Validation of spatiodemographic estimates produced through data fusion of small area census records and household microdata

Amy N. Rose and Nicholas N. Nagle

Abstract

Despite the increasing availability of current national censuses, these datasets are limited by their lack of small area demographic depth. At the same time, spatial microdata that include detailed demographic information are only available for limited geographies, thus limiting the complex analysis of population subgroups within and between small areas. Techniques such as Iterative Proportional Fitting have been previously suggested as a means to generate new data with the demographic granularity of individual surveys and the spatial granularity of small area tabulations of censuses and surveys. This article explores internal and external validation approaches for synthetic, small area, household- and individual-level microdata using a case study for Bangladesh. Using data from the Bangladesh Census 2011 and the Demographic and Health Survey, we produce estimates of infant mortality rate and other household attributes for small areas using a variation of an iterative proportional fitting method called P-MEDM. We conduct an internal validation to determine: whether the model accurately recreates the spatial variation of the input data, how each of the variables performed overall, and how the estimates compare to the published population totals. We conduct an external validation by comparing the estimates with indicators from the 2009 Multiple Indicator Cluster Survey (MICS) for Bangladesh to benchmark how well the estimates compared to a known dataset which was not used in the original model. The results indicate that the estimation process is viable for regions that are better represented in the microdata sample, but also revealed the possibility of strong overfitting in sparsely sampled sub-populations.

1. Introduction

Demographic information from censuses and surveys are used to support a wide range of decisions for public and private planning. For example, knowledge of the characteristics of a population in an area is critical to determine the need and feasibility of new programs including schools or community centers. Furthermore, changes in the size, distribution, and composition of a population will directly impact future planning of housing and infrastructure such as roads, water supply, and energy.

Users must choose between using publicly available tabulations from large scale, national censuses and surveys, or collecting individual-level data from custom surveys. National censuses and surveys offer a large sample size, and tabulations of relatively small areas, such as neighborhoods or communities, are often publicly available. Such small area estimates are important for understanding local variations in the distribution of population. Unfortunately, these tabulations may not contain the variables that are most relevant to a particular use, nor do they provide individual- and household-level detail that is necessary to understand human behaviors. In contrast, users may construct custom surveys to collect information about the relevant variables and to understand individual- and household-level behaviors. It is usually too expensive however, to construct surveys with a large enough sample size to understand small area variations.

Synthetic spatial microdata can be developed to fuse together information from census tabulations and individual survey microdata. Synthetic spatial microdata are unit record data that represent individuals or households at a small area level, and thus the methods to generate these data are part of the broader category of small area estimation techniques. The importance of the development of synthetic spatial microdata is two-fold: they allow for analysis of estimates of variables that are not available at a small area level, while simultaneously eliminating confidentiality concerns that are typical when dealing with microdata that reflects personal data. Furthermore, generating synthetic microdata is a way to create cross-tabulations that do not already exist in summary statistics.

Despite the existence of techniques to create such synthetic spatial microdata, the difficulty of validating their outputs limits their potential for use. Model outputs are useless to researchers, planners, and policymakers if those outputs are not reasonable representations of

Please cite this article as: Rose, A.N., & Nagle, N.N., Validation of spatiodemographic estimates produced through data fusion of small area census records and household microdata, Computers, Environment and Urban Systems (2016), http://dx.doi.org/10.1016/j.compenvurbsys.2016.07.006
the real world. Recent literature dealing with synthetic microdata highlights that validation of these data is still a shortcoming (Ballas & Clarke, 2001; Birkin, 2013; Edwards & Tanton, 2013; Morrissey & O’Donoghue, 2013; Ruther, Maclaurin, Leyk, Buttenfield, & Nagle, 2013; Williamson, Birkin, & Rees, 1998). The lack of finer spatial and demographic detail in census data is one of the primary motivators for creating the synthetic microdata in the first place, but is also the reason why validation is a difficult problem. There are rarely confirmatory data by which to validate against.

Simply describing the estimating method and reporting the inputs and outputs of the model are not good enough. Rigorous interrogation of the results must be attempted as to give the community of practice some confidence that the estimates are reliable. Voas and Williamson (2001) provide an excellent discussion of the many ways to test the fit of synthetic microdata estimates. Their point, which should be well taken by the larger community is that there is not one “best” method for measuring fit, but rather a give and take with regard to a variety of criteria including validity, ease of calculation, a known distribution, and familiarity to the user community.

There are two chief ways to approach the validation of small area estimation results. In internal validation, some of the input data are withheld from the model, and reserved from comparison with the outputs. In reality, these data would not be withheld, and the concern with this approach is that the errors of estimation may be different when these data are withheld versus when they are included. In external validation, the modeled estimates are compared to a data source that was not used in the model. In many cases, depending on the model and available data, it’s only possible to perform internal validation. However, attempts should be made to also externally validate modeled estimates if possible. This study examines methods by which to perform both internal and external validation, and considers issues associated with these validation measures, both in a general sense and specific to our case study. In the study we develop new microdata estimates for infant mortality at the District level, which currently do not exist. We do this using household and population characteristics from the 2011 Bangladesh Census as margins for which to scale data from the Bangladesh Demographic and Health Survey (DHS).

