 800 m of the household locations. The choice of proportion radius is both dependent on how geographically specific the framework analysis needs to be, and how accurate the analysis results need to be. As the radius decreases, the accuracy of the map will decrease. If the range of possible metric values forces the analysis to be too inaccurate for useful planning, four tactics may be employed: (1) additional survey data can be collected (in the case of DoD collected surveys), (2) complementary data can be found to augment the census or survey data, (3) subject-matter experts can be brought in to better refine the framework weights, and/or (4) the analysis can be performed at a coarser geographic scale, which will improve accuracy at the expense of precision. Table 3 demonstrates how the Inadequate Sanitation Indicator is defined by its four metrics. Each of the metric maps is multiplied by risk value raised to the power of its weight. All metric maps are then multiplied to create the indicator map. Risk values and weights are given ranges

to represent what is not known about the conditions of that location. Using Table 3 as an example, only an expert on Bangladesh sanitation would know the exact risk values and weights to apply. Analysts should increase the ranges the more uncertain they are of the true values. Calibration of risk values and weights can be performed in geographic areas with detailed knowledge of the population, whether based on civil affairs units, state partnership program collaboration, or trusted subject matter experts.

Table 3. Metrics for indicator. The indicator value is measured using a weighted product of metric values. Similarly, factors values are measured using a weighted product of indicator values, and so forth along the framework structure. Values and weights may be provided as a range (e.g., minimum and maximum). The wider the range, the less certain of the risk contribution. The tighter the range, the more certain of the risk contribution. Random values may also be inserted to account for unknown variables.

Indicator	Metric	Risk Value (min–max)	Weight (min–max)
Inadequate Sanitation	Sanitary Toilet Facility	1.0 – 1.0	0.8 – 1.0
	Shared Toilet Facility	0.2 – 0.7	0.4 – 0.6
	Perception of Sanitation Issues	0.3 – 0.9	0.3 – 0.3
	Unknown Inadequate Sanitation Metric	1.0 - 1.0	0.1 – 0.2

The spatial demographic data obtained through the spatial allocation process serve as quantitative metrics for indicators. For example, the state of the sanitation system is reflected by the answers to the survey question—the type of toilet facility, whether it is shared with other households, or whether sanitation is viewed as a serious issue.

Another example would be to determine the 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 (see Figure 7, page 26). 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 4, summing the weights in each survey case.

Table 4. Weights of survey responses indicating quality based on IPUMS responses.

Survey Question & 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)	- 0.1

We used a variation on a favorability function (Bonham-Carter 1995) to estimate relative resource deprivation:

$$I(\mathbf{s}) = F(\mathbf{s})_m / F(\mathbf{s})_h, F(\mathbf{s})_i = \prod_{N=1}^{n} X_{N,i}(\mathbf{s}), i = \{m, h\} \quad (3)$$

where

I(s) = the ratio of Muslim, m, household wealth at each location s in the study area to Hindu, h, household wealth

F(s)I = the favorability of ethnicity i evaluated as a continuous value between 0.0-1.0 at each location s in the study area

XN,I = the value at 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

4.5 Proportion map summary statistics

For each 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, which characterize such distributions, are then calculated from these proportion maps on a cell-by-cell basis. These summary statistics include minimum, maximum, median, average, first quartile, third 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, first quartile, or third 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 who have at least 12 years of education” would result in maps with the probability of attaining that 80% level using minimum proportion maps.

5 Discussion of Example Outputs

5.1 Geolocated survey case maps

After household density across the study area was determined, household cases were initially located to match the same characteristics fitted when weighting those cases. For the present 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, but increasing the criteria being fitted would 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.

Figure 5 shows a subset of one realization's households in the greater Dhaka metropolitan area. Realizations can be converted into Keyhole Markup Language (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 fit. Fitting six or more survey responses in rural areas inevitably led to less-precise fits for some responses. This effect was expected as results occurred in other similar algorithms recreating multiple statistics (Ehlschlaeger 2002). These Chittagong Division experiments, where the number of survey cases is much higher than the Dhaka realization subset, required computational loads greater than supportable by typical scientific workstations (Table 5).

Figure 5. A sampling of Bangladesh IPUMS survey cases' simulated locations from one realization as a KMZ file displayed in Google Earth (in Southern Dhaka). The process generates KML files for cartographic communication, GeoTIFF* files for Map Algebra, and comma-separated values (CSV) files for agent-based model inputs.

