5-1-2014

Developing Population Grid with Demographic Trait: An Example for Milwaukee County, Wisconsin

Wei Xu
University of Wisconsin-Milwaukee

DEVELOPING POPULATION GRID WITH DEMOGRAPHIC TRAIT: AN EXAMPLE FOR MILWAUKEE COUNTY, WISCONSIN

by

Wei Xu

A Thesis Submitted in

Partial Fulfillment of the

Requirement for the Degree of

Master of Science

in Geography

at

The University of Wisconsin–Milwaukee

May 2014

ABSTRACT
DEVELOPING POPULATION GRID WITH DEMOGRAPHIC TRAIT: A CASE STUDY FOR MILWAUKEE COUNTY, WISCONSIN

by

Wei Xu

The University of Wisconsin – Milwaukee, 2014
Under the Supervision of Professor Zengwang Xu

Population grids have been widely used in estimating population in environmental justice studies and emergency management. The currently available population grids are only for total population. There is an increasing need for population grids that have not only total population but also populations of different age, gender, race, and ethnicity. This study explores the methodology to develop these population grids endowed with the demographic characteristics. Areal interpolation methods are used to transfer total population at census blocks to the cells of the grid. Kernel density method is used to estimate the relative probability of population of different subgroups at the cells, and to disaggregate total population into subgroups. Population grids of black, white, and other, as well as populations of age 20 to 34 and age 65 and over, are derived for Milwaukee County, Wisconsin. Applications of the population grids demonstrate their potentials.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I am very much grateful to my academic advisor Professor Zengwang Xu for his insights and patience in helping me complete this thesis. His continued encouragements have been a great source of strength throughout my whole study here at University of Wisconsin-Milwaukee. I would also like to express my gratitude to Professor Changshan Wu, Professor Woonsup Choi and Professor Rina Ghose for being in my thesis advisory committee. Their helpful feedbacks are very much appreciated. Furthermore, I want to thank the faculty, staff and fellow graduate students in the department who have enriched my learning experience enormously. Finally, I would like to deliver special thanks to my family. My whole experience of studying abroad could not have been possible without their support and love.

1. Introduction

Population data are mainly released in censuses as aggregate data in areas such as states, counties, tracts, block groups and blocks. Analyses based on the population aggregate data assume that population densities are constant within the areal units and change abruptly from one unit to its neighbors. This will introduce errors in spatial analyses in which population estimates are needed in zones that are different from the census geography. For example, when estimating the people at risk in an air pollution emission area or a hurricane evacuation zone from population data at census geographical units (e.g., census tracts or blocks), the constant-density-assumption will introduce uncertainty in the population estimates, sometimes, in a significant level. Population grids, as an alternative, describe the spatial variation of population density at very fine resolution and locate people in space more accurately. As a result, the last two decades have witnessed a growing need of population grids as a preferable data to population distribution analysis (Bracken, 1991, 1994; Bracken and Martin, 1989, 1995; Honeycutt and Wojcik, 1990; Martin and Bracken, 1991; Moon and Farmer, 2001; Deichmann et al., 2001; Mennis, 2002, 2003; Xie, 2006; Bhaduri et al., 2007; Gallego, 2010).

Population grids are popularly used in representing population distribution due to several reasons. Firstly, they are not only capable of representing the heterogeneity of population distribution at the pixel level but can also capture the spatial continuity of population densities. This is different from the spatial discontinuity of choropleth maps, which are considered an inadequate portrayal of population distribution because (1) the generalization of areal units, acting as “low-pass spatial filters”, masks the heterogeneity of population distribution within them; (2) those boundaries of areal units have little logic

relationship to the spatial continuity of population densities (Goodchild, 1989; Langford and Unwin, 1994, p.22). Secondly, many population and housing data are released at census block or tract level; however, many analyses have to go down to the sub-block or sub-tract level. It is a nice feature if we can locate people at finer level than the census blocks. Thirdly, many different zonal systems (e.g., blocks, block groups, census tracts, counties and states) are used for population and other social economic data. Integration data from different zonal systems presents challenge to many social scientists and very much introduces additional errors. This is discussed by many scholars as the *incompatible zonal system problem* (Openshaw, 1983; Flowerdew and Green, 1989, 1990; Goodchild et al., 1993; Fisher and Langford, 1995). Population grids with high spatial resolutions are more flexible in this regard. If it has to be integrated with zonal systems, population estimates can be easily obtained by superimposing the zone boundaries on the grids (Martin and Bracken, 1991). Fourthly, area-based spatial analyses are usually prone to the *modifiable areal unit problem* (MAUP), which refers to the inconsistency in analytical results introduced by "modifiable" *zoning* and *scaling* in geographical units (Openshaw and Taylor, 1979; Openshaw, 1983; Fotheringham and Wong, 1991; Wong, 2004; Wu, 2004; Yang, 2005). Moreover, geographical units that are used in decennial censuses are designated to different scales and those units are often subject to changes. As a field-based model, the population grid treats population distribution as a continuous surface and enables demographical analysis to be independent of the arbitrary areal units, and hence is immune from the MAUP (Bracken, 1993; Mennis, 2003). Finally, locational and attribute errors can be more readily modelled in population (Goodchild, 1989; Martin and Bracken, 1991). Population grids

provide a fine-resolution, and easy-to-analyze representation of the spatial distribution of population. However, most of the existing population grids only account for total population counts and pay little attention to the variations of population distributions of different sex, age, race, and ethnicity.

Developing population grids is essentially an areal interpolation process because population data are transferred from source zones (e.g. census enumeration units) to target zones (regular sized cells). A number of areal interpolation methods have been developed to enable data transferring between different zonal systems, such as area weighting interpolation, pycnophylactic interpolation, dasymetric interpolation, and road network interpolation, etc. A major issue with these areal interpolation techniques is that their algorithms are mainly density based, i.e., they assume the population are uniformly distributed in the zones and apportion population according to their density. This is feasible for interpolating total population, but insufficient for interpolating population of demographic subgroups, as they might account for different proportions of the total population at different location. These interpolations perform poorly in estimating the sub-populations. In this study, I intend to combine areal interpolation and statistical modelling to develop high spatial resolution (30m×30m) population grids for Milwaukee County, Wisconsin that are able to represent not only the total population distribution but also the populations of different race and age groups.

2. Literature review

2.1. Existing population grids

There are a lot of efforts to develop population quadrilateral grids at both the global and local scales. At the global level, those grids include the Gridded Population of the World (GPW), LandScan Global Population Database, and the Global Rural Urban Mapping Project (GRUMP) (Bhaduri et al., 2007). The Gridded Population of the World (GPW) released its first version in 1995, which was the first effort to establish a global-level dataset representing nighttime population distribution (Tobler et al., 1995). The dataset was created by disaggregating population in national or subnational administrative units, depending on the highest spatial resolution available, into grids using a simple area weighting interpolation technique (Tobler et al., 1995; Center for International Earth Science Information Network [CIESIN] et al., 2004). The GPWv2 emphasized on improving the spatial resolution of input administrative boundaries and implemented Waldo Tobler's *smooth pycnophylactic interpolation* method on GPWv1 to get a smooth population grid (Tobler, 1979; Deichmann et al., 2001). The third version not only made the effort to keep acquiring higher resolution data for countries but also attempted to update the changes of administrative boundaries (Balk and Yetman, 2004). The Global Rural Urban Mapping Project (GRUMPv1) is a set of global population grids that are built on the GPWv3, with the same spatial resolution of 1km×1km at the equator. By incorporating nighttime light data produced by the U.S. National Geophysical Data Center (NGDC) in the population allocation process, the grids are able to distinguish urban and rural areas (CIESIN, 2004). In 1998, the United States Department of Energy's (USDOE) Oak Ridge National Laboratory (ORNL) released its first global population

database named LandScan Global, a population grid with cell size of approximately 1km ×1km at the equator. This dataset uses land cover, roads, slope, and nighttime lights as ancillary information to reallocate population counts (usually at provincial level) into grid cells (Bhaduri et al., 2002). The ORNL has been updating the database on an annual basis and the quality of the datasets is improved by utilizing finer resolution inputs in the modeling process. It is worth noticing that, unlike other global population grids, the LandScan Global integrated population movements in its model, so that the grids actually denote the "ambient or average population distribution over a 24-hour period" rather than nighttime or residential population distribution (Dobson et al., 2000; Bhaduri et al., 2002, p.35).

There also are continental level population grids to meet different needs such as environment, telecommunication, and public health issues. The GEOSTAT 1A project was implemented by EUROSTAT, a directorate branch of the European Commission, aiming at generating a 1km ×1km resolution grid-based population system for Europe by using the 2006 European censuses. Participating countries in this project include "Austria, Estonia, Finland, France, the Netherlands, Poland, Portugal and Slovenia" (European Forum for GeoStatistics [EFGS], 2012). For Asia, the National Center for Geographic Information and Analysis (NCGIA) and University of California, Santa Barbara (UCSB) published *the Corresponding (E. Asia) Population Density Map* and *the Corresponding (W. Asia) Population Density Map* in 1996, with the purpose of providing spatially referenced population databases that support applications including agricultural research and the analysis of environmental change (Deichmann, et al., 1996a, 1996b). These maps are also available in the grid format.

A number of population grids have been developed at the national level. To meet the need for more accurate population data within the United States, the ORNL expands the LandScan Global dataset to create the LandScan USA, a population density grid with a very high spatial resolution of 90m ×90m cell size. LandScan USA takes into account the movement of workers and students in the interpolation approach, and the dataset is capable of capturing population distribution over both space (the United States) and time (daytime and nighttime) (Bhaduri et al., 2007). The ORNL is also considering including the "transient population" such as tourists and business travelers in future releases to refine the dataset (Bhaduri et al., 2007, p.116). An intermediate-resolution population grid (approximately 450m ×450m) for the contiguous United States has been developed by the ORNL as well (Bhaduri et al., 2002). The U.S. Census Grids is a set of raster data that provide individual, household, and housing unit information for 1990 and 2000 across the county. The spatial resolution for the United States and Puerto Rico (Puerto Rico datasets are only available for 2000) is approximately 1km ×1km; while the resolution for 50 metropolitan areas is about 250m ×250m. In the dataset, the statistics regarding individuals include age, race, ethnicity, and so forth, but exclude sex (CIESIN, 2006). Countries other than the United States are also making efforts to create their own national population grids. A population density grid for Spain is developed using the dasymetric mapping approach by integrating high-resolution land cover data and census data (Goerlich and Cantarino, 2013). Also, disaggregated datasets from European grid were produced by the Austrian Institute of Technology (AIT) for European member countries (EFGS, 2012).

Most of the population grids mentioned above only use census population count in administrative units as the input to construct the grids, thus the grids portray nighttime or residential population distribution. LandScan Global and LandScan USA are exceptions because they incorporate population movements during the daytime. However, almost all of these population grids focus on the total number of the population rather than population with more demographic description, and the spatial resolutions of existing population grids with demographic traits are too coarse to explain the micro-level population distribution. With new interpolation techniques and new ancillary data being developed, it is foreseeable that population density grids with more demographic details will be produced.

2.2. Areal interpolation methods

Areal interpolation is referred to as the process of estimating the unknown values in the target zones given the known values in the source zones (Goodchild et al., 1980; Hawley and Moellering, 2005). There have been a large number of areal interpolation techniques developed and refined by demographers, statisticians and geographers in order to bridge incompatible zonal systems. For population estimation, those methods can be classified into two major categories: methods without ancillary data and methods with ancillary data (Hawley and Moellering, 2005; Wu et al., 2005; Xie, 2006).

Areal interpolation techniques without ancillary data estimate population in the target zones according to the change between source and target zones (Hawley and Moellering, 2005). The area weighting method (Lam, 1983), the “point-in-polygon”

method (Okabe and Sadahiro, 1997), and the pycnophylactic method (Tobler, 1979) are the most commonly used areal interpolation methods that do not need ancillary data. Other methods use ancillary data to distribute the weights of the values in source zones and to develop reliable target zone estimates, such as control zones (Flowerdew and Green, 1989; Goodchild et al., 1993), remote sensed imageries (McCleary, 1969; Langford et al., 1991; Fisher and Langford, 1995, 1996; Yuan et al., 1997; Langford, 2006, Deng and Wu, 2013), road network (Xie, 1995; Mrozinski and Cromley, 1999; Reibel and Bufalino, 2005), and other relevant demographic statistics such as housing units (Xie, 2006; Wu et al., 2008; Deng and Wu, 2013).

2.2.1. Areal interpolation methods without ancillary data

The simplest areal interpolation technique that does not need any ancillary data is the "point-in-polygon" method, which first determines a "representative point" for each source zone and assigns the value of the variable of interest in the source zone to the point. If the point is located in a target zone, the value of the point is assigned to the target zone through a point-location algorithm (Okabe and Sadahiro, 1997; Sadahiro, 2000). The biggest advantage of the method is the computation efficiency, which is essential if the numbers of source and target zones are very large (Okabe and Sadahiro, 1997). However, one of the problems of this method is that the selection of "representative points" can be rather random and the estimation results can be significantly different due to the algorithms used in the point selection process. For example, the points could be the centroids of source zones, the centers of source zones, the weighted centers of a certain variable, or even some arbitrary points (Okabe and

Sadahiro, 1997; Hawley and Moellering, 2005). The sizes of both source zones and target zones contribute variations in the selection of "representative points" as well (Okabe and Sadahiro, 1997). A similar implementation of the "point-in-polygon" method is the isopleth mapping technique. The method superimposes a grid on source zones and assigns the value of each source zone to a representative point. The values of grid cells are then interpolated from the representative points and averaged within source zones to obtain the source zone values (Lam, 1982). One of the problems with the method is that the populations in the target zones after interpolation might not be equal to those in the source zones even though the geographical boundaries are identical (Lam, 1982). This method generates fairly poor estimation results in terms of accuracy.

Another method is the area weighting method, or polygon overlay method (Markoff and Shapiro, 1973; White, 1978; Goodchild et al., 1980; Lam, 1983; Flowerdew and Green, 1989; Hawley and Moellering, 2005), or areal weighting method (Geogory, 2002; Geogory and Ell, 2006), or areal weighted method (Fisher and Langford, 1996), or simple areal interpolation (Flower and Green, 1993; Eicher and Brewer, 2001; Mugglin et al., 2000), as named in some literatures. The names of the methods vary but their principles are the same. This method uses the intersections between source zones and target zones to allocate the value of the variable of interest to target zones. The values in the overlapping areas are apportioned according to the proportion of the overlapping areas in the source zones (Goodchild et al., 1980; Lam, 1983; Fisher and Langford, 1995). This method assumes that the value of the variable of interest is evenly distributed in the source zones, which is counterfactual in most situations (Goodchild et al., 1980; Lam, 1983).