The remainder of this paper is structured as follows. Section 2 provides a brief background on techniques used for producing new small area estimates. Section 3 covers the motivation for and process of building the model for Bangladesh, including the selection of constraint variables. Model fitting and output will be described in Section 4, which will serve as a preface for the discussion of internal and external validation of the new microdata estimates in Section 5. Section 6 will conclude the paper with observations about the study and potential future work.

2. Background

Synthetic small area microdata are often calculated by methods that reweight a survey so as to reproduce known, aggregate data for small areas for which it was not designed to be representative. In essence, this modeling approach combines individual or household-level microdata for large spatial areas with spatially disaggregate data in order to create synthetic microdata estimates for small areas (Harding, Lloyd, Bill, & King, 2004; Taylor, Harding, Lloyd, & Blake, 2004).

A variety of techniques have been used to produce small area estimates and demographic characterizations in cases where this information was not collected as part of the national census, was collected but not reported due to privacy concerns, or was not available as cross-tabulations (Beckman, Baggerly, & McKay, 1996; Simpson & Tranner, 2005; Williamson et al., 1998; Wong, 1992). Of these, iterative proportional fitting (IPF) approaches have a long history of use, addressing a variety of issues including: voting behavior (Johnston & Pattie, 1993), individual travel patterns (Beckman et al., 1996), rural policy analysis (Ballas, Clarke, & Wiemers, 2006; Birkin & Clarke, 1988), and small area estimation (Leyk, Nagle, and Buttenfield, 2013; Simpson & Tranner, 2005; Wong, 1992).

2.1. 2.1 Iterative Proportional Fitting (IPF)

The Iterative Proportional Fitting (IPF) method is a well-established algorithm for aligning survey data with aggregate totals. IPF requires two datasets: one is an individual- or household-level microdataset, and the other is a dataset of known population subtotals or aggregates. Intuitively, IPF identifies weights for the microdataset so that the microdataset will be redistributed to the known totals. IPF works by iteratively adjusting an n-dimensional array until every dimension converges on the known margins. IPF can be viewed simultaneously as a mathematical scaling procedure (Deming & Stephan, 1940; Norman, 1999) as well as a procedure for creating disaggregated spatial data from spatially aggregated data (Wong, 1992). Birkin and Clarke (1988) provide an early demonstration of the utility of the IPF method in geographical research, and it is often used to overcome the lack of spatial or demographic detail in source data (Ballas, Clarke, & Turton, 1999, p. 23). IPF has been used to simulate entire national scale populations (Ballas et al., 2005), examine voting patterns (Johnston & Pattie, 1993), and to create synthetic populations in order to model the travel behavior of individuals (Beckman et al., 1996).

Wong (1992) tested the reliability of IPF results by taking a subset of his population data, treating it as the actual population, and drawing random samples from this subset. These samples were then fitted by the IPF procedure to produce population estimates. These estimates were then compared to the subset distribution and any discrepancies were attributed to random error effect. Through this process, Wong determined the method did in fact produce reliable estimates but could be improved through increased sample size. In the same paper, he argued for more extensive use of IPF in geographical research, particularly in light of studies (Fotheringham & Wong, 1991; Openshaw, 1984) that demonstrated that using areal unit data for drawing statistical inference is not justified considering the effects of the Modifiable Areal Unit Problem (MAUP).

Variations of IPF have been used in several contexts, and as clarified by Johnston and Pattie (1993), not always under the formal name of IPF. Specifically, early geographical work under the label of entropy maximizing procedures was done in the context of location-allocation (Wilson, 1971) and conducted to evaluate voting behavior (Johnston & Hay, 1983, 1984; Johnston, Hay, & Rumley, 1983, 1984; Johnston & Pattie, 1993), and small area estimation (Johnston & Pattie, 1993; Leyk, Buttenfield, & Nagle, 2013; Nagle, Buttenfield, Leyk, & Spielman, 2014; Ruther et al., 2013).

2.2. Penalized maximum entropy model (P-MEDM)

Recent work (Nagle, Buttenfield, Leyk, & Spielman, 2012; Nagle et al., 2014) formalized a penalized entropy maximizing approach geared toward small area estimation and particularly dasymetric mapping. Traditional maximum entropy approaches solve the model: \[ \text{max} - \sum_i \left( \frac{x_i}{d_i} \right) \log(w_i/d_i) \] subject to the constraints that the data reaggregate to the known margins, i.e., \[ \sum_{i \in A} w_i = \text{Pop}_A \] where \( w \) are the weights to be determined by IPF, \( d \) are prior survey weights and \( \text{Pop}_A \) are the known, marginal population totals. The IPF procedure estimates new weights so that the survey estimates are now consistent with the known population totals. The P-MEDM adjusts that maximum entropy to account for uncertainty in the population margins, and consequently, reduces overfitting problems that commonly plague IPF applications in sparse data problems. Furthermore, by accounting for the uncertainty throughout the model, a measure of quality can be produced for the final population estimates. The penalized maximum entropy model (P-MEDM) as
defined by Nagle et al. (2014)

\[
\max - \sum_{i} n \frac{w_i}{N} \log \left( \frac{w_i}{d_i} \right) - \sum_{i} \frac{e_i^2}{2\sigma_i^2}
\]

subject to relaxed population constraints

\[
\sum_{i} w_i = \text{Pop}_k + e_k
\]

for each constraint \(k\), where \(n\) is the survey sample size, \(N\) is the population size, and \(d_i\) is a prior estimate of the population \(w\) for sample \(i\) in area \(t\).