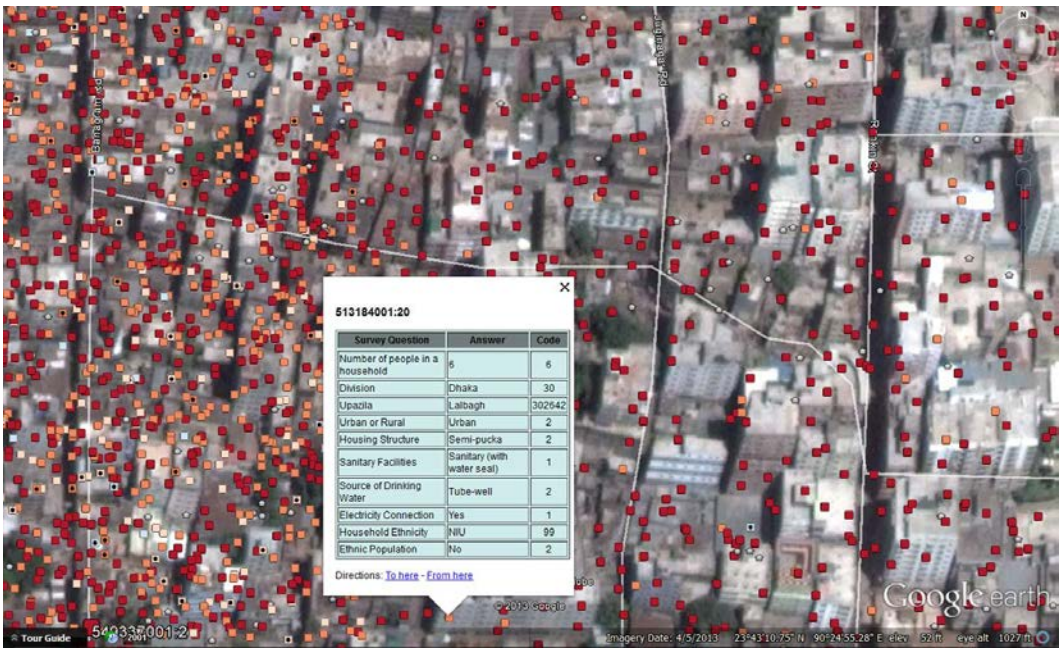


Table 5. Computational requirements for this methodology.

Process	Computer	Computation Time
Household Density Analysis	PC Scientific Workstation	Three 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	½ days

5.2 Survey response mapping (indicator maps)

The process described in this report generates three types of survey representations.

* GeoTIFF: a Tagged Image File Format that accepts the embedding of georeferencing metadata.

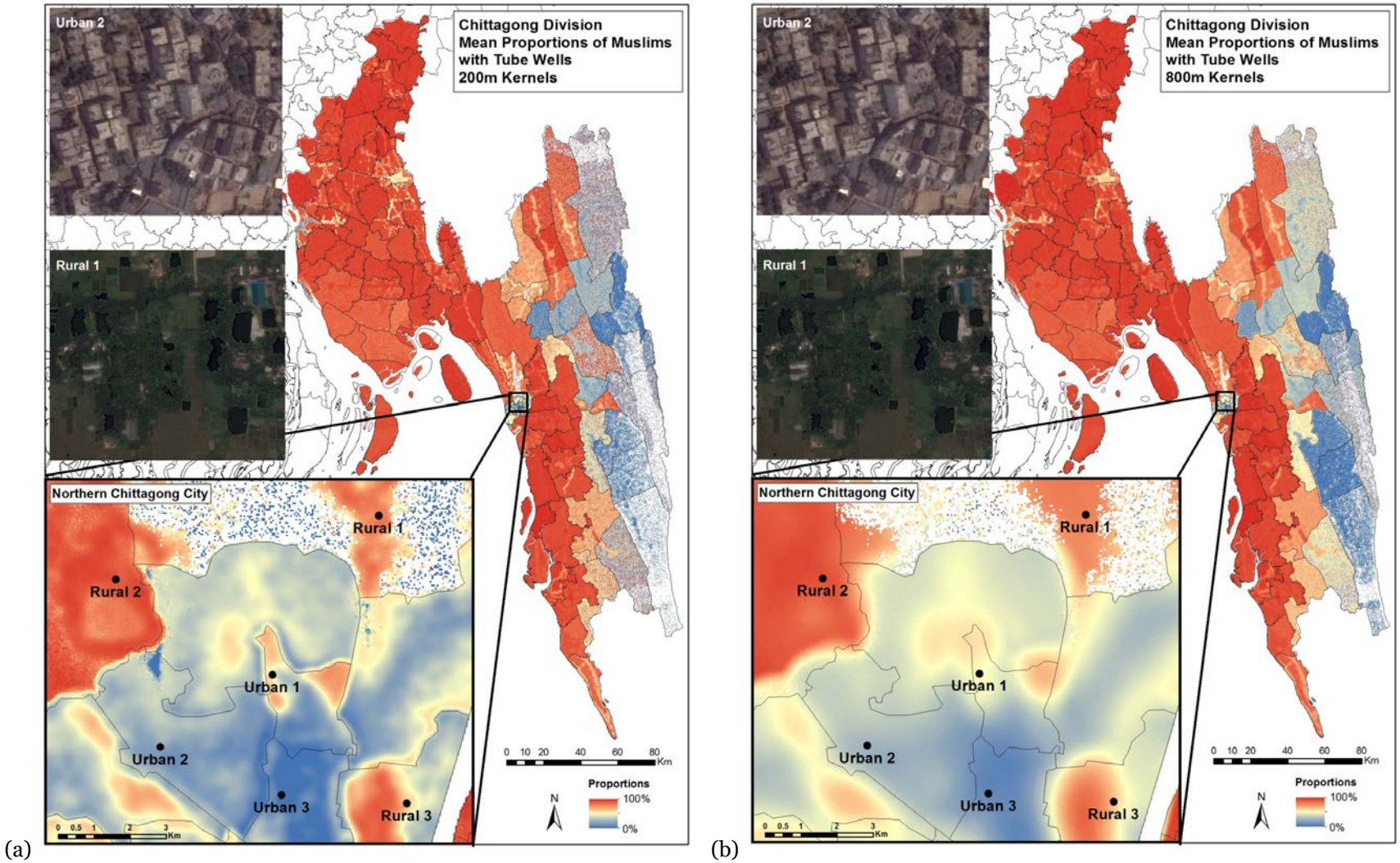
1. Multiple realizations of survey cases representing every household in the survey area can provide suitable inputs for agent-based modeling of socio-cultural behavior. By applying each realization of survey responses to an agent-based model via a bootstrapping process, model designers and users can see the variability of model results inherent from the uncertainties of the demographic input data. (See Figure 3, section 2.2 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, in section 4.4.2, 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.