By examining the previous areal interpolation methods, Tobler (1979) argues that the assumption that the value of the variable of interest is a constant in one single region might be plausible, but is very dubious when regions are interconnected because populations in neighboring regions do not always have sharp differences. To tackle this problem, he proposed the pycnophylactic method which assumes the existence of a continuous population density surface in the area of concern. A "smooth density function", which can be considered as a "value restraint" of predicted values, is employed to allow for "the effect of adjacent source zones" (Lam, 1983, p.140). This method, unlike the area weighting method, acknowledges the heterogeneity of the distribution of population within each source and target zone (Lam, 1983; Hawley and Moellering, 2005). The procedure is rather simple. First, a non-negative-value grid is generated to preserve the values of source zones and the value in each cell is determined by dividing the total value in the source zones by the number of cells in the zones. Then, the value in each cell is resampled with the values of the 4 neighboring cells. This smoothing process is reiterated until either the cell values no longer show significant difference compared to the cell values in the last iteration process or predicted source zone values do not have a significant different to actual source zone values (Hawley and Moellering, 2005). Acknowledged by many, one major advantage of this method is that it satisfies the *pycnophylactic property*, which means the magnitude of population in source zones is preserved in the interpolation process (Tobler, 1979; Lam, 1983; Reibel and Bufalino, 2005; Mennis, 2009). Rase (2001) later refined this method by introducing the *Triangulated Irregular Networks* (TIN) instead of the cells in creating the density surface because the TIN have the ability to follow the actual boundaries of source zones. These

areal interpolation methods are likely to yield poor estimation results because their fundamental assumption is that population within each source zone is homogenously distributed, although some methods add some refinements to this assumption.

2.2.2. Areal interpolation methods with ancillary data

Areal interpolation methods that use ancillary data take into account other useful information that is associated with the population distribution in the source zones. Flowerdew and Green (1989) proposed the "control zone" method to bridge incompatible zonal systems. Goodchild et al. (1993) improved this method by redefining control zones as a third set of areal units with constant densities. By overlapping source zones with control zones, the densities of control zones can be obtained and employed in calculating the values in the target zones.

Remote sensing data can be used as the ancillary data. Langford et al. (1991) developed a regression method to correlate different land use types with population densities. After a remote sensing imagery is classified into different land-use types according to their different spectral characteristics, the number of pixels of each land-use type in the source zones can be obtained. Then, a regression is performed based on the populations and the pixel numbers of land-use types in the source zones. The parameters (coefficients) of land-use types can then be used to estimate the populations in the target zones after the pixel number of each land-use type in the target zones is determined. The land cover data produced through the USGS's National Land Cover Dataset (NLCD) program can also be used as ancillary data (Cai et al., 2006; Reibel and Agrawal, 2007; Zandbergen and Ignizio, 2010). The regression in this method is performed globally,

meaning that the coefficients of all independent variables are constant across different regions in the study area. Yuan et al. (1997) argued that it was necessary to recognize the spatial non-stationarity of the correlation relationship between population densities and land cover types, and applied a geographically weighted regression (GWR) technique to improve the goodness-of-fit of the model. The estimation yields more accurate results than the regression models do.

An alternative to the regression method is the dasymetric method introduced by Wright (1936). In the case study of mapping population distribution in Cape Cod, he used topographic sheets to differentiate inhabitable and uninhabited areas, as well as the variations of densities in inhabitable areas. The general concept of taking advantage of other ancillary information has been widely employed in areal interpolation. When redistributing the populations in the source zones, the values are constrained by other source of information, such as the land cover data (Yuan et al., 1997; Eicher and Brewer, 2001; Holt et al., 2004; Langford, 2006; Briggs et al., 2007), street network (Xie, 1995; Mrozinski and Cromley, 1999; Reibel and Bufalino, 2005), parcel-based data (Maantay et al., 2007; Maantay et al., 2008; Deng and Wu, 2013), and reference maps (Langford, 2007). The commonly used dasymetric methods include the binary dasymetric method, the 3-class dasymetric method and the limiting variable method (Eicher and Brewer, 2001; Hawley and Moellering, 2005). The binary dasymetric method divides the study area into populated and unpopulated areas and allocates the total population to the populated areas (One example of the exclusionary regions is water bodies); the 3-class dasymetric method refines the classification scheme by dividing the population in the source zones into urban, agriculture/woodland, and forested categories with calibrated

weights; in the limiting variable method, land use types are assigned with the density threshold. If the population density in a polygon of a specific land use type exceeds the threshold, the threshold density is assigned to the polygon, and then the remaining population is redistributed by the same procedure (Eicher and Brewer, 2001; Hawley and Moellering, 2005; Longford, 2006; Tapp, 2010). The dasymetric method is also performed locally so that population in each source zone is allocated according to its own land use types and weights.

Another areal interpolation method with ancillary data is the road network method developed by Xie (1995). An overlaid network algorithm is used in the method to associate the distribution of population with road networks, assuming the existence of a positive correlation between the densities of the two. The road networks can be used as the weighting factor in three different ways: (1) the network length method uses the length of road segments in source zones to determine the population to be allocated in target zones; (2) the network hierarchical weighting method assumes that people are more likely to reside along neighborhood roads than along primary roads such as interstate highways, and it takes into account the various classes of road features, and uses a weight matrix of road networks to allocate population in source zones to target zones; (3) the network housing-bearing method takes advantage of the address ranges of the road segments to approximate the distribution of population (Xie, 1995; Voss et al., 1999; Reibel and Bufalino, 2005; Hawley and Moellering, 2005). A number of comparative analyses of areal interpolation techniques have showed that the road network methods outperform the areal weighting interpolation, the pycnophylactic method, and the land-

use regression method (Xie, 1995; Hawley and Moellering, 2005; Reibel and Bufalino, 2005).

Other areal interpolation methods are fairly flexible in terms of the ancillary data included. In "smart" interpolation methods, any information that is related to population distribution can be utilized to predict densities (Deichmann, 1996; Mennis and Hultgren, 2006). The term "smart" interpolation was first coined by Willmott and Matsuura (1995) when they used elevation and exposure information as weighting factors to interpolate annually averaged air temperature in the U.S. The method has also been widely used to make population density estimates (Dobson et al., 2000; Turner and Openshaw, 2001; Bhaduri et al., 2007). Conceptually, it refers to a multi-dimensional dasymetric model that uses finer resolution ancillary data (or indicator variables) as weighting factors to reallocate source zone populations to target zones. Mennis and Hultgren (2006) alternatively named the approach as "intelligent" dasymetric modelling. Datasets that can be used in "smart" interpolation range from rivers, roads, and settlements to elevation contours, wetlands, and neural networks. A weighting factor surface is created by the ancillary datasets, and populations in source zones are redistributed according to the weights associated with specific locations.

Some methods incorporate statistical analyses in the areal interpolation process. Flowerdew (1988) carried out a Poisson regression on a number of binary variables, with values of either presence or absence, to estimate target zones densities. Flowerdew and Green (1989) and Flowerdew et al. (1991) refined this method by introducing the Expectation-Maximization (EM) algorithm, which is an iterative modeling procedure originally developed by Dempster et al. (1977). By treating population in target zones as

missing data, the Expectation step in the EM algorithm computes the conditional expectation of missing values based on the values in source zones and the Maximization step constraints the results with maximum likelihood (Flowerdew and Green, 1989). Schroeder and Van Riper (2013) incorporated the GWR in the EM process, which allows the coefficients of control variables to vary among source zones. The hierarchical Bayesian method uses a probabilistic model based on available covariates in target zones to infer the population distribution in source zones (Handcock and Stein, 1993; Mugglin et al., 1999; Mugglin et al., 2000).

2.3. Demographic models in population estimation

Demographic models have the potentials to estimate population because population changes are usually registered or correlated to other demographic variables that are recorded. However, in contrast to the vast body of literature on the areal interpolation techniques aforementioned, the demographic models used in the population estimation are relatively limited. Major demographic methods include the Census Component Method II (CM-II) (U.S. Census Bureau, 1966), the Administrative Records method (Starsinic, 1974; Plane and Rogerson, 1994), the Ratio Correlation method (Martin and Serow, 1978), and the Housing Unit (HU) method (Smith and Lewis, 1980; Lo, 2003; Smith and Cody, 2004; Deng and Wu, 2013).

The CM-II method estimates the population for a post-censual year by adding the natural increase and net migration to the population counts in the base year. Vital statistics such as birth and death records over the period are used to generate the natural increase, and net migration is approximated from military records and school enrollments

data (Ghose and Rao, 1994). The administrative records method is designed to estimate the net migration component in the CM-II method. In this method, the net migration population under 65 years old is estimated through federal income tax form; while net migration of population age 65 and over is estimated using Medicare enrollment information (Plane and Rogerson, 1994; Deng and Wu, 2013). The ratio correlation method assumes that the population change in a spatial unit is correlated with the changes of a set of "coincident indicators" that are representative of the population in the unit (Martin and Serow, 1978, p.224). Thus, this relationship between population change and symptomatic variables can be used in a regression to adjust population estimates. Those "coincident indicators" range from school enrollment, tax return to car registration and workplace population (Smith and Mandell, 1984). However, the choice of "coincident indicators" is strongly restricted by the availabilities of those data at the geographic levels at which population is to be estimated. The housing unit method estimates population by multiplying the number of housing units with average persons per household, and adding the result with the population of group quarters (Smith and Lewis, 1980). The number of housing units within a study area can be estimated from utility records or building permits (Smith and Cody, 2004), or extracted from satellite imageries (Lo, 2003; Deng and Wu, 2013). The average persons per household can be obtained from sample survey or census data or demographic modeling. Moreover, it is possible to categorize the housing units and to apply different persons-per-household ratios to different categories (Wu et al., 2005). Population living in group quarters such as college residence halls, nursing homes, military barracks, and correction facilities are readily available in census

datasets, which help include the part of population who do not live in typical household arrangements in the estimation results.

2.4. Integrated approaches

With the increasing availability of various ancillary data and the growing capabilities of processing those data in the geographic information system (GIS), different methods and processes are being integrated to achieve more accurate interpolation results. For instance, the most recent advancements of dasymetric methods revolve around the inclusion of the Expectation-Maximization (EM) algorithm and exploring nontraditional ancillary information (Qiu and Cromley, 2013). Schroeder and Van Riper (2013) proposed the geographic weighted expectation-maximization (GWEM) algorithm to allow control unit densities to vary over space. Sridharan and Qiu (2013) incorporated the EM algorithm with housing unit volume derived from high-resolution remote sensed images to replace previous equal-sized control units with varying-size units. Griffith (2013) implemented the EM algorithm with a set of Poisson random variables to impute population counts as missing values. Bentley et al. (2013) improved the road network interpolation method by including cadastral lots as areal control units with the traditional road segments as linear control units. Deng and Wu (2013) expanded the housing unit method by introducing a geographic approach which extracted the housing unit numbers and average persons per household from high-resolution remote sensing imageries. Cai et al. (2006) used spatial interpolation methods to estimate age-sex proportion surfaces and incorporated these surfaces with total population in a binominal model to estimate age-sex specific populations in small areas.

As recent studies have shown, integration of areal interpolation method with demographic or statistical modeling has the potential to significantly improve the accuracy of small area population estimation. In my study, areal interpolation method and statistical modeling are combined to build the race and age specific population grids.

3. Research design

3.1. Study area

Milwaukee County, located in Southeast Wisconsin, is chosen as the case study area (see Figure 1). According to U.S. Census Bureau, the county has a total area of 3,080.89km^2 in 2010, of which 625.23 km^2 (20.29%) is land and 2,455.66km^2 (79.71%) is waster. In this study of population distribution, the water area that is essentially part of Lake Michigan is excluded. The county boundary, therefore, follows the outer blocks defined in the census, and the study area has 628.22 km^2 in total. As a part of the Milwaukee–Waukesha–West Allis metropolitan area and the most populated county in the state of Wisconsin, the racial composition of Milwaukee County has been fairly complicated. As of the 2010 Census, the county has a total population of 947,735, among which 574,656 (60.63%) are White alone, 253,764 (26.78%) are Black or African American alone, and 119,315 (12.59%) of others[1]. Regarding age distribution, 224,295 (23.67%) people are of age 20 to 34 and 132,381 (13.96%) people are of age 65 and over. These two age groups are selected as representatives of young and elderly populations, respectively. Figure 2-7 show the choropleth maps of total- and sub-populations at census block level. These maps generally depict how populations are distributed. Blocks in dark red have more populations; while blocks in light red have less.

[1] The race scheme used in this study is generalized from the original Census racial categories to three major racial categories: White alone, Black or African American alone, and other. The "other" category encompasses "American Indian and Alaska Native alone, Asian alone, Native Hawaiian and Other Pacific Islander alone, some other race alone, and of two or more races." According to the 2010 Census data for Milwaukee County, among the 119,315 of "other," 6,808 are American Indian and Alaska Native alone, 32,422 are Asian alone, 363 are Native Hawaiian and Other Pacific Islander alone, 51,429 are some other race alone, and 28,293 are of two or more races.

