CREATING TEMPORALLY CONSISTENT SMALL AREA CENSUS UNITS USING ADVANCED COMBINATIONS OF AREAL INTERPOLATION AND SPATIAL REFINEMENT: METHOD DEVELOPMENT AND ASSESSMENT

by

HAMIDREZA ZORAGHEIN

B.S., Khajeh Nasir Toosi University of Technology (KNTU), 2008

M.A., Khajeh Nasir Toosi University of Technology (KNTU), 2011

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirement for the degree of

Doctor of Philosophy

Department of Geography

2017

This thesis entitled:

Creating Temporally Consistent Small Area Census Units Using Advanced Combinations of Areal Interpolation and Spatial Refinement: Method Development and Assessment
written by Hamidreza Zoraghein
has been approved for the Department of Geography

_____

Professor Stefan Leyk (Chair)

_____

Professor Leiwen Jiang

Date_____

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Zoraghein, Hamidreza (Ph.D., Geography)

Creating Temporally Consistent Small Area Census Units Using Advanced Combinations of Areal Interpolation and Spatial Refinement: Method Development and Assessment

Dissertation directed by Associate Professor Stefan Leyk

The U.S. decennial census is an invaluable source to guide demographic analysis. It enumerates demographic characteristics within different levels of geography to protect privacy. Small statistical units such as census tracts and blocks in different points in time are indispensable to analyze regional and local trends of demographic characteristics. However, the linkage between census demography and those geographies mandates that their boundaries change from one census year to another to reflect underlying population changes. This inconsistency complicates studies of micro-scale nuanced demographic processes. Previous research efforts have aggregated inconsistent census geographies to larger comparable units or used areal interpolation to transfer demographic attributes from geographies of one census year (source zones) to geographies of another (target zones). The former disrupts the required resolution for micro-scale analysis while the latter is susceptible to errors.

This dissertation contributes analytical solutions to the above-mentioned persistent problem in enumerated data, typically used in demography, health sciences or economics. It combines spatial (dasymetric) refinement with areal interpolation methods to increase their accuracy in population estimation over time. This combination leads to a more precise allocation of population, which results in more reliable modeling for different configurations of target zones.

The research conducts comprehensive analyses involving various ancillary variables, namely the National Land-Cover Database (NLCD), the Global Human Settlement Layer (GHSL), parcels, buildings and ZTRAX®, to transfer different demographic attributes, namely total population, population by race and age structure and urban population from census tracts in 1990 and 2000 within census tract boundaries in 2010 across different geographic scales (county/state) and under various demographic settings (urban/rural). This constructs demographic estimates within temporally consistent small units over 10- and 20-year periods.

The outcomes of the research affirm the effectiveness of combining spatial refinement with areal interpolation for accurate multi-temporal demographic analysis. The application domain of the methodological advancements of this dissertation includes demography, risk assessments, resource allocation planning, crime analysis and economics, to name a few.

To my parents,

Ali and Mansoureh

Acknowledgments

First of all, I would like to express my greatest debt of gratitude to my advisor Professor Stefan Leyk for his great support throughout the research and also for taking the time and putting the effort in teaching and educating me about different aspects of research and academia. He has been a great role model to me and has provided constant encouragement throughout the course of my graduate studies. I am very grateful to him for directing me to become a better researcher.

I would also like to convey my deepest gratitude to my dissertation committee members, Professor Barbara Buttenfield who has been one of the best teachers I have ever had, Professor Carson Farmer and Professor Matt Ruther for their consistent support and great feedback, Professor Leiwen Jiang for not only being a wonderful supervisor to work with, but also a great source of encouragement and inspiration. I wish to express my gratitude to all the professors whose courses I have taken during my time at CU Boulder.

I would like to acknowledge the collaboration with Zillow Inc. based on a Data Access Agreement between the Regents of the University of Colorado and Zillow Inc. Through this agreement, I was able to make use of the Zillow's invaluable dataset (ZTRAX®), which was highly influential and opened new directions to my research that I will follow in the future. The source code for processing the data is available at https://github.com/spatialhistory/ztrax_analysis.

I would like to express my gratitude to my colleagues at CU-Boulder for the in-depth discussions that helped in shaping the work and their valuable comments. Thank you Alex, Galen and Mehran.

I want to extend a special thanks to my friends Romik, Farhad, Azadeh and Arash who made my time at CU-Boulder very memorable. We have traveled very long miles and have seen all colors of Colorado together. You have been my second family here and made living far away from home much easier for me.

Finally, the greatest thanks of all goes to my wonderful parents Ali and Mansoureh, my caring sisters, Mahdieh and Mahsa and my supporting brother, Vahid. I have always felt your prayers and great wishes, and nothing would have been possible without your support and love.

CONTENTS

TABLES

CHAPTER III

TABLE

CHAPTER IV

TABLE

CHAPTER V

TABLE

FIGURES

CHAPTER II

FIGURE

CHAPTER III

FIGURE

CHAPTER IV

FIGURE

CHAPTER V

FIGURE

CHAPTER VI

FIGURE

Chapter I

Introduction

1.1. General overview

The U.S. Census Bureau, like numerous counterparts in different countries, provides important datasets describing the population at a given point in time that are useful for applications such as demographic analyses, health studies, crime analysis, hazard/risk assessment and land-use planning, to name a few. It publishes the data over a hierarchy of geographical units – from as large as census regions and counties to small blocks – in aggregated forms to provide useful summary statistics while protecting privacy. There is also a great demand in different research communities to carry out multi-temporal demographic analysis in order to characterize changes in population characteristics and how these changes relate to land-use, infrastructure or ecological characteristics. However, those levels that are released as small "statistical" units (e.g., census tracts, block groups and blocks) are often not temporally consistent because they are sensitive to changes in their underlying population distribution in contrast to coarse resolution geographical and political units such as counties (Gregory 2002, Martin *et al.* 2002, Schroeder 2007). That is, if the number of people in a statistical small unit such as a census tract increases between two census

years, enumeration boundaries may have to be adjusted; i.e., a tract will likely be divided to maintain the population counts within a given range. On the other hand, if its population declines, it will likely be merged with its neighboring tracts to meet the minimum population requirement of a census tract. Census units can also be changed through boundary shifts other than merges or splits. The consequent inconsistency of small statistical units over time impedes the effectiveness of investigating temporal changes of demographic attributes at fine analytical scales, and studies have relied on highly aggregated data (Exeter *et al.* 2005, Barufi *et al.* 2012) or accepted that the resulting estimates may have large errors (Gregory 2002).

To characterize relevant demographic processes, the analyst needs the population data to be available for different points in time over long time periods and enumerated within temporally compatible small spatial units. Unfortunately, the availability of data with such characteristics is often limited. Spatial analytical methods such as areal interpolation have been developed in order to address the issue of incompatibility among various enumeration systems (e.g., Goodchild *et al.* 1993, Maantay and Maroko 2009) or changing enumeration systems over time (e.g., Schroeder 2007, Logan *et al.* 2014), but these methods are still sensitive to population changes and are error-prone. Therefore, there is an urgent need for more advanced areal interpolation methods that make use of inherent characteristics of the collected data beyond its initial analytical scale as well as integrate external data to construct temporally compatible small area units that consistently decrease estimation errors. Such advancements are crucial for the effective and reliable characterization of different nuanced and micro-scale demographic processes, whose benefits are not limited to only demography, but can also include other domains such as health geography, crime analysis, urban planning and risk assessment, to name few. This constitutes the main motivation of this dissertation.

Recent research efforts have improved regular areal interpolation frameworks through the development of alternative weighting mechanisms (Schroeder 2007, Schroeder and Van Riper 2013) and the integration of dasymetric refinement in temporal analysis of population estimates (Logan *et al.* 2014, Buttenfield *et al.* 2015, Ruther *et al.* 2015, Zoraghein *et al.* 2016, Schroeder 2017). This dissertation provides a comprehensive study, targeting the applicability of spatial refinement in enhancing regular areal interpolation methods in various contexts, and advances the previous research efforts in several ways. First, several different ancillary variables with varying spatial resolutions, attributions and availabilities are employed for dasymetric refinement to demonstrate its effectiveness for temporal analysis. Second, this study develops advanced methodological extensions of existing areal interpolation frameworks to further reduce estimation errors. Third, the performances of the proposed individual methods are investigated across different settings of rural/urban characteristics. Fourth, various selected demographic variables in addition to total population are analysed and modeled. This will examine whether consistent multi-temporal estimates at fine spatial resolutions for population subgroups related to race, age structure and urban residence can be produced reliably. Fifth, this dissertation also runs a case study on a natural hazard risk assessment, in which the proposed methodological frameworks are employed to derive reliable estimates of the population potentially exposed. If successful, such methods can be applied for exposure analysis in similar contexts including flood, erosion, fire, future climate change, etc.

As a central component in this dissertation, it is explored how the accuracy of current areal interpolation methods can be improved by dasymetric refinement, leveraging different types of ancillary variables that have some type of relationship to the underlying population distribution. The goal is to identify the variables or combinations that enable the estimation of different

demographic attributes within temporally consistent small area census units from 1990 to 2010, with the minimum estimation error. This approach is particularly timely in a time of increasing data availability, reflecting land-use, land-cover or settlement patterns with national, continental and even global coverages available for different points in time. It is prudent to identify the potential of these data products in applications such as multi-temporal demographic analysis to examine the applicability in data poor regions of the world. Following this strategy, this study assesses the effectiveness of several available ancillary variables associated with the population distribution, including a global dataset, the Global Human Settlement Layer (GHSL) (Pesaresi *et al.* 2016), as well as a national layer, the National Land-Cover Database (NLCD) (Vogelmann *et al.* 2001). The study also examines the extremely fine-resolution ZTRAX® data (Zillow 2017), a proprietary database that is available for this research, as well as parcels and building footprints. These different variables are incorporated in the dasymetric refinement here to fully examine their influence on the overall accuracy of multi-temporal demographic analysis (Chapters 3, 4 and 5).

Data integration also addresses the examination of different attributes of ancillary variables in addition to their geometric footprints, which extends the principle of dasymetric refinement using limiting ancillary variables to the use of related ancillary variables (Chapters 4 and 5). This approach explores the question if population estimation can become more accurate provided that ancillary variables characterize nuanced population characteristics. Composite approaches that combine different ancillary datasets to exploit their complementary effects on the dasymetric refinement process are also evaluated (Chapters 4 and 5). The performances of the proposed methodologies are explored for various geographic extents (county and state) (Chapters 3, 4, 5), time periods (1990-2010 and 2000-2010) (Chapters 3, 4, 5), demographic attributes (total population, population by race, age structure and urban residence) (Chapter 5), and residential

settings (urban and rural) (Chapter 4).

The aforementioned novel approaches of data integration and data composition seek to address the "differences of spatial support", as the fourth dimension of interoperability issues in the spatial analysis (Goodchild *et al.* 2005), in a temporal application domain, more effectively. This dissertation advances knowledge in existing concepts and methods of spatio-temporal demographic analysis at fine spatial resolutions while building a foundation to extend this analysis to long time periods and keeping estimation errors low. In addition to methodological improvements, the advanced areal interpolation frameworks have practical merits for applied science communities in geography, demography, urban planning and health research that are interested in studying the demographic evolution, by providing them with more accurate time series of different aspects of population at fine spatial resolutions. Moreover, generating consistent micro-scale estimates of population and its evolution provides important starting points for coupling them with environmental data to advance research in human-nature systems. This includes advancements in dynamic modeling systems of land-cover/land-use change for causal investigation of such relationships in complex systems. Importantly, the application domain of the proposed methodologies is not limited to the temporal interpolation of population. More broadly, any application in which researchers are interested in more accurate disaggregation of enumerated variables can benefit from the findings of this dissertation.

This dissertation also benefits non-academic communities and the public in general. It can inform the public about new insights of the evolution of different demographic aspects of the American society over time at a scale meaningful for demographic processes. This cannot be done using the current census data alone. Such insights may become important to better understand social impacts on population groups, environmental injustice issues or the effectiveness of resource

allocation programs that has the potential to improve planning efforts and may result in better outcomes for city and regional planners.

1.2. Research questions

This dissertation addresses various methodological challenges related to data integration as well as interpolation procedures required with the objective of improving the accuracy of areal interpolation methods using dasymetric refinement for multi-temporal analysis of population characteristics. The following research questions guide the research of this dissertation:

Research question 1*: How effectively can different types of ancillary variables be used for spatial refinement to systematically improve the accuracy of regular areal interpolation methods? What lessons can be learned about the applicability of different ancillary variables in different geographic and demographic settings?*

To answer this question, various ancillary variables are incorporated into the dasymetric refinement framework, and the results are compared. This includes globally and nationally available datasets such as GHSL and NLCD as well as local datasets such as ZTRAX®, building footprints and parcels. Error measures are calculated for each implementation, and the effectiveness of each of these different ancillary variables is evaluated.

Moreover, the selected ancillary variables offer distinct spatial resolutions. While the resolutions of the grid-based NLCD, converted ZTRAX® and GHSL are constant, parcels and building footprints have a certain variability in areal extents, with buildings being the most precise delineation of the human settlement. This allows a cross-resolution comparison to better understand the effectiveness of different ancillary variables employed in different interpolation methods.

Research question 2*: How can existing approaches of dasymetric refinements in multi-temporal demographic analysis be extended to also incorporate land-use related attributes of ancillary variables in addition to their geometric footprints? What is the gain in accuracy from such an extended integration?*

Expectation Maximization (EM) (Dempster *et al.* 1977, Flowerdew and Green 1994, Schroeder and Van Riper 2013) is the main framework used in this dissertation to utilize heterogeneous land-cover/land-use types of ancillary variables in addition to their geometric footprints, when applicable. The algorithm differentiates between land-cover/land-use types by assigning distinctive population density weights to them according to their likelihood of population residence. In other words, ancillary datasets are not employed as the *limiting variable* by solely constraining where the population resides; rather, they are treated as the *related variable* by amplifying or curtailing the likelihood of the population settlement (Leyk, Buttenfield, *et al.* 2013).

The performance of the algorithm is assessed across different geographic and demographic settings and compared to other dasymetrically refined areal interpolation methods. In this dissertation, EM is also improved by introducing the Enhanced Expectation Maximization (EEM) to tap into the potential accuracy gain of using related ancillary variables in a more complex fashion and more adequately.

Research question 3*: How can the effectiveness of dasymetric refinement be improved in rural areas typically associated with lower accuracy due to the lack of ancillary variables or the chronic under- or overestimation of population in such settings?*

Land-cover based datasets such as NLCD under-estimate developed lands in rural areas due to their misclassification errors. On the other hand, parcels over-estimate those lands because of typical large rural lots (Leyk *et al.* 2014). These issues, coupled with low population counts,

result in lower overall accuracy of rural population estimations (Tapp 2010). In this dissertation, a composite approach using NLCD, parcels and road networks is established to evaluate if higher accuracy levels can be achieved in rural areas by using the complementary refinement effects of the three datasets.

Research question 4*: How stable is the effect of spatial refinement across different geographic scales and in estimating additional demographic attributes other than total population? Does this framework allow for coupling between demographic and environmental data to analyze more complex relationships?*

The proposed methodologies are implemented in both county-scale and state-scale applications to investigate the robustness of combining dasymetric refinement with temporal areal interpolation in reducing population estimation errors, both among counties with distinct spatial and demographic characteristics and across different geographic scales.

Moreover, the enhanced methods are operationalized for demographic attributes other than total population counts, including population based on race, age structure and urban residence, to assess the stability of their efficacy in error reduction across various demographic estimates. Furthermore, the improved areal interpolation methodologies are coupled with an environmental dataset, i.e., flood zones in Massachusetts, to estimate potentially exposed population counts for different race and age-related sub-groups, based on tract-level data. The outcomes of this analysis are compared to more refined block-based estimated counts to evaluate the applicability of tract-based analyses where blocks are not available.

Research question 5*: By making the methodological frameworks operational for multiple demographic attributes and over different time-periods, what expectations exist related to the uncertainty inherent to the population estimates produced?*

As suggested by Schroeder (2007), several factors affect the uncertainty involved in temporal areal interpolation of population over census years. This dissertation estimates different demographic attributes over two time periods (1990-2010 and 2000-2010) at the census tract level and evaluates the magnitude of accuracy gains offered by the proposed methodologies. Particularly, the level of accuracy gains in temporal interpolation of population over the shorter time period (2000-2010), and presumably lower levels of inconsistency of census boundaries, is compared to that related to the longer period (1990-2010). Moreover, accuracy gains are compared between different demographic attributes with varying population counts. Such comparisons could shed light on what circumstances require the employment of the dasymetric refinement prior to areal interpolation, as the process is both time- and data-demanding.

1.3. Structure of the dissertation

Chapter two provides a literature review of the relevant applied and theoretical work for this dissertation and a detailed background on areal interpolation, dasymetric modeling and its applications in the temporal interpolation of population, found in recent research. Chapters three, four, and five present the different stages of the development and application of the proposed methodologies for the temporal estimation of population. Each of these three chapters is presented as a stand-alone paper that describes in-depth technical details, methodological components and experimental applications of the methods.

Chapter three details the conceptual and algorithmic foundation for interpolating total population estimates of census tracts in 2000 within census tract boundaries in 2010 to construct temporally consistent total population estimates at the small area tract level in Hennepin County, Minnesota. The methodological framework implemented in this chapter is a continuation of previous research efforts such as Buttenfield *et al.* (2015) and Ruther *et al.* (2015). However, parcel

footprints are used as the main ancillary variable to mitigate the underestimation artifact of using NLCD in rural settings (Leyk *et al.* 2014). First, regular areal interpolation methods such as Areal Weighting (AW) (Goodchild and Lam 1980), Pycnophylactic Modeling (PM) (Tobler 1979) and Target Density Weighting (TDW) (Schroeder 2007) are dasymetrically refined using parcels and NLCD separately. Second, their resulting accuracy measures are compared. This helps answer if either ancillary variable is dominantly more efficient across the algorithms, and if not, under which circumstances, one ancillary variable performs more accurately then the other.

This chapter treats both ancillary variables as the limiting variable, implying that population is distributed equally over the inhabitable land without considering different land-cover/land-use types that can have their own specifications of the human settlement.

Chapter four extends the conceptual foundation developed in Chapter three and addresses the associated challenges. It conducts three levels of spatially refined areal interpolation methods in five counties with distinct demographic features, namely Hennepin County, Minnesota, Mecklenburg County, North Carolina, Broward County, Florida, Hillsborough County, Florida and Worcester County, Massachusetts for the two time periods of 1990-2010 and 2000-2010. The first level of spatial refinement is the same as Chapter three. The second level uses EM to take into account varying population density weights associated with different housing characteristics of parcels. For instance, the population density weight assigned to condominiums differs from the one assigned to sing-family residences, and these variations are addressed by using the utilized algorithmic framework. EEM, which is an advancement of EM, is also introduced to offer a more reliable methodology to leverage the related ancillary variable. Finally, the third complementary spatial refinement, which is a composite approach making use of parcels, NLCD and road networks, is formulated to reduce population estimation errors in rural settings. The outcomes of

the methods are compared between the different study areas and over the two time durations.

Chapter five puts all the frameworks established in Chapters three and four in an application-oriented domain. This chapter pursues several goals, but all of them can be situated within one primary objective, which is to assess the stability and robustness of the proposed methodologies under varying circumstances. It first increases the geographic scale by carrying out the methods over the whole state of Massachusetts. It also employs multiple ancillary variables, including NLCD, GHSL, census-defined urban areas, parcels, building footprints and ZTRAX®, with distinctive characteristics in terms of their availability and spatial resolution, to objectively quantify the effectiveness of each dataset in improving the overall accuracy of the temporal estimation of population. Particularly, ZTRAX® is a unique and rich dataset, describing housing properties, provided by the Zillow® research group (Zillow 2017), whose potential for improved multi-temporal demographic analysis is explored in Chapter five.

In addition to total population, other demographic attributes including population counts by race, age structure and urban residence are input to the temporal interpolation context. The reliability of the census-defined urban areas in 1990, 2000 and 2010 in delineating urban lands is also investigated by the algorithmic frameworks. Finally, the proposed methodologies are tested in an environmental injustice application to investigate if certain sub-groups of population are disproportionately exposed due to elevated flood risks.

The numerous implementations of the enhanced methods tested in this chapter shed light on which methods are most effective under which circumstances. The outcomes provide opportunities to justify the employment of the dasymetric refinement prior to areal interpolation given their data and processing time entailments. They also point to the performances of GHSL and NLCD, which can be leveraged in data-poor regions as opposed to more precise datasets such

as parcels, building footprints and ZTRAX®.

Moreover, this chapter shows how tract-based population counts for different demographic sub-groups estimated to reside in flood zones mimic the corresponding block-based reference estimates in 1990 and 2000. This allows the extension of the analysis back to earlier census years, when census blocks had not been collected nationally, or other census systems that don't have similar fine-resolution enumeration units.

Chapters three, four and five document the formulation of the proposed methodologies and illustrate how they can be applied to different population interpolation tasks. A broad discussion of the findings of the three chapters is presented in ***Chapter six***, in which the different research questions posed above are revisited to systematically evaluate the research efforts and results of this dissertation. This is followed by conclusions for the dissertation.

Chapter II

Literature Review

This research is broadly situated in the literature on areal interpolation and dasymetric modeling in spatial analysis but has the main objective of contributing to the temporal demographic analysis using enumerated data that commonly suffer from boundary inconsistency over time. Therefore, first, the important and fundamental research efforts in areal interpolation and dasymetric modeling are briefly reviewed. Specific focus in this chapter is then given to discuss the different methodological ways of integrating these two modeling techniques to develop a technical framework for the temporal interpolation of population data that are initially inconsistent among different census releases.

2.1. Census Geography and Census Demography

The primary sources of spatio-demographic information in the United States are decennial censuses. The U.S. Census Bureau provides data as a series of summary text files labeled from 1 to 4. These text files provide information at different levels of spatial geography that include census blocks, as the smallest census unit, to as large as the entire United States (U.S. Census Bureau 2010a). Demographic characteristics are summarized and reported over geographies to protect

privacy, and can be based on 100% data (e.g., summary files 1 and 2) or sample data (e.g., summary files 3 and 4) (U.S. Census Bureau 2010b).

From the various geographical units provided by the U.S. Census, the two units that are used in this dissertation are census tracts and census blocks, and both are considered statistical small area units. Census tracts are small, relatively permanent statistical sub-divisions of a county or equivalent entity that are updated by local participants prior to each decennial census. Census tracts generally have a population size between 1200 and 8000 people, with an optimum size of 4000 people. The spatial size of census tracts varies widely depending on the density of settlement (U.S. Census Bureau 2012). On the other hand, census blocks are the smallest and most numerous unit by a large margin. For instance, in 2000 census data, there were over 8 million blocks, 39 times more than the next most numerous unit, block groups (Schroeder 2017). The nation-wide temporal availability and the number of demographic attributes offered by blocks are more limited than tracts (U.S. Census Bureau 2010b, Minnesota Population Center 2016).

To study local and regional trends of demographic characteristics, the demographic summary data enumerated over census small geographies are a vital resource. However, the linkage between census demography and census geography necessitates that small statistical units should change from census year to another to reflect underlying population changes. These boundary changes can be in the forms of split (population growth), merge (population decline) or other complex types (Zoraghein *et al.* 2016). For example, although census tracts are designed to be stable, almost 53% of 2000 census tracts – about 35000 out of 66000 – are not directly comparable with a 1990 census area (Schroeder 2007). This phenomenon complicates micro-scale multi-temporal analysis of demographic processes (Gregory 2002, Martin *et al.* 2002, Schroeder 2007, 2017). Thus analytical solutions to this persistent problem, such as areal interpolation and

dasymetric modeling, should be provided and enhanced.

2.2. Areal interpolation in spatial analysis

Areal interpolation is the process of transferring data aggregated over one set of areal units (source zones) to incongruent target zones (Lam 1983). Areal interpolation is specifically devised to address the change of support problem, which is concerned with inferences of values at or for locations different from those at which values have been originally observed (Gelfand *et al.* 2001), such as inconsistency between census units and hazard zones (e.g., Maantay and Maroko 2009, Mennis 2015), temporal inconsistency between census boundaries from different census years (e.g., Gregory 2002, Martin *et al.* 2002), and disaggregation of enumerated population to grid cells (e.g., Bhaduri *et al.* 2007, Dmowska and Stepinski 2017). More specifically, Mrozinski and Cromley (1999) describe different geometric situations to estimate unknown areal values that necessitate the implementation of areal interpolation methods; i.e., the missing data, alternative geography and intersection problem.

Figure 2.1 shows the two broad categories of areal interpolation methods as indicated by Lam (1983). According to Figure 2.1, the areal interpolation framework is divided into non-volume-preserving and volume-preserving methods.

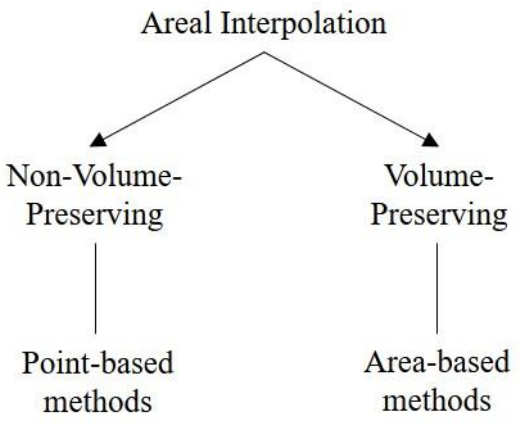

Figure 2.1. Hierarchical categorization of the areal interpolation framework

Point-based areal interpolation approaches first assign a representative control point to each source zone. Point interpolation schemes are then applied to interpolate the value at each grid node. Finally, the estimates of the grid points are averaged or aggregated within each target zone, yielding the final target-zone estimate. These approaches are not volume-preserving (pycnophylactic), meaning that if they are applied to a source zone, they don't necessarily calculate its exact value. In this study, the focus is placed on volume-preserving areal interpolation methods.

Areal Weighting (AW) (Goodchild and Lam 1980) is the most preliminary volume-preserving areal interpolation method, in which values of target zones are estimated proportionally based on areas of their intersections with source zones. The method's main drawback is the assumption that the attribute of interest is uniformly distributed within each source zone. Even though it performs poorly in most evaluations (Fisher and Langford 1995, Langford 2006, Kim and Yao 2010), it is frequently used because of its low data requirements and processing time.

Pycnophylactic Modeling (PM) (Tobler 1979), on the other hand, assumes the existence of a smooth volume-preserving population density surface and incorporates the densities of adjacent zones. The algorithm derives the density surface in an iterative process, which then can be

aggregated to any arbitrary set of target zones. The regular PM procedure is grid-based, extended to a surface representation based on a triangulated irregular network (TIN) by Rase (2001). Moreover, Kim and Yao (2010) combine PM with dasymetric modeling and report on accuracy gains in population estimation.

Schroeder (2007) introduces Target Density Weighting (TDW) as an areal interpolation method appropriate for temporal analysis of population data. The method extends "target count weighting", a term introduced by Schroeder (2007) for a technique described in Howenstine (1993) and Mugglin and Carlin (1998), which assumes that all target zones nest within source zones. TDW is designed to address the change of support problem in a temporal context, mainly used for establishing temporally consistent demographic estimates within misaligned units from different census years. It assumes that the ratios of population densities of atoms (intersections of source and target zones) to source zones remain constant over time. Based on previous studies, TDW often outperforms AW (Schroeder 2007, Schroeder and Van Riper 2013), suggesting that it is more reasonable to assume that the rate of change of the variable of interest is constant for atoms than to assume that the variable is homogeneously distributed within source zones.

Regression-based frameworks have also been developed for areal interpolation. Krivoruchko *et al.* (2011) report on the applicability of several areal kriging approaches, namely the Gaussian areal kriging, overdispersed Poisson areal kriging and binomial areal kriging models. Kyriakidis (2004) formulates areal interpolation as a geostatistical area-to-point kriging problem and Yoo *et al.* (2010) show how geostatistical methods are by design pycnophylactic. Other examples of using regression-based frameworks for areal interpolation point to the Geographically Weighted Regression (GWR) framework (Fotheringham *et al.* 1992) modified by Lin *et al.* (2011) in order to make it volume-preserving and the quantile regression by Cromley *et al.* (2012).

However, given the complexity and subjectivity typically involved in statistical techniques, previous research has shown that this type of methods tends to be outperformed by or does not offer much improvement over dasymetric modeling (Lin *et al.* 2011).

2.3. Dasymetric modeling as a special type of areal interpolation

As noted above, the main application of areal interpolation is to transfer population values from source zones to target zones. So-called "simple" areal interpolation methods do not exploit any supplementary data in the process (Okabe and Sadahiro 1997). However, population distribution can often be explained by other socioeconomic phenomena. Thus, incorporating associative ancillary variables can shed light on the underlying population distribution, thereby increasing the accuracy of areal interpolation (e.g., Flowerdew *et al.* 1991, Goodchild *et al.* 1993, Xie 1995, Eicher and Brewer 2001, Mennis 2003). Such integration will also alleviate the Modifiable Areal Unit Problem (MAUP) (Openshaw and Taylor 1979, Openshaw 1984), which is a special case of the Change Of Support Problem (COSP) (Anselin 2002). According to MAUP, spatial analysis using zonal data (e.g., census units) might lead to biased estimation outcomes depending on the configuration and size of the zones. One solution to alleviate this issue, which is typically applied in dasymetric modeling, is to transform the aggregated data enumerated within arbitrary units to those that may be more relevant in reflecting the spatial distribution of the variable of interest defined by ancillary variables (Wu and Murray 2005, Su *et al.* 2010).

Dasymetric modeling was first developed in the early 20th century as a cartographic technique aimed at addressing some of the issues associated with choropleth mapping (Semenov-Tian-Shansky 1928, Wright 1936). Fisher and Langford (1995) demonstrate how the principles of the technique can also be used to enhance areal interpolation algorithms. It assumes an association between the population distribution and ancillary data to create zones that are internally

homogenous and externally heterogeneous, and this usually reflects the underlying statistical surface more precisely (Eicher and Brewer 2001, Mennis and Hultgren 2006). Two types of ancillary data can be used in dasymetric modeling, namely the limiting ancillary variable and the related ancillary variable. The former determines where the population resides and where it is absent, which requires a recomputation of population density variables within spatially constrained areas and has been used in numerous research studies (e.g., Langford 2007, Kim and Yao 2010, Ruther *et al.* 2015). The latter has a more complex association with the population distribution and can both amplify and reduce the likelihood of population presence and population densities (Leyk, Buttenfield, *et al.* 2013, Nagle *et al.* 2014). One of the main challenges in employing related ancillary variables is the lack of available data that would allow the analyst to establish and utilize statistical relationships between such variables and population characteristics. Recently, important progress has been made on this topic and resulted in more accurate fine resolution population counts as well as more differentiated uncertainty analysis in different settings (Leyk, Nagle, *et al.* 2013, Schroeder and Van Riper 2013, Nagle *et al.* 2014)

Dasymetric modeling has gained popularity recently with the increased availability of spatial data from mapping agencies, open source data, and remote sensing data products (Mennis 2009, 2015, Langford 2013), as well as significant advancements in estimation techniques (Kim and Yao 2010, Leyk, Nagle, *et al.* 2013). Land-cover/land-use data is the most widely used ancillary variable in dasymetric modeling (Mennis 2003, Bian and Wilmot 2017, Dmowska and Stepinski 2017). There are numerous research efforts in the literature of dasymetric modeling, leveraging land-cover/land-use data and its products for population disaggregation (e.g., Mennis 2003, 2009, Reibel and Agrawal 2007, Linard *et al.* 2011, Buttenfield *et al.* 2015, Ruther *et al.* 2015, Dmowska and Stepinski 2017). The other types of ancillary data used in dasymetric

modeling include but are not limited to street networks (e.g., Xie 1995, Cromley and McLafferty 2002, Reibel and Bufalino 2005), Landsat TM imagery (Harvey 2002), imperviousness surfaces (e.g., Zandbergen and Ignizio 2010, Schroeder 2017), address points (e.g., Tapp 2010, Zandbergen 2011), high resolution satellite images (Lu *et al.* 2010, Ural *et al.* 2011, Lung *et al.* 2013), night-time lights (e.g., Zandbergen and Ignizio 2010, Wu *et al.* 2017), LiDAR (e.g., Dong *et al.* 2010, Qiu *et al.* 2010, Sridharan and Qiu 2013, Xie *et al.* 2015), tax parcel data (e.g., Maantay *et al.* 2007, Kar and Hodgson 2012, Mitsova *et al.* 2012, Jia *et al.* 2014, Jia and Gaughan 2016, Zoraghein *et al.* 2016), buildings (e.g., Wu *et al.* 2008, Calka *et al.* 2016), VGI (e.g., Bakillah *et al.* 2014, Lin and Cromley 2015, Geiß *et al.* 2016) and miscellaneous datasets composed of topography, land-cover/land-use and transportation layers (e.g., Bhaduri *et al.* 2007, Su *et al.* 2010).

The main task in dasymetric modeling is to establish correlations between ancillary and population data. The most preliminary and widely used approach is to assume a binary relationship between the two datasets (Langford 2013); i.e., population is not apportioned to presumably uninhabitable areas and instead is assigned homogeneously to the rest. The three-class and limiting variable methodologies are extensions to the binary variant, but still use subjective decisions to allocate population density weights to different classes of the ancillary dataset (Eicher and Brewer 2001). This limitation is partially overcome by empirical sampling (Mennis 2003, Mennis and Hultgren 2006) or statistical regression (Yuan *et al.* 1997, Harvey 2002, Langford 2006) frameworks. The former empirically samples representative units per ancillary data type and derives population density weights for different types using their corresponding units. The latter may produce negative population estimates and does not preserve data volumes, i.e., the pycnophylactic property, globally or locally. They usually require *a posteriori* coefficient

adjustments to maintain volume-preserving requirements. Flowerdew and Green (1994) adopt the Expectation Maximization (EM) algorithm (Dempster *et al.* 1977) to assign density weights, which incorporates ancillary data directly in an optimization process. There also exist examples using geostatistical methods such as co-kriging (e.g., Wu and Murray 2005, Liu *et al.* 2008), maximum entropy (Leyk, Nagle, *et al.* 2013) or a multi-layer multi-class framework described in Su *et al.* (2010) that attempt to establish associations between ancillary variables and population characteristics.

2.4. Combining dasymetric modeling and classical areal interpolation techniques for improved temporal estimation of population characteristics

Dasymetrically refined areal interpolation methods have recently been used in temporal contexts to construct compatible population estimates from initially misaligned census units with lower estimation errors. This has significant applications for geographers, demographers and urban planners interested in analyzing underlying demographic processes. In addition to the contextual applications of the approach, such spatially refined interpolations mitigate the MAUP issue to some degree because only inhabitable sub-areas of source and target zones are used, which are assumed to be more closely related to the spatial distribution of population than the unrefined zones.

Holt *et al.* (2004) use satellite-derived land-cover/land-use data to map population densities using dasymetric modeling for different time periods. They present three years of census data (1980, 1990 and 2000) in one set of common census tracts (1990) with a satisfactory level of accuracy. Schroeder (2007) formulates TDW specifically for temporal analysis of census data and produces estimates of compatible populations between the census years 1990 and 2000. Ruther *et al.* (2015) and Buttenfield *et al.* (2015) spatially refine incongruent census units using the National

Land-Cover Database (NLCD). They then apply common areal interpolation methods such as AW, TDW, PM and EM to different counties in the United States with varying population growth/decline patterns. They report on accuracy gains using dasymetric refinement for establishing consistent population estimates during the 1990-2010 time period. Zoraghein *et al.* (2016) extend this work, but use parcel footprints for dasymetric refinement – as a different ancillary variable with distinct potentials – and report on higher accuracy improvements in temporal analysis of population for 2000-2010. To generate a consistent longitudinal time series of population at the census tract level from 1970 to 2010, Logan *et al.* (2014) use two approaches, i.e., AW and a more sophisticated areal interpolation approach that takes into account the auxiliary distribution of population by blocks. They demonstrate the effectiveness of their proposed approach for large areas. Logan *et al.* (2016) report on the effectiveness of their approach formulated in Logan *et al.* (2014) and methodologies used in the National Historical Geographic Information Systems (NHGIS) in comparison to the preliminary AW approach for establishing consistent population estimates from 2000 to 2010 at the census tract level. Schroeder and Van Riper (2013) devise hybrid approaches of GWR, EM and TDW to exploit the complementary benefits of each separate method and show higher accuracy attainments by their developed methods in temporal estimation of census tract populations between 1970 and 1980. The development of hybrid approaches of the binary dasymetric modeling and TDW using composite ancillary datasets such as imperviousness data, road buffers and water polygons is further pursued by Schroeder (2017) to produce high-quality block-based estimates of 2000 census counts for 2010 units, and resulting accuracy improvements are documented.

All these research efforts stress the importance of generating compatible population data over time to analyze micro-scale demographic patterns over small census geographies. They have

been successful in improving the accuracy of multi-temporal population estimation compared to using regular areal interpolation methods. However, an in-depth research effort aiming to employ different combinations of available ancillary data sources and algorithmic frameworks to improve the accuracy of the interpolation process further for different aspects of census demography and under various geographical and demographical circumstances is still required.

Chapter III

Exploiting Temporal Information in Parcel Data to Refine Small Area Population Estimates

Abstract[1]

Temporal analysis of small-area demographic data commonly relies on areal interpolation methods to create temporally consistent and compatible areal units. In this study, cadastral (parcel) data are used to identify residential land and to dasymetrically refine census tracts, with the goal of achieving more accurate small-area estimates. The built date recorded for residential parcel units is used to create residential land layers for two different time points used in the areal interpolation. Three different areal interpolation methods are employed with and without dasymetric refinement, including areal weighting (AW), target density weighting (TDW) and pycnophylactic modeling (PM). The methods interpolate tract-level population counts in Hennepin County, Minnesota, in 2000 into census tract boundaries from the year 2010. The mean absolute error, median absolute

[1] This chapter was published as a journal paper:
Zoraghein, H., Leyk, S., Ruther, M., and Buttenfield, B.P., 2016. Exploiting temporal information in parcel data to refine small area population estimates. *Computers, Environment and Urban Systems*, 58, 19–28. DOI: 10.1016/j.compenvurbsys.2016.03.004

error, root mean square error and the 90th percentile of absolute error are calculated for each of the methods, and spatial variation in the interpolations are displayed in maps. Parcel-based refinements are also compared with refinements using the National Land Cover Dataset (NLCD).

