Spatially Referenced Synthetic Households for High-Resolution Population Modeling

Steven Rubinyi (corresponding author: steven.rubinyi@chch.ox.ac.uk)
Environmental Change Institute, University of Oxford, Oxford, United Kingdom
Urban, Disaster Risk Management, Resilience, and Land Global Practice, The World Bank, Washington, DC, United States
ORCID: 0000-0003-0631-6105

Jim W Hall (jim.hall@eci.ox.ac.uk)
Environmental Change Institute, University of Oxford, Oxford, United Kingdom

Johannes Gussenbauer (Johannes.Gussenbauer@gmail.com)
Statistics Austria, Vienna, Austria

Acknowledgements: This document is an output from the REACH Program funded by UK Aid from the UK Department for International Development (DFID) for the benefit of developing countries (Aries Code 201880). However, the views expressed, and information contained in it are not necessarily those of or endorsed by DFID, which can accept no responsibility for such views or information or for any reliance placed on them.

Spatially Referenced Synthetic Households for High-Resolution Population Modeling

This paper presents a methodology for generating spatially referenced synthetic households by combining population synthesis techniques for small area estimation with dasymetric population modeling. The methodology results in an output vector layer of households represented as points with unique structural and socioeconomic attributes which can be applied towards a better understanding of population dynamics, particularly those characteristics which are geographically influenced and may not conform to administrative boundaries linked to aggregated census statistics. To demonstrate the effectiveness of the methodology, the paper provides a case study in Dumuria Upazila, Bangladesh, where an independent household survey is used to validate the location of the synthetic households and associated socioeconomic attributes. The study found that 98 percent of all surveyed households were within 100 meters of at least one synthetic household and the median distance observed was only 18 meters. The study also found that the age distribution of synthetic populations was within five percent of the surveyed population for each of the five age bands considered.

Keywords: population modeling; population synthesis; geographic information systems; Bangladesh

1. Introduction

1.1.Anonymity and Geospatial Data

Population-based research can benefit significantly from detailed location information of the households or individuals studied (Olson et al., 2006). However, ensuring anonymity to a survey respondent is both essential for ensuring truthful responses and an ethical obligation for potentially sensitive data. Thus, most surveys and national censuses release their data for public usage in aggregate form, often paired with microdata that details out a fraction of anonymized responses for unique households (Arzberger et al., 2004). For example, the United States Census Bureau has disclosure avoidance requirements which restrict the release

of microdata for public use to aggregate populations of at least 100,000 (Gomez, 2017). Generally, higher level aggregates diminish the effectiveness of small area estimation techniques, mask localized spatial dependencies, and may lead to ecological fallacies at lower levels of aggregation (Paiva et al., 2014). Anonymized household and individual level releases of microdata can help provide a baseline for understanding the distribution of attributes collected within a geographic area but may not provide much further detail into the spatial variance of those attributes. To address this concern, geographically focused releases of microdata with higher resolution geographic indicators have been released, often with more restricted attributes. Initiatives such as the Integrated Public Use Microdata Series (IPUMS) have worked to populate a database of individual-level microdata samples, including from international census records. Efforts such as the Demographic and Health Survey (DHS) program, the Living Standards Measurement Study (LSMS), and the Multiple Indicator Cluster Surveys (MICS) have also helped to provide baseline socioeconomic information in developing countries. Although an improvement, the effectiveness of household or individual level spatial analysis and modeling remains limited without precise location data tied to features. Data have been used previously in this context but are typically dependent on either complete confidentiality or unique circumstances (Kramer et al., 2016).

The root of the challenge is an inherent trade-off between maximizing the usefulness of the data, and safeguarding anonymity. Anonymization aims to safeguard individuals, households, or larger groups of people from having potentially sensitive data attributed to them. This can include removal of information such as name, address, postal code, or any other attribute that could be used as a unique identifier. Assurances of anonymity are typically an important prerequisite for respondents to disclose potentially sensitive information. Removing location information is thus a key objective of anonymization, making spatially focused research particularly difficult to pursue without some sort of

workaround. Although government agencies such as the United States Census Bureau or the United Kingdom's Information Commissioner's Office have established guidelines for data anonymization, a global consensus does not exist on the level of anonymization required or the best technique to employ for each type of data (Brasher, 2018). This includes whether to release associated microdata for aggregate data sets, and in what type and quantity. Anonymization of data is therefore largely a qualitative judgment, leaving spatial analysts and modelers in a position where each survey or census requires a tailored approach to understand what level of information can be extracted within a comfortable margin of error, and which methods for downscaling are most appropriate to employ given the context of the methodology and modes of data release.

1.2.Geospatially Referenced Surveys

Increasingly, geographic coordinates of features are being collected during surveys and censuses. Cheap, GPS-enabled smartphones and free, open source data collection tools such as Open Data Kit (www.opendatakit.org) have enabled even the simplest of surveys to include a high-resolution spatial element. The efficacy of these techniques was exemplified during the 2013-2016 West Africa Ebola epidemic, where the analysis of data collected though GPS-enabled smartphones helped response teams better predict, contain and treat outbreaks (Tom-Aba et al., 2015). Governments, particularly in developing countries where robust and up to date address points are not available, are increasingly turning to quickly mapping geographic coordinates as a means for identification. For example, following the 2015 Nepal Earthquake, a census of all houses within earthquake-affected districts was conducted which linked attributes of households, damaged structures, and photos to a precise geographic location using Android tablets (Lallement et al., 2017). Although this proved incredibly useful for the Government of Nepal to mobilise recovery and track reconstruction

efforts at the household level, non-government organizations and the public were initially unable to access any data beyond simplified aggregates due to an inability to quickly determine an appropriate mode of anonymization and release.

The question of how to effectively anonymise geospatially referenced data is becoming increasingly important as the collection of respondent-level spatial information becomes easier to implement and more standardized into mainstream methodology. In situations where the objective is to analyse point data for individual or household data collected from geospatially referenced surveys, techniques such as geographical masking (Zimmerman et al., 1999) and spatial blurring (Cassa et al., 2008) can be employed to add layers of anonymity. Spatial blurring, for example, is a technique which moves point information a randomised distance from its true location. These techniques could theoretically be applied to a survey or census and would present the actual attributes collected by surveyors unless synthetic data are inserted as a replacement. Although access to geospatially referenced data would be ideal, unless involved in the collection process directly, the established norm is to only release data aggregated to a geographic resolution sufficiently diluted to ensure appropriate levels of anonymization. This may change in the future, but for the time being it means that the more common workaround would require the downscaling, or disaggregation of data into smaller areas of higher geographic resolution.

1.3.Geospatial Workarounds to Data Aggregates

1.3.1. Small Area Estimation Techniques

At a basic level, techniques for downscaling socioeconomic data, including small area estimation (SAE), can be used to improve the resolution of aggregates available for comparison (Ghosh and Rao, 1994). Aggregated units are typically linked to census blocks or

other administrative units but can also be linked to pre-drawn grid cells or other arbitrary boundaries. There is typically a trade-off between the number of attributes disclosed for each household and the geographic resolution in which the households can be identified. For example, an administrative level 2 microdata release may contain 20 attributes for each unique household, while an administrative level 4 microdata release may contain 10 attributes for each unique household. Usually the attributes that are being cut from the list are ones which are potentially sensitive if identified with the area of release.

Aggregate releases of data are often paired with microdata releases, which provide individual or household attributes for a percentage of households (typically up to 10 percent) within the aggregated geographic unit. Paired with aggregates, microdata can be applied to SAE methodologies for downscaling. In situations where official microdata directly taken from the census records are not released or where attributes not collected in the census are required, surveys collecting some of the same or similar attributes can be used instead, or in connection with official microdata for the purposes of SAE. This is exemplified through the Elbers, Lanjouw and Lanjouw (ELL) method of SAE to derive estimates of poverty rates, which require supplemental attributes related to welfare not typically collected in a national census (Elbers et al., 2003). In Bangladesh, the World Bank (2010b) uses the ELL method of SAE developed to derive estimates of poverty rates at the district and upazila level. This is done by linking parameter estimates from a consumption model developed based on the 2010 Household Income and Expenditures Survey (HIES) with the 2011 Population Census.

Although many unique methodologies have been proposed for SAE (e.g. Bishop et al., 2007; Ghosh and Rao, 1994; Whitworth, 2013), they typically fall under one of two categories: statistical regression-based techniques or spatial microsimulation techniques. Statistical SAE techniques, such as the ELL method, match explanatory variables from the

higher-level data set with those available from a separate data set in the smaller area of interest. This has proven a particularly robust technique for downscaling aggregates (i.e. administrative level 2 to administrative level 4) since results are inherently linked to confidence intervals around central point estimates. Statistical SAE techniques therefore concretely allow for uncertainty to be considered, and for downscaling to stop at a level where the analyst determines it is no longer of value. Although advancements have been made to calculate uncertainty using certain spatial microsimulation techniques, it largely remains a methodological weakness of the technique when used for SAE (Whitworth et al., 2017).

1.3.2. Population Synthesis

The primary advantage of using a spatial microsimulation technique is that it results in the generation of a synthetic population for each target area. These synthetic populations can be utilized in distinct forms of modeling, including bottom-up approaches for understanding complex geographic systems. Recent years have seen substantial advancements in these areas, particularly with geographic automata systems (Torrens and Benenson, 2005), agent-based modeling (ABM), and spatial microsimulation (Lovelace and Dumont, 2016). ABM and spatial microsimulation techniques have demonstrated their use for policy decision making, where the synthetic datasets can be used as inputs to simulation models that evaluate the ex-ante distributional effect of proposed policies or development programs. This has proven particularly effective in the transport sector, where agents are generated to be as reflective as possible of demographic structures and travel behavior of real people (Voellmy et al., 2001). Other sectors where ABM has demonstrated potential include coupled socio-ecological systems (Filatova et al., 2013), and health (Tracy et al., 2018). Synthetic populations are particularly valuable for applying ABM techniques to analyze individual or

household level interaction and decision-making. Improving the spatial resolution which populations can be disaggregated into agents therefore improves the level in which the agents are able to reflect the populations which they represent.

