Data Fusion Methods for Improved Demographic Resolution of Population Distribution Datasets

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Amy Nicole Rose
May 2015

Copyright © 2015 by Amy N. Rose
All rights reserved.

DEDICATION

To my parents
Thomas and Rosalie Nugent

To my husband
Shane Rose

ACKNOWLEDGEMENTS

The completion of my research and my degree would not have been possible without the support of my family, friends, and colleagues. To Shane: thank you for knowing exactly how to support me each and every day and for keeping me laughing even in my lowest moments. To my family: thank you for encouraging without nagging, but mostly for being my constant. Without all of you around me, I could not have had the strength to see this through. I would be remiss if I did not take this opportunity to thank Dr. Julian Ray for taking a chance on me very early in my career. I would certainly not be where I am today if he hadn't. Also, I am forever grateful to Eddie Bright for being my mentor at Oak Ridge National Laboratory. I cannot quantify the amount of professional knowledge and life lessons that he has taught me. I would also like to thank Dr. Budhendra Bhaduri for encouraging me to continue my education and for leading by example. Numerous other people deserve mention for continuing to encourage me, listening to me agonize during the low points of this process, and celebrating the (usually brief) high points; specifically I would like to say "thank you" to Dr. Robert Stewart, Jessica Moehl, Dr. Leanne Sulewski, Marie Urban, and Dr. Harini Sridharan. My utmost appreciation goes out to my entire graduate committee: Dr. Stephanie Bohon, Dr. Shih-Lung Shaw, and Dr. Micheline van Riemsdijk for their guidance both in my coursework and my research. Finally, I would like to wholeheartedly thank my advisor Dr. Nicholas Nagle for sharing his time, intellect, and wisdom with me; and for having the patience to see me through this long and sometimes painful process. I truly feel like I won the advisor lottery, and simply could not imagine going through this program without his direction.

ABSTRACT

Despite the availability of multiple global population distribution datasets, these datasets are limited by their lack of demographic depth. Although large area spatial datasets of population distributions currently exist, similar spatial representations of other demographic and socioeconomic characteristics are scarce. Spatial microdata that include detailed demographic information are rarely available for small areas, thus limiting the complex analysis of population subgroups. To address the lack of demographic resolution in existing population distribution datasets, a first step would be to develop large area microdata that can be attached to a country- or global-level population distribution dataset. This can be achieved by reweighting a national level sample so as to estimate the detailed socioeconomic characteristics of populations and households at a small area level. In essence, this modelling 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.

Methods to build synthetic spatiodemographic microdata are well documented in literature, yet these efforts have been implemented on limited geographic extent in data rich environments. More specifically, these methods have been tailored to fit specific local, regional, or national data sources with no plan or requirement for adaptation for other geography or data. To address this gap, this research will present a generalizable method for developing synthetic spatial microdata which in turn can be used to increase the demographic resolution of global population distribution data.

TABLE OF CONTENTS

INTRODUCTION 1
Background 1
Components of the Study 2
Study Area – Bangladesh 2
Infant Mortality 3
Research Outline 4
References 6
Appendix 8
CHAPTER I THE LANDSCAN GLOBAL POPULATION DISTRIBUTION PROJECT: CURRENT STATE OF THE ART AND PROSPECTIVE INNOVATION 9
Abstract 10
Introduction 10
Background 11
Population Distribution Methods 12
Methodological Considerations 14
Input Data Effects 14
Temporal Factors 15
Uncertainty Assessments 17
Current State of the Art 17
Gridded Population of the World 17
Global Urban-Rural Mapping Project 18
AfriPop 18
LandScan 19
Data and Methods 21
Future Innovations 28
Conclusion 29
References 31
CHAPTER II DEVELOPMENT OF COMPLETE MICRODATA THROUGH DATA FUSION: BANGLADESH AS A CASE STUDY 35
Abstract 36
Introduction 36
Background 37
Iterative Proportional Fitting (IPF) 38
Penalized Maximum Entropy Model (P-MEDM) 39
Data 40
2011 Bangladesh Census 40
Demographic and Health Survey 40
Constraining Variables 41
Theoretical Basis for Choosing Constraint Variables 42
Empirical Validation of Constraint Variables 42
Results 47

Descriptive Statistics....48
Mapping the Results....49
Infant Mortality Rate....50
Conclusion....54
References....55
CHAPTER III VALIDATION OF SPATIODEMOGRAPHIC ESTIMATES PRODUCED THROUGH DATA FUSION METHODS....59
Abstract....60
Introduction....60
Background....61
Output of the P-MEDM Algorithm for Bangladesh....63
Internal Validation....64
Error in Margin....64
Standardized Allocation Error....67
Comparing P-MEDM Estimates to Scaled DHS Estimates....70
External Validation....75
Multiple Indicator Cluster Survey (MICS)....75
Data Alignment....75
Variable of Interest....75
Results....77
Conclusion....81
References....82
CONCLUSION....85
Summary....85
Discussion....86
Future Research....87
Final Thoughts....88
References....90
VITA....91

LIST OF TABLES

Table A.1. Country population rankings for mid-year 2014 for the ten most populated countries, with mid-year 2050 populations and rankings included for those same countries.....8
Table A.2. Population density rankings as of mid-year 2014 for the top ten countries with the highest densities. 8
Table 1.1. Census history of selected nations (IPC 2008). 16
Table 2.1. Constraining variables chosen from both the DHS 2011 and the Bangladesh 2011 Census data.....43
Table 2.2. Sample size of Bangladesh DHS 2011 data with breakdown of variable categories.....44
Table 2.3. Results of logistic regression with Infant Mortality Rate as the dependent variable and for records where infant mortality was greater than zero.....47
Table 2.4. Example of P-MEDM output weights for Bangladesh. 48
Table 2.5. Root mean squared error (RMSE) for non-constraining variables for P-MEDM estimates and DHS estimates as compared to 2011 Census values. 49
Table 3.1. The distribution of householders from the 2011 Bangladesh Census listed by rural and urban designation and by Division.....63
Table 3.2. National level percent absolute error in margin of P-MEDM produced estimates as compared to margin totals taken from the 2011 Bangladesh Census. Female literacy and water source variables were not used as constraints. 65
Table 3.3. Comparison of Census and P-MEDM estimates of households for the variable home ownership by rural and urban designation. Results show that the largest error is with the estimates for urban designated households. 65
Table 3.4. Root mean squared error (RMSE) for P-MEDM estimates and DHS estimates for IMR as compared to the reported IMR from the MICS 2009. .78

LIST OF FIGURES

Figure 1.1. Creating a coefficient surface using ancillary variables. 13
Figure 1.2. Using the coefficient surface to redistribute population. 13
Figure 1.3. Evolution of LandScan. 20
Figure 2.1a. Constraining variables from 2011 Bangladesh Census show the spatial variation necessary for producing new microdata estimates. 45
Figure 2.1b. Constraining variables from 2011 Bangladesh Census show spatial variation. 46
Figure 2.2. Percent female literacy from Census 2011 as compared to percent female literacy estimated by the P-MEDM procedure at the District level. Mapped results show that the estimates match well to the Census data. 51
Figure 2.3. Absolute percent difference between P-MEDM estimates and actual District level data taken from the Census 2011 data for female literacy. 51
Figure 2.4. Percent of homes with tap water source from Census 2011 as compared to percent of homes with tap water source estimated by the P-MEDM procedure at the District level. Mapped results show that the estimates match well to the Census data. 52
Figure 2.5. Absolute percent difference between P-MEDM estimates and actual District level data taken from the Census 2011 data for tap water source. 52
Figure 2.6. Infant mortality rate as calculated from new P-MEDM small area estimates. 53
Figure 3.1. The relationship between the rate of homeownership and the error in margin for that variable by rural and urban classifications. 66
Figure 3.2. National level percent absolute error in margin of P-MEDM produced estimates as compared to margin totals taken from the 2011 Bangladesh Census. Home ownership variable is not shown. 67
Figure 3.3. 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. 69
Figure 3.4. Comparison of DHS raw estimates and P-MEDM reweighted estimates for three model variables for Bangladesh. 71
Figure 3.5. Comparison of DHS raw estimates and P-MEDM reweighted estimates for three model variables for rural and urban designations. 72
Figure 3.6. Comparison of DHS raw estimates and P-MEDM reweighted estimates for three model variables at the Division level. 73
Figure 3.7. Comparison of DHS raw estimates and P-MEDM reweighted estimates for three model variables at the District level. 74
Figure 3.8. District level infant mortality rates from the Bangladesh MICS 2009, and as calculated for the P-MEDM estimates. Districts are grouped by Division. 79
Figure 3.9. District level infant mortality rates from the Bangladesh MICS 2009, as calculated for the raw DHS estimates, and as calculated for the P-MEDM estimates at three different levels of variance. 80

INTRODUCTION

Background

Advances in remote sensing, dasymetric mapping techniques, and the ever-increasing availability of spatial datasets have markedly enhanced the resolution of global human population distribution databases. These datasets demonstrate an enormous improvement over the conventional use of choropleth maps to represent population distribution and are vital for an array of analysis and planning purposes including humanitarian response, disease mapping, risk analysis, and evacuation modeling (Dilley et al. 2005; Garb, Cromley, and Wait 2007; Morrow 1999; NRC 2007).

In many situations however, it is not sufficient to simply know the number of people in an area, but it is equally important to understand the demographic characteristics of the population. Lately, particular attention is being paid to the spatial dimension of these demographic analyses (Castro 2007; Entwisle 2007; Voss 2007). Unfortunately, spatial demographic studies often suffer from one of two conditions: 1.) the study is conducted at a coarse scale with a wide geographic extent, where the spatial component (and thus demographic variable of interest) is limited to a high level administrative enumeration unit, or 2.) the study is done at a scale which provides results at fine spatial and demographic resolution, but with limited geographic extent.

Surveys centered on population health and demographics have been ongoing for decades. These surveys are performed at a variety of geographic scales, but are rarely representative at small area levels. Some, such as the Demographic and Health Survey (DHS) and the Multiple Indicator Cluster Survey (MICS), have been completed for several years for numerous countries around the world. A broad range of individual and household attributes are collected through these surveys and are invaluable for understanding indicators of social progress such as child health. Although these data are a valuable resource for demographers, they still provide limited information on the small area population characteristics that undoubtedly vary across geographic space.

Recent efforts have focused on top-down approaches for higher resolution population distribution modeling as well as bottom-up methods aimed at demographic and sociocultural attribution of population data. Dasymetric modeling techniques in particular are moving toward a micro-level analytical framework in order to take advantage of the growing volume of information about population, whether pertaining to spatial distribution, temporal dynamics, or demographic characterization. One example of the expanding knowledgebase for population characterization is the Population Density Tables (PDT) developed by ORNL (Urban 2012). The PDT effort is aimed at capturing detailed local-area occupancy information for a large variety of facility types throughout the world.

Even though the focus is on population density, via observed or surrogate data for population count and facility area, these data are collected through open source pathways and are typically rich with fine resolution demographic data. These types of data, along with existing survey data, can be used to deepen our understanding of the demographic composition of the global population, particularly in data poor areas of the world. The immediate issue facing population researchers then is how to integrate these types of data with existing population distribution datasets that are available at a much higher spatial precision. Currently, there is no standard method by which to approach this integration.

Components of the Study

To address the lack of demographic resolution in existing population distribution datasets, this research will detail data fusion methods that link survey-based household-level demographic data to census-based population data and implement these methods as a proof of concept for a single country. I demonstrate the methodology within the context of the typically data-poor developing world where there is less availability of detailed, national scale demographic data compared to developed nations. The focus area for this research is Bangladesh, as it fits within the scope of a developing nation with expected spatial data challenges. The 2011 Bangladesh Census is used to provide population and demographic margins at the national and subnational level, and the 2011 Bangladesh Demographic and Health Survey (DHS) dataset provides household-level demographic data. For this case study, I will use proposed data fusion methods to specifically create District level infant mortality rates which currently do not exist via any other source, but yet could be vastly useful for demographic research.

Study Area – Bangladesh

Bangladesh is an important geographic area for demographic research for a number of reasons, primarily due to its situation as a developing country impacted by a large population. Population growth and urbanization are two important demographic challenges facing Bangladesh, particularly in light of the high level of poverty present in the country. Bangladesh continues to struggle with poverty and ranks near the bottom (146th) among all nations on the Human Development Index (HDI) as presented in the 2011 Human Development Report (UNDP 2011). The HPI is a composite measure of human development in a country which takes into account life expectancy (at birth and in total), educational attainment, and Gross National Income (GNI) per capita. Bangladesh also ranks highest among all countries in environmental deprivations among the multidimensionally poor, based on the lack of improved cooking fuel, drinking water, and sanitation (UNDP 2011).

According to the Population Reference Bureau (PRB 2014), Bangladesh ranks 8th in the world for total population in both 2014 and projected to 2050, and is currently ranked 8th in population density worldwide (see Tables A.1 and A.2 located in the Appendix). Not only is Bangladesh projected to continue to be one of the world's most populous countries, but the majority of the land area in the country is densely populated. While Bangladesh is currently listed 8th in population density worldwide, the argument could be made that it holds a place of more importance than the seven countries ranked higher. The majority of those seven countries are island nations or city-states, and all but one of the seven have a land area of less than 1,000 sq. km; this is in comparison to the nearly 144,000 sq. km. that make up the country of Bangladesh. The population pressure on the available land area in Bangladesh may become even greater in the future with the potential of shrinking land availability due to sea level rise as one aspect of climate change. It's estimated that by the year 2050, a projected 0.5 meter rise in sea level could result in Bangladesh losing over 10% of its land area, which would affect an estimated 15 million people (Wheeler 2011).

In addition to the large total population figure for the country, Bangladesh is experiencing rapid growth in urban areas, with the population estimated to grow yearly by 3.5% in the largest metropolitan areas (Fleischer, Lutz, and Schmidt 2010; Streatfield and Karar 2008). As counted in the 2011 Census, Bangladesh's total urban population is nearly 34 million people, and with projected population growth trends, this number has the potential to increase to over 100 million in this century (Streatfield and Karar 2008). Limited resources in rural areas, as well as the lack of work and the decrease in land cultivation, have driven many people into urban areas where they will have more opportunities. However, this sizeable and continuous rural to urban migration puts stress on the limited urban resources as well. This is particularly true in Dhaka where the annual growth rate of slum areas has been estimated to be around 7% (Fleischer, Lutz, and Schmidt 2010).

Extreme overcrowding in Bangladesh's urban areas, and particularly in slums, trigger the prevalence of airborne and waterborne diseases. The degraded health of urban slum dwellers can be chiefly attributed to overcrowding and lack of access to clean water and sanitation. These living conditions can foster negative outcomes for the population including stress due to crowding, insecurity due to lack of housing and land tenure, and various types of illegal or criminal activities (NIPORT 2013; Streatfield and Karar 2008). The lack of financial opportunities coupled with the high risk of disease can push populations living in slums even further into poverty.

Infant Mortality

Infant mortality is a widely accepted measure of not only the health status of children, but is also used more broadly as a gauge of the overall health of a

community. This indicator reflects the status of maternal health, the accessibility and quality of health care, socioeconomic status, and the availability of supportive services in the community. Infant mortality rates are used worldwide to assess the health and well-being of populations, and growing evidence suggests that higher infant mortality within a population is linked to that population's overall health and development across the life course. The population issues affecting Bangladesh now and into the future have a direct link to infant mortality. Poverty, lack of education, lack of adequate sanitation and water, and limited access to health care are all associated with poor health outcomes. Rapidly increasing populations, particularly with the potential for increased slum dwellings in densely populated urban areas can exacerbate these issues (Black, Morris, and Bryce 2003; Moran et al. 2009).

Reducing child mortality is one of the eight United Nations Millennium Development Goals (MGDs) which have a target date of 2015. Of particular note for this study, even as the rate of under-five deaths is decreasing globally, the proportion of deaths that occur during the first month after birth is increasing. As Black, Morris, and Bryce (2003) effectively illustrate, major causes of infant and child death, including socioeconomic conditions, vary geographically. Consequently, identifying this variation is critical in order to aid researchers, and ultimately policymakers, in adapting strategies to more effectively reduce infant mortality to meet this United Nations Millennium Development Goal.

Identifying the primary source of infant mortality is fundamental for reducing the instances of death, yet applying appropriate policies to address this issue is equally vital. In their global study of the socioeconomic determinants of infant mortality, Schell, Reilly, Rosling, Peterson, and Ekstrom (2007) use country level data to demonstrate that the comparative importance of major health determinants varies between socioeconomic strata. Given this finding, employing generalized health policies to address infant mortality is problematic. Furthermore, it’s reasonable to expect that the geographic variation of determinants exist not only at a country level, but would extend even further into subnational and local areas.

Given the importance of geographic variation with regard to both cause and remediation of infant mortality, the availability of demographic data at higher spatial resolutions will aid in the identification of local level triggers for infant mortality and allow for targeted responses to this concern.

Research Outline

In Chapter I, I present background on methods used to create population distribution datasets to provide context for this research. Over the last few decades, these methods have been developed and continually improved upon as new techniques and data emerge. Of particular importance has been the rapidly

growing availability of very high resolution satellite imagery. This development alone is an important breakthrough in global population mapping, not simply due to the high spatial resolution, but also because of the increasing temporal resolution of this imagery. Now more than ever, researchers are able to use this resource not just to map populations, but to monitor changes in locations of settlements, agriculture, and conflict. However, the limitation of using imagery for demographic characterization is obvious. This chapter is an important preface to Chapter II, as it makes the case that although it's important to continue to improve on the spatial and temporal resolution of large scale population distribution datasets, the current lack of demographic resolution cannot be ignored. Demographic information should not be confined to small scale studies, but rather addressed as a necessary input to a holistic picture of global population. The novelty of this content is the presentation of theoretical issues of implementing spatiotemporal representations of real world phenomena. The question of "what phenomena are we attempting to measure?" is an important one since the degree of geographical detail at which phenomena occur, the data we use to measure it, and the models we use to represent it can all vary greatly. Chapter I begins the dialog on how all this information can be integrated and reconciled in order to most accurately characterize human activity space.

The core of this dissertation is presented in Chapter II, where the data fusion methods themselves are discussed. Background is provided on the methods that will be used, and the requirements and description of the input datasets is reviewed to set the groundwork for the development of the model. The theoretical basis for choosing constraint variables for the model is one of the most critical aspects of the modeling process and thus is discussed at length both in the methods as well as the results.

Chapter III tackles the difficult question of validation of the data fusion output. At the most basic level, what this dissertation research is attempting to do is create demographic data that does not currently exist. However, these new data will only be useful to researchers and policymakers if there is some way to assess the quality of the model and assure users that the output data represent reasonably good estimates of the actual values. Validation is becoming a standard component of quantitative social science research, but is still scarce in spatial microsimulation literature. In this sense, this work is of particular importance in that it adds to the limited existing body of knowledge. Both internal and external validation is presented here, with particular attention paid to interpreting the results of the comparison of my output from Chapter II to an external dataset.

References

Black, R. E., Morris, S. S., & Bryce, J. (2003). Where and why are 10 million children dying every year?. *The Lancet*, *361*(9376), 2226-2234.

Castro, M. C. (2007). Spatial Demography: An Opportunity to Improve Policy Making at Diverse Decision Levels. *Population Research and Policy Review, 26*(5-6), 477-509.

Dilley, M., Chen, R. S., Deichmann, U., Lerner-Lam, A. L., Arnold, M., Agwe, J., Buys, P., Kjekstad, O., Lyon, B., & Yetman, G. (2005). Natural Disaster Hotspots: A Global Risk Analysis. Washington, D.C.: The World Bank.

Entwisle, B. (2007). Putting People Into Place. *Demography, 44*(4), 687-703.

Fleischer, A., Lutz, M., & Schmidt, J.-O. (2010). *Population Dynamics in Bangladesh*. Eschborn: GTZ.

Garb, J. L., Cromley, R. G., & Wait, R. B. (2007). Estimating Populations at Risk for Disaster Preparedness and Response. *Journal of Homeland Security and Emergency Management, 4*(1), Article 3.

Moran, A. C., Choudhury, N., Khan, N. U. Z., Karar, Z. A., Wahed, T., Rashid, S. F., & Alam, M. A. (2009). Newborn care practices among slum dwellers in Dhaka, Bangladesh: a quantitative and qualitative exploratory study. *BMC pregnancy and childbirth*, *9*(1), 54.

Morrow, B. H. (1999). Identifying and Mapping Community Vulnerability. *Disasters, 23*(1), 1-18.

National Research Council (NRC). (2007). *Tools and Methods for Estimating Population at Risk from Natural Disasters and Complex Humanitarian Crises*. Washington, D.C.: The National Academies Press.

National Institute of Population Research and Training (NIPORT), ICDDRB, MEASURE Evaluation, ACPR, USAID. (2013). *Bangladesh Urban Health Survey 2013 Preliminary Results***.** Retrieved December 4, 2014, from http://www.icddrb.org/images/stories/WhatWeDo/Research/CPUCC/uhs2013preliminaryreport_final_16oct2014.pdf.

Population Reference Bureau (PRB). (2014). Retrieved December 4, 2014, from http://www.prb.org/.

Schell, C. O., Reilly, M., Rosling, H., Peterson, S., & Ekström, A. M. (2007). Socioeconomic determinants of infant mortality: a worldwide study of 152 low-, middle-, and high-income countries. *Scandinavian journal of public health*,*35*(3), 288-297.

Streatfield, P. K., & Karar, Z. A. (2008). Population Challenges for Bangladesh in the Coming Decades. *Journal of Health, Population, and Nutrition*, *26*(3), 261–272.

United Nations Development Program (UNDP). (2011). *Human Development Report 2011*. New York: Oxford University Press.

Urban, M. L. (2012). "Developing a Global Data Infrastructure for Human Geography". Presented at The Association of American Geographers Annual Meeting, New York, NY.

Voss, P. R. (2007). Demography as a Spatial Social Science. *Population Research and Policy Review, 26*(5-6), 457-476.

Wheeler, D. (2011). *Quantifying Vulnerability to Climate Change: Implications for Adaptation Assistance*. Working Paper 240. Center for Global Development, Washington, DC. Retrieved December 4, 2014, from www.cgdev.org/content/publications/detail/1424759.

Appendix

Table A.1. Country population rankings for mid-year 2014 for the ten most populated countries, with mid-year 2050 populations and rankings included for those same countries.

Population Rank	Country	Population (mid-2014)	Population Projection (mid-2050)	Population Rank (2050)
1	China	1,364,072,000	1,311,782,000	2
2	India	1,296,245,000	1,656,919,000	1
3	United States	317,731,000	395,284,000	4
4	Indonesia	251,452,000	365,323,000	5
5	Brazil	202,769,000	226,348,000	7
6	Pakistan	193,979,000	348,007,000	6
7	Nigeria	177,542,000	396,509,000	3
8	Bangladesh	158,513,000	201,948,000	8
9	Russia	143,747,000	134,139,000	14
10	Japan	127,060,000	97,076,000	19

Source: Population Reference Bureau (PRB) Datafinder (www.prb.org)

Table A.2. Population density rankings as of mid-year 2014 for the top ten countries with the highest densities.

Density Rank	Country	Population (mid-2014)	Land Area (sq. km.)	Density (pop/sq. km)
1	Macao, SAR	620,666	26	23,872
2	Monaco	37,000	2	18,500
3	Singapore	5,487,400	683	8,034
4	Hong Kong, SAR	7,241,000	1,099	6,589
5	Bahrain	1,319,000	694	1,901
6	Malta	427,000	316	1,351
7	Maldives	369,900	298	1,241
8	Bangladesh	158,513,000	143,972	1,101
9	Channel Islands	162,000	194	835
10	Palestinian Territory	4,400,000	6,019	731

Source: Population Reference Bureau (PRB) Datafinder (www.prb.org)

CHAPTER I
THE LANDSCAN GLOBAL POPULATION DISTRIBUTION PROJECT: CURRENT STATE OF THE ART AND PROSPECTIVE INNOVATION

A version of this chapter was originally published by Amy N. Rose and Eddie Bright:

Amy N. Rose, Eddie Bright. 2014. “The LandScan Global Population Distribution Project: Current State of the Art and Prospective Innovation”. Proceedings of the Population Association of America Annual Meeting 2014, Boston, MA.