Results show that spatial refinement using residential parcels has the potential to improve the accuracy of areal interpolation for temporal analysis. Parcel-refined TDW out-performs the other tested methods, as well as the NLCD-refined TDW in this example. Parcel data identify residential land more reliably in rural areas. However, parcel units can have very large extents potentially biasing residential area delineation and population counts. Parcel-based refinement has the potential to further advance demographic change analysis over long time periods and large areas where the built date attribute is included in the dataset.

Keywords: Areal Interpolation; Spatial Refinement; Census Units; Temporal Analysis; Population

3.1. Introduction

Numerous pressing research questions in population studies, urban planning and public health require an in-depth understanding of demographic processes. Researchers must rely on data collected at predefined spatial resolutions in characterizing the process of interest which may operate at a different scale (Maantay *et al.* 2007, Giordano and Cheever 2010). In order to describe demographic change reliably, population data must be available for different points in time and enumerated within compatible spatial units. Such data rarely exist due to changes in enumeration boundaries, which often arise due to population growth or decline (Reibel and Agrawal 2007). Temporally incompatible spatial units impede demographic change analysis, and researchers have been developing analytical approaches for temporal analysis that attempt to solve such problems.

Research on small area population estimation at a single point in time has seen significant progress based on methods such as dasymetric modeling (Wright 1936, Eicher and Brewer 2001, Mennis and Hultgren 2006) or microsimulation (Birkin and Clarke 2011, Tanton *et al.* 2014). The problem of incompatible enumeration units in temporal analysis has recently found increasing interest, with much of this work focusing on methods of areal interpolation (Schroeder 2007, Schroeder and Van Riper 2013, Logan *et al.* 2014). These methods sometimes rely on assumptions concerning the spatial distribution of the population (or other demographic variable of interest) that do not hold true in many situations (Syphard *et al.* 2009).

The combination of dasymetric refinement with areal interpolation has been demonstrated as a promising way to improve the precision and accuracy of small area population estimates for temporal analysis. For example, Holt *et al.* (2004) showed that refining source zones using developed land cover classes in one point in time improved areal interpolation results considerably. Other studies have applied ancillary data such as road networks or land cover data to determine

weights that distinguish residential from non-residential areas in different points in time (Reibel and Bufalino 2005, Schroeder and Van Riper 2013). Dasymetric techniques have recently been examined to refine areal interpolation over multiple time periods. Buttenfield *et al.* (2015) and Ruther *et al.* (2015) compared different areal interpolation techniques with dasymetrically refined source zones using developed land classes from the National Land Cover Database (NLCD). Their results demonstrated that dasymetrically refined and unrefined methods performed differently for shorter and longer time periods, with lower error rates in most cases when dasymetric refinement was used. Buttenfield *et al.* (2015) showed that improvements can manifest in some areas whether the dasymetric method refines source or target estimates, or both. Ruther *et al.* (2015) included four different demographic settings in their study, and found distinct spatial and temporal variations in the error of estimates for different demographic conditions. High estimation errors were discovered in fast growing subregions, in areas that were more fully developed for both the source and target time periods, and in rural settings where NLCD is known to underestimate developed residential land. Mennis (2016) applied the principles of dasymetric mapping to spatiotemporal interpolation in an example of crime data analysis and indicated that the accuracy of estimates was significantly improved.

The present study employs cadastral (parcel) data to identify residential land, which is used to dasymetrically refine census tract population estimates from U.S. Decennial Censuses at two points in time prior to areal interpolation. The temporal information used to identify residential development – the date when a current residential building was built – is not a standard component in parcel data, but exists for many underlying county-level databases. Where the built year attribute is available, this study will demonstrate that it may be highly beneficial in providing a temporal dimension for a better understanding of the evolution of residential parcels. The parcel-refined

census tracts at each point in time are then employed in the interpolation to generate population estimates for 2000 within census tract boundaries from 2010 (target zones) such that spatially compatible units can be compared. An important motivation is that residential parcels likely represent populated places more reliably than developed land classes in regional or national land cover databases. This is particularly true in rural areas where remote sensing based classification does not reliably detect small areas of residential development. Hence the parcel-refined temporal interpolation is expected to result in reduced estimation errors in such settings when compared to non-refined and land cover refined scenarios. Three different areal interpolation methods are tested, with and without dasymetric refinement, and their estimation errors are evaluated using census block statistics.

3.2. Background

3.2.1. Areal interpolation

Areal interpolation is the process of transferring data aggregated over one set of areal units (source zones) to another (target zones) (Lam 1983). It can also be applied to apportion population from enumeration units for one time period into units for another time period, in order to achieve temporally consistent enumeration units (Schroeder 2007, Schroeder and Van Riper 2013). Several areal interpolation methods have been developed to date, including areal weighting (AW) (Goodchild and Lam 1980, Lam 1983), target count weighting (TCW) (a term introduced by Schroeder (2007) after the method presented by Howenstine (1993) and Mugglin and Carlin (1998)), pycnophylactic method (PM) (Tobler 1979), and target density weighting (TDW) (Schroeder 2007). All methods carry inherent assumptions. For example, AW assumes that population densities are constant within source zones. TCW assumes that target zones nest within source zones, and that the spatial distributions of the variable of interest and ancillary variable are

proportionally the same among target zones within each source zone. Such assumptions often do not hold, and there may be large estimation errors where significant changes in population counts occur. The utilization of ancillary (limiting) variables to guide areal interpolation for temporal analysis has been demonstrated as a promising avenue to reduce the problem of spatial unit incompatibility (Holt *et al.* 2004, Schroeder and Van Riper 2013). As mentioned above, an explicit dasymetric refinement using developed land classes prior to areal interpolation has been tested for several methods for temporal analysis to explore the influence on estimation accuracy for different time periods (Buttenfield *et al.* 2015) and across varying demographic settings (Ruther *et al.* 2015). The examination of alternative limiting variables and the application of dasymetric refinement to both source and target zones are promising directions to investigate. These two aspects will be tested in the present study.

3.2.2. Dasymetric modeling

A common application of census-defined areal units is to report their demographic variables. Areal interpolation methods may result in large estimation errors when spatial units do not reflect the inherent variability in the demographic variable of interest. Dasymetric modeling employs ancillary data to spatially refine the distribution of the variable of interest (Wright 1936). It assumes an association between the variable of interest and the ancillary data to create zones that are internally homogenous and externally heterogeneous, and that reflect the underlying statistical surface more reliably. Dasymetric modeling has gained popularity recently with the increased availability of spatial data from mapping agencies and remote sensing data products (Mennis 2009, 2015), as well as significant advancements in estimation techniques (Kim and Yao 2010, Leyk, Nagle, *et al.* 2013). The different types of ancillary data used in dasymetric modeling include road network density (Reibel and Bufalino 2005), Landsat TM imagery (Harvey 2002)

and cadastral data (Maantay *et al.* 2007). In addition, research has been carried out using combinations of different types of data, such as land-cover, imperviousness, road networks, and nighttime lights (Zandbergen and Ignizio 2010) or address points and parcels (Tapp 2010).

Early dasymetric methods relied on subjective decisions about the associations between ancillary and population data (Eicher and Brewer 2001). This limitation was partially overcome by sampling-based dasymetric methods as described in Mennis (2003) and Mennis and Hultgren (2006). Regression-based analyses (Yuan *et al.* 1997, Harvey 2002) may produce negative population estimates and do not preserve data volumes globally or locally. They usually require *a posteriori* coefficient adjustments to maintain volume-preserving requirements. Flowerdew and Green (1994) adopted the expectation/maximization (EM) algorithm (Dempster *et al.* 1977) for areal interpolation, which incorporates ancillary data directly in an optimization process that derives spatially refined population estimates. There also exist examples using geostatistical methods, such as area to point kriging (Kyriakidis 2004) and co-kriging (Wu and Murray 2005, Liu *et al.* 2008). Despite this methodological progress, some challenges remain, including the identification of residential areas in rural settings (Zandbergen and Ignizio 2010, Leyk *et al.* 2014) and the accurate estimation and validation of populations for various demographic attributes (Nagle *et al.* 2014).

The issue of rural residential area is of special relevance here. Many dasymetric mapping applications reported in the literature use land-cover classifications derived from Landsat remote sensing data with 30 m resolution as the limiting ancillary data for refining population distributions. In general, such ancillary datasets are more accurate in urban settings than in rural areas because in rural areas, residential units that are smaller than the pixel size of commonly used remotely sensed imagery cannot be detected. This results in underestimations of rural populated

places relative to urban places (Leyk *et al.* 2014). On the other hand, residential parcels are expected to depict residential developed land more reliably. However, rural parcels can have large areal extents, or non-residential buildings, and there is a need to explore how this may impede or even contradict the spatial refinement process. This paper will examine the role and benefits of parcel data for improving temporal small area estimation of population.

3.3. Study area, data and pre-processing steps

The methods of this paper were tested in Hennepin County, Minnesota. This study area includes both the highly populated city of Minneapolis in the eastern part and sparsely populated rural portions in the west. The variation in population density makes this area an ideal case study to evaluate the performance of each method under different conditions (Figure 3.1).

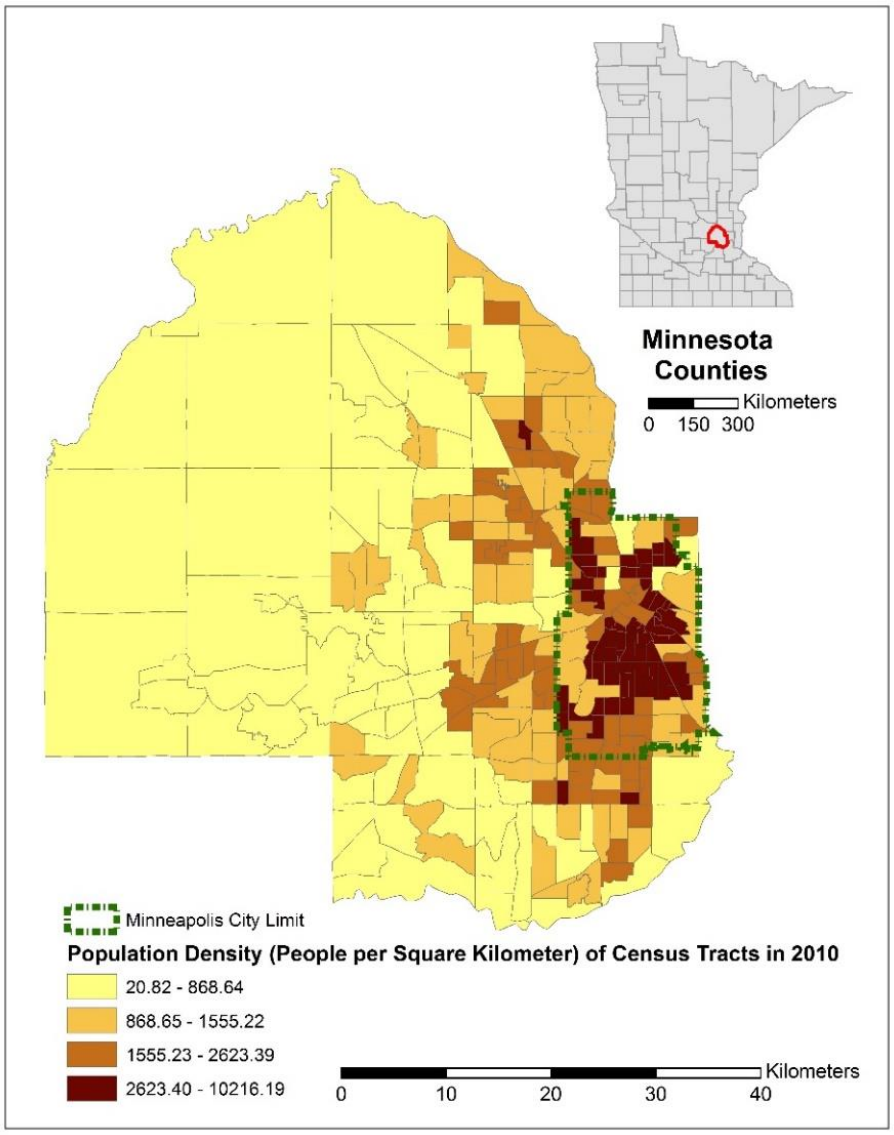

Figure 3.1. Study area and its location in Minnesota.

The aim of this study was to interpolate the population enumerated within 297 census tracts in 2000 into the 298 tract boundaries of the 2010 Census to create consistent and comparable units across time. Changes in census tract boundaries between 2000 and 2010 were divided into two categories: unchanged tracts (less than 900 m^2 difference) and changed tracts (more than 900 m^2 difference) from Census 2000. The 900 m^2 threshold is based on the extent of one pixel with 30 m resolution. By relying on this cut-off value, all boundary changes greater than 900 m^2 were

incorporated. It should be noted that some minor boundary changes may be related to the production process of TIGER/Line data by the Census Bureau. However, in this study, a rather conservative approach that takes into account changes in size as small as 900 m^2 was chosen to ensure that all possible changes were included in the accuracy assessment. Based on this criterion, 129 of the 2010 Census tracts were categorized as changed from 2000.

Census block data from 2000 (17,367 blocks) was used for validation. Comparison of block statistics aggregated within tracts to the interpolated tract values allowed calculation of different error metrics such as mean absolute error (MAE), median absolute error, 90^{th} percentile of absolute errors and root mean square error (RMSE). All population count data and census boundaries were retrieved from the U.S. Census Bureau website (http://factfinder.census.gov and https://www.census.gov/cgi-bin/geo/shapefiles2010/main, respectively).

The parcel dataset is available through the Hennepin County Open Data Portal (http://www.hennepin.us/gisopendata). The dataset is compiled monthly by the Hennepin County GIS Office from parcel geometry that is created and maintained by the Hennepin County Resident and Real Estate Services Survey Division and contains rich attribution on each property in the county. The 16 parcel types that indicated residential use – including single family housing unit, townhouse, triplex, nursing home, apartment, different types of condominiums, cooperatives, farm homesteads, and low income housing unit – were generalized into a single "residential" category. For this paper, the most salient attribute in the parcel dataset is the year when a building was established, with dates ranging from 1843 to 2010 in Hennepin County. This attribute was used to identify buildings that were built before 2000 and those that were built before 2010, to create snapshots of the distribution of residential developed land at two different points in time. It is important to acknowledge that the given date does not necessarily indicate the initial construction

activity in a particular parcel but rather the construction date of the currently existing building in the parcel. Moreover, the parcel record does not attribute dates when a structure was torn down or destroyed (e.g., by flooding, fire, or other accident). This problem will likely become more serious when longer time periods (more than 10 years) are covered, as the proportion of rebuilt structures – and possible changed building types – may be higher. Furthermore, because residential development was used as a limiting variable, differences in housing characteristics were not considered, and these differences could be neglected for the present study.

For some parcel types, such as condos, multiple spatially coincident polygons existed in the database. This occurs because the database stores one record per ownership, and a single condo structure typically includes multiple owners. For example, if 20 people own 20 units in a condo, the record corresponding to the geography of that condo is repeated 20 times in the database. Because the parcel data was used as a limiting ancillary variable in this study, only one instance of these repetitive records was retained. Furthermore, parcels with very small area and a built-year of "0" were considered noise and removed. Based on the above preprocessing steps, the final residential parcel datasets for the years 2000 and 2010 contained 296,367 and 317,292 records, respectively.

3.4. Methods

Three areal interpolation methods – AW, TDW, and PM – were used to estimate populations enumerated in census tracts in 2000 (source zones) within census tract boundaries of the 2010 Decennial Census (target zones) to create consistent spatial units across time. For the AW and PM methods, residential parcels were applied as limiting ancillary data to dasymetrically refine the enumerated population data within source zones prior to areal interpolation. This refinement was completed for both points in time for TDW. In the following sections, each method

is described in more detail.

3.4.1. Areal Weighting (AW)

3.4.1.1. AW-unrefined

The AW method estimates source population in target zone boundaries based on the overlapping area between source and target zones (i.e., intersections or "atoms"). An underlying assumption is that the population is uniformly distributed within a source tract (Equation 3.1):

$$\mathrm{pop_{st}} = (^{\mathrm{Area_{st}}}/_{\mathrm{Area_s}}) \times \mathrm{pop_s} \tag{3.1}$$

$\mathrm{Area_{st}}$ is the overlapping area between source zone s and target zone t, $\mathrm{Area_s}$ is the source zone area, $\mathrm{pop_s}$ is the source zone population, and $\mathrm{pop_{st}}$ is the population assigned to the atom. The population of target zone t is then simply calculated by summing up the population counts of all the atoms within it.

3.4.1.2. AW-refined

Dasymetrically refining source zones prior to areal interpolation is supported by residential parcel data, and modifies the underlying assumption as follows: population is homogenously distributed within the residential parcels of a source tract, and no population is assigned to non-residential parcels. This assumption is expected to be more realistic and allows more precise reapportionment of population. Residential parcels in 2000 were used to identify populated parts of each source zone and atom to modify Equation 3.1 as follows:

$$\mathrm{pop_{st}} = (^{\mathrm{Ref\_Area_{st}}}/_{\mathrm{Ref\_Area_s}}) \times \mathrm{pop_s} \tag{3.2}$$

$\mathrm{Ref\_Area_{st}}$ is the area of the residential parts of atom st, and $\mathrm{Ref\_Area_s}$ is the area of the

residential parcels in source tract s. Only enumeration geometry is used from the target census 2010. The "Built date" attribute was used to select and extract residential parcels existing prior to 2000.

3.4.2. Target Density Weighting (TDW)

3.4.2.1. TDW-unrefined

Schroeder (2007) introduced TDW as an areal interpolation method appropriate for temporal analysis of census data. The method extends "target count weighting", a term introduced by Schroeder (2007) for a technique described in Howenstine (1993) and Mugglin and Carlin (1998), which assumes that all target zones nest within source zones. To overcome this limitation, TDW makes two assumptions. First, within a source zone, the spatial distribution of the variable of interest *Y* among atoms is assumed to be proportionally the same as the distribution of an ancillary variable *Z* (Schroeder 2007). For example, if population is distributed in a 2:1 ratio between two tracts in 2010, it can be assumed that this ratio was the same between the two areas in 2000. The second assumption states that the density of *Z* in any atom equals the density of *Z* in the corresponding target zone:

$$z_{st}/\text{Area}_{st} = z_t/\text{Area}_t \tag{3.3}$$

where z_{st} and z_t indicate the ancillary variable Z for atom st and target tract t, respectively; and Area_{st} and Area_t are the corresponding areas. The inclusion of the second assumption might introduce some errors because the density of Z in a target zone might not be the same as the density in the atoms within it. However, this assumption is necessary to ensure that TDW is more robust and flexible than TCW. The accuracy of TDW is affected by the second assumption only when target zones intersect multiple source zones; otherwise, its accuracy only relies on the validity of

the first assumption (Schroeder 2007). In this study, the ancillary variable Z is the population distribution in 2010, and the variable of interest Y is the population distribution in 2000. Population in 2000 within 2010 census tract boundaries is estimated as follows:

$$y_t = \sum_s y_{st} = \sum_s \frac{(Area_{st}/Area_{t)} \times z_t}{\sum_\tau (Area_{s\tau}/Area_\tau) \times z_\tau} \times y_s \quad (3.4)$$

where y_t is the variable of interest for target zone t, y_{st} is the variable of interest for atom st, and y_s is the variable of interest for source zone s. The term τ is a target zone index, independent of t, which is defined for each target tract intersecting source zone s. As Equation 3.4 suggests, y_{st} is calculated based on the proportional distribution of the ancillary variable Z among atoms, and y_t is determined by aggregating all y_{st} values intersecting the target tract.

Based on previous studies, TDW often outperforms AW (Schroeder 2007, Schroeder and Van Riper 2013), suggesting that it is more reasonable to assume that the rate of population change is constant for atoms than to assume that population is homogeneously distributed within source zones.

3.4.2.2. TDW-refined

Refined TDW uses only residential areas within both source and target tracts. This refinement necessitates that the underlying assumptions of unrefined TDW be modified. In a first step, source and target zones are spatially refined using the areas occupied by residential parcels with an indicated built year before 2000 and 2010, respectively. Then TDW is applied to these refined areas. First, the refined areas within atoms in 2000 and 2010 are derived, separately. Next, it is assumed that the ratio of the refined population densities of atoms to the refined population densities of source zones remains the same in both years. It is important to note that the densities depend on the proportions of residential land within atoms and source zones, and these proportions

change over time. Based on this assumption, the refined population densities of atoms in 2000 can be calculated, while taking into account the adjusted second assumption in TDW that is the population density in 2010 in any refined atom equals the density in the corresponding target zone (i.e., the refined proportion of that zone). Finally, the population within target zones in the source year can be derived by aggregating the population values of the atoms within them. This adjustment of areas and population densities is expected to improve the estimation accuracy based on recent studies using land cover data for refinement in the source census year (e.g., Holt *et al.* 2004, Ruther *et al.* 2015).

3.4.3. Pycnophylactic Method (PM)

3.4.3.1. PM-unrefined

The PM-unrefined method assumes the existence of a smooth density function and incorporates the densities of adjacent zones. The density function must be pycnophylactic, i.e., volume-preserving: it must reproduce the original value of a source zone if applied to it. This function is defined as follows. Let p_k be the population of zone k, $Area_k$ the area of zone k, Den_{ij} the density in cell ij, and α the area of a cell. The following equations fulfil the pycnophylactic condition (Lam 1983):

$$\sum_{ij} \alpha Den_{ij} q_{ij}^{k} = p_k \tag{3.5}$$

$$\sum_{ij} \alpha q_{ij}^{k} = Area_k \tag{3.6}$$

$$\sum_{k} q_{ij}^{k} = 1 \tag{3.7}$$

In these equations, q_{ij}^{k} equals 1 if the cell ij belongs to zone k, and 0 otherwise. A cell is assigned to the zone that occupies the maximum area of that cell. Equation 3.7 also guarantees that

each cell ij belongs to no more than one zone. The smooth density function is obtained by minimizing the sum of squared gradients in x and y directions (i.e., Dirichlet's integral):

$$\iint((\delta z/\delta x)^2 + (\delta z/\delta y)^2)\delta x\delta y \quad (3.8)$$

In order to reach a smooth density function that preserves the volume for a population mapping application, the following process is iterated:

1. Overlay source zones with a raster grid of finer resolution than the source zones;
2. Divide the population of each source zone by the number of cells falling within it to assign an equal initial population to each cell;
3. Apply a smoothing function that replaces the initial value of each cell by the mean of its neighborhood;
4. Sum the modified values in each zone and compare the result to the original value of the zone;
5. Adjust the modified values so that they sum to the original value of the zone;
6. Repeat steps 3, 4, and 5 until a pre-defined stopping criterion is satisfied;
7. When the stopping criterion is reached, and the system is stable, aggregate cell values to any desired set of target zones.

This study applied the aforementioned procedure to census tracts in 2000 and interpolated population values to census tract boundaries in 2010. A circular neighborhood with a radius of 25 cells (750 meters) was used for both the unrefined and refined pycnophylactic methods following recent research results in related dasymetric mapping studies (e.g., Kim and Yao 2010, Ruther *et al.* 2015). The number of iterations was set to 25, at which there would not typically be any noticeable difference between the final estimated population values and the values from the

previous run.

3.4.3.2. PM-refined

Kim and Yao (2010) proposed the dasymetrically refined pycnophylactic method for non-temporal small area estimation. This hybrid method created smooth surfaces dependent on the neighborhood of each cell, and was defined over dasymetrically refined areas, thus allowing more precise depiction of populated areas and neighborhood relations. The refined pycnophylactic method in this study uses the same iterative process as the unrefined pycnophylactic method. However, instead of dividing the population of each source zone by the number of all cells within it, the method divides the zone population by the number of all cells comprising residential parcels within it. To implement this method, the census tracts and residential parcels in 2000 were rasterized, the number of residential parcel cells in each source tract was calculated, and the tract population was divided by this number. In this study, a constraint was applied, after applying the smoothing function (step 3 above), to keep non-residential cells "unpopulated" following each iteration. Thus, the final population surface was not as smooth as that described by Kim and Yao (2010), but non-residential cells remained "unpopulated" until the algorithm converged. After the pycnophylactic iterative process reached a stable surface, all cell values in each target zone were aggregated to compute target zone estimates. As was the case in the unrefined version, the system stabilized after 25 iterations to create the final population surface.

3.4.4. Validation method

In order to validate the methods, population counts in 2000 were derived for each 2010 tract boundary using 2000 census blocks, a finer resolution census unit often employed in validation efforts for small area estimation (e.g., Reibel and Bufalino 2005, Reibel and Agrawal 2007, Schroeder 2007, Tapp 2010). Because blocks in 2000 were not completely nested in tract

boundaries of 2010 in some instances, AW was used to reapportion block populations to different subparts, assuming that population was distributed homogeneously within blocks. The use of AW in census blocks is assumed to be reliable because of the very small size of census blocks. After obtaining block-measured (validation) and estimated (interpolated) values for target tract boundaries, four statistics were used to evaluate the accuracy of each method. These statistics were Mean Absolute Error (MAE), median absolute error, 90th percentile of absolute errors, and Root Mean Square Error (RMSE). MAE was calculated by averaging the absolute differences between estimated and measured values of target tracts. Median absolute error was determined by taking the value of the 50th percentile of absolute errors. RMSE was calculated based on absolute differences between estimated and measured values of target tracts by taking the square root of the mean of squared differences.

3.5. Results and discussion

3.5.1. Evaluating results using block statistics

Table 3.1 shows the areal interpolation error estimates based on changed census tracts for each method.

Table 3.1. Error measures of unrefined and refined methods for changed target tracts.

Method	MAE	Median Absolute Error	RMSE	90th Percentile Error	Standard Deviation
AW	138	36	321	496	290
Refined AW	148	7	402	568	374
TDW	101	27	230	236	207
Refined TDW	70	6	177	180	162
PM	144	34	321	522	287
Refined PM	119	5	312	357	288

As can be seen, refining AW does not result in better overall accuracies, and the MAE and

RMSE measures are higher for refined AW than for unrefined AW. This indicates a decrease in the overall accuracy in refined AW results. However, the median absolute error for refined AW is lower than for unrefined AW. This indicates that refined AW produces more accurate estimates for the first half of target zones, ordered by absolute error estimates. More quantitatively, refined AW produces estimates that have equal or lower absolute errors than AW for 92 of the total 129 changed target tracts. But some large estimation errors exist in the right tail of the absolute error distribution and cause lower overall accuracy measures.

One possible explanation for these findings can be found in the limiting ancillary data type used here, and the assumption that all parcels contain the same type of residential unit. Areal extents of parcel units can be very large in sparsely populated rural settings. On the other hand, densely populated multi-unit buildings (apartments and condominiums) can be located in relatively small parcel units, mostly in more urban settings. Therefore, large sparsely populated parcels are incorrectly assigned higher population estimates than small densely populated counterparts when only their areas are utilized. This problem becomes even more serious if, in two unrefined areas with a similar areal extent, one encompasses mainly sparsely populated units and the other mostly densely populated buildings (Figure 3.2). The bottom target tract in Figure 3.2 includes highly populated (multi-unit) parcels and the top one includes large sparsely populated (single-family) parcels. Residential parcels of the bottom tract cover less area than those of the top one, while the unrefined target tracts are very similar in extent. The total area of the residential parcels in the bottom target tract is 0.97 km^2, and the population count of that tract is 3809 based on Census 2000 blocks. These numbers are 2.2 km^2 and 2952, respectively for the top target tract. Without considering the building type, more population from the source tract is erroneously assigned to the top target tract (overestimation) while fewer people are assigned to the bottom target tract

(underestimation). In unrefined AW, this assignment error is less pronounced because of the similar areal extents of the unrefined areas; and thus the prediction error is lower for unrefined AW than for refined AW.

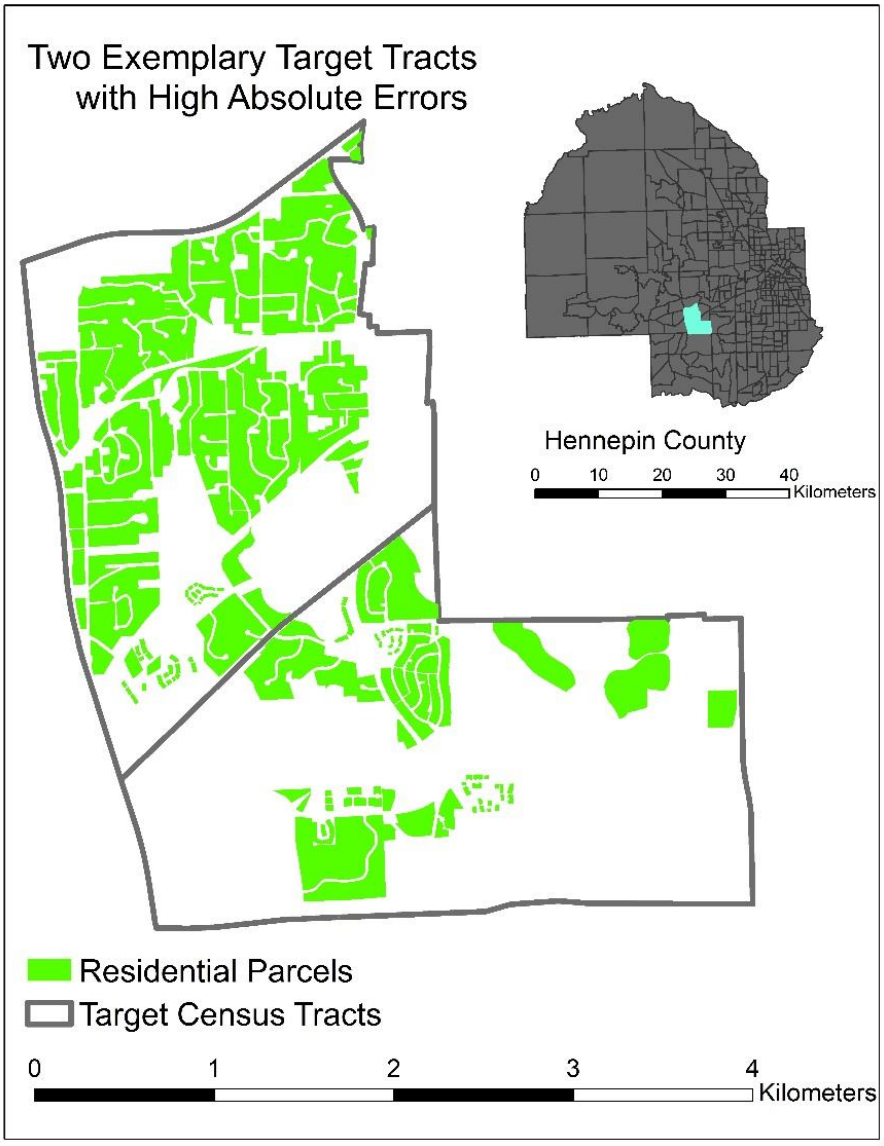

Figure 3.2. Two target tracts with different types of residential parcels.

Table 3.1 also shows that TDW has generally lower errors than AW in both the unrefined and refined model versions, indicating that the assumptions underlying TDW are more realistic for the study area. Refined TDW shows the lowest error estimates for changed tracts in most

situations. Moreover, while refined AW shows inconsistent performance in Table 3.1, refined TDW shows a consistent refinement effect for all error measures. One reason for this superior performance could be that the problematic assumptions of TDW are successfully mitigated in some part through parcel-based refinement across a relatively short time period. The spatial refinement adjusts for densities among residential areas and represents a more reasonable assumption than both refined AW and unrefined TDW. Moreover, refined TDW is the only method that incorporates residential data from both time periods.

Refined PM also consistently reduces the error measures relative to unrefined PM. This consistent refinement effect of the method shows that excluding non-residential cells and only running the pycnophylactic algorithm on the more realistic population distribution creates a more reliable population surface. Consequently, aggregation of cell values of this refined population surface within target zones results in more accurate population estimates than those from unrefined PM. While most errors were consistently higher than those of TDW and refined TDW, refined PM shows the lowest median absolute error among all methods tested, which suggests the interesting potential of this method.

The maps in Figure 3.3 illustrate the distribution of absolute errors across target zones. The maps confirm the above described findings visually. The absolute errors are lower for refined AW than for unrefined AW for small to medium error classes. However, the existence of a few target tracts in the center and southern parts of the study area with large absolute errors is the main reason for lower overall estimation accuracy of refined AW. Moreover, absolute errors are higher for rural areas in unrefined AW than in refined AW. This pattern is manifested by smaller absolute errors in target tracts away from Minneapolis in the refined AW map. The visual differences between unrefined TDW and refined TDW are not as noticeable as in the AW maps, but rural target tracts

have consistently lower absolute errors in refined TDW than in unrefined TDW. The absolute error map of refined TDW shows that not only are small absolute errors lower than in other maps, but target tracts with high absolute errors are also less frequent. The absolute error maps of unrefined and refined PM also demonstrate that the absolute errors are lower in refined PM than unrefined PM, particularly in rural census tracts away from Minneapolis.

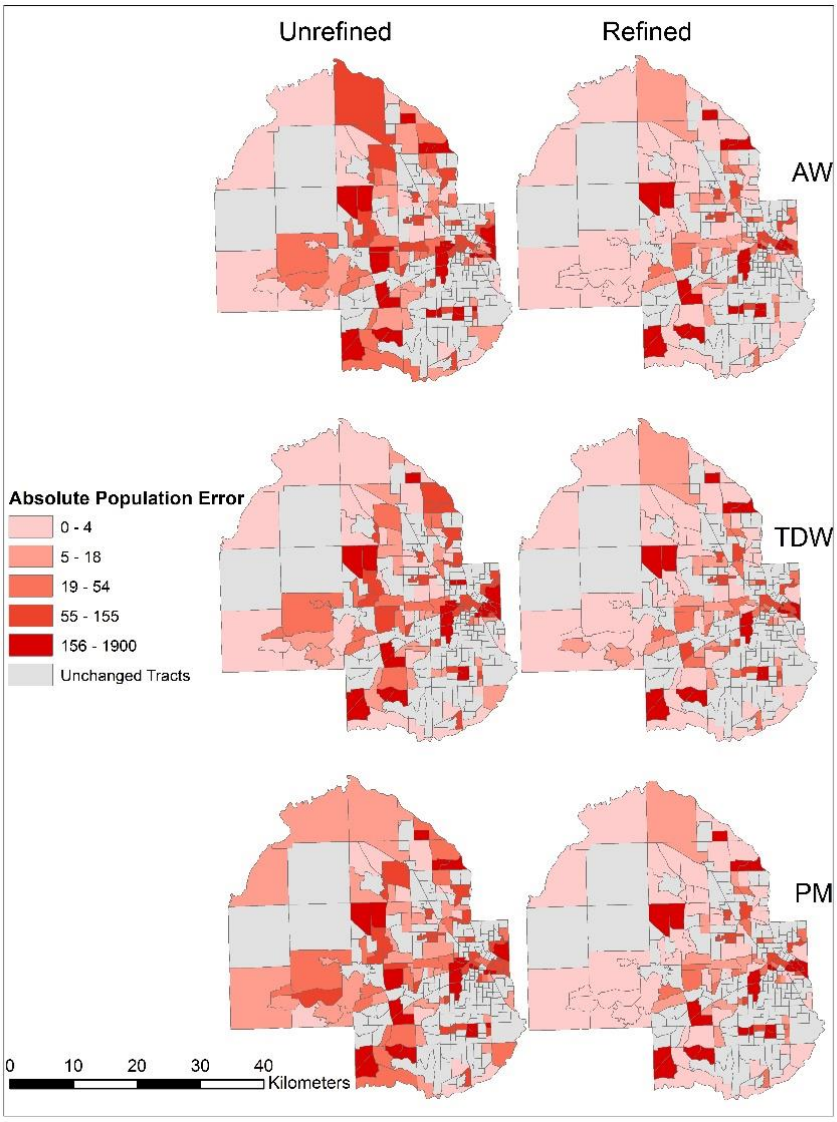

Figure 3.3. Absolute error maps of parcel-refined methods for changed target tracts in comparison to unrefined methods.

Figure 3.4 shows which method was the most accurate method in each changed target tract. AW and PM do not appear because there was no target zone in which either of these methods was most accurate. The map shows that refined TDW is the most accurate method for the majority of target zones. However, there exists a noticeable number of census tracts in which a method other than refined TDW has the lowest absolute error.