Creating a dataset of synthetic populations through spatial microsimulation techniques requires the application of algorithms to extract information from various data sources to construct an entirely new dataset with the required attributes of households (Munnich and Schurle, 2003). Techniques for spatial microsimulation SAE can be clustered into three methodologies: 1) iterative proportional fitting (IPF); 2) combinatorial optimization; and, 3) generalized regression reweighting (Whitworth et al., 2017). Although each method for spatial microsimulation SAE differs in substantial ways, they all "fit" the microdata to the multidimensional aggregates and generate representative synthetic populations. The raw synthetic population data can be useful on its own or be re-aggregated into the smaller area of interest. Standardized methodologies can be employed directly through statistical software, such as the R package simPop, an open source data synthesizer based on a modular object-oriented concept (Templ et al., 2017). Spatial microsimulation techniques for SAE first present a top-down approach, and then shift to a bottom-up approach. Regardless of which method for SAE is employed, the output remains aggregated in a spatial sense. While the release of anonymized microdata paired with aggregated data sets enables data scientists to investigate the relationships between collected attributes and increases the resolution in which SAE techniques can estimate spatial distribution, the spatial component often remains lost at a higher resolution.

1.3.3. Dasymetric Population Modeling

While SAE focuses on downscaling individual or household attributes, spatial population modeling focuses on the spatial distribution of individuals or households over a population

surface. The primary difference between the two is that spatial population modeling focuses solely on the geographic distribution of the population over a discrete surface, while SAE techniques focus on the downscaling of population characteristics (e.g. socioeconomic attributes). As a result, both techniques rely heavily on the disaggregation of census data but approach it from a distinct perspective. Dasymetric modeling, a branch of areal interpolation where boundaries between ancillary variables are discrete, is often employed for spatial disaggregation of population data. In many cases, there is little agreement on which subset technique to employ due to an inherent subjective element to determining the quality and importance of available spatial datasets, as well as the output resolution required. However, even the most basic methods of areal interpolation, such as areal weighting, hold an advantage over choropleth mapping for gaining insights into the spatial variation of populations (Wright, 1936; Goodchild and Lam, 1980; Lam, 1983).

Access to increasingly high-resolution free and open spatial data sets derived from earth observation (e.g. Sentinel and Landsat products) as well as improvements in post-processing techniques unlocked through machine learning have accordingly led to increasingly accurate dasymetric population modeling which produces data sets available at a global scale (Nieves et al., 2017). An example of a global population dataset is the Gridded Population of the World Layer (currently GPWv4), which provides 1-kilometer resolution population estimates derived from a proportional allocation gridding algorithm (Doxsey-Whitfield et al., 2015). WorldPop is another example, which produces country-level layers at 100-meter resolution using a weighting scheme based on a random forest, nonparametric predictive model which improves incrementally as more spatial covariates are introduced (Stevens et al., 2015). The High-Resolution Settlement Layer (HRSL) has produced 30-meter resolution population data sets for 18 countries applying an approach for identification of building footprints which uses convoluted neural networks (Tiecke et al., 2017). While each

method has its own relative strengths, the important takeaway from progress over the past decade is that the resolution in which we can accurately predict the distribution of populations spatially has been incrementally improving.

Geostatistical modeling approaches are often applied in fields such as epidemiology (see the Malaria Atlas Project (www.malariaatlas.org) and the Institute for Health Metrics and Evaluation (www.healthdata.org)), and to better understand socioeconomic characteristics of populations in developing countries (see WorldPop (www.worldpop.org)). Alegana et al. (2018) show how geostatistical methods can be effective: a Bayseyian item response theory framework is used to better understand treatment seeking behaviour for children with fever by triangulating household survey data with georeferenced public health facility data. Utazi et al. (2020) also use a geostatistical modeling approach to produce 1-kilometer resolution estimates for coverage of measles vaccine in Nigeria.

While certainly an effective approach, geostatistical methods can be complemented by techniques that generate spatially referenced synthetic populations for usage in agent-based modeling. Several recent efforts have focused on producing spatially referenced synthetic household data. Thomson et al., (2018) use simPop and ArcGIS to derive a spatially referenced synthetic population in Namibia's Oshikoto region. RTI, an independent, non-profit research institute, has produced a synthetic household population database for 116 million households in the United States (RTI, 2013). The primary contributions of this paper are: (i) to present a novel methodology for the generation of spatially referenced synthetic households by combining spatial microsimulation and dasymetric modeling techniques; and, (ii) to present a case study that includes field validation from a georeferenced survey.

Methodology

Our proposed methodology combines spatial microsimulation techniques for generating

synthetic populations with dasymetric modeling techniques for determining population distribution to produce a dataset of spatially referenced synthetic households in the form of point data with attached attributes. Our methodology entails three discrete steps: 1) generation of synthetic households inclusive of selected attributes to the lowest reasonable level of aggregation; 2) spatial disaggregation of household units using dasymetric modeling techniques; and, 3) pairing of synthetic households with location of household units based on the population distribution identified in the dasymetric model.

1.4.Defining the source areas and associated household aggregates

We start by defining l non-overlapping source areas A_i: $i = 1, \dots, I$ where

$$\bigcup_{i=1}^{I} A_i = A$$

Each A_i is known to contain n_i households. Source areas would typically pertain to discrete administrative units (e.g. districts, states, etc.), and the number of households within each source area would typically be informed by a population census where data aggregates are made publicly available as attributes linked to those discrete administrative units.

1.5.Generate synthetic households

The methodology we use for the population synthesis uses the framework provided by the R package simPop (Templ et al. 2017) which is an open-source and flexible synthesizer based on a modular object-oriented concept. It relies on aggregate tables for population margins and a microdata sample, both ideally from a population census, where the microdata sample contain at least all variables available in the aggregated tables. The steps for generating the synthetic population can be summarized as follows:

- Calibrate survey weights in the microdata to fit population margins

- Extrapolate the microdata using alias sampling to generate the synthetic population
- Model variables of interest, household or personal variables, on the microdata and "predict" them for each person in the synthetic population
- Use simulated annealing (SA) to the synthetic population to known population margins

SA is the heuristic procedure used to approximate complex optimization problems (Templ et al., 2017). The SA algorithm has shown to outperform other methods, including deterministic re-weighting and conditional probabilities (Harland et al. 2012).

1.6.Calibration

If the available microdata does not already have suitable sampling weights they need to be created or fitted to known population margins. Ideally one should use the available aggregate tables and constructing suitable sampling weights. For the calibration technique we use iterative proportional fitting (IPF) (Deming and Stephan 1940), also known as matrix raking.

1.7.Extrapolate microdata

Next, we define the household structure of the synthetic population, using variables such as age, sex, geographic areas and combinations thereof. Variables for geographic areas must at least include the non-overlapping source areas $A_i: i = 1, \dots, I$ or a more detailed sub-grouping of those. With these basic variables, including a personal and household identifier, the structure of the synthetic population is simulated using the calibrated survey weight and alias sampling (Walker 1977). The resulting synthetic population contains each of the variables selected for defining the household structure. Additional variables can be added to the synthetic population using modeling techniques.

1.8.Model variables

The use of synthetic population greatly increases with having a richer set of variables available for each unit. To enrich our synthetic population, one can use models fitted to the survey microdata to predict additional variables onto each unit in the synthetic population using. Those can be personal variables as well as household variables. Let $\mathbf{v} = (v_1, v_2, \dots, v_p)$ be a vector of variables available for some or all units in the microdata as well as the synthetic population and y_j a variable available in the microdata. We fit a model f depending on the set of variables in $\mathbf{v}$ to all units j in the microdata where y_j is non-missing:

$$y_j = f((v_1, v_2, \dots, v_p)_j, \theta) + \epsilon \quad \forall j: y_j \text{ non} - \text{missing}$$

with ϵ as a random error term and θ as a set of parameters which are estimated by fitting the model. The variable y is then predicted for each unit i from the synthetic population using $\hat{y}_i$

$$\hat{y}_i = f((v_1, v_2, \dots, v_p)_i, \theta) \quad .$$

The choice of model mainly depends on the type of variable, for instance logistic regression for binary variables or linear regression for continuous models. The choice of dependent variables should be chosen based on expert knowledge and tested using out of sample prediction techniques like k-fold cross-validation.

The synthetic population now holds $u_i, i = 1, \dots, N_u$ people within N_h households $h_j, j = , \dots, N_h$. Each person has a set of variables $\mathbf{x} = (x_1, \dots, x_{\tilde{q}})$ which have either been simulated through the household structure or predicted using a fitted model. Population margins derived from this synthetic data set might deviate from known population aggregates. This can easily be the case for any of the population margins derived using a predicted variable since the underlying models are simplifications of the underlying true mechanisms and they also include random error terms. Thus, the synthetic population needs to be adjusted to fit specified population margins. The main idea of the procedure is to duplicate the synthetic

population k times, then choose a sub-population thereof which fulfils each of the population margins up to a pre-specified error ϵ. Since this is a combinatorial optimization problem we use simulated annealing, as implemented in R package simPop, to solve it. The simulated annealing procedure starts by selecting each of the units in the original synthetic population and the swapping households at random between the set of already selected households and non-selected households, e.g. the $k-1$ duplicates of the synthetic population. The procedure stops if the margins of the selected units satisfy the supplied population margins or the number of iterations or time elapsed reaches a maximum.

1.9.Applying dasymetric mapping on the synthetic population

The synthetic population now holds $h_i, i = 1, \ldots, n_i$ households in each source area A_i. Each unit u_j in each household has a vector of attributes $\mathbf{v} = (v_1, v_2, \ldots, v_q)$ which can be associated either with the entire household (e.g. source area A_i of household, house structure type, etc.) or to individuals within the household (e.g. age, sex, etc.). The aim of dasymetric mapping is to enhance the geographic resolution of the synthetic population, e.g. mapping each household first to higher resolution grid cells, or eventually to unique x and y coordinates.

1.10. Defining sets of spatially disaggregated synthetic households within each source area into smaller target cells

For dasymetric population modeling, we first partition the source area A into many smaller target cells of equal area $C_j : j = 1, \ldots, m$. Each cell C_j is then allocated to one and only one source area A_i. There will inherently be cells which intersect the boundary of a source area to which the modeler must define a rule for determining singular allocation. For example, one could assign a target cell C_j to the source area A_i if the centroid of C_j, γ_j lies in A_i, $\gamma_j \in A_i$.

Thus, each A_i is is composed of m_i cells which we label $C_{i,j}: j = 1, \dots, m_i$. We then define a density variable $d_{i,j}$, with:

$$d_{i,j} \geq 0 \quad \wedge \quad \sum_{j=1}^{m_i} d_{i,j} = 1$$

The density variable may be obtained by combining a binary masking variable $b_{i,j}: b_{i,j} \in \{0,1\}$ with some other variables $z_{i,j}$ which contain allocated values of a specific attribute for each $C_{i,j}$. A binary variable should indicate whether people are living in a grid cell, such as the existence of housing infrastructure. For example, if housing infrastructure exists in target cell $C_{i,j}$ then $b_{i,j} = 1$, if not then $b_{i,j} = 0$.