The model creates three types of demographic results:

1. Simulated locations of every person and household
2. Plausible realizations of survey responses
3. Maps that summarize the distribution of realization results as box plot variables (McGill, Tukey, and Larsen 1978)

Figure 6 demonstrates the correlation between tube well water access and urbanicity with tube well water access proportions following rural areas. Tube wells are significant in humanitarian aid/disaster relief (HA/DR) planning due to the increased likelihood of water contamination during flood events.

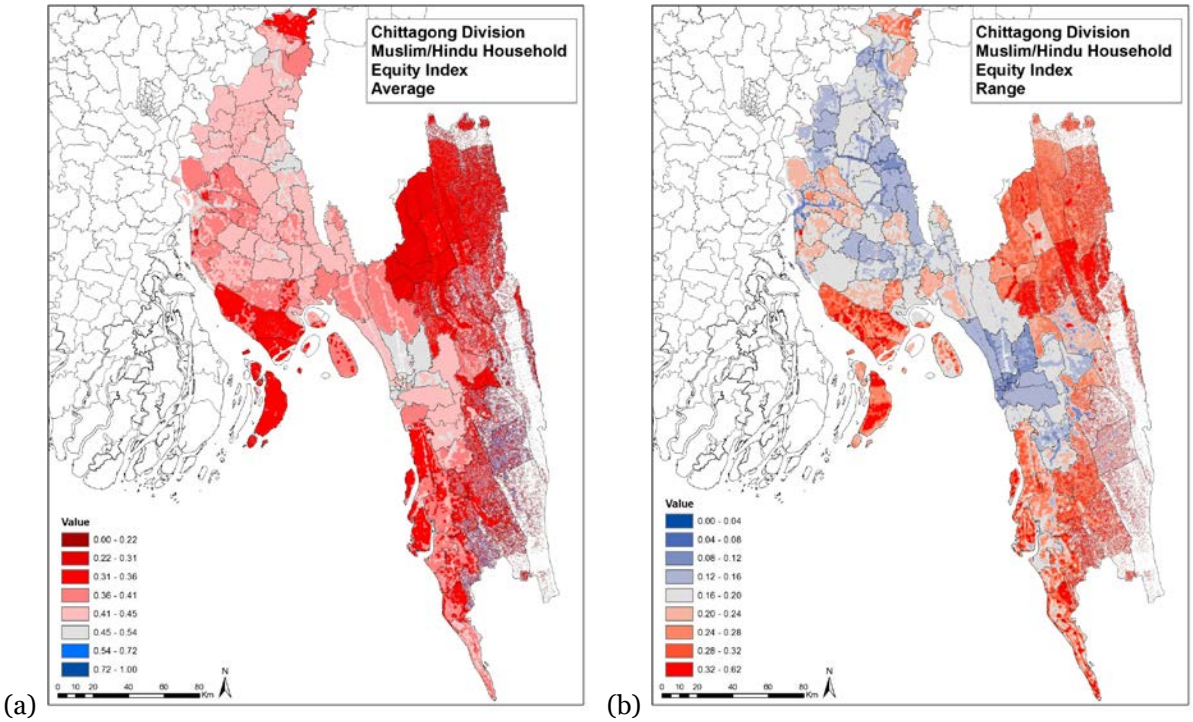
Figure 6. Bangladesh IPUMS survey probability density surface realizations
for Muslim households getting their water via tube wells with kernel radii of 200 m (left, a) and 800 m (right, b).



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 7a are where Muslims are substantially wealthier as compared to Hindus, and red is the opposite extreme.

Figure 7. Muslim/Hindu household equity index average with 800 m kernel radius (left, a); and range of the index, encompassing the spread of index values across all realizations, providing a measure of the application uncertainty at every location in the study area (right, b).



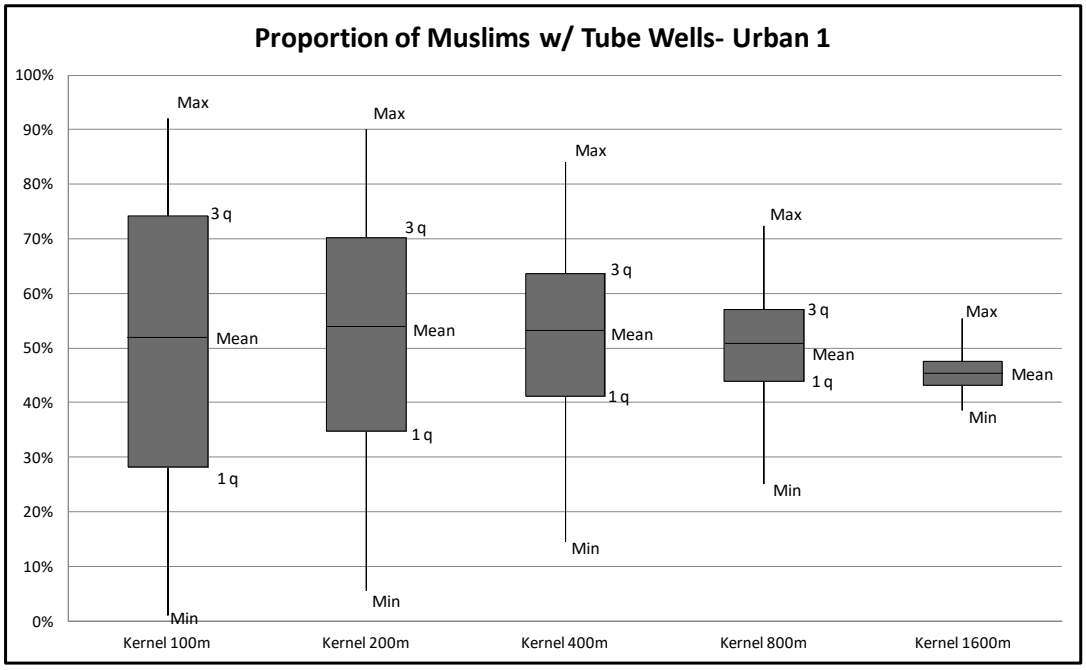
(a) (b)