No assumptions about the membership of sample records to geographic areas are made, and instead the model relies only on the constraints for \(\sigma^2\). The pycnophylactic constraints (Tobler, 1979) in the P-MEDM are relaxed in order to account for the error between the true and estimated populations. The uncertainty associated with the constraints is explicitly defined in the model as part of the penalty factor \(\sum_k \frac{e_i^2}{2\sigma_i^2}\). Therefore, if the P-MEDM output exactly fits a population constraint, then the error will be zero; conversely if the constraint is not fit exactly, there will be an estimated error \(e_i\) effectively penalizing the maximum entropy solution.

In contrast with traditional IPF approaches, the P-MEDM approach requires knowledge of the variance \(\sigma^2\) of the constraining margins, and also allows the output margins to deviate from the input margins by an error \(e_i\). This allowable error will be smaller when the input margins are more precise. An advantage of this approach in the small area situation is that it is less prone to overfitting in sparse data situations such as those that commonly occur in small area estimation.

3. Bangladesh case study

Bangladesh, like other developing nations, faces a number of population challenges including increasing urbanization, adolescent population growth, maternal mortality and morbidity, and HIV/AIDS as an epidemic (CPD, 2003). These issues require a decidedly demographic lens by which to understand the population, which makes Bangladesh an important geographic area to examine for this research (Fig. 1).

The population of Bangladesh in 2011 was approximately 150 million persons. The administrative geography of Bangladesh in 2011 comprised seven administrative divisions, which were subdivided into 64 districts, which ranged in size from approximately 380,000 persons (Bandarban) to 11,800,000 persons (Dhaka). Our individual- and household-level data come from the 2011 Bangladesh Demographic and Health Survey (DHS 2011). The Demographic and Health Survey (DHS) is a nationally representative sample survey designed to provide information on basic national indicators of social progress including fertility, childhood mortality, contraceptive knowledge and use, maternal and child health, nutritional status of mothers and children, awareness of AIDS, and domestic violence. As part of the DHS program, nearly 300 surveys for over 90 countries have been performed since 1984 (Measure DHS, http://www.measuredhs.com/). The DHS Program is authorized to distribute, at no cost, unrestricted survey data files for legitimate academic research (http://www.dhsprogram.com/).

The Bangladesh DHS is a two-stage, stratified cluster sample, containing about 18,000 households, that is designed to be representative for the country as a whole, for urban and rural areas separately, and for each of Bangladesh’s seven administrative divisions. The DHS 2011 survey is not designed to be representative of the 64 smaller districts of Bangladesh however, and it is for these smaller districts that we are most interested in developing detailed demographic estimates. We note that the DHS 2011 survey data can be obtained with geocodes, however, this geographic detail does not help to produce estimates that are representative of 64 smaller, unplanned-for administrative districts.

3.2. Multiple Indicator Cluster Survey (MICS)

For the external validation discussed in Section 5, we will be using Multiple Indicator Cluster Survey (MICS) data for Bangladesh. The MICS is a household survey developed by UNICEF in the mid-1990’s to assist countries in filling data gaps for monitoring the situation of children and women. The most recent MICS in Bangladesh was conducted in 2009. The Bangladesh MICS was designed to provide estimates on indicators on the condition of children and women for a variety of geographic aggregations including urban and rural areas, at the national, district, and sub-district levels. Sub-districts were used as the primary sampling domains, and sample weights were used for reporting national and district level results. The planned sample for the MICS was 300,000 households of which 299,842 were interviewed successfully for a household response rate of 99.9% (BBS and UNICEF, 2010).

3.3. The census and the constraining variables

Producing synthetic data with IPF or P-MEDM requires marginal tabulations with which to constrain the survey data. We use district-level tabulations from the 2011 Bangladesh Census as constraints on the synthetic data estimation. There are many possible tabulations that are available for use, however, and we must choose a subset for consideration. In this section, we review the theoretical considerations to guide this section, and then discuss the empirical validation of our chosen constraints.

3.3.1. Theoretical basis for choosing constraint variables

The most obvious consideration in selecting variables is that the variable needs to appear in both the survey data and the census. It would be impossible to determine how to adjust the survey weights if the variable...
is in only one data set and not the other. The more substantive theoretical considerations are that 1) the variable should be correlated with the specific attribute of interest, and relatedly, 2) that there should be some spatial variation within the constraining variable. It is important to select constraint variables that are as closely correlated as possible with the purposes of the new microdata (Chin & Harding, 2006), and thus the choice of these constraint variables was dependent on the presence of a reasonably good correlation between the constraint variables and the variable that is ultimately being mapped. Ideally, these constraining variables should represent the underlying spatial heterogeneity of population characteristics (Rutherford et al., 2013; Simpson & Tranner, 2005).

The ability to produce reliable estimates that recreate the spatial allocation of the input data is a clear requirement when choosing constraining variables, but consideration must also be given to the role variables will play in the validation process. It can be assumed that constraint variables, as well as any non-constraint variables that are highly correlated with the constraint variables, will be reasonably estimated. However, since the variable(s) of interest – those that do not already exist but we seek to estimate – are not used in the model, we can only assume that they will be reliably estimated if they are strongly correlated with the input variables.