Amy Rose was the primary author of this article, and was responsible for writing the text of this article in its entirety. Eddie Bright contributed historical knowledge of input datasets that had been used in LandScan, particularly under the sections “Vector Spatial Data” and “Raster Spatial Data”, and also contributed Figure 1.3.

Abstract

Advances in remote sensing, dasymetric mapping techniques, and the ever-increasing availability of spatial datasets have enhanced global human population distribution databases. These datasets demonstrate an enormous improvement over the conventional use of choropleth maps to represent population distribution and are vital for analysis and planning purposes including humanitarian response, disease mapping, risk analysis, and evacuation modeling. Dasymetric mapping techniques have been employed to address spatial mismatch, but also to develop finer resolution population distributions in areas of the world where subnational census data are coarse or non-existent. One such implementation is the LandScan Global model which provides a 30 arc-second global population distribution based on ancillary datasets such as land cover, slope, proximity to roads, and settlement locations. This work will review the current state of the LandScan model, future innovations aimed at increasing spatial and demographic resolution, and situate LandScan within the landscape of other global population distribution datasets.

Introduction

LandScan is now in its fifteenth year of production, a milestone which warrants a long overdue examination of this dataset; including its origins, data inputs, ongoing innovations, and future pathways. Although LandScan is a widely distributed dataset, a frequent critique is the lack of documentation of and transparency into the process by which the data are developed each year. Over the course of the last 15 years, LandScan has undergone numerous changes with regard to input data sources. Furthermore, the basic methodology used to produce LandScan continues to be refined based on the increased availability of good quality national level datasets, as well as new and innovative imagery processing techniques. This paper highlights the input data issues and availability, recent advances in the LandScan modeling process, and the plans for the future. Future requirements and the appropriate strategy to meet those needs are critical; the necessity of population distribution datasets is apparent

through the continued funding and production of such efforts. Often the similarities and differences between each of these efforts is questioned and so this paper examines other datasets that are frequently likened to LandScan.

High resolution population data are a necessary requirement for research and analysis in two primary contexts: prevention and response situations. For prevention, the requirement focuses on developing capabilities to improve the response to critical events over time and increase the resilience of local populations. For response situations, there is often a shorter-term need for information that can aid in estimating damage and providing relief to affected areas (NRC 2007, 109). The LandScan population distribution database is explicitly developed to address both of these contexts by helping to identify populations at risk, whether from natural or man-made events. The goal is not to specifically quantify a certain number of people in a particular geographical location, but rather to provide a more realistic population distribution for consequence assessments than is afforded strictly by census counts.

The uses of LandScan data are varied and extensive. LandScan has significantly enhanced the utility and impact of various applications in areas including counter-terrorism, homeland security, emergency planning and management, consequence analysis, epidemiology, exposure analysis, and urban sprawl detection. National and international organizations including the United Nations (UN), the World Health Organization (WHO), the Food and Agricultural Organization (FAO), and several federal agencies in the U.S. and other countries currently employ these data in their analyses. In all, Oak Ridge National Laboratory (ORNL) has distributed LandScan Global population data to over 750 different organizations spread across the world (Bhaduri et al. 2002).

Background

Population data are vital for an array of analysis and planning purposes and is a basic component of humanitarian response efforts (NRC 2007). Unfortunately, those analyses are often hampered by the reality that population source data (e.g. census information) rarely conform to the spatial extent of analysis regions (Goodchild, Anselin, and Deichmann 1993). To mitigate the problem of disparate spatial data assimilation, a surface representing population distribution is often employed for spatial analyses (Fotheringham and Rogerson 1993). Numerous methodologies have been devised to create population surfaces including simple areal interpolation (Tobler 1979), allometric and regression models (Lo and Welch 1977), and dasymetric models (Wright 1936; Deichmann 1996; Eicher and Brewer 2001; Mennis and Hultgren 2006). The following section discusses these methods.

Population Distribution Methods

Bracken and Martin (1989) developed surface representations of demographic data for census enumeration areas in the United Kingdom using a centroid-kernel method. Using this point-based method, population is assigned to the centroids of the enumeration areas. A moving kernel window is positioned over each centroid and population is allocated to cells falling within the window based on weights of other centroids falling within the window. This method uses a weighting based on a distance decay function so that closer centroids have more weight than those farther away. While this does allow for some areas of the resulting surface to contain zero population, it also assumes that population density decreases further away from the centroid.

The largest issue with a point-based interpolation method is that generally the total value within each source enumeration area is not preserved. To address this, areal interpolation methods were developed (Goodchild and Lam 1980; Flowerdew and Green 1992; Goodchild, Anselin, and Deichmann 1993; Fisher and Langford 1996). Simple areal interpolation uses the geometric properties of the source area to determine the proportion of the total value that will be allocated to that area. Essentially, an area-weighted function is used where the area proportion serves as a weight for population distribution. However, the assumed homogeneity of an area, especially with regard to population distribution is problematic (Lam 1983).

The dasymetric method was originally developed by Benjamin Semenov-Tian-Shansky in the early 1900's in his proposal to produce the 'Dasymetric Map of European Russia' (Petrov 2012). However, this fact is often overlooked in literature, owing instead the popularization of this method to J.K. Wright by way of his population density mapping of Cape Cod (Wright 1936). Wright felt that choropleth maps, with their uniform distribution, did not provide a realistic representation of population within enumeration units, so he used his local knowledge to develop a method to refine the population densities (Fisher and Langford 1996). In this method, partitions are iteratively created to disaggregate general zones to detailed zones of population density while preserving the population counts of the original zones. Building on this methodology, more robust dasymetric methods have been developed recently. Since population is commonly related to other factors including land use and transportation networks, it stands to reason that those features would be important inputs in the development of representative population distribution surfaces. These so-called "intelligent" dasymetric methods use ancillary variables to guide the redistribution of the population (Flowerdew, Green, and Kehris 1991; Flowerdew and Green 1992; Mennis and Hultgren 2006). Figures 1.1 and 1.2 below show an example of the dasymetric mapping process.

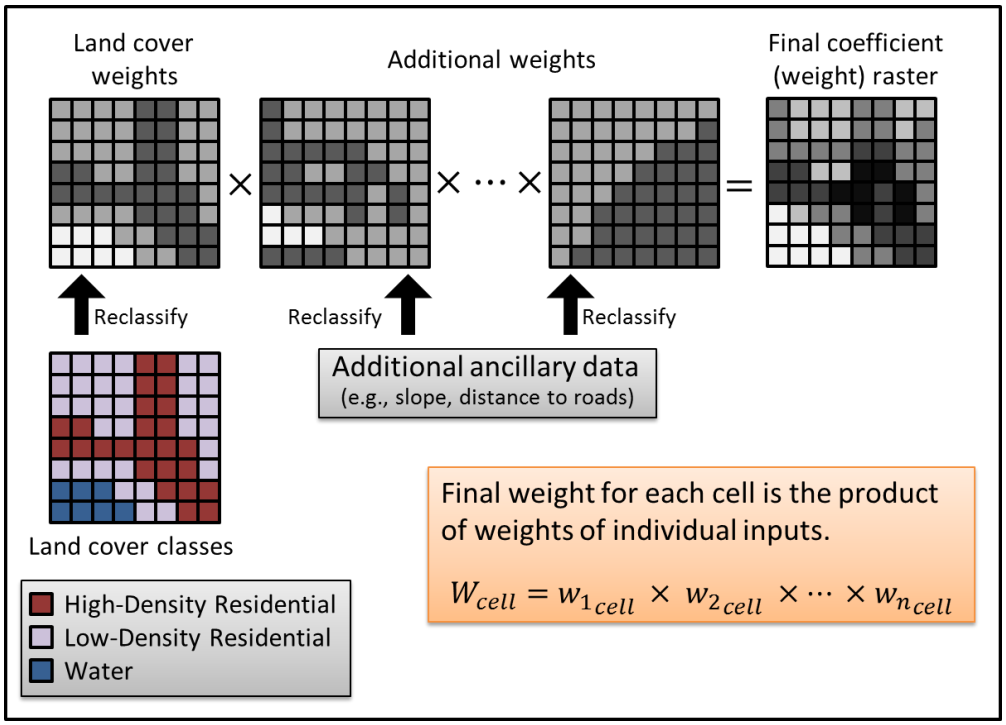

Figure 1.1. Creating a coefficient surface using ancillary variables.

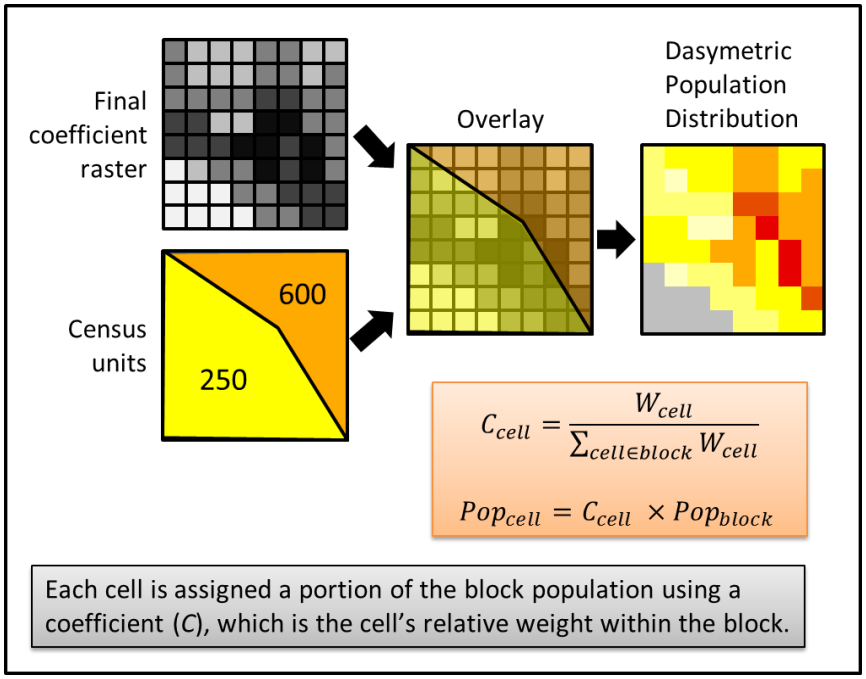

Figure 1.2. Using the coefficient surface to redistribute population.

Methodological Considerations

The application of a particular population allocation approach should consider not just methodology, but also the theoretical issues associated with implementing spatiotemporal representations of real world phenomena. When developing population distribution surfaces, identifying what phenomena we are attempting to measure is critical since the degree of geographical detail at which phenomena occur, the data we use to measure it, and the models we use to represent it can greatly vary. Population occurs at one scale, input data are available at a different scale, and the output population distribution at yet another scale. How can all this information be integrated and reconciled in order to most accurately characterize human activity space?

The foundation of population distribution methods is the reliance on spatial information that either directly describes population at or within a given place, or provides a reasonable proxy for human activity. At the most basic level, point or areal features with an associated population count are required. Additional features including land cover, slope, or transportation infrastructure that can be used to inform the distribution can help refine the geographical detail of the output. Regardless, uncertainty will still manifest in the output, particularly due to both spatial and temporal misalignment of input data.

The increasingly high geographical detail that is required for useful analytical output makes global population distribution datasets with a very high spatial resolution extremely desirable as input for modeling. In some areas of the world, particularly developed nations, very detailed census and spatial data are available, so that creating population distributions with high geographic detail is not difficult. However, most of the world lacks in either good quality spatial data or recent (and accurate) census data. This lack of data however should not be a barrier to producing population distribution datasets, but rather the implications of data limitations should be acknowledged so that users of population distribution datasets are aware of how these issues may affect the outcome of their own analyses.

Input Data Effects

The annual update of LandScan, as well as other global population distribution datasets, requires that subnational boundaries for every country are examined with a critical eye. Often subnational boundaries coincide with natural features or circumscribe urban areas. The spatial accuracy and precision of administrative boundaries can be evaluated by overlaying the data on high resolution geo-rectified imagery. Some boundary datasets are topologically correct but not spatially accurate, with errors commonly exceeding multiple kilometers. Often subnational administrative boundaries have been greatly generalized for cartographic purposes which may result in villages, towns, and even portions of larger cities being located into the wrong administrative region.

The resolution, or more specifically the scale and detail, of administrative unit boundaries and the population associated with them is the single most important input into a dasymetric model – or any population distribution model. The idea is that the smaller the administrative units, the more accurate any population distribution model within those units will be. This is a reasonable assumption in most cases, but it also assumes that the boundaries are spatially accurate. Furthermore, population associated with administrative boundaries is typically census or residential population counts. Using these population counts to develop a finer resolution population distribution for small boundary units within an urban area may not accurately reflect the true human activity signature in that area since it will not account for other potential activity outside of residences.

Temporal Factors

Population distribution datasets are most useful when the date of representation aligns with the analyses being performed. Regardless of the temporal domain of the input data used to create the population distribution surface, the output data are either implicitly or explicitly capturing a snapshot in time of the actual distribution of humans on the earth's surface. The nature of population dynamics prevents any census from being truly representative of where people are located. However, population distribution methods aim to capture the population at a given time.

Universally, population distribution methods use some type of "official" population counts whether census year counts, intercensal estimates, or registry information. Given this, it's important to note that censuses can vary greatly in their regularity and level of execution (Table 1.1). Particularly for developing countries, even if scheduled, censuses may not be conducted regularly and in some cases have been lacking for multiple decades.

It is important to consider the implications of using outdated census information for developing any population distribution datasets. First, for many countries migration – either voluntary or forced – can have an enormous impact on the population dynamics of an area. For example, recent civil conflicts in neighboring countries Sudan (and areas that are now South Sudan), Ethiopia, and Somalia have produced an associated internal displacement, as well as cross-border displacement with each other in what essentially constituted a population swap. Of these countries, Ethiopia conducted its most recent census in 2007, Sudan in 2008 – prior to South Sudan becoming an independent state, and the last official census conducted in Somalia was in 1987. Therefore, using these numbers to develop any population distribution dataset must involve some type of validation of current conditions, and applied as corrections to either the input datasets or the output itself.

Table 1.1. Census history of selected nations (IPC 2008).

Country or Area	Round of Population Census						
	1950 Round (1945-54)	1960 Round (1955-64)	1970 Round (1965-74)	1980 Round (1975-84)	1990 Round (1985-94)	2000 Round (1995-2004)	2010 Round (2005-2014)
Benin	--	--	--	3/20-30/79 (F)	2/15/92	2/11/02	2012 (S)
Botswana	5/7/46	4/1/64	8/31/71	8/12-26/81	8/14-23/91	8/17-26/01	2011 (P)
Burundi	--	--	--	8/15-16/79 (F)	8/15-16/90	--	8/08 (S)
Cameroon	--	--	--	4/9/76 (F)	4/14-28/87	--	11/11/05
Cape Verde	12/15/50	12/15/60	12/15/70	6/1-2/80	6/23/90	6/16-30/00	2010 (P)
Chad	--	--	--	--	4/15/93 (F)	--	--
Comoros	--	9/7/58 (F)	7/66-9/66	9/15/80	9/15/91	9/03	2013 (S)
Cote d'Ivoire	--	--	--	4/30/75 (F)	3/1/88	11/21/98	2008 (S)
Ethiopia	--	--	--	5/9/84 (F)	10/11-27/94	--	5/29/07
Gabon	--	10/8/60-5/61 (F)	6/1/69-6/70	8/1-31/80	7/1-31/93	12/31/03	2014 (S)
Gambia, The	--	4/17/63 (F)	4/21/73	4/15/83	4/15/93	4/15/03	2013 (S)
Ghana	2/1/48 (F)	3/20/60	3/1/70	3/11/84	--	3/26/00	2010 (S)
Kenya	2/25-8/23/48 (F)	8/15-16/62	8/24-25/69	8/25/79	8/24/89	8/24/99	2009 (S)
Liberia	--	4/2/62 (F)	2/1/74	2/1-14/84	--	--	3/21-27/08
Libya	7/31/54 (F)	7/31/64	7/31/73	7/31/84	--	8/95	4/15-5/1/06
Madagascar	--	--	--	1/26-8/18/75 (F)	8/1-19/93	--	2009 (S)
Morocco	--	6/18/60 (F)	7/20/71	9/3-21/82	9/4/94	9/1/04	9/14 (S)
Mozambique	9/21/50	9/5/60	12/15/70	8/1/80	--	8/1-15/97	8/1-15/07
Namibia	--	9/6/60 (F)	5/6/70	8/26/81	10/21/91	8/27-9/11/01	2011 (P)
Niger	--	--	--	10/7-11/6/77 (F)	5/10-24/88	5/20-6/18/01	2011 (S)
Nigeria	7/52-6/53	11/5-8/63	--	--	11/27-29/91		3/21/06
Rwanda	--	--	--	8/15-16/78 (F)	8/15-16/91	8/16-30/02	2012 (S)
Somalia	--	--	--	2/7-20/75 (F)	11/86-2/87	--	--
Sudan	--	7/1/55-9/2/56 (F)	4/3/73	2/1/83	4/15/93	--	4/22-5/6/08
Swaziland	5/7/46	7-8/56	5/24/66	8/25/76	8/25/86	5/12/97	5/12/07
Togo	--	--	3/1-4/30/70 (F)	11/22/81	--	--	--
Zimbabwe	--	4/10-5/20/62 (F)	4/21-5/11/69	8/18/82	8/18/92	8/18/02	2012 (S)

(F) First full modern census taken (P) Projected based on pattern of census dates (S) Scheduled; not yet taken or known if taken

Uncertainty Assessments

Quantifying the accuracy of population distribution data has been the focus of several recent publications (Hall, Stroh, and Paya 2012; Mondal and Tatem 2012; Tatem et al. 2011), yet the reality is that the true accuracy of these datasets are difficult if not impossible to measure due not only to the lack of independent ground-truth data, but also as a result of spatiotemporal population dynamics continually taking place at a variety of scales.

Current State of the Art

More recently, dasymetric mapping techniques have been employed not just to address spatial mismatch, but also to develop finer resolution population distributions in areas of the world where subnational census data are coarse or non-existent. One such implementation is the LandScan Global dataset which provides a 30 arc-second global population distribution based on ancillary datasets such as land cover, slope, proximity to roads, and settlement locations (Dobson et al. 2000). LandScan is just one of the available high-resolution global population datasets that has been developed in recent years. Projects like the Gridded Population of the World (GPW), the Global Rural Urban Mapping Project (GRUMP), and AfriPop are all public domain datasets available for research and analysis use. Of these, AfriPop is a recent addition, initiated in 2009 with an aim of producing population distribution maps for the whole of Africa. While each of these population distribution efforts relies on some of the same datasets as inputs to their model, they are methodologically dissimilar. The next section discusses each of these population distribution models in turn.

Gridded Population of the World

The Gridded Population of the World (GPW) project, originally produced at the National Center for Geographic Information Analysis (NCGIA) in 1995 (Tobler et al., 1995), can be considered the first noteworthy attempt to generate a consistent spatial global population dataset. GPW was developed with the purpose of providing a (nighttime) global population dataset at a scale required to perform analysis at the country or global level. GPW has been successively updated in 2000 (Deichmann et al., 2001) and 2005 (Balk and Yetman, 2004) by The Center for International Earth Science Information Network (CIESIN) at Columbia University. GPW is simple, areal-weighted redistribution (i.e. proportional allocation) of census population from their census boundaries (or administrative polygons) to a uniform quadrilateral grid. The population totals in GPW datasets are based solely on administrative boundary data and population estimates associated with those administrative units. Since that first release, successive releases of GPW have incorporated larger numbers of administrative units which are used to guide the population redistribution. However, no effort is made to model population distribution, and no ancillary data were used to predict

population distribution or revise the population estimates. The only assumption made was that population is uniformly distributed within each administrative unit.

The principal drawbacks of the GPW dataset are the coarse resolution of 2.5 arc-minutes and the lack of any modeling of population distribution within administrative units. The latter is particularly significant in that the result is evenly distributed populations across any given administrative unit. This is unlikely to represent a realistic population distribution, especially within large units with significant variation in land cover characteristics. Although GPW utilizes administrative unit data at the finest resolution available, this by far is not the resolution of human settlement on the ground.

Global Urban-Rural Mapping Project

The Global Rural-Urban Mapping Project (GRUMP) began around 2004 building on the methodology employed by GPW in order to provide a framework of urban and rural area (nighttime) populations. The GRUMP population grid represents a nighttime population at a 30 arc-second resolution. GRUMP employs urban proportional reallocation based on land area using nighttime lights, and in some cases the Digital Chart of the World (DCW). Although areal weighting is used to redistribute population from polygons to grid, it uses urban polygons as well as administrative polygons and makes assumptions about the parameters of the assignment in those urban areas. GRUMP relies on accurate subnational population projections as well as the usefulness of the nighttime lights dataset. The update frequency of GRUMP is not published, and no update has been produced since the initial release of the dataset in the mid-2000s.

AfriPop

The AfriPop project was initiated in 2009 in order to provide spatial data on African (nighttime) population distributions to support epidemiological modeling. Although at the present time this is not a global population distribution dataset, related projects, AsiaPop and AmeriPop, were launched in 2012 and 2013 respectively. Initial development and release of country level population datasets has been swift, although updates to the data have not yet been undertaken. AfriPop is promoted as a 100m resolution dataset which uses an area-weighted reallocation method to distribute population counts. The AfriPop methodology combines census boundaries and population, settlement points, and land cover information in a semi-automated process to produce population distribution maps (Linard et al. 2012). Specifically for the population distribution weighting, 30 meter LandSat Enhanced Thematic Mapper (ETM) is the primary input, and there is limited use of varying scale vector data indicating populated places.

The AfriPop methodology relies on accurate subnational census data (population counts) as well as accurate open source settlement locations. However, particularly in the case of Africa, these are not always available. Similarly to

GPW, AfriPop considers greater numbers (and smaller sizes) of subnational administrative areas to be a primary factor in improving the resolution of the output data. However, simply increasing the number of administrative areas will not improve the actual resolution of the output population distribution if those boundaries do not have a high level of geographical detail on their own. Since the output population distribution is fractional rather than integer, AfriPop generally has few areas of no population. That is, if an administrative unit has at least one person, that person will be divided up over the entire area. While population allocated to most areas may or may not be accurate indications of human activity space, one of the limitations of AfriPop is that it depicts population density as being greater in urban peripheral areas and villages rather than urban centers.

LandScan

Created in 1998, LandScan is a global population database depicting an ambient (24-hour average) population distribution. It was conceived as an improved resolution global population distribution database for estimating populations at risk. The LandScan methodology disaggregates subnational census information through a suite of novel and dynamically adaptable algorithms using spatial data, imagery derived spatial products, and manual corrections. LandScan exploits spatial data and imagery analysis technologies in a multi-variable dasymetric modeling approach (Dobson et al. 2000). LandScan data represent an average, or ambient, population that integrates diurnal movements and collective travel habits into a single measure (Dobson et al. 2000). Since natural or man-made emergencies may occur at any time of the day, the goal of LandScan is to develop a population distribution surface in totality, not just the locations of where people sleep. LandScan data are analogous to mapping biological habitat where the species total environment (e.g. nests, feeding areas, travel pathways, density gradients and boundary conditions) are considered (Guisan and Thuiller 2005).

Because of the ambient nature of LandScan, care should be taken with direct comparisons of LandScan data with other population distribution surfaces. Furthermore, since LandScan incorporates new spatial data and imagery into the distribution algorithms for each new version, comparing different versions of the dataset on a cell to cell basis may result to misleading conclusions. While some of the differences between LandScan versions are due to recently developed urban or suburban expansion, there are many cases where a village identified with high resolution imagery may have existed for hundreds of years, but was never represented in various spatial data products. Figure 1.3 illustrates the differences between the original version of Landscan from 1998 and a later version, both of which are at a 30 arc-second resolution.