Gray areas represent where wealth is approximately equal. It can be clearly seen where zones in which one religious group is wealthier than its counterpart in that 1,600 m diameter area. Each of the grid cells on this map is 50 m, although precise results could be obtained at resolutions as fine as 15 m. There are two sources of error for this application: those emerging from the data and those rooted in understanding the application. Table 3 (section 4.4.2) represents the uncertainty modeling parameters in terms of the modeler's knowledge of the relative value placed on each population attribute relative to the highest-quality attribute. (Since the application modelers have never lived in Bangladesh and do not have first-hand knowledge of the quality of each survey response, they gave a wide range

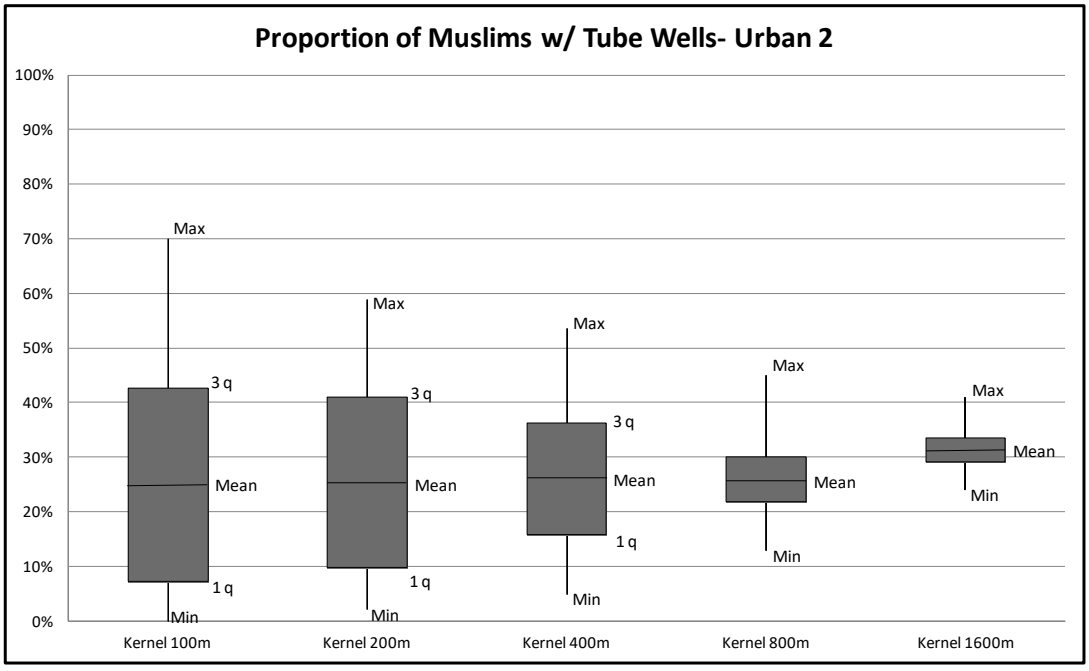
of relative values for each response.) Figure 9 shows the range of application results using the maximum variation of parameter values. Blue areas, which are both urban and very densely populated, usually have high-quality infrastructure and little variation between minimum and maximum parameter estimates. Rural and lightly populated upazilas have greater variation in parameter estimates. In Dhaka Division, there is little variation between minimum and maximum parameter estimates. There is a clear distinction between the wealth of differing religious population. Hindus tend to have greater wealth in areas with government housing, whereas productive agricultural areas favor Muslim wealth. Chittagong Division has no regions favoring Muslim or Hindu wealth; there, the variation in parameter estimates is large.

Figure 8 shows the variation of survey responses at different spatial scales to help illustrate uncertainty arising 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 represent zones where either better data are needed to build the demographic model or the analysis should be performed using a longer kernel radius. From a decision maker's point of view, the IPUMS responses from most neighborhoods in Chittagong have enough demographic information to determine the level of inequity if they believed that Hindus there perceive their neighborhood to encompass homes within 800 m of their own dwellings. However, application results are so varied from neighborhoods smaller than 800 m in both urban and rural areas of Chittagong Division that decision makers cannot be confident that their analysis results are correct. These results were unexpected by the demographic modelers, who assumed that the more densely urban areas would have smaller ranges of values. (When they had performed a similar analysis for Dhaka Division, a small sampling of urban and rural locations indicated greater variation in rural areas.) If the decision makers needed higher-quality data, they could include constraint maps of utilities or other survey-specific household attributes, and then 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.