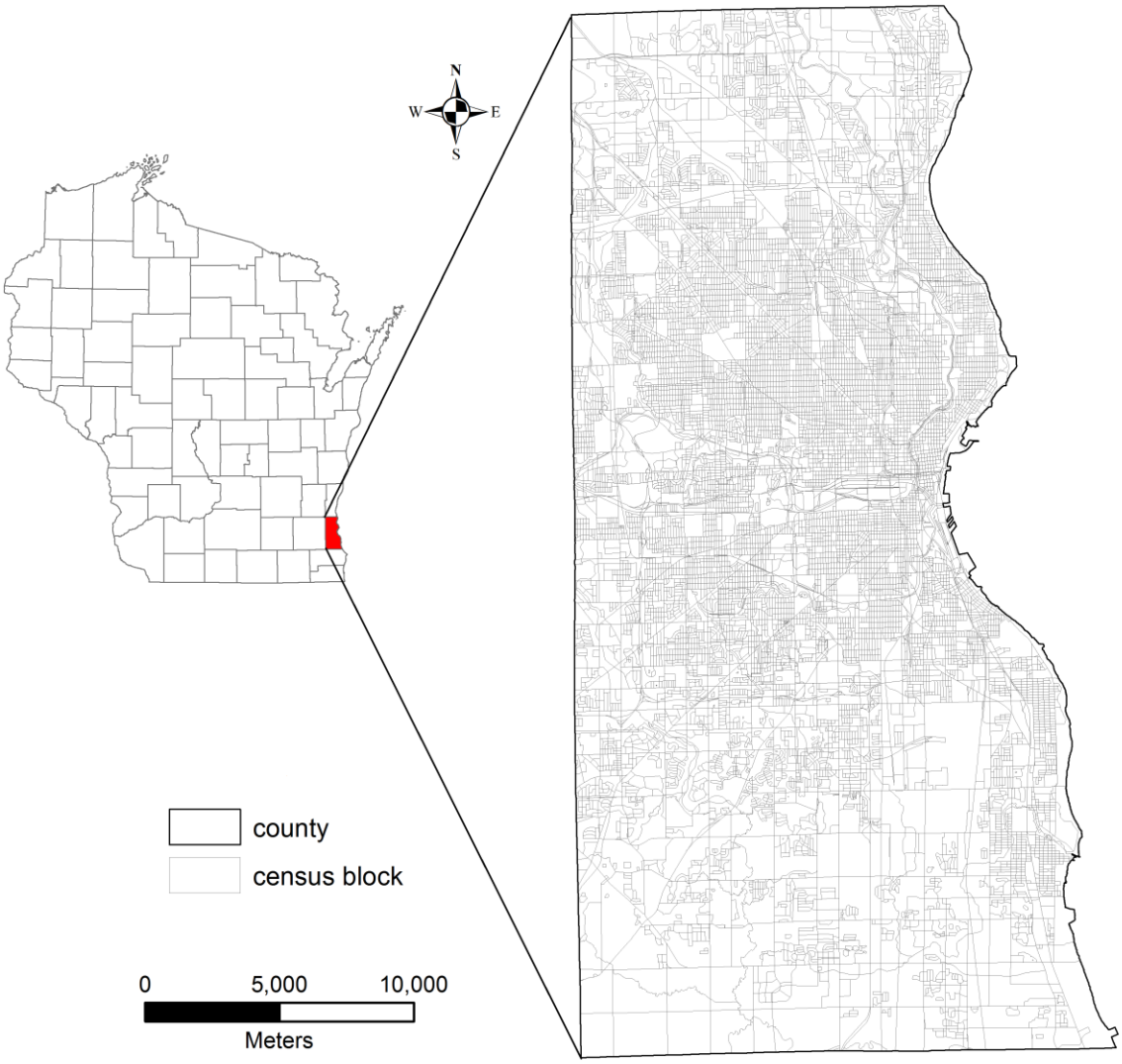


Figure 1. The location of the study area, Milwaukee County, WI.

To further examine the spatial distribution of the populations, Anselin's Local Moran's *I* is calculated to reveal the spatial clusters and outliers of high and low populations. Figure 8-13 is a set of maps that show the significance of local Moran's *I* statistics for total- and sub-populations, respectively. Census blocks that are shaded grey are statistically insignificant, meaning populations in these locations are in a random spatial pattern. Other census blocks shaded with different colors are associated with cluster types: Census blocks of HH denote a statistically significant cluster of high populations; HL denotes that blocks with high populations are surrounded by blocks with low populations; LH denotes that blocks with low populations are surrounded by blocks with high populations; and LL denotes a statistically significant cluster of low populations (ESRI, 2013). As shown in Figure 8-13, total- and sub-populations have rather different cluster patterns. Total population is rather dispersed in general, even though there are a number of significant clusters; White population is mostly clustered in the south and east side of the county; Black population is clustered in the northwest side of downtown; Other population is mostly clustered in the area closer to the south side of downtown; population between age 20 and 34 clusters in the same location as Other population does and the east side of the county; population of age 65 and over clusters in suburb areas. Different from the choropleth maps in Figure 2-7, these maps show not only how populations are spatially distributed but also the relationship between neighboring blocks.

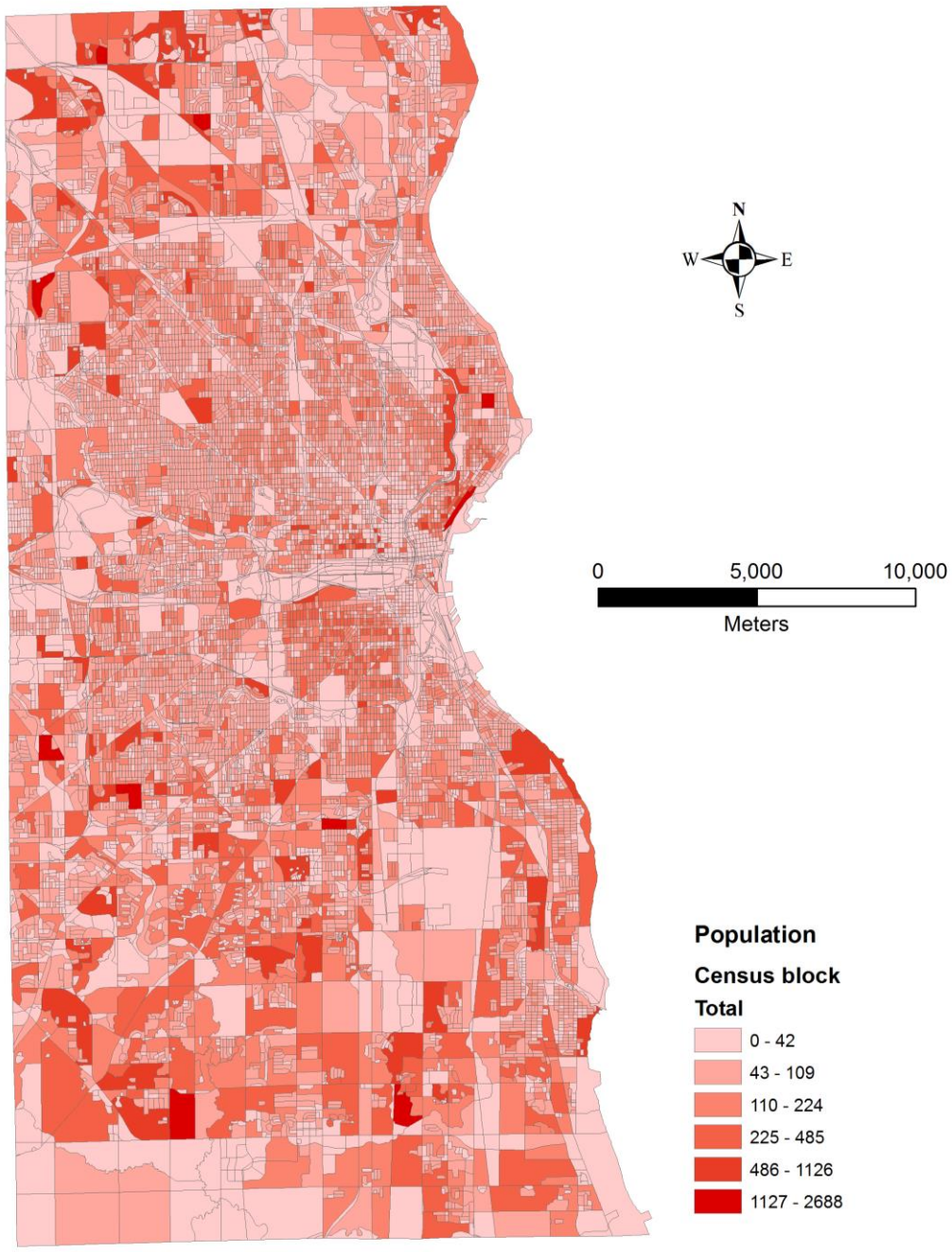


Figure 2. Choropleth map of census blocks representing the distribution of total population.

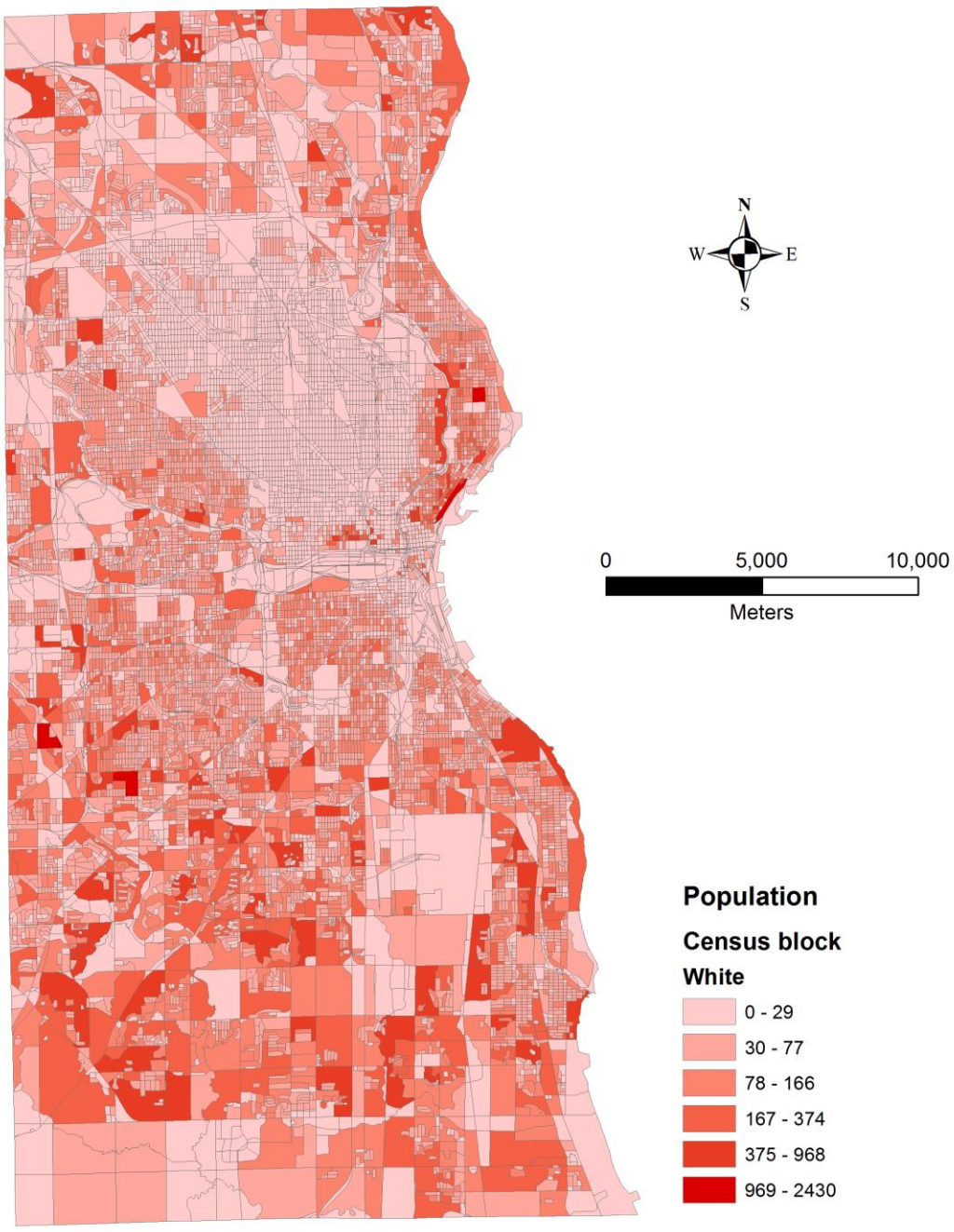


Figure 3. Choropleth map of census blocks representing the distribution of white population.

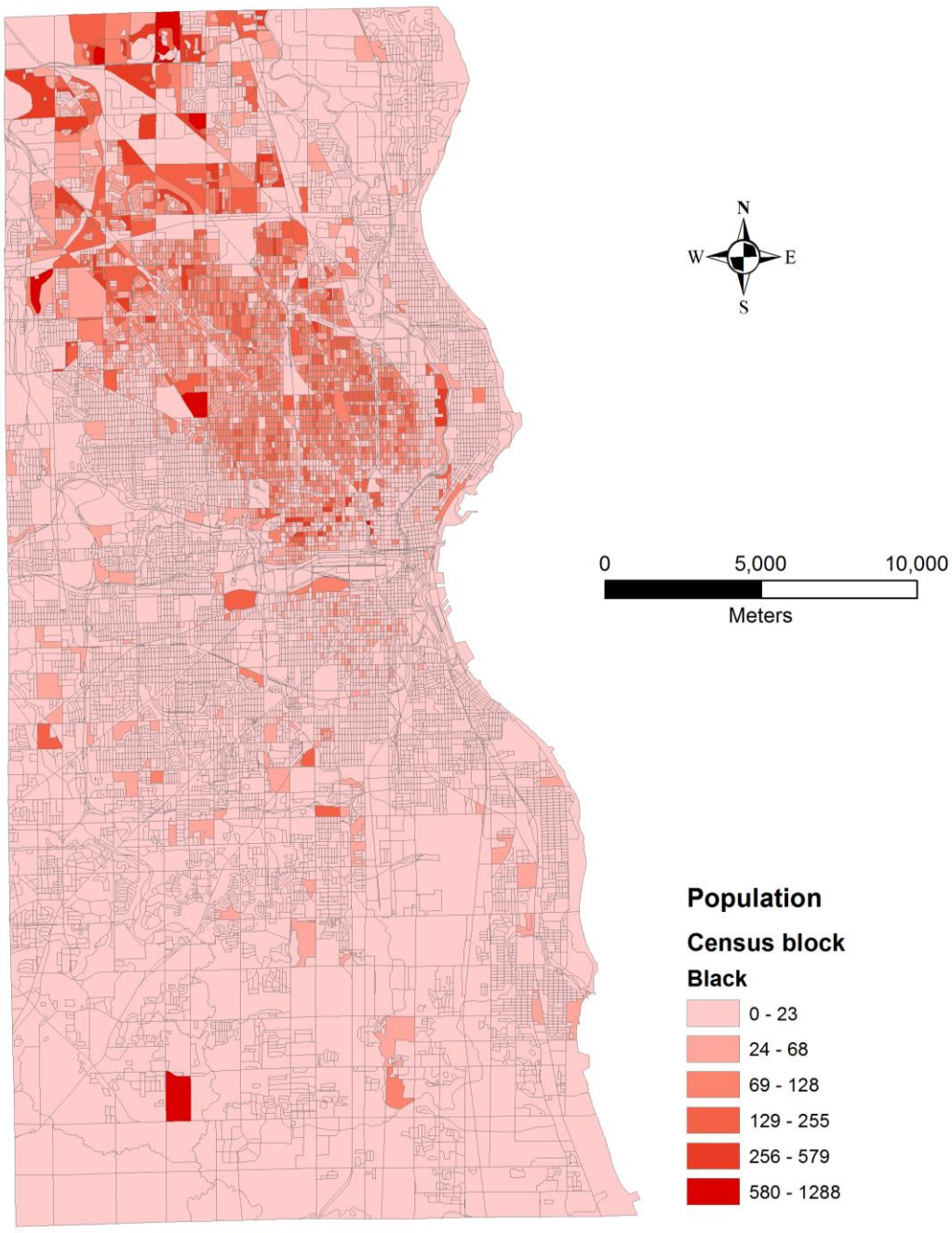


Figure 4. Choropleth map of census blocks representing the distribution of black population.