The density variable $d_{i,j}$ can be calculated through

$$d_{i,j} = \frac{b_{i,j} z_{i,j}}{\sum_{j=1}^{m_i} b_{i,j} z_{i,j}}$$

The expected number of households, $\overline{N}_{i,j}$, in target cell $C_{i,j}$ is then defined by

$$\overline{n}_{i,j} = d_{i,j} n_i \quad .$$

Since $\overline{N}_{i,j}$ is not necessarily an integer value we define an integer realization of the number of households in $C_{i,j}$ by

$$\hat{n}_{i,j} = \lfloor \overline{n}_{i,j} \rfloor + \delta$$

where $\lfloor x \rfloor$ denotes the largest integer smaller or equal x and $\delta_{i,j}$ a random variable taking the values 0 or 1.

Within each target area A_i the realizations of the sum of random variables $\delta_{i,j}, j = 1, \dots, m_i$ must equal $n_i - \sum_{j=1}^{m_i} \lfloor \overline{n}_{i,j} \rfloor$ thus

$$\delta_{i,j} = \begin{cases} 1 & \text{if} \quad x_{i,j} \leq x_{(t)} \\ 0 & \text{otherwise} \end{cases}$$

$$t = n_i - \sum_{j=1}^{m_i} \lfloor \overline{n}_{i,j} \rfloor$$

with $(x_{i,j}, \dots, x_{i,m_i})$ a vector of random values drawn from a uniform distribution $U(0,1)$ and $(x_{(1)}, \dots, x_{(m_i)})$ the order statistic of these randomly generated values.

1.11. Adding coordinates to synthetic households

The synthetic households $h_i, i = 1, \dots, N_h$ can then be disaggregated to an individual target cell $C_{i,j}$ thus creating spatially-referenced synthetic households $\hat{h}_{i,j}, i = 1, \dots, n_{i,j}$. This can be done randomly if spatial dependencies of attributes are insignificant or not considered for the analysis. The random selections must consider that x and y coordinates of households may not overlap. Using random selection can be justified easier in cases where the target cells $C_{i,j}$ cover relatively small areas (e.g. less than 100 square meters). If specific attributes show spatial dependencies, hybrid techniques can be deployed which integrate spatial covariates to weigh the spatial distribution of the synthetic households h_i (Monteiro et al. 2017). The next step entails creating point values for the spatially-referenced synthetic households $\hat{h}_{i,j}$ within each target cell $C_{i,j}$. The precise x and y coordinates of the point values for spatially-referenced synthetic households $\hat{h}_{i,j}$ can be generated randomly within the target cell $C_{i,j}$ or can be weighted or snapped to a spatial covariate (e.g. building footprint centroids). The new spatially-referenced synthetic households $\hat{h}_{i,j}$ are assigned spatial attributes x and y in addition to those joined from the synthetic household h_i such that for each $\hat{h}_{i,j}$ holds variable vector $\mathbf{v} = (v_1, v_2, \dots, v_q, x, y)$.

2. Case study of Dumuria Upazila, Bangladesh

To demonstrate the efficacy of these methods, we look at three Unions in Bangladesh: 1) Bhandar Para; 2) Sahas; and, 3) Sarappur. All three are located within Dumuria Upazila, Khulna District. Taken together, the three unions are contiguous, rural and located primarily within Polder 29 in the Bangladesh Coastal Zone.

In total, our study area is 83.96 square kilometers and contains 51,497 residents within 12,697 households. We define our study area as all areas covered by the 2011 National Census polygons for our three selected Unions (administrative level 4). A summary of key attributes for each Union can be found in Table 1 and a map of our study area can be found in Fig. 2. Our three Unions contain a combined 71 Mauzas (administrative level 5). The average Mauza in the study area is 1.18 square kilometers. However, there is a large variation in size, with the largest being 6.69 square kilometers and the smallest 0.15 square kilometers. The 2011 National Census only covers "populated Mauzas", of which 59 Mauzas within the study area are classified.

Table 1 Area, Population, and Household Data for Bhandar Para, Sahas, and Sarappur Unions

	Area (sq. km)	Population	Households
Total	***83.96***	***51,497***	***12,697***
Bhandar Para	34.89	16,749	4,077
Sahas	25.29	18,647	4,498
Sarappur	23.78	16,101	4,122

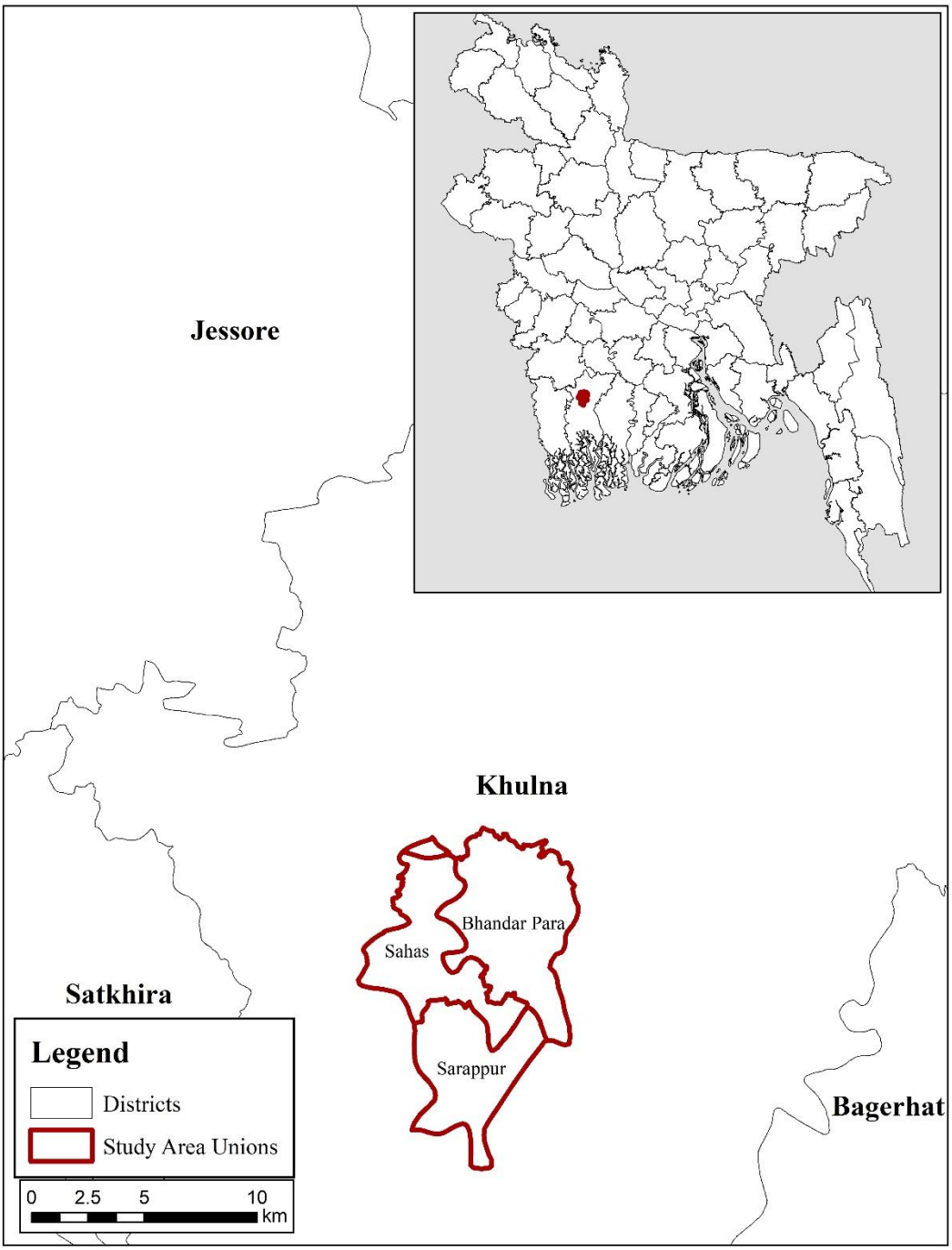


Fig. 1 Map of Study Area Unions

Our study area was selected in consultation with the REACH Program, a global research program aimed at improving water security for the poor. The REACH Program has a sub-project within Polder 29 which includes a georeferenced household survey. Polder 29 provides an interesting case for further usage of our output dataset for agent-based modeling. Polders can be modeled as enclosed systems within the context of broader biophysical

conditions in which the interactions between flood events, salinity, water infrastructure, agricultural production and income can be tested (Borgomeo et al. 2017). Each of these inherently has a spatial element within a polder system, and the output dataset from this study may provide more insight into those spatial heterogeneities. Taken together, our study area presents an opportunity to apply our methodology in a well-mapped and uniquely studied coastal system which provides the means for on-the-ground verification of results.

2.1.Data

For census aggregates, our study utilizes data from the 2011 National Census, produced by the Bangladesh Bureau of Statistics (BBS). Census data were used both in the dasymetric mapping segment and the population synthesis segment of our study. For dasymetric mapping, the Union polygons provided the basis for spatial disaggregation of households, while the Mauza polygons provides a means for verification through reaggregation of households into lower-level administrative boundaries. For population synthesis, microdata from the 2011 National Census released through the Integrated Public Use Microdata Series (IPUMS) for Dumuria Upazila were used to extrapolate household composition and attributes.

Our study area was also the target of a "mapathon", which took place over a period of five hours on August 25, 2017 by volunteers mobilized from OpenStreetMap (OSM) Bangladesh. Specific features mapped included over 22,000 building footprints, 200 roads, 100 waterways, and 300 land-use polygons. These features, particularly building footprints, are important for the dasymetric population modeling section of our study, which uses the layer as a binary mask for determining areas of potential habitation.

Lastly, this study relies on a first phase survey by conducted by the REACH Program in Polder 29 from December 17, 2017 to January 7, 2018. The survey collected attributes for

707 households within our study area, representing 5.6 percent of all households. For the survey, households were selected randomly within each Mauza to ensure a representative distribution, and GPS coordinates were logged of each household location. Individual household locations were used to validate the output dataset from this study.

A summary of data layers used in the analysis can be found in Table 2.

Table 2 Data Layers Used in Dumuria Upazila Analysis

Type	Description	Source	Pop Synth	Pop Mod	Field Ver
National Census Aggregates/ Shapefiles	2011, administrative levels 3 (Upazila), 4 (Union) and 5 (Mauza)	Bangladesh Bureau of Statistics (BBS)			
National Census Microdata	2011, administrative level 3 (Upazila)	BBS			
Building Footprints	2017, manually traced	OpenStreetMap (OSM)			
Polder 29 REACH Survey	2017 – 2018, 707 households within study area	REACH Survey			

2.2.Analysis and Results

2.2.1. Population Synthesis

To illustrate the functionality of our model for generating spatially referenced synthetic households, we rely on the R package simPop for population synthesis. The workflow used can be viewed in Fig. 2.