3.3.2. Empirical validation of constraint variables

For this exploratory research, a single household-level attribute of interest was selected from the DHS data to test the feasibility of developing complete microdata. The model focused on child health, specifically looking at infant mortality. Previous literature suggests that infant mortality in Bangladesh is strongly correlated with other aspects of child health, all of which are highly correlated with the education level of the mother and overall household wealth and socioeconomic status (DHS, 2013, p. 115). The corresponding fields in the DHS and Census data that were matched to create the constraints are shown in Table 1.

Literacy is widely acknowledged to benefit the individual and society and is associated with a number of positive outcomes for health and nutrition, particularly for women (DHS, 2013, p. 34). In the DHS findings, literacy is highly correlated with age, and varies notably by both division and urban/rural designation. However, literacy was comparable for both men and women. Therefore, the use of female literacy combined with other education indicators should be a reasonable proxy for the overall education level of the mother.

The household recode provided data on household size, water source, electricity, and housing tenure. National-level summaries of these variables from the DHS2011 are listed in Table 2.

In order to be informative, the constraint data must display spatial heterogeneity, that is, that the constraint values must differ from area to area. The only spatial indicators in the DHS microdata are Division and region. The only spatial indicators in the DHS microdata are Division and region.

4. Model fitting and output

The inputs for the computational model are prior survey weights and a set of control totals or tables for selected attributes and geographic regions, as described in the previous section. We identified attributes that could be matched between the DHS and the Bangladesh Census, and selected those attributes that were correlated with infant mortality rate at the household level. These matchable attributes and their sample counts from the DHS are listed in Table 2. The Bangladesh Census published tables of these attributes at the District level and stratified by urban/rural, and we selected these tables as the control tables. We then reserved (omitted) female literacy and water source from the set of control tables. While these variables are correlated with infant mortality rate, and are thus potentially good predictors for small area estimation of infant mortality, we omitted them so that they were available for use in later validation.

We fit the model separately for each Division (administrative level 1), and produced estimates for each District within (administrative level 2). Example model outputs are shown in Table 3. For each household record, we obtain a sample weight representing the estimated number of instances of that household in each of the Districts within the Division. These weights are the results of the computational model trying to match the control totals, while not deviating too much from the prior survey weights. Using these weights, we can use the original DHS to obtain any desired District-level quantity simply by aggregating the DHS using the District-specific weights. For example, even though infant mortality rate is not estimated at the District-level by the Bangladesh Census, we can use this method to produce District-level estimates from DHS (combined with other Census Data through the P-MEDM model).

4.1. Comparison of P-MEDM estimates to census

To fundamentally assess whether the P-MEDM procedure produced reasonable estimates, non-constraining variable estimates were mapped and compared to the same variables taken from the Census name and a rural/urban flag, and in combination only produce 14 unique spatial variations of the data. All of these variables were determined to have spatial variation in the Census Data, as demonstrated in Fig. 2, and are thus candidates for effective constraints.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Constraining variables chosen from the NIPORT, Mitra and Associates, and ICF International (2013) Census data.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indicator</td>
<td>DHS 2011</td>
</tr>
<tr>
<td>Education</td>
<td>Recode of age/sex/in school</td>
</tr>
<tr>
<td>Socioeconomic</td>
<td>Recode of age/sex/in school</td>
</tr>
<tr>
<td>Socioeconomic</td>
<td>Recode of occupation: women</td>
</tr>
<tr>
<td>Socioeconomic</td>
<td>Recode of source of drinking water</td>
</tr>
<tr>
<td>Demographic</td>
<td>Has electricity</td>
</tr>
<tr>
<td>Demographic</td>
<td>Owns household</td>
</tr>
<tr>
<td>Spatial</td>
<td>Calculated mean size of household</td>
</tr>
<tr>
<td>Spatial</td>
<td>Type of residence</td>
</tr>
<tr>
<td>Spatial</td>
<td>Region of residence</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Sample size of Bangladesh DHS 2011 data with breakdown of variable categories.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Household members</td>
<td>Male</td>
</tr>
<tr>
<td>Household members</td>
<td>Female</td>
</tr>
<tr>
<td>Females in school</td>
<td>Under 15</td>
</tr>
<tr>
<td>Females in school</td>
<td>Age 15+</td>
</tr>
<tr>
<td>Literate females</td>
<td>11,568</td>
</tr>
<tr>
<td>Employed females</td>
<td>Field</td>
</tr>
<tr>
<td>Age of householders</td>
<td>Industry</td>
</tr>
<tr>
<td>Age of householders</td>
<td>Service</td>
</tr>
<tr>
<td>Age of householders</td>
<td>Age 0–4</td>
</tr>
<tr>
<td>Age of householders</td>
<td>Age 5–9</td>
</tr>
<tr>
<td>Age of householders</td>
<td>Age 10–14</td>
</tr>
<tr>
<td>Age of householders</td>
<td>Age 15–19</td>
</tr>
<tr>
<td>Age of householders</td>
<td>Age 20–24</td>
</tr>
<tr>
<td>Age of householders</td>
<td>Age 25–29</td>
</tr>
<tr>
<td>Age of householders</td>
<td>Age 30–39</td>
</tr>
<tr>
<td>Age of householders</td>
<td>Age 40–49</td>
</tr>
<tr>
<td>Age of householders</td>
<td>Age 50–59</td>
</tr>
<tr>
<td>Age of householders</td>
<td>Age 60+</td>
</tr>
</tbody>
</table>

Please cite this article as: Rose, A.N., & Nagle, N.N., Validation of spatiodemographic estimates produced through data fusion of small area census records and household microdata, Computers, Environment and Urban Systems (2016), http://dx.doi.org/10.1016/j.compenvurbsys.2016.07.006
2011 data. Both female literacy and tap water source estimates were comparable to those from the Census data at the District level. Although urban designation was used as a constraining variable, spatial data defining rural and urban boundaries are not available and thus, mapping results at that finer level of spatial detail was not possible.