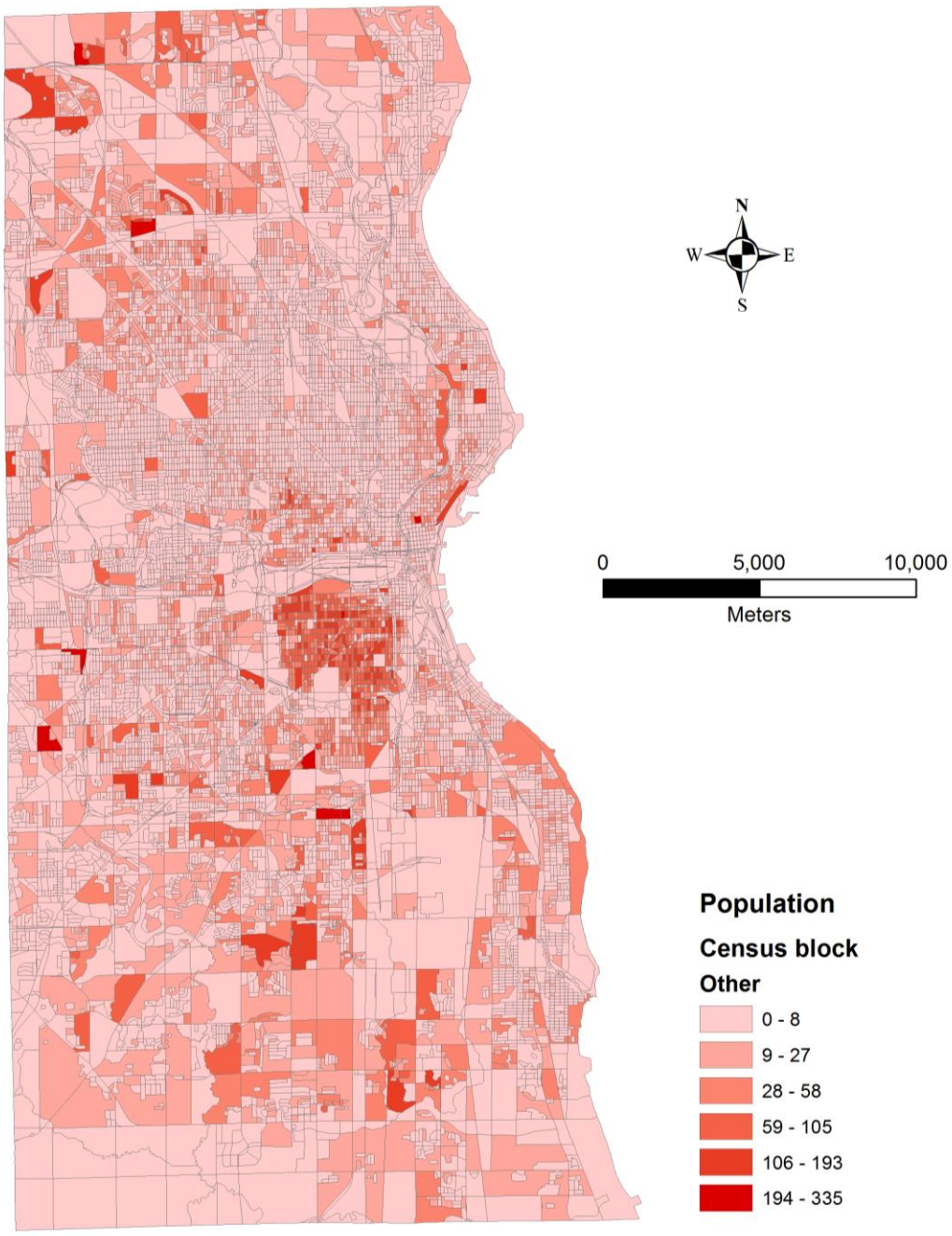


Figure 5. Choropleth map of census blocks representing the distribution of other population.

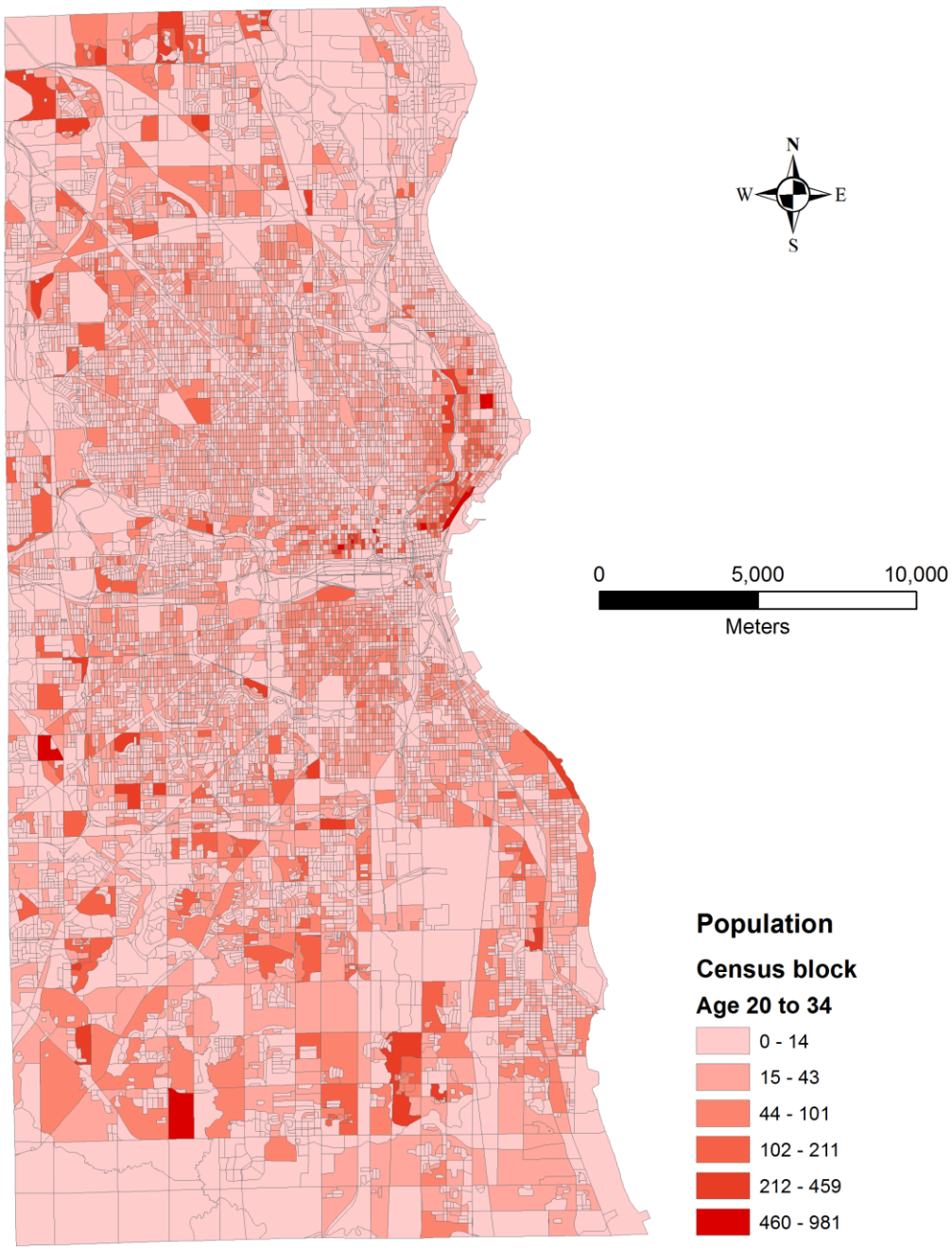


Figure 6. Choropleth map of census blocks representing the distribution of population of age 20 to 34.

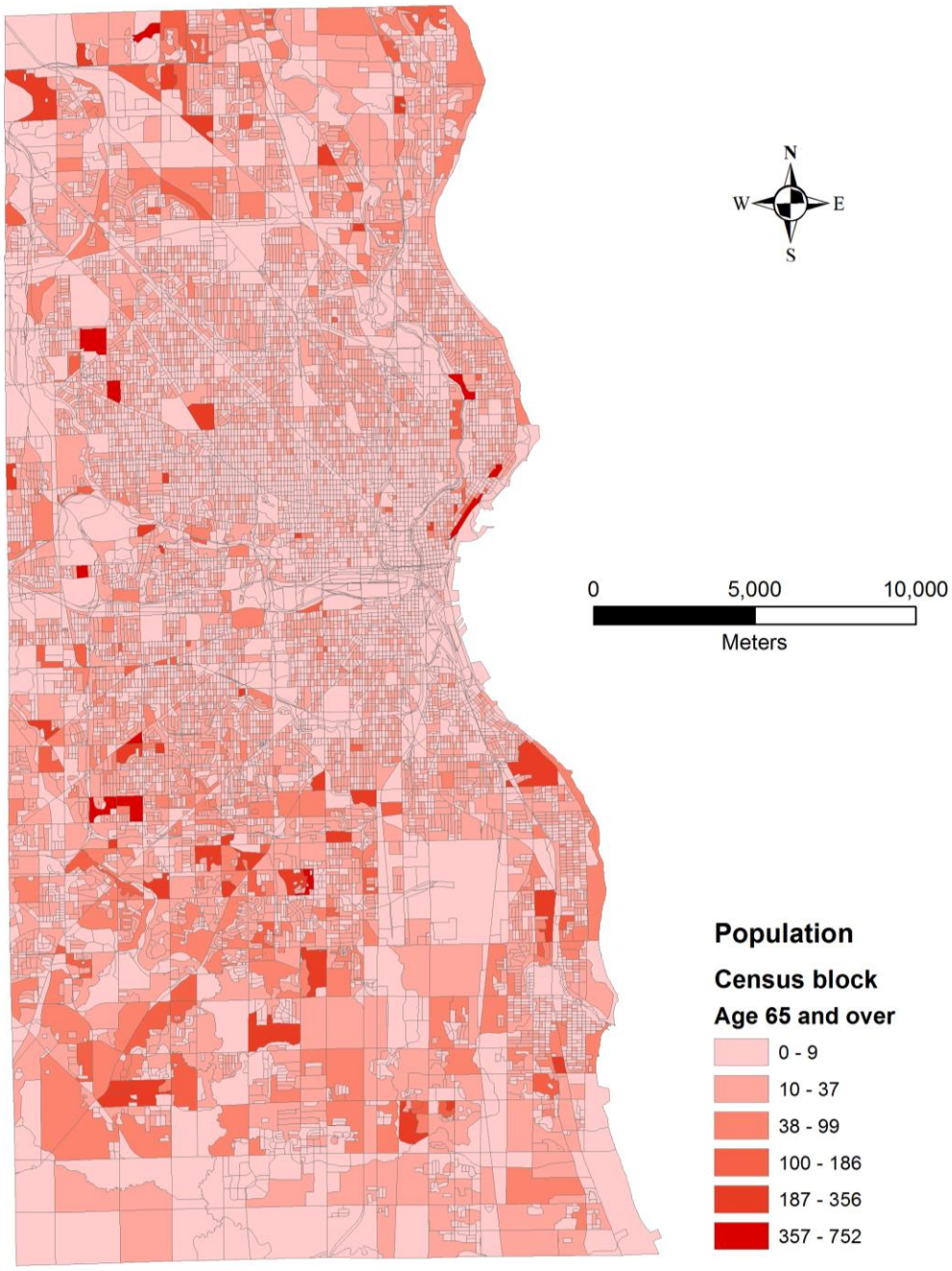


Figure 7. Choropleth map of census blocks representing the distribution of population of age 65 and over.

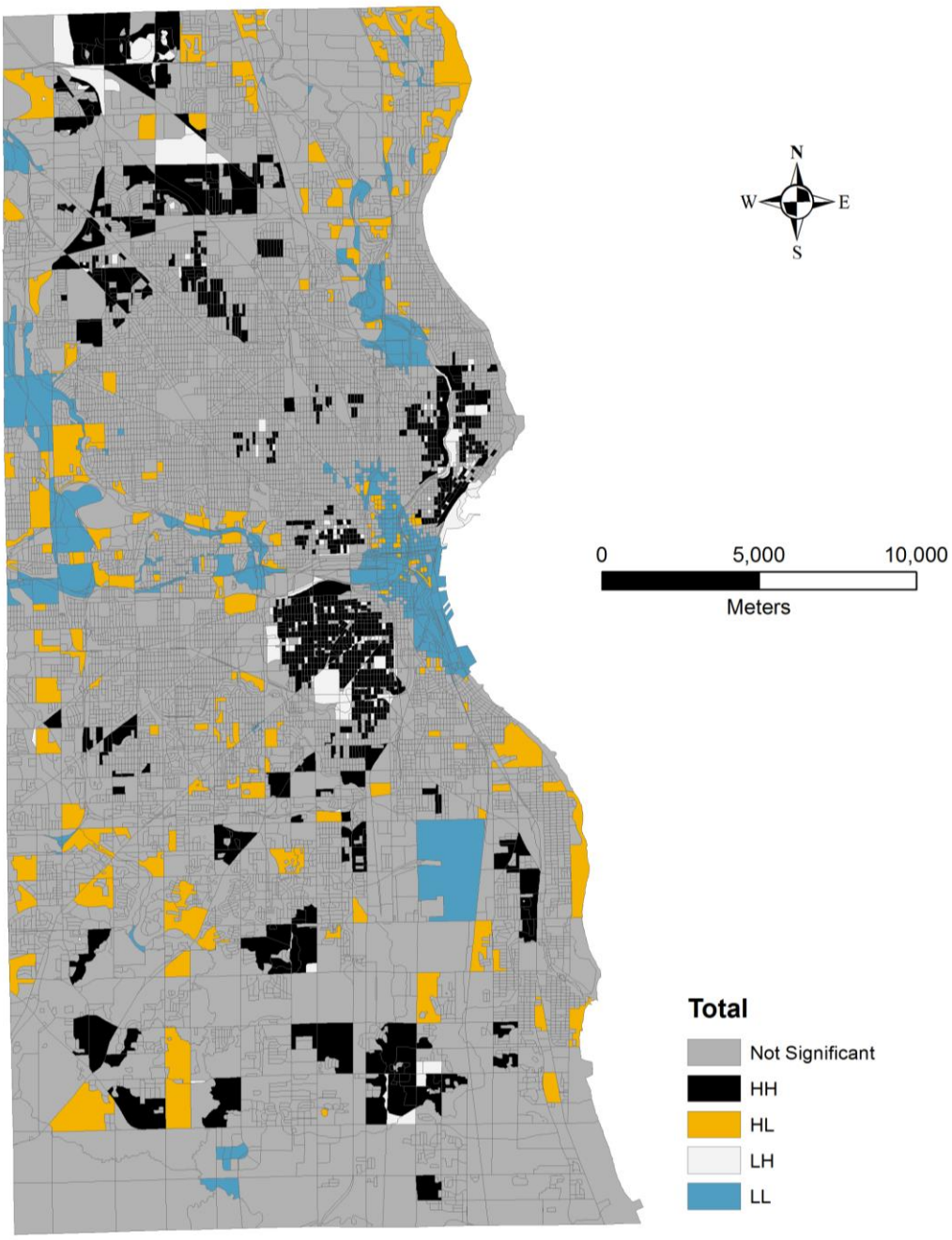


Figure 8. Cluster and outlier analysis (Anselin Local Moran's I) of total population.