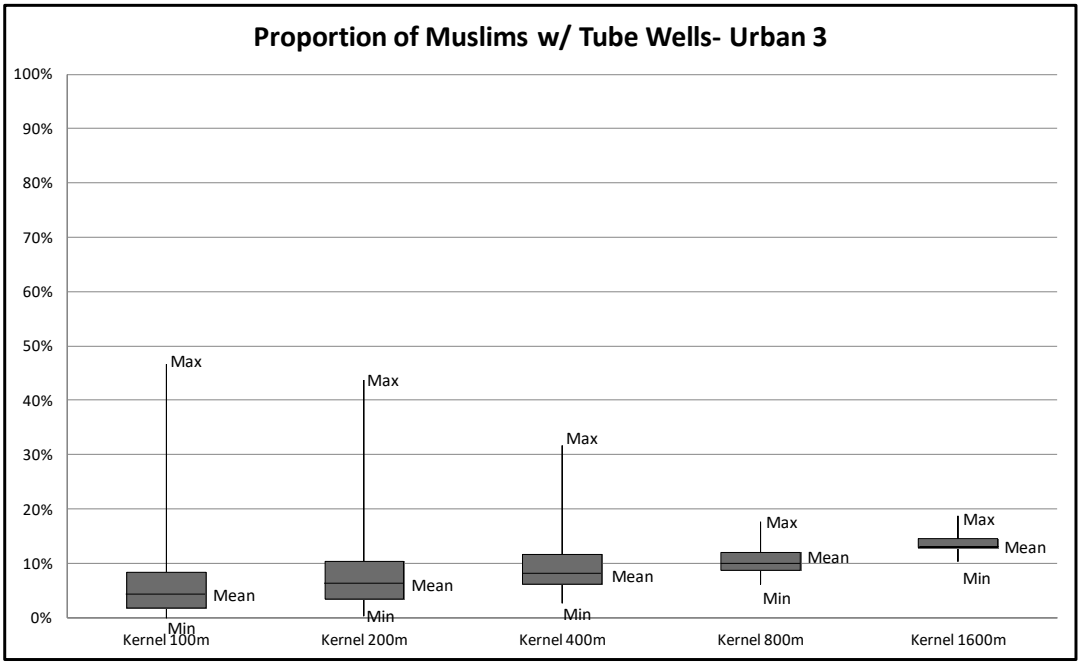
Figure 8. Bangladesh IPUMS survey variability across realizations for Muslims with access to water from tube wells with kernel radii of 100 m, 200 m, 400 m, 800 m, and 1600 m for three urban and three rural locations. Generally speaking, kernel radii of 800 m or more gave reasonably accurate results in denser urban and rural areas, but not in places with low household density.



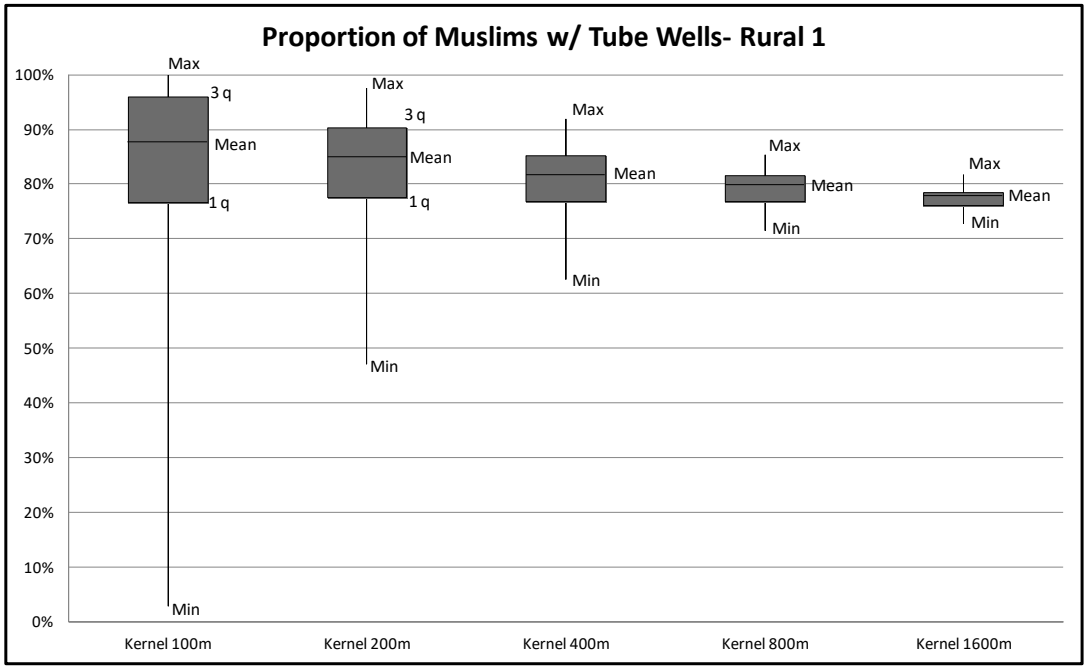
(a)