As shown in Fig. 3, the spatial allocation of the P-MEDM estimates for female literacy were comparable to the Census data totals. The chart in Fig. 4 breaks down these differences numerically showing very minimal differences for all Districts. Similar results for tap water source are shown in Figs. 5 and 6. These results show that not only were the estimates very close numerically to the actual Census data, but also confirms that the spatial heterogeneity present in the Census data was in fact recreated through the P-MEDM process.

5. Validation of new microdata estimates

In this section, we present the main findings of the paper, describing a variety of internal and external validation tests we were able to perform on the model estimates. In this context, we describe internal validation as those comparisons between the model output and other estimates from the DHS and Bangladesh Census. While the DHS and the Bangladesh Census were used in the model, our internal validation is conducted by comparing the model outputs with information from these data that was not directly used in fitting the model. Yet, there is a possibility that these comparisons are not truly independent, as we can expect a strong degree of correlation among any data from the same source. Thus, we also report on external validation results, in which we compare our estimates of IMR with estimates produced from the Multiple Indicator Cluster Survey (MICS), a large, independent survey that was not used in the modeling process (mics.unicef.org).

5.1. Internal validation

5.1.1. District-level comparison of P-MEDM with raw DHS estimates

In the first of the internal validations, we estimate the difference between our district-level estimates and district-level Bangladesh Census to the difference between the raw DHS district-level estimates and the district-level Bangladesh Census. The DHS data are published with district geocodes, however the DHS methodology is not designed to be representative at the district level. Nonetheless, it is technically possible to produce district-level estimates from the DHS, and we use this comparison as an initial check, to verify that the P-MEDM approach to weighting produces more reasonable results than the raw DHS weights. It may be the case that the raw DHS weights happen to be good
estimates at the district-level, even though they were not designed to be so, and thus, that P-MEDM is unnecessary. This validation allows us to evaluate this consideration.

We make this comparison based on the water source and female literacy attributes that were withheld from the model fitting procedure. These attributes are correlated with the selected control tables, but were not included as control tables. To scale the DHS estimates, the household sample weights were used to weight all household records, and then these records were scaled using a multiplier to assure the total household count per District was equal to the Census household counts for that District. As expected, the RMSE for the District level DHS estimates showed a poor fit to the Census benchmarks (Table 4). However, the P-MEDM results showed a much better fit as interpreted by the RMSE. The P-MEDM estimates have an RMSE that is typically 1000 times smaller than that of the raw DHS data.

5.1.2. Validating the spatial distribution of the estimates

Internal validation should help determine whether the model accurately recreated the spatial variation of the input data, how each of the variables performed overall, and how the estimates compare to the initial margins – particularly for the non-constraining variables. Internal validation of the P-MEDM estimates was conducted by calculating the Standardized Allocation Error (SAE) in order to compare the P-MEDM estimates to Census tables for small areas.

The standardized allocation error (SAE) (Anderson, 2013; Ballas et al., 1999, 2005; Ruther et al., 2013; Williamson et al., 1998) was used to compare the model allocations to the actual census counts at multiple geographic aggregations. The SAE can be used to evaluate how well each variable was allocated over multiple geographic aggregations $GA$:

$$\text{SAE} = \frac{\sum (P_i - C_i)}{\sum C_i}$$

where $P_i$ is the P-MEDM population estimate for area $GA_i$ and $C_i$ is the census population for area $GA_i$. The result is a positive or negative value that can be intuitively interpreted as an underestimation or overestimation. Previous studies, however, do not clearly indicate what the acceptable bounds of SAE values should be. Some studies suggest an absolute SAE less than 20% may be appropriate, although the range should depend on the data (Ballas et al., 1999; Ruther et al., 2013; Smith, Clarke, & Harland, 2009). Since the SAE for constraint variables should approach zero, here we look at the SAE only for non-constraining variables.

The SAE was calculated at the Division level, for rural/urban breakdown, at the District level, and at the geographic level used in the P-MEDM process (District plus rural/urban breakdown). For all of these aggregations, the absolute SAE was less than 1% for the non-constraining variables (female literacy and water source). However, there were some interesting results that emerged at the District level.
These results are shown in Fig. 7 where negative values indicate under-allocation and positive values indicate over-allocation.

Looking at the results, it appears that all of the non-constraining variables were well allocated with no allocation errors over 0.25% or under −0.5%. In general, the literacy estimates are too small, indicating an underestimation of literacy in the DHS relative to the Census, and Tap Water Source estimates are too large, indicating an overestimation of this quantity in the DHS relative to the Census.