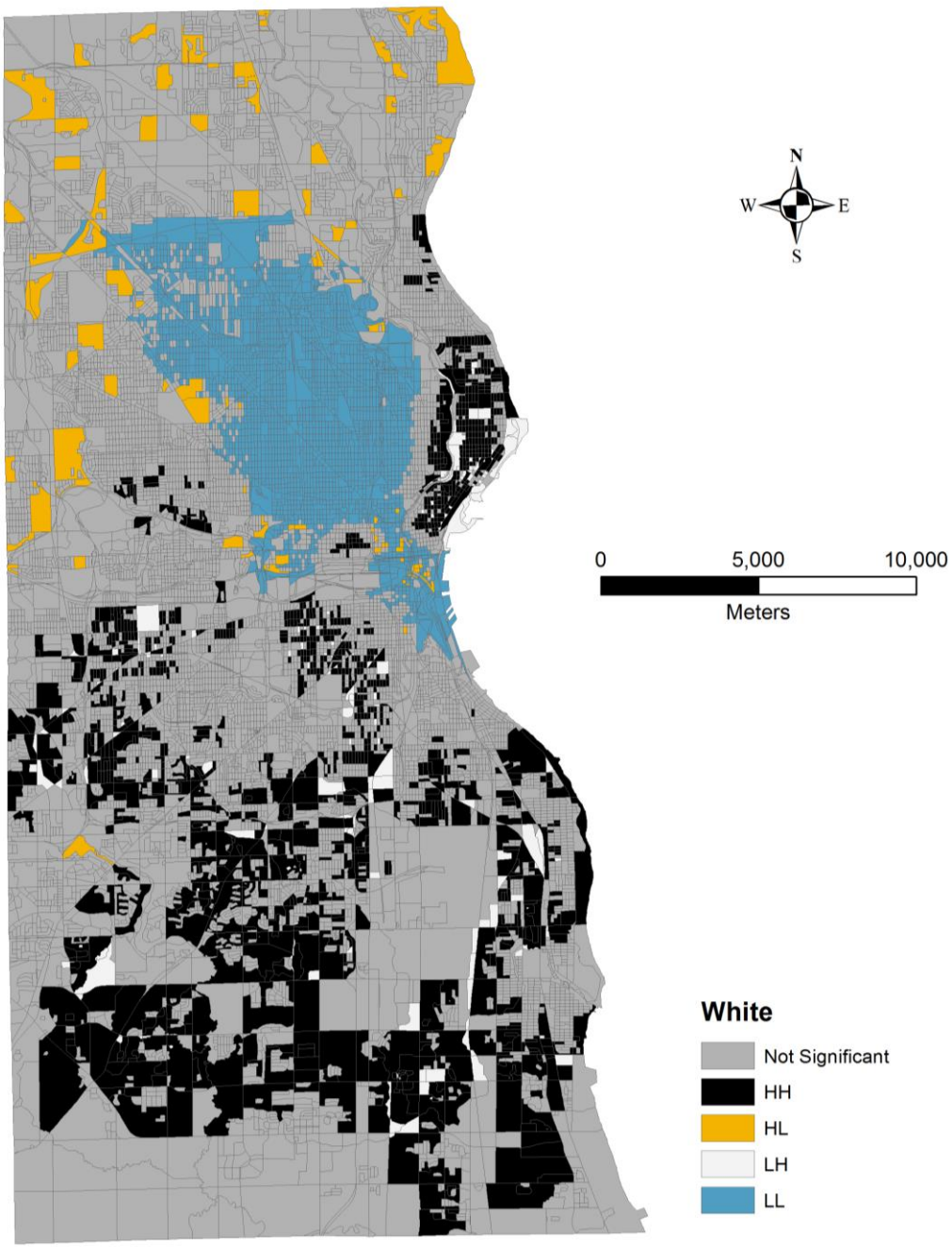


Figure 9. Cluster and outlier analysis (Anselin Local Moran's I) of white population.

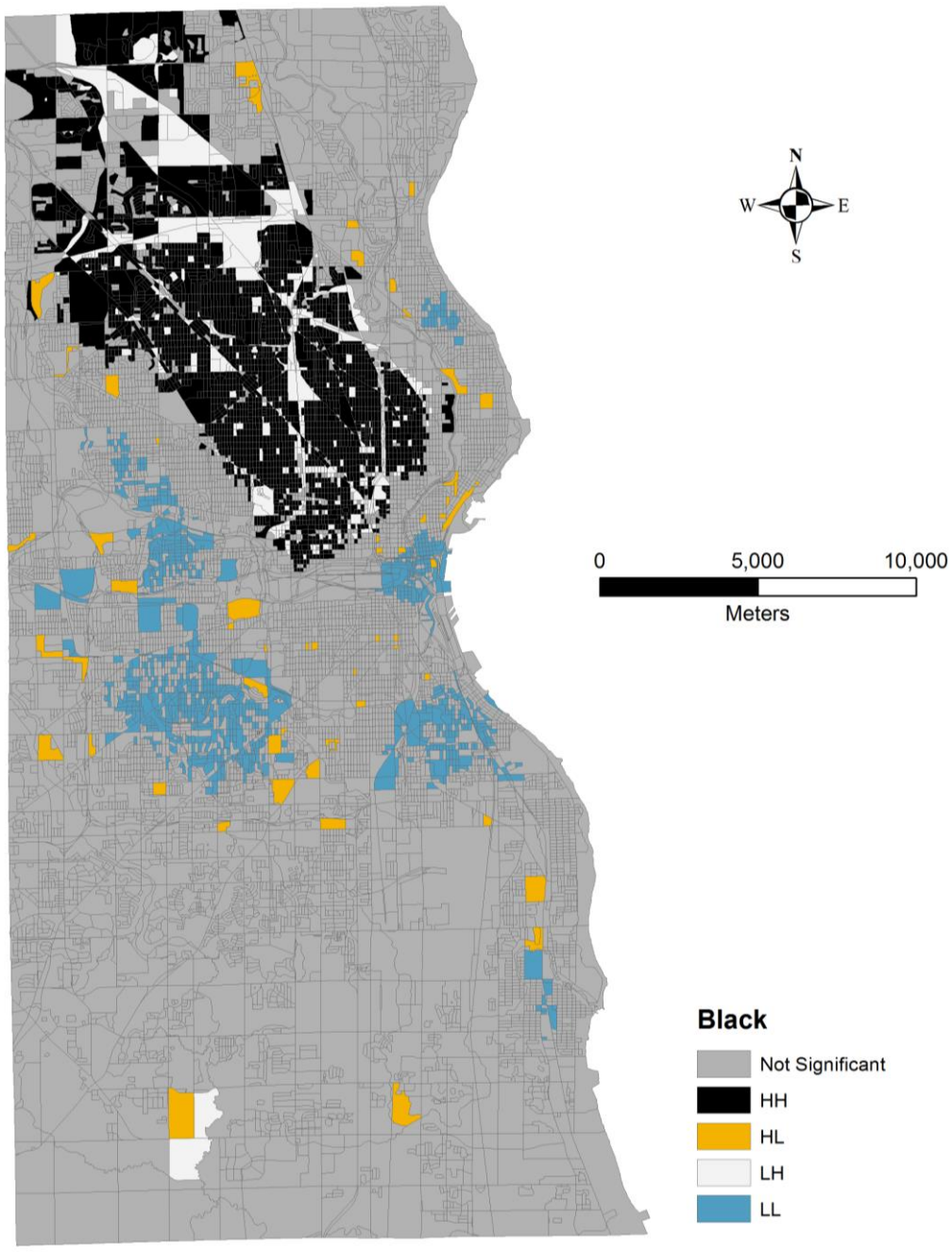


Figure 10. Cluster and outlier analysis (Anselin Local Moran's I) of black population.

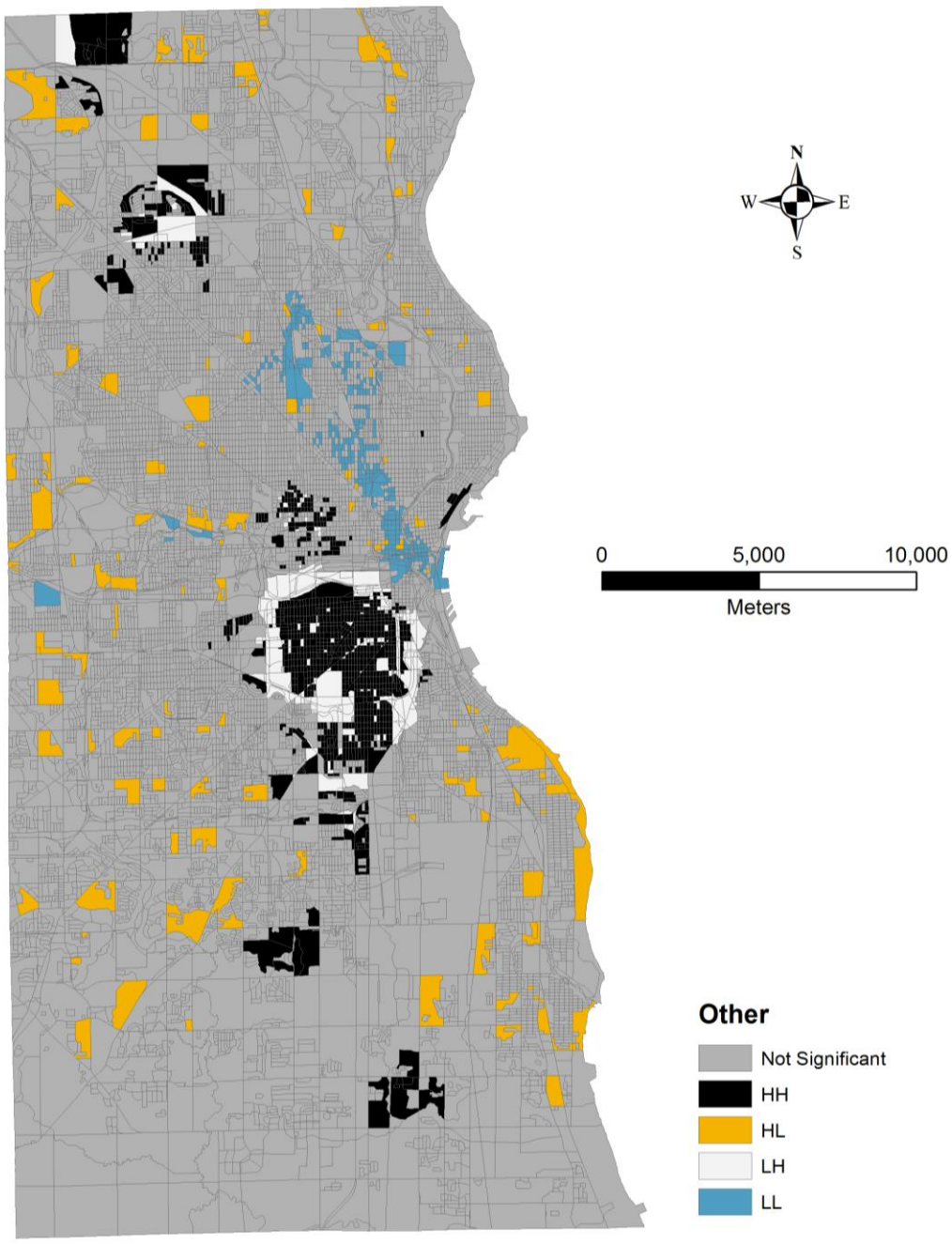


Figure 11. Cluster and outlier analysis (Anselin Local Moran's I) of other population.

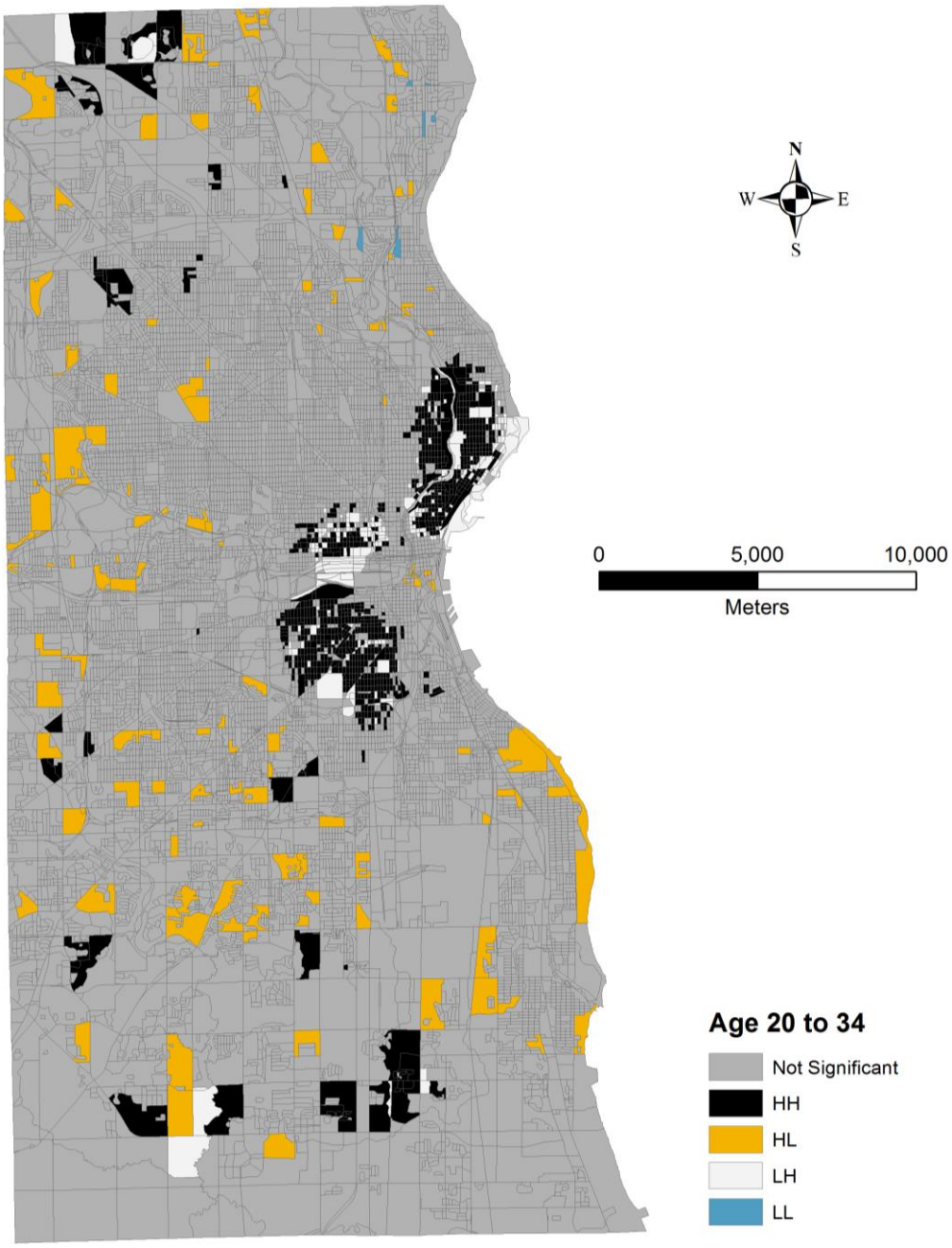


Figure 12. Cluster and outlier analysis (Anselin Local Moran's I) of population of age 20 to 34.