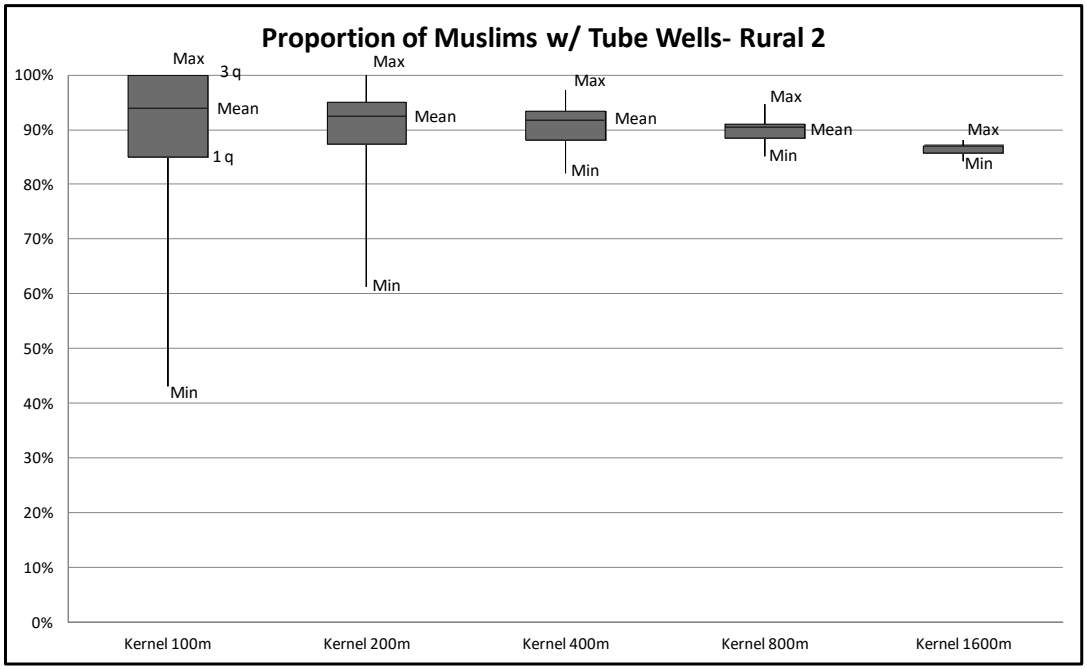
(b)


Proportion of Muslims w/ Tube Wells- Urban 3
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Max
Mean
Min
Kernel 100m
Kernel 200m
Kernel 400m
Kernel 800m
Kernel 1600m

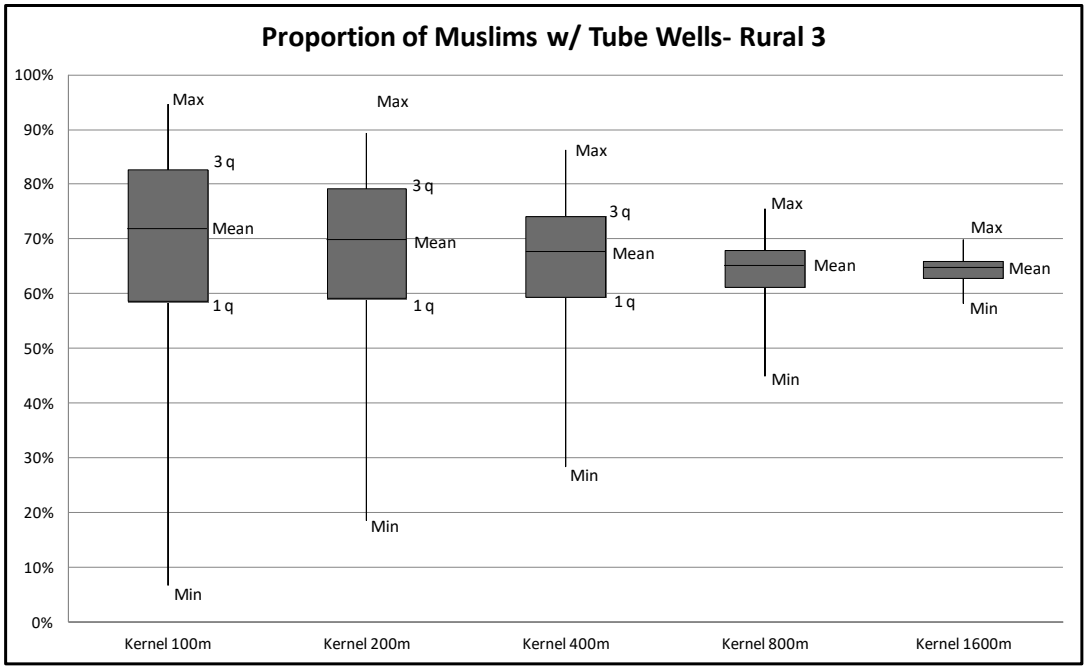
(c)


Proportion of Muslims w/ Tube Wells- Rural 1
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Max
3 q
Mean
1 q
Min
Kernel 100m
Kernel 200m
Kernel 400m
Kernel 800m
Kernel 1600m

(d)