It is informative to inspect one of the “unusual” values of the P-MEDM estimates. While the literacy estimates underestimate the census estimates, there is a single District for which the fertility estimate is too high: Bandarban District in the Chittagong Division. Even though the allocation error is very small (0.0006%), it is an unusual P-MEDM estimate because it is the only female literacy estimate that is too high. Inspection of the P-MEDM output shows that female literacy was only over-allocated to the rural part of the District. This district, which is located in a very hilly area in southern Bangladesh bordering Myanmar (Burma), is the most remote and least populated district in Bangladesh. A review of the Census data shows that the rural section of Bandarban has the lowest female literacy rate of all districts at 19.8%, 6% lower than the next highest district. The over-allocation highlights an important aspect of P-MEDM: since P-MEDM draws on similar households across all Districts within a Division, it can underestimate or overestimate unusual or atypical values in the population.

5.2. External validation

External validation of a model compares the estimates with exogenous data that are considered to represent a standard for comparison. Going back to the purpose of this research, the need exists to produce estimates for variables that are not available for small areas. The fact that these variables don’t exist is exactly what makes external validation difficult. Furthermore, much like the estimates the model produces, the external data used for estimate comparisons are likely subject to sampling and non-sampling errors. Despite these issues, external validation should be attempted when possible. For our purposes, the Multiple Indicator Cluster Survey (MICS) estimates will be used as if they are the actual observed values in order to compare the P-MEDM estimates and the DHS raw estimates.

5.2.1. Data alignment

An immediate issue with the Bangladesh 2009 MICS data was the temporal mismatch with both the DHS and the Census data. During the period between when the MICS was conducted (2009) and when
the DHS and Census were conducted (2011), a new division was created from existing districts in Bangladesh. Rangpur became Bangladesh’s 7th division on January 25, 2010, and was created from the northern eight districts of the Rajshahi Division (Rangpur, Dinajpur, Kurigram, Gaibandha, Nilphamari, Panchagarh, Thakurgaon, and Lalmonirhat). Since the MICS was conducted prior to this, geographic area coding still reflects the old divisions and districts. To account for this, these eight districts were recoded as part of the Rangpur district during processing of the MICS data prior to using it for validation.

5.2.2. Variable of interest

Ideally, as part of the external validation, the original variable of interest – infant mortality – should be compared. For this validation, infant mortality rate (IMR) was calculated as the number of deaths per 1000 live births for ages 0–11 months. The IMR was calculated using the synthetic cohort method employed by DHS for survey final reports (Rutstein & Rojas, 2006, 90–94).

To calculate the infant mortality rate, the new P-MEDM household weights were used to replicate birth records by household from the DHS birth recode table. Since the reweighting was done at the Division level, households in any given District within that Division could be used to produce new weights for any other District in that Division. Once the birth records were replicated, IMR was calculated using births and deaths occurring two to seven years prior to the DHS survey.

Table 4
Root mean squared error (RMSE) for non-constraining variables for P-MEDM estimates and DHS estimates as compared to 2011 Census values.

<table>
<thead>
<tr>
<th>Variable</th>
<th>RMSE P-MEDM estimates</th>
<th>RMSE DHS estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female literacy</td>
<td>0.047</td>
<td>10.916</td>
</tr>
<tr>
<td>Tap water</td>
<td>0.002</td>
<td>4.559</td>
</tr>
<tr>
<td>Tubewell water</td>
<td>0.015</td>
<td>11.934</td>
</tr>
<tr>
<td>Other water</td>
<td>0.004</td>
<td>4.106</td>
</tr>
</tbody>
</table>

Fig. 7. Standard allocation error at the District level for each of the non-constraining variables used in the P-MEDM. Division groupings are shown in varying colors.
date. This time period was chosen because of the temporal alignment with the MICS survey date and the reference period used in the MICS IMR calculations. In addition to calculating IMR using the new estimates, IMR was also calculated based on the raw DHS estimates using the same reference period.

The infant mortality rates that have been published for the Bangladesh 2009 MICS were derived using the indirect (Brass) method. As stated in the UNICEF (2010) report (p. 14):

“The most robust values of that method are numbers based on information from women aged 25-29 and 30-34 years, concerning the number of children born and the number who survive. That information, using a computer program (Offive) and applying model West Life tables, gives an average estimate for a reference period of five years before the survey date (2004).”

Many indirect methods of estimating fertility are based on the $P/F$ ratio method first proposed by Brass (1964), where $P$ is the average parity (cumulative lifetime fertility) of a cohort of women up to a particular age, and $F$ is a close approximation of cumulative current fertility up to that same age. In addition to Brass’ contribution to indirect methods, the synthetic cohort life table approach used by DHS, as well as other cohort-period fertility rate calculation methods, are interpreted in a similar way to the Brass method. This does not mean however that indirect methods and direct methods are comparable. The use of the Brass indirect estimation method is an immediate issue with using the MICS 2009 IMR to benchmark the P-MEDM derived IMR estimates as the two methods are not comparable in their derivation data. However, in the absence of other available comparable datasets or another suitable proxy of infant mortality, the use of the MICS IMR is regarded as the best source of exogenous data.