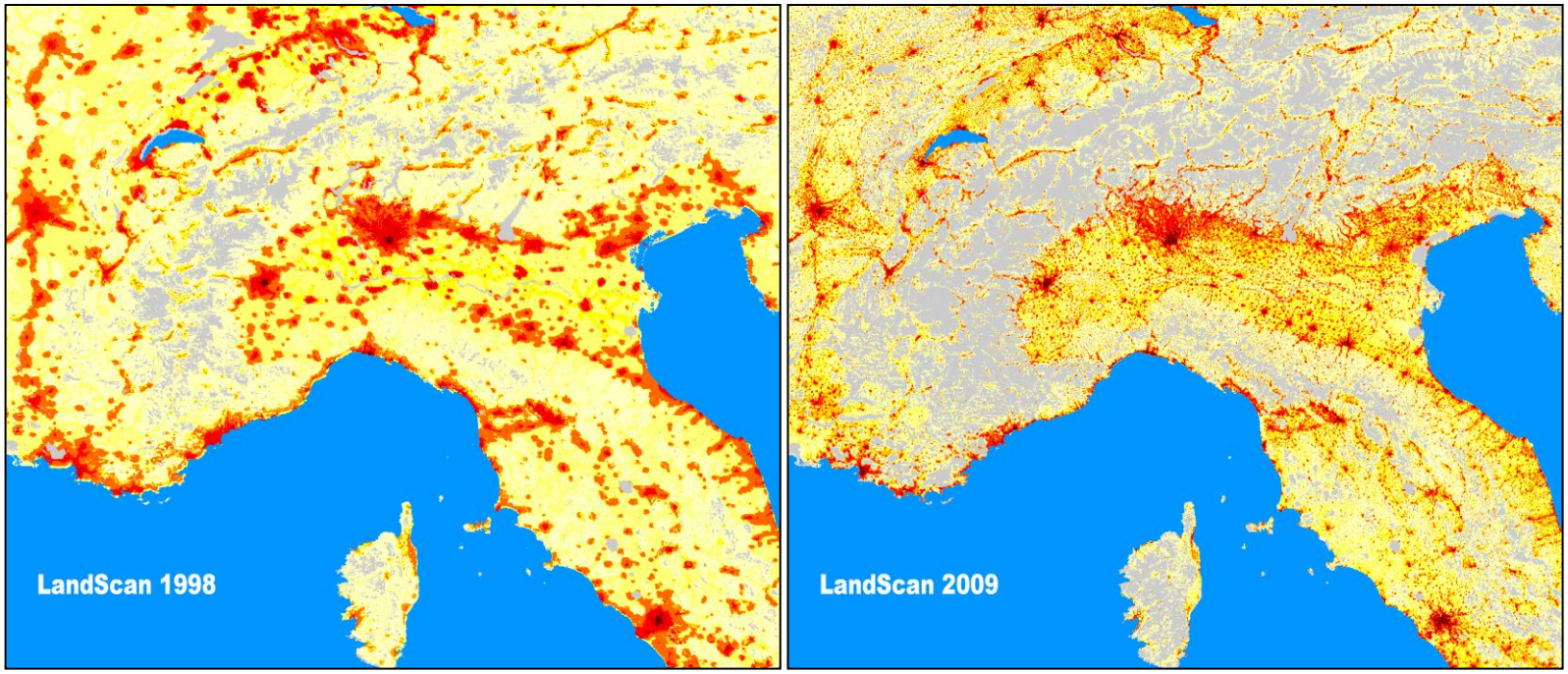

Figure 1.3. Evolution of LandScan.

Data and Methods

Census Data

LandScan, like other previously discussed datasets, relies on accurate subnational census data as the basis for population distribution. The LandScan development team expends significant effort each year to collect population censuses or estimates for each country from a myriad of open sources. While recent census tabulations are preferred, censuses can vary greatly in their regularity and level of execution. As such, official intercensal estimates, registration data, or other proxies may be used. Regardless of how subnational population counts are obtained, at the country level the population is always normalized to the mid-year national estimates provided through the CIA World Factbook (https://www.cia.gov/library/publications/the-world-factbook/).

Once subnational population numbers have been acquired, those population counts are matched to their corresponding administrative boundaries. All LandScan data meet a pycnophylactic condition (i.e. mass preserving). The LandScan distribution algorithm uses a normalization process that ensures that the population counts distributed throughout each administrative unit will reflect exactly the number reported for that bounded area. After the initial population distribution, the procedure determines if either more or fewer people need to be added to the calculation of each administrative area in order to total to the subnational numbers. If more population is needed in an area, then population is added to the cells with the greatest likelihood coefficient; if less population is needed, then population is subtracted from the cells with the least likelihood coefficient. This procedure is repeated for each administrative area in a country. Since administrative units are used as spatial controls for population distribution, the spatial and attribute accuracy of administrative boundaries are integral considerations for the LandScan model.

Subnational Administrative Boundaries

A required input for the development of any population distribution dataset is administrative boundaries, yet the availability of accurate digital subnational administrative boundaries remains a problem for many nations of the world. Accurate subnational boundary information has been identified as a significant need for population research and emergency response (NRC 2007). The United Nations Geospatial Information Working Group (UNGIWG) began work on a Second Administrative Level Boundary (SALB) dataset to globally map second level administrative boundaries. Due to differences in the quality of the documents and data compiled for the SALB, the spatial data layers of the database are more suitable for thematic mapping rather than precise representation or modeling. Another effort, the Global Administrative Unit Layers (GAUL) by the Food and Agriculture Organization (FAO) of the United Nations

within the European Commission Food Security Programme, aims at “compiling and disseminating the most reliable spatial information on administrative units for all countries in the world” (FAO 2010). Unfortunately, many countries are not willing to share detailed subnational boundary information (Lauber 2007). Furthermore, much of the data volunteered in these repositories lack adequate spatial fidelity for accurate representations at 30 arc-second cell resolution. Numerous digital versions of subnational boundary information exist, but only the minority possesses the spatial accuracy desired for input into the LandScan model.

The annual update of LandScan requires that subnational boundaries for every country are examined with a critical eye. Often subnational boundaries coincide with natural features or circumscribe urban areas. The spatial accuracy and precision of administrative boundaries can be evaluated by overlaying the data on high resolution geo-rectified imagery. Some boundary datasets are topologically correct but not spatially accurate, with errors commonly exceeding multiple kilometers. Often subnational administrative boundaries have been greatly generalized for cartographic purposes which may result in villages, towns, and even portions of larger cities being located into the wrong administrative region. Digital representations of boundaries with inadequate detail create topological incongruities. For example, a city alongside a meandering river coincident with a boundary may be placed on the incorrect side of a generalized line segment. The census population that is associated with the city would not be reflected in the output because the erroneous representation would position the city in the adjacent administrative region. In effect, the population of one administrative region would be underrepresented and the population of the adjacent administrative region would be overrepresented. In other cases, either purposely or by accident, digital boundaries may simply be spatially incorrect.

Inconsistencies between census information and the concurrent subnational administrative boundaries are common. For many nations, subnational administrative boundaries or other census enumeration areas are dynamic. Often administrative boundaries move, merge, or change names for a variety of reasons. Recent census information may not be temporally synced to the existing boundary files. That is, the name of a region remains the same, but the area it represents may have changed. These types of discrepancies can be difficult to detect by automated quality assurance algorithms.
The administrative unit level by which the census data are distributed by the LandScan algorithm varies considerably in size and spatial precision from country to country. The number of administrative units per nation is considered in the LandScan model parameterization process. Nations with few, but very large administrative areas require a different weight in the model parameters to allocate representative populations to their appropriate locations. Generally smaller administrative boundaries lead to better population distribution – if the

boundaries are spatially accurate. Small administrative areas that are spatially misplaced actually induce population distribution errors. To mitigate this, where possible, LandScan algorithms merge poor sub-province boundaries to the province level and distribute the entire province population according to the population likelihood locations determined by the model rather than constrict population distributions to incorrect locations. Very small administrative or enumeration areas equivalent to U.S. Census blocks or block groups have unintended consequences for modeling an ambient population. Since the populations associated with census tables are places of residence, commercial and industrial areas may have zero or very low populations associated with them. Thus the output would be reflective of a residential only population distribution instead of an ambient population distribution.

If they display reasonable geographic detail for mapping at 30 arc-seconds, the lowest level administrative units available are used in the weighting process, but they are not used to construct the pycnophylactic constraints. Given the intent to represent the distribution of persons, not simply the distribution of persons at night, using the lower level administrative units would not provide an accurate representation. Take for example a U.S. Census block level illustration; using the block unit which contains the Willis Tower in Chicago would yield an output population distribution with no people in it. Although this is a reasonable residential population representation, it does not in any way depict the true population activity of that block. Using higher level administrative units (larger areas) also has the benefit of smoothing out the noise in the distribution process in so that the weights are essentially a trend line.

Vector Spatial Data

Vector spatial data used in the LandScan model are chosen for their usefulness in predicting possible population distributions. The vector map (VMap) series distributed through the National Geospatial-Intelligence Agency (NGA) includes global coverage for a set of spatial features including road and rail networks, hydrologic features and drainage systems, utility networks, airports, selected elevation contours, international boundaries, populated places, and geographical names. VMap data are created by extracting spatial features from hardcopy maps at a consistent scale, converting these features into a digital representation of their locations and attributes, and finalizing the output to the Vector Product Format (VPF). The initial version of LandScan incorporated the VMap Level 0 series data. The primary source for the VMap0 data was the 1:1,000,000 scale Operational Navigation Chart (ONC) series co-produced by the military mapping authorities of Australia, Canada, United Kingdom, and the United States. The absolute horizontal accuracy of VMap0 data for all features derived from ONCs is 2,040 meters at 90 percent circular error.

At the time of the initial release of LandScan, VMap0 was the only global coverage available. Subsequent versions of LandScan incorporated NGA's VMap Level 1 data as they became available. VMap1 is a medium resolution database derived primarily from 1:250,000 Joint Operation Graphic hardcopy sources. Vector information collected from a 1:250,000 Class 1 source may be expected to be accurate to within 125 meters while certain other sources may result in accuracy of 500 meters on the ground (NGA 1995). For selected areas, LandScan employed both VMap Level 2 data (based on 1:50,000 scale maps) and Urban Vector Map UVMAP data (based on 1:7,500 – 1:25,000 scale maps). These higher resolution input data result in increased spatial precision and an increased quantity of various spatial features. However, VMAP products are not without problems. VMAP products are based on hardcopy maps created by various producers at different dates and therefore feature details are not necessarily consistent from one map to the next and features may not match at tile edges. The date of the original map compilations captured within a VMAP tile may, in some cases, be quite old, creating population distribution anomalies especially for rapidly developing urban areas. Vector data layers used in the LandScan modeling algorithms include roads, populated areas (urban boundaries), and populated points (towns and villages). Each data layer serves as an indicator of likely population locations. Spatial feature representation and accuracy is markedly improved from VMAP0 to VMAP1 to VMAP2 for all features used as inputs into the LandScan modeling algorithms.

Given their intricate spatial patterns, digital coastlines require a very high resolution to represent coastal features accurately. Coastal areas are dynamic landscapes; shorelines change and coastal islands may grow, shift positions, or disappear entirely. In some parts of the world, such as Dubai or parts of Yemen, new land is being created on the coast which is essentially pushing the coastlines seaward. Popular coastline databases may lack both the spatial resolution and the update frequency to capture these complexities, and instead, intersect current land features thereby potentially missing populated areas. For this reason, LandScan extends all coastal boundaries several kilometers seaward to ensure all shore and small island features are encapsulated within an administrative unit boundary. Instead of a vector shoreline, land cover data and imagery are used to capture populated areas along the coast. That being said, most subnational boundary data do not have enough geographic detail to allow for even finer resolution population distributions than LandScan that are currently being pursued. Errors of omission and commission are common problems with using poor quality boundary data, particularly near coastlines. Figure 1.4 shows an example of this in Dar es Salaam, Tanzania.

Raster Spatial Data

Elevation and slope are incorporated into the LandScan modeling algorithms as most human settlements occur at low to moderate elevations on flat to gently

sloping terrain (Cohen and Small 1998; Dobson et al. 2000). However, every continent has instances of settlements at high elevations. For some nations (e.g. Eritrea and Yemen) coastal regions may be particularly inhospitable for human habitation and settlements occur at higher elevations. Although on occasion humans will frequent the very highest elevations on the planet, there are elevation levels above which consistent habitation is impractical. Since settlements can occur at all but the very highest elevations, the LandScan model does not evaluate elevation other than to exclude population distributions at elevations over 18,000 ft.

The initial version of LandScan incorporated the DTED0 global elevation dataset into the population distribution model. Using DTED0, a single slope value was calculated for each cell by examining the elevation of the eight neighboring cells. Thus for approximately each square kilometer a single slope value was calculated and given a weight in the model according to the settlement patterns of each country. Calculating an average slope value for each 30 arc-second cell may mask many small, relatively flat areas within the cell that are suitable for habitation. Subsequent versions of LandScan processed NGA's Digital Terrain Elevation Data (DTED1). DTED 1 has near global land area coverage with resolution of 3 arc-seconds.

Compared to DTED0, the DTED1 data have 100 times more elevation values for each output cell. The elevation data from each DTED data tile were extracted, converted to raster format, projected to a Universal Transverse Mercator (UTM) projection to calculate slope values in appropriate units, and the calculated slope values were re-projected to the original geographic projection. Rather than average these high resolution slope values into an aggregated 30 arc-second cell slope value, a slope suitability ranking index was assigned to each 3 arc-second cell. By processing the higher resolution elevation data into discrete slope categories, cells with very high slope values do not disproportionately skew the average slope of an aggregated cell. An average of these 3 arc-second slope index values is calculated for each 30 arc-second cell. This methodology produces a slope suitability index by capturing the percentage of land that is preferable for habitation within each 30 arc-second cell. This discrete slope index technique accounts for some of the small areas with relatively gentle slope such as narrow valleys or knolls within an overall steep landscape. More recent versions of LandScan integrated the Shuttle Radar Topography Mission (SRTM1) elevation data. The SRTM data also have a resolution of 3 arc-seconds with an extent in latitude from 60 degrees North to 56 degrees South. The SRTM data were processed in a fashion similar to that used for the DTED1 data, and SRTM2 data were processed for limited areas.

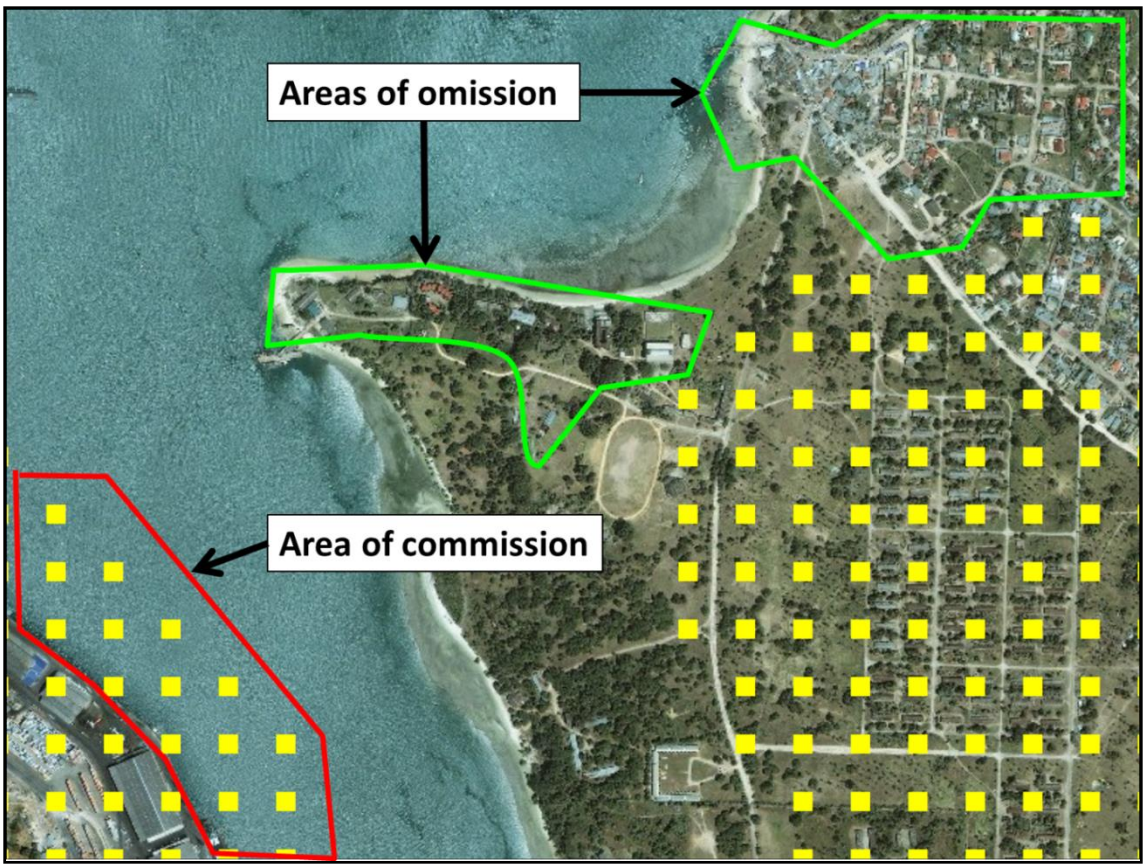

Figure 1.4. Example of errors of omission and commission that can arise when using subnational administrative boundaries with incorporated coastlines that lack appropriate geographic detail to develop very high resolution population distributions. Yellow squares mark the centerpoint of a 100m resolution population distribution.

Land cover is a crucial data layer for modeling population distribution (Tian et al. 2005, Mubareka et al. 2008). Signatures of anthropogenic activity are observed in all but the most remote places on earth. Dedicated land cover or land use classes such as urban, developed, residential, or even agricultural, reinforce the importance of human modifications on the landscape. LandScan analyzes each land cover dataset by comparing the data for each nation to settlement characteristics observed in coincident high resolution imagery. Relative weights assigned to different land cover types generally follow a logical progression (Urban > Agriculture > Grass/Forest > Desert > Ice/Water). Absolute weighting values are adjusted for different settlement patterns, data completeness, and spatial accuracy of the land cover data.

The values for the weights are amended considerably from one area to another due to cultural settlement patterns, economic activity, and history. For example, prime agricultural lands in the U.S. generally have low population densities whereas principal agricultural regions in developing nations generally exhibit much higher densities. Even many medium resolution land cover databases do not capture thousands of smaller towns, rural villages, or kampongs, that have population densities rivaling that found in many developed cities. The land cover data associated with these agricultural areas receive higher population likelihood in the distribution model, not only for the missing villages, but also to represent the potential number of workers in fields.

High Resolution Imagery Analysis

Currently, image archives ingest terabytes of data per day (National Academies Press 2006). High resolution imagery is employed in every phase of the LandScan population distribution process. At the outset, high resolution imagery is used to identify settlement patterns and building characteristics, but is also used to evaluate the accuracy and precision of the different spatial data layers used in the models as well as to adapt the weighting factor for each layer in the model algorithms. Preliminary model output is superimposed on high resolution imagery to verify relative population distributions and magnitude. As new spatial data are received, iterative modifications to variable weights in the likelihood coefficient file are made and the distribution algorithms are re-calculated.

High resolution imagery is used to create or modify existing spatial data layers, especially to update or refine the land cover data related to urban boundary delineations. To speed processing of vast image archives, an automated urban boundary delineation algorithm based on texture and edge information extracted from high-resolution panchromatic images has been developed. Feature vectors using statistical features derived from gray level co-occurrence matrices (GLCM) and local edge pattern (LEP) matrices were extracted and deployed in a parallel computing environment to speed computation time (Cheriyadat et al. 2007). Additional automated extraction of urban features is accomplished by implementing a parallel algorithm for a Gabor filter based multi-resolution representation of panchromatic images to compute the texture features at different scales and orientations. The Gabor filter representation has been widely used for texture analysis as it can be shown that it minimizes the joint two-dimensional uncertainties in space and time (Vijayaraj, Bright, and Bhaduri 2007).

Positional or attribute errors and anomalies are to be anticipated in large volumes of disparate spatial data. LandScan includes a manual verification and modification process to improve the spatial precision and relative magnitude of the population distribution. Imagery analysts identify obvious population distribution errors and create an additional spatial data layer of population

likelihood coefficient modifications to correct or mitigate input data anomalies. Derived land cover such as that from Thematic Mapper data can not reveal urban properties such as building densities or building heights that can be readily inferred with visual inspection using high resolution imagery. As a result, a large number of modifications are made to urban areas and urban extents.

Future Innovations

Although it may seem that dissecting the data and methodologies used in the creation of the LandScan population distribution database is simply a metadata exercise, it is an important start for understanding the limitations, as well as potential for improvement of these data, including new directions for research. Despite the extensive use of Geographic Information Systems (GIS) to map spatial patterns and distributions of population, much less research has been undertaken with globally mapping the sociocultural properties of place including human experience, social hierarchies, and power relations. It is yet to be determined whether LandScan is an appropriate or plausible vehicle for this type of critical theoretical extension; the scale of the output alone suggests that it would be impractical, if not impossible, to foray into this realm using the current "top-down" approach. However, some recent research has been done using LandScan as the starting point for developing more complex population models. Specifically, Fernandez et al. (2010) experimented with building a credible synthetic population for Afghanistan consisting of 31 million social atoms. While the research centered on data and computational feasibility, the outcome showed promise in using social theories to model regions at the individual level.

The current spatial resolution of LandScan is appropriate for a global dataset and is more than adequate for use in national and subnational analyses. However, there is a continued effort to improve the resolution without sacrificing the availability of annual updates. To this end, work has been done in the area of automating both the extraction and characterization of human settlements in a high performance computing (HPC) environment (Cheriyadat et al. 2007). The Settlement Mapper Tool (SMT) developed by ORNL rapidly delineates and characterizes settlements using high resolution imagery. The application of this to LandScan is twofold. First, the extraction of human settlements will allow for a baseline understanding of settlement area vs. non-settled area across the globe. Second, the characterization allows a deeper dive by using low-level image features (i.e., edges, lines, textures, etc.) to characterize types of settlement structures (Graesser et al. 2012). In a remote sensing environment, neighborhoods can be classified as formal or informal using both spatial and spectral characteristics (Hofmann et al. 2008). While formal settlements tend to have larger building sizes, regular street patterns, and more homogeneous building materials, informal settlements are more likely to consist of small buildings, narrow streets, and heterogeneity of materials.

Another component of the ongoing efforts to improve the spatial, temporal, and demographic dimensions of LandScan, is the Population Density Tables (PDT). PDT is a set of national-level databases, which model density by the sociocultural activities and associated facility spaces that define normal patterns of life. Such a modeling exercise involves automated collection, and subsequent fusion of information from open, published sources including academic journals, official government statistics, websites of humanitarian organizations, corporate and university web pages, buildings databases, tourism brochures, reported field observations, and NGO reports. The result is a collection of data that provide a baseline, spatiotemporal and demographic snapshot of facility occupancy levels for nighttime and daytime periods at a higher spatial resolution. Activity spaces, which include a variety of buildings and open air facilities, are characterized by functional categories which include residential, cultural institutions, retail, commercial, and transportation spaces. Population density estimates for night and day capture a normal workweek and routinely scheduled periodic and episodic events. The PDT database also contains other population density values representing weekends, holidays, special events, and seasonal population distribution scenarios.

In light of efforts such as PDT and SMT, the advancement of population distribution models has moved beyond numerical input and toward attempting to quantify qualitative data. In the same sense that expert or local knowledge can inform intelligent dasymetric mapping, so too can the fusion of SMT and PDT type outputs produce a richer understanding of not just population distribution, but also the nature of settlement and the characteristics of the people and activities that take place within that settlement. The challenge and current focus then is to facilitate the convergence of remote sensing techniques and qualitative sociological data and methods in order to produce a richer understanding of the population and human activity of a particular area at very high spatial, temporal, and demographic resolutions.