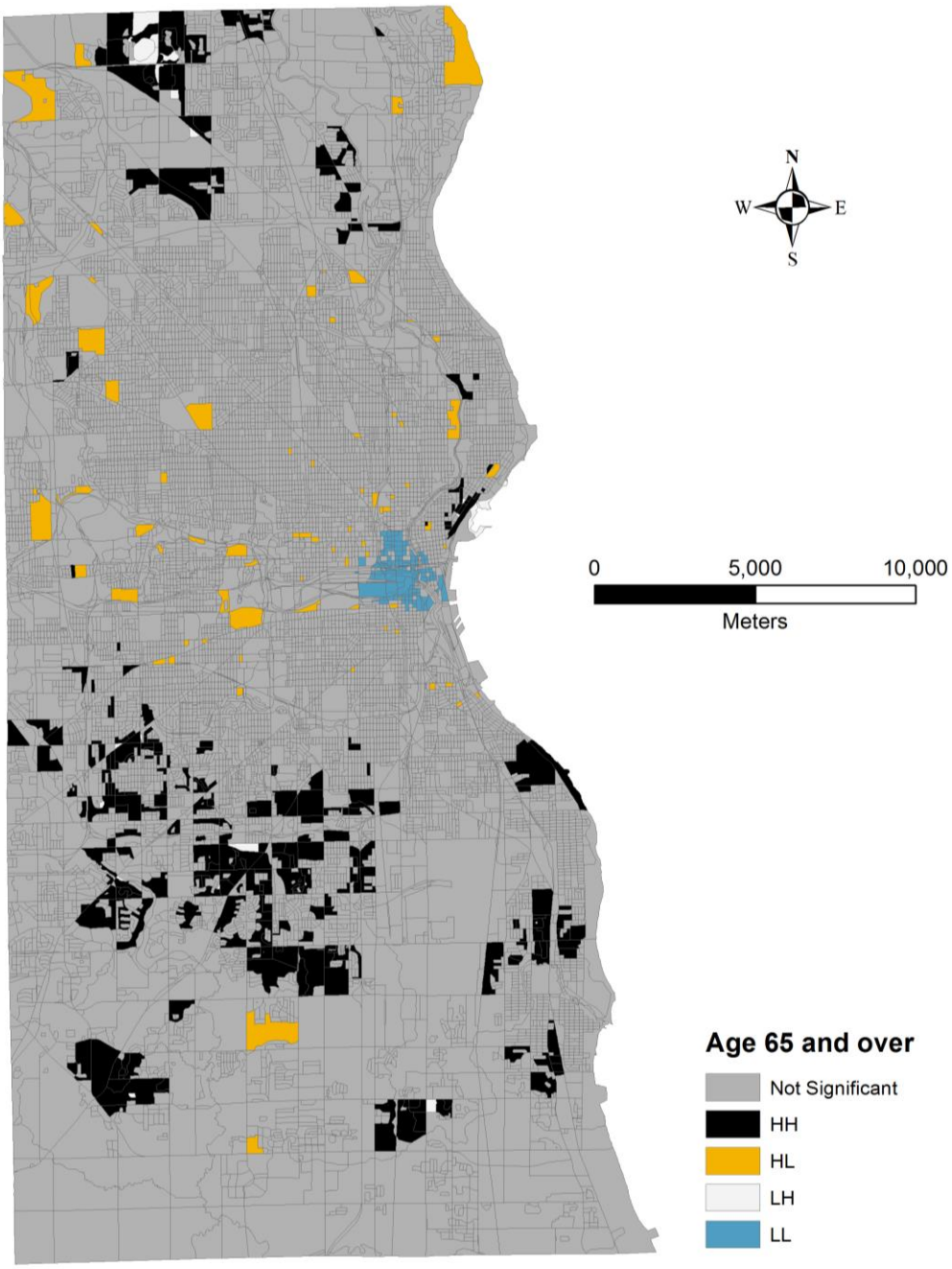


Figure 13. Cluster and outlier analysis (Anselin Local Moran's I) of population of age 65 and over.

3.2. Methodology

The method in this study consists of two steps. The first step copes with the estimation of total population count for each grid cell, while the second one is a process of "splitting" the total population counts into demographic groups. As discussed in the literature review, areal interpolation techniques are useful for estimating total population. Empirical studies have shown that dasymetric interpolation methods generally obtain better population estimation results (see Fisher and Langford, 1996; Eicher and Brewer, 2001; Mennis, 2003, 2009). In this study, "smart" interpolation method is chosen to make total population estimates. It is because, compared to traditional dasymetric interpolation method that mainly uses land use as ancillary data, "smart" interpolation is a multi-dimensional dasymetric interpolation method that takes advantage of a variety of ancillary data to approximate more realistic population distribution. In addition, this method has been previously employed to develop the widely used population grids (e.g. LandScan Global, LandScan USA). After the total population grid is interpolated, six kernel density surfaces, denoting the occurrence likelihood of total- and sub-populations, are generated. The ratios between cell values of sub-population density surfaces and total population density surface are used to apportion the total population residing in each grid cell into different demographic groups.

3.2.1. Total population estimation

The choice of the ancillary data (indicator variables) is critical in "smart" interpolation because it affects how the weights of grid cells will be assigned. In cases of developing generic interpolation models, the choice of variables should be relatively small

(Deichmann, 1996), but when interpolation is of a very high spatial resolution, more variables are preferred. In the production of LandScan Global (spatial resolution of 1km×1km), indicator variables include "road proximity, slope, land cover, and nighttime light" (Dobson et al., 2000, p.853). LandScan USA, a finer resolution (90m×90m) population grid, uses land cover, proximity to road networks, slope, landmarks, parks, schools, prisons, airports and water bodies as ancillary data. This study takes the similar approach as LandScan USA does with indicator variables being customized for the study area.

The first important variable included in the "smart" interpolation is land cover data. As briefly discussed in the literature review section, land cover types are proved to have strong correlation with population distribution. High-resolution (30m×30m) land cover data of Milwaukee County in 2006 is downloaded as a subset of USGS's NLCD (source: http://www.mrlc.gov/nlcd06_data.php). More updated land cover data can be acquired by classifying latest remote sensing imageries of the area. However, since errors will be inevitably introduced in the classification process, either by taking a supervised or unsupervised classification approach, I choose to use the latest land cover/land use data that are already released and verified by USGS for data accuracy and the ease of implementation. Fry et al. (2011) gives a detailed report on the completion of the 2006 NLCD. The original NLCD for Milwaukee County includes 15 land cover categories, including 4 urban land cover types (i.e., Developed Open Space, Developed Low Intensity, Developed Medium Intensity, and Developed High Intensity). Figure 14 illustrates the land cover data for Milwaukee County in 2006. Based on the original NLCD, the land use of Milwaukee County is reclassified into two types: inhabited and

uninhabited. The reason of choosing the binary dasymetric mapping over the 3-class dasymetric method and the limiting variable method is that there has been very little evidence that the latter two outperform the former (Kim and Yao, 2010; Longford, 2006, 2007; Qiu et al., 2012). Urban land use categories are assigned as the inhabited, while other land use categories are assigned as the uninhabited. No population will be assigned to the grid cells that are classified as uninhabited. Figure 15 shows the reclassified binary land use data for Milwaukee County.

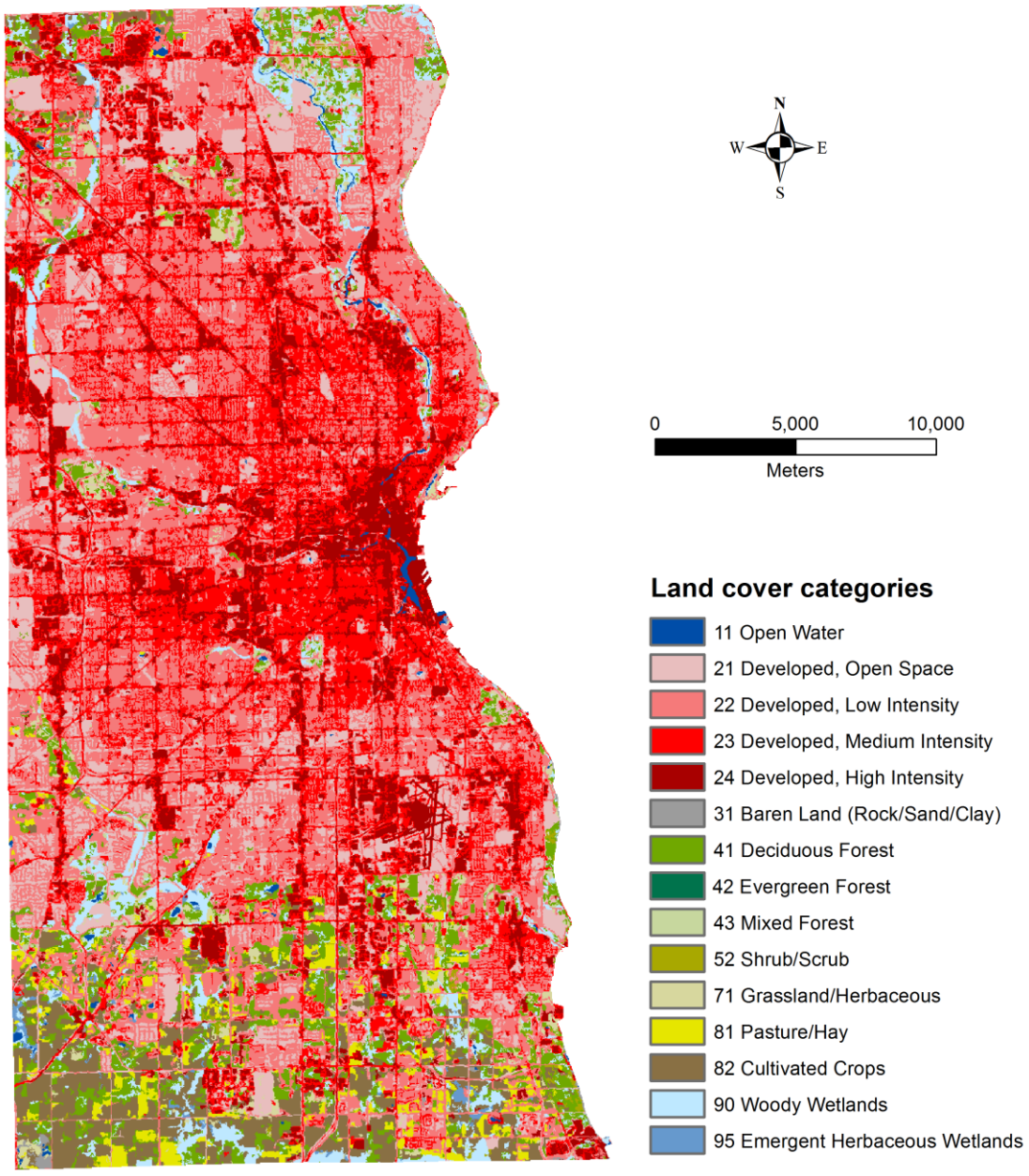


Figure 14. Original land cover data for Milwaukee County obtained from USGS's NLCD program, 2006.

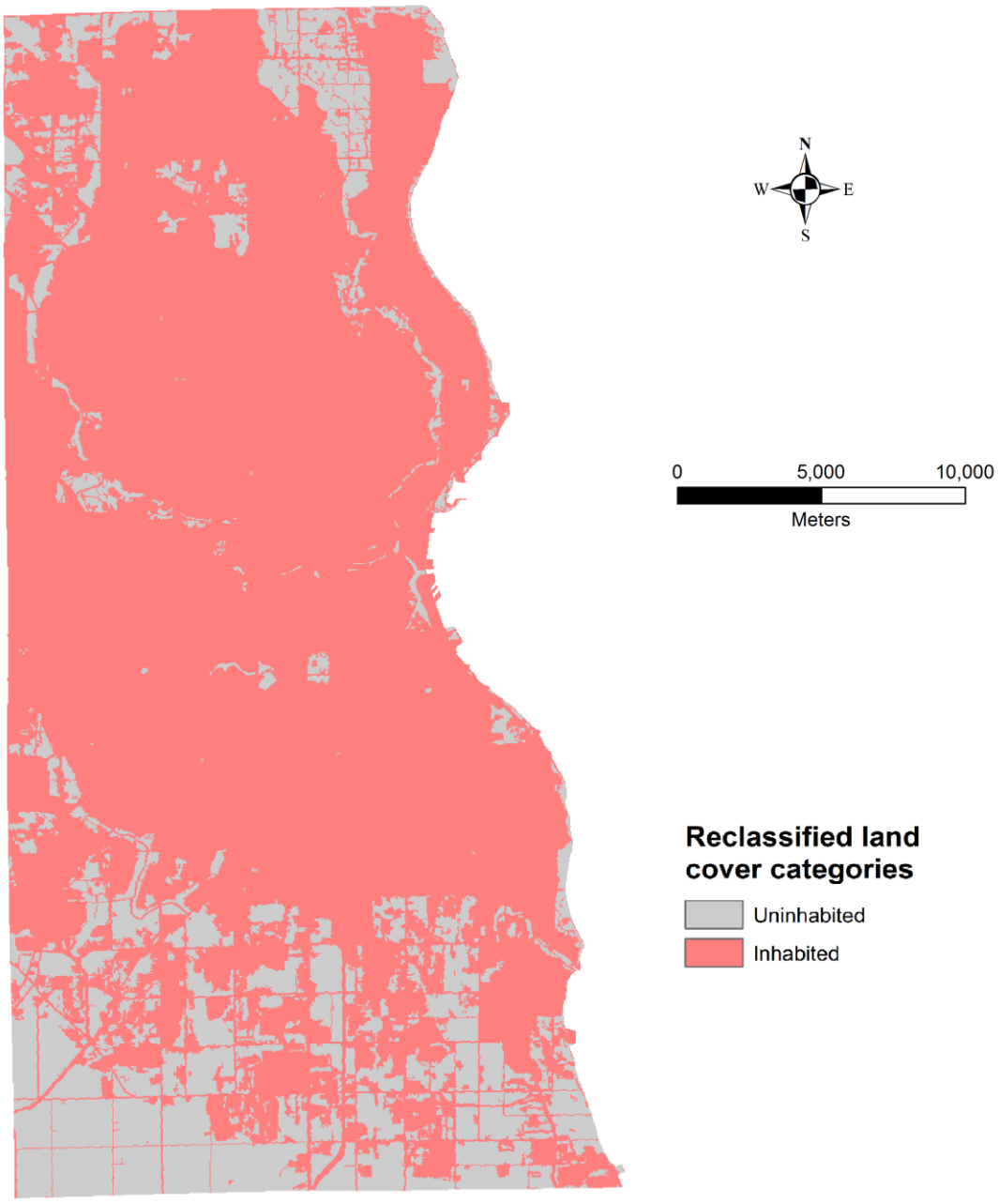


Figure 15. Reclassified land cover data using a binary classification scheme.