5.2.3. Comparison of P-MEDM and MICS IMR estimates

In the P-MEDM algorithm, a variance term is assigned to account for varying quality of the input data. The variance in this sense is the penalty term or uncertainty about the estimates, in that if the variance of the input datasets is small then the penalty on errors will be large. On the other hand, if the input datasets have a high variance, then the penalty on errors will be low. In order to perfectly replicate the original population constraints for Bangladesh, the variance would be set to zero. However, this is unrealistic since the constraints themselves are imprecise. If we try too hard to fit data that are not precisely known, the result could be overfitting the model. This is the benefit of using the P-MEDM algorithm, as these constraints can be relaxed. Typically the variance term can be obtained from published margins of error for input datasets. However, in this case, there is no published margin of error for the Bangladesh Census. Therefore, the variance of each constraint was taken to be $N^p P^*(1 - P)$, where $N$ is the number of households in a division, and $P$ is the proportion of households in each constraint, i.e., $P = constraint \ total / household \ count$. For the estimates with smoothing, the variance was increased by a factor of $10^2$ and $20^2$ (i.e. the standard deviation was increased by 10 and 20).

Plotting the results against the MICS IMR (Fig. 8) there are three very clear outliers in the Chittagong District, as well as a large overestimation of the IMR for all the districts in the Rajshahi Division. As discussed in the previous section where P-MEDM estimates were compared to raw DHS estimates, there were ten urban subsections of districts that were not sampled in the DHS data. Three of these districts: Bandarban, KahAGRchhari, and Rangamati; are the three Chittagong Division areas that show up in Fig. 8 as outliers when compared to the MICS estimates. Interestingly, these are the three districts that make up what is known as the Chittagong Hill Tracts in southeastern Bangladesh. The Chittagong Hill Tracts vary considerably in demographic makeup from the rest of Bangladesh in that it is home to eleven indigenous groups rather than Bengalis that populate most of the rest of Bangladesh (IWGIA, 2014).

For these initial results, the variance for the model run solution shown in Fig. 8 was set to 5, which may be too restrictive and does not account for the true uncertainty inherent in the Bangladesh Census. To further investigate the fit of these data, two additional P-MEDM runs were performed; one with a variance term of 10 and one with a variance term set to 20. The results are shown in Fig. 9. Here, the outliers are still apparent, but the overall fit improves as the variance is increased.

The effect that the outliers have on the overall fit of the P-MEDM estimates to the MICS estimates are quantified in Table 5. The root mean squared error (RMSE) was calculated for each set of estimates; DHS, P-MEDM ($\sigma^2 = 5$), P-MEDM ($\sigma^2 = 10$), P-MEDM ($\sigma^2 = 20$); as compared to MICS. As shown here, the P-MEDM estimates tend to be more stable than the raw DHS estimates, but are tenuous in the worst case. It is important to note that the P-MEDM estimates are also less variable within a Division than both the MICS and the raw DHS estimates, since the P-MEDM allows a household from the survey data to be used as a representative household for a District other than where it was sampled from, provided that it is still used within the proper Division.

Although the scenario where estimates were produced using a higher variance term seems to yield a better fit, it’s still unknown whether the value chosen was the best one. Increasing the variance in the model appears to stabilize the estimates by better accounting for the uncertainty in the constraints. However, in the absence of a published margin of error by which to choose the variance, more scenarios may be useful to quantify how the model behavior changes with changes to the variance. It may also be useful to adjust the variance term for each constraint since the size of each population group may change. For example, when evaluating data for household water source, the target population is all households for a given geographic area. However, when evaluating female literacy, the target population is now only the female population of a certain age. In the case of infant mortality, the target population can become even smaller since only women who have had children are being considered. In these cases, it may be prudent to evaluate variance by the acceptable margin of error for each household or individual level characteristic.

6. Conclusion

In this study, we have conducted internal and external validation of a method for producing small area estimates from a sample survey. While the internal validation results are encouraging, the external validation results are mixed. A large problem with the external validation here arises from the difficulty in comparing infant mortality between two very different surveys, and concerns about the accuracy of both data. Nonetheless, from the results published here, we conclude, that it is possible to produce accurate small area estimates, under the important provision that the analyst understands the data and sampling method and can recognize where the method is likely to fail. In particular, the method is likely to fail when producing any estimate of a small or rare population that is not adequately sampled. No amount of reweighting can correct for these problems. When used for downscaling adequately sampled subgroups, the small area estimation methods appear to work well. When used to downscale poorly sampled groups, or to produce estimates for “unusual” geographic places, the downsampling methods can fail – and fail dramatically. Fortunately, we think that the practitioner who is knowledgeable in the survey design may be able to identify these problematic instances before performing such an analysis. We also note that the method we chose (P-MEDM) was chosen because of its reported robustness to uncertainty and many constraints. We expect that the problems mentioned here may be more extreme when other methods are used such as IPF.
The results must be caveated with the reality that the DHS data are a small sample relative to the actual population and household count of Bangladesh. It’s important to note this, particularly since many of the variables used in the model were taken only from the female portion of the sample. Bangladesh is one of the top ten most populated countries in the world, with the 2011 recording approximately 144 million individuals and 32 million households. Comparatively, the DHS sample size is 83,731 individuals (51% female) and 17,141 households. This translates to the DHS being a less than 0.1% sample of individuals and households. Even so, one of the main points of this research is to begin to develop methods in the absence of data rich environments that can provide a more detailed demographic look at an entire country. Using the P-MEDM procedure allows reasonable estimates to be produced even for areas where sample data are sparse or non-existent by associating that area with similar areas where the sample is large enough. Estimates can be produced at multiple geographic levels to attenuate the effect of small sample sizes. Ultimately, the new microdata generated using P-MEDM will still reflect the quality of the original survey microdata.