Conclusion

What's missing from LandScan and other global population distribution datasets is the demographic depth that is most useful for analysis and policy formation. So while it can be argued that LandScan is enormously useful for addressing questions about the magnitude of population at risk in cases of natural disasters or conflict, it does not address the more complex relationships and characteristics of these populations that are tied to long term policy questions of health, education, and poverty.

The need for finer resolution spatiodemographic data for analysis is well documented in the literature. Analyses of health, accessibility, and poverty issues, of emergency response planning, and of political stability are all examples of where both population distribution and characterization must be key

components, particularly in the developing world. In response to this need, work has begun to improve the demographic depth of LandScan beyond simply the age/sex breakdown that is currently available to LandScan users.

Linking demographic survey data such as the Demographic and Health Survey (DHS) to a population distribution database would unquestionably enable the use of both datasets in a wider variety of studies. Furthermore, a generalizable methodology for providing this linkage to LandScan Global rather than for specific cities or countries would provide a particularly useful dimension to the LandScan data, as well as a possible pathway for many researchers to replicate for their particular needs. To this end, current work is being undertaken to develop a prototype data fusion methodology to link DHS data to census data for a single country as a proof of concept, as well as to document challenges that arise with regard to uncertainty and validation issues.

References

Balk, D., & Yetman, G. (2004). The Global Distribution of Population: Evaluating the gains in resolution refinement. CIESIN, Columbia University, N.Y.

Bhaduri, B. L., Bright, E. A., Coleman, P. R., & Dobson, J. E. (2002). LandScan: Locating People is What Matters. *Geoinformatics, 5*(2), 34-37.

Bracken, I., & Martin, D. (1989). The generation of spatial population distributions from census centroid data. *Environment and Planning A, 21*(4), 537-543.

Cheriyadat, A., Bright, E., Potere, D., & Bhaduri, B. (2007). Mapping of settlements in high-resolution satellite imagery using high performance computing. *GeoJournal, 69*(1-2), 119-129. doi: 10.1007/s10708-007-9101-0

Cohen, J. E., & Small, C. (1998). Hypsographic demography: The distribution of human population by altitude. *Proceedings of the National Academy of Sciences, 95*, 14009-14014.

Deichmann, U., Global Resource Information Database National Center for Geographic Information Analysis Consultative Group on International Agricultural Research (1996). *A Review of Spatial Population Database Design and Modeling*: National Center for Geographic Information and Analysis.

Deichmann, U., Balk, D., & Yetman, G. (2001). Transforming Population Data for Interdisciplinary Usages: From census to grid. CIESIN, Columbia University, NY.

Dobson, J., Bright, E., Coleman, P., Durfee, R., & Worley, B. (2000). LandScan: A global population database for estimating populations at risk. *Photogrammetric Engineering and Remote Sensing, 66*(7), 849-857.

Eicher, C. L., & Brewer, C. A. (2001). Dasymetric Mapping and Areal Interpolation: Implementation and Evaluation. *Cartography and Geographic Information Science, 28*(2), 125-138.

Fernandez, S. J., Rose, A. N., Bright, E. A., Beaver, J. M., Symons, C. T., Omitaomu, O. A., & Jiao, C. (2010, 20-22 Aug. 2010). *Construction of Synthetic Populations with Key Attributes: Simulation Set-Up While Accommodating Multiple Approaches within a Flexible Simulation Platform.* Paper presented at the 2010 IEEE Second International Conference on Social Computing (SocialCom).

Fisher, P., & Langford, M. (1996). Modeling Sensitivity to Accuracy in Classified Imagery: A Study of Areal Interpolation by Dasymetric Mapping. *The Professional Geographer, 48*(3), 299-309. doi: 10.1111/j.0033-0124.1996.00299.x

Flowerdew, R., & Green, M. (1992). Developments in areal interpolation methods and GIS. *The Annals of Regional Science, 26*(1), 67-78. doi: 10.1007/bf01581481

Flowerdew, R., Green, M., & Kehris, E. (1991). USING AREAL INTERPOLATION METHODS IN GEOGRAPHIC INFORMATION SYSTEMS. *Papers in Regional Science, 70*(3), 303-315. doi: 10.1111/j.1435-5597.1991.tb01734.x

Fotheringham, A. S., & Rogerson, P. (1993). GIS and spatial analytical problems. *International Journal of Geographical Information Science, 7*(1), 3-19. doi: 10.1080/02693799308901936

Goodchild, M. F., Anselin, L., & Deichmann, U. (1993). A framework for the areal interpolation of socioeconomic data. *Environment and Planning A, 25*, 383-397.

Goodchild, M. F., & Lam, N. S.-N. (1980). Areal interpolation: a variant of the traditional spatial problem. *Geo-Processing, 1*, 297-312.

Graesser, J., Cheriyadat, A., Vatsavai, R. R., Chandola, V., Long, J., & Bright, E. (2012). Image Based Characterization of Formal and Informal Neighborhoods in an Urban Landscape. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 5*(4), 1164-1176.

Guisan, A., & Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecology Letters, 8*(9), 993-1009. doi: 10.1111/j.1461-0248.2005.00792.x

Hall, O., Stroh, E., & Paya, F. (2012). From Census to Grids: Comparing Gridded Population of the World with Swedish Census Records. *The Open Geography Journal, 5*, 1-5.

Hofmann, P., Strobl, J., Blaschke, T., & Kux, H. (2008). Detecting informal settlements from QuickBird data in Rio de Janeiro using an object based approach. In T. Blaschke, S. Lang & G. Hay (Eds.), *Object-Based Image Analysis* (pp. 531-553): Springer Berlin Heidelberg.

Lam, N. S.-N. (1983). Spatial Interpolation Methods: A Review. *The American Cartographer, 10*(2), 129-149.

Lauber, W.G. (2007). Political Geography and Emergency Relief. In *Tools and Methods for Estimating Population at Risk from Natural Disasters and Complex Humanitarian Crises*. Washington, D.C.: The National Academies Press.

Linard, C., Gilbert, M., Snow, R. W., Noor, A. M., & Tatem, A. J. (2012). Population distribution, settlement patterns and accessibility across Africa in 2010. *PLoS One, 7*(2), e31743. doi: 10.1371/journal.pone.0031743

Lo, C. P., & Welch, R. (1977). Chinese Urban Population Estimates. *Annals of the Association of American Geographers, 67*(2), 246-253.

Mennis, J., & Hultgren, T. (2006). Intelligent Dasymetric Mapping and Its Application to Areal Interpolation. *Cartography and Geographic Information Science, 33*(3), 179-194.

Mondal, P., & Tatem, A. J. (2012). Uncertainties in Measuring Populations Potentially Impacted by Sea Level Rise and Coastal Flooding. *PLoS One, 7*(10), e48191. doi: 10.1371/journal.pone.0048191

Mubareka, S., Ehrlich, D., Bonn, F., & Kayitakire, F. (2008). Settlement location and population density estimation in rugged terrain using information derived from Landsat ETM and SRTM data. *International Journal of Remote Sensing, 29*(8), 2339-2357. doi: 10.1080/01431160701422247

National Geospatial-Intelligence Agency (1995). MILITARY SPECIFICATION VECTOR MAP (VMap) Level 1. Retrieved September 27, 2013, from http://earth-info.nga.mil/publications/specs/printed/VMAP1/vmap1.html.

National Research Council (2007). *Tools and Methods for Estimating Population at Risk from Natural Disasters and Complex Humanitarian Crises*. Washington, D.C.: The National Academies Press.

Petrov, A. (2012). One Hundred Years of Dasymetric Mapping: Back to the Origin. *The Cartographic Journal, 49*(3), 256-264. doi: 10.1179/1743277412y.0000000001

Priorities for GEOINT Research at the National Geospatial-Intelligence Agency. (2006). The National Academies Press.

Tatem, A., Campiz, N., Gething, P., Snow, R., & Linard, C. (2011). The effects of spatial population dataset choice on estimates of population at risk of disease. *Population Health Metrics, 9*(1), 1-14. doi: 10.1186/1478-7954-9-4

Tian, Y., Yue, T., Zhu, L., & Clinton, N. (2005). Modeling population density using land cover data. *Ecological Modelling, 189*(1-2), 72-88. doi: 10.1016/j.ecolmodel.2005.03.012

Tobler, W. (1979). Smooth Pycnophylactic Interpolation for Geographical Regions. *Journal of the American Statistical Association, 74*(367), 519-530.

Tobler, W., Deichmann, U., Gottsegen, J., & Maloy, K. (1995). The Global Demography Project (D. o. Geography, Trans.): National Center for Geographic Information and Analysis.

Vijayaraj, V., Bright, E. A., & Bhaduri, B. L. (2007, 23-28 July 2007). *High resolution urban feature extraction for global population mapping using high performance computing.* Paper presented at the Geoscience and Remote Sensing Symposium, 2007. IGARSS 2007. IEEE International.

Wright, J. K. (1936). A Method of Mapping Densities of Population: With Cape Cod as an Example. *Geographical Review, 26*(1), 103-110.

CHAPTER II
DEVELOPMENT OF COMPLETE MICRODATA THROUGH DATA FUSION: BANGLADESH AS A CASE STUDY

Abstract

Despite the increasing availability of current national censuses, these datasets are limited by their lack of small area demographic depth. In this context, defined small areas are considered as areas where population and demographic information are below the scale of reliable measurement, but can vary in geographic extent. Although large area spatial population datasets are available, similar small area representations of other demographic and socioeconomic characteristics are scarce. 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. To address the lack of demographic resolution in existing population distribution datasets a first step would be to develop large area microdata that can be attached to a country- or global-level population distribution dataset. This can be achieved by reweighting a national level sample so as to estimate the detailed socioeconomic characteristics of populations and households at a small area level. In essence, this modelling 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. Methods to build synthetic spatiodemographic microdata are well documented in literature, yet these efforts have been implemented in places with limited geographic extent that are also data rich environments. More specifically, these methods have been tailored to fit particular local, regional, or national data sources with no plan or requirement for adaptation for other geography or data. To address this gap, this research presents a generalizable method for developing synthetic spatial microdata which in turn can be used to increase the demographic resolution of global population distribution data.

Introduction

In the public sector, planners rely on demographic information to support a wide range of decisions, including the identification of suitable locations to provide community services to meet local needs. 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. In the private sector, demographic information is equal to consumer information, and the characteristics of a population are used to predict market behavior. Location and demand are key to identifying profitable retail site locations based on the expected growth of specific segments of the population, developing marketing strategies designed to advertise products to different segments of the population, tracking the growth or decline of existing markets, including the diffusion of innovations (Cui et al. 2012), and determining

what products are most likely to be successful among different population groups.

Surveys aimed at collecting information such as travel behavior, socioeconomic status, health, and housing characteristics are commonly conducted for small segments of population. Household or individual samples are collected from a single neighborhood, city, region, or country to provide very detailed information concerning the activities and characteristics of a population. However, although these data contain considerable depth, they lack breadth in that they are not available for a full population. Synthetic spatial microdata can be developed to fill this data gap. 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. Methods to build synthetic spatiodemographic microdata are well documented in literature, yet these efforts have been implemented on limited geographic extent in data rich environments (Ballas, Clarke, and Wiemers 2006; Vidyattama and Tanton 2010). More specifically, these methods have been tailored to fit particular local, regional, or national data sources with no plan or requirement for adaptation for other geography or data. The significance then of this research is that it is the first attempt at developing synthetic spatial microdata using generalizable methods in order to increase the demographic resolution of global population distribution data in a data poor environment - namely the developing world.

Background

A vital but often overlooked component of population distribution datasets is the demographic characteristics of those populations. Typically, only age/sex breakdowns are available with these spatial datasets, and often only at the first administrative unit level. Unfortunately, this severely limits the usefulness of these datasets for certain types of analyses. Significant improvement to the demographic component of these datasets could be seen through the use of spatial microsimulation methods which can be used to create synthetic or simulated microdata estimates of demographic characteristics. Spatial microsimulation can be conducted by reweighting a national level sample so as to estimate the detailed socioeconomic characteristics of populations and households at a small area level. 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 (Taylor et al. 2004; Harding et al. 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 (Wong 1992; Beckman, Baggerly, and McKay 1996; Williamson, Birkin, and Rees 1998; Simpson and Tranmer 2005). Of these, iterative proportional fitting (IPF) approaches have a long history of use to address a variety of issues including voting behavior (Johnston and Pattie 2003), individual travel patterns (Beckman et al. 1996), rural policy analysis (Birkin and Clarke 1988; Ballas et al. 2006) and small area estimation (Wong 1992; Simpson and Tranmer 2005; Leyk, Nagle, and Buttenfield 2013).

Iterative Proportional Fitting (IPF)

The Iterative Proportional Fitting (IPF) method is well established, with the approach of iteratively adjusting an n-dimensional matrix until every dimension converges on predefined margins. IPF can be viewed as both a mathematical scaling procedure (Deming and Stephan 1940, Norman 1999) as well as a procedure for creating disaggregated spatial data from spatially aggregated data (Wong 1992). The earliest users of IPF, Deming and Stephan (1940), employed IPF to estimate joint distributions of population across certain variables. 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, and Turton, 1999, pg. 23).

In its simplest implementation, the IPF procedure adjusts a two-dimensional matrix iteratively until the row sums and column sums equal predefined marginal values (Birkin and Clarke 1988; Wong 1992). However, other studies have utilized multidimensional matrices by choosing a larger number of constraining variables for the model. Multidimensional IPF approaches have been used to simulate entire national scale populations (Ballas et al. 2005), examine voting patterns (Johnston and 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 (Openshaw 1984; Fotheringham and Wong 1991) 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 and Hay 1983, 1984; Johnston, Hay, and Rumley 1983, 1984; Johnston and Pattie 1993), and small area estimation (Johnston and Pattie 1993; Leyk, Buttenfield, and Nagle 2013; Nagle et al. 2014; Ruther et al. 2013).

Penalized Maximum Entropy Model (P-MEDM)

Maximum entropy models are constrained by prior information that is known about the data, but makes no assumptions about what is not known. Recent work (Nagle et al. 2012; Leyk, Nagle, and Buttenfield 2013; Nagle et al. 2014) formalized an entropy maximizing approach geared toward small area estimation and particularly dasymetric mapping. The penalized maximum entropy model (P–MEDM) as defined by Nagle et al. (2014)

$$max\ -\sum_{it}\frac{n}{N}\frac{w_{it}}{d_{it}}\ log\left(\frac{w_{it}}{d_{it}}\right)-\sum_{k}\frac{e_k^2}{2\sigma_k^2}$$

subject to relaxed *pycnophylactic* constraints

$$\sum_{it\in k} w_{it} = \widehat{Pop}_k + e_k$$

for each constraint *k*.

where n is the sample size, N is the population, and d_{it} is the 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 fit. 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 $\sum_k \frac{e_k^2}{2\sigma_k^2}$. 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_k effectively penalizing the maximum entropy solution.

The benefit of the P-MEDM approach is that it accounts for uncertainty in the input data, and in fact carries a quantifiable uncertainty forward through the model to the solution. In a practical sense, this allows the model to be flexible

and adapt to varying quality of input data. For example, small subpopulations may have large margins of error associated with them, but the P-MEDM approach can include estimates of these populations without impacting the model itself.

Moehl (2014) evaluated the performance of the P-MEDM using U.S. Census American Community Survey (ACS) Public Use Microdata samples (PUMS) and ACS five year summary files for Jefferson County, KY. This evaluation was conducted not only to assess the robustness of the procedure, but also to test the feasibility of scaling this model to a national level. The results indicated that the P-MEDM not only performed well when fitting the marginal totals, but also displayed the stability necessary to implement at a national scale.

Data

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 distribution and dynamics, which makes Bangladesh an important geographic area to examine for this research.

2011 Bangladesh Census

The 2011 Bangladesh Census was recently released from the Bangladesh Bureau of Statistics (http://www.bbs.gov.bd/home.aspx), and is available only as tables embedded in PDF format for each district level of Bangladesh. For this research, population and household counts, as well as aggregated population and household characteristics at a variety of enumeration levels were extracted from these tables and reformatted as one table listing the enumeration hierarchy and associated data row-wise. The hierarchy of enumeration areas is as follows:

- 7 Divisions ranging in population size (approximate) from 8M (Barisal) to 48M (Dhaka)
- 64 Districts (Zila)
- 545 Subdistricts (Upazila/Thana)
- Enumeration units are further dissected to Unions, Mouzas (Wards), and Villages, all of which have Census population counts and selected household and population attributes associated with them.

Demographic and Health Survey

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 is a two-stage, stratified cluster sampling. In the first stage, every enumeration area in the country is assigned a measure of size equal to the number of households or the population in that area. Then, in each stratum, a sample of these enumeration areas is selected independently with the probability proportional to the measure of size. The survey was designed to produce representative results for the country as a whole, for the urban and the rural areas separately, and for each of Bangladesh's seven administrative divisions.

The attractiveness of the DHS data for this research is two-fold. First, the extensive reach of the program makes the application of a methodology using any country DHS dataset more likely to be generalizable to a wider geographic area. Second, DHS data are a well-known resource and are widely used in a variety of academic and policy research areas. As of 2012, over 1,000 analyses using DHS data have been published in over 200 journals (Fabic et al. 2012). It is reasonable to expect therefore, that providing a more spatially explicit extension of this dataset would be immediately useful to a large number of researchers.

Although DHS surveys have been conducted for nearly three decades, GIS and spatial analysis with these data has only been largely limited to the last decade (for example see Chin et al. 2011, Soares Magalhaes and Clements 2011, Kazembe and Namangale 2007, Balk et al. 2004, Gemperli et al. 2004, Elbers, Lanjouw, and Lanjouw 2003). Although DHS are a nationally, and in some cases provincially, representative sample, they are not statistically representative estimates of the demographic characteristics at the sub-provincial level, so most of these studies have combined census data with DHS data in order to either derive explanatory covariates for household characteristics or to carry out small area estimation (Mansour, Martin, and Wright 2012). However, as noted by Mansour et al. (2012), none of these works have endeavored to discuss the problems associated with spatially linking the DHS clusters with other datasets.

Constraining Variables

The Bangladesh DHS 2011 survey data only have high level geographical locations reported for them. To address this, new spatial microdata can be created through a reweighting process in order to account for the spatial variation of the data. An essential part of a microsimulation model is the reweighting process where the survey microdata are used as a basis for calculating new person or household weights according to how representative they are for each small area. The weights are calculated and adjusted until the known marginal

distribution of the population of the small area is matched by the newly weighted survey microdata.

Theoretical Basis for Choosing Constraint Variables

Multiple constraint variables will be used to reweigh the DHS survey data in order to build a spatial microsimulation model. The constraint variables must be available in both the Census data and the DHS survey data in order to be able to calculate small area weights. It is important to select constraint variables that are as closely correlated as possible with the purposes of the new microdata (Chin and Harding 2006), and thus the choice of these constraint variables will depend on the presence of a reasonably good correlation between the constraint variables and the variable that is ultimately being mapped. A robust simulation model should result in very low error margins between the simulated and actual data. The number of constraint variables used will also be limited by computational cost and the possibility of non-convergence if too many variables are used (O'Donoghue et al. 2013).

The goal is to not only produce estimates for subpopulations, but also more importantly, to estimate the spatial allocation of these household or individual characteristics. In that the spatial disaggregation into new small area estimates is an important outcome, careful consideration should be used when selecting constraining variables for the model. Ideally, these constraining variables should represent the underlying spatial heterogeneity of population characteristics (Simpson and Tranmer 2005; Ruther et al. 2013).

Not only should constraining variables be chosen for their potential in producing reliable estimates that recreate the spatial allocation of the input data, but consideration should 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.

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.

The Bangladesh Demographic and Health Survey 2011 Final Report (http://dhsprogram.com/publications/publication-FR265-DHS-Final-Reports.cfm, pg. 115) reports on the socioeconomic differentials that affect infant mortality in

Bangladesh. In general, 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. First, infant mortality varies significantly by division. Furthermore, women's education is inversely related to infant mortality. A child's risk of dying is also associated with the socioeconomic status of the household. This information, in combination with the availability of comparable variables in the 2011 census data, was used to inform the choice of variables that were used for this research. These variables were not only used as linking variables, but also to understand the household structure. For constraint variables, the corresponding fields in the DHS and Census data are shown in Table 2.1.

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, pg. 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 householder literacy combined with other education indicators should be a reasonable proxy for the overall education level of the mother.

Table 2.1. Constraining variables chosen from both the DHS 2011 and the Bangladesh 2011 Census data.

Indicator	DHS 2011	Bangladesh 2011 Census
Education	Recode of Age/Sex/In School	Under Age 15 School Attendance: Female
	Recode of Age/Sex/In School	Age 15 - 29 School Attendance: Female
	Literacy: Female	Literacy: Female
Socioeconomic	Recode of Occupation: Women	Employment Field of Activity: Female
	Recode of Source of Drinking Water	Source of Drinking Water
	Has Electricity	Electricity Connection
	Owns Homestead	Housing Tenancy Status
Demographic	Calculated Mean Size of Household	Average Size of Household
Spatial	Type of Residence	Rural/Urban
	Region of Residence	Administrative Division

Variables were chosen for two purposes: for linking the DHS data to the Census data via the P-MEDM procedure, and for the relationship to the variable of interest. For this research, data from both the household and the individual tables were used. The individual recode and household member recode from the DHS provided data on female literacy, education, and employment as well as the variables required for combining and transforming data into an infant mortality value. The household recode provided data on household size, water source, electricity, and housing tenure. These variables are listed in Table 2.2.

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 name and a rural/urban flag, and in combination only produce 14 unique spatial variations of the data. Spatial variation in the Census data necessary to produce new microdata estimates are shown in Figure 2.1a and Figure 2.1b.

Table 2.2. Sample size of Bangladesh DHS 2011 data with breakdown of variable categories.