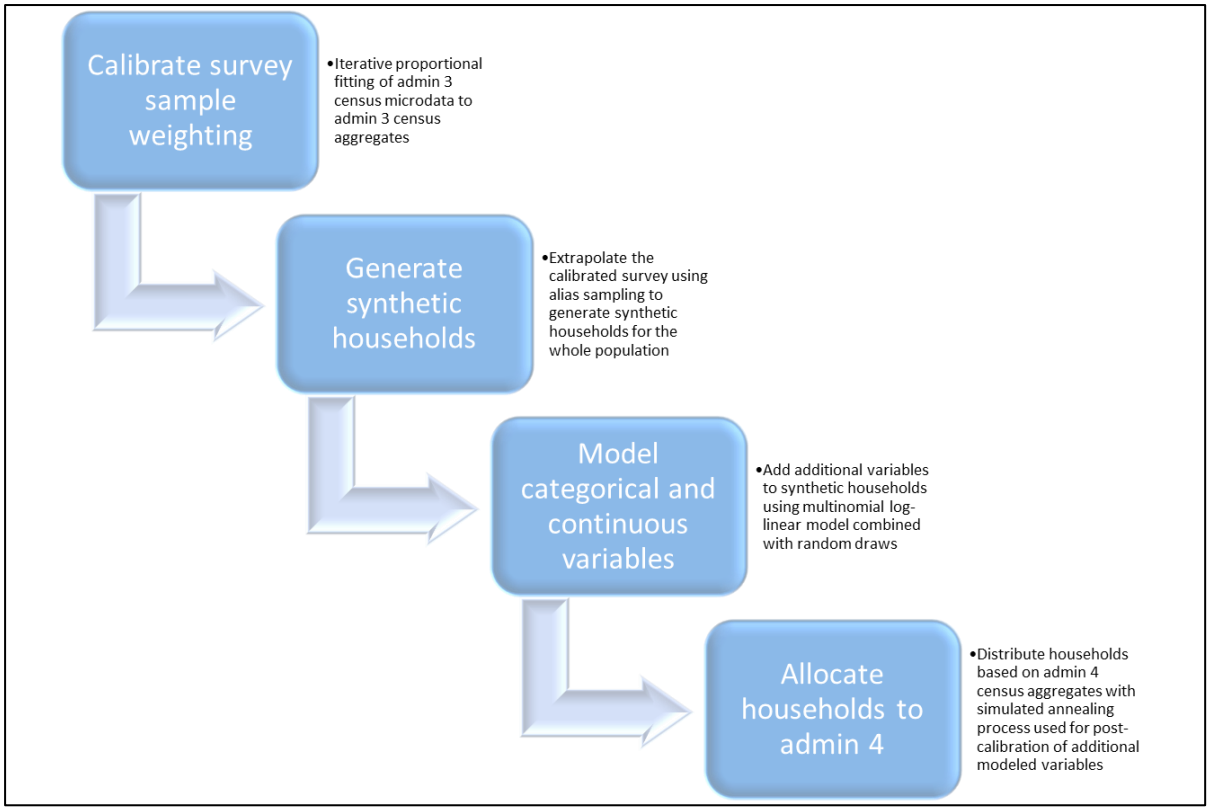


Fig. 2 Workflow for Simulation of the Population with SimPop

We start by calibrating the sample survey microdata for Dumuria Upazila to its census aggregate data using iterative proportional fitting, implemented in the R-package surveysd. Next, we define the household structure of the synthetic population by extrapolating the calibrated survey sample using alias sampling (Walker 1977) with attributes for household and personal identifiers, household size and sampling weight. Because Dumuria is the only Upazila considered in our analysis we do not need to consider strata. Next, we separately model continuous and categorical variables of individuals within each household. The only continuous variable we consider in our analysis is age. For categorical variables, we consider sex (male or female) and relation to head of household (spouse_partner, child, other_relative, or non_relative). For both, we use simPop's built-in functions to employ a multinomial log-linear model combined with random draws (Templ *et al.* 2017).

We next employ simPop's function for allocating households into smaller areas, which

allows us to place our synthetic households into discrete Unions as target areas $A_i: i = 1, \ldots, I$ within Dumuria Upazila. This process uses a simulated annealing algorithm, which entailed an iterative process of swapping households for optimal combinations based on our known margin of total population within each Union. Our final output is synthetic population data for all Unions within Dumuria Upazila, although only three are considered further in our analysis. Selected combined results for the three Unions considered can be viewed in Figs. 3 and 4.

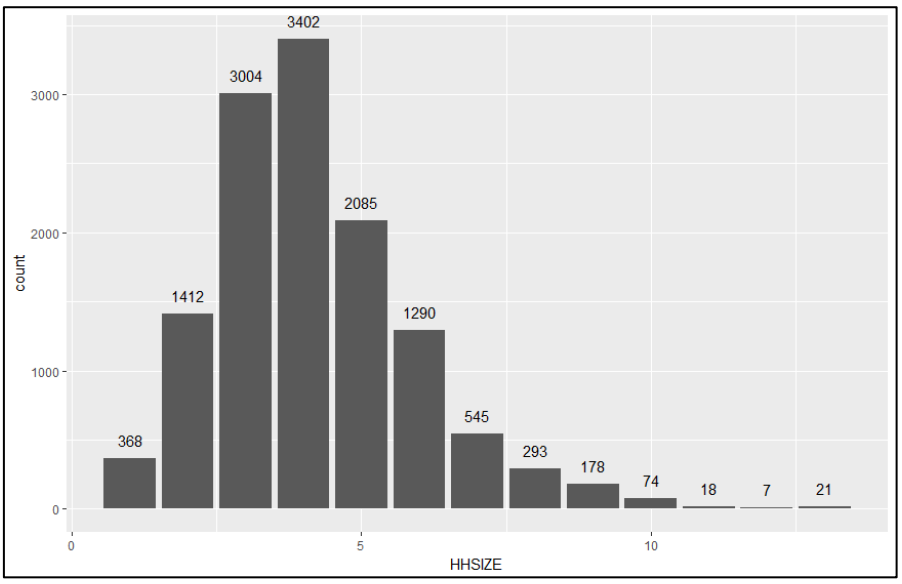


Fig. 3 Combined distribution of household size within study area

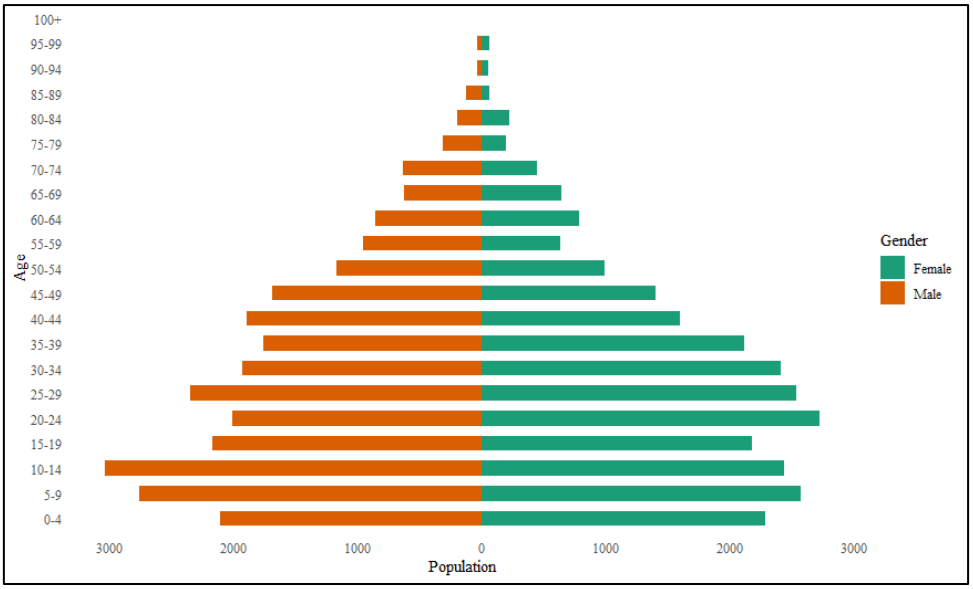


Fig. 4 Combined synthetic population pyramid for study area

2.2.2. Dasymetric Population Modeling

For dasymetric population modeling, we start by applying a 30-meter resolution grid across the study area. We selected 30-meter resolution to align with the resolution presented by the HRSL. We then overlay building footprint polygons and calculate the area within each cell covered by a building footprint polygon. The total number of households within each Union (our target area A_i) are then allocated to cells $C_{i,j}$ within that Union using the building footprint polygons as a binary mask and the building footprint area within each cell as a proxy for density. To calculate our density variable, we divide the footprint area within each cell by the sum of building footprint areas within each Union and multiply by the population of the Union to get a real number of households per cell $\hat{n}_{i,j}$. From there, we separate the integer from the decimal part and classify all integers as households. Decimal parts are then used for weighted random allocation of remaining unallocated households within each Union. The resulting layer is of 30-meter resolution gridded households, where each cell $C_{i,j}$ is associated with an attribute for $\hat{n}_{i,j}$households. Lastly, $\hat{n}_{i,j}$ random points were created within each cell $C_{i,j}$ using the Create Random Points tool in ArcMap 10.6. The process for achieving

modeled household points can be seen in Fig. 5.