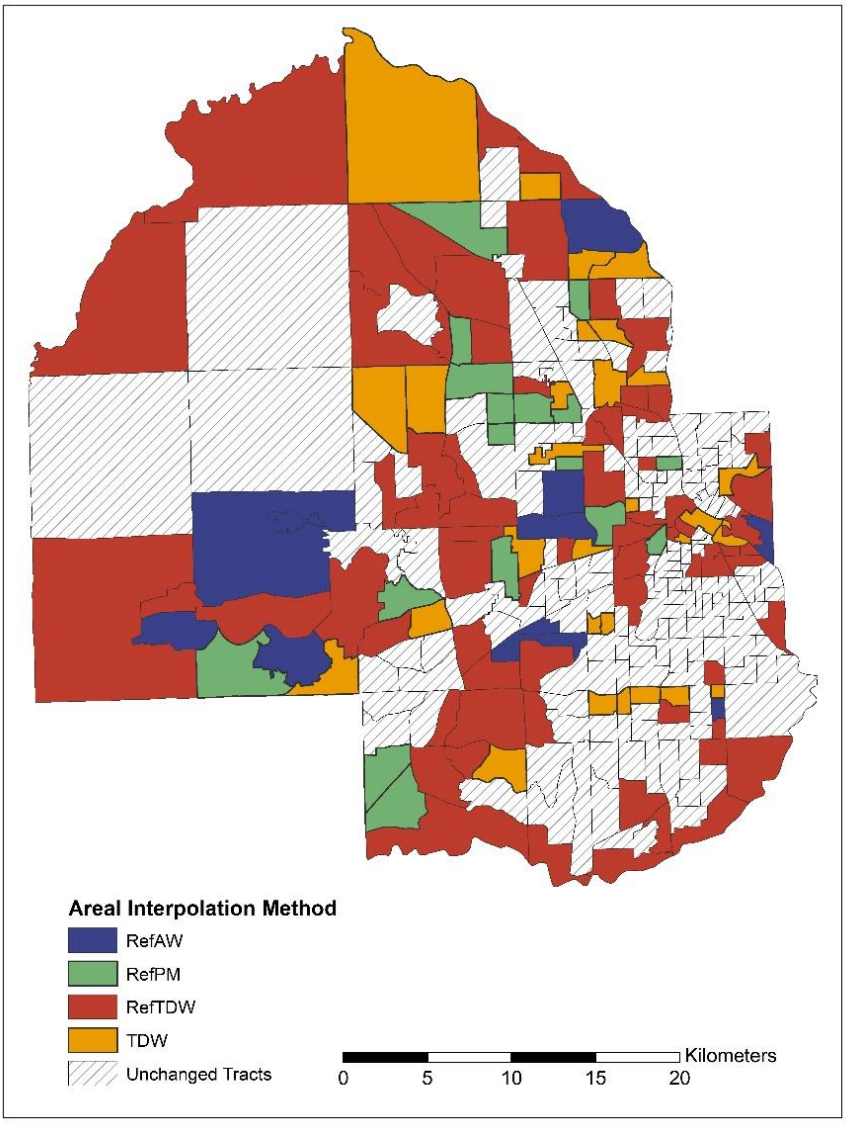

Figure 3.4. Map of the most accurate method in each changed target tract.

3.5.2. Comparing results to models using land cover based refinement

In order to better evaluate the reported temporal interpolation errors, the results described above were compared to model outcomes that were based on a different ancillary variable for dasymetric refinement, namely developed land categories derived from the 2001 National Land Cover Database (Homer *et al.* 2007). Table 3.2 shows the estimation errors for the dasymetrically refined models using developed land for the same changed tracts.

Table 3.2. Error measures of NLCD-refined methods for changed target tracts.

Method	MAE	Median Absolute Error	RMSE	90th Percentile Error	Standard Deviation
Refined AW	115	21	232	445	201
Refined TDW	89	18	197	247	176
Refined Pycnophylactic	130	26	280	492	248

When comparing results between Tables 3.1 and 3.2, it can be seen that NLCD-refined AW has lower MAE, RMSE, and 90th percentile error than parcel-refined AW. However, the median absolute error of parcel-refined AW is much lower than NLCD-refined AW. This indicates that while parcel-refinement decreases the absolute error values of the first half of error-ordered tracts, the existence of a few tracts with high absolute errors at the right tail of the absolute error distribution causes the overall measures to be higher. In contrast, NLCD-refined TDW has higher values for all error measures than parcel-refined TDW. This indicates that TDW assumptions using parcel refinement are more realistic with respect to block-level population counts than using NLCD refinement in the study area. Finally, all error measures except RMSE are lower in parcel-refined PM than in NLCD-refined PM. Thus the population estimates of the target tracts are generally more accurate in parcel-refined PM than in NLCD-refined PM in this study.

Figure 3.5 shows the absolute error maps of NLCD-refined methods for changed tracts. The maps confirm the results outlined in the previous paragraph. Moreover, the absolute error values of large rural census tracts are generally higher in NLCD-refined AW, TDW and PM maps than their parcel-refined counterparts, a pattern manifested by large rural target tracts in the northern and western parts of the study area.

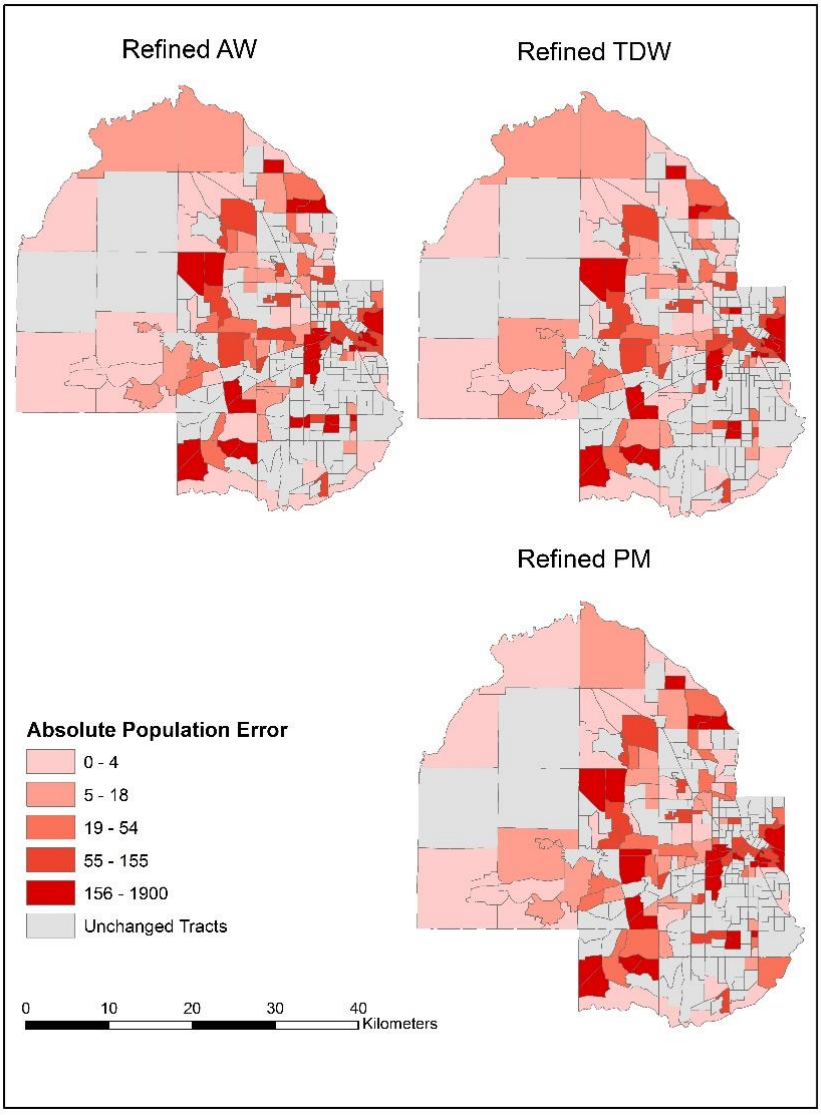

Figure 3.5. Absolute error maps of NLCD-refined methods for changed target tracts.

3.6. Conclusions and future directions

The incorporation of ancillary data for dasymetric refinement prior to areal interpolation for temporal analysis has great potential, as has been shown in recent studies (Holt *et al.* 2004, Ruther *et al.* 2015). The employment of parcel data as a limiting ancillary variable has the potential for distinct improvements relative to land cover data for three reasons. First, cadastral data generally depict residential land parcels more reliably in rural areas. Second, parcel data is updated more frequently than is national land cover, to accommodate tax records, zoning variations, etc. Third, when parcel data contain information about the built date of a residential building, one can approximate residential development for different points in time.

However, the use of residential parcels as the limiting variable comes with several challenges. The three areal interpolation methods demonstrate differences that appear to be based on the choice of ancillary data, with AW displaying better results for parcel data when small errors are present, and better results for land cover data elsewhere. TDW displays consistently better results for parcel-based refinement. PM shows better results for parcel ancillary data based on three metrics. Further research for study areas in other demographic conditions (population growth or fast decline) could be undertaken to assess whether method-produced differences occur in other situations. Alternatively, interpolations could be limited to rural or urban areas to assess where these errors are most pronounced, for each method.

In rural areas, the use of residential parcels as the ancillary dataset enables the reliable identification of residential development where remote sensing-based land-cover products such as NLCD may misclassify small patches of developed land. However, in line with results found by Leyk *et al.* (2014), parcel refinement tends to overestimate the areal development extent in rural settings. Figures 3.3 and 3.5 suggest that parcel-refined methods result in lower absolute errors

than NLCD-refined methods for rural target tracts. However, to reach such a conclusion with confidence, more study areas have to be included to establish reliable quantitative measures. Future research will compare parcel and NLCD based refinements for temporal analysis in more detail and for various study areas by dividing them into urban and rural target zones. This will better delineate the performance of the areal interpolation methods in these different settings and the role of the different ancillary variables.

Residential parcels possess different housing types, ranging from single family units to multi-level condominiums or apartment buildings; this information has been ignored in this study. Using residential parcel extents as a limiting variable implies that population is redistributed equally over all residential entities without considering these different housing types. This simplification can cause biased estimates through the spatial refinement process and may mislead the analysis in some situations resulting in increased estimation errors in some affected areas.

Finally, there is inherent uncertainty due to the temporal slicing of residential parcels, as the built-year attribute may not reflect the date of the first construction for those parcels in which a building had been torn down and later rebuilt, or for a single family parcel that is converted to multi-family use. The date commonly refers to the main structure in a parcel and does not reflect extensions or remodeling. It is acknowledged that this could be a source of uncertainty if the built-year attribute is handled differently in other study areas. Furthermore, in some urban areas the type of land use could have changed over time, as for example in the development of mixed commercial – residential land use parcels in some expanding urban areas.

Disadvantages related to urban – rural distinctions are mainly caused by the large extent of rural parcels. A processing step that achieves refinement of these rural residential parcel units using additional ancillary variables would be a natural solution to this problem. Moreover, since

building types are closely related to population density, the inclusion of a building type attribute in dasymetric refinement has the potential to further decrease estimation errors in temporal analysis. Such an extension would transition parcel data from a binary limiting ancillary variable (residential or non-residential) to a "related" ancillary variable and allow the incorporation of different categories of residential type that may be related to varying population counts.

Future research will focus on the inclusion of residential unit types (e.g., multi-unit or single family), addressing the problem of large areal extents through additional spatial refinement, and the extension of the time period considered. Alternative approaches such as the Expectation Maximization algorithm (Dempster *et al.* 1977) for residential parcels will also be explored. The method has been shown promising when applied to land cover data in recent research (Schroeder and Van Riper 2013, Ruther *et al.* 2015). As Figure 3.4 implies, no method has the most accurate estimates for all target tracts. Therefore, a hybrid method that utilizes the complementary effects of different methods may result in more accurate estimates (similar to Schroeder and Van Riper (2013)). This idea constitutes yet another area for further research. Finally, if access to the temporal attribute can be facilitated more broadly, temporal analysis as described in this paper would have great potential to be deployed to long time periods over large areas. Such a temporal component can potentially further extend time periods covered in recent studies on population change analysis such as 1970-2010 in Logan *et al.* (2014), 1980-2010 in Schroeder and Van Riper (2013), and 1990-2000 in Schroeder (2007).

Acknowledgements

This research is funded by the National Science Foundation: “Collaborative Research: Putting People in Their Place: Constructing a Geography for Census Microdata”, Project BCS-0961598 awarded to University of Colorado – Boulder. Access to parcel data for Hennepin County was provided courtesy of Minnesota Population Center, and their data support is gratefully acknowledged.

Chapter IV

Consistent Population Estimation within Changing Census Boundaries: Enhancing Interpolation Frameworks through Dasymetry

Abstract[1]

To assess micro-scale population dynamics effectively, demographic variables need to be available over temporally consistent small area units such as census tract. However, fine-resolution census boundaries often change between survey years and methodological solutions to transfer population counts aggregated over source zones in one census year to target zone boundaries from another census year are in high demand.

This research proposes advanced areal interpolation methods that incorporate dasymetric refinement using ancillary variables to create consistent population counts in 1990 and 2000 (source zones), respectively, within tract boundaries of the 2010 census (target zones), for five counties in the US. The counties are characterized by different demographic processes and include Hennepin County, Minnesota, Mecklenburg County, North Carolina, Broward County, Florida,

[1] This chapter was submitted as a journal paper to the International Journal of Geographic Information Science (IJGIS) by Hamidreza Zoraghein and Stefan Leyk in March 2017.

Hillsborough County, Florida and Worcester County, Massachusetts.

Three levels of spatial (or dasymetric) refinement of source and target zones are evaluated in this study. First, residential parcel boundaries are used as a limiting ancillary variable prior to regular areal interpolation for temporal analysis including Areal Weighting (AW) and Target Density Weighting (TDW). Second, Expectation Maximization (EM) and Enhanced EM (EEM) leverage housing types of residential parcels, their area and density measures, as related ancillary variables to incorporate nuanced and more complex associations between different types of residential parcels and population estimates. Finally, a third refinement strategy that aims at mitigating the overestimation effect of large residential parcels in rural areas, uses road buffers and developed land cover classes from the National Land Cover Database (NLCD) as additional conditional ancillary variables. Different validation metrics, namely Mean Absolute Error (MAE), median absolute error, Root Mean Square Error (RMSE) and 90% percentile of absolute error are computed using census block statistics in 1990 and 2000 to evaluate the effectiveness of each method.

Results suggest great potential of all three levels of spatial refinement in reducing estimation errors for all five counties. Refined TDW and EEM are generally the best-performing methods from the first and second spatial refinement levels with EEM being superior for the longer time period of 1990 to 2010. The third refinement is generally effective in reducing the estimation errors in rural areas. The results allow for a first insight on the potential accuracy that could be achieved in varying geographic and demographic settings using different combined approaches if comparable ancillary data would become available nationwide. Such improved consistent population estimates over long time periods would provide the basis for more advanced demographic research, comparing population characteristics over time at fine spatial resolution.

Keywords: Areal Interpolation; Dasymetric Modeling; Population Estimation; Spatial Refinement; Spatial Analysis

4.1. Introduction

The ability to create precise spatial distributions of demographic variables such as total population or population density is critical for future efforts in sustainable development. Demographic estimates are vital for urban planning and resource management decisions, including the allocation of food and medical supplies, access to public services, transportation and regional development (Su *et al.* 2010). Moreover, population is an important factor, driving land-use/land-cover change, such as urbanization, deforestation, or afforestation, which affect environmental conditions such as air pollution, soil erosion and water quality (Li and Lu 2016). Although large-scale population products such as the Gridded Population of the World (GPW) (Balk and Yetman 2004), LandScan Global (Dobson *et al.* 2000) and LandScan USA (Bhaduri *et al.* 2007) already exist, they have rather coarse resolution and underlie error-prone population allocation approaches. Thus, these data products do not eliminate the need for establishing more detailed population distributions using regional datasets and advanced algorithms in order to address local management issues (Su *et al.* 2010, Jia *et al.* 2014) and better reflect operational scales of the demographic processes of interest. Moreover, to be able to implement effective measures of development and growth trends, there is an urgent need in fine-resolution population distributions that are compatible over time. Due to changing enumerations (Schroeder 2007) among census surveys, this remains a persistent challenge in the field. However, recent research has begun to invest in the development of analytical solutions based on areal interpolation to produce fine-resolution population distributions for multiple points in time within temporally compatible analytical units (Schroeder 2007, Schroeder and Van Riper 2013). While the integration of dasymetric refinements in such temporal analysis has shown great potential (Ruther *et al.* 2015, Zoraghein *et al.* 2016) these attempts are often constrained to the use of limiting ancillary variables. More research is needed to identify methodological solutions to make more effective

use of advanced dasymetric modeling techniques for spatiotemporal interpolation that also employ related ancillary variables in order to make use of nuanced relationships between ancillary variables and population attributes for more accurate estimation.

Dasymetric mapping is a special type of areal interpolation, employing ancillary datasets correlated with the variable of interest to map the variable from a set of aggregated source zones in a choropleth map to a set of target zones that reflect its actual distribution more precisely (Wright 1936, Eicher and Brewer 2001, Mennis 2003, Langford 2006). The methodology, which is covered extensively in the literature, is commonly used to downscale population from large aggregated census units to smaller target zones, and has many applications, including crime analysis (Mennis 2016), environmental health justice (Maantay *et al.* 2007, Mennis 2015) and historical population estimation (Holt *et al.* 2004, Buttenfield *et al.* 2015, Ruther *et al.* 2015, Pavía and Cantarino 2016) to name a few. With widespread availability of various remote sensing products and growing analytical capabilities in geospatial software tools, dasymetric mapping has become a common spatial analytical approach for population reallocation and demographic small area estimation (Mennis 2009). Examples of ancillary variables employed in dasymetric mapping include but are not limited to land-cover (Mennis and Hultgren 2006), road networks (Reibel and Bufalino 2005), reflectance data from high resolution satellite imagery (Alahmadi *et al.* 2015), imperviousness surfaces (Zandbergen and Ignizio 2010, Li and Lu 2016), address points (Tapp 2010) and cadastral data (Maantay *et al.* 2007, Jia *et al.* 2014, Jia and Gaughan 2016).

The main focus of the current paper is on the integration of more advanced dasymetric refinement in spatio-temporal interpolation of census data building on recent research. This recent research has demonstrated how the performance of areal interpolation methods such as areal weighting (AW) or target density weighting (TDW) can be improved by employing spatial

refinements through dasymetric mapping. Dasymetrically refined areal interpolation for temporal analysis essentially transfers demographic variables from one set of source zones at one point in time to a set of target zones at a second point in time, and makes use of ancillary variables such as land cover data and residential parcels to geometrically adjust these zones. This refinement creates time series of more accurate population estimates over consistent small census units but still shows large errors in regions with rapid or unexpected population growth (Buttenfield *et al.* 2015, Ruther *et al.* 2015, Zoraghein *et al.* 2016). The present study aims to enhance the idea of dasymetric refinement in spatiotemporal interpolation of demographic variables by operationalizing related ancillary variables and thus creating temporally consistent population estimates with low interpolation error. The product benefits disciplines such as demography, geography, economics, political science and sociology. The methodology is applied in five U.S. counties for both 1990-2010 and 2000-2010 time periods to transfer total population from census tracts in 1990 and 2000 (source zones) to census tract boundaries in 2010 (target zones), respectively.

4.2. Study area and data

4.2.1. Study area

The five counties used in this study to test the methods represent different geographic and demographic settings that can be characterized by the urban/rural proportion of population, the extent and the population growth rate. They include Hennepin County, Minnesota, Mecklenburg County, North Carolina, Broward County, Florida, Hillsborough County, Florida and Worcester County, Massachusetts. Figure 4.1 depicts the five selected counties as well as their census tracts in 2010 and the boundaries of their seats.

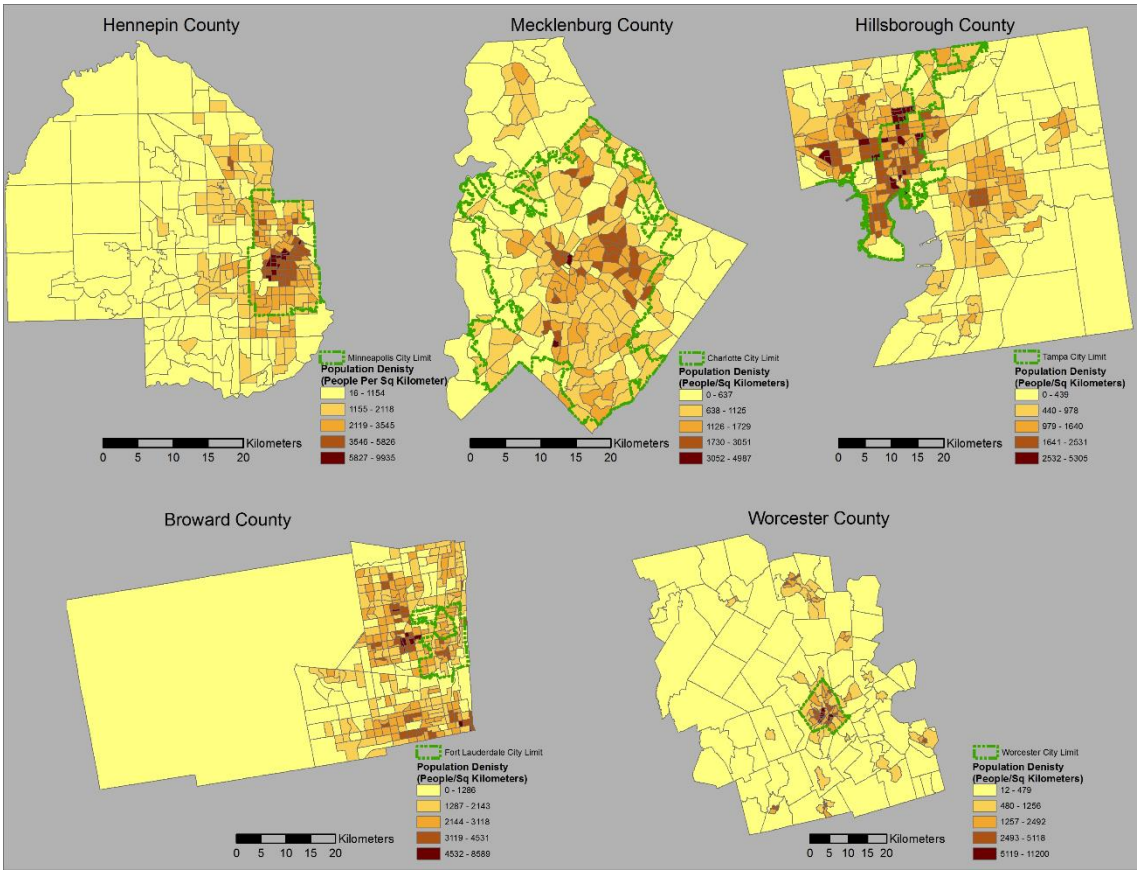

Figure 4.1. The study area and target census tracts.

Hennepin County includes the urban region of Minneapolis in the east and rural areas in the west. According to the U.S. Census, a low total population growth rate of 11% (from 1,032,431 to 1,146,195) has been observed between 1990 and 2010.

Mecklenburg County includes the urban region of Charlotte at the center, which covers the majority of the county area, and rural census tracts on the fringe. The total population growth rate in this fast-growing county is 80%, an increase from 511,433 to 919,628 during the 1990-2010 time period.

Hillsborough County contains urban census tracts of Tampa in the west and rural areas in the east. Densely populated census tracts in this county are not limited to the city limits of Tampa

but appear scattered all over the county except in the eastern parts. The population of the county has observed a rather fast growth rate of 47%, increasing from 834,027 to 1,229,226 during the 1990-2010 time period.

Most of Broward County is composed of census tracts with medium to high population density in the west. There is also a large sparsely populated census tract in the east. The population of the county has grown by 39%, an increase from 1,255,462 to 1,748,066 between 1990 and 2010.

Worcester is the county with the highest proportion of sparsely populated census tracts with only a few more densely populated tracts within the city limits of Worcester at the county center. The population of the county has grown by 12% and increased from 709,705 to 798,552 during the 1990-2010 time period.

The selected counties include different proportions of high and low population density areas, making them ideal case studies to evaluate the performance of each method under different conditions. More specifically, Mecklenburg, Hillsborough and Broward represent counties with fast population growth whereas Worcester and Hennepin indicate counties with medium/low population growth rate. Testing the methods across these different population development histories will help better understand methods performance in different demographic settings.

4.2.2. Data

The boundaries of census tracts in 1990, 2000 and 2010 along with their total population counts found in the census summary files are the focus in this study. Census blocks represent the smallest enumeration units published by the Census. Therefore, blocks in 1990 and 2000 as well as their population values are used as reference data to evaluate the estimated total population counts at the tract level. The tract-level and block-level population values and boundaries for 1990

were retrieved from the National Historical Geographic Information System (NHGIS) (Minnesota Population Center 2016) whereas population counts and boundaries for 2000 and 2010 were extracted from U.S. Census Bureau (2010) and U.S. Census Bureau (2016), respectively.

Three ancillary variables associated with the distribution of total population are used in this study. They include residential parcels of each study area accessed from the corresponding county or state GIS data portal (Hennepin County GIS 2016, Mecklenburg County GIS 2013, University of Florida GeoPlan Center 2016, MassGIS 2016), the National Land Cover Database (NLCD) in 1992, 2001 and 2011 (Multi-Resolution Land Characteristics 2016) and TIGER/Line road networks in 2000 and 2010 (U.S. Census Bureau 2016). NLCD is a Landsat based national land cover dataset at 30m resolution. Its primary objective is to provide nationally complete, current, consistent, and public domain information on the nation's land cover. The dataset presents different land cover types in different classes (Homer *et al.* 2015).

4.3. Methods

The spatio-temporal areal interpolation methods in this study are divided into three categories according to different modes of spatial refinement and the number of ancillary variables used. Starting with the use of limiting ancillary variables (first spatial refinement), the second category uses related ancillary variables while the third category implements an a posteriori refinement in rural settings.

4.3.1. First spatial refinement using limiting ancillary variables

The first level of spatial refinement in spatio-temporal interpolation is employed through the use of limiting ancillary variables such as developed land cover classes in the NLCD or residential parcel footprints. Well-known methods such as AW (Goodchild and Lam 1980) and TDW (Schroeder 2007) are adjusted to sub-areas of source and target zones delineated by

residential parcels of the five study areas following the approach described in Zoraghein *et al.* (2016) for comparison to the next level of spatial refinement.

AW is the most basic areal interpolation method and assumes the population density is constant within source zones. The method estimates source population in target zone boundaries based on the overlapping area between source and target zones (i.e., intersections or "atoms"). The population of each target zone is then simply calculated by summing up the population counts of all the atoms within it.

Spatially refining source zones prior to areal interpolation is supported by different ancillary variables and modifies the underlying assumption as follows: population is homogenously distributed within the developed land of a source zone, and no population is assigned to non-developed parts. This assumption is expected to be more realistic and generally results in more precise reapportionment of population counts.

Schroeder (2007) introduced TDW as an areal interpolation method appropriate for temporal analysis of census data. TDW is based on the assumption that the spatial distribution of population densities among atoms within a source zone in the source year remains proportionally the same over time. For example, if population density is distributed in a 2:1 ratio between two atoms in 2010, it is assumed that this ratio was the same in 2000.

Based on previous studies, TDW often outperforms AW (Schroeder 2007, Schroeder and Van Riper 2013), suggesting that it is more reasonable to assume that the ratio of population densities of atoms in one source zone remains constant than to assume that population is homogeneously distributed within source zones.

Refined TDW employs developed/built-up areas within both source and target zones. This

refinement implies that the underlying assumption of unrefined TDW be modified. In a first step, source and target zones are spatially refined using the developed areas labeled by the ancillary variable. Then TDW is applied to these refined areas under the assumption that the ratio of refined population densities of atoms to refined population densities of source zones remains the same over time. While refined AW uses developed areas only in the source year, refined TDW incorporates this refinement in both the source and target years.

4.3.2. Second spatial refinement using related ancillary variables

The above spatial refinement does not differentiate between different types or densities of residential units such as low-density single-family parcels as compared to high-density condos. Since it is well-known that the relationships between population and ancillary variables are not binary in nature, an approach to employ ancillary data to reduce and amplify the likelihood of population and estimates of population density would have great potential to further improve the accuracy of such estimates. This type of association is addressed below by incorporating related ancillary variables into the dasymetric refinement for spatio-temporal interpolation based on the Expectation Maximization algorithm.

4.3.2.1. EM with control zones based on residential types

Expectation Maximization (EM) (Dempster *et al.* 1977) can be used as an iterative process to optimize population density weights under different conditions defined by the ancillary variable, thereby offering an appropriate framework for implementing the second spatial refinement.

The EM algorithm provides a robust framework for model fitting and maximum likelihood estimation in settings of incomplete data. First, the expectation (E) step “completes” the data by computing the conditional expectation for missing data, given a set of observed data and estimated model parameters. The maximization (M) step then fits the model, estimating model parameters

by maximum likelihood given the "complete" data from the E step. A feedback loop between E and M steps is established and repeated until convergence (Schroeder and Van Riper 2013).

Flowerdew and Green (1994) demonstrated how the EM algorithm can be applied in areal interpolation applications, and Ruther *et al.* (2015) applied EM in temporal interpolations using land cover classes as ancillary data to define control zones. In this study, control zones are defined by residential parcels that have the same housing type and then used to calculate the population density weight for each control zone. This approach is justified by the expectation that different housing characteristics can be related to varying average population densities, and accounting for such variation is expected to improve the resulting estimation within target zones.

In the E step, the algorithm estimates the values of $\widehat{y_{sc}}$, i.e., the population counts for the intersections between source zone s and control zone c:

$$\widehat{y_{sc}} = y_s \left(\widehat{\lambda_c} A_{sc} \Big/ \sum_k \widehat{\lambda_k} A_{sk} \right) \tag{4.1}$$

Where y_s is the population count of source zone s, $\widehat{\lambda_c}$ is the estimated density of control zone c, A_{sc} is the area of intersection between s and c, and k is a second control zone index, independent of c to reflect all control zones intersecting s. The first E step is essentially similar to AW and assumes equal weights for all housing types. Then, the M step re-estimates all λ_c values using the equation below:

$$\widehat{\lambda_c} = \left(\sum_s \widehat{y_{sc}} \right) \Big/ A_c \tag{4.2}$$

The estimates of $\widehat{\lambda_c}$ from the M step are used to re-estimate $\widehat{y_{sc}}$ in the next E step, which is followed by another M step, and so on until the system converges. The algorithm stops when the

maximum absolute difference between the current population weights and those calculated from the previous run is less than 0.001. Finally, $\widehat{y_{sc}}$ values are used to calculate the population count for target zone t ($\widehat{y_t}$):

$$\widehat{y_t} = \sum_s \sum_c {(A_{tsc}\widehat{y_{sc}})}/{A_{sc}} \quad (4.3)$$

Where A_{tsc} is the area of the intersection between target zone t, source zone s, and control zone c.

4.3.2.2. Enhanced EM based on more homogeneous control zones

EM assumes that the population density is constant within each control zone. However, this assumption can become problematic. For example, if the residential parcels of the same type that form a control zone are diverse in area or number of units, the assumption of constant population density for the whole control zone becomes unrealistic. This research introduces Enhanced EM (EEM) to address this issue.

EEM categorizes control zones into more similar and homogeneous sub control zones based on the area of units and unit density criteria. First, the approach identifies those residential-type control zones with the highest number of residential parcels to guarantee a sufficient number of parcels per sub control zone. For those control zones that show the highest variation in the area measure of their underlying parcels, sub-control zones are created based on area quartiles. For some control zones such as condominiums, the number of units per parcel can be derived. Therefore, a unit density measure can be computed by dividing the number of units by the area of the encompassing parcel and then used for creating sub-control zones but this time based on quartiles of unit density. This new set of more homogeneous sub control zones is input to the EM algorithm as described above.

For each study area and time period, EEM is simulated over different combinations of eligible control zones, testing different numbers of combined type-area and type-density categories to identify the optimal solution. The best-performing model is the one that minimizes error metrics that will be introduced in Section 3.4. Figure 4.2 illustrates the workflow of EEM.

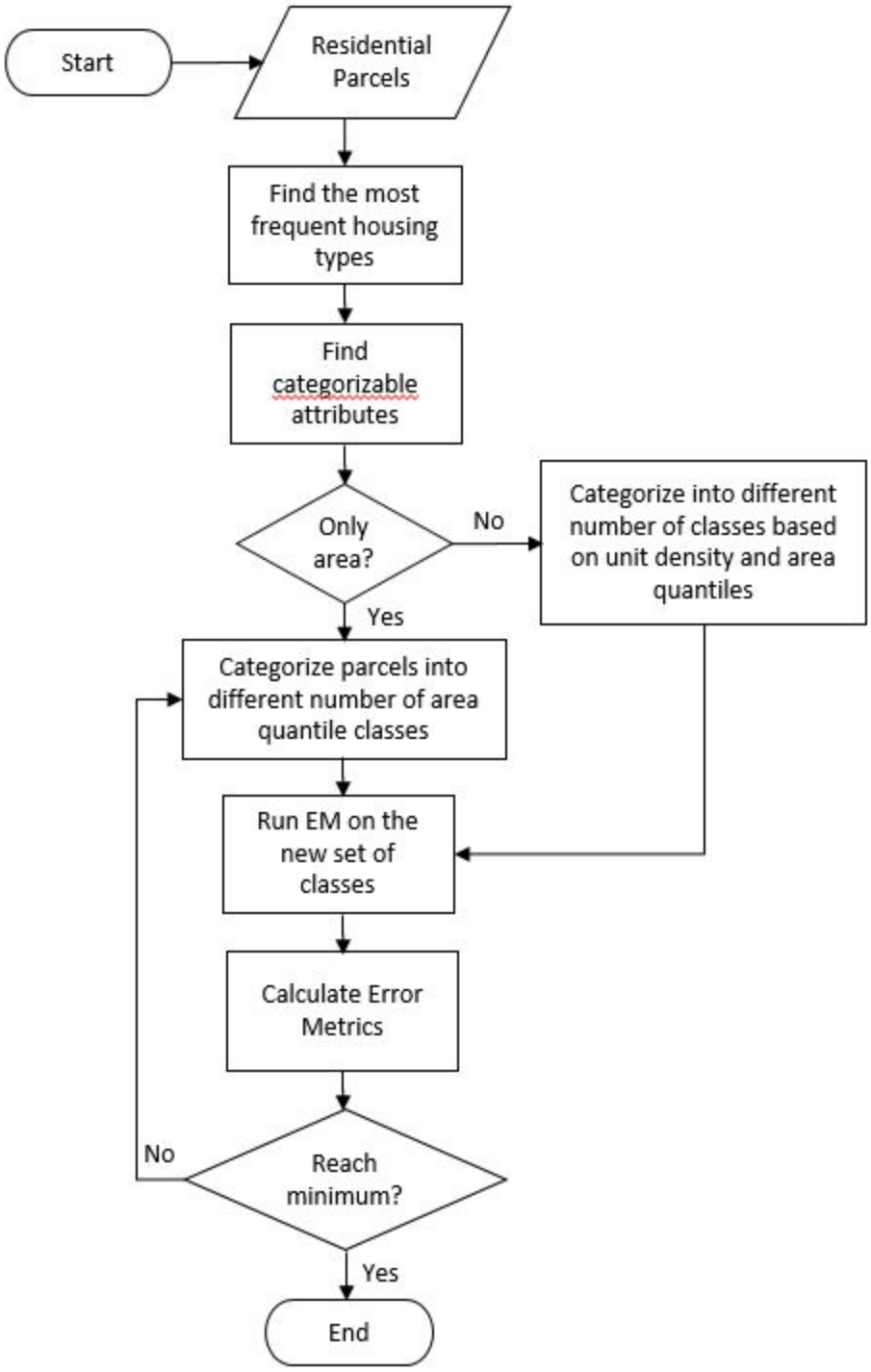

Figure 4.2. Workflow of EEM.

4.3.3. Third spatial refinement using complementary ancillary variables

The third refinement strategy is not confined to only residential parcels and specifically targets rural settings, where large residential parcels are known to overestimate developed land

area while developed land cover classes commonly underestimate development (e.g., Leyk *et al.* 2014). To mitigate these effects, this approach leverages additional complementary ancillary variables such as NLCD-derived developed classes and road buffer zones, assuming that population most likely resides where developed land can be found, or if developed land cannot be found, population would be expected close to roads. NLCD databases published in 1992, 2001 and 2011, which approximately match the three census years, and the available TIGER road networks in 2000 and 2010 are employed to derive additional ancillary variables. The largest residential rural parcels, i.e., the largest 10%, are selected and refined as follows: if a residential parcel contains developed land as classified by the NLCD, only those instances are used for spatial refinement thereby geometrically adjusting the residential parcel. If no developed land exists, the intersection between the parcel and road buffers (using 50m buffer distance) is used to spatially constrain the area of the parcel. Figure 4.3 demonstrates the process of the third spatial refinement. Residential developed land is defined by classes 21 and 22 in NLCD 1992 and classes 21, 22 and 23 in NLCD 2001 and 2011, following recommendations in other studies (e.g., Ruther *et al.* 2015).

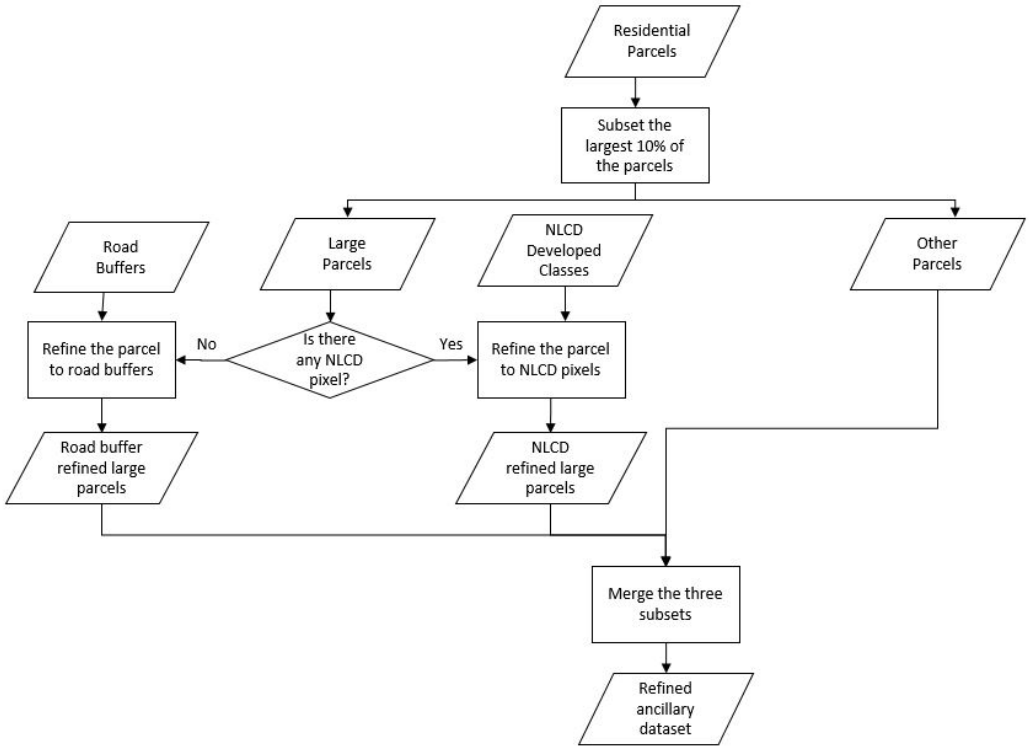

Figure 4.3. Workflow of the third spatial refinement.