As a supplementation of the reclassified NLCD, other ancillary data also serve as exclusion zones to approximate the true population distribution. These datasets may include parks, water bodies, and census blocks with zero population (see Figure 16). In developing population grid by age and sex, Cai et al. (2006) used area landmarks (e.g., airports, parks, cemeteries, educational institutions, sports stadium, etc.) as part of the exclusion zones. In the production of LandScan USA, polygon landmarks are also employed as one of the indicator variables in the cell weighting process. In this study, by cross-examining the area landmarks in Milwaukee County and their spatially corresponding census blocks, it is found that many landmark polygons are populated. Some even have especially high population densities. Considering unpopulated landmark polygons are overlapped with zero population census blocks, the whole category of area landmarks is excluded in the weighting equation to improve the input data quality. Then, the exclusion zones will be spatially overlaid with the reclassified NLCD to obtain the refined binary land use data. Those grid cells that are classified as uninhabited land cover types are assigned zero weights. Similarly, grid cells that fall inside areas that are parks, water bodies and zero population census blocks are assigned zero weights as well; while urban land cover cells are assigned positive weights.

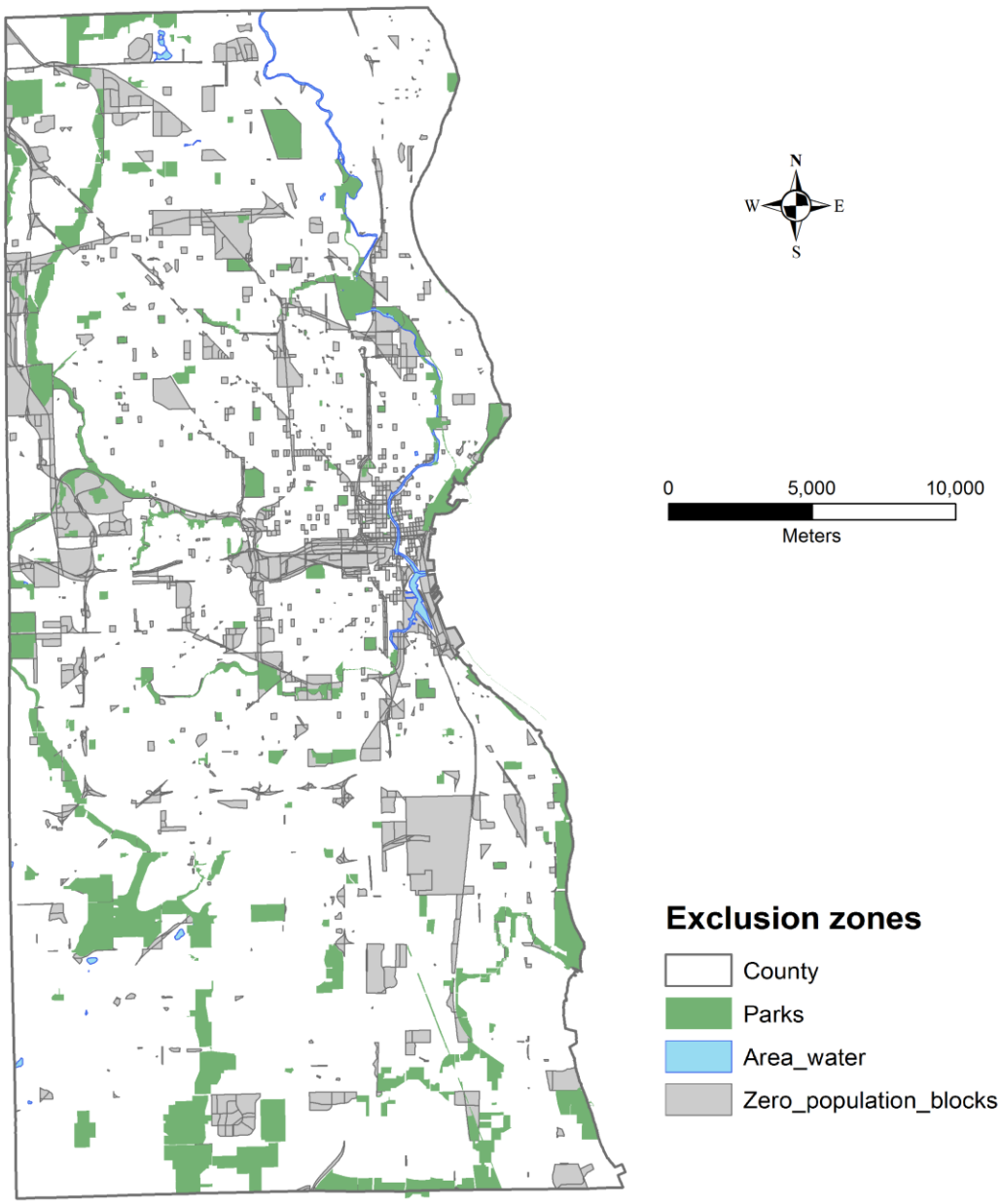


Figure 16. "Exclusion zones", including parks, water bodies, and census blocks with zero population.

In addition, proximity to road networks is also included in the weighting process. It has been proved that residential houses are often located along street segments (Xie, 1995, 2006). In a number of comparative analyses of areal interpolation methods, road network interpolation has generated the relative better interpolation results. One noteworthy issue here is that street segments of different hierarchies have different levels of correlation to residential densities. To be more specific, people are more likely to live along neighbourhood roads rather than highways. In this study, I exclude all U.S., interstate, state highways from the road network, and only the proximity to neighbourhood roads is considered as an input in weighting grid cells. The distribution of selected neighbourhood roads is shown in Figure 17. The proximity to road networks is measured by the distance between each grid cell to its nearest road segment. The distance can be acquired using the *Near Analysis* function in ArcGIS 10.1. To reduce the computation burden, the calculation is confined within a searching radius of 1000m around each grid cell.

Figure 17. Neighborhood roads.

The LandScan USA model suggests that slope angle is negatively correlated to population density. Common sense also suggests that people are more likely to build houses on a plain surface instead of a steep one. The elevation data (see Figure 18) for Milwaukee County in 2010 is downloaded from USGS's National Elevation Dataset (NED) (source: https://lta.cr.usgs.gov/NED) and transformed into a 30m×30m slope surface using the *Spatial Analysis* extension in ESRI's ArcGIS 10.1 (see Figure 19). The slope values will be included in the total population interpolation model as an important indicator variable. An issue appearing here is that 54,102 cells out of all 698,065 grid cells have slope value of 0 and that some highly populated blocks consist of only these cells. Such small values will result in a zero value of the range of relative weights of grid cells, and thus grid cells within these blocks will be allocated zero population. To address this problem, I selected all grid cells that have 0 slope value and assigned them with the value of 0.1. This approach ensures that the population of census blocks will be allocated to the inhabited grid cells with zero slope values.

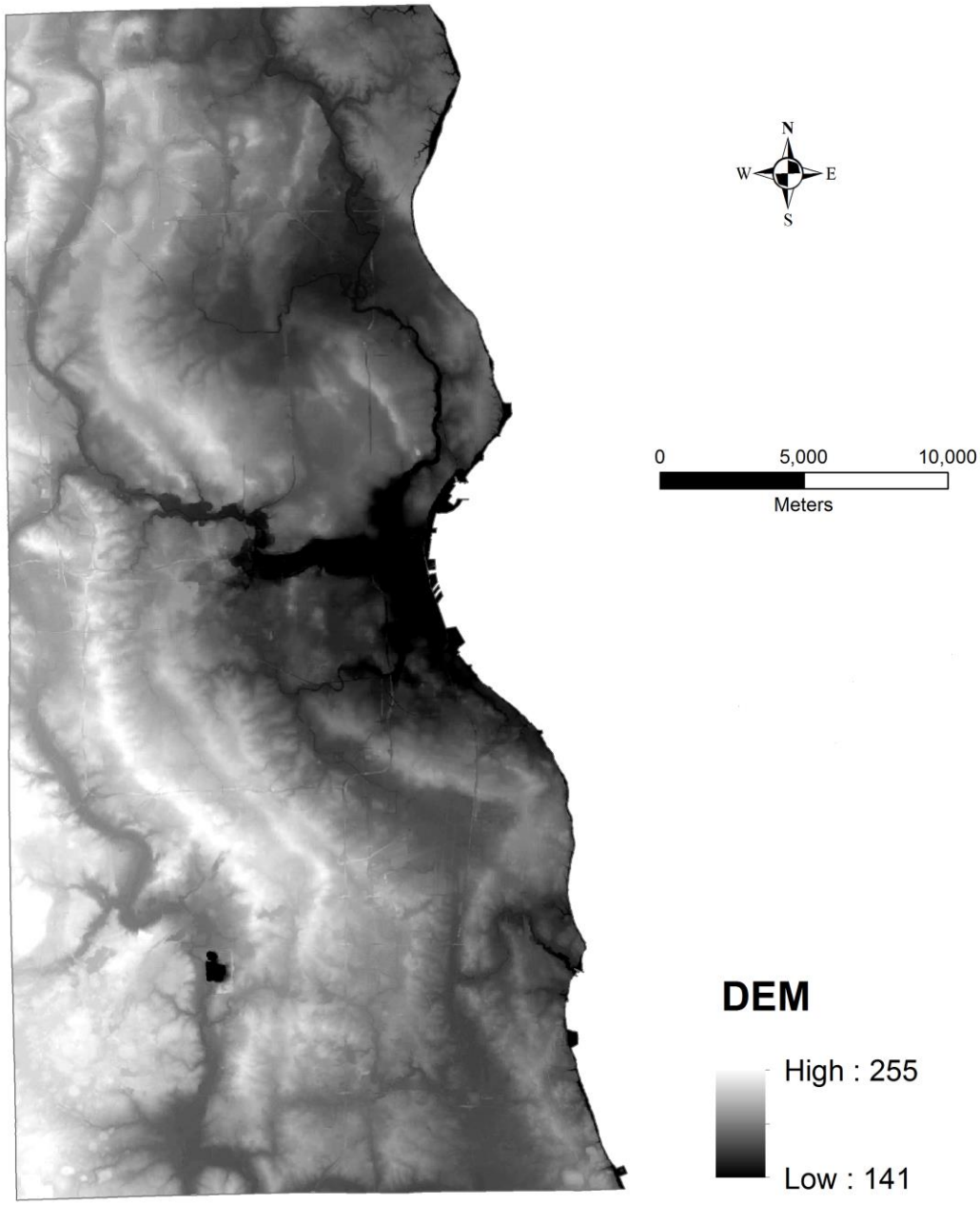


Figure 18. Digital Elevation Model (DEM) of Milwaukee County, 2010.

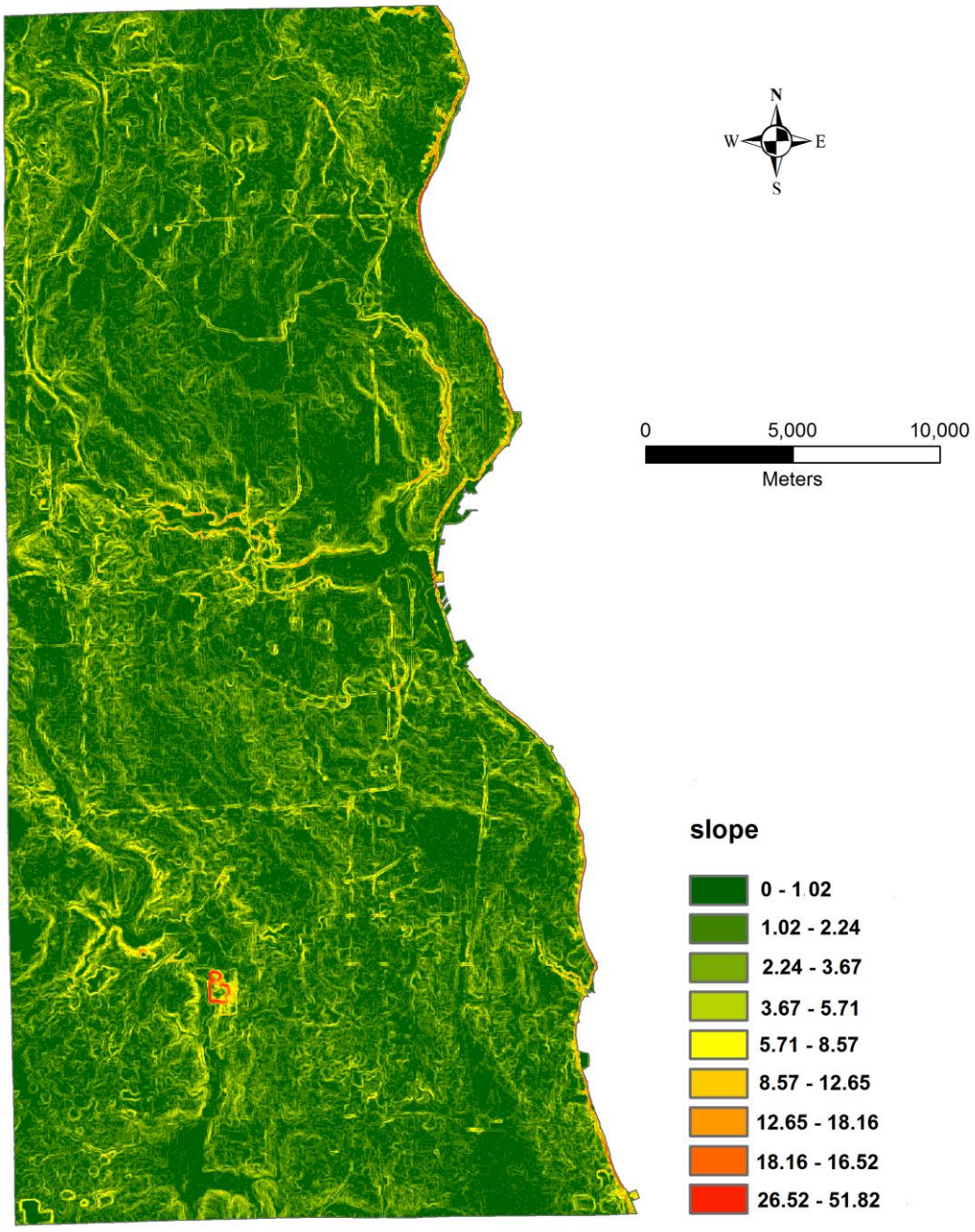


Figure 19. Slope surface generated from DEM.