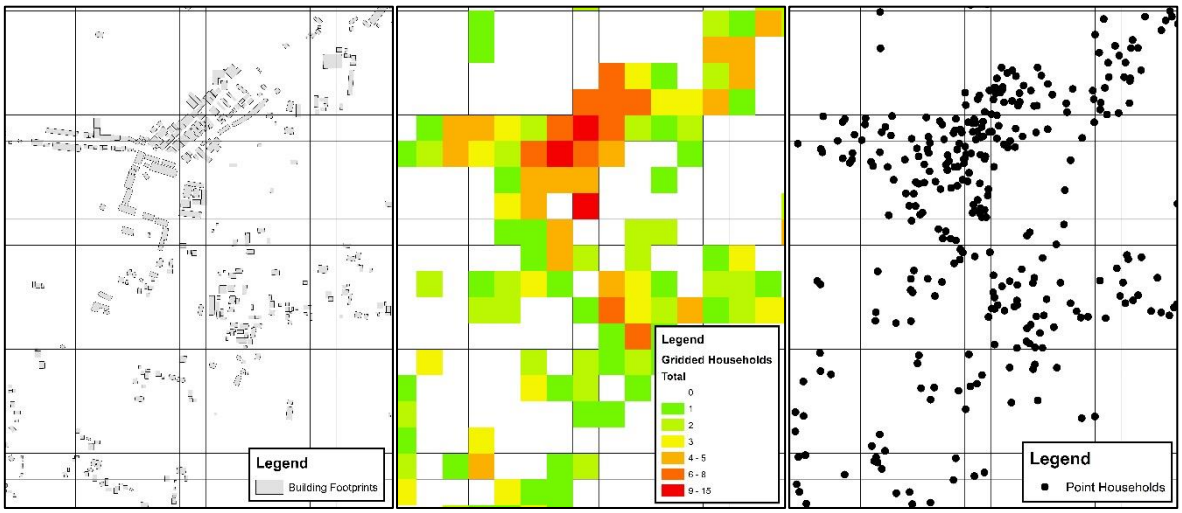
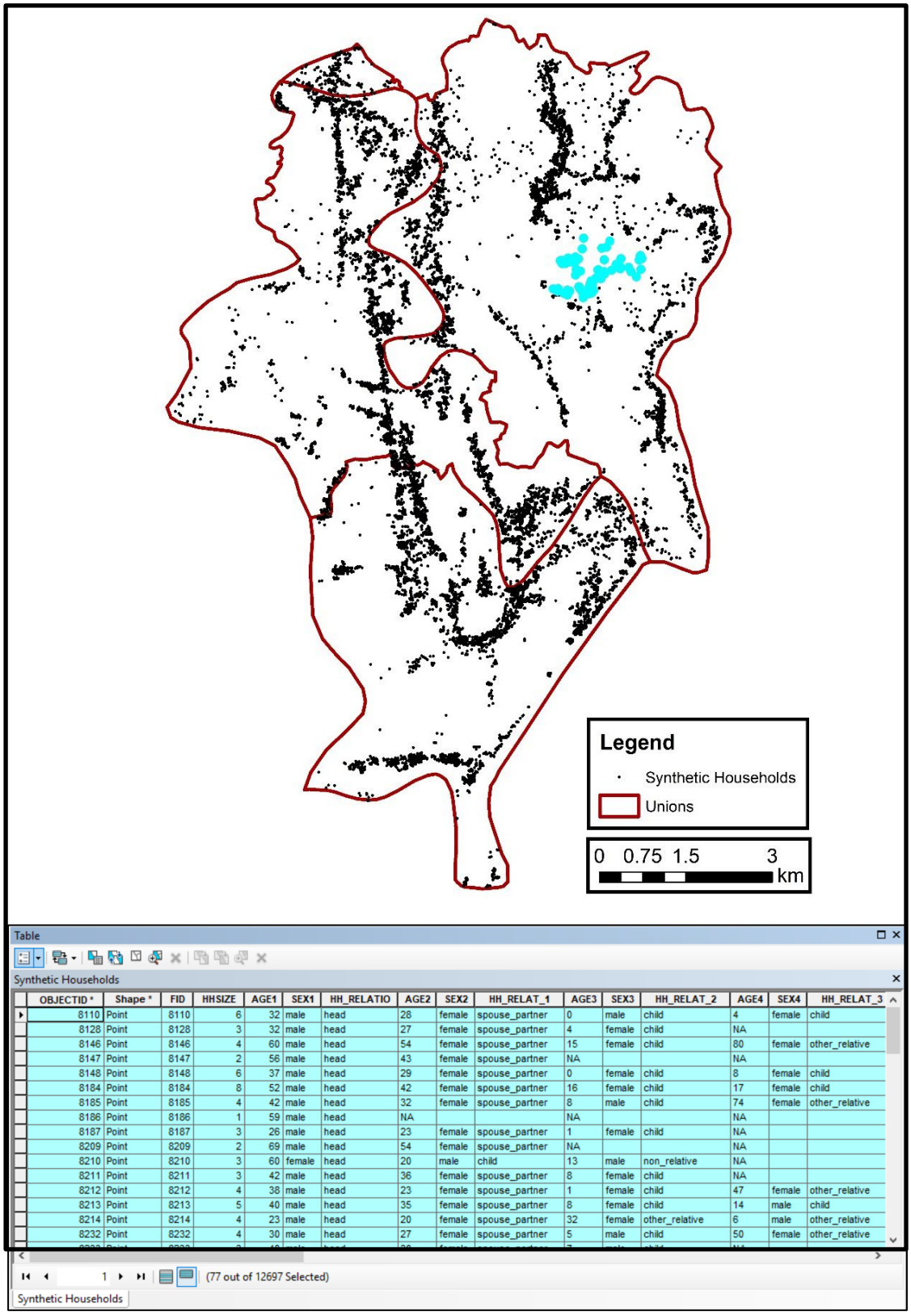


Fig. 5 30-meter resolution grid cells overlain by building footprints (left), 30-meter resolution modeled gridded households (center), and modeled household points (right) in Sahas Union, Khulna District

2.2.3. Spatially Referenced Synthetic Households

The last step entails joining the tabular synthetic households with the point features. We first randomize the order of the synthetic households then assign each a unique identification attribute. We then execute a join (table to layer) which matches each point with a synthetic household by matching the unique identification attributes of both. This is done for each of the three Unions. After which, each synthetic household point feature contains attributes such that $\mathbf{v} = (v_1, v_2, \ldots, v_q, x, y)$. The results can be seen in Fig. 6.



OBJECTID *	Shape *	FID	HHSIZE	AGE1	SEX1	HH_RELATIO	AGE2	SEX2	HH_RELAT_1	AGE3	SEX3	HH_RELAT_2	AGE4	SEX4	HH_RELAT_3
8110	Point	8110	6	32	male	head	28	female	spouse_partner	0	male	child	4	female	child
8128	Point	8128	3	32	male	head	27	female	spouse_partner	4	female	child	NA		
8146	Point	8146	4	60	male	head	54	female	spouse_partner	15	female	child	80	female	other_relative
8147	Point	8147	2	56	male	head	43	female	spouse_partner	NA			NA		
8148	Point	8148	6	37	male	head	29	female	spouse_partner	0	female	child	8	female	child
8184	Point	8184	8	52	male	head	42	female	spouse_partner	16	female	child	17	female	child
8185	Point	8185	4	42	male	head	32	female	spouse_partner	8	male	child	74	female	other_relative
8186	Point	8186	1	59	male	head	NA			NA			NA		
8187	Point	8187	3	26	male	head	23	female	spouse_partner	1	female	child	NA		
8209	Point	8209	2	69	male	head	54	female	spouse_partner	NA			NA		
8210	Point	8210	3	60	female	head	20	male	child	13	male	non_relative	NA		
8211	Point	8211	3	42	male	head	36	female	spouse_partner	8	female	child	NA		
8212	Point	8212	4	38	male	head	23	female	spouse_partner	1	female	child	47	female	other_relative
8213	Point	8213	5	40	male	head	35	female	spouse_partner	8	female	child	14	male	child
8214	Point	8214	4	23	male	head	20	female	spouse_partner	32	female	other_relative	6	male	other_relative
8232	Point	8232	4	30	male	head	27	female	spouse_partner	5	male	child	50	female	other_relative

Fig. 6 Point features for spatially referenced synthetic households and sample from accompanying attribute table for selected features (in light blue)

2.3.Validation

2.3.1. Field validation of household locations through the REACH Survey

To assess the performance of our results, we benefit from access to the spatially referenced REACH survey. The REACH survey contains 707 households comprised of 3,124 persons within our study area in comparison to our 12,697 synthetic households comprised of 25,964 persons. We first compare the age distribution of the surveyed population with that of our synthetic population. Although our synthetic population provided real numbers 0 – 99, the survey only provided categorical information which we used to aggregate our synthetic populations. Although the synthetic populations were skewed towards a younger population, results in all categories are within five percent of the surveyed population. The results can be seen in Fig. 7.

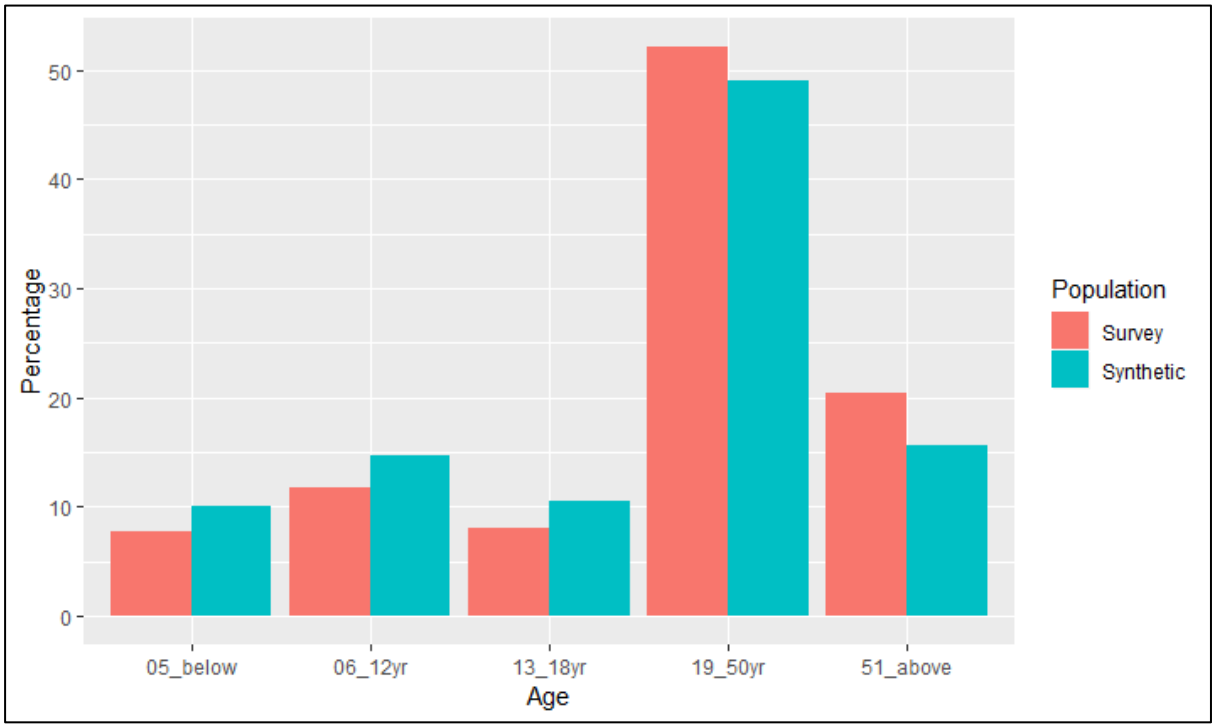


Fig. 7 Age distribution of surveyed vs synthetic populations

We then use the geolocation of the households surveyed to validate the location of our synthetic households. We do this by calculating for each surveyed household the distance to

the nearest synthetic household. Although the analysis reveals a few distinct outliers, 98 percent of all surveyed households were within 100 meters of at least one synthetic household, and the median distance observed was only 18 meters. This pattern can be clearly detected in Fig. 8.