Once residential parcels are spatially refined using developed land and road buffers as limiting ancillary variables, the resulting dataset is input to AW, TDW and EM to form the third refinement method. It is expected that the accuracy of the methods to estimate population will improve because they are adjusted to more precise locations of likely human settlement in rural settings where errors are commonly high due to the large extent of residential parcels. This third spatial refinement cannot be implemented with EEM in its current form since area and density attributes are used for population weighting (i.e., as related variable) in this method and must not be modified a posteriori. Integrating such a refinement in EEM requires a priori implementation so that the changed area and density measures can be used for simulation and optimization. The results in this study will provide some indication of the potential benefit of this adjustment.

4.3.4. Validation

The validation of the estimated tract-level results for each census year is done using census block statistics. After transferring population estimates from source zones (tracts in 1990 and 2000, respectively) to target zones (2010 tract boundaries), each 2010 census tract can be linked with its estimated population counts in 1990 and 2000. These estimates for target zones in 1990 and 2000 are compared to population counts of census blocks in 1990 and 2000 aggregated to the target zone boundaries. Blocks are the finest resolution enumeration units used by the Census Bureau and are based on a full population count and thus very useful for validation efforts. Different error measures are calculated such as the Mean Absolute Error (MAE), median absolute error, Root Mean Square Error (RMSE) and 90% percentile of absolute error. These error measures and error distributions can be compared across methods to characterize and evaluate the performance of the described methods. For example, MAE and RMSE measures illustrate the overall behavior of the estimation error and are sensitive to outliers, while the median absolute error and 90% percentile of absolute error can be used to describe the upper end of the error distribution and placement of extreme absolute error values.

4.4. Results

Tables 4.1 to 4.5 show the absolute errors for each of the methods described for the two time periods (1990-2010 and 2000-2010) and each of the five counties. The last column in these tables called “Refinement Level” shows if and what type of spatial refinement is applied for each method (grey-tone encoded).

Table 4.1. Absolute error measures of unrefined and refined methods in Hennepin.

Method	MAE	Median Absolute Error	RMSE	90th Percentile Error	Refinement Level
1990-2010					
AW	219	57	487	646	Unrefined
Refined AW	158	47	342	429	First
Modified Refined AW	148	50	310	464	Third
TDW	201	56	420	650	Unrefined
Refined TDW	132	47	294	368	First
Modified Refined TDW	117	48	234	306	Third
EM	203	60	402	616	Second
Modified EM	139	58	262	391	Third
EEM	110	44	222	270	Second
2000-2010					
AW	58	0	214	97	Unrefined
Refined AW	53	0	248	16	First
Modified Refined AW	44	0	201	17	Third
TDW	36	0	127	88	Unrefined
Refined TDW	24	0	95	19	First
Modified Refined TDW	21	0	86	17	Third
EM	49	0	217	24	Second
Modified EM	31	0	159	14	Third
EEM	18	0	89	23	Second

Table 4.2. Absolute error measures of unrefined and refined methods in Mecklenburg.

Method	MAE	Median Absolute Error	RMSE	90th Percentile Error	Refinement Level
1990-2010					
AW	546	346	832	1477	Unrefined
Refined AW	384	212	624	974	First
Modified Refined AW	261	130	427	684	Third
TDW	387	255	575	829	Unrefined
Refined TDW	263	168	407	588	First
Modified Refined TDW	220	129	360	493	Third
EM	443	240	743	1094	Second
Modified EM	251	125	412	628	Third
EEM	178	91	297	455	Second
2000-2010					
AW	613	213	1012	1728	Unrefined
Refined AW	465	210	793	1294	First
Modified Refined AW	290	135	502	796	Third
TDW	330	138	531	931	Unrefined
Refined TDW	309	116	720	808	First
Modified Refined TDW	228	115	490	548	Third
EM	464	240	741	1294	Second
Modified EM	246	141	394	630	Third
EEM	183	79	309	478	Second

Table 4.3. Absolute error measures of unrefined and refined methods in Broward.

Method	MAE	Median Absolute Error	RMSE	90th Percentile Error	Refinement Level
1990-2010					
AW	1016	584	1699	2460	Unrefined
Refined AW	610	310	1012	1499	First
Modified Refined AW	623	322	978	1667	Third
TDW	630	309	1075	1566	Unrefined
Refined TDW	312	197	489	738	First
Modified Refined TDW	365	194	609	907	Third
EM	499	256	854	1235	Second
Modified EM	481	284	764	1140	Third
EEM	374	207	576	914	Second
2000-2010					
AW	560	47	1654	1619	Unrefined
Refined AW	282	29	538	845	First
Modified Refined AW	290	17	555	863	Third
TDW	151	14	375	439	Unrefined
Refined TDW	100	13	204	324	First
Modified Refined TDW	101	13	197	333	Third
EM	227	33	457	709	Second
Modified EM	232	28	468	700	Third
EEM	151	14	297	449	Second

Table 4.4. Absolute error measures of unrefined and refined methods in Hillsborough.

Method	MAE	Median Absolute Error	RMSE	90th Percentile Error	Refinement Level
1990-2010					
AW	633	300	1000	1807	Unrefined
Refined AW	357	178	615	901	First
Modified Refined AW	321	168	569	765	Third
TDW	431	234	667	1230	Unrefined
Refined TDW	293	148	560	726	First
Modified Refined TDW	327	143	909	688	Third
EM	425	220	674	1118	Second
Modified EM	334	157	550	894	Third
EEM	276	114	500	697	Second
2000-2010					
AW	574	5	2513	1741	Unrefined
Refined AW	231	0	501	862	First
Modified Refined AW	228	0	491	776	Third
TDW	178	3	363	689	Unrefined
Refined TDW	130	0	298	426	First
Modified Refined TDW	121	0	287	401	Third
EM	213	0	493	679	Second
Modified EM	162	0	380	606	Third
EEM	133	0	313	477	Second

Table 4.5. Absolute error measures of unrefined and refined methods in Worcester.

Method	MAE	Median Absolute Error	RMSE	90th Percentile Error	Refinement Level
1990-2010					
AW	275	52	598	892	Unrefined
Refined AW	221	51	446	803	First
Modified Refined AW	146	46	298	379	Third
TDW	147	77	287	299	Unrefined
Refined TDW	138	56	272	290	First
Modified Refined TDW	129	45	266	326	Third
EM	290	72	596	845	Second
Modified EM	155	46	382	319	Third
EEM	129	46	288	269	Second
2000-2010					
AW	184	9	530	568	Unrefined
Refined AW	150	8	402	748	First
Modified Refined AW	101	7	301	240	Third
TDW	37	13	74	104	Unrefined
Refined TDW	37	6	100	103	First
Modified Refined TDW	36	7	110	89	Third
EM	184	8	473	745	Second
Modified EM	75	6	196	297	Third
EEM	40	5	112	87	Second

Figures 4.4 and 4.5 show maps of the absolute errors of the two best-performing methods found for each region and time period, but focusing on the first and second spatial refinement scenarios. In almost all cases except in Worcester for the time period 2000-2010, refined TDW and EEM are the two best performing methods. Because third refinement methods require more specific analysis and interpretation with a focus on rural settings, they will be evaluated, separately.

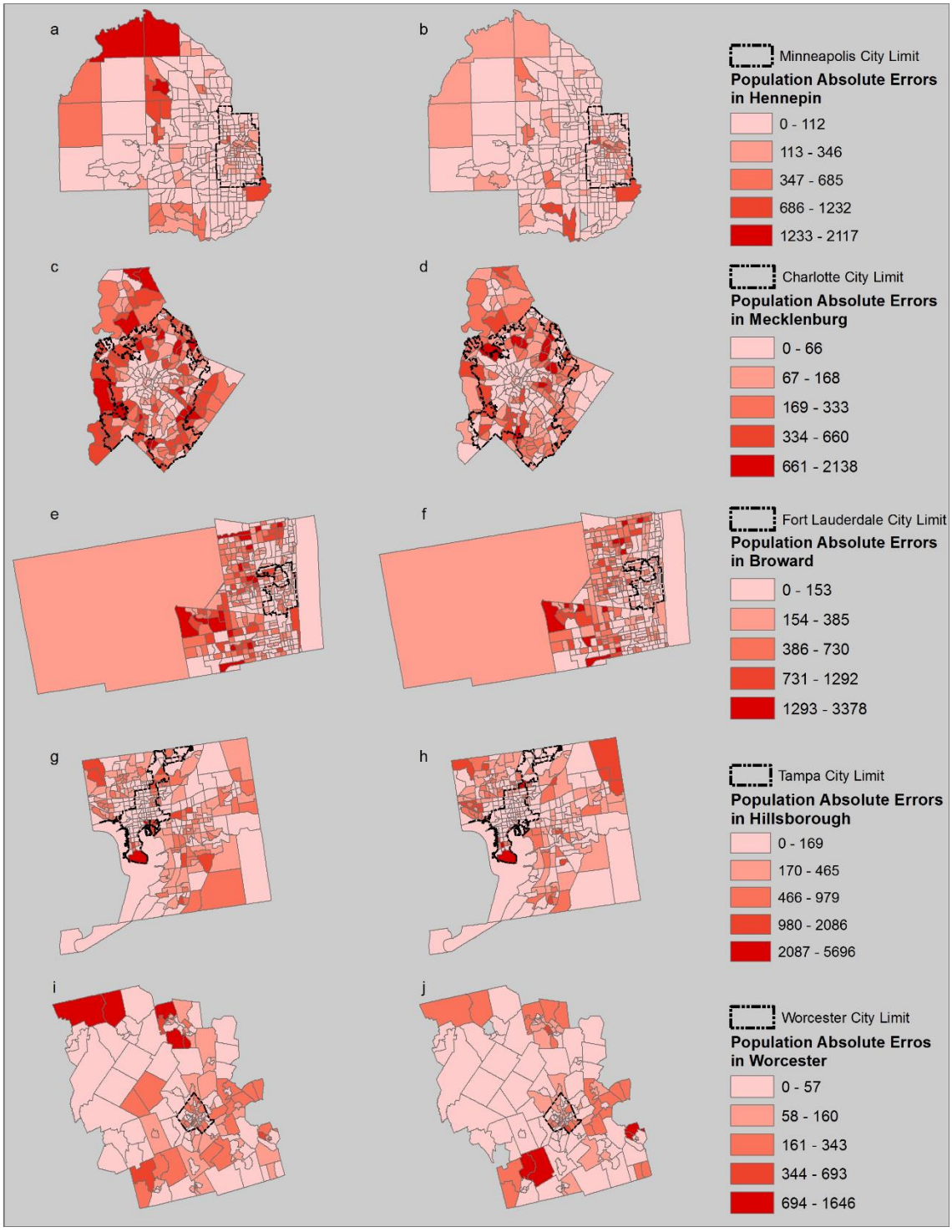

Figure 4.4. Error maps of the five counties (1990-2010), (a) Hennepin: Refined TDW, (b) EEM, (c) Mecklenburg: Refined TDW, (d) EEM, (e) Broward: EEM, (f) Refined TDW, (g) Hillsborough: Refined TDW, (h) EEM, (i) Worcester: Refined TDW, (j) EEM.

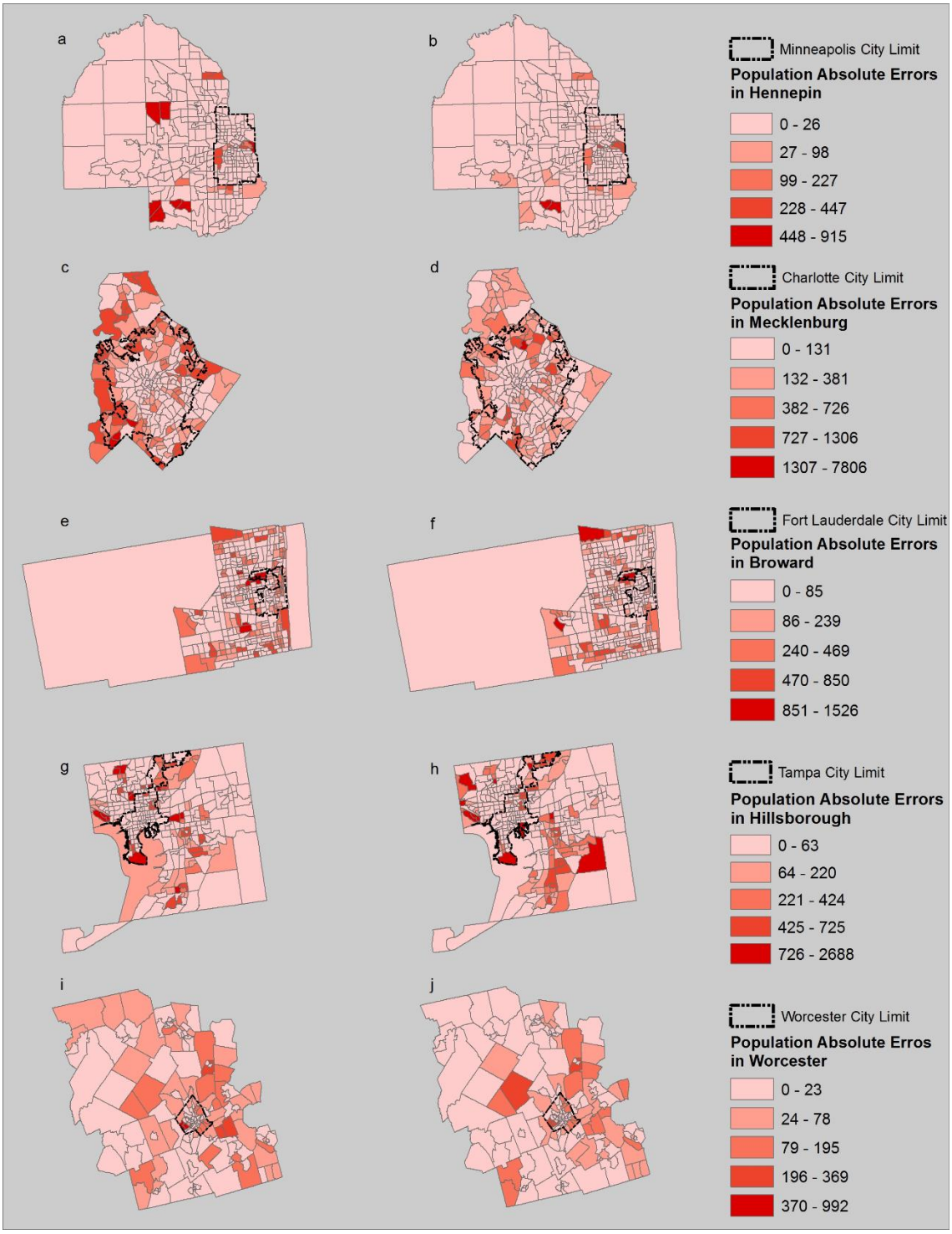

Figure 4.5. Error maps of the five counties (2000-2010), (a) Hennepin: Refined TDW, (b) EEM, (c) Mecklenburg: Refined TDW, (d) EEM, (e) Broward: EEM, (f) Refined TDW, (g) Hillsborough: EEM, (h) Refined TDW, (i) Worcester: TDW, (j) Refined TDW.

Figures 4.6 and 4.7 depict the derived population maps in 1990 and 2000 from the two best

performing methods of the first and second spatial refinement scenarios compared to the population maps resulting from aggregating block population counts to target tract boundaries in Hennepin County and Mecklenburg County. These maps visualize the agreement between the spatially refined population estimates in 1990 and 2000, respectively, and the corresponding ground-truth choropleth population maps based on block level aggregates.

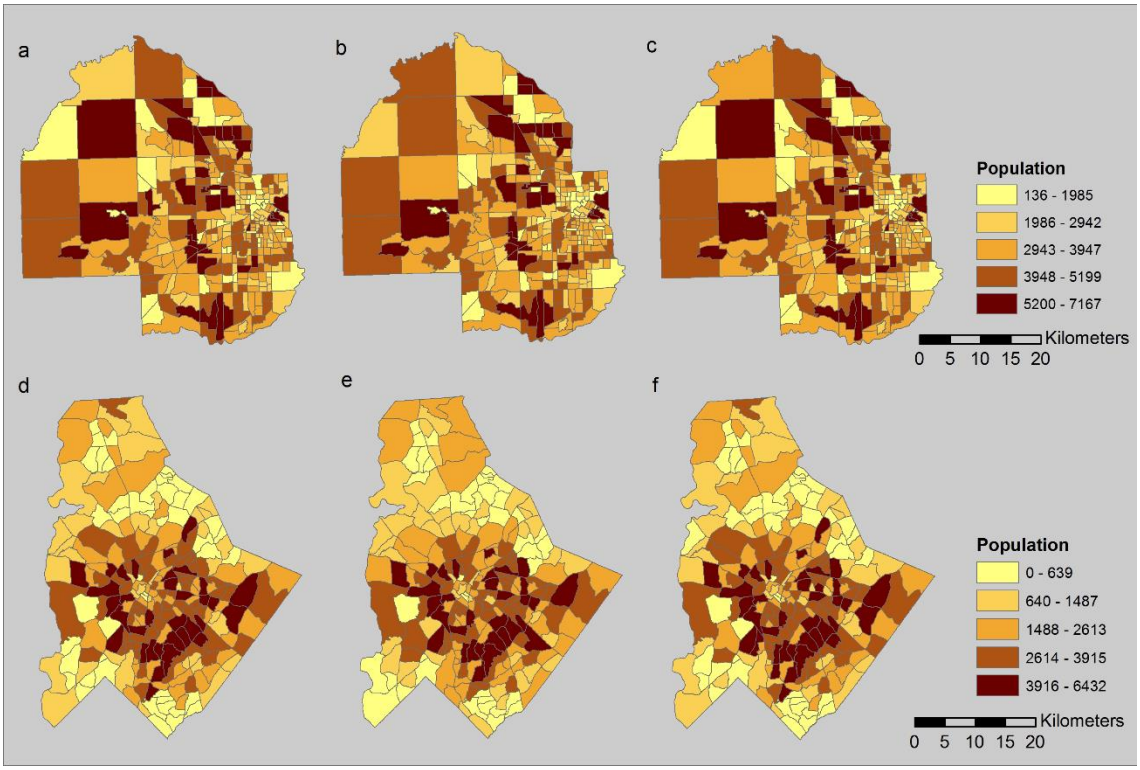

Figure 4.6. Population maps in 1990 at the target tract level, (a) Hennepin: block-aggregated, (b) Refined TDW, (c) EEM, (d) Mecklenburg: block-aggregated, (e) Refined TDW, (f) EEM.

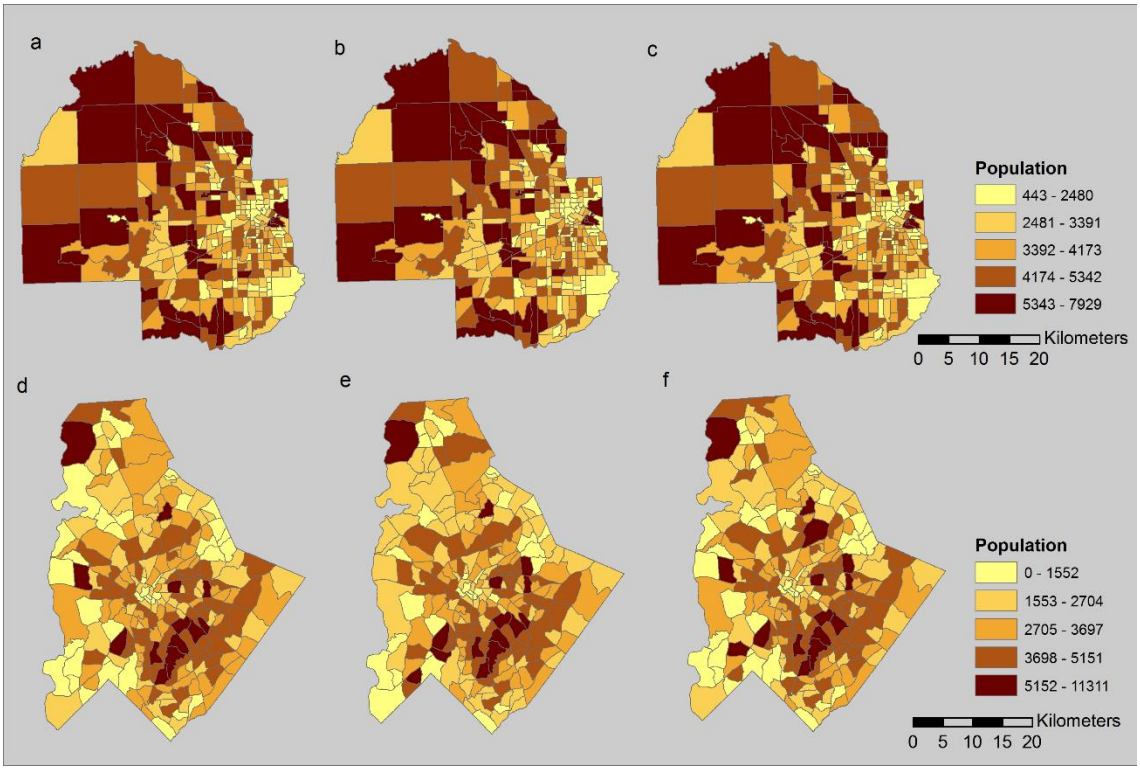

Figure 4.7. Population maps in 2000 at the target tract level, (a) Hennepin: block-aggregated, (b) Refined TDW, (c) EEM, (d) Mecklenburg: block-aggregated, (e) Refined TDW, (f) EEM.

To evaluate how third spatial refinement methods perform in comparison to their first or second refinement equivalents in rural areas, Table 4.6 presents the overall absolute errors (i.e., MAE and RMSE) produced by all refined methods when applied to only rural target tracts in each study area. It is expected that the third spatial refinement methods are effective in rural areas because they further refine large parcels more frequently located in rural parts. To identify rural tracts, the number of rural households indicated by Census is divided by the total number of households, and all tracts with a proportion of rural population greater than 10% are treated as rural. Broward is excluded because this highly urbanized county includes only 1 rural target tract. Other thresholds for identifying rural tracts (i.e., 5% and 15%) were also tested and yielded similar orders of results. Based on the utilized threshold, there exist 9, 10, 38 and 58 rural target tracts in Hennepin, Mecklenburg, Hillsborough and Worcester, respectively.

Table 4.6. Absolute errors for all refined methods applied to rural tracts in four study areas.

Method	Refined AW	Modified Refined AW	Refined TDW	Modified Refined TDW	EM	Modified EM
Hennepin: 1990-2010						
MAE	282	153	575	360	209	137
RMSE	405	243	966	613	331	198
Hennepin: 2000-2010						
MAE	3	0	1	1	1	1
RMSE	6	1	2	2	2	1
Mecklenburg: 1990-2010						
MAE	841	355	573	443	616	264
RMSE	1106	447	694	510	857	346
Mecklenburg: 2000-2010						
MAE	1003	246	577	284	819	208
RMSE	1276	328	678	340	1003	349
Hillsborough: 1990-2010						
MAE	494	353	312	273	507	398
RMSE	754	512	470	437	733	561
Hillsborough: 2000-2010						
MAE	303	237	137	111	186	154
RMSE	730	620	342	286	425	334
Worcester: 1990-2010						
MAE	238	102	147	122	268	178
RMSE	495	195	300	284	579	520
Worcester: 2000-2010						
MAE	173	54	36	35	137	49
RMSE	442	131	73	105	316	125

Figures 4.8 and 4.9 demonstrate the effectiveness of third refinement methods in comparison to their first or second refinement counterparts, particularly in rural areas, for both time periods. This is visualized at the target tract level, and rural tracts are emphasized in the maps. Although the emphasis of the third refinement methods is on rural tracts, the absolute error values of other tracts might change as well due to the existence of some large parcels in urban areas.

If a third refinement method results in a lower error for a target tract compared to the first or second refinement, that tract is shown in green. For example, in Figure 4.8(a), all green tracts represent those for which modified refined AW leads to lower absolute errors than refined AW in

estimating the population in 1990 within target tract boundaries from 2010 in Hennepin County. Tracts shown in orange and grey indicate those where refined AW outperforms modified refined AW and results in equal error measures, respectively.

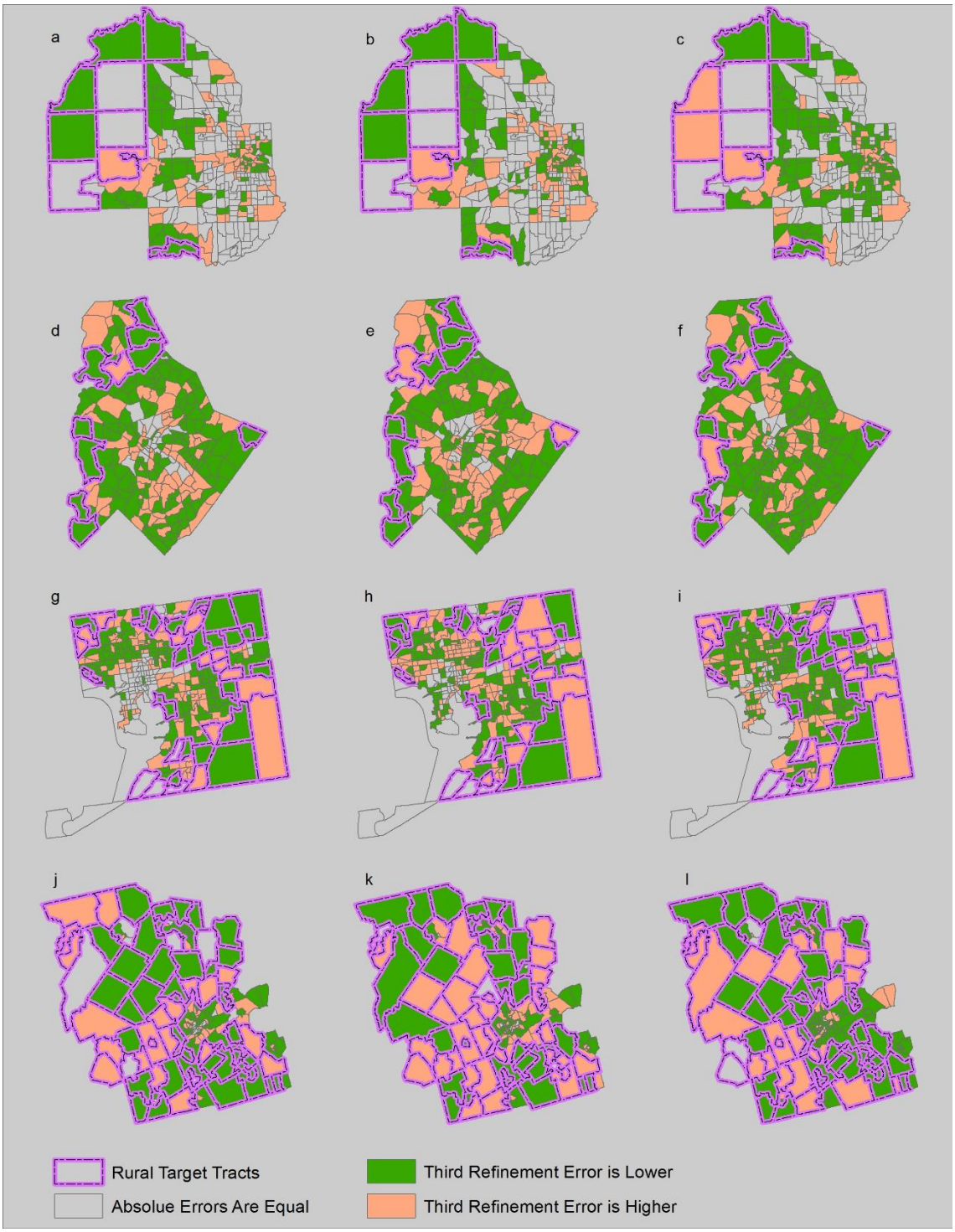

Figure 4.8. Third spatial refinement methods in comparison to their first or second refinement equivalents in 1990-2010, (a) Hennepin: AW, (b) TDW, (c) EM, (d) Mecklenburg: AW, (e) TDW, (f) EM, (g) Hillsborough: AW, (h) TDW, (i) EM, (j) Worcester: AW, (k) TDW, (l) EM.

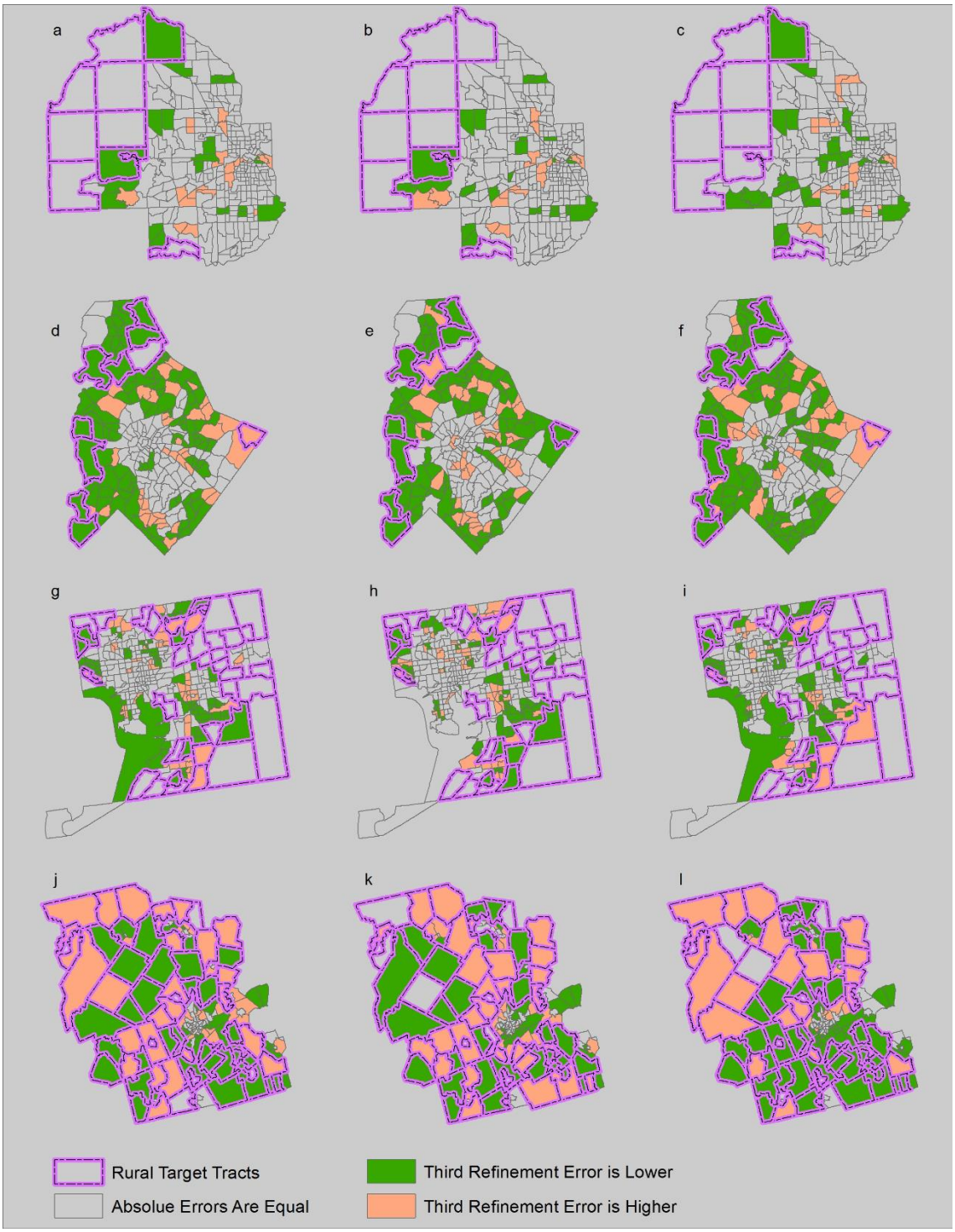

Figure 4.9. Third spatial refinement methods in comparison to their first or second refinement equivalents in 2000-2010, (a) Hennepin: AW, (b) TDW, (c) EM, (d) Mecklenburg: AW, (e) TDW, (f) EM, (g) Hillsborough: AW, (h) TDW, (i) EM, (j) Worcester: AW, (k) TDW, (l) EM.

4.5. Discussion and conclusions

As Tables 4.1 to 4.5 imply, the first spatial refinement often reduces the absolute error measures in comparison to regular implementations of AW and TDW. This re-confirms the great potential of using residential parcels for the spatial refinement in temporal interpolation of population as described in earlier research (Zoraghein *et al.* 2016). Notably, the quality of the ancillary variable and the degree to which underlying assumptions hold greatly influence the effectiveness of spatial refinement. For instance, the overestimation of developed land through residential parcel units can be one major reason why population can be allocated disproportionally in refined AW. In refined TDW, that overestimation may invalidate the underlying assumption of proportionally equal population densities within source unit boundaries at different points in time, leading to biased estimates and increased errors. The very few instances in which errors are higher in the first spatial refinement methods than their unrefined counterparts can be attributed to those issues. For example, the reason why the RMSE of refined TDW is higher than TDW in Mecklenburg County for 2000-2010 relates to the existence of very few target tracts with high absolute errors in the last 10% of the error distribution, possibly pointing to the above-mentioned problem.

Although EM uses the housing type attribute of residential parcels in addition to their geometric footprints for refinement, the computed error measures are rather high. A possible reason could be that there is considerable variation within the individual residential control zones across a county, which cannot be reflected by this method. However, the accuracy level of EEM is much higher than EM as verified in all 10 cases. Consequently, the refinement of residential control zones based on area and density measures appears to be a useful approach to reflect and approximate this within-class variation, ultimately resulting in further error reduction.

This demonstrates the efficacy of EEM as an accurate method for temporal interpolation of population especially over longer time periods where the assumptions of refined TDW begin to fall apart. For 1990-2010, EEM is the best performing method in all study areas except Broward. For 2000-2010, however, neither EEM nor refined TDW dominates as the most accurate method over the five study areas. For shorter time periods, the assumptions of refined TDW appear to be realistic enough, explaining its robust performance in Hillsborough, Broward and Worcester Counties. However, in regions where the assumptions of refined TDW are not reliable even over short time periods, EEM can be more accurate than refined TDW such as in Mecklenburg and Hennepin Counties. EEM is computationally expensive, and running the simulation to find the best combination of eligible control zones and optimum number of categories per control zone is time-demanding. Nevertheless, the current results suggest that the methodology carries great potential to be implemented in cases where the accuracy levels of conventional refined methods are not satisfactory, which can especially be true over long time periods. This is an extremely important point as EEM is expected to be more efficient once parcel-derived demographic or property databases become available for larger regions that contain more consistent, broadly defined residential classes, thus eliminating the need for expensive simulation and exploratory analysis. If these large-scale parcel and property databases carry consistent temporal information for residential development, this will open new possibilities to model small area population estimates over very long time periods, nationally, once historical census data become also available. Another advantage of EEM over refined TDW is that EEM is pycnophylactic; that is, it preserves the population value of source zones as opposed to refined TDW.

According to Figures 4.4 and 4.5, target tracts with higher absolute errors are more frequent and pervasive outside highly urbanized areas and city boundaries. This effect is consistent with

the explanations provided before, suggesting that the overestimation of developed land through large residential parcels in less urbanized areas contributes to decreased accuracy of spatial refinement in those areas.

Figures 4.6 and 4.7 demonstrate that the population maps in 1990 and 2000 resulting from the two best performing methods closely match the ground-truth maps. The derived population estimates of almost all target tracts from the two methods, particularly EEM, and their corresponding ground-truth values belong to the same population categories in both 1990 and 2000.

Third spatial refinement methods increase the accuracy of first and second refinement methods in most cases as can be observed in Tables 4.1 to 4.5. The main focus of this additional refinement step, however, is on rural settings where land cover databases typically underestimate developed land, and parcel units overestimate residential area. Table 4.6 demonstrates that the improvement effect of the third spatial refinement is more consistent in rural target tracts. In almost all cases, the overall error measures of that set of methods are lower, and in some cases, the superiority is substantial. The reason why the error metrics of Hennepin for 2000-2010 are very low is because the boundaries of rural tracts are almost unchanged during the period. Only in Worcester for time period 2000-2010, the overall absolute error metrics of a third spatial refinement method is lower than its first refinement counterpart (i.e., modified refined TDW vs. refined TDW). Figures 4.8 and 4.9 display that the absolute error values of the majority of rural target tracts are reduced by the third spatial refinement in many cases. The only cases where such an effect is not observed are TDW for 1990-2010 and 2000-2010 in Worcester and EM for 2000-2010 in Hillsborough. However, according to Table 4.6, even for two of those three cases, the MAE values of rural target tracts are decreased, meaning that error reduction is greater for those

tracts where the third spatial refinement is effective. These slight performance variations can be explained by the quality of ancillary variables used for the third refinement. Both developed land classes in NLCD and road network are imperfect data sources and have locational errors as well as classification issues that can impact the population estimation, resulting in some higher errors in some parts of the study areas.

Future research will assess the combination of different refinement methods. Although, that approach may be computationally expensive and more complex, each of the utilized methods has its own strengths for different circumstances, justifying the initial efforts for the development of a hybrid approach. Moreover, modified EEM can be implemented by running the simulation over different combinations of housing types and geometrically modified developed lands to examine its effectiveness, especially in rural areas and over long time periods. The incorporation of new global data products such as the recently introduced Global Human Settlement Layer (GHSL) that represents built-up land for four points in time (1975-2014) (Pesaresi *et al.* 2016), represents a promising research avenue to apply similar approaches to areas that are less data-rich.

Acknowledgements

This research is funded by the National Science Foundation: “Collaborative Research: Putting People in Their Place: Constructing a Geography for Census Microdata”, Project BCS-0961598 awarded to University of Colorado – Boulder.