Householders		
Variable	**Categories**	**Count**
Household Members		***83,731***
	Male	40,661
	Female	43,070
Females in School		10,584
	Under 15	8,111
	Age 15+	2,473
Literate Females		11,568
Employed Females		2,693
(Field)	Agriculture	189
	Industry	1,475
	Service	1,029
Age of Householders	Age 0-4	9,336
	Age 5-9	9,881
	Age 10-14	9,713
	Age 15-19	8,605
	Age 20-24	7,658
	Age 25-29	6,842
	Age 30-49	19,103
	Age 50-59	5,773
	Age 60-64	2,240
	Age 65+	4,580

Households		
Variable	**Categories**	**Count**
Households		***17,141***
(Size)	1 person	260
	2 persons	1,300
	3 persons	2,807
	4 persons	4,060
	5 persons	3,449
	6 persons	2,292
	7 persons	1,251
	8+ persons	1,722
Water Source	Tap Water	1,983
	Tube Well	14,754
	Other	404
Has Electricity		10,496
Owns Home		16,093

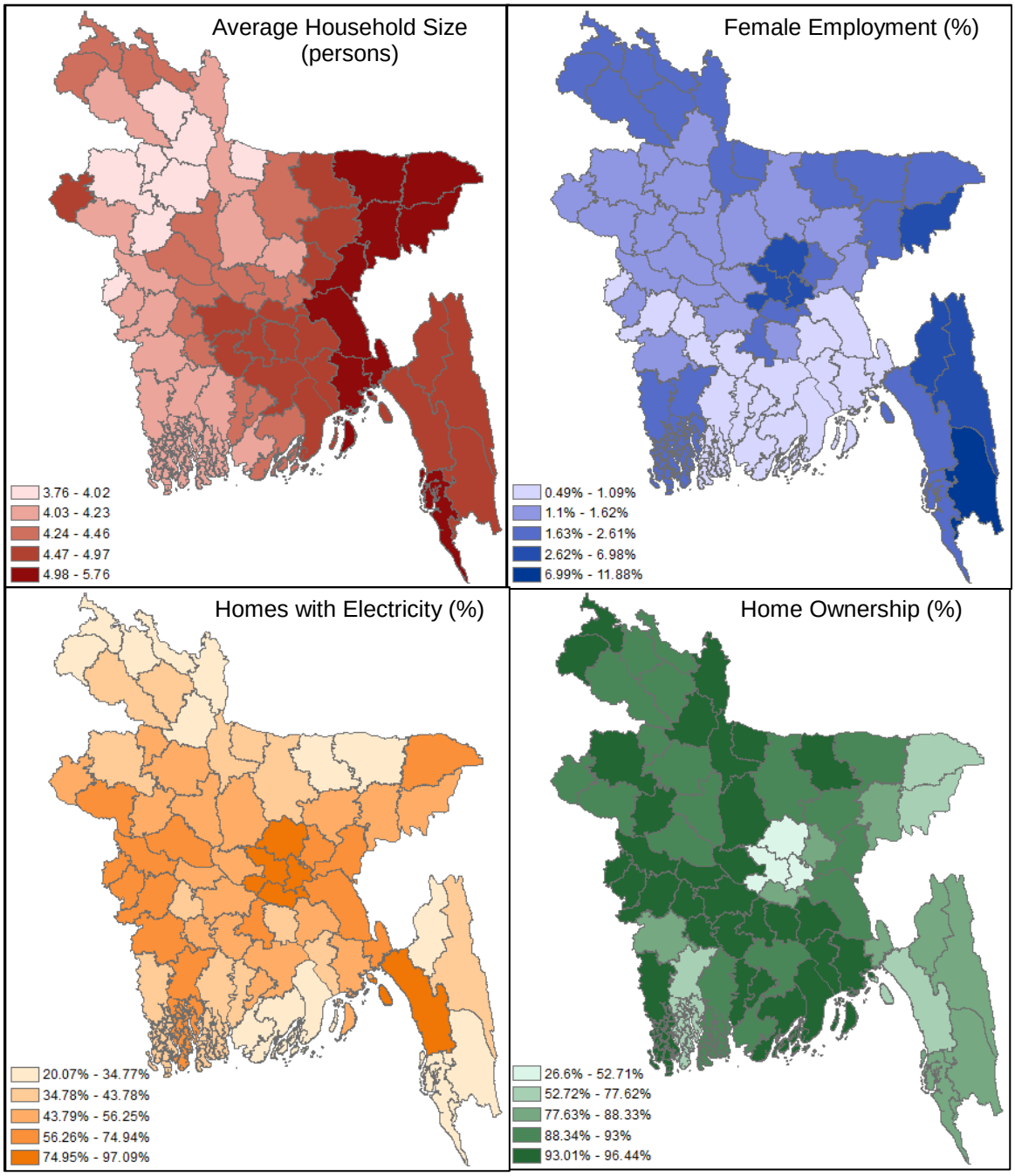

Figure 2.1a. Constraining variables from 2011 Bangladesh Census show the spatial variation necessary for producing new microdata estimates.

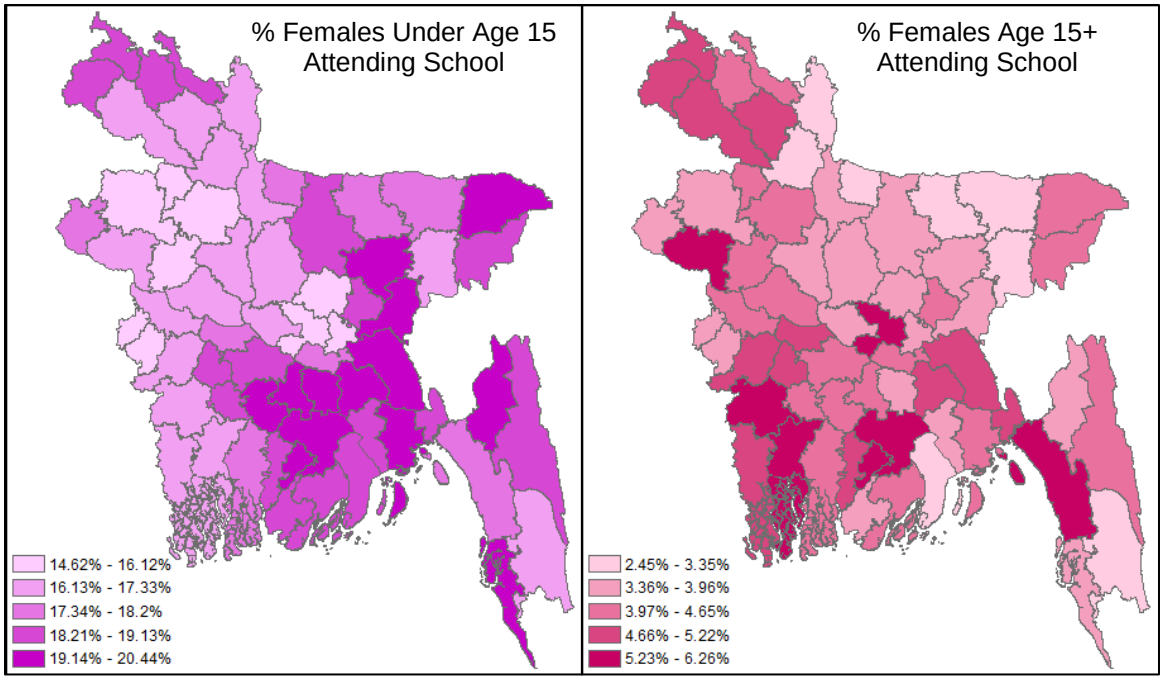

Figure 2.1b. Constraining variables from 2011 Bangladesh Census show spatial variation.

In order to verify that the variables chosen for the P-MEDM procedure are reasonable predictors of the variable of interest (infant mortality), a full regression model was run using the variables specified by the DHS report as having the greatest effect on infant mortality. Here, regression is not used to measure how infant mortality varies based on predictors such as measures of wealth and education, but rather to verify that those predictors do in fact have an effect on infant mortality.

For this analysis, logistic regression was used since the predictor variables can take any form; there is no assumption about the distribution of these variables. That is, they don't have to be normally distributed, linearly related, or of equal variance within each group. The relationship between the response and predictor variables is not a linear function in logistic regression but rather expressed as

$$\widehat{Y}_i = \frac{e^u}{1 + e^u}$$

Where $\widehat{Y}_i$ is the estimated probability that the *i*th case is in a category and u is the regular linear regression $a + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i$

Although ultimately used in the P-MEDM procedure, the variables indicating households with tubewell or other water source were left out of the regression due to multicollinearity with tap water source. Results of the regression are shown in Table 2.3.

Table 2.3. Results of logistic regression with Infant Mortality Rate as the dependent variable and for records where infant mortality was greater than zero.

InfMortRate	Coef.	Std. Err.	z	P>z	95% Conf.	
Urban	-0.0456467	0.0547779	-0.83	0.405	-0.1530094	0.0617161
Householders	0.0463714	0.0099599	4.66	0.000	0.0268503	0.0658925
FemLit	-0.0090317	0.0007841	-11.52	0.000	-0.0105685	-0.0074949
FemSchUnd15	-0.0022047	0.0010393	-2.12	0.034	-0.0042418	-0.0001677
FemSch15+	0.0010615	0.0016747	0.63	0.526	-0.0022208	0.0043439
FemEmp	0.0028015	0.0009985	2.81	0.005	0.0008445	0.0047586
TapWater	-0.0031245	0.0008676	-3.60	0.000	-0.004825	-0.0014239
Electricity	-0.0024532	0.0004886	-5.02	0.000	-0.0034109	-0.0014955
Owns	0.0003334	0.0009297	0.36	0.720	-0.0014888	0.0021556

The fitted logistic model then is

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -1.268 - .046urb + .046hh - .009femlit - .002femschund15 + .001femsch15p + .003fememp - .003tap - .002elec + .000owns$$

where $\hat{p}$ is the estimated probability of infant mortality.

Although small, the effects of some of the chosen variables are significant. Specifically, female literacy, females under 15 years old attending school, households with tap water source, and households with electricity are directly related to infant mortality rate, while household size and female employment are inversely related with infant mortality. The p-value of the chi-square is 0.000, indicating that, as a whole, the model is statistically significant. Significant effects of the predictor variables are indicated by the values in the 95% confidence intervals.

Results

Tables for each of the seven Divisions in Bangladesh were built from the DHS recode tables. These tables contained a record for each household and values for each of the variables listed in Table 2.2. Margin tables were created from the

Census by aggregating population and household characteristics at the District level using the same variables as those included from the DHS. Prior to running the P-MEDM, female literacy and water source attributes were eliminated as constraining variables and thus removed from all input tables. These variables were removed to allow for post-model run comparison of non-constraining variable estimates to Census distributions.

Each Division was modeled as a separate process to constrain the results geographically. Household-level estimates within a Division were produced across Districts so that a specific household representation was imputed for each District as varying weights (Table 2.4). A more in-depth description of the process of microdata weighting and allocation can be found in Leyk, Buttenfield, and Nagle (2013).

Table 2.4. Example of P-MEDM output weights for Bangladesh.

Household record	District level imputed weights				
	District 10791	District 10040	District 10061	District 10091	District 10421
2 103	12.62	581.68	50.15	63.99	18.49
2 107	0.69	49.34	3.27	9.73	0.85
2 112	1.05	31.51	19.45	21.54	2.45
2 116	67.65	124.85	320.65	67.22	69.43
2 121	0.03	3.17	1.29	0.97	0.15
2 125	20.10	310.58	49.85	51.07	19.59
2 129	4.95	84.30	25.38	31.47	4.92

The P-MEDM output household-level weights were multiplied by each of the household-level attributes (female literacy, female employment, school attendance, water source, etc.) to arrive at portions of the population that each weighted household represents. In order to do basic comparisons to the Census data, these weighted households were then aggregated so that totals could be mapped at the District level.

Descriptive Statistics

Following Simpson and Tranmer (2005), to obtain a measure of the overall accuracy of the P-MEDM estimates, the root mean square error (RMSE) was calculated by finding the difference between the actual data (Census) and the estimates for each area and dividing by the number of records multiplied by the number of areas. The square root of the entire calculation is then taken in order to interpret the results as average differences. By squaring the differences,

larger weights are given to larger errors in effect creating a higher sensitivity to outliers.

$$RMSE = \sqrt{\sum (estimate - actual)^2 / n}$$

The RMSE is calculated for female literacy and household water source variables for both the P-MEDM estimates and the DHS estimates. The DHS estimates for Bangladesh are only representative at the country and Division level. In order to assert that the P-MEDM estimates would not only closely fit the Census data, but also allow for representative estimates at lower geographic levels than is afforded by DHS, scaled estimates were created from the DHS data and compared to the P-MEDM estimates as well as the Census data.

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 2.5). However, the P-MEDM results showed a good fit as interpreted by the RMSE.

Table 2.5. Root mean squared error (RMSE) for non-constraining variables for P-MEDM estimates and DHS estimates as compared to 2011 Census values.

	RMSE	
	P-MEDM Estimates	DHS Estimates
Female Literacy	0.047873661	10.91674535
Tap Water	0.002431115	4.559284611
Tubewell Water	0.015288892	11.93444206
Other Water	0.004807795	4.106551414

Mapping the Results

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 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 Figure 2.2, the spatial allocation of the P-MEDM estimates for female literacy were comparable to the Census data totals. The chart in Figure 2.3 breaks down these differences numerically showing very minimal differences for all Districts. Similar results for tap water source are shown in Figures 2.4 and 2.5. 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.

Infant Mortality Rate

Ultimately, the reason for producing new household estimates using P-MEDM is to be able to improve the demographic resolution of population datasets. Demographic information that is not collected as part of the Census, not reported due to privacy concerns, or not available as joint distributions can be imputed as to arrive at reasonable estimates of these data.

Infant mortality, the variable of interest for this case study, was not collected as part of the Bangladesh Census. However, via P-MEDM, this variable can now be estimated at the subnational level. 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 date.[1]

As shown in Figure 2.6, infant mortality varies spatially and is lowest in predominantly urban areas, such as Dhaka and Chittagong. Although developing these small estimates for infant mortality where they previously did not exist is noteworthy, it's also important that these estimates be further evaluated by comparing them to either a comparable exogenous measure of infant mortality that is representative at the same level of geographic aggregation, or by other datasets that are known measureable proxies for infant mortality.

---

[1] To calculate IMR, modifications were made to Stata code originally written by Dr. Thomas Pullum, Director of Research, Measure DHS Project, to calculate mortality rates for Nigeria 2006 DHS.

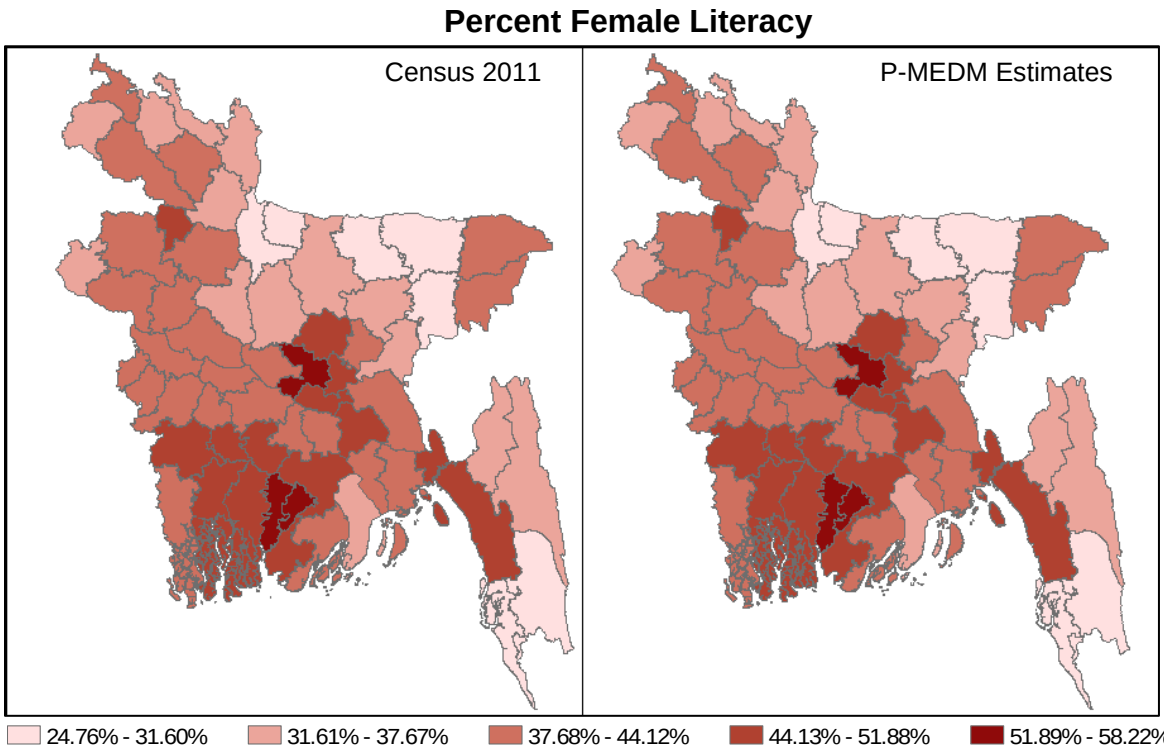

Figure 2.2. Percent female literacy from Census 2011 as compared to percent female literacy estimated by the P-MEDM procedure at the District level. Mapped results show that the estimates match well to the Census data.

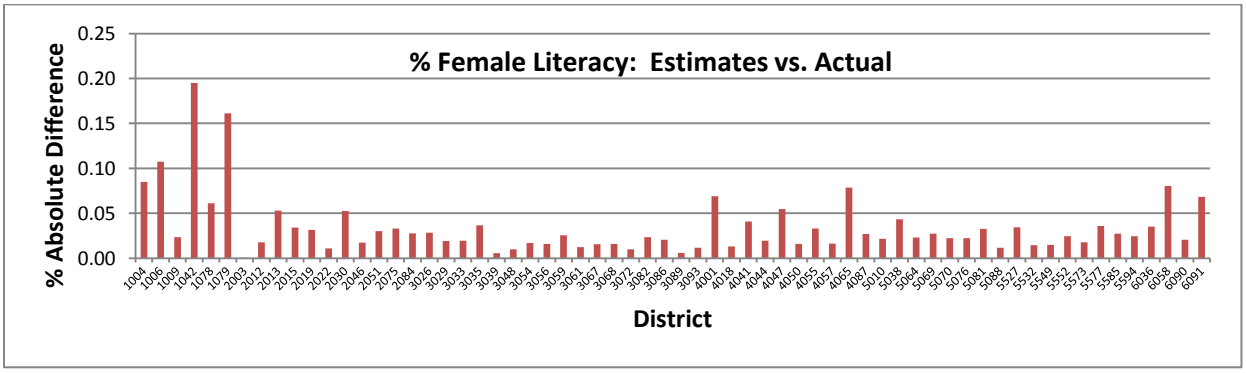

Figure 2.3. Absolute percent difference between P-MEDM estimates and actual District level data taken from the Census 2011 data for female literacy.

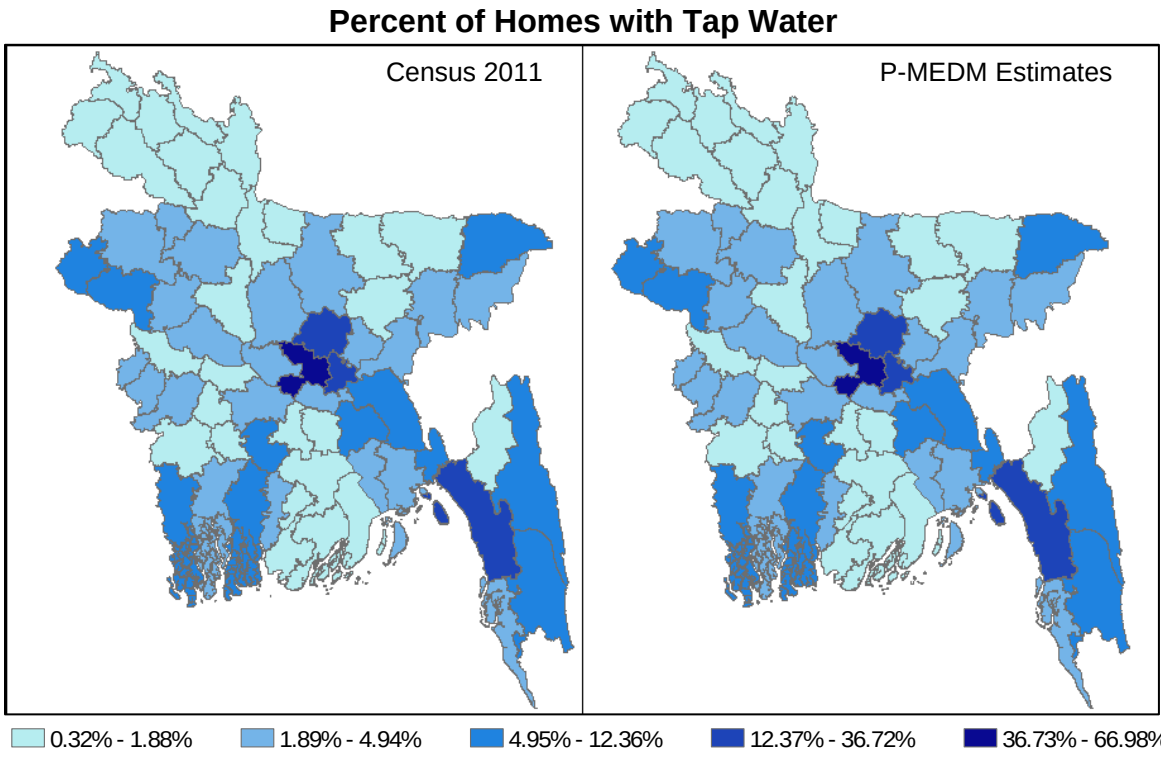

Figure 2.4. Percent of homes with tap water source from Census 2011 as compared to percent of homes with tap water source estimated by the P-MEDM procedure at the District level. Mapped results show that the estimates match well to the Census data.

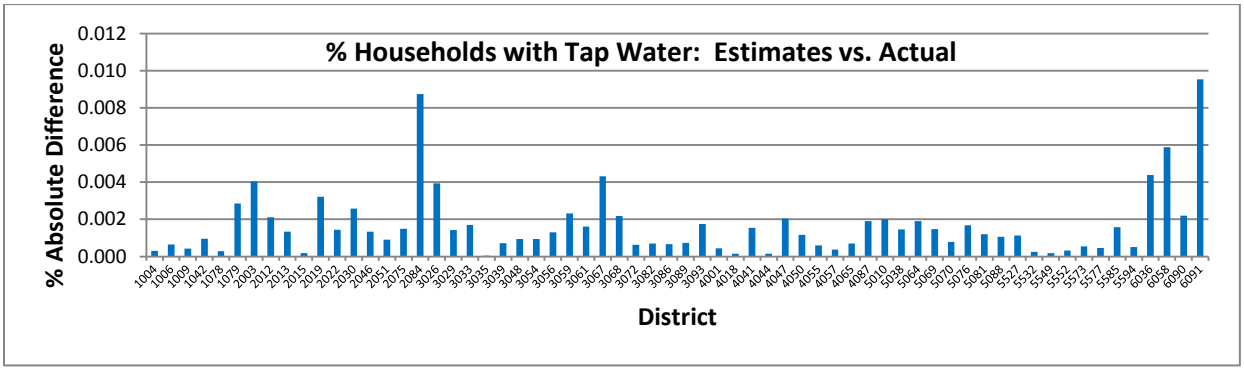

Figure 2.5. Absolute percent difference between P-MEDM estimates and actual District level data taken from the Census 2011 data for tap water source.

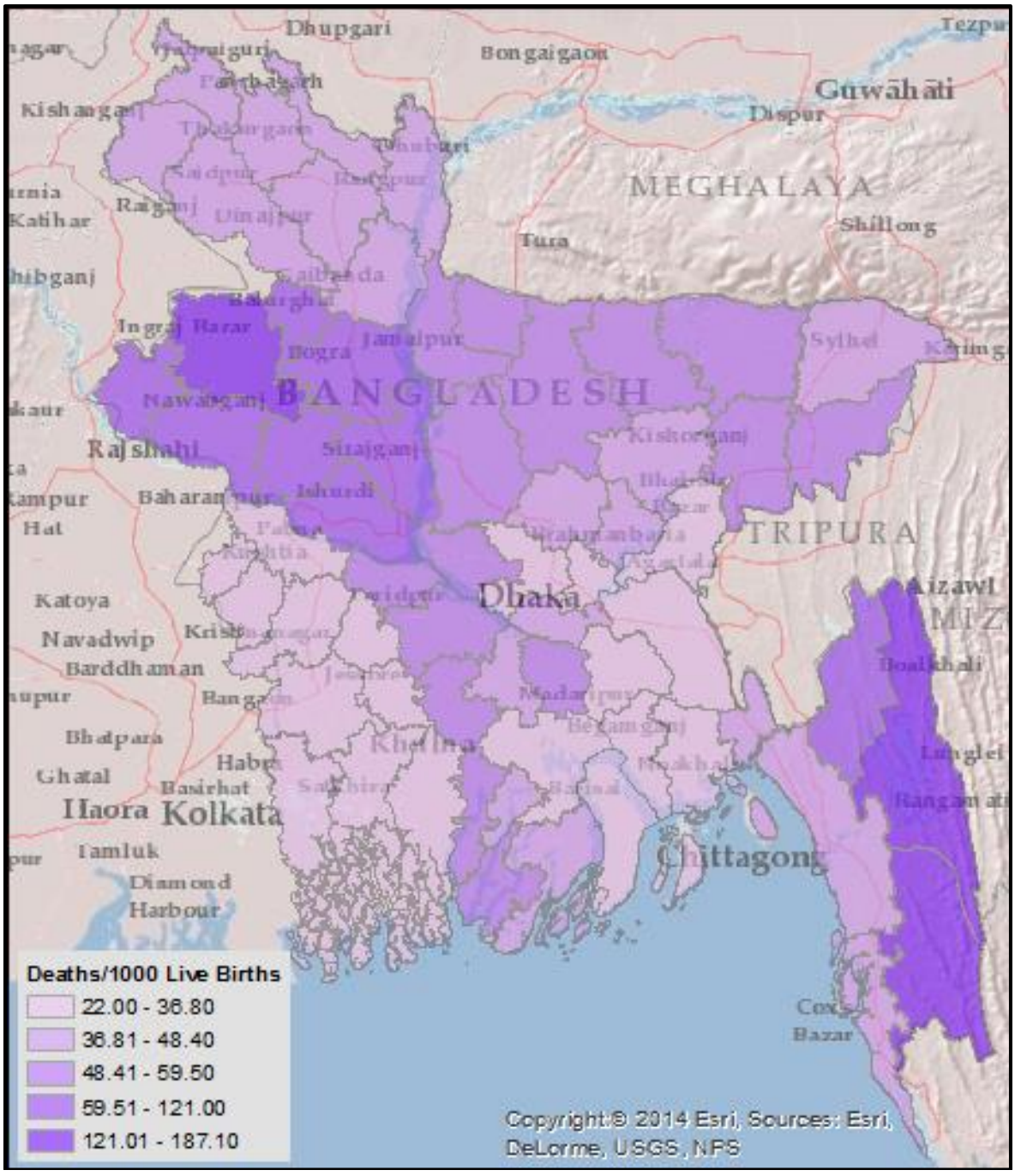

Figure 2.6. Infant mortality rate as calculated from new P-MEDM small area estimates.