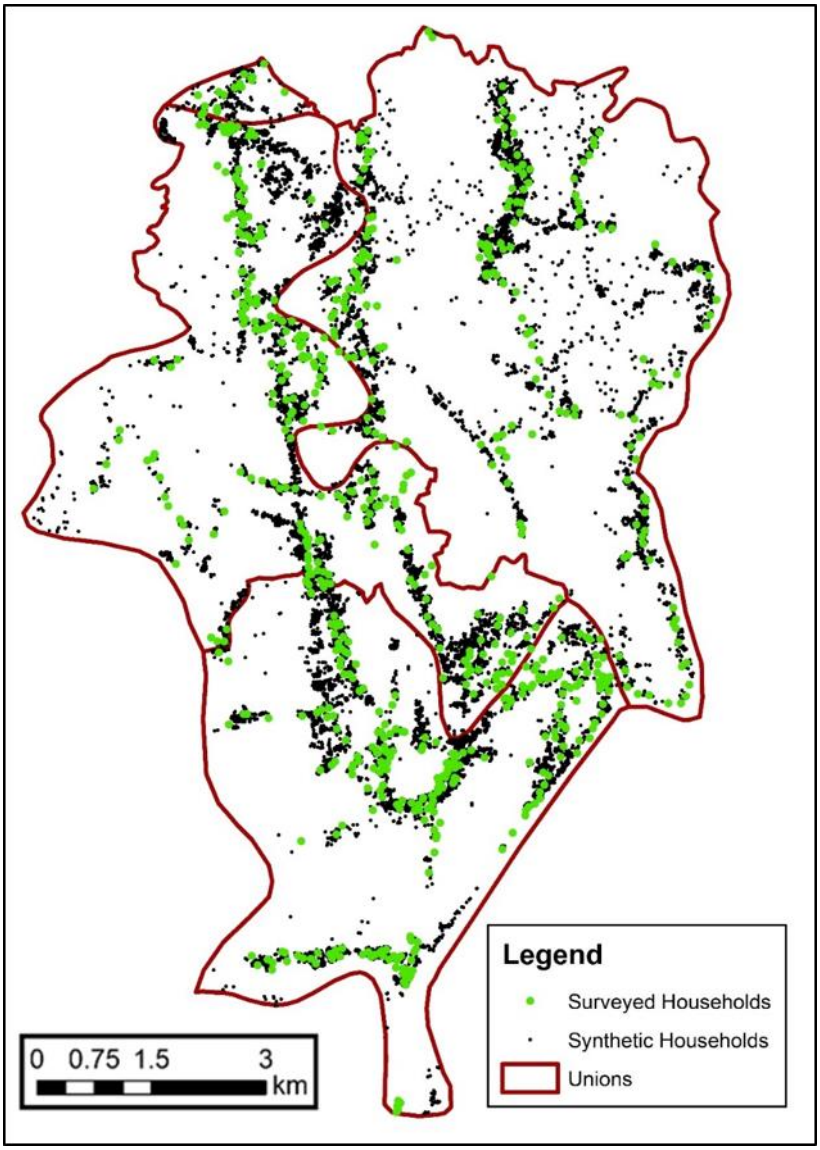


Fig. 8 Point features for spatially referenced synthetic households overlaid by point features for surveyed households

2.3.2. Performance through RMSE and MAE of lower-level administrative aggregates

Our model spatially disaggregates synthetic households at the Union level. To validate our results, we reaggregate our synthetic households (households predicted) at the Mauza level

and compare results to the census aggregates (households observed). However, we only do this for the 27 Mauzas within Sahas and Sarappur Unions as there are noticeable errors in the census aggregates for Bhandar Para Union visible in Fig. 9. Sitarampur Mauza in particular shows clear patterns of settlement despite being classified as unpopulated per the 2011 National Census. These errors were not observed within Mauzas of Sahas and Sarappur Unions.

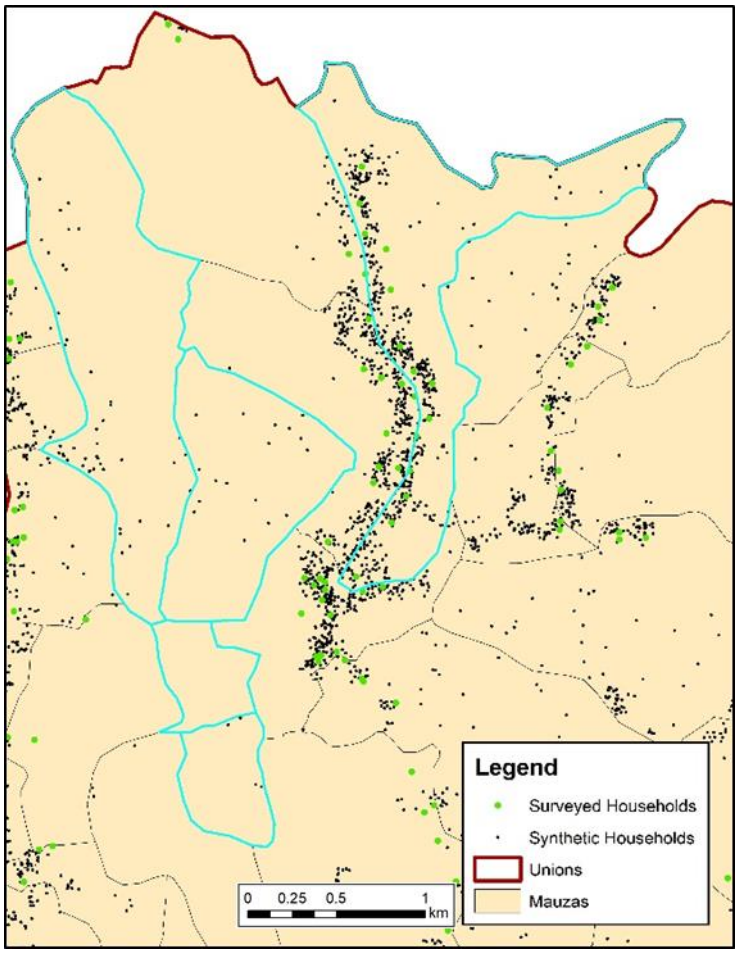


Fig. 9 Selected Mauzas within Bhandar Para Union are designated as "NF BBS 2011" per the 2011 National Census and are listed as unpopulated

The performance of our model can be seen in Fig. 9. We generate the root mean square error (RMSE) and mean absolute error (MAE) of lower-level administrative aggregates. We also express each of the errors as a percentage of the average number of households per Mauza

tested.

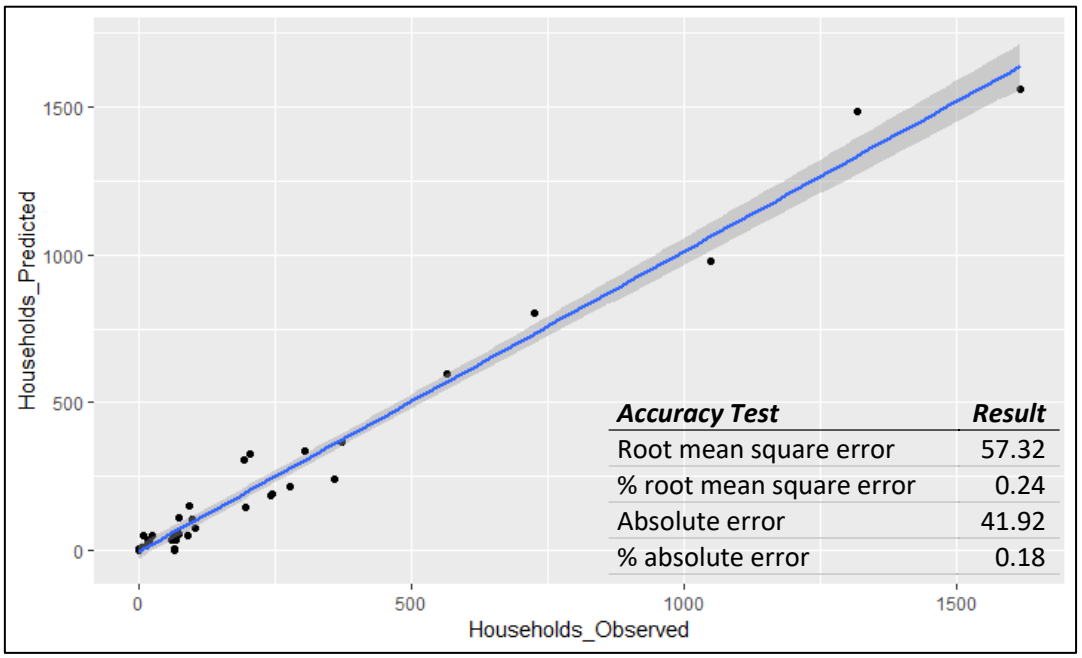


Fig. 9 Linear model of households observed vs households predicted with 95 percent confidence interval and results of accuracy tests for 27 Mauzas with Sahas and Sarappur Unions

3. Discussion

3.1.Introducing a Paired Approach for Generating Geospatially Referenced Synthetic Household Agents

The methodology proposed in this paper effectively combines spatial microsimulation techniques for generating synthetic populations with dasymetric modeling techniques for determining the spatial distribution of populations to produce a dataset of spatially referenced synthetic households in the form of point data with attached attributes. This represents the natural convergence between two prominent geospatial workarounds for better understanding population dynamics without having full access to primary information. While this technique is exceedingly difficult to verify at lower resolutions (perhaps why it has yet to be done), our study has the advantage of being able to draw upon a spatially-referenced survey of

households to verify both modeled location information and associated household attributes in a contained geographic area.

Although the circumstances at first appear unique, they will become increasingly accessible with advancements in technology and proliferation of publicly available microdata and geospatially referenced survey data. The techniques used for high-resolution dasymetric mapping in particular are increasingly achievable at higher accuracy and at a larger scale. At the same time, the increased prevalence of higher resolution microdata releases and advancements in spatial microsimulation techniques have enhanced the ability in which we can confidently predict socioeconomic characteristics of those mapped populations. Efforts such as IPUMS and other data repositories have facilitated this by encouraging the open release of microdata and cataloguing its availability for public usage. Still, public releases of census microdata only cover a small percentage of the total individuals and households for the aggregated unit (typically up to 10 percent), thus making population synthesis necessary to generate attributes for the entire population reflective of its general composition.