Chapter V

Comprehensive Multi-Faceted Multi-Temporal Demographic Estimation: Enhancement of Areal Interpolation Using Spatial Refinement and Diverse Ancillary Variables

Abstract

This research evaluates the performance of areal interpolation coupled with spatial refinement to estimate different demographic attributes, namely total population and population sub-groups based on race, age structure and urban residence, within consistent census tract boundaries from 1990 to 2010 in Massachusetts. This allows the study of the nuanced and micro-scale evolution of different aspects of population, which is very complicated according to the current temporally incompatible small area census geographies. Various ancillary variables, including the Global Human Settlement Layer (GHSL), the National Land-Cover Database (NLCD), parcels, building footprints and the proprietary ZTRAX® dataset are utilized for spatial refinement, and their effectiveness for improving the accuracy of multi-temporal population estimations is evaluated. The outcomes of these analyses are then employed for an environmental injustice application to first estimate the counts of different population sub-groups living within flood zones of the state, and second to assess if certain racial or age-related sub-groups are

disadvantaged. The results of the analyses of this research show the effectiveness of using areal interpolation enriched by spatial refinement for more reliable multi-temporal estimations of population although the level of the effectiveness depends on the utilized ancillary variables, the demographic attribute and the time duration of the application. This research also shows the effectiveness of using advanced areal interpolation methods for risk assessment and environmental injustice applications based on tract-level census demographies, which can be extended to earlier historical census years or other census systems that do not offer units as small as blocks. The findings of this research also point to the potential of using the proposed methods to redefine urban lands and population objectively over time.

Keywords: Areal Interpolation; Demographic Analysis; Spatial Refinement; Urban; Environmental Injustice Assessment

5.1. Introduction

The effectiveness of using enhanced areal interpolation methods, leveraging dasymetric modeling (Wright 1936, Eicher and Brewer 2001, Mennis 2003) in temporal interpolation of demographic variables is reported by different researchers (e.g., Buttenfield *et al.* 2015, Ruther *et al.* 2015, Zoraghein *et al.* 2016, Schroeder 2007, 2017, Schroeder and Van Riper 2013, Logan *et al.* 2016). The ultimate objective of those works is to create temporally consistent total population estimates within non-coterminous census units from different census years with minimum error. This is accomplished by identifying the census boundaries of one census year, i.e., target year, as "target zones". Population counts recorded for census boundaries used in other census years, i.e., "source zones" are then transferred or redistributed to those target zones. All the above-mentioned works report that such refinement techniques result in reductions in absolute population estimation errors.

The common areal interpolation methods that are used for temporal estimation of population are Areal Weighting (AW) (Goodchild and Lam 1980), Target Density Weighting (TDW) (Schroeder 2007), Pycnophylactic Modeling (PM) (Tobler 1979) and Expectation Maximization (Dempster *et al.* 1977, Flowerdew and Green 1994). Enhancement using dasymetric refinement entails that the methods be applied to only inhabitable sub-areas of source and target zones identified by ancillary variables to benefit from a more precise delineation of the underlying statistical surface of population distribution. The theoretical frameworks of refined AW, refined TDW and EM are explained in Ruther *et al.* (2015) and Zoraghein *et al.* (2016), and Kim and Yao (2010) formulate the methodological implementation of refined PM.

All areal interpolation methods show some level of estimation errors, and their performance varies inevitably with specific conditions (Zandbergen and Ignizio 2010). Moreover,

the performance of spatially refined areal interpolation methods also can vary according to the ancillary data employed (Langford 2013). Recent studies argue that exceedingly complex spatially refined areal interpolation methods may not offer much accuracy improvement over existing methods and can even increase estimation errors (Lin *et al.* 2011, 2013, Schroeder and Van Riper 2013, Lin and Cromley 2015b). Therefore, it is important to focus on various ancillary variables – both publically available and commercial – to better understand their effectiveness in enhancing the currently established areal interpolation methods for temporal estimation of population.

Regardless of whether the purpose of areal interpolation in demographic data is to downscale population from a set of coarse zones to a set of finer-resolution target units (e.g., Bhaduri *et al.* 2007, Dmowska and Stepinski 2017) or to create temporally consistent population estimates among incompatible census units (Schroeder 2007, Buttenfield *et al.* 2015, Ruther *et al.* 2015, Zoraghein *et al.* 2016), the most commonly modeled demographic variable is total population. Rarely have other variables been analyzed or evaluated, which represents a persistent shortcoming in demographic applications.

Incorporating other attributes in addition to total population in such modeling efforts sheds light on various new aspects of demographic distributions and their changes over time. For example, to better understand the underlying trends of different subgroups of the population related to race, age, and urban residence, those demographic estimates need to be examined over temporally consistent fine-scale census units such as census tracts that suffer from often reported incompatibilities (Gregory 2002, Martin *et al.* 2002, Schroeder 2007). If such high-resolution depictions of different sub-populations can be provided with confidence, this would result in a more refined and differentiated portrayal of population distribution and reveal important nuances of where and under what circumstances the population lives. One example in that context is the

exposure assessment.

The objectives of this paper are threefold. First, different ancillary variables are used in order to enhance regular areal interpolation methods, including AW, TDW and EM to estimate tract-level demographic variables for the whole state of Massachusetts in 1990 and 2000 (source zones) within and compatible to target tract boundaries from the 2010 Census. These models will create temporally consistent time-series of population estimates across the three census years. The ancillary variables comprise both readily and freely available datasets such as the National Land-Cover Database (NLCD) and Global Human Settlement Layer (GHSL) and those that are either proprietary or not readily available such as tax parcels of the state, its building footprints as well as ZTRAX® records. This part of the analysis evaluates how different ancillary variables of varying spatial resolution influence the accuracy of regular areal interpolation methods in a temporal context. It also examines the limitations of accuracy improvement by using localized often unavailable ancillary variables over nationally or globally publicly available datasets.

Second, the establishment of consistent time series of population estimates is extended to and evaluated for other sub-groups of population related to age, race and urban residence. This analysis can reveal the level of consistency of accuracy improvements due to spatial refinements in temporal analysis across different demographic attributes. The successful construction of consistent micro-scale time series of different aspects of population would be beneficial for applications pertaining to health studies, crime analysis, hazard/risk assessment, land-use planning, or environmental impacts assessment.

Third, in order to place the results of the previous two steps within an application-oriented context, a risk assessment is conducted to evaluate exposure levels of various racial and age-related sub-groups of population to the flood hazard in Massachusetts. Spatially refined areal interpolation

methods by the most reliable ancillary variables selected from the previous steps are employed to estimate counts of people from different demographic sub-groups who reside within areas of elevated flood hazard. Based on such improved sub-population distributions, it is examined through an environmental justice lens, if the minority populations are potentially more exposed than the majority. Previous studies in the field report an increase in risk and exposure to disasters for communities of color in the United States (Fothergill *et al.* 1999). This study examines whether the same patterns can be revealed from the above-described refined methodological frameworks.

5.1.1. Background: Dasymetric refinement for improved exposure assessments

Various ancillary variables have been used in the literature so far for population downscaling in dasymetric modeling. Land-cover/land-use is still the most widely used ancillary variable (Wright 1936, Mennis 2003, 2009, Reibel and Agrawal 2007, Linard *et al.* 2011, Buttenfield *et al.* 2015, Ruther *et al.* 2015, Dmowska and Stepinski 2017). High resolution satellite images constitute another possible ancillary dataset in population disaggregation (Lu *et al.* 2010, Ural *et al.* 2011, Lung *et al.* 2013). Fine-resolution images are input to Object Based Image Analysis (OBIA) (Blaschke 2010) for extracting buildings as the smallest unit of settlement (Wang *et al.* 2016). The range of other employed ancillary variables is extensive and spans datasets such as LiDAR (Dong *et al.* 2010, Qiu *et al.* 2010, Sridharan and Qiu 2013, Xie *et al.* 2015), tax parcel data (Maantay *et al.* 2007, Kar and Hodgson 2012, Mitsova *et al.* 2012, Jia *et al.* 2014, Jia and Gaughan 2016, Zoraghein *et al.* 2016), street networks (Reibel and Bufalino 2005, Su *et al.* 2010), impervious surfaces (Zandbergen and Ignizio 2010, Schroeder 2017), address points (Tapp 2010, Zandbergen 2011), buildings (Wu *et al.* 2008, Calka *et al.* 2016) and Volunteered Geographic Information (VGI) (Bakillah *et al.* 2014, Geiß *et al.* 2016).

Vulnerability is generally defined as the potential for loss of life or property due to hazards

(Hazards and Vulnerability Research Institute 2014) and is typically assessed through estimating exposure, impact or damage (Cutter *et al.* 2009, Burton 2010). Exposure is considered a highly tangible component of risk. It comprises the assets in terms of people, properties, infrastructure, or economic activities potentially affected by a hazardous event (Schneiderbauer and Ehrlich 2004, Geiß and Taubenböck 2013).

In order to carry out an unbiased analysis of exposure to natural hazards in an application, the analyst requires highly detailed population distributions within the area under study, both in terms of resolution and demographic characteristics. Several examples of combining dasymetric modeling with risk analysis that can be found in the literature typically focus on issues such as environmental injustice, exposure assessment and evacuation preparedness in the event of a natural disaster, to name a few. Maantay and Maroko (2009) use their Cadastral-based Expert Dasymetric System (CEDS) (Maantay *et al.* 2007) to estimate the level of impact on racial sub-groups by 100-year flooding in New York City. Wu *et al.* (2017) leverage ancillary datasets including night-time lights, road networks and land-cover/land-use for the disaggregation of asset values in China to improve current disaster risk assessment frameworks. Geiß *et al.* (2016) employ remote sensing data and VGI (Goodchild, 2007) to improve the estimation of crucial exposure components, including the number of buildings and population counts at fine resolution. Bian and Wilmot (2017) use dasymetric modeling to investigate if risk-prone and disadvantaged people are sufficiently served by the current evacuation facilities in New Orleans.

5.2. Study area and data

5.2.1. Study area

The below-described analyses are conducted for the whole state of Massachusetts for several reasons. First, different state-wide datasets that can be used as ancillary variables for

dasymetric modeling are publicly available. This includes parcels as well as building footprints. Second, the proprietary ZTRAX® database is available for this research, and unlike in some other states, the completeness of this database is very high. Third, although Massachusetts is a small state, its population size is relatively high, with highly variable population densities ranging from the densely populated Boston metropolitan region in the east to sparsely populated western parts of the state. For these reasons, Massachusetts represents an excellent benchmark study area for future national-scale extensions of the study.

However, some shortcomings related to the study area selection are noteworthy. First, there is relatively a low level of racial diversity across the state, possibly limiting the validity of the environment injustice analysis. While according to the U.S. Census Bureau (2010), white and black populations constitute 80.4% and 6.6% of the total population of the state of Massachusetts, respectively, those percentages are 66.2% and 27.9% in South Carolina and 59.7% and 30.5% in Georgia. Nevertheless, the aforementioned reasons justify the use of Massachusetts as a convenient study area at this stage.

5.2.2. Data

5.2.2.1. Census data

The focus of this study are census tracts in 1990, 2000 and 2010, along with their demographic attributes including total population and population subgroups based on race (white/black), age (under18, under 65, above 65) and urban residence. The tabular summary files and geometric boundaries in 1990 were extracted from The National Historical Geographic Information System (NHGIS) data portal (Minnesota Population Center 2016). Census tract boundaries in 2000 and 2010 were accessed as TIGER/Line® (U.S. Census Bureau 2016) and their demographic attributes were retrieved from the American FactFinder download center (U.S.

Census Bureau 2010b). Census block boundaries – as the smallest census aggregation unit – in 1990 and 2000 as well as their demographic attributes were extracted from the same sources as their coincident tracts and used to validate interpolated tract-level demographic estimates for 1990 and 2000, respectively.

5.2.2.2. Publically available national or global ancillary variables

The NLCD land-cover product, which is derived from Landsat imagery and published at a resolution of 30m, provides nationally complete, current, consistent, and publicly available information on the nation's land-cover. In this study, the NLCD layers of the three temporally closest vintages to the census years of interest, 1990, 2000 and 2010, are employed for spatial refinement. They are NLCD 1992 (Vogelmann *et al.* 2001), NLCD 2001 (Homer *et al.* 2007) and NLCD 2011 (Homer *et al.* 2015) and commonly refer to a range of years prior to and after the year indicated by the product release. NLCD includes multiple classes, representing different land-cover and land-use types. The classes 21, 22 and 23 in NLCD 1992 and 21, 22, 23 and 24 in 2001 and 2011 label developed land with varying degrees of development intensity. In this study, development masks are created based on different combinations of the developed classes in different years and then used in spatial refinement at those points in time in order to determine the optimum combination for each period.

The GHSL dataset represents global spatial information about the human presence on the planet over time. The data layers are generated using evidence-based analytics and knowledge based on new spatial data mining technologies. In this study, the Landsat-based fine resolution (38m) version of GHSL is used. It classifies built-up land from before 1975 to 2014 (Pesaresi *et al.* 2016). GHSL built-up layers that are approximately coincident with the three census years were used (GHSL epochs of 1990, 2000 and 2014). Some drawbacks in using GHSL are the temporal

mismatch between the latest GHSL epoch and the latest census year, the low levels of classification accuracy in rural settings (Leyk *et al.* n.d.), as well as the assumption that development cannot return to non-development, which might not necessarily be the case in some areas. Regardless of these caveats, the dataset provides a unique and recently available global depiction of human settlement, which is very convenient for spatial refinement and could potentially be applied in data-poor regions.

5.2.2.3. Local or commercial ancillary variables

Tax parcels of Massachusetts except those in the City of Boston are available per township (MassGIS 2016). Therefore, parcels of all of the townships were downloaded and merged together, then combined with the City of Boston's parcel data created by its Assessing department (BostonGIS 2016) to form a complete state-wide dataset. Land-use classes in parcel records typically indicate the existence of developed areas at the lot level, and can thus be considered a promising ancillary variable (Zoraghein *et al.* 2016). However, their size presents a high level of diversity, stretching from very small lots in highly urbanized areas to extremely large ones in rural locations (Leyk *et al.* 2014).

Building footprints represent the smallest achievable unit of residence for spatial refinement. Fortunately, the state-wide building footprints of Massachusetts are publicly available and were utilized in this study for dasymetric refinement (MassGIS 2017). The dataset has been created using the conjunction of fine resolution imagery from DigitalGlobe®, LiDAR data and parcel boundaries.

This research also employs the proprietary ZTRAX® dataset for the first time as another ancillary variable for spatial refinement, courtesy of the Zillow Company (Zillow 2017). It contains a multitude of housing and property-related information as well as longitude and latitude

of the geometric centroid of encompassing parcels. The dataset is nation-wide with varying levels of completeness and attribution.

To have a consistent set of attributes for parcels, buildings and ZTRAX® housing records, the standardized set of attributes of ZTRAX® records are assigned to encompassing parcels and buildings located inside those parcel boundaries using the spatial join operation. This approach has some advantages. First, it eliminates the inconsistency between attributes of parcels in Boston and those in the remaining areas of the state, which is a result of being compiled by two different sources. Second, it creates a high level of consistency in land-use attributes between the three datasets. Third, it increases the flexibility of using parcels and buildings in other areas, where parcel records do not provide land-use information. In those instances, as long as only geometric footprints of parcels and buildings are ready, the spatial refinement using them can proceed because the ZTRAX® data is consistently and nationally available.

The built-year attribute indicating when the main building within a parcel has been built and the land-use class attribute defining the category of the building were extracted from the ZTRAX® data. The first attribute was used to temporally match the three ancillary variables with the census year while the second attribute determined the relevant records to be utilized for creating ancillary masks.

5.3. Methods

AW, TDW, EM – with and without spatial refinement – are implemented to transfer total population and population sub-groups based on age (under 18, under 65 and above 65), race (white and black) and urban residence from source census tracts in 1990 and 2000 to target tract boundaries in 2010 in order to create temporally consistent tracts for the 1990-2010 and 2000-2010 time periods, respectively. The underlying assumptions of the methods as well as their

mathematical foundations are described recently (e.g., Schroeder 2007, Schroeder and Van Riper 2013, Ruther *et al.* 2015, Zoraghein *et al.* 2016).

Enhanced EM (EEM) introduced and explained in Zoraghein and Leyk (n.d.) is also implemented. EEM makes EM more robust by first dividing its initial control zones into more homogeneous sub-control zones according to area and unit-density quantiles, and then performing the EM algorithmic framework on the new set of more homogenized zones. However, this method requires computationally complex simulations using different numbers of classes and subcategories of residential units. Given the size of the study area and the required processing time, only few such simulations can be examined. Thus, based on prior experience in testing this method, the five most frequent classes with the highest area diversity and the class "condominium" are each divided into seven more homogeneous sub-groups according to quantiles of area and unit density, respectively.

5.3.1. Spatial refinement for total population and population sub-groups

Spatial refinement is based on the ancillary variables introduced in Sections 2.2.2 and 2.2.3 and their different combinations. For example, the selection of NLCD developed classes is not confined to those suggested in Ruther *et al.* (2015), and different combinations of developed classes of NLCD at different points in time are explored for creating spatial masks. Moreover, NLCD developed classes are combined with built-up land depictions derived from GHSL to create composite spatial masks.

ZTRAX® records in Massachusetts are already categorized into 245 land-cover/land-use types, of which 41 classes that indicate inhabited or populated lands are extracted. The point locations of all these selected records are spatially joined with parcel boundaries and buildings contained in these parcels. These integrated data layers are used for the spatial refinement of the

demographic enumerated data. Spatial masks for refinement are also created solely based on the selected ZTRAX® point locations. To do so, these point features are rasterized using a target resolution of 30m to make the ancillary dataset comparable to NLCD.

The different spatially refined areal interpolation methods leverage the different ancillary variables as binary masks (limiting ancillary variables) but also as related ancillary variables) to derive temporally consistent estimates of total population and population sub-groups based on race (white/black) and age (under 18/under 65/ above 65) in 1990 and 2000 within 2010 census tract boundaries, respectively.

5.3.2. Spatial refinement for estimating changes in urban land and urban population

The definition of urban lands and urban population represents one of the most persistent challenges in demography and urban geography and typically underlies complex processes. The U.S. Census Bureau uses various criteria based on population, identification of designated places, land-use and road segments, among others, to identify urban lands, which are delineated by layers of Urbanized Areas (UAs) and Urban Clusters (UCs) (Department of Commerce 2011, U.S. Census Bureau 2011). One well-known problem in using these layers is the change in the underlying definition of what is urban. In fact, in addition to the temporal incompatibility in the small-area enumeration units, the concepts of urban lands and consequently urban population change over time. This makes studying the evolution of these two complex phenomena extremely challenging. Nevertheless, given the ubiquitous and growing trends of urbanization, worldwide, and the limited knowledge about such processes, research efforts to model urban lands and population reliably and consistently over time and across regions are essential and have important implications in domains such as planning, policy making and resource allocation.

The census-defined urban areas of Massachusetts in 1990, 2000, and 2010 were available

and accessed from the NHGIS portal (Figure 5.1). The purpose of this part of the analysis is two-fold. First, by using these urban land layers for spatial refinement, it can be assessed how reliably they represent the areas in which urban population resides and thus enclose the underlying statistical surface of urban population. This might also aid in deriving and testing ideas on how the urban definitions can be further refined and improved in a data-driven approach. Second, the performance of other ancillary variables, as possible surrogates for census defined urban areas, can be investigated using insights from this analysis.

To serve the first purpose, the census defined urban areas are initially treated as another ancillary variable for dasymetric refinement (Wei *et al.* 2017) and integrated into the areal interpolation methods to estimate temporally consistent urban population values from 1990 to 2010 and 2000 to 2010 at the census tract level, respectively. Next, the urban area layers in each year are further refined by the aforementioned ancillary variables, and for each composite refinement, the performances of the methods are compared. It is noteworthy to mention that a larger number of land-use classes from the ZTRAX® data should be selected to cover land-use types that are not residential but indicate urban land (e.g., commercial or industrial).

For the second purpose, all the ancillary variables are employed for spatial refinement but they are not limited to be situated within the census defined urban areas. The outcomes of these models are then evaluated, and the level of the reliability of each combination of ancillary variables to mimic the urban areas are presented.

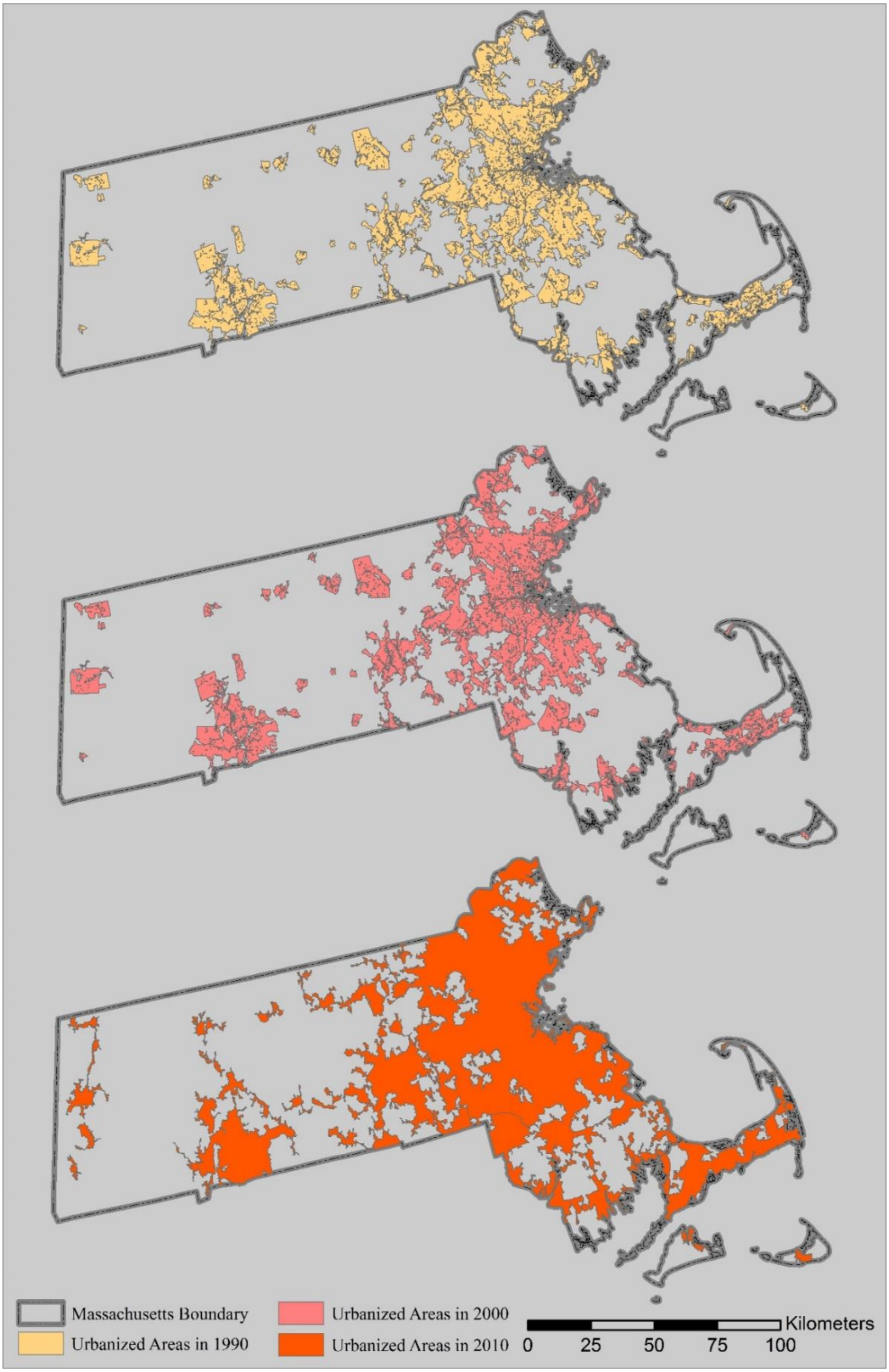

Figure 5.1. Census-defined urban areas in Massachusetts in 1990, 2000 and 2010.

5.3.3. Enhanced areal interpolation methods for improved risk assessment over time

The National Flood Hazard Layer (NFHL) dataset, containing 100-year and 500-year flood zones, which has been updated by the Federal Emergency Management Agency (FEMA) was accessed (MassGIS 2014). 100-year and 500-year flood zones represent areas with a 1% and 0.2% annual chance for flooding. This study examines how risk assessment analysis can leverage refined AW, EM and EEM to improve the estimation of the potentially exposed total population and population sub-groups (age and race) residing within hazardous flood zones in 1990 and 2000. In other words, the above-described areal interpolation methods transfer population values from source tracts in 1990 and 2000 within flood zones (i.e., target zones). Refined TDW cannot be utilized because it also requires population counts in target zones, which are not available.

Examining census tracts as source zones presents three advantages. First, census tracts are more historically and spatially available than blocks, and thus allow for longer historical vintages. Second, existing comparable enumeration units in other countries can also be implemented in similar analyses, indicating the potential applicability of such methods to other census systems. Third, blocks can be used independently for validation to assess which tract-based method produces the most accurate and realistic delineation of the population at risk. Figure 5.2 depicts an overview of flood zones in Massachusetts.

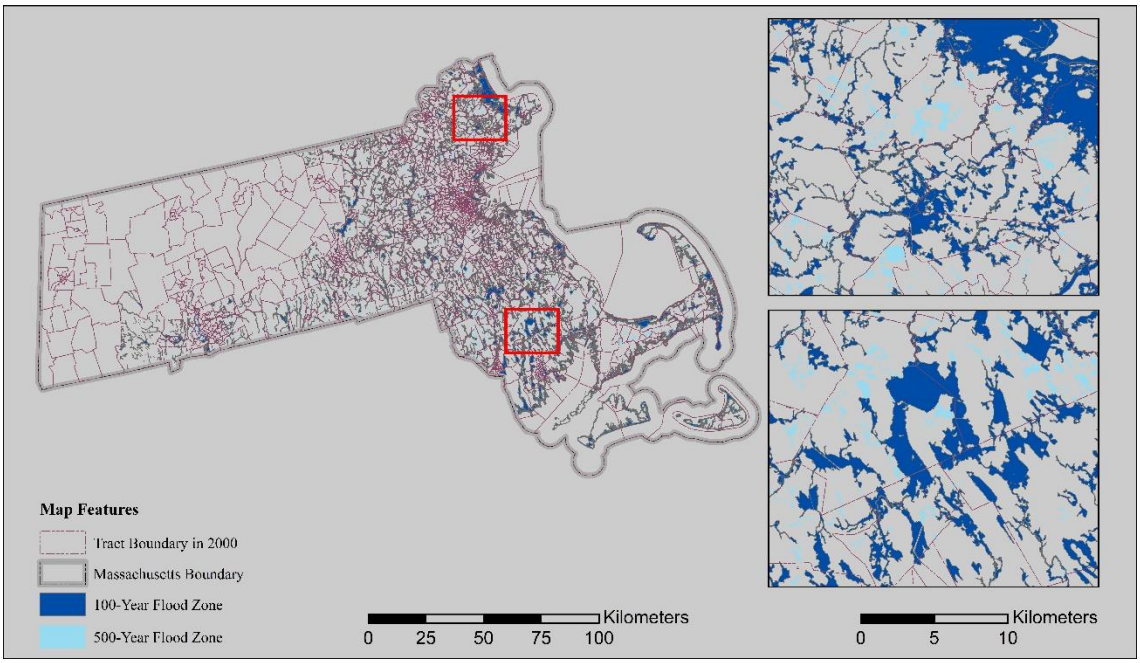

Figure 5.2. Flood zones and tract boundaries from Census 2000 in Massachusetts.

To identify the most accurate method, a benchmark estimation of the potentially exposed population is required. This benchmark is derived by using EEM employing building footprints to transfer population values from blocks to flood zones. EEM is selected because it is generally more accurate than refined AW (Zoraghein and Leyk n.d.), and buildings are chosen for dasymetric refinement because the dataset is the finest resolution delineation of human settlement available.

Population counts of the most accurate methods are also input to assess the risk of exposure for different sub-groups. In this part of the analysis, answering this question is specifically of interest: are certain sub-groups of population more exposed to risk than expected? The expected values for different sub-groups are derived by assuming that their population is distributed within flood zones proportionally; i.e., the ratio of the exposed count of a population sub-group to the total exposed population is the same as the ratio of that population sub-group to the total population of the state.

5.3.4. Validation

The validation of the estimated tract-level results for each census year is done using census block statistics, as is often done in tract-level analysis and dasymetric modeling (e.g., Buttenfield *et al.* 2015, Ruther *et al.* 2015, Zoraghein *et al.* 2016). After transferring population estimates from source zones to target zones, each 2010 census tract can be linked with its estimated population counts in 1990 and 2000. These estimates for target zones in 1990 and 2000 are compared to population counts of census blocks in 1990 and 2000 aggregated to target zone boundaries. In addition to total population, population subgroups in this study are reported over both census tracts and blocks, thereby enabling blocks for validation.

Different error measures are calculated such as the Mean Absolute Error (MAE), median absolute error, Root Mean Square Error (RMSE) and 90% percentile of absolute errors. These error measures and error distributions can be compared across methods to characterize and evaluate the performance of the described methods. For example, MAE and RMSE measures illustrate the overall behavior of estimation errors and are sensitive to outliers whereas the median absolute error and 90% percentile of absolute error can be used to describe the upper end of the error distribution and placement of extreme absolute error values.

5.4. Results

This section describes some selected results to reflect the most relevant outcomes among a large number of model runs. To reduce complexity, different implementations of refined AW are excluded as it is typically found to be the least effective approach whose accuracy often ranks the lowest among refined methods.

5.4.1. Spatial refinement for total population and population sub-groups

In this section, maps of absolute errors for population estimates based on TDW are shown.

TDW is the most accurate unrefined areal interpolation method and the best-performing refined method when employing different types of ancillary variables. These absolute errors result from estimating the demographic attribute in 1990 and 2000 within census tract boundaries in 2010 (target zones) (Figures 5.3, 5.5 and 5.7). Furthermore, population maps are shown, which depict the estimated demographic attribute itself in 1990 and 2000 within target tract boundaries, resulting from the above-mentioned unrefined and refined TDW in comparison to block-aggregated attributes used as references (Figures 5.4, 5.6 and 5.8).

To demonstrate outcomes for population sub-groups based on race and age classes, the respective maps are shown for white population (Figures 5.5 and 5.6) and the number of people aged under 65 (Figures 5.7 and 5.8).

As shown in Figure 5.3, the best-performing refined methods for estimating total population during the 1990-2010 and 2000-2010 time periods are TDW refined by building footprints and TDW refined by rasterized ZTRAX® points, respectively. Figure 5.4 shows how these total population estimates compare to block-aggregated references.

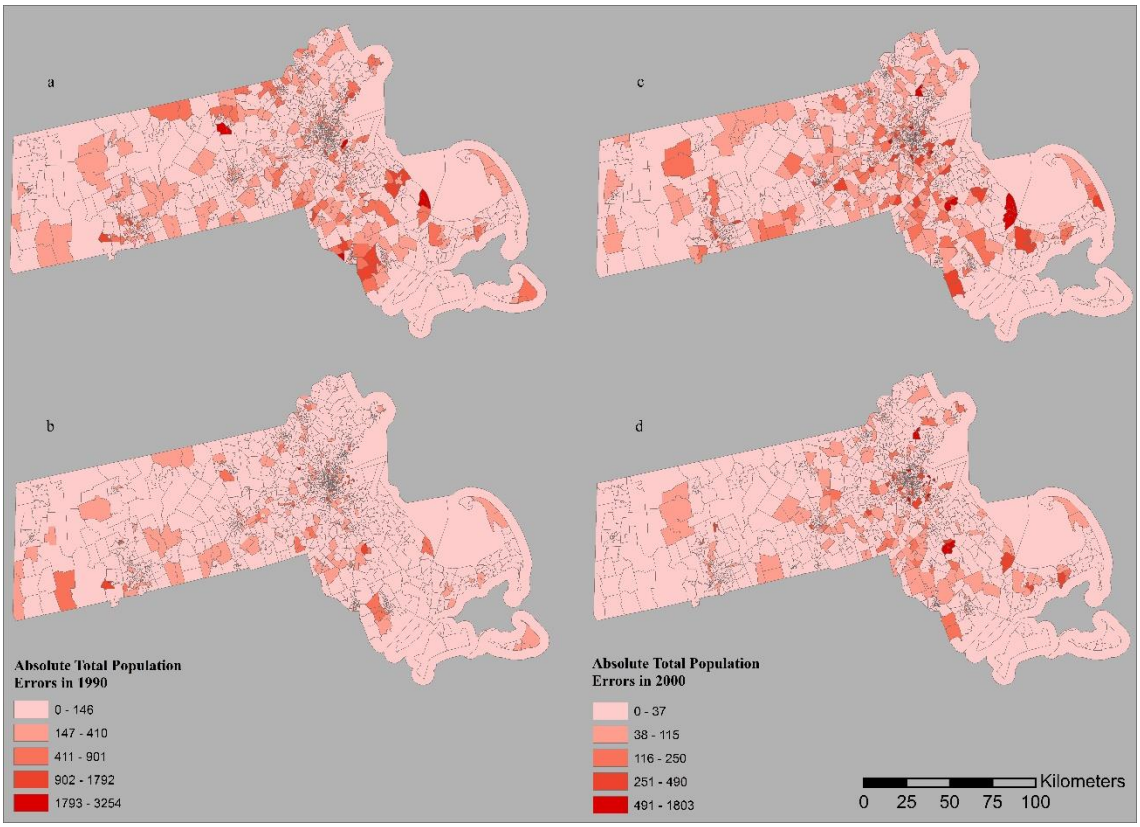

Figure 5.3. Absolute error maps of total population estimates at the target zone level: (a) in 1990 based on TDW, (b) in 1990 based on TDW refined by buildings, (c) in 2000 based on TDW and (d) in 2000 based on TDW refined by ZTRAX®[1].

[1] The dataset was provided by the Zillow Inc. (https://www.zillow.com/research/data/), and the source code for its process is at https://github.com/spatialhistory/ztrax_analysis.

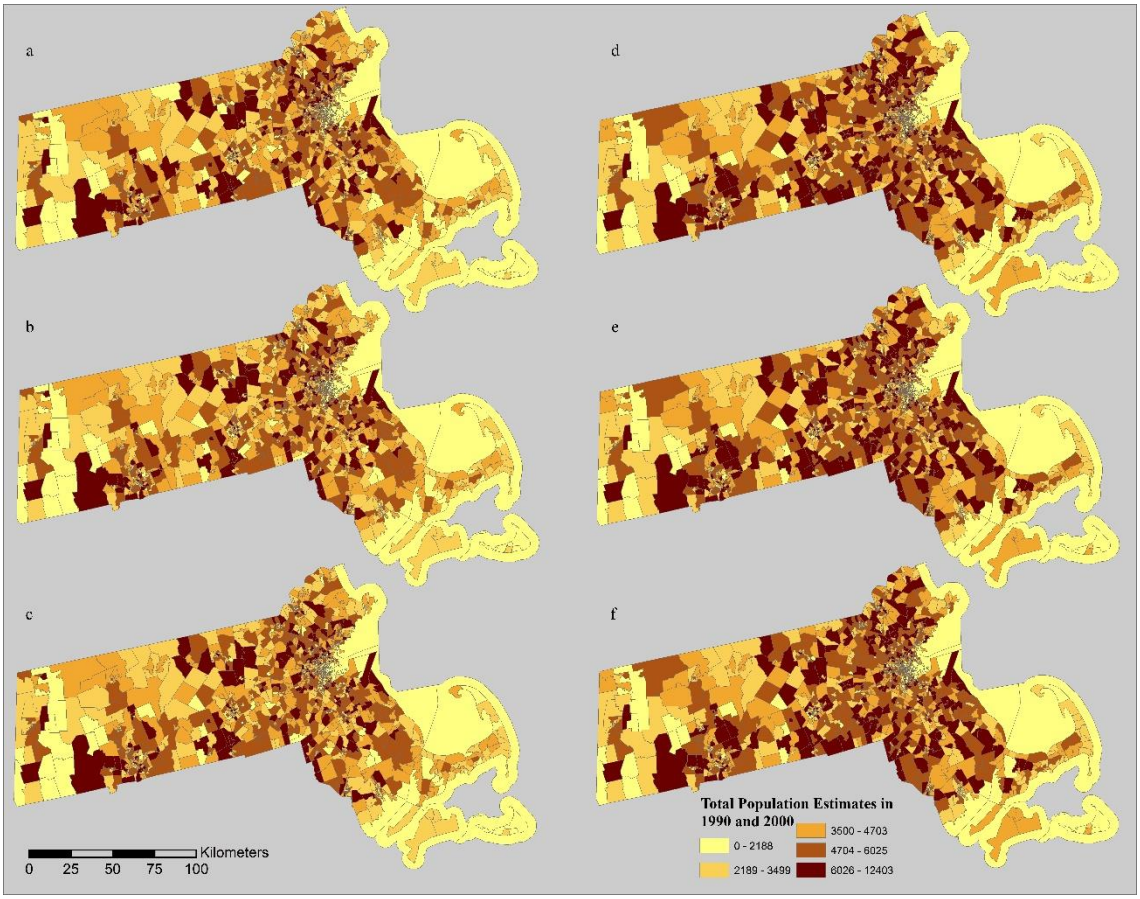

Figure 5.4. Total population maps at the target zone level: (a) in1990 based on TDW, (b) in 1990 based on TDW refined by buildings, (c) in 1990 based on block aggregation, (d) in 2000 based on TDW, (e) in 2000 based on TDW refined by ZTRAX®[1], and (f) in 2000 based on block aggregation.

Similar to the presentation of the results for total population, Figure 5.5 illustrates the absolute error maps of white population estimates. The best-performing methods for the 1990-2010 and 2000-2010 time periods are the same as those for total population. Figure 5.6 shows the maps of white population estimates.

[1] The dataset was provided by the Zillow Inc. (https://www.zillow.com/research/data/), and the source code for its process is at https://github.com/spatialhistory/ztrax_analysis.