Both internal validation and external validation must be undertaken not simply to test the results of this case study, but also to contribute to an understudied, but critical body of knowledge required to inform future models. There have been a limited number of validation studies performed with regard to spatial microsimulation models in general,

Please cite this article as: Rose, A.N., & Nagle, N.N., Validation of spatiodemographic estimates produced through data fusion of small area census records and household microdata, Computers, Environment and Urban Systems (2016), http://dx.doi.org/10.1016/j.compenvurbsys.2016.07.006
and consequently even fewer specific to entropy maximizing procedures. One reason for this is the limited opportunities for validation of estimates due to the lack of available external data for comparison. This lack of confirmatory validation continues to be a limitation for using these methods in research (Edwards & Tanton, 2013; Ruther et al., 2013; Williamson et al., 1998). In this sense, this exploratory work can guide future research, which must include reasonable approaches for validation, particularly in data poor environments. Validating the new household and individual estimates produced for Bangladesh was an essential but non-trivial exercise, and the results underscore the importance of both internal and external validation. Although the internal validation performed well, the external validation uncovered a potential overfitting of the model due to a variance term that likely was not taking into account the appropriate level of uncertainty in the Bangladesh Census data.

An important component of the P-MEDM algorithm is the relaxing of the pycnophylactic constraints in order to account for uncertainty in the population estimates that are used as constraints. For the Bangladesh example, even though population counts directly from the Bangladesh Census were used as constraints in the model, these constraints are not precise, but are actually estimates themselves. Therefore, trying to preserve these population totals across joint distributions must be approached from the perspective of fitting without overfitting. Since the P-MEDM relaxes constraints, unusual household characteristics, as well as small numbers of households for an area won’t necessarily make convergence difficult, but it could skew the new household weights. This is where the trade-off between uncertainty and biased estimates is most evident. In the Bangladesh case study presented here, a small sample size, particularly with respect to female-only indicators and infant mortality, and areas with unordinary demographic characteristics had the effect of creating more noise in the estimates. Further restricting the strength of the validation effort is the limited availability of an exogenous dataset that can be used for comparison. This does not mean however that the results were poor, but rather that more work must be done to validate and refine the model where possible.

As discussed previously, a major limitation with the external validation is that the MICS 2009 raw data were not available at the time of this research. Although the raw data for the BBS and UNICEF (2007) were available, information on infant mortality that could be used to directly compare estimates was not included. Any error that was introduced into the validation due to the mismatch between IMR calculation methods is unknown at this time. To address this, it would be useful to perform the same external validation once the MICS 2009 raw data are available. By using the published MICS IMR rather than calculating it in the same manner as was done for the P-MEDM estimates, differences in the district level figures could stem from differences in direct and indirect methods to calculate infant mortality. Presumably the synthetic cohort life table approach, as was used to calculate IMR for the P-MEDM estimates, could be used to calculate the IMR based on the 2009 MICS data. Completing this exercise could eliminate any error associated with the difference between the two IMR calculations.

Specific to the P-MEDM process, more investigation is needed as to how to appropriately tune the variance term when none is available a priori. The research done here showed that the initial variance term used was likely too restrictive and did not account for the true uncertainty inherent in the Bangladesh Census. Additional runs of the algorithm with adjusted variance terms showed an improved fit as the variance was increased, although outliers were still present. It would be worthwhile to further investigate the sensitivity of a global variance term, but also experiment with using different variance terms for each variable. Each variable realistically has a different level of error associated with it, and in theory, assigning specific variance terms to each variable would account for this varying uncertainty.

Acknowledgments

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-000OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

References


Please cite this article as: Rose, A.N., & Nagle, N.N., Validation of spatiodemographic estimates produced through data fusion of small area census records and household microdata, Computers, Environment and Urban Systems xxx (2016) xxx–xxx

Table 5

<table>
<thead>
<tr>
<th>RMSE</th>
<th>All Districts</th>
<th>3 Districts held out</th>
<th>3 District + Rajshahi held out</th>
</tr>
</thead>
<tbody>
<tr>
<td>P-MEDM ($\sigma^2 = 5$)</td>
<td>34.62</td>
<td>24.75</td>
<td>11.72</td>
</tr>
<tr>
<td>P-MEDM ($\sigma^2 = 10$)</td>
<td>29.32</td>
<td>18.65</td>
<td>10.86</td>
</tr>
<tr>
<td>P-MEDM ($\sigma^2 = 20$)</td>
<td>20.91</td>
<td>11.44</td>
<td>9.91</td>
</tr>
<tr>
<td>DHS</td>
<td>19.46</td>
<td>19.17</td>
<td>19.02</td>
</tr>
</tbody>
</table>


Please cite this article as: Rose, A.N., & Nagle, N.N., Validation of spatiodemographic estimates produced through data fusion of small area census records and household microdata, Computers, Environment and Urban Systems (2016), http://dx.doi.org/10.1016/j.compenvurbsys.2016.07.006

Dhaka, Bangladesh and Calverton, Maryland. USA: NIPORT, Mitra and Associates, and ICF International.