While SAE techniques can increase the resolution in which we understand the spatial distribution of attributes, they are still restricted to aggregates so would not pick up on clusters and patterns of settlement within those aggregated areas. Thus, despite an inability to effectively differentiate household attributes spatially within the finest level of aggregation, it is still a worthwhile exercise to disaggregate those households in line with spatial patterns of development. This is particularly true if only interested in attributes which would not have strong spatial dependencies within a small administrative unit, such as household size within predominantly rural areas, or composition of sexes within a household. For spatial attributes with stronger potential spatial dependencies (e.g. wealth, water access, etc.) the spatial distribution can be further refined within the administrative unit by incorporating spatial

modeling techniques which consider factors such as the spatial autocorrelation of selected attributes, dependencies related to additional known covariates (e.g. poorer households may live further from water access) or additional georeferenced surveys which can unlock spatial dependencies for attributes collected which are also contained within the census data. When analysis is combined with other spatial data sets (e.g. flood risk, groundwater contamination, etc.), even a limited number of attributes can prove quite useful when households are fixed to a geographic location.

While recent efforts have been made to better approach challenges of geographic partitioning for data anonymization, none have resolved the fundamental issue that providing a spatial reference compromises the anonymity of potentially sensitive data (Croft et al. 2017). Certain applications, particularly spatial ABM modeling and spatial microsimulation stand to benefit heavily from household-level spatial information. For these applications, spatially-referenced synthetic populations could provide a workable solution for preserving security of potentially sensitive data. Spatial ABMs require a modeled landscape to capture spatial heterogeneities of inputs, which our proposed method can provide. These modeled landscapes are often comprised dynamic networks of interacting agents, such as in an ecosystem, city, or financial market (Grimm et al. 2006). The challenge in spatial ABMs is to capture the interplay between discrete representations of space and network or diffusion models, so the more detail available on the discrete representations the better. Past models have done this through aggregated administrative zones (Smajgl and Bohensky 2013), division of "city area types" (Gaube et al. 2013), high-resolution land use (Sun and Muller 2013), and various other criteria. The approach that we propose in this paper advances the resolution dramatically, by specifying the precise location of each household within the synthetic population set through a novel method which incorporates the latest techniques of dasymetric population modeling. Using our proposed method, discrete representations could thus include individual

households.

The primary advantages of distributing the populations spatially into discrete households are two-fold: 1) it allows for spatial patterns of settlement to be considered for spatial analysis, effectively breaking down barriers of potentially arbitrary administrative units; and, 2) it allows for spatial ABM of populations into complex ecosystems, including the interaction with other mapped physical features including roads, rivers, economic centers, etc. The outputs of our study provide a ground-truthed first take at producing spatially referenced synthetic households using only census aggregates, household microdata, and gridded population data; all of which are readily available for most countries in the world.

3.2.Applying our model to enhance local decision-making capabilities

The output dataset of spatially referenced synthetic households can be used as an input for further spatial analysis and agent-based modeling. An example of this at a basic level in the Bangladesh Coastal Zone would be to use a roads layer to calculate travel times of households to cyclone shelters. This would be helpful for prioritization of new cyclone shelter construction, as well as planning of outlet roads. Similar accessibility layers could be produced for access to schools or health centers.

Flood risk is another domain where our output datasets may provide utility. To exemplify this, we overlaid our spatially referenced synthetic households layer with a layer showing areas that have been inundated at least once from 1985-2010 (Winsemius et al. 2013). We found that roughly 15 percent of households lie in a flood prone area despite 36 percent of the study area lying in a flood prone area. It makes conventional sense that human settlement would avoid flood prone areas; however, with only administrative boundaries and census aggregates it is difficult to come to a substantiated conclusion. We can take this a step further

to analyze information derived from household and individual attributes. From our output layer, we can show that approximately 3,000 children live within flood prone areas, with children represented in 86 percent of all exposed households.

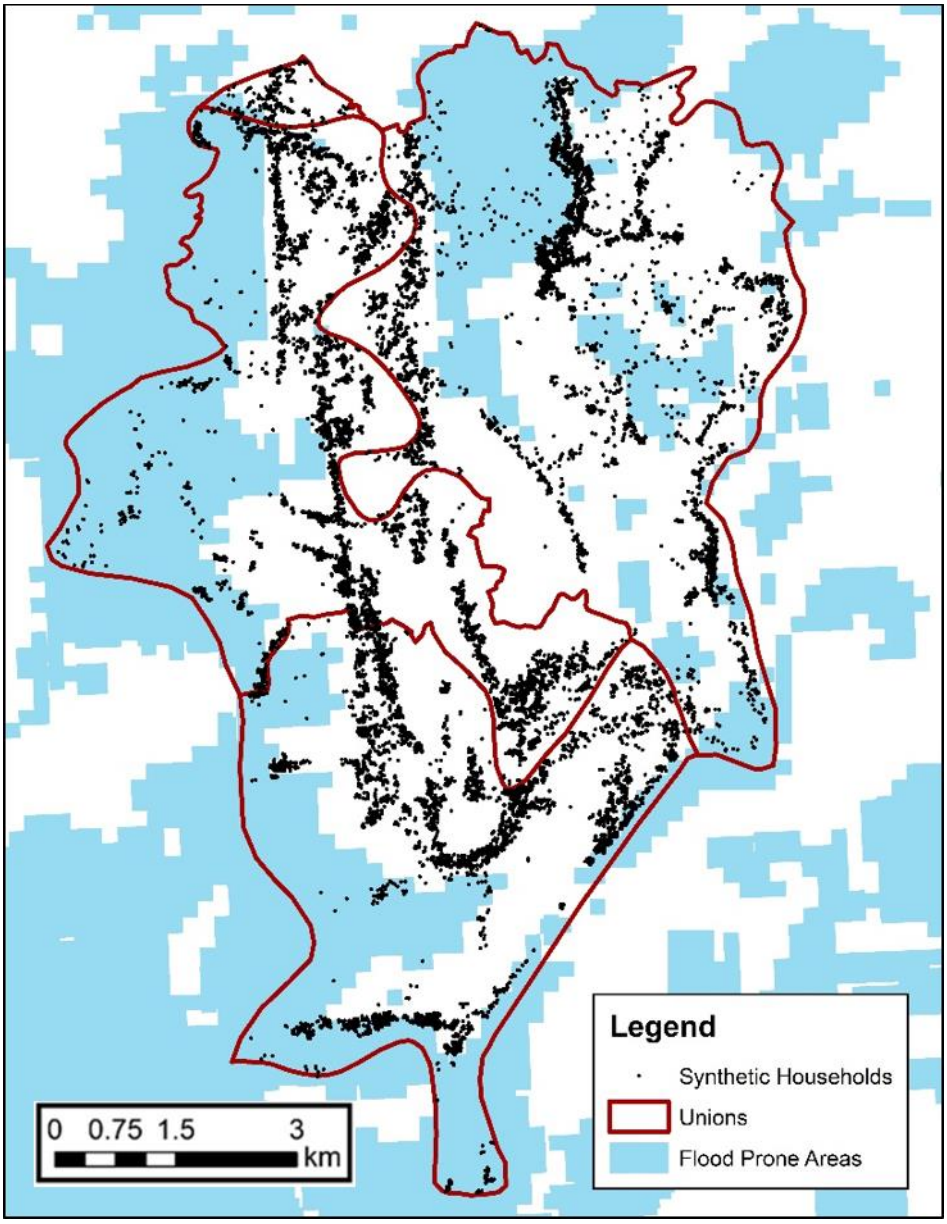


Fig. 11 Overlay of areas that have been inundated at least once on the full Dartmouth Flood Observatory record (1985-2010) and spatially referenced synthetic households

This information can be helpful to improve analysis of biophysical characteristics on human settlement patterns and their corresponding socioeconomic attributes. Agent-based modeling

can then be employed to analyze the movements and effects of flood events on household agents. Each household can thus be viewed as either a georeferenced and fixed automaton or a non-fixed automaton depending on the attributes considered (e.g. house structure is a spatially fixed asset whereas value attached to the individual household members would be consider non-fixed). This would be helpful for analyzing the impacts on welfare of households affected by disasters.

3.3.Data-driven advancements

Our model is designed to be flexible to advancements in technology on multiple fronts. Dasymetric population modeling depends heavily on earth observation derived data. Increasingly, this data is becoming available at higher-resolutions and at lower costs. Free satellite data from the European Union's Copernicus Program combined with better post-processing techniques which rely heavily on machine learning have led to more and better information at scale. Machine learning techniques will also help to close the gap on modeling socioeconomic characteristics. At a more localized level, the use of drones has proliferated to enable quicker and better access to high-resolution earth observation data including LiDAR where 3D representations and detailed classification of features can help to further refine physical characteristics of structures which may be indicative of the socioeconomic characteristics of those households living within. This can be supplemented using street view imagery to further classify structural attributes of features such as building structure type and façade, or even associating the brands and models of cars within a neighborhood to household income (Gebru et al. 2017).

Our study utilizes survey data for verification purposes. However, another area of advancement lies in the application of open data such as OpenStreetMap, street view imagery and other surveyed data as training sets for machine learning (Diou et al. 2018). Our model

allows for expansion into this area as it provides a top-down then bottom-up approach. This is particularly true for spatially-dependent socioeconomic characteristics where we may have a good sense of the statistical distribution of attributes at an aggregate level but require further spatial information to distribute synthetic populations appropriately. Our study avoids confronting this challenge by only focusing on less spatially-dependent attributes such as household size, gender and age composition. However, going forward this is a critical area of growth which should be further considered.

3.4.Limitations

In this study, we do not project demographic changes for households listed in the 2011 National Census for 2018 (the year of the REACH survey). However, from the 2001 National Census to the 2011 National Census, the population of Khulna District (administrative level 2) remained relatively steady, decreasing slightly from 2.38 million to 2.32 million (Bangladesh Bureau of Statistics, 2015). This study focuses primarily on validating the precise location of spatially referenced synthetic households, which are unavailable from the 2011 National Census data. Thus, although there may be slight changes from 2011 to 2018 in total population within our study area (e.g. negative population growth due to broader urbanization trends), we do not see this as a major limitation. The larger limitation is that due to lower fertility rates, the age structure of the population will have shifted: the United Nations estimates that the population aged 0-14 will have reduced from 29.3 percent in 2010 to 26.8 percent in 2020, while the population aged 15-64 will have increased from 65.6 percent in 2010 to 68.0 percent in 2020 (UN DESA, 2017). Furthermore, while the output layer show promise, there is still room to improve the types of attributes included in the analysis by incorporating additional methods for distributing synthetic population across the study area. This is particularly important for potentially spatially heterogeneous attributes

such as household welfare, which may be better informed by including additional spatial covariates such as those related to house structure type, access to water, or type of toilet (see Thomson et al., 2018), and machine learning methods which use survey data as training sets.