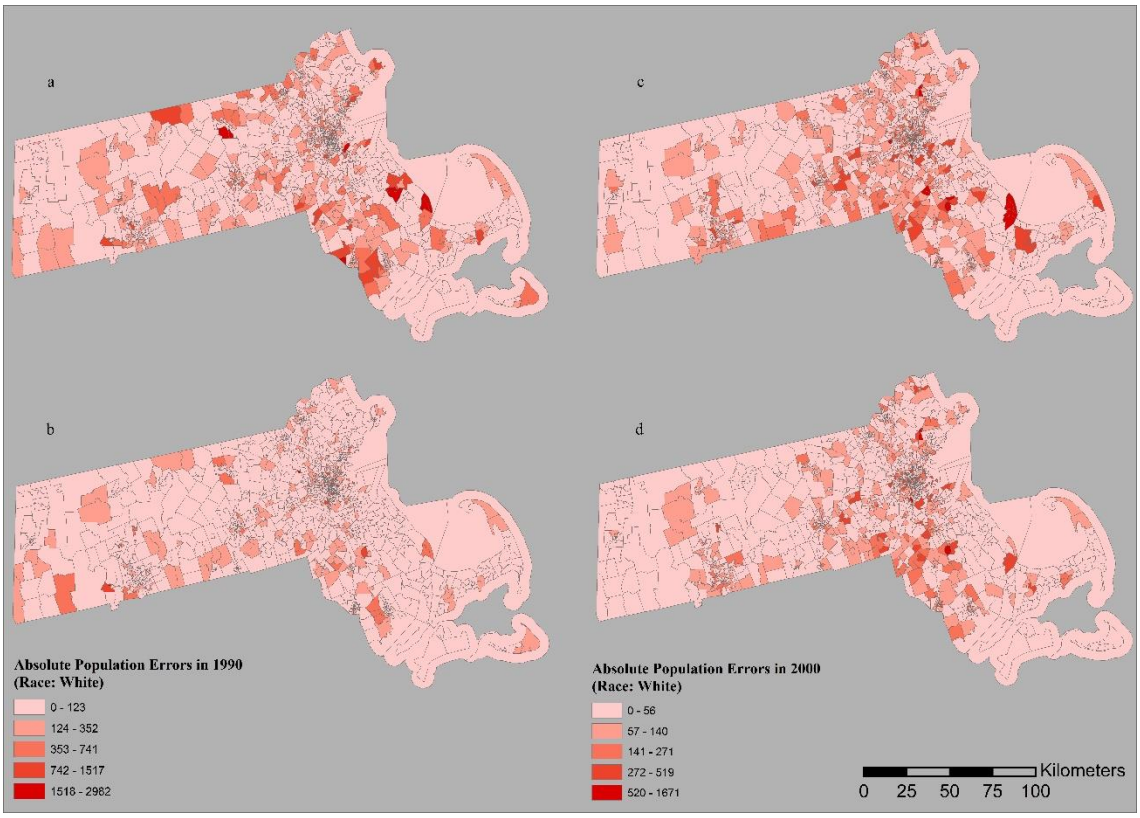

Figure 5.5. Absolute error maps of white population at the target zone level: (a) in 1990 based on TDW, (b) in 1990 based on TDW refined by buildings, (c) in 2000 based on TDW, and (d) in 2000 based on TDW refined by ZTRAX®[1].

[1] The dataset was provided by the Zillow Inc. (https://www.zillow.com/research/data/), and the source code for its process is at https://github.com/spatialhistory/ztrax_analysis.

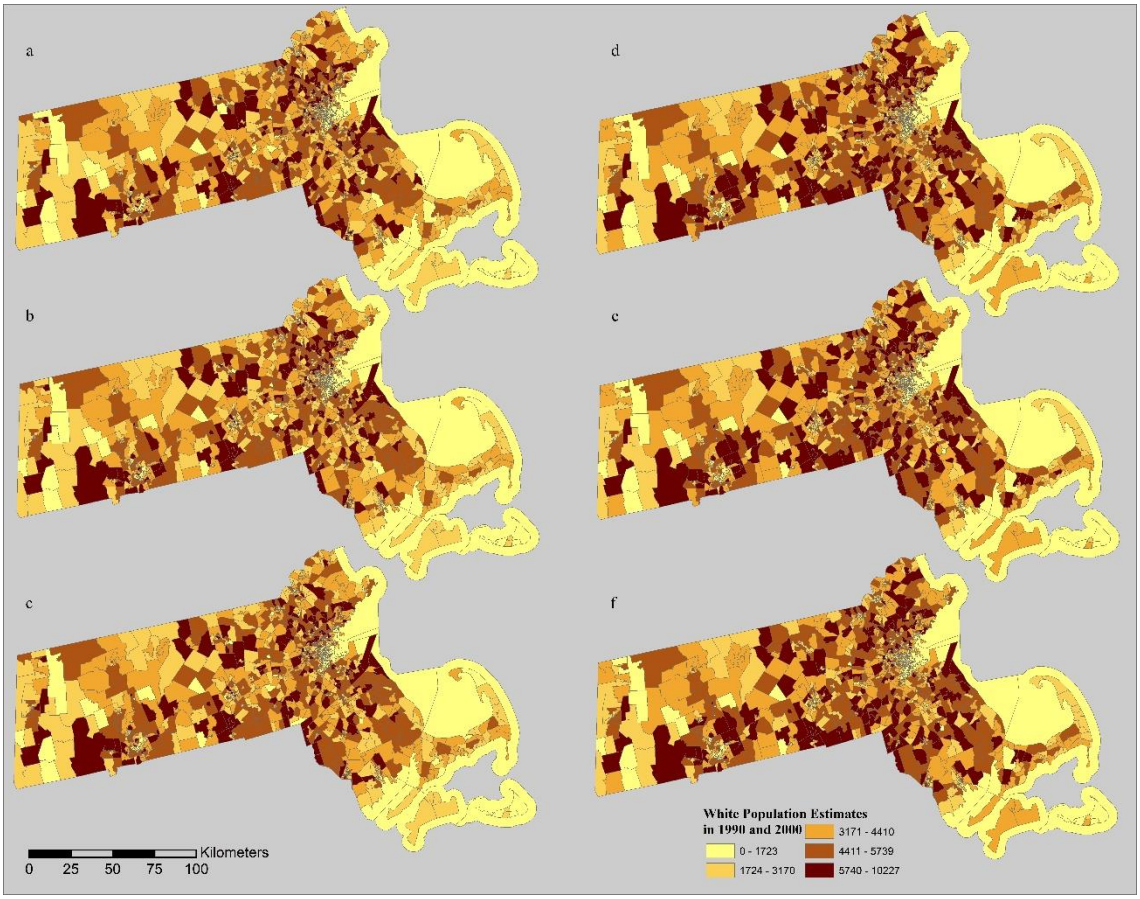

Figure 5.6. Maps of white population at the target zone level: (a) in 1990 based on TDW, (b) in 1990 based on TDW refined by buildings, (c) in 1990 based on block aggregation, (d) in 2000 based on TDW, (e) in 2000 based on TDW refined by ZTRAX®[1], and (f) in 2000 based on block aggregation.

Figure 5.7 shows the absolute error maps of population estimates aged under 65. The best-performing methods for the 1990-2010 and 2000-2010 time periods are also identical to those for total population. Figure 5.8 shows the maps of estimates of the population sub-group.

[1] The dataset was provided by the Zillow Inc. (https://www.zillow.com/research/data/), and the source code for its process is at https://github.com/spatialhistory/ztrax_analysis.

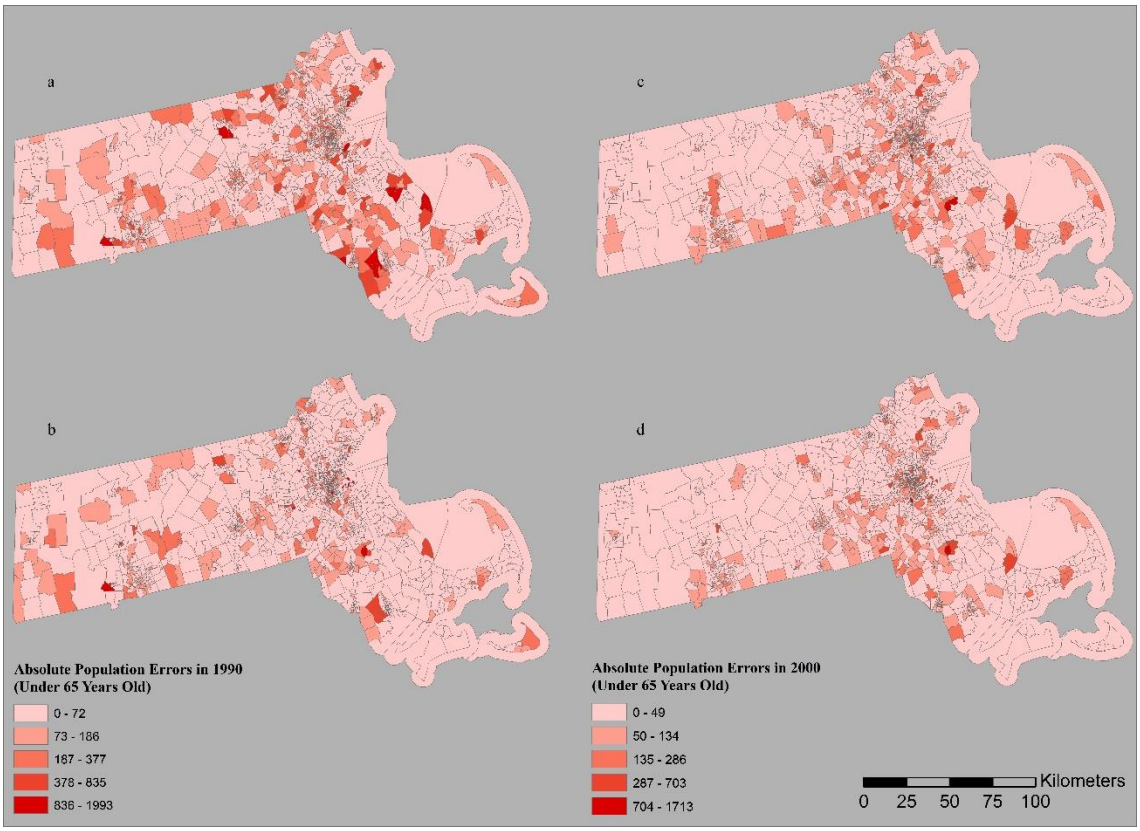

Figure 5.7. Absolute error maps of population aged under 65 at the target zone level: (a) in 1990 based on TDW, (b) in 1990 based on TDW refined by buildings, (c) in 2000 based on TDW and (d) in 2000 based on TDW refined by ZTRAX®[1].

[1] The dataset was provided by the Zillow Inc. (https://www.zillow.com/research/data/), and the source code for its process is at https://github.com/spatialhistory/ztrax_analysis.

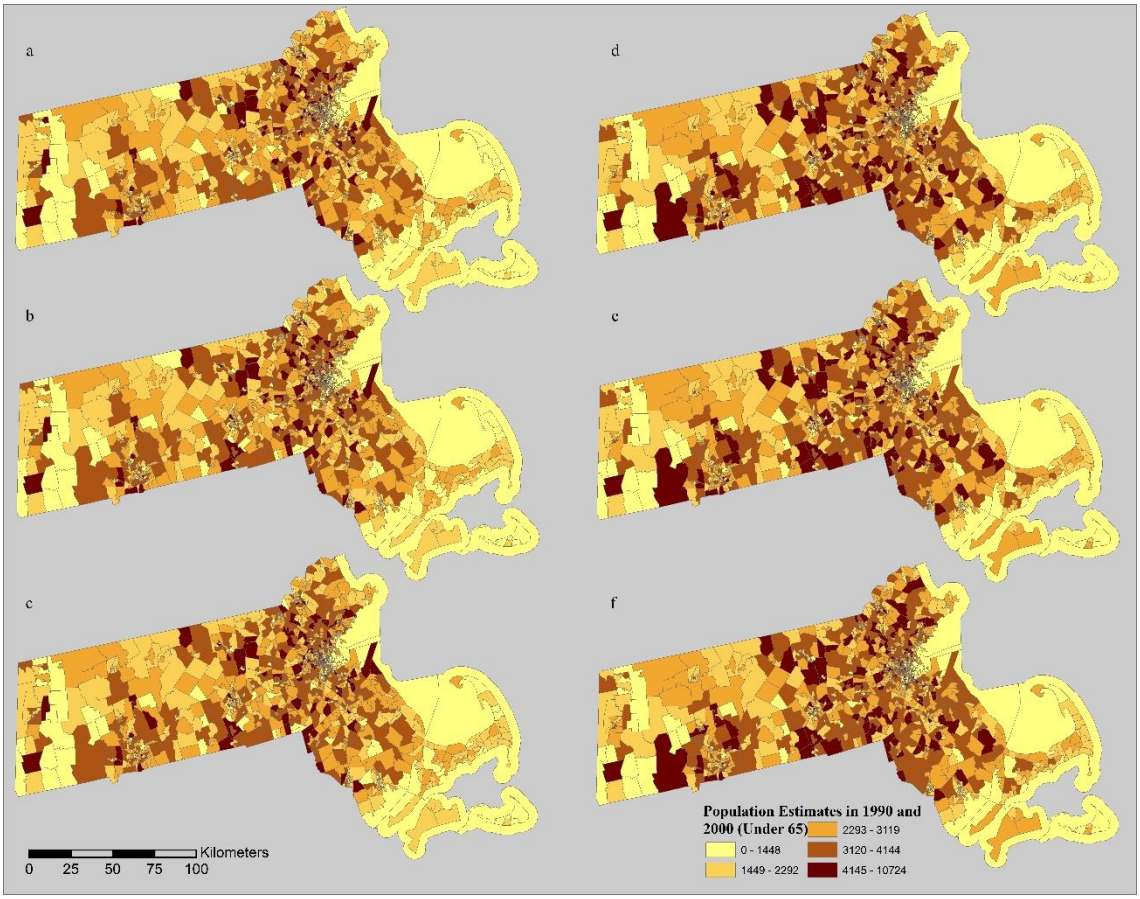

Figure 5.8. Maps of population aged under 65 at the target zone level: (a) in 1990 based on TDW, (b) in 1990 based on TDW refined by buildings, (c) in 1990 based on block aggregation, (d) in 2000 based on TDW, (e) in 2000 based on TDW refined by ZTRAX®[1] and (f) in 2000 based on block aggregation.

The maps in Figures 5.9 and 5.10 show the population-normalized absolute errors pertaining to the above-mentioned demographic attributes for the 1990-2010 and 2000-2010 time periods, respectively. For normalization, the absolute errors for estimating the three demographic attributes shown in Figures 5.3, 5.5 and 5.7 are divided by the block-aggregated values of the respective year. Normalized errors are generally between 0 and 1, but can exceed 1 if the absolute error for a target tract is higher than its reference value, which can be observed in tracts with small

[1] The dataset was provided by the Zillow Inc. (https://www.zillow.com/research/data/), and the source code for its process is at https://github.com/spatialhistory/ztrax_analysis.

populations. Normalized error distributions provide a more objective comparison between target tracts across study areas and between different time periods.

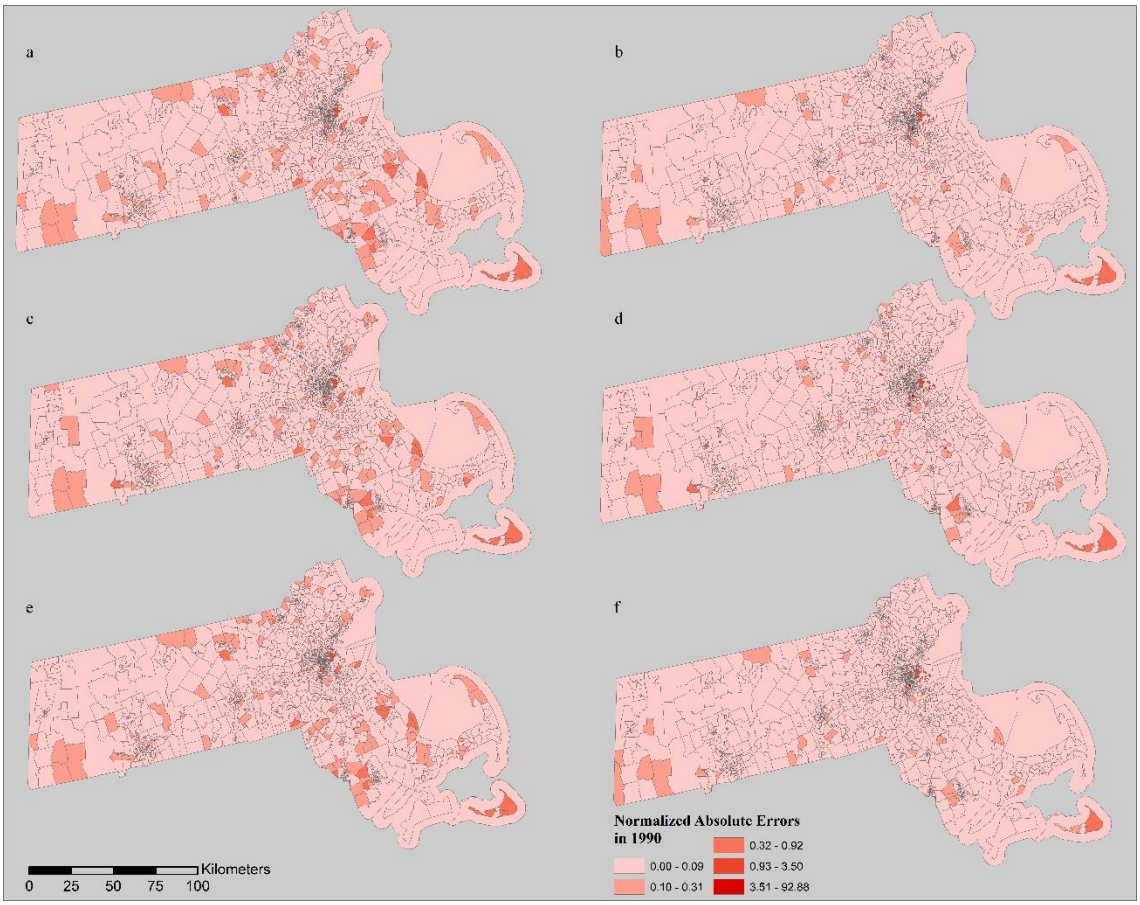

Figure 5.9. Normalized absolute error maps in 1990 for: (a) total population using TDW, (b) for total population using TDW refined by buildings, (c) population aged under 65 using TDW, (d) population aged under 65 using TDW refined by buildings, (e) white population using TDW, and (f) white population using TDW refined by buildings.

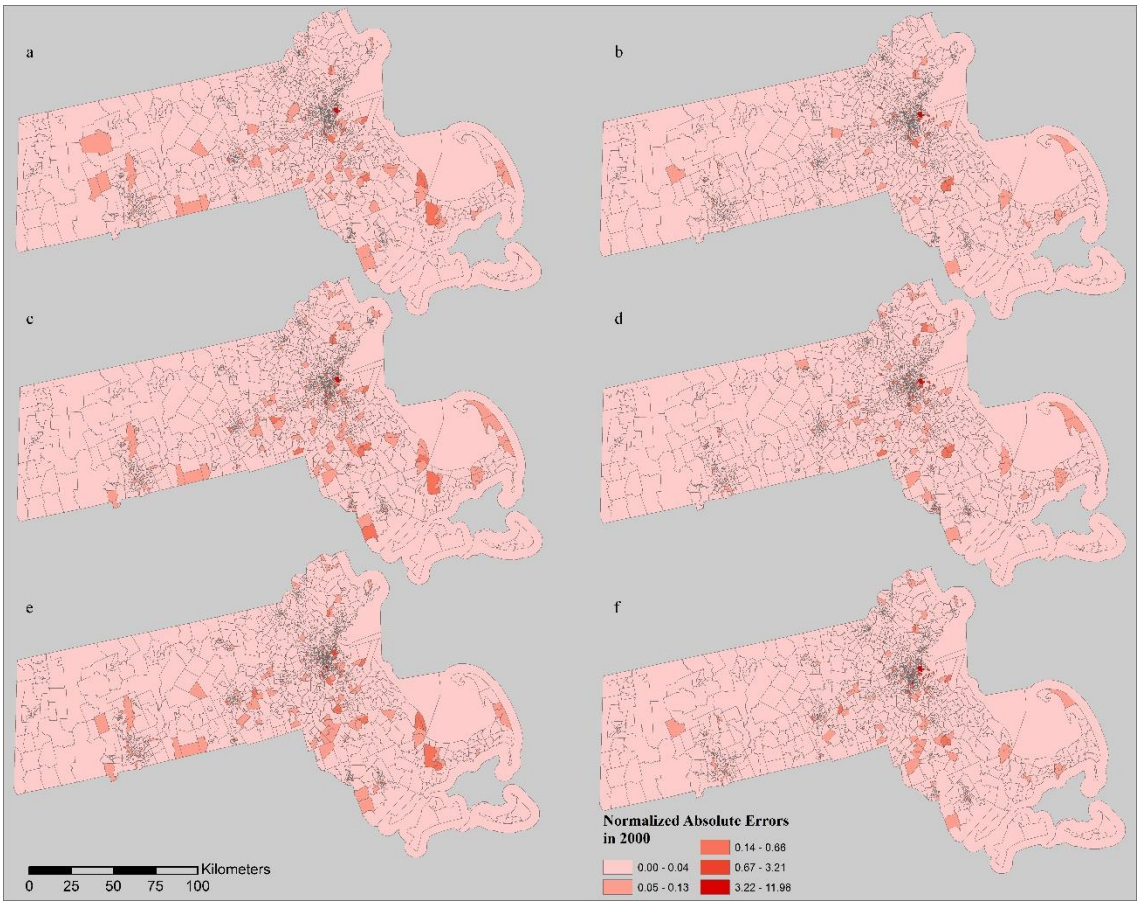

Figure 5.10. Normalized absolute error maps in 2000 for: (a) total population using TDW, (b) for total population using TDW refined by ZTRAX®[1], (c) population aged under 65 using TDW, (d) population aged under 65 using TDW refined by ZTRAX®, (e) white population using TDW, and (f) white population using TDW refined by ZTRAX®.

The above figures show only the results of TDW and the best-performing refined TDW method, depending on the demographic attribute and year. An extensive summary of the performance of the different methods in estimating demographic attributes from 1990 and 2000 within 2010 target tract boundaries can be found in Tables 5.1 to 5.4. These tables include four error metrics associated with estimates of total population, race (white), race (black), and age

[1] The dataset was provided by the Zillow Inc. (https://www.zillow.com/research/data/), and the source code for its process is at https://github.com/spatialhistory/ztrax_analysis.

(under 65), respectively, and thus allow a more detailed comparison between the different methods according to the two time periods and for various demographic attributes.

Table 5.1. Absolute error measures pertaining to total population estimates.

	MAE	Median Abs Error	RMSE	90th Percentile Abs Error	Ancillary Variable
Method	Time Period: 1990-2010				
AW	376	66	842	1252	None
TDW	162	75	312	388	None
RefTDW	168	64	327	454	NLCD
RefTDW	136	67	239	336	GHSL
RefTDW	143	75	247	359	NLCD-GHSL
RefTDW	131	57	260	313	Parcels
RefTDW	**101**	**49**	**207**	**232**	**Buildings**
RefTDW	104	44	246	240	ZTRAX[1]
EM	174	45	377	531	NLCD
EM	353	77	773	1141	Parcels
EM	154	50	346	409	Buildings
EM	162	49	391	438	ZTRAX
EEM	158	51	386	387	Parcels
EEM	147	48	334	368	Buildings
	Time Period: 2000-2010				
AW	340	11	1449	1084	None
TDW	55	12	135	150	None
RefTDW	47	7	129	127	NLCD
RefTDW	55	8	151	168	GHSL
RefTDW	51	7	143	154	NLCD-GHSL
RefTDW	51	6	171	129	Parcels
RefTDW	43	5	150	97	Buildings
RefTDW	**41**	**5**	**144**	**89**	**ZTRAX**
EM	142	7	447	366	NLCD
EM	191	6	621	569	Parcels
EM	87	5	300	229	Buildings
EM	92	4	320	238	ZTRAX
EEM	96	5	355	247	Parcels
EEM	82	5	281	218	Buildings

[1] The dataset was provided by the Zillow Inc. (https://www.zillow.com/research/data/), and the source code for its process is at https://github.com/spatialhistory/ztrax_analysis.

Table 5.2. Absolute error measures pertaining to white population estimates.

	MAE	Median Abs Error	RMSE	90th Percentile Abs Error	Ancillary Variable
Method	Time Period: 1990-2010				
AW	343	55	772	1184	None
TDW	146	63	292	359	None
RefTDW	127	56	230	327	NLCD
RefTDW	120	58	214	301	GHSL
RefTDW	128	66	223	321	NLCD-GHSL
RefTDW	117	50	233	284	Parcels
RefTDW	**87**	**42**	**178**	**204**	**Buildings**
RefTDW	91	37	216	214	ZTRAX[1]
EM	203	54	417	649	NLCD
EM	304	65	667	990	Parcels
EM	127	41	271	330	Buildings
EM	130	39	308	344	ZTRAX
EEM	132	43	303	332	Parcels
EEM	120	43	256	308	Buildings
	Time Period: 2000-2010				
AW	313	53	1281	837	None
TDW	77	44	131	180	None
RefTDW	71	38	126	175	NLCD
RefTDW	78	41	142	194	GHSL
RefTDW	74	38	136	178	NLCD-GHSL
RefTDW	75	39	157	172	Parcels
RefTDW	68	36	138	156	Buildings
RefTDW	**66**	**36**	**133**	**152**	**ZTRAX**
EM	143	42	346	345	NLCD
EM	182	41	479	464	Parcels
EM	95	37	209	232	Buildings
EM	94	37	211	243	ZTRAX
EEM	101	39	246	224	Parcels
EEM	90	36	191	212	Buildings

[1] The dataset was provided by the Zillow Inc. (https://www.zillow.com/research/data/), and the source code for its process is at https://github.com/spatialhistory/ztrax_analysis.

Table 5.3. Absolute error measures pertaining to black population estimates.

	MAE	Median Abs Error	RMSE	90th Percentile Abs Error	Ancillary Variable
Method	Time Period: 1990-2010				
AW	17	3	65	33	None
TDW	11	3	30	24	None
RefTDW	11	3	32	26	NLCD
RefTDW	11	3	31	24	GHSL
RefTDW	11	3	33	26	NLCD-GHSL
RefTDW	11	2	35	27	Parcels
RefTDW	11	2	39	24	Buildings
RefTDW	11	2	36	26	ZTRAX[1]
EM	14	2	51	27	NLCD
EM	17	2	62	34	Parcels
EM	15	2	50	29	Buildings
EM	16	2	53	31	ZTRAX
EEM	15	2	52	29	Parcels
EEM	15	2	51	32	Buildings
	Time Period: 2000-2010				
AW	23	2	94	41	None
TDW	10	2	32	25	None
RefTDW	9	2	31	25	NLCD
RefTDW	10	2	33	24	GHSL
RefTDW	10	2	32	25	NLCD-GHSL
RefTDW	9	2	30	24	Parcels
RefTDW	9	1	32	24	Buildings
RefTDW	10	2	31	25	ZTRAX
EM	15	2	50	30	NLCD
EM	17	2	73	31	Parcels
EM	15	2	56	28	Buildings
EM	15	2	56	30	ZTRAX
EEM	15	2	57	30	Parcels
EEM	15	2	55	31	Buildings

[1] The dataset was provided by the Zillow Inc. (https://www.zillow.com/research/data/), and the source code for its process is at https://github.com/spatialhistory/ztrax_analysis.

Table 5.4. Absolute error measures pertaining to estimates of population aged under 65.

	MAE	Median Abs Error	RMSE	90th Percentile Abs Error	Ancillary Variable
Method	Time Period: 1990-2010				
AW	248	42	578	797	None
TDW	110	49	210	267	None
RefTDW	99	45	182	248	NLCD
RefTDW	97	46	180	246	GHSL
RefTDW	101	50	184	252	NLCD-GHSL
RefTDW	94	39	196	224	Parcels
RefTDW	**76**	**33**	**164**	**180**	**Buildings**
RefTDW	78	31	198	182	ZTRAX[1]
EM	150	41	319	456	NLCD
EM	231	48	536	712	Parcels
EM	112	33	279	281	Buildings
EM	116	30	313	295	ZTRAX
EEM	116	34	313	287	Parcels
EEM	110	33	275	292	Buildings
	Time Period: 2000-2010				
AW	245	44	934	735	None
TDW	63	33	114	151	None
RefTDW	60	31	114	140	NLCD
RefTDW	65	33	127	154	GHSL
RefTDW	62	32	123	144	NLCD-GHSL
RefTDW	62	32	134	136	Parcels
RefTDW	59	30	125	130	Buildings
RefTDW	**57**	**29**	**120**	**128**	**ZTRAX**
EM	121	34	304	294	NLCD
EM	151	35	434	373	Parcels
EM	90	33	236	215	Buildings
EM	92	32	259	214	ZTRAX
EEM	97	32	290	222	Parcels
EEM	87	32	229	215	Buildings

5.4.2. Spatial refinement for urban population

Urban population estimates for both 1990 to 2010 and 2000 to 2010 time periods are derived based on two underlying assumptions. First, it is assumed that no representation of urban

[1] The dataset was provided by the Zillow Inc. (https://www.zillow.com/research/data/), and the source code for its process is at https://github.com/spatialhistory/ztrax_analysis.

footprints exists in running the temporal interpolation. Second, the census-defined urban areas in 1990 and 2000 are used, respectively, as well as additional ancillary variables for further refinement. Consequently, the maps of absolute errors in estimating urban population (Figure 5.11) cover three implementations in both 1990 and 2000, namely TDW, the best performing method refined by the ancillary variable extending outside census-defined urban areas, and the best performing method using the ancillary variable limited within the urban areas. Moreover, Figure 5.12 shows the urban population maps based on those approaches in 1990 and 2000.

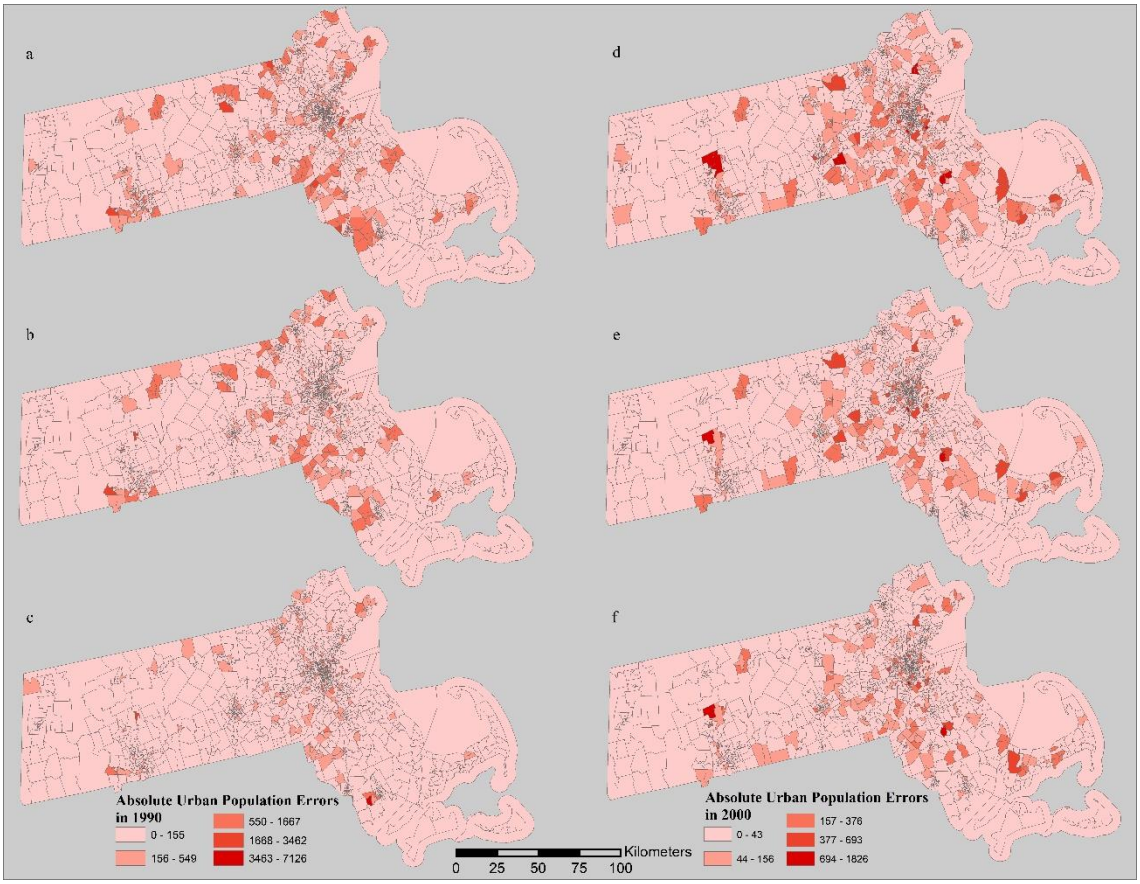

Figure 5.11. Absolute error maps of urban population at the target zone level: (a) in 1990 based on TDW, (b) in 1990 based on TDW refined by ZTRAX®[1] (not limited to urban areas), (c) in 1990 based on TDW refined by ZTRAX® (limited to urban areas) as well as (d) in 2000 based on TDW, (e) in 2000 based on TDW refined by ZTRAX® (not limited to urban areas), (f) in 2000 based on TDW refined by ZTRAX® (limited to urban areas).

[1] The dataset was provided by the Zillow Inc. (https://www.zillow.com/research/data/), and the source code for its process is at https://github.com/spatialhistory/ztrax_analysis.

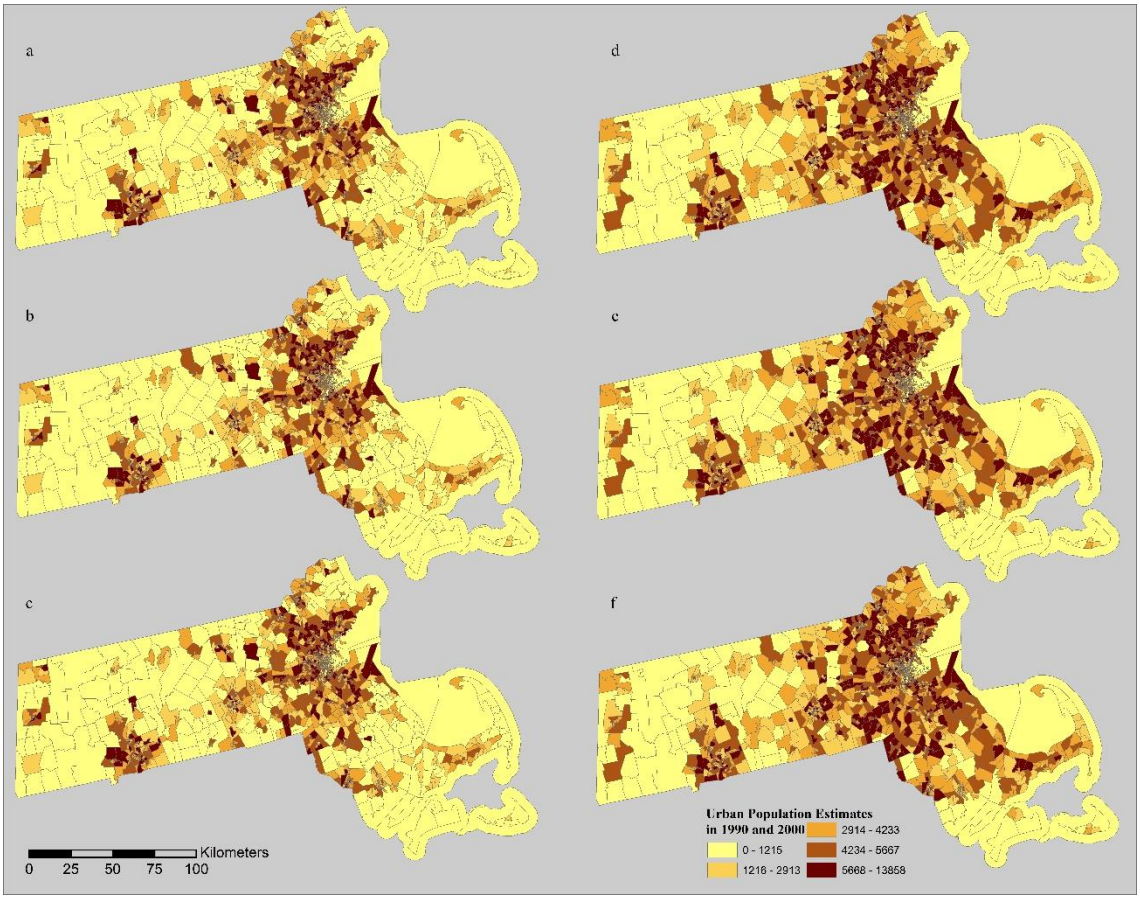

Figure 5.12. Resulting maps of urban population at the target zone level: (a) in 1990 based on TDW, (b) in 1990 based on TDW refined by ZTRAX®[1] (not limited to urban areas), (c) in 1990 based on TDW refined by ZTRAX® (limited to urban areas) as well as (d) in 2000 based on TDW, (e) in 2000 based on TDW refined by ZTRAX® (not limited to urban areas), (f) in 2000 based on TDW refined by ZTRAX® (limited to urban areas).

Tables 5.5 and 5.6 provide a detailed view on the error distributions of the different methods for interpolating urban population values in 1990 and 2000 within census tract boundaries in 2010, respectively. Each table is composed of two parts: the upper part shows the error estimates when additional refinement is employed within the census-defined urban areas, whereas in the lower part, the refinement is employed inside and outside the urban areas. These two tables show

[1] The dataset was provided by the Zillow Inc. (https://www.zillow.com/research/data/), and the source code for its process is at https://github.com/spatialhistory/ztrax_analysis.

how the accuracy measures of one areal interpolation method can change depending on whether the method is applied to an individual ancillary variable or its intersection with the census defined urban areas in the respective year. As shown in Tables 5.5 and 5.6, the most accurate results are achieved when TDW is refined to the composite census-defined urban areas and ZTRAX® data, in both 1990 and 2000.