Conclusion

Overall, the results of this exploratory work show promise in terms of improving the demographic resolution of spatially explicit population datasets for large geographic areas. Preliminary visual inspection of estimates produced through the P-MEDM procedure show that non-constraining variables mapped quite well to the same variables as reported in the Census data. Furthermore, infant mortality, a variable that is not collected as part of the Bangladesh Census, can now be mapped at the subnational level.

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 detail 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.

The results suggest that implementing this process for other areas is a viable path forward considering that there are a number countries where both census and survey data, whether DHS or otherwise, are available. However, there is a clear need to perform robust validation procedures to fully evaluate the reliability of the P-MEDM estimates. 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, 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 and Tanton 2013, Ruther et al. 2013, Williamson, Birkin, and Rees 1998). In this sense, this exploratory work can guide future research, which must include reasonable approaches for validation, particularly in data poor environments.

References

Balk, D., Pullum, T., Storeygard, A., Greenwell, F., & Neuman, M. (2004). A spatial analysis of childhood mortality in West Africa. *Population, Space and Place, 10*(3), 175-216.

Ballas, D., Clarke, G., Dorling, D., Eyre, H., Thomas, B., & Rossiter, D. (2005). SimBritain: a spatial microsimulation approach to population dynamics. *Population, Space and Place, 11*(1), 13-34.

Ballas, D., Clarke, G., & Turton, I. (1999). *Exploring Microsimulation Methodologies for the Estimation of Household Attributes*. Paper presented at the 4th International conference on GeoComputation, Virginia, USA.

Ballas, D., Clarke, G. P., & Wiemers, E. (2006). Spatial microsimulation for rural policy analysis in Ireland: The implications of CAP reforms for the national spatial strategy. *Journal of Rural Studies, 22*(3), 367-378.

Beckman, R. J., Baggerly, K. A., & McKay, M. D. (1996). Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice, 30*(6), 415-429.

Birkin, M., & Clarke, M. (1988). SYNTHESIS - a synthetic spatial information system for urban and regional analysis: methods and examples. *Environment and Planning A, 20*, 1645–1671.

Center for Policy Dialogue (CPD) 2003. "Bangladesh's Population Policy: Emerging Issues and Future Agenda." Retrieved October 4, 2014, from http://cpd.org.bd/pub_attach/unfpa23.pdf.

Chin, B., Montana, L., & Basagana, X. (2011). Spatial modeling of geographic inequalities in infant and child mortality across Nepal. *Health Place, 17*(4), 929-936.

Chin, S.-F., & Harding, A. (2006). Regional Dimensions: Creating Synthetic Small-area Microdata and Spatial Microsimulation Models. *Online Technical Paper - TP33*.

Cui, X., Kim, H. K., Liu, C., Kao, S.-C., & Bhaduri, B. L. (2012). Simulating the household plug-in hybrid electric vehicle distribution and its electric distribution network impacts. *Transportation Research Part D: Transport and Environment, 17*(7), 548-554.

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, 11*(4), 427-444.

Edwards K. L., & Tanton, R. (2013). *Spatial Microsimulation: A Reference Guide for Users*: Springer.

Elbers, C., Lanjouw, J. O., & Lanjouw, P. (2003). Micro–Level Estimation of Poverty and Inequality. *Econometrica, 71*(1), 355-364.

Fabic, M. S., Choi, Y., & Bird, S. (2012). A systematic review of Demographic and Health Surveys: data availability and utilization for research. *Bulletin of the World Health Organization, 90*, 604-612.

Fotheringham, A. S., & Wong, D. W. S. (1991). The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A, 23*(7), 1025-1044.

Gemperli, A., Vounatsou, P., Kleinschmidt, I., Bagayoko, M., Lengeler, C., & Smith, T. (2004). Spatial Patterns of Infant Mortality in Mali: The Effect of Malaria Endemicity. *American Journal of Epidemiology, 159*(1), 64-72.

Harding, A., Lloyd, R., Bill, A., & King, A. (2004). Assessing Poverty and Inequality at a Detailed Regional Level: New Advances in Spatial Microsimulation: World Institute for Development Economic Research (UNU-WIDER).

Johnston, R. J., & Hay, A. M. (1983). Voter Transition Probability Estimates: An Entropy-Maximizing Approach. *European Journal of Political Research, 11*(1), 93-98.

Johnston, R. J., & Hay, A. M. (1984). The geography of ticket-splitting: a preliminary study of the 1976 elections using entropy-maximizing methods. *Professional Geographer, 36*(2), 201-206.

Johnston, R. J., A.M. Hay, and D. Rumley (1983). Entropy-Maximizing Method for Estimating Voting Data: A Critical Test. *Area 15,* 35-41.

Johnston, R. J., A. M. Hay, and D. Rumley (1984). On Testing for Structural Effects in Electoral Geography, Using Entropy-Maximizing. *Environment and Planning A, 16*, 233-40.

Johnston, R. J., & Pattie, C. J. (1993). Entropy-maximizing and the iterative proportional fitting procedure. *Professional Geographer, 45*(3), 317.

Kazembe, L. N., & Namangale, J. J. (2007). A Bayesian multinomial model to analyse spatial patterns of childhood co-morbidity in Malawi. *Eur J Epidemiol, 22*(8), 545-556.

Leyk, S., Buttenfield, B.P., and Nagle, N. (2013). Modeling ambiguity in Census microdata allocations to improve demographic small area estimates. *Transactions in Geographic Information Science* 17(3): 406-425.

Leyk, S., Nagle, N., & Buttenfield, B.P. (2013). Maximum entropy dasymetric modeling for demographic small area estimation. *Geographical Analysis* 45(3), 285-306.

Mansour, S., Martin, D., & Wright, J. (2012). Problems of spatial linkage of a geo-referenced Demographic and Health Survey (DHS) dataset to a population census: A case study of Egypt. *Computers, Environment and Urban Systems, 36*(4), 350-358.

Moehl, J. J. (2014). *Comparing Models of Demographic Subpopulations*. (Master of Science), The University of Tennessee, Knoxville, TN.

Nagle, N.N., Buttenfield, B.P., Leyk, S., & Spielman, S.E. (2012). *An uncertainty-informed penalized maximum entropy dasymetric model*. Presented at the 7th International Conference on Geographic Information Science (GIScience 2012), Columbus, OH, September 18-21, 2012.

Nagle, N. N., Buttenfield, B. P., Leyk, S., & Spielman, S. (2014). Dasymetric Modeling and Uncertainty. *Annals of the Association of American Geographers, 104*(1), 80-95.

National Institute of Population and Training, Mitra and Associates, & Measure DHS (2013). Bangladesh DHS, 2011 - Final Report. Retrieved September 24, 2014, from http://dhsprogram.com/publications/publication-FR265-DHS-Final-Reports.cfm.

Norman, P. (1999). *Putting Iterative Proportional Fitting on the Researcher's Desk*. University of Leeds. (Unpublished).

O'Donoghue, C., Ballas, D., Clarke, G., Hynes, S., & Morrissey, K. (2013). *Spatial Microsimulation for Rural Policy Analysis*: Springer.

Openshaw, S. (1984) The Modifiable Areal Unit Problem. Vol. 38. CATMOG - Concepts and Techniques in Modern Geography. Norwich: Geo Books.

Ruther, M., Maclaurin, G., Leyk, S., Buttenfield, B., & Nagle, N. (2013). Validation of spatially allocated small area estimates for 1880 Census demography. *Demographic Research, 29*, 579-616.

Simpson, L., & Tranmer, M. (2005). Combining Sample and Census Data in Small Area Estimates: Iterative Proportional Fitting with Standard Software. *The Professional Geographer, 57*(2), 222-234.

Soares Magalhaes, R. J., & Clements, A. C. (2011). Mapping the risk of anaemia in preschool-age children: the contribution of malnutrition, malaria, and helminth infections in West Africa. *PLoS Med, 8*(6), e1000438.

Taylor, E., Harding, A., Lloyd, R., & Blake, M. (2004). Housing unaffordability at the statistical local area level: new estimates using spatial microsimulation. *Australasian Journal of Regional Studies, 10*(3), 279-300.

Tobler, W. (1979). Smooth Pycnophylactic Interpolation for Geographical Regions. *Journal of the American Statistical Association, 74*(367), 519-530.

Vidyattama, Y., & Tanton, R. (2010). Projecting small area statistics with Australian spatial microsimulation model (SpatialMSM). *Australasian Journal of Regional Studies, 16*(1), 99+.

Williamson, P., Birkin, M., & Rees, P. H. (1998). The estimation of population microdata by using data from small area statistics and samples of anonymised records. *Environment and Planning A, 30*(5), 785-816.

Wilson, A. G. (1971). A family of spatial interaction models, and associated developments. *Environment and Planning, 3*(1), 1-32.

Wong, D. W. S. (1992). The Reliability of Using the Iterative Proportional Fitting Procedure. *Professional Geographer, 44*(3), 340-348.

CHAPTER III
VALIDATION OF SPATIODEMOGRAPHIC ESTIMATES PRODUCED THROUGH DATA FUSION METHODS

Abstract

Data fusion methods, or more specifically, spatial microsimulation models, can be used to estimate previously unknown data at the small area level. However, validating the output of these models can be problematic and is frequently overlooked in the literature. Without proper vetting of synthetic small area estimates, these data cannot be used in research and analysis with confidence. The principal barrier to validating small area estimation models is that it is rarely possible to directly validate the output variables since they are likely being estimated due to the fact that these data do not already exist. This chapter describes both internal and external validation approaches for synthetic microdata. Estimates of individual and household level attributes for Bangladesh were produced through a variation of an iterative proportional fitting method called P-MEDM. Using this method, Census 2011 data and the 2011 Demographic and Health Survey for Bangladesh were linked and new small area estimates of household attributes were produced. Internal validation was conducted to 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. The P-MEDM estimates were also compared to 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. Although the results suggest that implementing this process for other areas is viable, the validation uncovered potential overfitting of the model indicating that additional modeling and validation work is needed.

Introduction

In any modeling effort, validation is both a requirement and a challenge. Model outputs are useless to researchers, planners, and policymakers if those outputs are not reasonable representations of the real world. Validation of a model of any sort requires a statistical method by which to validate as well as benchmark data to validate against. However, recent literature dealing with synthetic microdata highlights that validation of these data is still a shortcoming (Ballas and Clarke 2001; Birkin 2013; Edwards and Tanton 2013; Morrissey and O'Donoghue 2013; Ruther et al. 2013; Williamson, Birkin, and 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 ways to approach validation of small area estimation results. The first is to internally validate the modeled estimates against the actual data; either the constraint variables used in the model or non-constraining variables if they are available. The second is to perform external validation of the modeled estimates by comparing them to a data source that was not used in the model. Depending on the model and available data, sometimes it’s only possible to perform internal validation, but attempts should be made to externally validate modeled estimates if possible.

In a previous study, certain household and population characteristics from the 2011 Bangladesh Census were used as margins for which to scale data from the Bangladesh Demographic and Health Survey (DHS) to using the penalized maximum entropy model (P–MEDM) as developed by Nagle et al. (2014). The motivation for and process of building that model, including selection of constraint variables, are discussed elsewhere. However, what was not covered in that work was the validation of the estimates produced by the P-MEDM process. To address this, the following sections will cover both the internal validation and external validation of the estimates that were produced for the Bangladesh case study.

Background

In order to develop new estimates that are accurate, strong correlations must be present between the input variables and the variable of interest. Internal consistency of the model can be assessed through evaluation of constraint variables, and by comparing model performance when experimenting with varying non-constraining variables (Melhuish, Blake, and Day 2002; Morrissey and O’Donoghue 2013; Ruther et al. 2013).

Although a critical part of developing synthetic microdata is internal and external validation processes, there is no single goodness of fit measure or validation procedure that can be applied to all models (Williamson, Birkin, and Rees 1998). Some measures of model performance emphasize absolute differences between the estimates and actual data, while others assess relative distributions. As a result, synthetic estimates may appear to have a good fit using one measure but not another. Also, the model fit may be good overall, but deficient for a particular variable or small area. Assessing the goodness of fit is further complicated by the fact that the tables of observed and expected results are often sparse, where counts are at or close to zero (Beckman, Baggerly, and McKay 1996; Wong

1992). Previous endeavors (Birkin and Clarke 1988; Williamson 1992) have used the z-statistic to test whether differences between the actual and estimated values are significantly different. However, because the z-statistic is calculated using proportional rather than absolute differences, without modification it can produce misleading results.

As stated, external validation of synthetic microdata is a particularly challenging endeavor. In the absence of actual "true" values to compare the estimates to, other methods must be employed which provide some level of confidence that estimates are reasonable. One method that has been used in other studies is spatial aggregation (Edwards and Clarke 2013; Edwards et al. 2011; Farrell, Morrissey, and O'Donoghue 2013; Morrissey, Clarke, and O'Donoghue 2013; Tanton et al. 2011). That is, aggregating the estimated values to a coarser geographic level at which comparable data is available. For example, suppose data on literacy were available in the census at the first administrative unit level (i.e. provinces or states), but not at lower levels. In this case, literacy could be estimated at a lower geographic level (i.e. districts or counties), and then aggregated to the higher administrative level and compared to the data from the census. Ideally, the margins used to create the estimates and the data used for comparison would not be from the same source. This procedure can be used to evaluate whether the new estimates are generally correct, but cannot assess the accuracy of the spatial distribution at the small area level.

When an exogenous dataset that is temporally consistent with the new estimates is not available to use as a benchmark, backcasting techniques can be used to confirm that a model has produced reasonably good estimates (Birkin 2013; Birkin and Malleson 2010). In general, to evaluate the robustness of the estimation procedure, estimates are created going back in time, as opposed to the current state or even future state. These estimates can then be compared to a known historical dataset. A variation of backcasting validation was performed by Ruther et al. (2013), where the entire estimation method was based on and then validated against the full count population from the 1880 U.S. Census, which is available at a fine spatial resolution since it privacy is no longer a concern for this historical dataset. The major criticism of backcasting is that even if the comparison to historical data shows the estimates to be of good quality, there is no guarantee that the same process that produced those estimates could produce reasonable estimates for a current or future system.

In addition to validation of the absolute estimate values, the spatial distribution of the estimates should be considered as well. As tested by Johnston and Pattie (1991), IPF procedures produce reliable estimates of the relative spatial variation of the estimated variables; that is, absolute values may vary considerably, but relative values (and therefore relative variation) do not. This is an important consideration when the goal is to spatially disaggregate known data.

Output of the P-MEDM Algorithm for Bangladesh

Infant mortality rates were estimated at the small area level for the entirety of Bangladesh using the penalized maximum entropy dasymetric model (P–MEDM) developed by Nagle et al. (2014). Household and population characteristics from the 2011 Bangladesh Census were used as margins for which to scale data from the Bangladesh Demographic and Health Survey (DHS) in order to create new household weights, which in turn could be used to create new estimates. Each Division (administrative level 1 unit) was modeled as a separate process to constrain the results geographically. Household-level estimates within a Division were produced across Districts (administrative level 2 unit) so that a specific household representation was imputed for each District as varying weights. This means that a single household from the survey data may 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. This allows for reasonable estimates to be produced even when samples within a District are sparse, by borrowing from similar areas (Simpson and Tranmer 2005).

Initially, all variables having some correlation to infant mortality rate were used in the P-MEDM process. However, this introduces a significant issue when attempting internal validation. That is, by using all of these variables to constrain on, the model estimates should be very close to the margin totals. To account for this, performance of the model was evaluated by holding out some variables from the P-MEDM; specifically female literacy and household water source. The model inputs and outputs were stratified by Division, District, and rural/urban designation. It's important to note that there are approximately three times the amount of rural householders/households as there are urban household members/households and nearly fifty percent are located in the Dhaka division (Table 3.1).

Table 3.1. The distribution of householders from the 2011 Bangladesh Census listed by rural and urban designation and by Division.

	Householders	**Households**
Rural	110,480,514	24,671,590
Urban	33,563,183	7,502,040
Division		
Barisal	8,325,666	1,862,841
Chittagong	28,423,019	5,626,310
Dhaka	47,424,418	10,849,315
Khulna	15,687,759	3,739,779
Rajshahi	18,484,858	4,486,829
Rangpur	15,787,758	3,817,664
Sylhet	9,910,219	1,790,892
Total	**144,043,697**	**32,173,630**

Internal Validation

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 will be conducted by using three separate measures. First, the error in margin will be evaluated for all variables included in the P-MEDM process, regardless of whether they were used as constraining variables or not. Next, the Standardized Allocation Error (SAE) will be calculated in order to compare the P-MEDM estimates to initial Census constraints. Finally, the new estimates will be compared to the original DHS estimates.

Error in Margin

The error in margin measure is a goodness of fit measure designed to assess the allocation of individual variables for the entirety of the study area (Ruther et al. 2013). To calculate the error in margin, the total number of households for a particular variable in the Bangladesh Census was compared to the total estimated households representing that same variable, calculated at the country level. The difference between these two household counts was normalized by the total number of households for that variable in the Census, and then converted to a percentage. Although the error in margin measure does not provide any information about the allocation at the District or Division level, any sizeable differences between the actual household counts and the P-MEDM estimated counts will pinpoint for which variables model did not perform well.

It is clear from simply looking at the error in margin analysis at the country level that a single variable – home ownership – had a large margin of error relative to the other variables. With regard to home ownership, the model overestimated the number of households that owned their home by approximately four million households or roughly 16% (Table 3.2). Looking further at the rural/urban breakdown of this variable, we see that the poor model estimates for home ownership are predominantly for urban households (Table 3.3). Critical inspection of the homeownership estimates in urban geographic areas shows that the largest cities in Bangladesh; Dhaka, Chittagong, Gazipur, Khulna, Narayanganj, and Sylhet; have by far the largest errors in margin.

Table 3.2. National level percent absolute error in margin of P-MEDM produced estimates as compared to margin totals taken from the 2011 Bangladesh Census. Female literacy and water source variables were not used as constraints.

Variable	Absolute Error in Margin (%)	Variable	Absolute Error in Margin (%)
Male	0.0052%	Female Literacy	0.0680%
Female	0.0052%	Fem Under15 School	0.0004%
1 person HH	0.0488%	Fem 15+ School	0.0032%
2 person HH	0.0163%	Fem Employ	0.0158%
3 person HH	0.0033%	Age0-4	0.0068%
4 person HH	0.0056%	Age5-9	0.0087%
5 person HH	0.0050%	Age10-14	0.0091%
6 person HH	0.0057%	Age15-19	0.0080%
7 person HH	0.0041%	Age20-24	0.0054%
8+ person HH	0.0117%	Age25-29	0.0002%
Tap Water	0.0262%	Age30-49	0.0010%
Tubewell Water	0.0054%	Age50-59	0.0022%
Other Water	0.1258%	Age60-64	0.0106%
Electricity	0.0020%	Age65+	0.0030%
Owns Home	15.6867%		

Table 3.3. Comparison of Census and P-MEDM estimates of households for the variable home ownership by rural and urban designation. Results show that the largest error is with the estimates for urban designated households.

	Census	P-MEDM Estimates	Difference	% Error in Margin
Rural	22,835,461.00	23,689,000.15	853,539.15	3.74
Urban	3,774,650.00	7,095,362.24	3,320,712.24	87.97

There are only six rural/District combinations where the homeownership rate falls below 90%. In comparison, the urban/District combinations showed homeownership rates ranging from a high of 90% to a low of 19%. Within this range, only 6 of these areas had a rate of less than 50%. These are also the areas that had a much higher error in margin as compared to other urban/District areas.

Figure 3.1. shows the relationship between the rate of homeownership and the error in margin for that variable by rural and urban classifications. As shown in Figure 3.2, the model performed well for all other variables, including female literacy and household water source which were not used as constraining variables. The highest error in margin was for the water source – "other" water variable. Still, at roughly 0.13%, this measurement is quite small.

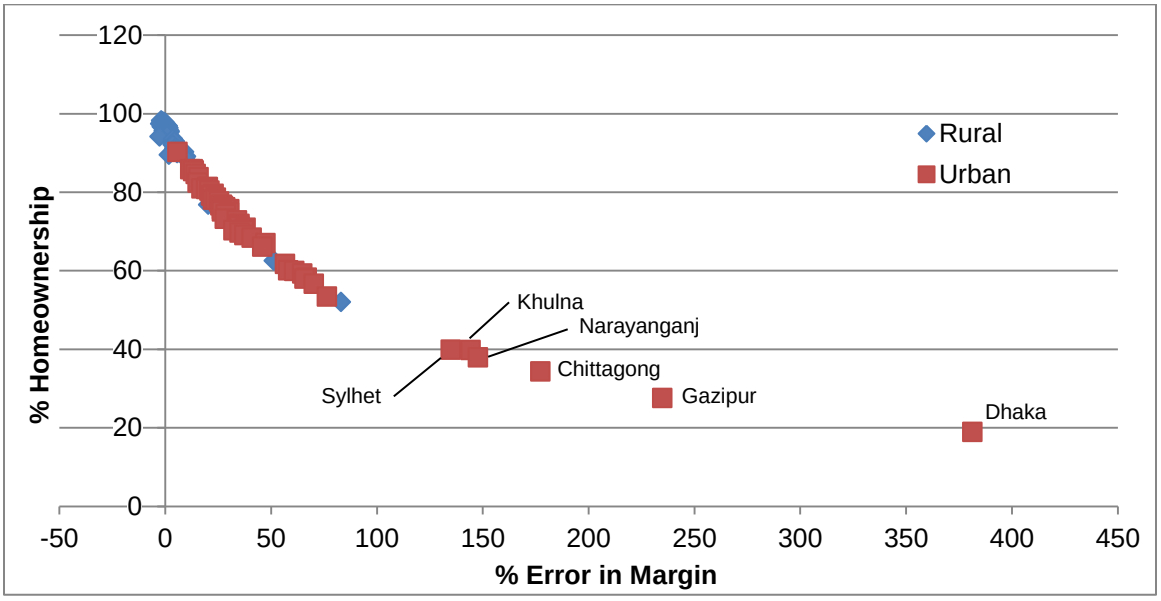

Figure 3.1. The relationship between the rate of homeownership and the error in margin for that variable by rural and urban classifications.