4. Conclusions

Our study presents a methodology to generate spatially referenced synthetic households for population modeling. We combine spatial microsimulation techniques to generating synthetic households with dasymetric modeling techniques to inform that spatial distribution of those households. In our case study of Dumuria Upazila, Bangladesh, we highlight the efficacy of our techniques through on-the-ground verification of location and attribute information and re-aggregation into lower-level administrative boundaries.

Distributing populations spatially into discrete households provides advantages even where only limited attributes are included. Population is inherently distributed unevenly across a geographic surface, and our model allows for synthetic households to be represented spatially much better than being associated with potentially arbitrary administrative units. This isn't particularly important for ABM that is focused on populations within aggregate areas. However, new branches of ABM which focus on spatial relationships between agents as individual automata can benefit tremendously. Individual household agents can allow for a deeper understanding of the interactions between households and complex geographic systems.

Our method does present some potential challenges and opportunities in terms of privacy considerations. Advancements in spatial modeling and machine learning techniques are enabling us to know much more about geographic surfaces and the socioeconomic characteristics of populations who inhabit them. While this could present a tremendous

opportunity to improve our insights into everything from poverty reduction programs to infrastructure decision-making, it also requires careful consideration form the side of citizen privacy rights. It is not unreasonable to expect that one could take a small training set of surveyed household data and extrapolate it to generate accurate and potentially sensitive information related to socioeconomic characteristics such as household income or potential to default on a housing loan. Indeed, some of this may be unavoidable should census micro-data truly be replaced by targeted georeferenced household surveys used as training sets. Our model does not take it that far but does shed light into the future direction of population modeling.

References

Armstrong, M.P., Rushton, G. & Zimmerman, D.L. 1999, "Geographically masking health data to preserve confidentiality", *Statistics in medicine,* vol. 18, no. 5, pp. 497-525.

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., Moorman, D., Uhlir, P. & Wouters, P. 2004, "Promoting access to public research data for scientific, economic, and social development", *Data Science Journal,* vol. 3, pp. 135-152.

Bishop, Y.M., Fienberg, S.E. & Holland, P.W. 2007, *Discrete multivariate analysis: theory and practice*, Springer Science & Business Media.

Borgomeo, E., Hall, J.W. & Salehin, M. 2018, "Avoiding the water-poverty trap: insights from a conceptual human-water dynamical model for coastal Bangladesh", *International Journal of Water Resources Development,* vol. 34, no. 6, pp. 900-922.

Cassa, C.A., Wieland, S.C. & Mandl, K.D. 2008, "Re-identification of home addresses from spatial locations anonymized by Gaussian skew", *International journal of health geographics,* vol. 7, no. 1, pp. 45.

Croft, W.L., Shi, W., Sack, J. & Corriveau, J. 2017, "Comparison of approaches of geographic partitioning for data anonymization", *Journal of Geographical Systems,* vol. 19, no. 3, pp. 221-248.

Deming, W.E. & Stephan, F.F. 1940, "On a least squares adjustment of a sampled frequency table when the expected marginal totals are known", *The Annals of Mathematical Statistics,* vol. 11, no. 4, pp. 427-444.

Diou, C., Lelekas, P. & Delopoulos, A. 2018, "Image-Based Surrogates of Socio-Economic Status in Urban Neighborhoods Using Deep Multiple Instance Learning", *Journal of Imaging*, vol. 4, no. 11, pp. 125.

Doxsey-Whitfield, E., MacManus, K., Adamo, S.B., Pistolesi, L., Squires, J., Borkovska, O. & Baptista, S.R. 2015, "Taking advantage of the improved availability of census data: a first look at the gridded population of the world, version 4", *Papers in Applied Geography,* vol. 1, no. 3, pp. 226-234.

Elbers, C., Lanjouw, J.O. & Lanjouw, P. 2003, "Micro–level estimation of poverty and inequality", *Econometrica,* vol. 71, no. 1, pp. 355-364.

Filatova, T. 2013, "Changing climate, changing behavior: modeling housing markets responses", *Modeling Social Phenomena in Spatial Context,* vol. 2, pp. 47.

Gaube, V., Remesch, A. & Smetschka, B. 2016, "Planning, Residential Decisions and Energy Use in Vienna" in *Social Ecology* Springer, , pp. 489-503.

Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E.L. & Fei-Fei, L. 2017, "Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States", *Proceedings of the National Academy of Sciences,* vol. 114, no. 50, pp. 13108-13113.

Ghosh, M. & Rao, J. 1994, "Small area estimation: an appraisal", *Statistical science,* vol. 9, no. 1, pp. 55-76.

Gomez, J. 2017, *Introduction to the American Community Survey Public Use Microdata Sample (PUMS) Files Webinar*. Available: https://www.census.gov/programs-surveys/acs/news/events/pums-2017.html.

Goodchild, M.F. & Lam, N.S. 1980, *Areal interpolation: a variant of the traditional spatial problem*, Department of Geography, University of Western Ontario London, ON, Canada.

Grimm, V., Berger, U., Bastiansen, F., Eliassen, S., Ginot, V., Giske, J., Goss-Custard, J., Grand, T., Heinz, S.K. & Huse, G. 2006, "A standard protocol for describing individual-based and agent-based models", *Ecological Modelling*, vol. 198, no. 1-2, pp. 115-126.

Kramer, A.M., Pulliam, J.T., Alexander, L.W., Park, A.W., Rohani, P. & Drake, J.M. 2016, "Spatial spread of the West Africa Ebola epidemic", *Royal Society open science*, vol. 3, no. 8, pp. 160294.

Lallemant, D., Soden, R., Rubinyi, S., Loos, S., Barns, K. & Bhattacharjee, G. 2017, "Post-disaster damage assessments as catalysts for recovery: A look at assessments conducted in the wake of the 2015 Gorkha, Nepal, earthquake", *Earthquake Spectra*, vol. 33, no. S1, pp. S435-S451.

Lam, N.S. 1983, "Spatial interpolation methods: a review", *The American Cartographer*, vol. 10, no. 2, pp. 129-150.

Lovelace, R. & Dumont, M. 2016, *Spatial microsimulation with R*, Chapman and Hall/CRC.

Münnich, R. & Schürle, J. 2003, "On the simulation of complex universes in the case of applying the German Microcensus".

Nieves, J.J., Stevens, F.R., Gaughan, A.E., Linard, C., Sorichetta, A., Hornby, G., Patel, N.N. and Tatem, A.J., 2017. Examining the correlates and drivers of human population distributions across low-and middle-income countries. *Journal of the Royal Society interface*, *14*(137), p.20170401.

Olson, K.L., Grannis, S.J. & Mandl, K.D. 2006, "Privacy protection versus cluster detection in spatial epidemiology", *American Journal of Public Health*, vol. 96, no. 11, pp. 2002-2008.

Paiva, T., Chakraborty, A., Reiter, J. & Gelfand, A. 2014, "Imputation of confidential data sets with spatial locations using disease mapping models", *Statistics in medicine*, vol. 33, no. 11, pp. 1928-1945.

Smajgl, A. & Bohensky, E. 2013, "Behaviour and space in agent-based modelling: poverty patterns in East Kalimantan, Indonesia", *Environmental modelling & software*, vol. 45, pp. 8-14.

Sun, Z. & Müller, D. 2013, "A framework for modeling payments for ecosystem services with agent-based models, Bayesian belief networks and opinion dynamics models", *Environmental modelling & software*, vol. 45, pp. 15-28.

Templ, M., Meindl, B., Kowarik, A. & Dupriez, O. 2017, "Simulation of synthetic complex data: The R package simPop", *Journal of Statistical Software*, vol. 79, no. 10, pp. 1-38.

Thomson, D. R., Kools, L., & Jochem, W. C. (2018). Linking synthetic populations to household geolocations: a demonstration in Namibia. *Data*, 3(3), 30.

Tiecke, T.G., Liu, X., Zhang, A., Gros, A., Li, N., Yetman, G., Kilic, T., Murray, S., Blankespoor, B. & Prydz, E.B. 2017, "Mapping the world population one building at a time", *arXiv preprint arXiv:1712.05839*.

Tom-Aba, D., Olaleye, A., Olayinka, A.T., Nguku, P., Waziri, N., Adewuyi, P., Adeoye, O., Oladele, S., Adeseye, A. & Oguntimehin, O. 2015, "Innovative technological approach to Ebola virus disease outbreak response in Nigeria using the open data kit and form hub technology", *PloS one*, vol. 10, no. 6, pp. e0131000.

Torrens, P.M. & Benenson, I. 2005, "Geographic automata systems", *International Journal of Geographical Information Science*, vol. 19, no. 4, pp. 385-412.

Tracy, M., Cerdá, M. & Keyes, K.M. 2018, "Agent-based modeling in public health: current applications and future directions", *Annual Review of Public Health*, vol. 39, pp. 77-94.

UN DESA. (2017). World population prospects 2017.

Utazi, C. E., Wagai, J., Pannell, O., Cutts, F. T., Rhoda, D. A., Ferrari, M. J., . . . Adeniran, A. (2020). Geospatial variation in measles vaccine coverage through routine and campaign strategies in nigeria: Analysis of recent household surveys. *Vaccine*,

Voellmy, A., Vrtic, M., Raney, B., Nagel, K. & Axhausen, K.W. 2002, "Status of TRANSIMS implementation for Switzerland", *Arbeitsberichte Verkehrs-und Raumplanung*, vol. 109.

Walker, A.J. 1977, "An efficient method for generating discrete random variables with general distributions", *ACM Transactions on Mathematical Software (TOMS)*, vol. 3, no. 3, pp. 253-256.

Whitworth, A. 2013, "Evaluations and improvements in small area estimation methodologies".

Whitworth, A., Carter, E., Ballas, D. & Moon, G. 2017, "Estimating uncertainty in spatial microsimulation approaches to small area estimation: A new approach to solving an old problem", *Computers, Environment and Urban Systems*, vol. 63, pp. 50-57.

Winsemius, H.C., Van Beek, L., Jongman, B., Ward, P.J. & Bouwman, A. 2013, "A framework for global river flood risk assessments".

World Bank 2010, "Poverty Maps of Bangladesh 2010".

WorldPop 2017, *Bangladesh 100m Population, Version 2*. Available: https://www.worldpop.org/geodata/summary?id=94.

Wright, J.K. 1936, "A method of mapping densities of population: With Cape Cod as an example", *Geographical Review*, vol. 26, no. 1, pp. 103-110.