Table 5.5. Absolute error measures pertaining to urban population estimates in 1990 within 2010 tract boundaries.

	MAE	Median Abs Error	RMSE	90th Percentile Abs Error	Ancillary Variable
Method	Limited to Urban Areas				
AW	353	58	835	1146	None
TDW	166	62	337	437	None
RefTDW	187	50	523	444	Urban Areas
RefTDW	146	44	365	387	NLCD
RefTDW	142	51	370	322	GHSL
RefTDW	140	50	347	359	NLCD-GHSL
RefTDW	258	59	1402	486	Parcels
RefTDW	107	43	281	257	Buildings
RefTDW	**95**	**38**	**282**	**226**	**ZTRAX[1]**
EM	132	31	313	390	NLCD
EM	132	33	307	371	NLCD-GHSL
EM	199	47	477	585	Parcels
EM	127	40	320	317	Buildings
EM	132	37	332	347	ZTRAX
EEM	121	41	318	267	Parcels
EEM	124	41	315	304	Buildings
	Not Limited to Urban Areas				
RefTDW	152	50	298	427	NLCD
RefTDW	143	58	276	363	GHSL
RefTDW	142	53	278	407	NLCD-GHSL
RefTDW	176	61	363	459	Parcels
RefTDW	121	47	245	305	Buildings
RefTDW	119	42	268	295	ZTRAX
EM	188	48	405	599	NLCD
EM	169	43	363	513	NLCD-GHSL
EM	346	69	793	1184	Parcels
EM	196	50	442	593	Buildings
EM	195	44	464	626	ZTRAX
EEM	172	48	412	480	Parcels
EEM	185	49	419	538	Buildings

[1] The dataset was provided by the Zillow Inc. (https://www.zillow.com/research/data/), and the source code for its process is at https://github.com/spatialhistory/ztrax_analysis.

Table 5.6. Absolute error measures pertaining to urban population estimates in 2000 within 2010 tract boundaries.

	MAE	Median Abs Error	RMSE	90th Percentile Abs Error	Ancillary Variable
Method	Limited to Urban Areas				
AW	322	11	1393	1006	None
TDW	60	10	152	164	None
RefTDW	76	8	237	170	Urban Areas
RefTDW	49	4	147	126	NLCD
RefTDW	59	6	184	148	GHSL
RefTDW	57	5	171	148	NLCD-GHSL
RefTDW	438	6	3438	610	Parcels
RefTDW	46	4	138	119	Buildings
RefTDW	**40**	**4**	**122**	**98**	**ZTRAX[1]**
EM	128	5	428	322	NLCD
EM	138	5	446	402	NLCD-GHSL
EM	142	4	546	317	Parcels
EM	77	2	281	194	Buildings
EM	84	3	302	202	ZTRAX
EEM	88	3	349	189	Parcels
EEM	74	3	269	184	Buildings
	Not Limited to Urban Areas				
RefTDW	51	6	145	143	NLCD
RefTDW	57	7	164	166	GHSL
RefTDW	55	6	157	159	NLCD-GHSL
RefTDW	457	8	3576	636	Parcels
RefTDW	48	4	138	137	Buildings
RefTDW	45	4	134	122	ZTRAX
EM	139	6	442	390	NLCD
EM	142	6	450	436	NLCD-GHSL
EM	192	5	633	514	Parcels
EM	101	4	340	244	Buildings
EM	99	3	343	284	ZTRAX
EEM	95	5	353	255	Parcels
EEM	93	4	310	270	Buildings

5.4.3. Enhanced areal interpolation for improved natural hazards risk assessment

Tables 5.7 and 5.8 present the estimated population sub-groups based on race and age

[1] The dataset was provided by the Zillow Inc. (https://www.zillow.com/research/data/), and the source code for its process is at https://github.com/spatialhistory/ztrax_analysis.

living in designated flood zones using AW, refined AW, EM and EEM. AW is added to show the excessive and exaggerated population counts that would be calculated without any refinement on census tracts. No actual population counts exist for flood zones that can be used as references. Therefore, in this application, census blocks in 1990 and 2000 are used, and their relevant population counts are refined and transferred to flood zones using EEM refined by building footprints to calculate benchmark estimates of potentially exposed populations.

Table 5.7. Estimated exposed population sub-groups based on race in 1990 and 2000 within 2010 tract boundaries.

	1990		2000		
Method	White	Black	White	Black	Ancillary Variable
Block-level Benchmark	215051	7213	208512	8774	Buildings
Tract-level AW	781452	22402	783798	28789	None
RefAW	425569	12433	434762	15901	Parcels
RefAW	**235903**	**7868**	**230047**	**9548**	**Buildings**
RefAW	277503	8933	278701	10781	ZTRAX[1]
EM	461037	12318	471498	15712	Parcels
EM	241443	7606	235501	9234	Buildings
EM	286118	8724	286443	10483	ZTRAX
EEM	317630	9992	316827	11895	Parcels
EEM	245600	7720	239823	9319	Buildings

[1] The dataset was provided by the Zillow Inc. (https://www.zillow.com/research/data/), and the source code for its process is at https://github.com/spatialhistory/ztrax_analysis.

Table 5.8. Estimated exposed population sub-groups based on age in 1990 and 2000 within 2010 tract boundaries.

	1990			2000			
Method	-18	-65	+65	-18	-65	+65	Ancillary Variable
Block-level Benchmark	53619	151633	35840	56899	155561	37346	Buildings
Tract-level AW	190591	539056	117769	213359	562081	126418	None
RefAW	105339	294199	65731	118779	312022	71885	Parcels
RefAW	**60823**	**163723**	**39472**	**64822**	**169158**	**41223**	**Buildings**
RefAW	70881	191577	44637	77995	201262	47843	ZTRAX[1]
EM	110103	318716	69416	124932	336851	75208	Parcels
EM	61326	169572	40096	65137	175618	41628	Buildings
EM	71919	198889	45193	78897	208891	48276	ZTRAX
EEM	80060	224833	49391	86615	234554	52922	Parcels
EEM	61386	173978	40401	64991	180675	42053	Buildings

To determine if any population sub-group is disproportionately distributed within flood zones, an expected value for each sub-group and year should be calculated based on broader general distributions of the population within more aggregated regions. Thus, the ratio of each sub-group to the total population in the whole study area is calculated and multiplied by the estimated potentially exposed population within designated flood zones resulting from the benchmark analysis. These values indicate if the estimated population counts (sub-groups) within flood zones deviate from the expected share of the total population. For example, if the ratio of the observed affected population to the expected count for a sub-group is greater than 1, this indicates that the population is more exposed than it would be expected based on the total population. Such observations could indicate disadvantaged population groups due to their elevated exposure to natural hazards.

[1] The dataset was provided by the Zillow Inc. (https://www.zillow.com/research/data/), and the source code for its process is at https://github.com/spatialhistory/ztrax_analysis.

Figure 5.13 shows the ratios of observed to expected population subgroups (race, age-structures) in both 1990 and 2000. The ratios are calculated using both the benchmark approach and the best-performing method, which is determined based on the similarity of its results to benchmark estimates. In all cases, this was AW refined by building footprints. The figure visualizes which population sub-groups are estimated to reside within flood zones more than expected. The dashed line marks the case where the populations of sub-groups would be distributed within flood zones completely proportionately.

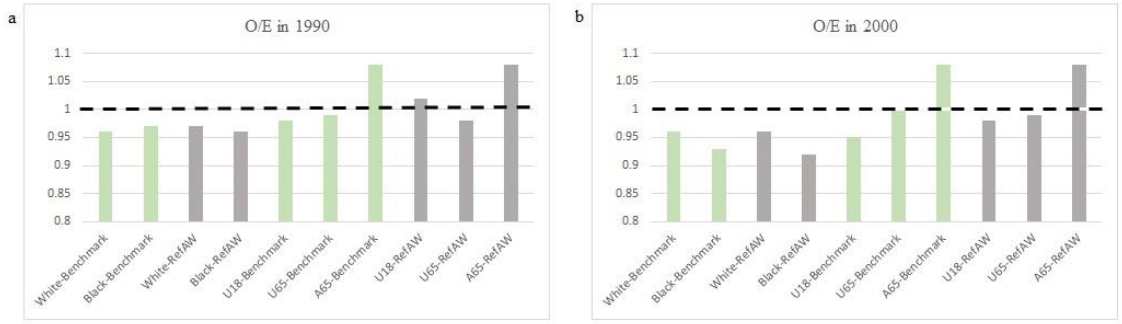

Figure 5.13. Observed/expected ratios for race and age subgroups in (a) 1990 and (b) 2000.

5.5. Discussion and conclusions

5.5.1. Multi-temporal estimates of demographic attributes

The above results show that the spatially refined interpolation methods produce improved estimates of the demographic attributes in most cases, particularly in interpolating population counts in 1990 within target tract boundaries of the 2010 Census. Error measures of refined AW methods, which are not included in the result tables, show higher errors but similar patterns and are even more consistent than the refined TDW approaches. Nonetheless, the results allow a critical reflection on the refinement effect, which sometimes results in increased errors. In general, estimates based on a refined interpolation method can be associated with lower absolute errors. However, the existence of a few target zones with excessive errors can result in lower overall

accuracy measures. For example, the MAE, RMSE and 90th percentile of absolute errors are higher in the refined TDW using NLCD than in the regular, unrefined TDW when estimating total population counts in 1990 at the target tract level (Table 5.1) whereas the median absolute error is lower. This suggests that large errors exist in the right tail of the error distribution, causing the overall error measures to rise. As discussed in Schroeder (2007), the uncertainty level has a positive statistical relationship with initial population counts and dissimilarity between boundaries of source and target zones. Thus, when interpolating attributes of small sub-groups such as black population (Table 5.3) or population aged above 65 in the present study, or transferring demographic attributes from boundaries in 2000 to those in 2010, which often do not change drastically (the bottom parts of Tables 5.1 to 5.4), the uncertainty of the process is expected to be lower, thereby reducing the visible potential gain of spatial refinement. It should also be noted that exogenous factors related to particular attributes might affect their estimation accuracy in areal interpolation approaches, regardless of the effectiveness of spatial refinement. For example, if the current framework will include 2020 in future applications, the existence of baby boomers, as a large demographic cohort, can lead to inflated distributions of the population group aged above 65 in 2020. This can consequently result in biased temporal estimations of the population group, especially using a method such as TDW – with and without refinement – that assumes ratios of population densities remain the same over time. Therefore, all these factors should be considered when interpreting the results of the temporal interpolation of enumerated attributes.

The quality of the ancillary variable also influences the gain in accuracy of spatial refinement (Langford 2013). For example, the developed classes of NLCD for refinement were initially selected following Ruther *et al.* (2015), but resulted in higher absolute error measures in NLCD-based refined methods. Subsequently, different combinations of the developed classes

among the different releases were tested to determine the most accurate results. In this study, the classes 21, 22 and 23 from the NLCD 1992, and 21, 22, 23 and 24 from the NLCD versions of 2001 and 2011, respectively, result in the highest accuracy. The achieved high performance of the methods that include the higher density developed classes of NLCD may be related to densely populated areas such as Boston, where distributions of high population densities are modeled more reliably by doing so. The TDW refined by GHSL built-up areas result in a lower level of accuracy compared to other refinement strategies for 2000-2010 (Tables 5.1 to 5.4). As a possible explanation, Landsat-based GHSL has lower levels of classification accuracy in rural settings, a direct consequence of mixed pixel problems in the remote sensing imagery (Leyk *et al.* n.d.). Also, GHSL models built-up areas additively, assuming that built-up areas in each epoch are the summation of the areas from the previous epoch plus those built in between, which propagates classification errors through the different time periods. Furthermore, the combination of ancillary variables for spatial refinement does not necessarily lead to improved performance, especially when one of the participating variables is associated with higher errors, as indicated by the combinations of NLCD and GHSL in Tables 5.1 to 5.4.

Building footprints and ZTRAX®, two ancillary variables that are available locally and as a national proprietary data product, respectively, perform generally well in reducing the absolute error measures using refined TDW across different demographic attributes and time periods (Tables 5.1 to 5.4). For example, according to Table 5.1, the spatial refinement of TDW using building footprints and ZTRAX® leads to 23% and 21% MAE reductions, respectively, relative to the next most accurate refined TDW implementation for 1990-2010. Parcels, however, as another local variable, do not demonstrate comparable accuracy gains, probably due to their overestimation effect in rural areas discussed in Leyk *et al.* (2014). The performance of ZTRAX®, indicated in

Tables 5.1, 5.2 and 5.4, is promising for future studies because the dataset is available and semantically consistent nationally, as opposed to buildings and parcels, which are released by county and city authorities often with varying land-use classes and housing attributes. While the ZTRAX® data is not free, such proprietary data sources become increasingly available for scientific purposes. However, for attributes with small enumerated counts over source and target zones, and applications over short time periods, the supremacy of local and proprietary ancillary datasets over the publically available national (NLCD) and global (GHSL) data layers is diminished (Table 5.3 and the bottom parts of Tables 5.1-5.4). This suggests that the benefits of incorporating such datasets are fully tapped when the analysis covers longer time periods.

Although EM leverages other attributes of ancillary variables such as land-use classes in addition to their geometric footprints (Schroeder and Van Riper 2013), the resulting accuracy level in this study is low compared to the different refined TDW approaches, especially when utilizing parcels (Tables 5.1 to 5.4). This finding may be related to the assumption that each control zone can be associated with one population density weight regardless of the underlying area diversity within the control zone and ignoring possible non-stationarity effects. EEM, on the other hand, attempts to address the issue and successfully reduces the absolute error measures of EM (Zoraghein and Leyk n.d.), particularly when using parcels as the ancillary variable. For example, while the MAE values pertaining to the total population estimation using parcels are reduced by 55% and 50% in 1990 and 2000, respectively, they drop by only 5% and 6% when using buildings (Table 5.1). This indicates that the issue of area diversity within buildings is much less severe than within parcels. Moreover, the improvement effect of EEM is not substantial when population counts over source and target zones are small (Table 5.3). It should be noted that EM is not applicable to GHSL because each epoch represents a binary partition of built-up and non-built-up.

Furthermore, EEM cannot be operationalized using the NLCD and ZTRAX® datasets as ancillary variables because these are inherently grid-based data layers.

According to Tables 5.1 to 5.4, the accuracy of each EEM implementation is always lower than its counterpart refined TDW. EEM employing buildings is not effective, and its accuracy level is similar to EM. For parcels, however, one can expect that EEM may outperform refined TDW over longer time periods, as suggested by the current patterns observed in the tables. First, the accuracy attainment of TDW over AW is substantially higher for the shorter time period than for the longer one, implying the high performance of TDW over short durations. For example, the ratio of the MAE of TDW to AW is 0.43 for 1990-2010 while the value drops to 0.16 for 2000-2010 (Table 5.1). Second, the accuracy improvement of refined TDW compared to EEM decreases over time. For instance, the ratio of the MAE of EEM to refined TDW using parcels drops from 1.88 in 2000-2010 to 1.21 in 1990-2010, implying that the assumptions of refined TDW are less reliable as the time period is extended. Those two observations suggest that EEM has the potential to outperform refined TDW for longer historical applications.

One possible reason why EEM has lower accuracy levels than refined TDW in this state-based analysis can be related to the spatial non-stationarity (the spatial variation of statistical relationships across regions). Although EEM is assumed to be applied on more homogeneous control zones for assigning population density weights, each zone can have varying regimes of population density across the study area, making one population density weight per control zone unrepresentative. Employing solutions that allow population density weights of individual control zones to vary spatially can help clarify if non-stationarity is indeed an issue in EEM. It is noteworthy to mention that unlike refined TDW, EEM is pycnophylactic.

Absolute error maps of the selected demographic attributes (Figures 5.3, 5.5 and 5.7)

visualize the improvement effect of the most accurate methods compared to TDW and coincide with the outcomes summarized in Tables 5.1, 5.2 and 5.4. This improvement effect can be seen in the fewer target zones with high absolute errors when applying the best performing methods compared to those from TDW in both 1990 and 2000 and across all the tested demographic attributes. The same patterns are observed in Figures 5.9 and 5.10 that depict normalized absolute errors, with the advantage that error estimates can be compared over time and across regions. According to Figures 5.9 and 5.10, the normalized errors for most demographic attributes are less than 9% and 4% in 1990 and 2000, respectively.

Figures 5.4, 5.6 and 5.8 illustrate the visual agreement between the choropleth maps of the estimated demographic attributes using the best-performing methods at the target zone level in 1990 and 2000, those from TDW, and those from block aggregation. The maps can uniquely demonstrate how each selected demographic attribute has evolved within fine-resolution consistent units over 10 and 20 years, respectively, when combined with the actual choropleth maps of census tracts in 2010. This time series of the demographic evolution represents the major outcome of this multi-temporal estimation framework. The application of alternative demographic attributes and their validation provide some insights of the value of applying this framework to other attributes at the tract level that are not available at the block level, thereby extending the application of this research to an extensive range of demographic sub-populations.

5.5.2. Multi-temporal estimates of urban population

While the problem of modeling urban population in and of itself is an extremely complex one, the majority of the conclusions drawn for the other demographic attributes above, can also be applied to the urban population estimation in 1990 and 2000 within target tract boundaries in 2010, according to Tables 5.5 and 5.6 and Figures 5.11 and 5.12. However, there exist additional findings

specific to the urban population estimation that deserve some attention, which also provide important insights into the more general conceptual framing of the term urban and the understanding of urban population and urban lands, pointing to open research avenues in urban geography and demography.

Refinement using the census-defined urban areas in 1990, 2000 and 2010 does not reduce the absolute error measures in comparison to TDW, as indicated by Tables 5.5 and 5.6. This outcome confirms that these areas do not represent the underlying statistical surface of the urban population reliably and justifies the need for further spatial refinement. The U.S. Census employs all land-use types that have urban characteristics, such as commercial and industrial districts, for delineating urban lands whereas the urban population is a subset of the total residential population. Due to this dichotomy between the two concepts, the census-defined urban areas overestimate the spatial distribution of the urban population.

It should be noted that EM cannot be applied to the census-defined urban areas because there are no categories inside these areas that could be related to varying urban population density weights in a reasonable way.

According to Tables 5.5 and 5.6, when the census-defined urban areas are further refined by an ancillary variable (i.e., a combined spatial refinement), the associated error measures are generally lower compared to using the urban areas alone. This indicates that the further refinement causes the modified urban areas to more precisely represent the spatial distribution of the urban population as a subset of the total residential population. For example, the refined TDW using the urban areas and ZTRAX® combined for spatial refinement shows reductions in MAE by 49% and 47%, compared to utilizing only the urban areas for 1990-2010 and 2000-2010, respectively. However, even if no representations of urban lands existed, the employment of ancillary variables

alone would have the ability to depict those areas where the urban population resides, reliably. This is indicated by the lower absolute error measures in the bottom parts of Tables 5.5 and 5.6 in comparison to the regular methods. This can be considered a promising initial research effort to define temporally consistent urban extents and population distributions using data-driven approaches, which potentially will result in a more uniform understanding of how urbanization has evolved at fine spatial resolution and over long periods of time.

Different combinations of the developed classes of NLCD were also tested. The combinations that result in the highest level of accuracy are the classes 21, 22 and 23 from the NLCD 1992, and 22, 23 and 24 from the NLCDs 2001 and 2011. These observations suggest that excluding the low density class of 21 from the recent NLCDs results in more precise delineations of the urban lands in the years 2000 and 2010.

According to Tables 5.5 and 5.6, the absolute error measures for estimating urban population using TDW and EM refined by parcels is exceedingly high. For delineating urban lands, additional land-use types are included, which in turn aggravates the overestimation issue when using parcels. However, EEM manages to mitigate the issue.

Figure 5.11 illustrates the gradual absolute error reductions within target zones in 1990 and 2000, matching visually the outcomes shown in Tables 5.5 and 5.6. Not only does Figure 5.12 illustrate how the derived and block-aggregated urban population maps compare to each other, it also depicts the most reliable trend of urbanization from 1990 to 2000 at the target zone level, which can also be accompanied by the actual urban population map in 2010 to portray a 20-year trend of urbanization over consistent units in a unique way. Such trends and multi-temporal patterns will be of great use to better understand processes such as urban sprawl.

5.5.3. Multi-temporal estimation of exposed population groups to flood risk

Tables 5.7 and 5.8 show that the spatially refined methods leveraging the fine resolution ancillary variables such as buildings and ZTRAX® result in the highest accuracy when estimating the potentially exposed sub-populations (broken down by race and age classes) in different years. Using parcels for spatial refinement clearly overestimates the exposed population counts, presumably ensuing from the less precise delineation of residential lands by the ancillary variable in rural settings. The issue is attenuated by utilizing EEM although the overestimation problem still persists. In this study, refined AW using buildings generally leads to the lowest deviances with the benchmark estimates (i.e., block-based populations deriving from EEM using buildings).

With the exception of the population group of age > 65, there is no indication of unexpectedly high concentrations of certain population subgroups within flood zones in 1990 and 2000, as shown in Figure 5.13. The ratios of the observed counts – using the benchmark estimates and the outcomes of the best-performing areal interpolation method – to the expected values (based on the state-level distributions) are lower than 1 in both years and for almost all the demographic attributes. For the age group > 65, however, the observed count is 8% higher than expected, suggesting that the group is exposed to the flood hazard, disproportionately, possibly increasing its vulnerability.

The results imply that there is no issue of the environmental injustice regarding race, i.e., both white and black populations are distributed proportionately within flood zones. However, in order to reach a concrete conclusion, more contextual factors should be considered such as a chronology of the historic patterns of residential settlement and segregation, different levels of industrial development along the waterfront, recent and historic landfilling of coastal wetland areas, gentrification in certain areas of the state, and cultural changes over the years concerning

the desirability of living along the waterfront and therefore within flood zones (Maantay and Maroko 2009). Furthermore, diversity in Massachusetts is at low levels compared to other states, and other groups are not considered in this study. Future research will target other states to tackle these questions in more detail.

It should be noted that ratios around 1 assume that distributions are proportionally equal to the whole state. Of course, it could be argued that among those whose ratios are lower than 1, some might be more advantaged than the others depending on how far their ratios are lower than 1. A similar order for disadvantaged groups can also be established.

5.5.4. Final general remarks and conclusions

This study shows how spatial refinement can improve the accuracy of regular areal interpolation methods in constructing time-series of demographic estimates within consistent small area census units. By incorporating demographic attributes other than total population, micro-scale patterns of demographic change can be detected. The temporal analyses do not depend on census blocks and utilize them for validation purposes. Thus, the analyses can be replicated for applications involving longer time periods, where the availability of blocks is limited, or for data-poor regions. They can also estimate more sensitive demographic attributes only aggregated over tracts with expectations of the inherent estimation error.

This research also shows that the difference in performance between regular and refined methods on one hand, and methods refined by publically available ancillary variables and those by local variables, on the other hand, is more pronounced when population counts are high or long time periods are incorporated. If repeated over more heterogeneous demographic attributes or longer time periods, an experimental rule may be established about the necessity of spatial refinement and the types of ancillary variables that have the most potential.

Moreover, the case study of applying enhanced areal interpolation methods to evaluate the exposure status of race- and age-related population sub-groups residing in flood risk zones illustrates the power of this analytical framework in a risk assessment context. The outcomes show that if larger enumeration units such as census tracts (instead of the finest-resolution blocks) are incorporated, the reliability of population estimates is still comparable to block-based results. This demonstrates the potential to expand the applicability of similar risk assessment analyses to applications where the availability of fine-resolution enumeration units is limited, or earlier census years in the Unites States, in which block-based population values were not collected.

Future research can be pursued in different directions. First, the extent of the analyses can be expanded both temporally and spatially to attain robust and conclusive insights about the applicability and performance of various methods. Second, the environmental justice analysis can be applied to more racially and ethnically diverse study areas such as the southern states of the United States to obtain a more comprehensive basis of substantive interpretations. Finally, the success of the refinement of the census-defined urban areas in better representing the urban land shows the great potential of leveraging data-driven approaches to establish consistent definitions of the urban land and population through time.

Acknowledgements

This research is funded by the National Science Foundation: “Collaborative Research: Putting People in Their Place: Constructing a Geography for Census Microdata”, Project BCS-0961598 awarded to University of Colorado – Boulder.

The work is funded, in part, by the National Science Foundation award #1416860 to the City University of New York, the Population Council, the National Center for Atmospheric Research and the University of Colorado. Innovative Seed Grant funding from the University of Colorado as well as a development grant received from its Population Center (CUPC) at the Institute of Behavioral Science, are acknowledged.

The ZTRAX® dataset is provided thanks to the collaboration with Zillow Inc. (https://www.zillow.com/research/data/) based on a Data Access Agreement between the Regents of the University of Colorado and Zillow Inc. The source code for processing the data is available at https://github.com/spatialhistory/ztrax_analysis.

Chapter VI

Discussion and Conclusions

6.1. Discussion

This dissertation has contributed solutions to the persistent problem of spatial inconsistency between small area enumeration units used for different census surveys. This inconsistency results from existing linkages between census demography and census geography maintained at small area units; i.e., boundaries of these units get updated in each census year to reflect population changes that have occured. This is one of the most persistent problems in the temporal analysis of local and regional trends in demographic, housing, economic and health-related characteristics (Gregory 2002, Martin *et al.* 2002, Schroeder 2007, 2017). This dissertation has shown the effectiveness of spatial refinement in improving the accuracy involved in multi-temporal demographic analysis. For validation, census blocks in each year were aggregated to target zone boundaries and formed benchmark (model) values. Notably, however, the validation approach has some drawbacks. First, although census blocks are generally very small, they can occasionally be very large, especially in rural areas, making the assumption of the homogeneous population distribution within them problematic. One solution to this issue is to apply spatial refinement to blocks prior to validation. However, as reported by previous researchers, validation

results with and without spatial refinement on blocks are almost similar (Ruther *et al.* 2015), and for this reason, this approach was not followed in this dissertation. Second, the demographic variables used in this research are supposed to represent 100% of counts. However, this might not always be the case given the costly process of collecting census data and proportions of non-responsiveness. Because the objective was to do a cross comparison between the performances of different methods in each year, this issue was not seen as a problem as long as benchmark values were the same for all methods.

Accuracy gains by combining typical areal interpolation methods with spatial (or dasymetric) refinement were analyzed for a plethora of ancillary variables (Chapters 3, 4 and 5), different geographic extents (Chapters 3, 4 and 5), two time periods (Chapters 4 and 5), urban and rural settings (Chapter 4), and different demographic attributes (Chapter 5). The three scientific papers (Chapters 3, 4 and 5) presented the different stages of the development and application of the enhanced areal interpolation methodologies for creating temporally consistent population estimates at the census tract level with decreased estimation errors. Chapter five also investigated a first effort in establishing associations between demographic and environmental data to carry out improved potential exposure assessments by identifying populations residing in designated flood zones in different points in time that could potentially face environmental injustice. Overall, the findings of this research affirm the effectiveness of the idea of systematically redistributing population within inhabitable areas to then improve typical areal interpolation methods for temporal analysis. The approach mitigates the issue of the Modifiable Areal Unit Problem (MAUP) (Openshaw 1984) often observed in spatial analysis using arbitrarily defined areal units that contain inhabitable and uninhabited areas and thus non-homogeneous spatial distributions of population. In general, the approach provides more flexibility and an analytical solution to

researchers in different disciplines working on summary statistics enumerated over fixed geographical units that may change over time. This includes researchers who work with databases on public health, well-being, veterinary topics, agricultural censuses, employment statistics or economic surveys at similar levels of demographic data collections. Thus, the solutions proposed in this dissertation have the potential to advance the thematic research of different natures and across disciplines.

In this chapter, the research questions posed in Chapter one are systematically revisited to discuss the results found in the dissertation and evaluate the methodological frameworks developed in this project. Conclusions for the dissertation summarizing its general findings are then presented, followed by future works.

Research question 1*: How effectively can different types of ancillary variables be used for spatial refinement to systematically improve the accuracy of regular areal interpolation methods? What lessons can be learned about the applicability of different ancillary variables in different geographic and demographic settings?*

In this dissertation, the Global Human Settlement Layer (GHSL), National Land-Cover Database (NLCD), parcels, buildings, road-related variables, ZTRAX® and various combinations of these data were tested as ancillary variables for the spatial refinement of the input enumerated data that are then integrated with areal interpolation to produce spatially consistent units with reduced estimation errors. It should be mentioned that each ancillary variable is a model of reality, i.e., the spatial distribution of human settlement. Depending on how well it models reality, its integration into spatial refinement can affect the effectiveness of the approach. For example, a remote-sensing based product such as NLCD might misclassify large proportions of residential or developed lands, especially in rural areas (Leyk *et al.* 2014). Or, GHSL, as another remote-sensing

based ancillary variable, assumes that built-up areas in each epoch are the summation of built-up areas from the previous epochs, disregarding developed patches that have been torn down or changed to other land-cover types. It also does not distinguish between residential and non-residential land-use types (Pesaresi *et al.* 2016). Such imperfections need to be addressed in the interpretation of results to ensure that the reader understands that each additional ancillary variable might also introduce some additional uncertainty, contributing to the complex composition of error sources that will be propagated through the system.

Chapter three applied residential parcels and a selection of the NLCD developed classes to estimate total population values in 2000 at the target census tract level (i.e., Census 2010) in Hennepin County, Minnesota. The results showed that although the effectiveness of spatial refinement could be limited within some target zones due to the quality of the ancillary variable (Langford 2013), spatial refinement using each ancillary variable generally leads to more accurate multi-temporal total population estimates within target units. The sporadic higher error measures of the tested refined methods compared to the unrefined ones may be related to the existence of a few target zones with high absolute errors in the right tail of the error distribution. The spatial refinement using parcels performed better than using NLCD although this superiority is not consistent within all target zones and across all the utilized methods.

Chapter four applied parcels and a composite ancillary dataset of parcels, NLCD and road buffers to estimate total population counts in 1990 and 2000 at the consistent target census tract level (i.e., Census 2010) in five demographically different counties. The results re-confirmed the effectiveness of using spatial refinement to reduce error measures in multi-temporal estimations of population. The spatial refinement using parcels mostly reduced the absolute error measures compared to unrefined implementations of the tested areal interpolation methods. However, the

overestimation of developed lands through large residential parcels in rural settings (Leyk *et al.* 2014) that in turn invalidated the underlying assumptions of the utilized methods resulted in a few target zones with excessive absolute errors. This caused rare instances characterized by higher absolute error measures for a spatially refined method. The spatial refinement using the composite dataset reduced the error measures when compared to using only parcels. However, the improvement effect was not always present, especially in more urbanized counties such as Broward County, Florida. The composite spatial refinement was specifically devised for addressing the issue of the overestimation of developed lands by parcels in rural areas, and its effectiveness will be evaluated when discussing findings for research question three.

Chapter five demonstrated that spatial refinement was effective in increasing the accuracy of multi-temporal demographic estimations for the whole state of Massachusetts and for both 1990-2010 and 2000-2010, especially when fine resolution and precise ancillary variables such as building footprints and ZTRAX® were used. The global GHSL and national NLCD as well as local parcels performed generally well in reducing the absolute error measures. However, that effectiveness varied according to the quality of the dataset and the method used. For example, the cases in which GHSL and NLCD did not enhance the accuracy of the unrefined Target Density Weighting (TDW) can be related to the inherent misclassification problems of the datasets (NLCD (Smith *et al.* 2002, Leyk *et al.* 2014), GHSL (Leyk *et al.* n.d.)) and their inability to capture small patches of developed lands. Furthermore, the reason why Expectation Maximization (EM) using parcels produced large error measures could be explained by the excessive area variability existing within units of the same class in the dataset. These issues were mitigated through the incorporation of the fine resolution buildings and ZTRAX® datasets, especially in refining TDW, by tapping into spatial refinement more effectively. The degree of the effectiveness of the different ancillary

variables over the two time periods and for the different tested demographic attributes will be described while discussing research question five below.

Overall, the findings of the research question one show that spatial refinement increases the accuracy of areal interpolation methods for multi-temporal demographic estimations. However, the application domain is not necessarily limited to multi-temporal settings and can also include more accurate transfers of various enumerated attributes such as economic or health-related variables from different source zones to target zones specific to the application.

Research question 2*: How can existing approaches of dasymetric refinements in multi-temporal demographic analysis be extended to also incorporate land-use related attributes of ancillary variables in addition to their geometric footprints? What is the gain in accuracy from such an extended integration?*

EM (Dempster *et al.* 1977, Flowerdew and Green 1994, Schroeder and Van Riper 2013) was the algorithmic foundation employed to incorporate relevant datasets as related ancillary variables (Leyk, Buttenfield, *et al.* 2013) by exploiting their different land-cover/land-use characteristics. The algorithm assigns a distinct population density weight to each land-cover/land-use class (control zone) in an iterative process and thus overcomes the use of limiting ancillary variables that only allow the differentiation between populated and unpopulated lands.

The results reported in Chapter four unexpectedly showed that using EM on parcels did not produce highly accurate results. The method produced either the highest or the second highest absolute error measures among the tested methods including refined Areal Weighting (AW) and refined TDW for the two time periods and across all the selected counties. This low level of accuracy could be explained by the high variability in the areal extents of underlying parcels of the same land-use type constituting individual control zones. This variability eroded the

fundamental assumption of constant population density values within control zones and resulted in unrealistic estimation results. EEM was implemented to make EM more robust by identifying more homogeneous sub-control zones. The method resulted in substantial accuracy gains over EM in all cases. EEM was the most accurate method for multi-temporal population estimations in four of the five counties tested for the longer time period. It also provided the most accurate results in two counties for the shorter time period. The remarkable performance of EEM demonstrated the effectiveness and applicability of incorporating additional attributes of ancillary variables especially over longer time periods.

Chapter five also conducted EM and EEM on relevant ancillary variables. The results demonstrated that EM, as expected, performed relatively poorly, especially using NLCD and parcels due to their misclassification issues and extensive area variations, respectively. EEM was only applicable for parcels and buildings. The outcomes presented in Chapter five showed that EEM always improved the accuracy of EM using parcels. However, EEM was not effective when employed on buildings because of the much lower and less significant area variability existing within the dataset (i.e., house footage vs. property acreage). The observed trends implied that the inclusion of additional attributes of ancillary variables, especially parcels, in the form of EEM, could outperform refined TDW over longer historical time periods.

Refined TDW is a relatively accurate spatially refined areal interpolation method, specifically tailored for multi-temporal applications. However, the findings related to research question two suggest that whenever the ancillary variable can be integrated as a related variable to incorporate nuanced population densities across control zones, algorithmic foundations such as EM and EEM should be tested. This is especially the case over longer time periods when the underlying assumptions of refined TDW become less reliable.

Research question 3*: How can the effectiveness of dasymetric refinement be improved in rural areas typically associated with lower accuracy due to the lack of ancillary variables or the chronic under- or overestimation of population in such settings?*

Prior research has shown that dasymetric models based on land-cover data yield more accurate results if they are further restricted to areas within road buffers (Lin *et al.* 2013, Lin and Cromley 2015b, Schroeder 2017). Chapter four adopted a similar approach to specifically increase the estimation accuracy in rural target zones. Typically, such rural areas are associated with high estimation errors for several reasons; e.g., the under- and overestimation of developed lands by land-cover products such as NLCD and parcels, respectively (Leyk *et al.* 2014). The proposed approach further refined large rural parcels by using instances of the selected NLCD developed classes and designated road buffers. This new complementary spatial refinement was conducted for the five study areas in Chapter four and tested for both all target zones and only rural target zones. The outcomes showed that the error reduction due to the introduction of the complementary spatial refinement was more consistent over rural zones, indicating the effectiveness of the approach to address excessive error measures typically observed in these settings. Therefore, the composite refinement can be considered a viable solution for the long-lasting issue of high demographic estimation errors in rural areas.

Research question 4*: How stable is the effect of spatial refinement across different geographic scales and in estimating additional demographic attributes other than total population? Does this framework allow for coupling between demographic and environmental data to analyze more complex relationships?*

Spatial refinement generally increased the accuracy of multi-temporal demographic estimates within the selected counties and states, as shown in Chapters three, four and five. The

performance of spatially refined areal interpolation methods on selected demographic attributes, however, varied depending on the attribute of interest, as indicated in Chapter five. While the levels of accuracy gain in estimating counts of total population, white population, population aged under 65 and urban population remained almost similar, accuracy improvements could not be observed for the temporal estimation of black population counts. This could be related to the initially low counts of the sub-group in large parts of the state. It should also be noted that the accuracy of the temporal estimation of the age attribute can also be affected by exogenous processes such as the emergence of baby boomers as a relatively large demographic cohort. For example, if the time period in future applications includes 2020, the population aged above 65 will have excessive and disproportionate counts compared to earlier census years, thereby affecting the temporal accuracy of an areal interpolation method such as TDW that is based on constant ratios of population densities over time. This and similar issues are subject to future research efforts.

Moreover, the performance of individual ancillary variables could vary based on the attribute of interest. For example, refined TDW using parcels produced excessively high absolute error measures for estimating urban population counts compared to the other ancillary variables. This was because the selection of parcels for the temporal estimation of urban population was not limited to residential land-use types. Instead, all types that could presumably indicate developed lands were selected. This seemingly intensified the overestimation issue of parcels, culminating in a poor performance by the ancillary variable for multi-temporal estimations of urban population.

Chapter five provided important insights about the effectiveness of using census-defined urban areas to better model urban population. Urban population is a complex demographic attribute because of a well-known high degree of inconsistency in its definition over time. In other words, in addition to the temporal incompatibility in the small-area enumeration units, the concepts

of urban population and lands differ over time. Nonetheless, given the growing trends of urbanization, research efforts to model urban lands and population reliably represent an extremely urgent need in different domains including planning, policy-making and resource allocation. According to the results of Chapter five, unexpectedly, the census-defined urban areas of Massachusetts in 1990, 2000 and 2010 do not reflect the spatial distribution of urban population reliably. The U.S. Census incorporates all land-use types that can be urban such as commercial and industrial districts for delineating urban lands whereas urban population is a subset of the total residential population. Therefore, there is a persistent incompatibility between these two concepts, often resulting in an overestimation of the spatial distribution of urban population by the census-defined urban areas. For this reason, these areas were further refined by using additional ancillary variables to define a new set of urban lands that represent the spatial distribution of urban population presumably more precisely. The further refinement showed great potential to delineate urban residential lands reliably. The individual ancillary variables such as building footprints and ZTRAX® also demonstrated valuable potential to be used as a surrogate for urban land representations. These findings can be seen as successful initial experiments to advance existing research on urban analysis and will be used for future data-driven experiments aiming to create consistent and precise delineations of urban lands and urban population over time. It is hoped that these insights will help to model the evolution of urban populations and urban lands in a more unified and semantically-compatible way.