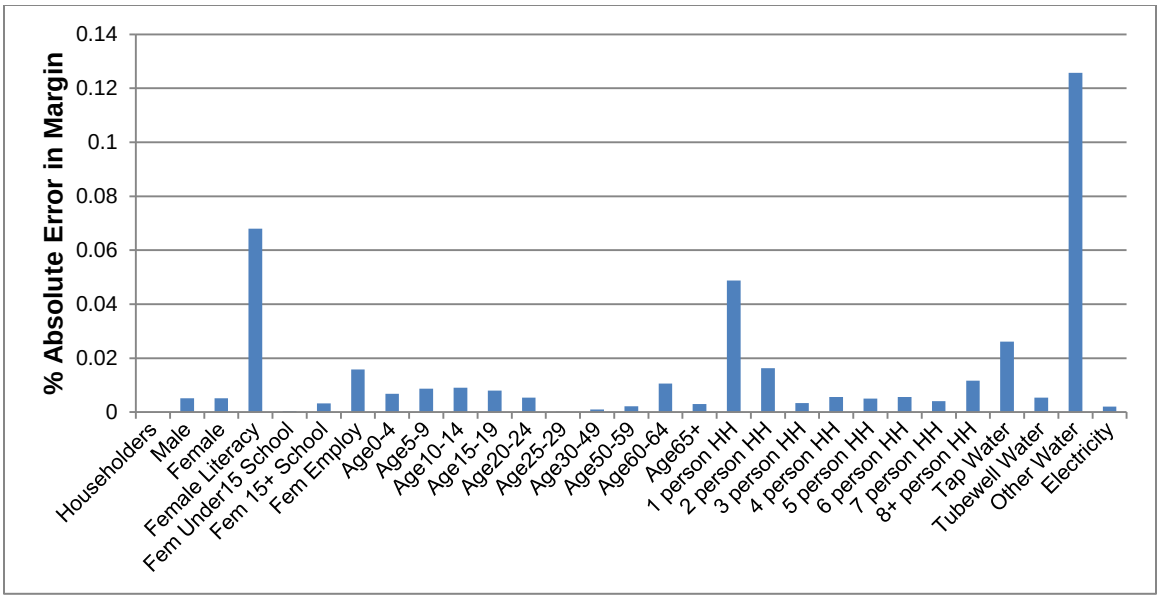

Figure 3.2. National level percent absolute error in margin of P-MEDM produced estimates as compared to margin totals taken from the 2011 Bangladesh Census. Home ownership variable is not shown.

Standardized Allocation Error

The standardized allocation error (SAE) (Anderson 2013; Ballas, Clarke, and Turton 1999; Ballas et al. 2005; Ruther et al. 2013; Williamson, Birkin, and Rees 1998) was used to compare the model allocations to the actual populations at multiple geographic aggregations. As opposed to the error in margin which only looks at how the model fit each variable for the entire study area, the SAE can be used to evaluate how well each variable was allocated over multiple geographic aggregations. This measure is referred to as the standardized absolute error in some literature, and in some cases evaluated as the total absolute error (TAE). Although TAE and SAE are very similar approaches, in the SAE the population for each variable is used as the denominator rather than the total population. This has the effect of more closely stating the size of the error of each variable. It should be clear that minimizing the difference between the actual values and the estimated values is the objective, yet other previous studies do not clearly indicate what the acceptable bounds of SAE values should be. Some studies suggest an SAE between -20% and 20% (or less than 20% if using absolute values), although the range may depend on the data (Ballas, Clarke, and Turton 1999; Ruther et al. 2013; Smith, Clarke, and Harland 2009). Since the SAE for

constraint variables should approach zero, here I 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 all non-constraining variables. However, there were some interesting results that emerged at the District level. These results are shown in Figure 3.3 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%. However, there is a single District level over-allocation for female literacy. Even though it's a small allocation error, since it's the only over-allocation for this variable it warrants further inspection. The Bandarban District in the Chittagong Division shows an overestimation of %0.0006, but looking specifically at the rural and urban difference, 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 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 then makes sense since the P-MEDM process draws on similar households in other Districts within a Division to arrive at the allocation of variables.

Although the SAE is useful for evaluating the difference between the model estimates and actual data, it doesn't provide any information as to whether these differences are statistically significant.

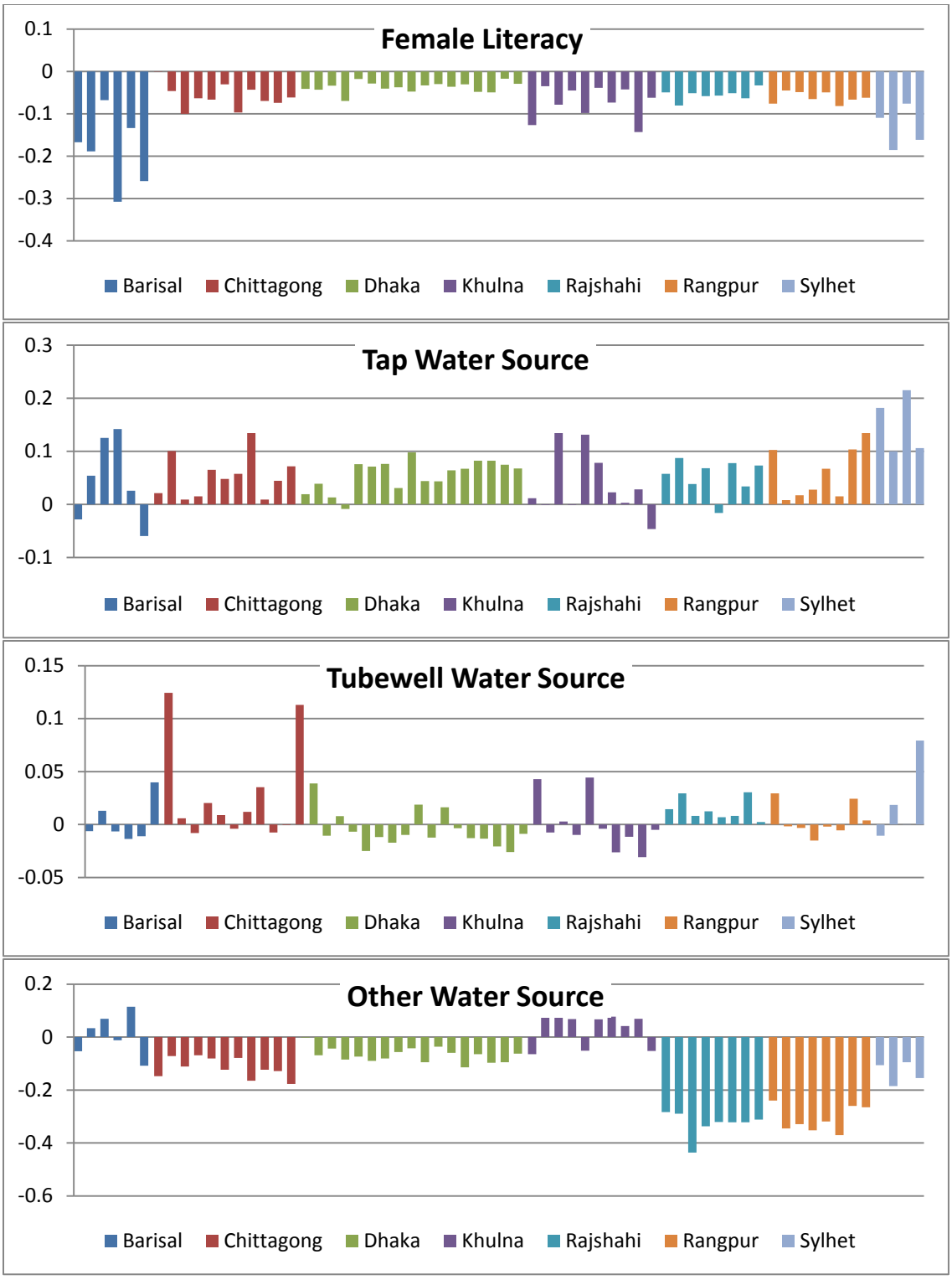

Figure 3.3. Standard allocation error at the District level for each of the non-constraining variables used in the P-MEDM. Division groupings are shown in varying colors.

Comparing P-MEDM Estimates to Scaled DHS Estimates

The Bangladesh DHS was designed to produce representative results for the country as a whole, for the urban and the rural areas separately, and for each of Bangladesh's seven administrative divisions. Since the P-MEDM produced estimates not only at these levels but also at lower geographic levels, I will compare the weighted DHS estimates for the country, divisions, districts, and urban/rural designations to the P-MEDM estimates at these same levels. What this should show is the estimates being comparable at the level for which the DHS estimates are representative, but not at lower levels. For this comparison, female literacy rate, households with tap water source, and households with electricity were used as the comparison variables since they had the most significant correlations with the variable of interest (infant mortality rate), as well as representative of both constraining and non-constraining variables in the model.

To scale the surveyed households to the total households for the geographic areas for which DHS samples were taken from, I used the total number of households from the 2011 Census for a given geographic area and then divided that by the sum of the DHS household weights for all DHS sample records within that same geographic area. This produced a scaling factor which was then multiplied by the original DHS household weight for each DHS household record:

$$\frac{\sum C_i}{\sum w_i} * \ w_i$$

Where C_i are the households in the Census for geographic area i, and w_i are the raw DHS household weights for geographic area i.

This created a new household weight for each DHS household record which could then be used to scale the DHS sample to the total number of households as reported in the 2011 Bangladesh Census.

DHS sample records were available for all districts, but not for all rural/urban designations within each district. For that reason, only P-MEDM estimates for district/rural/urban combinations where DHS samples existed were used in this comparison. There are 64 districts in Bangladesh, and when split based on rural or urban designation, results in a total of 128 geographic areas. Of these, 10 areas fell into the category of not having DHS samples available. Although according to the 2011 Census, all of these districts have areas classified as urban, they are all predominantly rural. Of these areas, three are located in the Chittagong Division, six in the Dhaka Division, and one in the Khulna Division.

Comparisons for national, division, and rural/urban designations, for which DHS estimates are representative, are shown in Figures 3.4 – 3.6. At all of these

levels, households with tap water as well as households with electricity are comparable for both sets of estimates. Female literacy estimate comparisons become more tenuous at increasingly lower geographic levels, although this can be generally attributed to the nearly 15% difference between the DHS estimates and P-MEDM estimates for urban areas. This urban-based variance is what percolates into the district level female literacy estimates.

As expected, Figure 3.7 shows that the DHS and P-MEDM estimates are not comparable at the district level. Even though we know that the DHS estimates are not necessarily representative at this level, this comparison cannot offer explanation as to whether the P-MEDM estimates are reasonably representative either. The comparison is simply used to show that the P-MEDM algorithm produced estimates that are different from the DHS estimates, which are known to not be representative at the district level. In order to assess whether these new estimates are an acceptable depiction of the true ground condition, an exogenous dataset must be employed as a comparison benchmark.

The following sections will describe the external validation process whereas the new estimates will be used to calculate the infant mortality rate for each district. The results will be examined alongside the district level infant mortality as reported in the Bangladesh 2009 Multiple Indicator Cluster Survey.

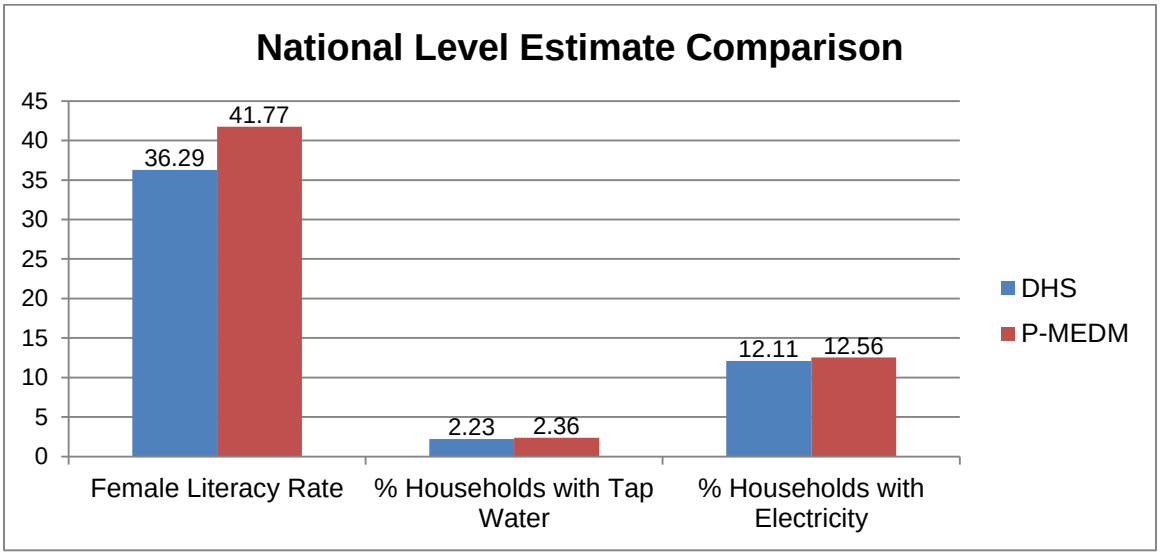

Figure 3.4. Comparison of DHS raw estimates and P-MEDM reweighted estimates for three model variables for Bangladesh.

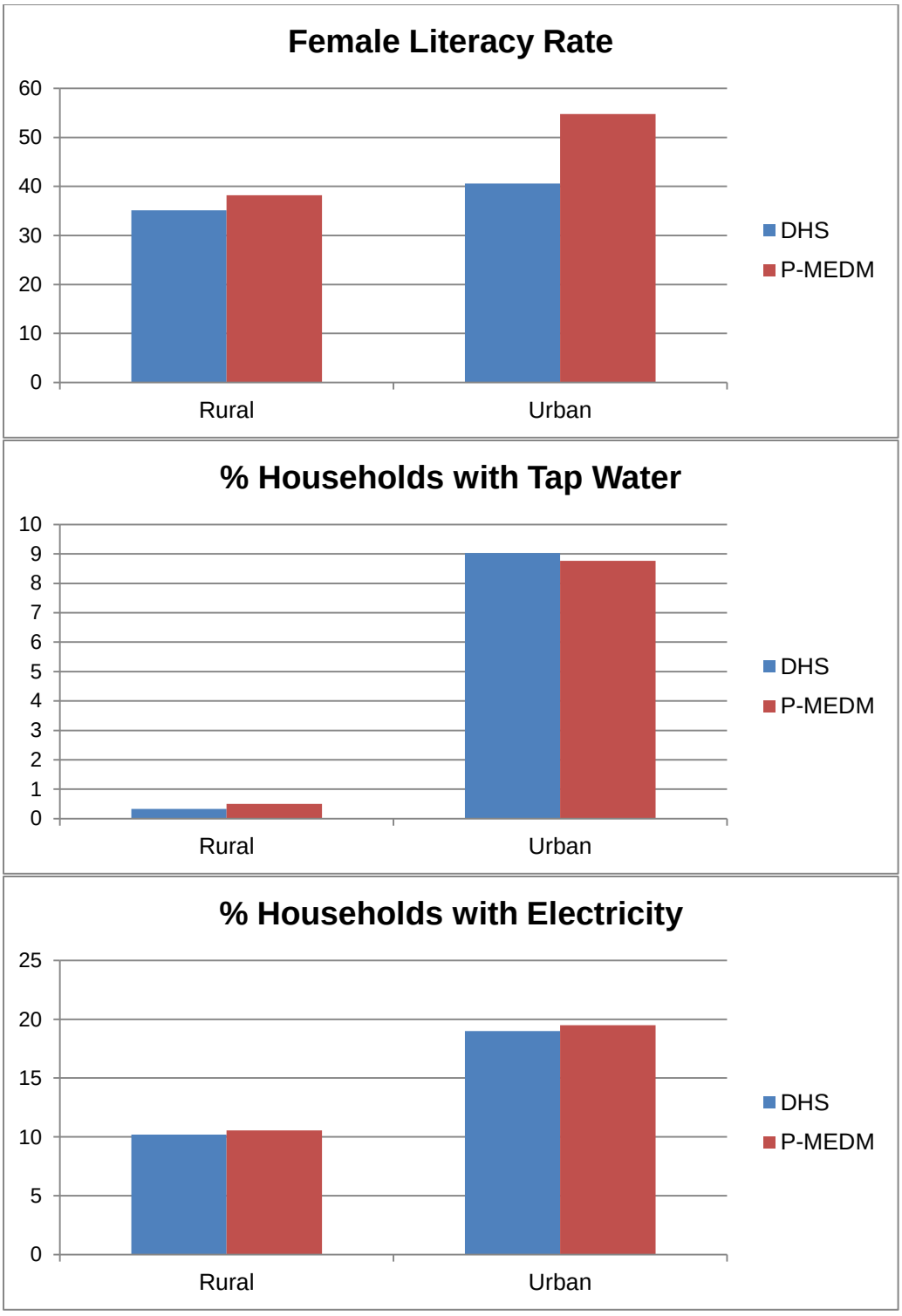

Figure 3.5. Comparison of DHS raw estimates and P-MEDM reweighted estimates for three model variables for rural and urban designations.

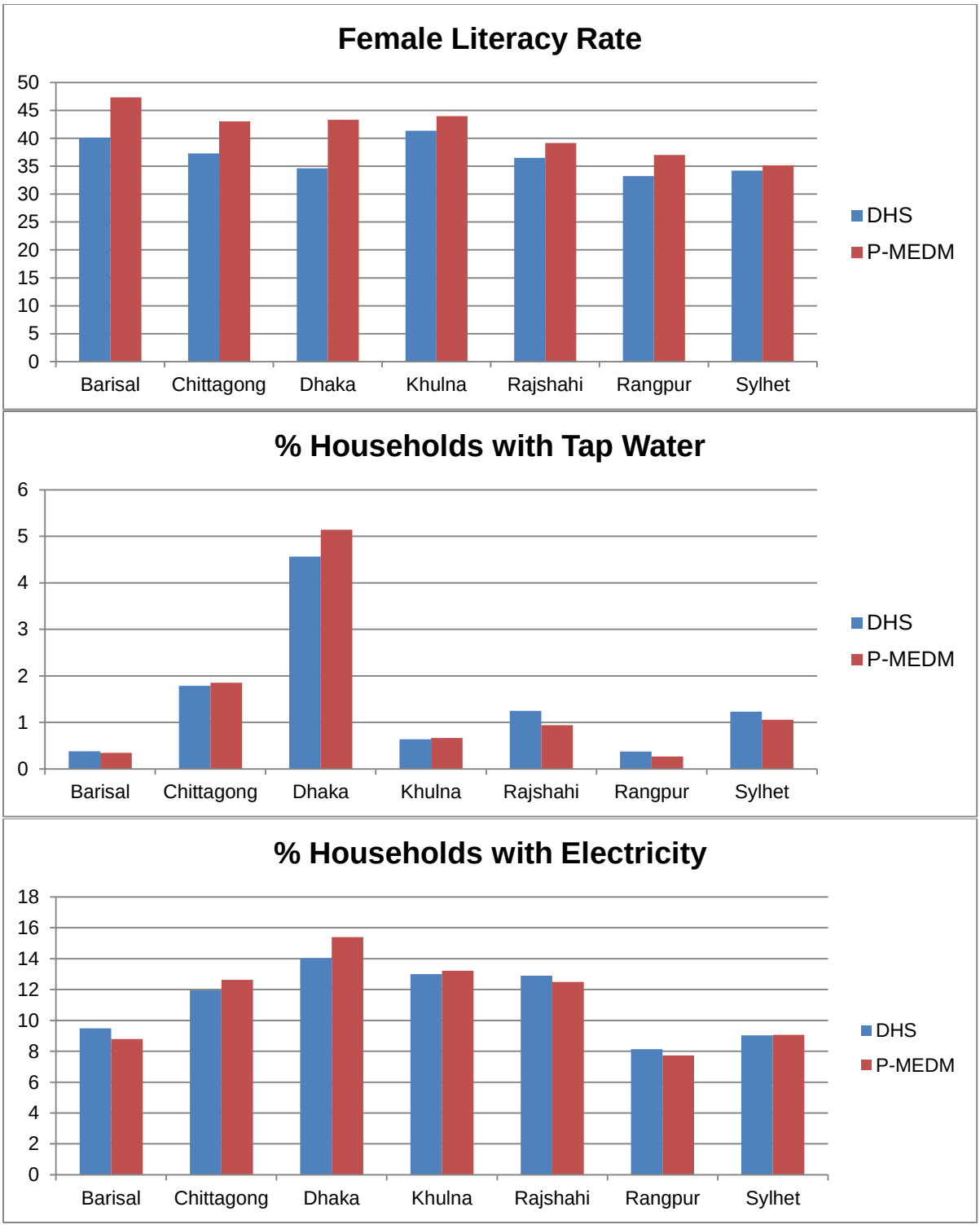

Figure 3.6. Comparison of DHS raw estimates and P-MEDM reweighted estimates for three model variables at the Division level.

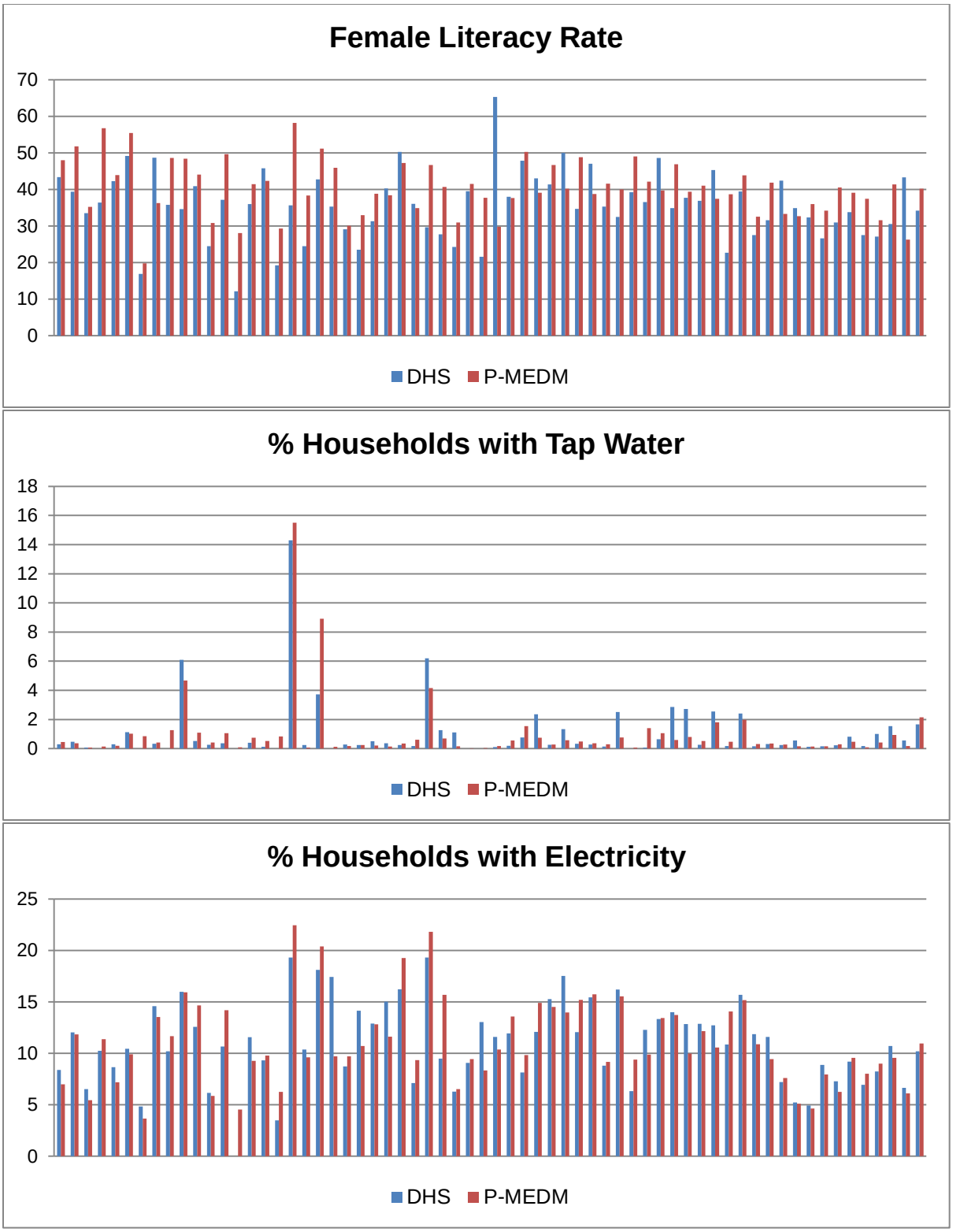

Figure 3.7. Comparison of DHS raw estimates and P-MEDM reweighted estimates for three model variables at the District level.