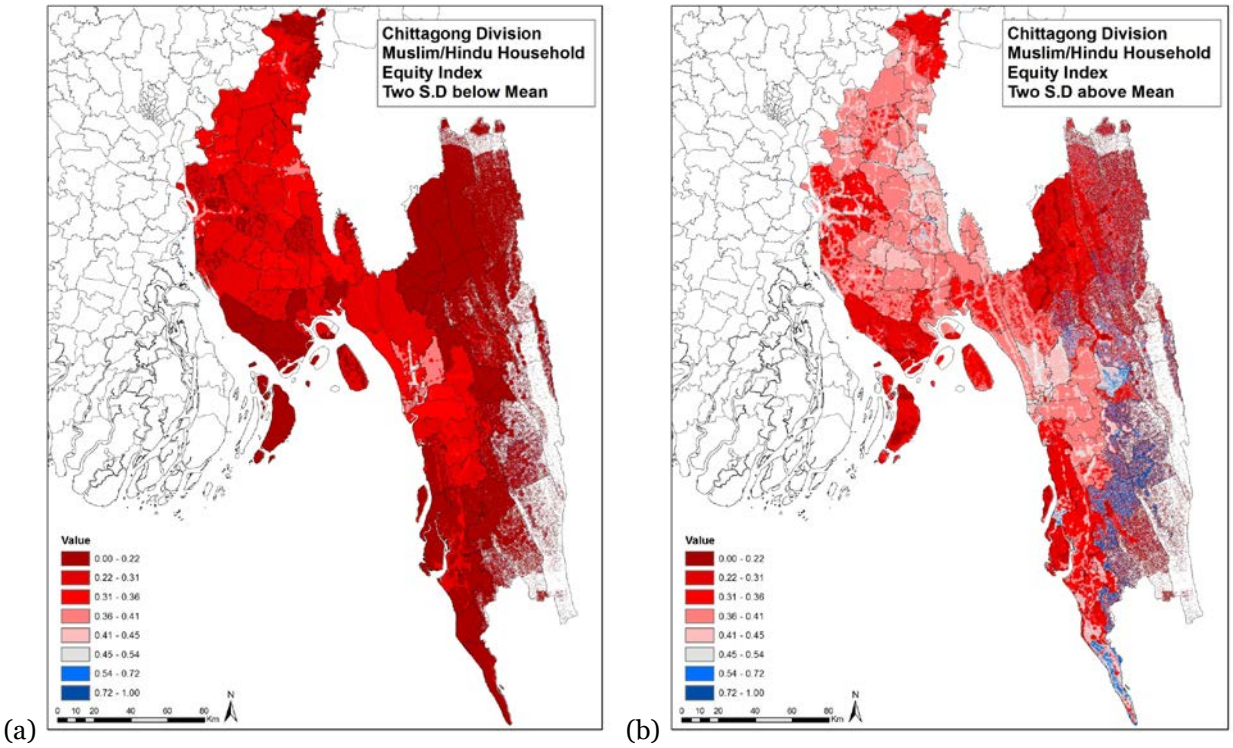

Proportion of Muslims w/ Tube Wells- Rural 2
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Max
3 q
Mean
1 q
Min
Kernel 100m
Kernel 200m
Kernel 400m
Kernel 800m
Kernel 1600m

(e)


Proportion of Muslims w/ Tube Wells- Rural 3
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Max
3 q
Mean
1 q
Min
Kernel 100m
Kernel 200m
Kernel 400m
Kernel 800m
Kernel 1600m

(f)

Figure 9. Estimated data error rates with 800 m kernel radius, shows two deviations below the mean of Muslim/Hindu household inequity (left (a)); and two deviations above the mean (right (b)). Red areas have higher wealth in Muslim households, blue areas have higher wealth in Hindu households, and gray areas have equal wealth among religious households.



5.3 Conclusion

The model used to produce the example outputs described above demonstrates the ability to convert raw authoritative data into many high-resolution maps in order to provide a relative measure of risk. As this methodology has been transitioned into our Humanitarian Crisis Framework (publication in preparation), each level of the framework is represented as a map containing the range of possible risk values created via a Monte Carlo process, realizing different risk values and weights for each instantiation of the computation model. By including *unknown* metric/indicator/factor/condition map representations as well as ranges for risk values and weights, a transparent and easily communicated representation of errors and uncertainties of both the conceptual framework model and the data quality are presented.

6 Summary

The results of this research verify the development and functionality of a data-conflation methodology that can be used to produce sociocultural indicator maps. This methodology can be adapted to work within the context of various frameworks to generate improved sociocultural indicator maps that can help multidisciplinary project teams produce better-informed analyses, plans, policy recommendations, etc., than possible using conventional methods.

The ability to convert any survey into neighborhood-scale survey-response maps offers many benefits for survey designers, sociocultural analysts, and decision makers. A technology for mapping multiple sociocultural indicators is a critical capability for enabling non-experts to understand the complexities represented by demographic maps. One aspect of such complexities is that each input data layer in demographic maps contains various types of errors and uncertainties resulting from factors constraining the various survey process, collection methods, data resolution, etc. Traditional measures of error in the various input layers were developed for the producers of data, but they are mostly of no value to end users for purposes of resolving the errors before analysis. A goal of the conflation method described here is to retain the data errors and uncertainties and combine them in ways that improve the utility of the information for decision makers. In this study, for example, Figure 9 (see page 31) presents the spread of application results across a range of four standard deviations, quickly communicating where in Chittagong Division insufficient data were collected to provide useful results at neighborhood sizes of 1,600 m in diameter.

Another benefit of this technique is that any combination of survey responses can be represented as a surface map of any scale. This capability avoids ecological inference fallacies that may occur when a user predicts individual variable probabilities for one map scale based on the probabilities applying to a map of a different scale. The software developed for this technique generates both GeoTIFF grids and KML files for each survey response desired. The software is designed so that once the demographic model is developed, the entire process is automated. 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 Bang-

ladesh IPUMS survey generates 120 single-response maps, with any number of multiple-survey responses possible. The example outputs in Chapter 5 required a short Linux shell script.

Another potential benefit of this process is the ability for survey developers to form a more accurate understanding of how many survey cases need to be collected for the surveys to be useful for their intended users. The maps created provide precise measures of where information is more useful and less useful, with error bars provided at all locations. As more samples are collected, of course, the error bars become smaller. Also, survey collectors could better apportion the subpopulations to minimize the error bars associated with the survey responses of most interest in the study.