Chapter five also investigated the effectiveness of spatial refinement in an environmental injustice application. The counts of the potentially exposed population broken by racial and age categories living in flood zones of Massachusetts were derived for 1990 and 2000. Moreover, experiments were conducted to evaluate if certain population sub-groups were disadvantaged in

each year. The estimated counts demonstrated the potential of the proposed methodologies in transferring different demographic attributes aggregated within census units to independently mapped flood zones. The spatial refinement using buildings and ZTRAX® mimicked the validation benchmark estimates reasonably well. This benchmark was calculated by applying spatial refinement to block-based population counts. These results were also compared to the analysis using only parcels and to typical unrefined AW, which overestimated the exposed populations substantially. The described outcomes from this tract-based analysis demonstrate the applicability and flexibility of the approach to generate more reliable estimates of the exposed population in earlier census years to attain valuable historical patterns of possible environmental injustice issues, and how they changed.

One of the main objectives of this type of analysis is to evaluate social or environmental injustice through the evaluation of potential exposure levels of different populations (e.g., racial groups) to natural or industrial hazards. This has broad and outreach applications for other fields such as sociology aiming to assess if different minority groups systematically have lower levels of quality of life. Although Massachusetts is not an ideal case study given the existing racial homogeneity, the high availability of data in the state drove the efforts to develop the required analytical framework that can be later implemented in other states with high racial diversity.

Research question 5*: By making the methodological frameworks operational for multiple demographic attributes and over different time-periods, what expectations exist related to the uncertainty inherent to the population estimates produced?*

Chapters four and five carried out spatially refined methods for two time periods (1990-2010 and 2000-2010), and Chapter five also implemented the methods on other demographic attributes or population sub-groups. These analyses provided insights about the uncertainty sources

influencing the performances of the methods under varying conditions.

Chapter four demonstrated that refined TDW was more effective in error reduction compared to unrefined TDW over the longer 1990-2010 time period in all the five counties. The ratios of the overall error measures, namely Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), of refined TDW to TDW were generally lower for the longer time period than for the shorter. This signified that spatial refinement for TDW was more effective in error reduction when applied over longer time periods. The pattern was not as dominant for refined AW, and that comparison could not be made for EM because the algorithm inherently uses spatial refinement. Chapter 5 also showed the same pattern for estimates of total population, white population, population aged under 65 and urban population; i.e., the ratios of the overall error measures of the best-performing refined TDW to TDW were lower for 1990-2010 than for 2000-2010. This pattern, shown in Chapters four and five, indicates that spatial refinement can be more effective when the uncertainty of areal interpolation is higher due to more frequent boundary changes (Schroeder 2007).

Although refined TDW was the most accurate method in a majority of the configurations tested in this dissertation, the findings and observed patterns in Chapters four and five indicated that EEM could outperform refined TDW and rank as the most accurate method with the extension of the time period. These findings and patterns suggest that the underlying assumptions of refined TDW tend to become less valid over longer time periods, thereby reducing the reliability and robustness of the method for the temporal interpolation of population.

Chapter five also presented the effectiveness of spatial refinement for interpolating black population, a demographic attribute with typically low population counts within census units in Massachusetts. As shown in Chapter five, spatial refinement regardless of the utilized ancillary

variable, did not perform well in error reduction for the demographic attribute, implying that the effectiveness of spatial refinement could diminish when the uncertainty of areal interpolation is low due to small population counts of the demographic attribute (Schroeder 2007).

It should also be noted that the comparisons made between the methods of this dissertation were according to absolute error measures. On the one hand, there might be target zones with high population counts whose absolute errors are also high. On the other hand, target zones might also exist with low population counts whose absolute errors are low. However, the ratios of absolute errors to population counts could be lower for the former. Normalizing absolute errors by population counts will address this issue and provide a more comprehensive picture of the error behavior. Although this dissertation does not include normalized errors, they were calculated, and method comparisons based on them led to similar interpretations.

The findings of research question five provide insights about the effectiveness of spatial refinement. These findings are valuable because spatial refinement is a costly process, both in terms of time and data demands. Therefore, it should be incorporated in areal interpolation only when it can be effective. According to the current findings, spatial refinement has its greatest potential when the application includes long historical time periods, and the demographic attribute has grown rapidly with high counts within enumeration units.

6.2. Conclusions and future work

Summary data enumerated over small area units from national censuses are a vital resource for studies of local and regional demographic trends. However, the linkages between census demography and census geography necessitate that boundaries of small statistical units change over time to reflect underlying demographic processes such as population growth and decline. Thus, studying local and regional demographic trends using summary data is frequently

complicated by temporal boundary changes in census small area units. This dissertation aimed to derive demographic estimates within temporally consistent small area census units with minimum estimation errors. The successful estimation of such small area estimates allows more advanced studies of micro-scale demographic processes, and how they may interrelate and interact with environmental ones.

The outcomes of this study demonstrate that spatial refinement coupled with typical areal interpolation methods is generally effective in reliably estimating demographic attributes within temporally consistent small area census units. The effectiveness of spatially-refined areal interpolation methods was tested for two time periods, various ancillary variables, geographic scales extending from county to state and different target demographic attributes. Overall, the substantial estimation accuracy gains offered by implementing spatial refinement in areal interpolation were confirmed across these different configurations. Figure 6.1 demonstrates an example of how total population counts enumerated over initially misaligned census units from three census years (1990, 2000 and 2010) compare to its estimates over consistent units, in a sub-region of Mecklenburg County, North Carolina. As shown in Figure 6.1, the analysis of nuanced micro-scale changes of population based on census tracts is complicated by the inconsistency of boundaries among the three census years (left panels). However, a spatially refined areal interpolation method (EEM on parcels) makes such analysis possible by transferring total population counts of the first two years to the fine-resolution geographical units, namely census tract boundaries in 2010 (right panels). In fact, not only are the changes in census geography addressed, precise and fine-scale patterns in census demography can also be presented without bias.

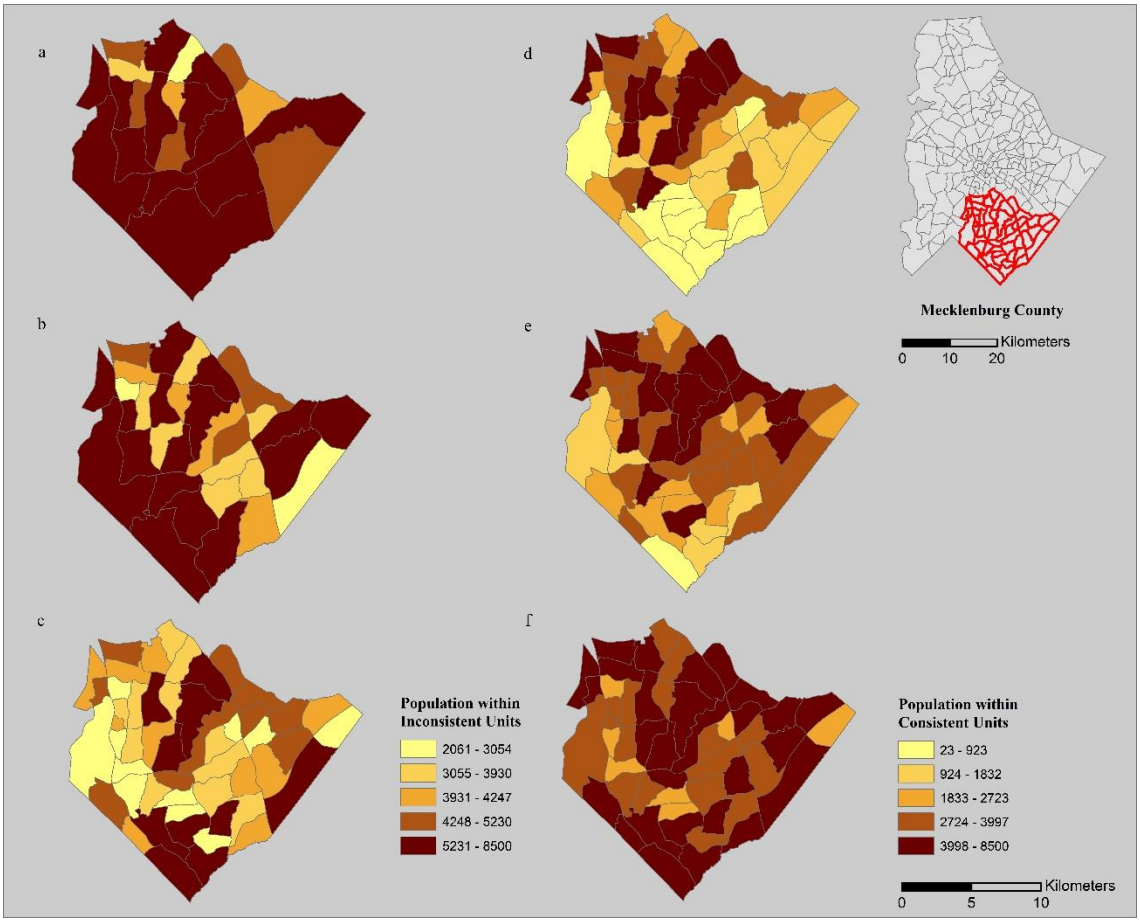

Figure 6.1. Population maps within inconsistent census tracts: (a) in 1990, (b) in 2000, (c) in 2010, and within consistent target census tracts: (d) in 1990, (e) in 2000 and (f) in 2010.

In this dissertation, source and target zones were tract boundaries from the Decennial Censuses 1990, 2000, and 2010. For most of the analysis, the tracts in 2010 were used as target zones as they tend to be the most refined units. Although census blocks are the smallest enumeration units, they were used for validation in this research. Besides the obvious reason of being able to carry out the validation per se, this strategy also allows for using the methodological advancements of this dissertation in numerous other applications. First, the temporal dimension of historical demographic data sources at the tract level in the United States can extend to earlier census years (i.e., prior to 1990), in which nationally available census block data were not yet available. Second, additional demographic attributes collected over tracts but not blocks can be

incorporated into the analysis by assuming that derived accuracy measures and discussed uncertainty insights are also applicable to those attributes. It is important to note that the census tract level carries numerous demographic attributes and thus provides a broad base of applicability to different population sub-groups. Third, the findings of this research, especially those related to using GHSL and NLCD, can also be carried over to data-poor regions, where local ancillary variables are not available. Nevertheless, the proposed methods in such regions should be used with caution. The quality of large-scale global and national datasets changes over time following advancements in data acquisition, algorithms and related technologies. As long as a global or national ancillary variable is not subject to systematic quality issues such as misclassification or under-estimation, especially in earlier years, it has the potential to be implemented for more accurate temporal estimation of demographic attributes in data-poor regions. Another issue in data-poor regions needing attention is the typical lack of fine-resolution census geographies such as census blocks for validation. One solution to this issue is to aggregate the finest available enumeration units to larger arbitrarily defined boundaries and use the initial units for validation. However, this analytical approach can be subject to MAUP (Openshaw 1984). Another solution is to assume that accuracy is improved by spatial refinement and to use the smallest units as source and target zones. Fourth, the application domain of the proposed methodologies can extend beyond temporal demographic analysis and evaluate more sophisticated human environment interactions such as exposure to hazards, either in the past (Maantay and Maroko 2009) or in the future (Jones and O'Neill 2016).

Future research will implement the methodologies of this dissertation to longer time periods including projections forward, more study areas and unexplored demographic attributes. Not only will such analyses lead to a comprehensive and quantitative uncertainty evaluation, it

will also provide the foundations necessary to advance the current methodologies further and devise hybrid frameworks that exploit the complementary features of individual methods given different conditions and settings. For example, the higher performance of the composite approach particularly in rural areas or different levels of accuracy offered by refined TDW and EEM indicate the potential of devising complementary hybrid approaches that are versatile for complex applications. Moreover, the ZTRAX® data used in this dissertation showed great potential to accurately refine areal interpolation methods. The dataset is semantically more consistent than raw parcel records across the country although the issue of missing data is severe in some parts. Nevertheless, it offers a unique opportunity to apply the methods of this dissertation to the whole United States and over long time periods. The algorithmic and technological advancements required for this national-scale analysis will constitute another direction for the future research. Finally, the outcomes of this dissertation showed that census-defined urban areas could be refined and redefined using dasymetrically refined approaches to better (or more consistently) reflect urban lands and population. The development of a data-driven approach based on the current analyses of this dissertation that can consistently and reliably delineate urban lands and population over time will be another topic for the future research.

BIBLIOGRAPHY

Alahmadi, M., Atkinson, P., and Martin, D., 2015. Fine spatial resolution residential land-use data for small-area population mapping: a case study in Riyadh, Saudi Arabia. *International Journal of Remote Sensing*, 36 (17), 4315–4331.

Anselin, L., 2002. Under the hood issues in the specification and interpretation of spatial regression models. *Agricultural economics*, 27 (3), 247–267.

Bakillah, M., Liang, S., Mobasheri, A., Jokar Arsanjani, J., and Zipf, A., 2014. Fine-resolution population mapping using OpenStreetMap points-of-interest. *International Journal of Geographical Information Science*, 1–24.

Balk, D. and Yetman, G., 2004. The global distribution of population: evaluating the gains in resolution refinement. *In*: *Center for International Earth Science Information Network (CIESIN)*. New York: Columbia University.

Barufi, A.M., Haddad, E., and Paez, A., 2012. Infant mortality in Brazil, 1980-2000: A spatial panel data analysis. *BMC public health*, 12 (1), 181–195.

Bhaduri, B., Bright, E., Coleman, P., and Urban, M., 2007. LandScan USA: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics. *GeoJournal*, 69 (1–2), 103–117.

Bian, R. and Wilmot, C.G., 2017. Measuring the vulnerability of disadvantaged populations during hurricane evacuation. *Natural Hazards*, 85 (2), 691–707.

Birkin, M. and Clarke, M., 2011. Spatial microsimulation models: A review and a glimpse into the future. *In*: J. Stillwell and M. Clarke, eds. *Population dynamics and projection methods*. Dordrecht: Springer, 193–208.

Blaschke, T., 2010. Object based image analysis for remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*.

BostonGIS, 2016. PARCELS 2016 DATA FULL [online]. Available from: http://bostonopendata-boston.opendata.arcgis.com/datasets/f3d274161b4a47aa9acf48d0d04cd5d4_0 [Accessed 10 May 2017].

Burton, C.G., 2010. Social Vulnerability and Hurricane Impact Modeling. *Natural Hazards Review*, 11 (2), 58–68.

Buttenfield, B.P., Ruther, M., and Leyk, S., 2015. Exploring the impact of dasymetric refinement on spatiotemporal small area estimates. *Cartography and Geographic Information Science*, 42 (5), 449–459.

Calka, B., Bielecka, E., and Zdunkiewicz, K., 2016. Redistribution population data across a regular spatial grid according to buildings characteristics. *Geodesy and Cartography*, 65 (2), 149–162.

Cromley, E.K. and McLafferty, S.L., 2002. *GIS and public health*. New York: Guilford Press.

Cromley, R.G., Hanink, D.M., and Bentley, G.C., 2012. A Quantile Regression Approach to Areal Interpolation. *Annals of the Association of American Geographers*, 102 (4), 763–777.

Cutter, S.L., Emrich, C.T., Webb, J.J., and Morath, D., 2009. Social vulnerability to climate variability hazards: A review of the literature. *Final Report to Oxfam America*, 5, 1–44.

Dempster, A., Laird, N., and Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39 (1), 1–38.

Department of Commerce, 2011. Federal Register. *Federal Register*, 76 (164), 53030–53043.

Dmowska, A. and Stepinski, T.F., 2017. A high resolution population grid for the conterminous United States: The 2010 edition. *Computers, Environment and Urban Systems*, 61, 13–23.

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

Dong, P., Ramesh, S., and Nepali, A., 2010. Evaluation of small-area population estimation using LiDAR, Landsat TM and parcel data. *International Journal of Remote Sensing*, 31 (21), 5571–5586.

Eicher, C.L. and Brewer, C.A., 2001. Dasymetric mapping and areal interpolation:

Implementation and evaluation. *Cartography and Geographic Information Science*, 28 (2), 125–138.

Exeter, D., Boyle, P.J., Feng, Z., Flowerdew, R., and Scheirloh, N., 2005. The creation of 'consistent areas through time' (CATTs) in Scotland, 1981-2001. *Population Trends*, 119, 28–36.

Fisher, P. and Langford, M., 1995. Modelling the errors in areal interpolation between zonal systems by Monte Carlo simulation. *Environment and Planning A*, 27 (2), 211–224.

Flowerdew, R. and Green, M., 1994. Areal interpolation and types of data. *In*: S. Fotheringham and P. Rogerson, eds. *Spatial analysis and GIS*. London: Taylor & Francis, 121–145.

Flowerdew, R., Green, M., and Kehris, E., 1991. Using areal interpolation methods in geographic information systems. *Papers in Regional Science*, 70 (3), 303–315.

Fothergill, a, Maestas, E.G., and Darlington, J.D., 1999. Race, ethnicity and disasters in the United States: a review of the literature. *Disasters*, 23 (2), 156–173.

Fotheringham, A.S., Brunsdon, C., and Charlton, M., 1992. *Geographically weighted regression: the analysis of spatially varying relationships*. Chichester: John Wiley & Sons.

Geiß, C., Schauß, A., Riedlinger, T., Dech, S., Zelaya, C., Guzmán, N., Hube, M.A., Arsanjani, J.J., and Taubenböck, H., 2016. Joint use of remote sensing data and volunteered geographic information for exposure estimation: evidence from Valparaíso, Chile. *Natural Hazards*, 1–25.

Geiß, C. and Taubenböck, H., 2013. Remote sensing contributing to assess earthquake risk: from a literature review towards a roadmap. *Natural Hazards*, 68 (1), 7–48.

Gelfand, A., Zhu, L., and Carlin, B., 2001. On the change of support problem for spatio-temporal data. *Biostatistics*, 2 (1), 31–45.

Giordano, A. and Cheever, L., 2010. Using dasymetric mapping to identify communities at risk from hazardous waste generation in San Antonio, Texas. *Urban Geography*, 31 (5), 623–647.

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

Goodchild, M., Kyriakidis, P., Schneider, P., and Sifuentes, J., 2005. Uncertainty and interoperability: the areal interpolation problem. *In*: *Fourth International Symposium on Spatial Data Quality (ISSDQ 05)*. Beijing.

Goodchild, M. and Lam, N.S.N., 1980. Areal interpolation: a variant of the traditional spatial problem. *Geo-processing*, 1, 297–312.

Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69 (4), 211–221.

Gregory, I.N., 2002. The accuracy of areal interpolation techniques: standardising 19th and 20th century census data to allow long-term comparisons. *Computers, Environment and Urban Systems*, 26 (4), 293–314.

Harvey, J.T., 2002. Estimating census district populations from satellite imagery: Some approaches and limitations. *International Journal of Remote Sensing*, 23 (10), 2071–2095.

Hazards and Vulnerability Research Institute, 2014. Social vulnerability index [online]. Available from: http://artsandsciences.sc.edu/geog/hvri/sovi®-0 [Accessed 7 May 2017].

Hennepin County GIS, 2016. Hennepin County Geographic information systems [online]. *Hennepin County*. Available from: http://www.hennepin.us/gisopendata [Accessed 3 Dec 2016].

Holt, J.B., Lo, C.P., and Hodler, T.W., 2004. Dasymetric estimation of population density and areal interpolation of census data. *Cartography and Geographic Information Science*, 31 (2), 103–121.

Homer, C., Dewitz, J., Fry, J., Coan, M., Hossain, N., Larson, C., Herold, N., McKerrow, A., VanDriel, J.N., and Wickham, J., 2007. Completion of the 2001 national land cover database for the conterminous United States. *Photogrammetric Engineering and Remote Sensing*, 73 (4), 337–341.

Homer, C.G., Dewitz, J.A., Yang, L., Jin, S., Danielson, P., Xian, G., Coulston, J., Herold, N.D., Wickham, J.D., and Megown, K., 2015. Completion of the 2011 National Land Cover

Database for the conterminous United States-Representing a decade of land cover change information. *Photogrammetric Engineering and Remote Sensing*, 81 (5), 345–354.

Howenstine, E., 1993. Measuring demographic change: the split tract problem. *The professional geographer*, 45 (4), 425–430.

Jia, P. and Gaughan, A.E., 2016. Dasymetric modeling: A hybrid approach using land cover and tax parcel data for mapping population in Alachua County, Florida. *Applied Geography*, 66, 100–108.

Jia, P., Qiu, Y., and Gaughan, A.E., 2014. A fine-scale spatial population distribution on the High-resolution Gridded Population Surface and application in Alachua County, Florida. *Applied Geography*, 50, 99–107.

Jones, B. and O'Neill, B.C., 2016. Spatially explicit global population scenarios consistent with the Shared Socioeconomic Pathways. *Environmental Research Letters*, 11 (8), 84003.

Kar, B. and Hodgson, M.E., 2012. A process oriented areal interpolation technique: a coastal county example. *Cartography and Geographic Information Science*, 39 (1), 3–16.

Kim, H. and Yao, X., 2010. Pycnophylactic interpolation revisited: integration with the dasymetric-mapping method. *International Journal of Remote Sensing*, 31 (21), 5657–5671.

Krivoruchko, K., Gribov, A., and Krause, E., 2011. Multivariate areal interpolation for continuous and count data. *Procedia Environmental Sciences*, 3, 14–19.

Kyriakidis, P., 2004. A geostatistical framework for area to point spatial interpolation. *Geographical Analysis*, 36 (3), 259–289.

Lam, N.S.N., 1983. Spatial interpolation methods: a review. *The American Cartographer*, 10 (2), 129–150.

Langford, M., 2006. Obtaining population estimates in non-census reporting zones: An evaluation of the 3-class dasymetric method. *Computers, Environment and Urban Systems*, 30 (2), 161–180.

Langford, M., 2007. Rapid facilitation of dasymetric-based population interpolation by means of

raster pixel maps. *Computers, Environment and Urban Systems*, 31 (1), 19–32.

Langford, M., 2013. An evaluation of small area population estimation techniques using open access ancillary data. *Geographical Analysis*, 45 (3), 324–344.

Leyk, S., Buttenfield, B.P., Nagle, N.N., and Stum, A.K., 2013. Establishing relationships between parcel data and land cover for demographic small area estimation. *Cartography and Geographic Information Science*, 40 (4), 305–315.

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

Leyk, S., Ruther, M., Buttenfield, B.P., Nagle, N.N., and Stum, A.K., 2014. Modeling residential developed land in rural areas: A size-restricted approach using parcel data. *Applied Geography*, 47, 33–45.

Leyk, S., Uhl, J.H., Balk, D., and Jones, B., n.d. Assessing the accuracy of multi-temporal built-up land layers across rural-urban trajectories in the United States. *Submitted to Remote Sensing of Environment*.

Li, L. and Lu, D., 2016. Mapping population density distribution at multiple scales in Zhejiang Province using Landsat Thematic Mapper and census data. *International Journal of Remote Sensing*, 37 (18), 4243–4260.

Lin, J., Cromley, R., and Zhang, C., 2011. Using geographically weighted regression to solve the areal interpolation problem. *Annals of GIS*, 17 (1), 1–14.

Lin, J. and Cromley, R.G., 2015a. Evaluating geo-located Twitter data as a control layer for areal interpolation of population. *Applied Geography*, 58, 41–47.

Lin, J. and Cromley, R.G., 2015b. A local polycategorical approach to areal interpolation. *Computers, Environment and Urban Systems*, 54, 23–31.

Lin, J., Cromley, R.G., Civco, D.L., Hanink, D.M., and Zhang, C., 2013. Evaluating the use of publicly available remotely sensed land cover data for areal interpolation. *GIScience & Remote Sensing*, 50 (2), 212–230.

Linard, C., Gilbert, M., and Tatem, A.J., 2011. Assessing the use of global land cover data for guiding large area population distribution modelling. *GeoJournal*, 76 (5), 525–538.

Liu, X.H., Kyriakidis, P.C., and Goodchild, M.F., 2008. Population density estimation using regression and area to point residual kriging. *International Journal of Geographical Information Science*, 22 (4), 431–447.

Logan, J.R., Stults, B.J., and Xu, Z., 2016. Validating population estimates for harmonized census tract data, 2000–2010. *Annals of the American Association of Geographers*, 106 (5), 1013–1029.

Logan, J.R., Xu, Z., and Stults, B.J., 2014. Interpolating U.S. Decennial Census Tract Data from as Early as 1970 to 2010: A Longitudinal Tract Database. *The Professional Geographer*, 66 (3), 412–420.

Lu, Z., Im, J., Quackenbush, L., and Halligan, K., 2010. Population estimation based on multi-sensor data fusion. *International Journal of Remote Sensing*, 31 (21), 5587–5604.

Lung, T., Lübker, T., Ngochoch, J.K., and Schaab, G., 2013. Human population distribution modelling at regional level using very high resolution satellite imagery. *Applied Geography*, 41, 36–45.

Maantay, J. and Maroko, A., 2009. Mapping urban risk: Flood hazards, race, & environmental justice in New York. *Applied Geography*, 29 (1), 111–124.

Maantay, J., Maroko, A.R., and Herrmann, C., 2007. Mapping population distribution in the urban environment: The cadastral-based expert dasymetric system (CEDS). *Cartography and Geographic Information Science*, 34 (2), 77–102.

Martin, D., Dorling, D., and Mitchell, R., 2002. Linking censuses through time: problems and solutions. *Area*, 34 (1), 82–91.

MassGIS, 2014. MassGIS Data - FEMA National Flood Hazard Layer [online]. Available from: http://www.mass.gov/anf/research-and-tech/it-serv-and-support/application-serv/office-of-geographic-information-massgis/datalayers/nfhl.html [Accessed 14 May 2017].

MassGIS, 2016. MassGIS Data Download - Level 3 Assessor's Parcels [online]. *Office of Geographic Information (MassGIS), Commonwealth of Massachusetts, MassIT*. Available

from: http://www.mass.gov/anf/research-and-tech/it-serv-and-support/application-serv/office-of-geographic-information-massgis/datalayers/download-level3-parcels.html [Accessed 3 Dec 2016].

MassGIS, 2017. MassGIS Data - Building Structures [online]. *Office of Geographic Information (MassGIS), Commonwealth of Massachusetts, MassIT*. Available from: http://www.mass.gov/anf/research-and-tech/it-serv-and-support/application-serv/office-of-geographic-information-massgis/datalayers/structures.html [Accessed 11 May 2017].

Mecklenburg County GIS, 2013. Data - Open Mapping - Mecklenburg County GIS [online]. Available from: http://maps.co.mecklenburg.nc.us/openmapping/data.html [Accessed 3 Dec 2016].

Mennis, J., 2003. Generating Surface Models of Population Using Dasymetric Mapping. *The Professional Geographer*, 55 (1), 31–42.

Mennis, J., 2009. Dasymetric mapping for estimating population in small areas. *Geography Compass*, 3 (2), 727–745.

Mennis, J., 2015. Increasing the accuracy of urban population analysis with dasymetric mapping. *Cityscape*, 17 (1), 115–126.

Mennis, J., 2016. Dasymetric spatiotemporal interpolation. *The Professional Geographer*, 68 (1), 92–102.

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

Minnesota Population Center, 2016. National Historical Geographic Information System: Version 11.0 [Database]. *University of Minnesota*.

Mitsova, D., Esnard, A.-M., and Li, Y., 2012. Using enhanced dasymetric mapping techniques to improve the spatial accuracy of sea level rise vulnerability assessments. *Journal of Coastal Conservation*, 16 (3), 355–372.

Mrozinski, R.D. and Cromley, R.G., 1999. Singly- and doubly-constrained methods of areal Interpolation for vector-based GIS. *Transactions in GIS*, 3 (3), 285–301.

Mugglin, A.S. and Carlin, B.P., 1998. Hierarchical Modeling in Geographic Information Systems: Population Interpolation over Incompatible Zones. *Journal of Agricultural, Biological, and Environmental Statistics*, 3 (2), 111–130.

Multi-Resolution Land Characteristics, 2016. Multi-Resolution Land Characteristics Consortium (MRLC) [online]. Available from: https://www.mrlc.gov/index.php [Accessed 3 Dec 2016].

Nagle, N., Buttenfield, B., Leyk, S., and Spielman, S., 2014. Dasymetric modeling and uncertainty. *Annals of the Association of American Geographers*, 104 (1), 80–95.

Okabe, A. and Sadahiro, Y., 1997. Variation in count data transferred from a set of irregular zones to a set of regular zones through the point-in-polygon method. *International Journal of Geographical Information Science*, 11 (1), 93–106.

Openshaw, S., 1984. *The Modifiable Areal Unit Problem*. Concepts and Techniques in Modern Geography (CATMOG). Norwitch: GeoBooks.

Openshaw, S. and Taylor, P., 1979. A million or so correlation coefficients. *In*: N. Wrigley, ed. *Statistical Methods in the Spatial Sciences*. London: Pion, 127–144.

Pavía, J. and Cantarino, I., 2016. Can Dasymetric Mapping Significantly Improve Population Data Reallocation in a Dense Urban Area? *Geographical Analysis*.

Pesaresi, M., Ehrlich, D., Ferri, S., Florczyk, A., Freire, S., Halkia, M., Julea, A., Kemper, T., Soille, P., and Syrris, V., 2016. *Operating procedure for the production of the Global Human Settlement Layer from Landsat data of the epochs 1975, 1990, 2000, and 2014*. JRC Technical Report; European Commission, Joint Research Centre, Institute for the Protection and Security of the Citizen: Ispra, Italy.

Qiu, F., Sridharan, H., and Chun, Y., 2010. Spatial autoregressive model for population estimation at the census block level using LIDAR-derived building volume information. *Cartography and Geographic Information Science*, 37 (3), 239–257.

Rase, W.-D., 2001. Volume-preserving interpolation of a smooth surface from polygon-related data. *Journal of Geographical Systems*, 3 (2), 199–213.

Reibel, M. and Agrawal, A., 2007. Areal interpolation of population counts using pre-classified land cover data. *Population Research and Policy Review*, 26 (5–6), 619–633.

Reibel, M. and Bufalino, M.E., 2005. Street-weighted interpolation techniques for demographic count estimation in incompatible zone systems. *Environment and Planning A*, 37 (1), 127–139.

Ruther, M., Leyk, S., and Buttenfield, B.P., 2015. Comparing the Effects of an NLCD-derived Dasymetric Refinement on Estimation Accuracies for Multiple Areal Interpolation Methods. *GIScience & Remote Sensing*, 52 (2), 158–178.

Schneiderbauer, S. and Ehrlich, D., 2004. *Risk, hazard and people's vulnerability to natural hazards. A review of definitions, concepts and data*. Joint Research Centre, European Commission, EUR 21410.

Schroeder, J.P., 2007. Target density weighting interpolation and uncertainty evaluation for temporal analysis of census data. *Geographical Analysis*, 39 (3), 311–335.

Schroeder, J.P., 2017. Hybrid areal interpolation of census counts from 2000 blocks to 2010 geographies. *Computers, Environment and Urban Systems*, 62, 53–63.

Schroeder, J.P. and Van Riper, D.C., 2013. Because Muncie's densities are not Manhattan's: Using geographical weighting in the EM algorithm for areal interpolation. *Geographical analysis*, 45 (3), 216–237.

Semenov-Tian-Shansky, B., 1928. Russia: territory and population: a perspective on the 1926 census. *Geographical Review*, 18 (4), 616–640.

Smith, J., Wickham, J., Stehman, S., and Yang, L., 2002. Impacts of patch size and land-cover heterogeneity on thematic image classification accuracy. *Photogrammetric Engineering and Remote Sensing*, 68 (1), 65–70.

Sridharan, H. and Qiu, F., 2013. A spatially disaggregated areal interpolation model using light detection and ranging-derived building volumes. *Geographical Analysis*, 45 (3), 238–258.

Su, M.D., Lin, M.C., Hsieh, H.I., Tsai, B.W., and Lin, C.H., 2010. Multi-layer multi-class dasymetric mapping to estimate population distribution. *Science of the Total Environment*, 408 (20), 4807–4816.

Syphard, A.D., Stewart, S.I., Mckeefry, J., Hammer, R.B., Fried, J.S., Holcomb, S., and Radeloff, V.C., 2009. Assessing housing growth when census boundaries change.

International Journal of Geographical Information Science, 23 (7), 859–876.

Tanton, R., Williamson, P., and Harding, A., 2014. Comparing two methods of reweighting a survey file to small area data. *International Journal of Microsimulation*, 7 (1), 76–99.

Tapp, A.F., 2010. Areal interpolation and dasymetric mapping methods using local ancillary data sources. *Cartography and Geographic Information Science*, 37 (3), 215–228.

Tobler, W.R., 1979. Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association*, 74 (367), 519–530.

U.S. Census Bureau, 2010a. Standard Hierarchy of Census Geographic Entities [online]. Available from: https://www2.census.gov/geo/pdfs/reference/geodiagram.pdf [Accessed 27 May 2017].

U.S. Census Bureau, 2010b. American FactFinder - Download Center [online]. Available from: https://factfinder.census.gov [Accessed 10 May 2017].

U.S. Census Bureau, 2010c. QuickFacts Massachusetts [online]. Available from: https://www.census.gov/quickfacts/table/PST045216/25 [Accessed 10 May 2017].

U.S. Census Bureau, 2011. Differences Between the Census 2000 and 2010 Census Urban Area Criteria [online]. Available from: http://www2.census.gov/geo/pdfs/reference/ua/2000_2010uadif.pdf [Accessed 26 Feb 2017].

U.S. Census Bureau, 2012. Geographic Terms and Concepts - Census Tract [online]. Available from: https://www.census.gov/geo/reference/gtc/gtc_ct.html [Accessed 27 May 2017].

U.S. Census Bureau, 2016. TIGER/Line Shapefiles and TIGER/Line Files [online]. *Census.gov, Maps and Data, TIGER Products*. Available from: https://www.census.gov/geo/maps-data/data/tiger-line.html [Accessed 26 Nov 2016].

University of Florida GeoPlan Center, 2016. The Florida Geographic Data Library - FGDL [online]. *University of Florida GeoPlan Center*. Available from: http://www.fgdl.org/download/ [Accessed 3 Dec 2016].

Ural, S., Hussain, E., and Shan, J., 2011. Building population mapping with aerial imagery and GIS data. *International Journal of Applied Earth Observation and Geoinformation*, 13 (6), 841–852.

Vogelmann, J.E., Howard, S.M., Yang, L., Larson, C.R., Wylie, B.K., and Van Driel, N., 2001. Completion of the 1990s National Land Cover Data Set for the conterminous United States from Landsat Thematic Mapper data and ancillary data sources. *Photogrammetric Engineering and Remote Sensing*, 67 (6), 650–662.

Wang, S., Tian, Y., Zhou, Y., Liu, W., and Lin, C., 2016. Fine-Scale Population Estimation by 3D Reconstruction of Urban Residential Buildings. *Sensors*, 16 (10), 1755.

Wei, C., Taubenböck, H., and Blaschke, T., 2017. Measuring urban agglomeration using a city-scale dasymetric population map: A study in the Pearl River Delta, China. *Habitat International*, 59, 32–43.

Wright, J., 1936. A method of mapping densities of population: With Cape Cod as an example. *Geographical Review*, 26 (1), 103–110.

Wu, C. and Murray, A.T., 2005. A cokriging method for estimating population density in urban areas. *Computers, Environment and Urban Systems*, 29 (5), 558–579.

Wu, J., Li, Y., Li, N., and Shi, P., 2017. Development of an asset value map for disaster risk assessment in China by spatial disaggregation using ancillary remote sensing data. *Risk analysis*.

Wu, S., Wang, L., and Qiu, X., 2008. Incorporating GIS building data and census housing statistics for sub-block-level population estimation. *The Professional Geographer*, 60 (1), 121–135.

Xie, Y., 1995. The overlaid network algorithms for areal interpolation problem. *Computers, Environment and Urban Systems*, 19 (4), 287–306.

Xie, Y., Weng, A., and Weng, Q., 2015. Population estimation of urban residential communities using remotely sensed morphologic data. *IEEE Geoscience and Remote Sensing Letters*, 12 (5), 1111–1115.

Yoo, E., Kyriakidis, P.C., and Tobler, W., 2010. Reconstructing population density surfaces

from areal data: A comparison of Tobler's pycnophylactic interpolation method and area-to-point kriging. *Geographical Analysis*, 42 (1), 78–98.

Yuan, Y., Smith, R.M., and Limp, W.F., 1997. Remodeling census population with spatial information from Landsat TM imagery. *Computers, Environment and Urban Systems*, 21 (3), 245–258.

Zandbergen, P.A., 2011. Dasymetric Mapping Using High Resolution Address Point Datasets. *Transactions in GIS*, 15 (SUPPL. 1), 5–27.

Zandbergen, P.A. and Ignizio, D.A., 2010. Comparison of dasymetric mapping techniques for small area population estimates. *Cartography and Geographic Information Science*, 37 (3), 199–214.

Zillow, 2017. Zillow Data [online]. Available from: https://www.zillow.com/research/data/ [Accessed 11 May 2017].

Zoraghein, H. and Leyk, S., n.d. Consistent Population Estimation within Changing Census Boundaries: Enhancing Interpolation Frameworks through Dasymetry. *Under Review by International Journal of Geographical Information Science*.

Zoraghein, H., Leyk, S., Ruther, M., and Buttenfield, B.P., 2016. Exploiting temporal information in parcel data to refine small area population estimates. *Computers, Environment and Urban Systems*, 58, 19–28.