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 my purposes, the Multiple Indicator Cluster Survey (MICS) estimates will be used as if they are the actual observed values in order to compare the PMEDM estimates and the DHS raw estimates.

Multiple Indicator Cluster Survey (MICS)

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, known as Progotir Pathey, 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% (MICS 2010).

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 exiting 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.

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 1,000 live births for ages 0 – 11 months. The IMR was calculated using the synthetic cohort method employed by DHS for survey final reports (Rutstein and 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 date.[1] 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 (pg. 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 (Qfive) 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.

[1] To calculate IMR, modifications were made to Stata code originally written by Dr. Thomas Pullum, Director of Research, Measure DHS Project, to calculate mortality rates for Nigeria 2006 DHS.

Results

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 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 $N * p * (1 - p)$, 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 (Figure 3.8) 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, Kahgrachhari, and Rangamati; are the three Chittagong Division areas that show up on in Figure 3.8 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 than 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).

From these initial results, the variance for the model run solution shown in Figure 3.8 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 Figure 3.9. 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 3.4. 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.

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.

Table 3.4. Root mean squared error (RMSE) for P-MEDM estimates and DHS estimates for IMR as compared to the reported IMR from the MICS 2009.

	RMSE		
	All Districts	3 Districts Held Out	3 District + Rajshahi Held Out
P-MEDM ($\sigma^2 = 5$)	34.62	24.75	11.72
P-MEDM ($\sigma^2 = 10$)	29.32	18.65	10.86
P-MEDM ($\sigma^2 = 20$)	20.91	11.44	9.91
DHS	19.46	19.17	19.02

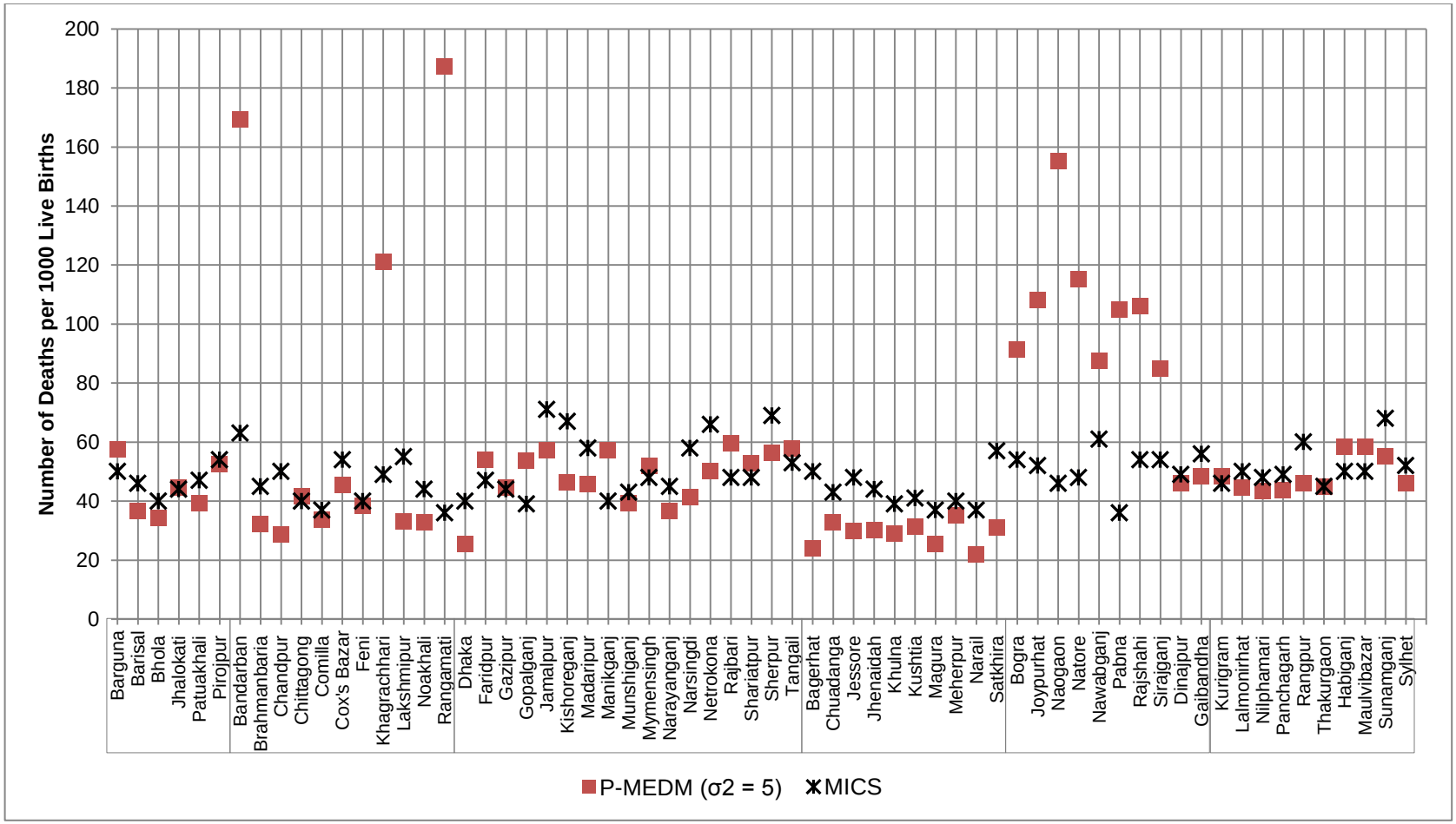

Figure 3.8. District level infant mortality rates from the Bangladesh MICS 2009, and as calculated for the P-MEDM estimates. Districts are grouped by Division.

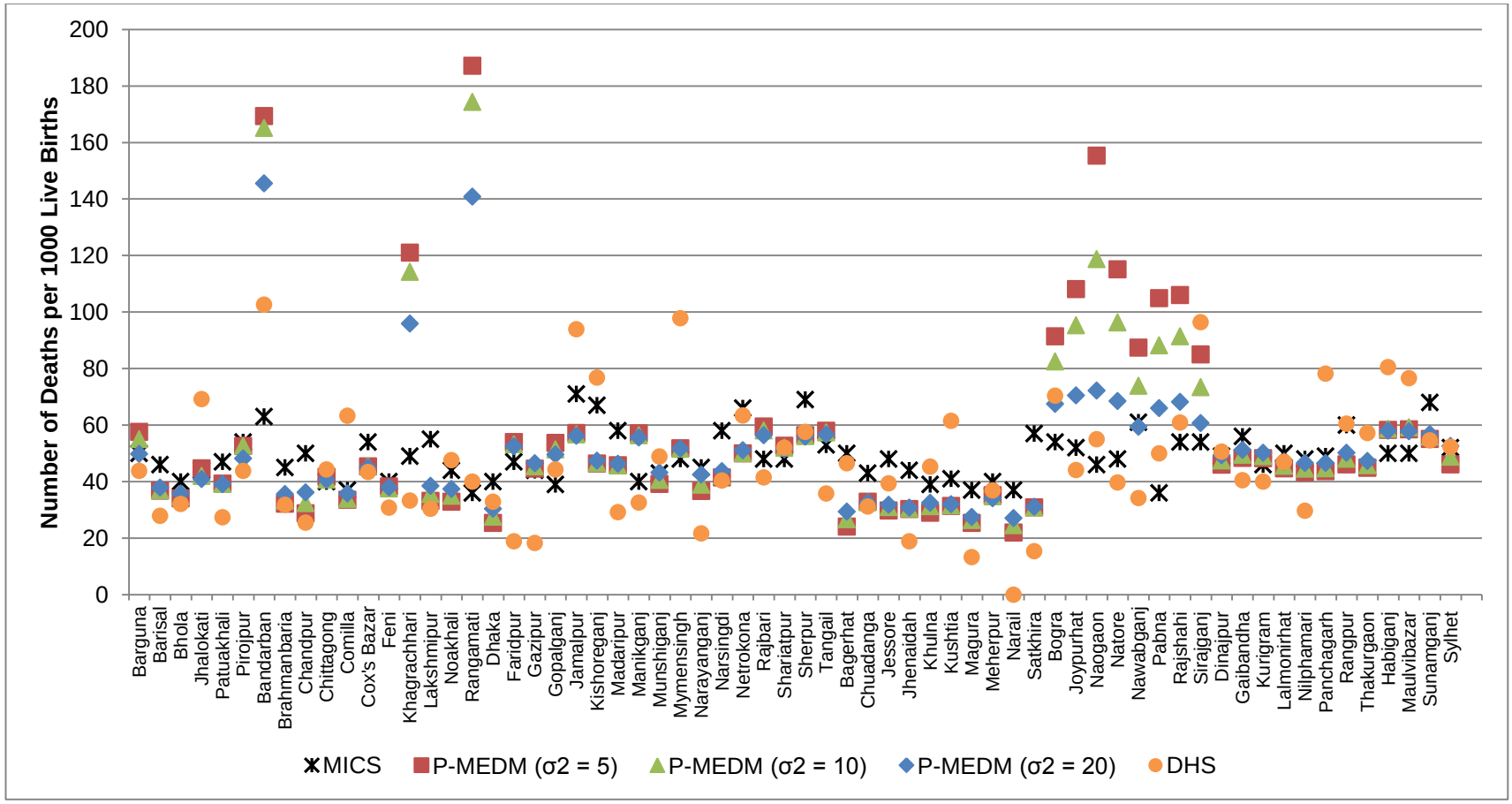

Figure 3.9. District level infant mortality rates from the Bangladesh MICS 2009, as calculated for the raw DHS estimates, and as calculated for the P-MEDM estimates at three different levels of variance.

Conclusion

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.

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 MICS 2006 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, 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.

References

Anderson, B. (2013). Estimating Small-Area Income Deprivation: An Iterative Proportional Fitting Approach. In R. Tanton & K. Edwards (Eds.), *Spatial Microsimulation: A Reference Guide for Users* (Vol. 6, pp. 49-67): Springer Netherlands.

Ballas, D., & Clarke, G. P. (2001). Modelling the local impacts of national social policies: a spatial microsimulation approach. *Environment and Planning C: Government and Policy, 19*(4), 587-606.

Ballas, D., Clarke, G., Dorling, D., Eyre, H., Thomas, B., & Rossiter, D. (2005). SimBritain: a spatial microsimulation approach to population dynamics. *Population, Space and Place, 11*(1), 13-34.

Ballas, D., Clarke, G., & Turton, I. (1999). *Exploring Microsimulation Methodologies for the Estimation of Household Attributes*. Paper presented at the 4th International conference on GeoComputation, Virginia, USA.

Beckman, R. J., Baggerly, K. A., & McKay, M. D. (1996). Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice, 30*(6), 415-429.

Birkin, M. (2013). Challenges for Spatial Dynamic Microsimulation Modelling. In R. Tanton & K. L. Edwards (Eds.), *Spatial Microsimulation: A Reference Guide for Users* (pp. 223-245). Dordrecht: Springer.

Birkin, M., & Clarke, M. (1988). SYNTHESIS -- a synthetic spatial information system for urban and regional analysis: methods and examples. *Environment and Planning A, 20*(12), 1645-1671.

Birkin, M., & Malleson, N. (2010). *An Investigation of the Robustness of a Dynamic Microsimulation Model of Urban Neighbourhood Dynamics*. Paper presented at the North American Regional Science Council (NARSC), Denver.

Brass, W. (1964). Uses of Census and Survey Data for the Estimation of Vital Rates. In: *African Seminar on Vital Statistics*. (Addis Ababa, Dec. 14–19, 1964).

Edwards, K. L., & Clarke, G. (2013). SimObesity: Cominatorial Optimisation (Deterministic) Model. In R. Tanton & K. L. Edwards (Eds.), *Spatial Microsimulation: A Reference Guide for Users* (pp. 69-86). Dordrecht: Springer.

Edwards, K. L., & Tanton, R. (2013). Validation of Spatial Microsimulation Models. In R. Tanton & K. L. Edwards (Eds.), *Spatial Microsimulation: A Reference Guide for Users* (pp. 249-258). Dordrecht: Springer.

Edwards, K. L., Clarke, G. P., Thomas, J., & Forman, D. (2011). Internal and External Validation of Spatial Microsimulation Models: Small Area Estimates of Adult Obesity. *Applied Spatial Analysis and Policy, 4*(4), 281-300.

Farrell, N., Morrissey, K., & O'Donoghue, C. (2013). Creating a Spatial Microsimulation Model of the Irish Local Economy. In R. Tanton & K. L. Edwards (Eds.), *Spatial Microsimulation: A Reference Guide for Users* (pp. 105-126). Dordrecht: Springer.

International Work Group for Indigenous Affairs (IWGIA). (2014). Indigenous peoples of the Chittagong Hill Tracts. Retrieved September 16, 2014, from http://www.iwgia.org/regions/asia/the-chittagong-hill-tracts-.

Johnston, R. J., & Pattie, C. J. (1991). Evaluating the use of entropy-maximising procedures in the study of voting patterns: 1. Sampling and measurement error in the flow-of-the-vote matrix and the robustness of estimates. *Environment and Planning A, 23*(3), 411-420.

Melhuish, T., Blake, M., & Day, S. (2002). An Evaluation of Synthetic Household Populations for Census Collection Districts Created Using Optimisation Techniques. *Australasian Journal of Regional Studies, 8*(3), 369-387.

MICS. (2010). Bangladesh: Monitoring the Situation of Children and Women. Progotir Pathey 2009. Multiple Indicator Cluster Survey, Volume I: Technical Report. Dhaka: Prepared by Bangladesh Bureau of Statistics and UNICEF.

Morrissey, K., Clarke, G., & O'Donoghue, C. (2013). Linking Static Spatial Microsimulation Modelling to Meso-scale Models: The Relationship Between Access to GP Services and Long-Term Illness. In R. Tanton & K. L. Edwards (Eds.), *Spatial Microsimulation: A Reference Guide for Users* (pp. 127-144). Dordrecht: Springer.

Morrissey, K., & O'Donoghue, C. (2013). Validation Issues and the Spatial Pattern of Household Income. In C. O'Donoghue, D. Ballas, G. Clarke, S. Hynes & K. Morrissey (Eds.), *Spatial Microsimulation for Rural Policy Analysis* (pp. 87-102). Heidelberg: Springer.

Nagle, N. N., Buttenfield, B. P., Leyk, S., & Spielman, S. (2014). Dasymetric Modeling and Uncertainty. *Annals of the Association of American Geographers, 104*(1), 80-95.

Rutstein, S. O., & Rojas, G. (2006). Guide to DHS Statistics: Demographic and Health Surveys Methodology. Calverton, Maryland: ORC Macro.

Ruther, M., Maclaurin, G., Leyk, S., Buttenfield, B., & Nagle, N. (2013). Validation of spatially allocated small area estimates for 1880 Census demography. *Demographic Research, 29*, 579-616.

Simpson, L., & Tranmer, M. (2005). Combining Sample and Census Data in Small Area Estimates: Iterative Proportional Fitting with Standard Software. *The Professional Geographer, 57*(2), 222-234.

Smith, D. M., Clarke, G. P., & Harland, K. (2009). Improving the synthetic data generation process in spatial microsimulation models. *Environment and Planning A, 41*(5), 1251-1268.

Tanton, R., Vidyattama, Y., Nepal, B., & McNamara, J. (2011). Small area estimation using a reweighting algorithm. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 174*(4), 931-951.

United Nations Children's Fund (UNICEF). (2010). Understanding Urban Inequalities in Bangladesh: A prerequisite for achieving Vision 2021. Bangladesh.

Voas, D., & Williamson, P. (2001). Evaluating Goodness-of-Fit Measures for Synthetic Microdata. *Geographical & Environmental Modelling, 5*(2), 177-200.

Williamson, P. (1992). *Community Health Care Policies for the Elderly: A Microsimulation Approach.* (Doctor of Philosophy), University of Leeds, Leeds.

Williamson, P., Birkin, M., & Rees, P. H. (1998). The estimation of population microdata by using data from small area statistics and samples of anonymised records. *Environment and Planning A, 30*(5), 785-816.

Wong, D. W. S. (1992). The Reliability of Using the Iterative Proportional Fitting Procedure. *Professional Geographer, 44*(3), 340-348.

CONCLUSION

Summary

The full body of this research presented the case for improving the demographic depth of existing population distribution datasets through data fusion methods. The current state of the art for global population distribution datasets was described in Chapter I, not merely to provide background on those datasets, but also to underline the fact that although a great deal of effort has been spent on improving the spatial and even temporal resolution of these datasets, little progress has been made in improving the demographic resolution.

Chapter II described a process by which Demographic and Health Survey data and Census data for Bangladesh were linked via a modified iterative proportional fitting algorithm (P-MEDM) developed by Nagle et al. (2014). Overall, the results of this exploratory work show promise in terms of improving the demographic resolution of spatially explicit population datasets for large geographic areas. Preliminary visual inspection of estimates produced through the P-MEDM procedure show that non-constraining variables mapped quite well to the same variables as reported in the Census data. Furthermore, infant mortality, a variable that is not collected as part of the Bangladesh Census, can now be mapped at the subnational level. 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. 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 detail 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.

The results suggest that implementing this process for other areas is a viable path forward considering that there are a number countries where both census and survey data, whether DHS or otherwise, are available, yet performing robust validation procedures is critical to fully evaluate the reliability of the P-MEDM estimates. Chapter III presented a full validation of the estimates produced in Chapter II. Internal validation revealed that the new estimates closely matched the original constraints for all variables, although home ownership had a large margin of error relative to the other variables. Further investigation showed that the poor model estimates for home ownership are predominantly for urban households specifically in the six largest urban areas in Bangladesh. 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 can create difficulties in producing reasonable 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.

Discussion

Despite the issues uncovered through the external validation, the results from the Bangladesh case study make it clear that there is potential for using the P-MEDM process to develop new small area microdata for other areas of the world. However, more work needs to be done in order to provide users with confidence in the resulting estimates. Since the eventual goal is to be able to scale this method globally, the process to some degree must approach a generalizable method. Although it was critical to identify issues including sensitivity to small sample sizes and input data uncertainty, the single case study of Bangladesh limits the amount of insight that can be used moving forward. Additional case studies, varying in geographic area, input data quality, and input data sources are a requirement before any additional judgment can be offered as to the widespread usefulness of this approach. It's possible that for many countries there will be small areas that cannot be reasonably estimated, whether due to unordinary demographic characteristics or small sample sizes. However, considering the potential that P-MEDM shows in data poor areas facing societal challenges, it's worthwhile to pursue this research.

Imagine the potential inquiries that could be conducted if researchers had more complete small area microdata available to them. In the case study presented in this research, infant mortality rate at the District level was produced via the P-MEDM process. These data were previously unavailable at this geographic

level, making it impossible for researchers to understand the geographic variation of infant mortality and gain insight into potential correlations with other socioeconomic indicators at a more granular level. The importance of geographic variation with regard to both cause and remediation of infant mortality cannot be overstated. Having demographic data at higher spatial resolutions can help identify local level triggers and targeted responses. A more focused (rather than global) approach to the issue of infant mortality would be a better use of resources as well as result in more effective outcomes. For example, UNICEF has supported persistent efforts at the local level over the last two decades to train community healthcare workers, resulting in sharp declines in both maternal and child mortality in Bangladesh. This type of targeted approach is viable only if the socioeconomic conditions and health status is known or can be well approximated for these small areas.

Future Research

Ultimately, the goal of this research was to provide a proof of concept that the P-MEDM method for data fusion could not only work in data poor environments, but also provide new data that could be applied to population distribution datasets. To meet that first goal, care was taken to evaluate estimates as-is rather than "fitting" them to whole number household counts. The process of implementing integer households from the estimates (Ballas et al. 2005) transforms weights into whole numbers. It is yet to be determined whether whole number households are the appropriate basis for which to link to population distribution datasets. However, it is expected that this type of integerization process would reduce variation within geographic areas. There was no methodological requirement for the research undertaken in Chapters II and III to allocate whole number households to each small area. Instead, fractional households were allowed so that the P-MEDM process could be more directly validated. Beyond this case study and validation there is a need for an in-depth investigation into the integerization of estimates with the purpose of linking household and individual demographic attributes to population distribution datasets.

A clear outcome of this dissertation research is that validation of new demographic estimates is a very difficult proposition. As shown in Chapter III, although internal validation may indicate a good model fit, the need for ways to reliably externally validate model estimates is still a gap that should be addressed in the very near term. This dissertation points out that validation of models and outputs of those models is often overlooked, attempted with minimal effort, or avoided altogether. Few very robust validation attempts in the realm of spatial microsimulation are documented in the literature. Those that are make clear that there are no standard methods for validation, but rather many methods that each provides a different measurement of quality and model fit.

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.

Final Thoughts

This research presented a case study for Bangladesh where there is arguably a lack of high resolution demographic data available. Even so, as shown in this research, the availability of census and survey data can allow researchers to fill the data gap where no high resolution spatiodemographic data exist. Although Bangladesh represents a data poor case study, there are other geographic areas throughout the world that not only haven't had a census conducted in recent times, but also do not have available survey microdata. Take for example, Somalia, Democratic Republic of the Congo (DRC), and Eritrea, all of which have not had a census in decades. Somalia has not had a population census conducted since 1987, and no plans to conduct one have been made due to ongoing conflict across the country. After Eritrea gained its independence, a national census was scheduled to take place in 1998 for the first time since 1984. However, continued conflict with Ethiopia caused it to be postponed. DRC experienced multiple waves of upheaval in the 1990's including political and social crises in the early part of the decade and armed conflicts in the latter part of the decade.

It's a catch 22 for places like Somalia, Eritrea, and DRC: a census is critical to assess the population composition and distribution in order to better provide resources such as infrastructure and social services. However, it is very difficult to mobilize technical, financial, and human resources necessary to conduct a census in post-conflict situations where humanitarian assistance still remains a priority. The intersection of this is no clearer than with the situation surrounding the current Ebola outbreak that threatens entire countries in west Africa. This outbreak has now reached DRC (www.who.int/csr/don/2014_09_10_ebola/en, last accessed 9/21/14), and has brought into sharp focus the limitations of spatial demographic information This current health crisis should serve as a strong reminder to the research community that in the truly global environment that we live in, we can no longer continue to address issues in spatial demography strictly in limited geographic extents. Future research in this area must also consider ways to address areas that do not have data resources, as these areas

are, in many cases, those that face the most critical needs in terms of health, social services, and infrastructure.

References

Ballas, D., Clarke, G., Dorling, D., Eyre, H., Thomas, B., & Rossiter, D. (2005). SimBritain: a spatial microsimulation approach to population dynamics. *Population, Space and Place, 11*(1), 13-34.

Nagle, N. N., Buttenfield, B. P., Leyk, S., & Spielman, S. (2014). Dasymetric Modeling and Uncertainty. *Annals of the Association of American Geographers, 104*(1), 80-95.

VITA

Amy Rose currently works as a Research Scientist in the Geographic Information Science & Technology (GIST) Group at Oak Ridge National Laboratory in Oak Ridge, TN (http://www.ornl.gov/sci/gist/index.shtml). Prior to joining the GIST Group, Amy worked in both public and private sector organizations implementing GIS technology in environmental, logistics, and transportation planning projects. She is currently co-developer of ORNL's LandScan Global Population Distribution Program. LandScan, a high-resolution global population distribution database, is the standard database used by the Department of Defense for estimating population at risk. In addition, she is the co-developer of LandScan USA, a very high-resolution population database for the U.S. including diurnal and demographic variations. Amy earned a Master of Science degree in Geography with a graduate minor in Logistics and Transportation from The University of Tennessee in 2001.