References

Bailey, T., and Gatrell, A. *Interactive Spatial Data Analysis*. Volume 413. Essex: Longman Scientific and Technical, 1995.

Bangladesh Bureau of Statistics. 2012. Population and Housing Census 2011. Socio-economic and Demographic Report, Bangladesh National Series, Volume 4.

Bonham-Carter, G. F. 1995. Geographic Information Systems for Geosciences. Oxford: Pergamon.

Cobb, M. Chung, M., Foley III, H., Petry, F., and Shaw, K. 1998. "A Rule-based Approach for the Conflation of Attributed Vector Data." *GeoInformatica* 2 (1998): 7-35.

Dudík, M., Philips, S. J., Schapire, R. E. 2004. Performance guarantees for regularized maximum entropy density estimation. Proceedings of the 17th Annual Conference on computational Learning Theory.

Ehlschlaeger, C. "Representing multiple spatial statistics in generalized elevation uncertainty models: moving beyond the variogram." *International Journal of Geographical Information Science* 16 no. 3 (2002): 259-285.

Ehlschlaeger, C. "Incorporating Second-order Properties for Cluster Detection Analysis and Agent Based Modeling." GeoComputation 2005 Conference Proceedings, August 2005. http://www.geocomputation.org/2005/, accessed 11/29/2015.

Elith J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., Yates, C. J. "A Statistical Explanation of MaxEnt for Ecologists." *Diversity and Distributions* 17 (2011): 43-57.

Goldstein, M., Candau, J., Clarke, K. 2004. Approaches to simulating the "March of Bricks and Mortar". *Computers, Environment and Urban Systems* 28:125–147.

Jaynes, E. T. "Information Theory and Statistical Mechanics." *Phys. Review* 106 (1957): 620-30.

Kalton, G. "Standardization: A Technique to Control for Extraneous Variables." *Applied Statistics* 17 (1968): 118-36.

Kalton, G. Compensating for missing survey data. Institute for Social Research, Ann Arbor, 1983.

Kish, L. "Weighting: why, when, and how?" Proceedings of the Survey Research Methods Section American Statistical Association 1990, pp 121-30.

Lozar, Robert C., Scott A. Tweddale, Charles R. Ehlschlaeger, Carey L. Baxter, and Jeffrey A. Burkhalter. 2018. Testing Maximum Entropy Analysis to Define Population Distributions. ERDC/CERL TR-18-22.

McGill, R., Tukey, J. W. Larsen, W. A. Variations of Box Plots. *American Statistician* 32 no. 1 (1978): 12-16.

O'Sullivan, D., Unwin, D. "The pitfalls and potential of spatial data." *Geographic Information Analysis* (2003): 33-53.

Phillips, S. J., Anderson, R. P., Schapire, R. E. "Maximum entropy modeling of species geographic distributions." *Ecological Modelling* 190 no. 3-4 (2006): 231-59.

Stevens, F.R., Gaughan, A.E., Linard, C., Tatem, A.J. "Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data." *PLoS one* 10, no. 2 (2015): e0107042. doi:10.1371/journal.pone.0107042.

Tobler, W., Deichmann, U., Gottsegen, J., and Maloy, K. World population in a grid of spherical quadrilaterals. *International Journal of Population Geography* 3 (1997): 203-225.

U.S. Congress, Committee on Government Reform. Subcommittee on the Census. Census Bureau's proposed American Community Survey (ACS): Hearing before the Subcommittee on the Census of the Committee on Government Reform, House of Representatives, One Hundred Seventh Congress, first session, June 13, 2001.

Westervelt, J., Bendor, T., Sexton, J. "A technique for Rapidly Forecasting Regional Urban Growth." *Environment and Planning B: Planning and Design* 38 no.1 (2011): 61-81.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
September 2018	Final	

4. TITLE AND SUBTITLE
Conflating Survey Data into Sociocultural Indicator Maps

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
622784T41

6. AUTHOR(S)
Charles R. Ehlschlaeger, Jeffrey A. Burkhalter, Natalie R. Myers, Carey L. Baxter, Matthew D. Hiett, Ellen R. Hartman, Scott A. Tweddale, James D. Westervelt, Robert C. Lozar, Yizhao Gao, Dandong Yin, Marina V. Drigo, and David A. Brown

5d. PROJECT NUMBER
P2 458304

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Engineer Research and Development Center (ERDC)
Construction Engineering Research Laboratory (CERL)
PO Box 9005
Champaign, IL 61826-9005

8. PERFORMING ORGANIZATION REPORT NUMBER
ERDC/CERL TR-18-32

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of the Assistant Secretary of the Army for Acquisition, Logistics, and Technology
103 Army Pentagon
Washington, DC 20314-1000

10. SPONSOR/MONITOR'S ACRONYM(S)
ASA(ALT)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release. Distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This report 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 used to combine any number of survey responses to stochastically weight survey cases and to simulate survey-case locations in a study area. Through these methods, known errors from each input source can be retained. By keeping the individual survey case 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 using data and output maps for Chittagong Division, 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 for each input source via the automatically generated summary statistics created at the application's geographic scale.

15. SUBJECT TERMS
Sociology, Military; Cities and towns; Bangladesh; Geographic information systems; Demographic surveys

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER (include area code)
Unclassified	Unclassified	Unclassified	UU	44	