Each ancillary data layer assigns grid cells with relative values, and all relative values of each cell are multiplied to obtain the cumulative value. The weights of grid cells that fall inside a census block will be determined by their cumulative values proportionally. The calculation of cumulative values is formulated as Equation 1:

$$V_i = LC_i \times RD_i \times ZP_i \times PK_i \times WT_i \times SL_i \quad (1)$$

where *i* is the number of grid cell within each census block; *V* is the cumulative value for the cell; *LC* is the value for land cover (1: Urban; 0: non-Urban); *RD* is the proximity to road network (the distance between the grid cell to nearest road segment); *ZP* is the value for zero population blocks (1: outside; 0: within); *PK* is the value for parks (1: outside; 0: within); *WT* is the value for water bodies (1: outside; 0: within); SL_i is the slope value at the location of grid cell *i*.

One of the issues is that the cell weights are not positively correlated with their cumulative values; indeed, grid cells with smaller cumulative values should be assigned with larger weights because common sense suggests that people are more likely to concentrate in places that are less steep and nearer to roads. A reverse normalization is necessary to assign cell weights correctly. Within each census block, the cumulative weight values of grid cells are normalized, as follows:

$$W_i = \begin{cases} \frac{V_{i/max} - V_i}{V_{i/max} - V_{i/min}} & \text{if } V_i \neq 0 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Where, within each census block, W_i is the normalized cumulative value of grid cell *i*; $V_{i/max}$ is the maximum cumulative value in the census block; V_i is the cumulative value of grid cell *i*; $V_{i/min}$ is the minimum cumulative value in the census block. Then, the

normalized cumulative values for individual grid cells that fall inside each census block are aggregated. Population value of each grid cell is calculated by multiplying census block population with the ratio of the cell's normalized cumulative value to the aggregated normalized cumulative values in that block, as follows:

$$P_i = P_{cbn} \times \frac{W_i}{\sum_1^m W_i} \quad (3)$$

where P_i is the population value of grid cell i; P_{cbn} is the population of n^{th} census block; W_i is the normalized cumulative value of grid cell i in n^{th} census block; m is the number of grid cells in n^{th} census block.

3.2.2. Sub-population estimation

The total population grid gives a high-resolution illustration of the residential pattern in the study area. However, this pattern might be different from the distribution of sub-population. The total population grid is estimated mainly by a density-based approach. This approach does not work for population of different racial and age groups, as some high population (total) density area may have low density of sub-populations. Citro (1998) emphasized that, with the progress in model developing and the boost of computation power, using statistical models would be of great potential in small area estimations. This study chooses to employ the point-based kernel density estimation to differentiate the distribution of different racial and age groups from the distribution of total population. Kernel density estimation is a commonly used non-parametric estimation method in statistics and it does not require any presumptive knowledge of the

probability density function (pdf) of the original data set (Zucchini et al., 2003). The kernel density function is an estimate of the probability of the spatial distribution. Silverman (1986, p.76) defines, for a two-dimensional data set X (x_1, x_2, …x_n), the kernel function $K_2(X)$ as:

$$K_2(X) = \begin{cases} 3\pi^{-1}(1 - X^T X)^2 & \text{if } X^T X < 1 \\ 0 & \text{othewise} \end{cases} \quad (4)$$

where the subscript of K denotes the number of dimensions of the data set. As a condition, this quadratic kernel density function satisfies:

$$\int_{R^2} K_2(X)\, dx = 1 \quad (5)$$

where R is the "window width" (Silverman, 1986, p.15) (or search radius) which defines the neighborhood around input data sets and the extent of smoothing iterations. The equation above implies that the probabilities of kernel densities sum up to 1 in the defined neighborhood. The output of the point-based *kernel density function* in ArcGIS 10.1 is a smoothly tapered surface generated based on the locations and values of original point features. In this study, representative point features are census block centroids with the attributes of total- and sub-populations. Six kernel density surfaces are created using the block centroids, the search radiuses, and the values of population counts as inputs. In the kernel density surfaces, for each kernel, cell values are the highest at the locations of the point features (block centroids) and decrease as the distances reach the search radius according to a distance-decay function (Bracken and Martin, 1989; ESRI, 2012a). The assumption that population distribution follows this distance-decay function may be easily subject to critics, but the kernel densities in this study are not directly transformed

into population quantities, instead, they are used to calculate the probabilities of sub-populations residing in each grid cell. Thus, the spatial interpolation of total population defines the distribution of population, while kernel density estimates obtain the proportions of sub-populations in a cell and define the distribution of sub-populations.

To calibrate the distance-based kernel density function in the sub-population estimation process, one should determine the appropriate width of the search window, which is related to the spatial autocorrelation in the population variables. If positive spatial autocorrelation exists, it indicates that people of the same demographic trait live nearer to each other and that the distance-decay function is conceptually valid. The results of Moran's *I* (Moran, 1950), one of the standard global spatial autocorrelation measurements, are calculated. The distance threshold used in calculating Moran's *I*s is the default distance that ensures every census block has at least one neighbor. Table 1 shows the Moran's *I* statistics for the total- and sub-populations. The table shows that the Moran's *I* values of total- and sub-populations and their Z-scores are all positive, and that their *p*-values are all statistically significant. It means that total- and sub-populations all have positive spatial autocorrelation. The table also shows that other population is the most spatially clustered while the population of age 65 and over has the relatively weakest spatial autocorrelation patterns.

Table 1. Moran's *I* statistics.

	Moran's I^{*}	Z-score[#]	p-value[+]
Total	0.075	65.955	0.000
White	0.137	121.326	0.000
Black	0.319	281.884	0.000
Other	0.446	392.968	0.000
Age 20 to 34	0.122	107.833	0.000
Age 65 and over	0.028	24.618	0.000

[*]Moran's *I*: +1means perfect correlation; -1means perfect dispersion; 0 means random spatial pattern.
[#]Z-score: using a 95% confidence level, Z-score that is larger than 1.96 or smaller than -1.96 indicates the existence of spatial autocorrelation. Positive Z-score means high and/or low values are more spatially clustered.
[+]*p*-value: <0.05, statistically significant.

Once the positive spatial autocorrelation is verified, the kernel density estimation can be executed on the populations. The appropriate width (*R*) of the smoothing window is essential in kernel density estimation (Silverman, 1986). Larger *R* in Equation 5 means more smoothing iterations and gets more generalized results, while smaller *R* means less smoothing iterations and gets more detailed estimation results. If the proper *R* is measured by the deviation between the aggregation of transformed population value from kernel density and the population value in each block, then the smaller the search radius the better. This is because, in kernel density estimation, populations generated by census block centroids with smaller "window width" are more likely to be constrained within the blocks. However, the problems with very small window width are that population kernel densities may be overly high in the areas around census block centroids, and that they cannot reflect the influence of neighboring census blocks.

The semivariogram, which quantifies the intensity of spatial autocorrelation based on values and distances of spatial features, is used to determine the proper *R* for total-and sub-population kernel density estimations (ESRI, 2012b). The semivariogram function is formulated as follows.

$$\gamma(S_i, S_j) = \frac{var(Z(S_i) - Z(S_j))}{2} \quad (6)$$

where S_i and S_j are two locations, $Z(S_i)$ and $Z(S_j)$ are the values of the two locations, $\gamma(S_i, S_j)$ is the semivariogram of the two locations, and *var* is the variance.

The semivariogram function plots the relationship between the distance (x-axis) and the variance of the value differences (y-axis) between all pairs of locations. If this relationship is plotted with a curve fitting the empirical data, the curve will rise as distance increases but level off at some point. The distance threshold at which a semivariogram model levels off is called the "range." What it means is that the intensity of spatial autocorrelation does not have significant difference any more once the distance between two locations goes beyond the threshold. The "range" determined by the semivariogram model is commonly used to find out the proper "window width" or "search radius" in spatial interpolations, such as *Kriging* and *Inverse Distance Weighting* (IDW) (Gotway et al., 1996; Ver Hoef et al., 2001).

Using the semivariogram function to decide the proper *R* in the kernel density estimation consists of three steps: (1) creating empirical semivariogram; (2) fitting a model to the empirical semivariogram; and (3) finding the threshold of distance beyond which the semivariogram values remain unchanged. In this study, there are 13,179 centroids generated from census blocks. The pairs of centroids are "binned" into groups

according to their distance and direction. The averages of distances and variances of centroid pairs in each bin will be treated as one single point in the empirical semivariogram plot. How the original centroid pairs are binned will have a great effect on the empirical semivariogram and thus the fitting of the semivariogram model. Two parameters need to be determined. The first one is the lag size, which determines the level at which the spatial autocorrelation will be analyzed. A bigger lag size means more centroid pairs grouped in each bin and less points in the semivariogram plot. An overly large lag size may filter out the spatial autocorrelation among the centroids that have short distances. When lag size is too small, it may result in the situation that sample centroids in the bins are too few to generate representative averages (ESRI, 2012b). The second parameter is the number of lags. As long as the lag size is restricted, the number of lags should be determined by spatial extent of original data to which the semivariogram is calibrated (Coombes, 2002). The default rule on how the number of lags should be selected is that the value of lag size multiplying the number of lags should be less than half of the longest distance between point pairs (ESRI, 2012c). Also, one can adjust the values and visually examine the fitness of the semivariogram model in many software.

To decide the proper lag size for the semivariogram, the *Average Nearest Neighbor* tool in ArcGIS 10.1 is used to calculate the average distance of closest centroid points, and the value (122.59m) is selected as the reasonable lag size. This is the default approach to calculate lag size in ArcGIS 10.1, and the method ensures that there are at least several points being grouped in each bin. In the original population dataset, there are six different population values, total- and five sub-populations, of each point. It is easily

conceivable that the spatial autocorrelations of total- and sub-populations should be different. For each of the six populations, with the fixed lag size, I experimented with the number of lags ranging from 10 to 25, and plotted the Root Mean Square Error (RMSE) generated in the cross-validation step. Figure 20 shows the RMSE of total- and sub-populations as the number of lags rises from 10 to 25. As one can see, six RMSE curves have different patterns, but each of them has a lowest point at which the RMSE is the smallest. For total and white population, the optimal number of lags is 20; for black population, the number of lags is 16; for other population, the number of lags is 13; for population of age 20 to 34, the number of lags is 22; and for population of age 65 and over, the number of lags is 20. Figure 21 shows the empirical semivariogram models with the lag size and the number of lags determined.

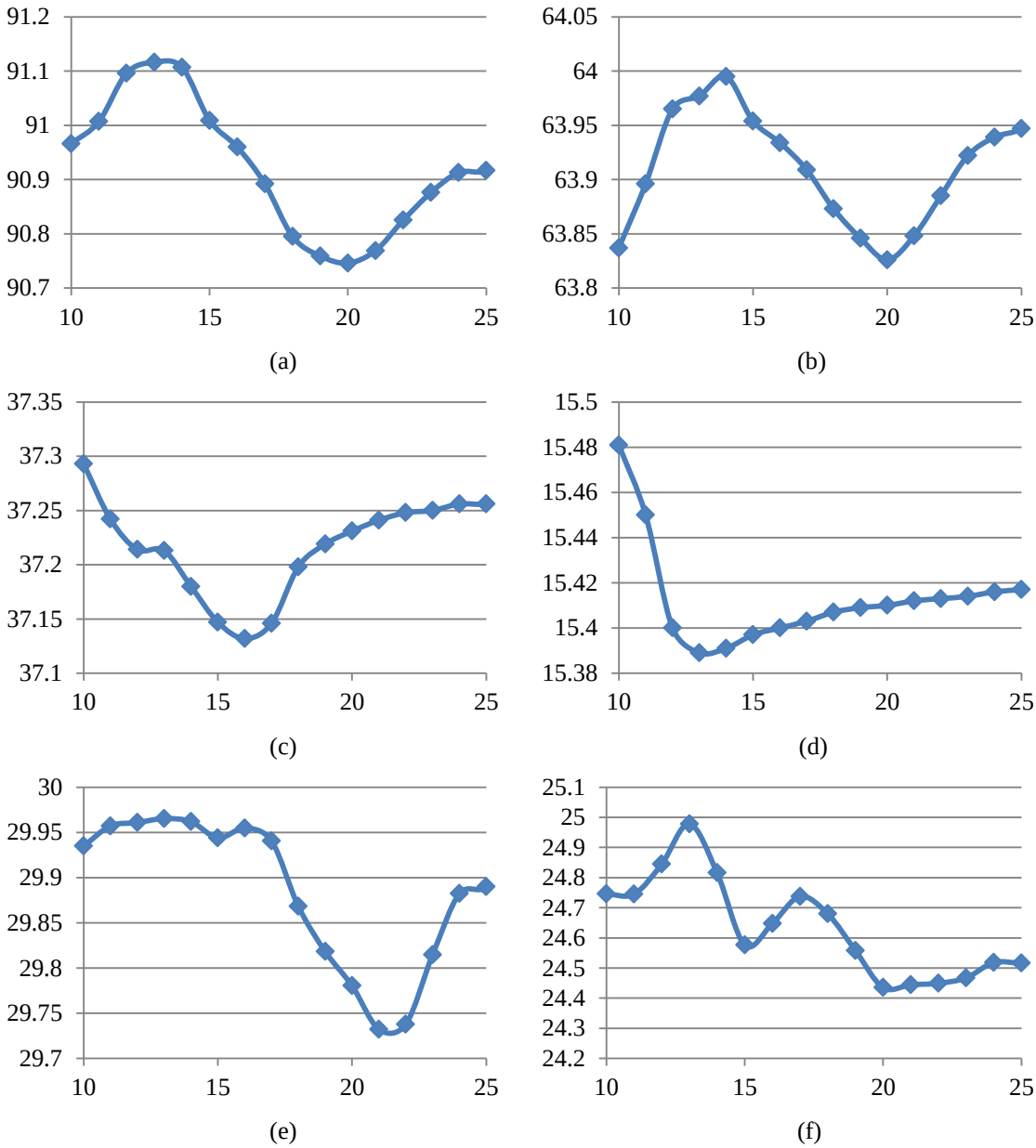


Figure 20. Plots of RMSE and the number of lags, with the fixed lag size of 122.59m: (a) Total; (b) White; (c) Black; (d) Other; (e) Age 20 to 34; (f) Age 65 and over.

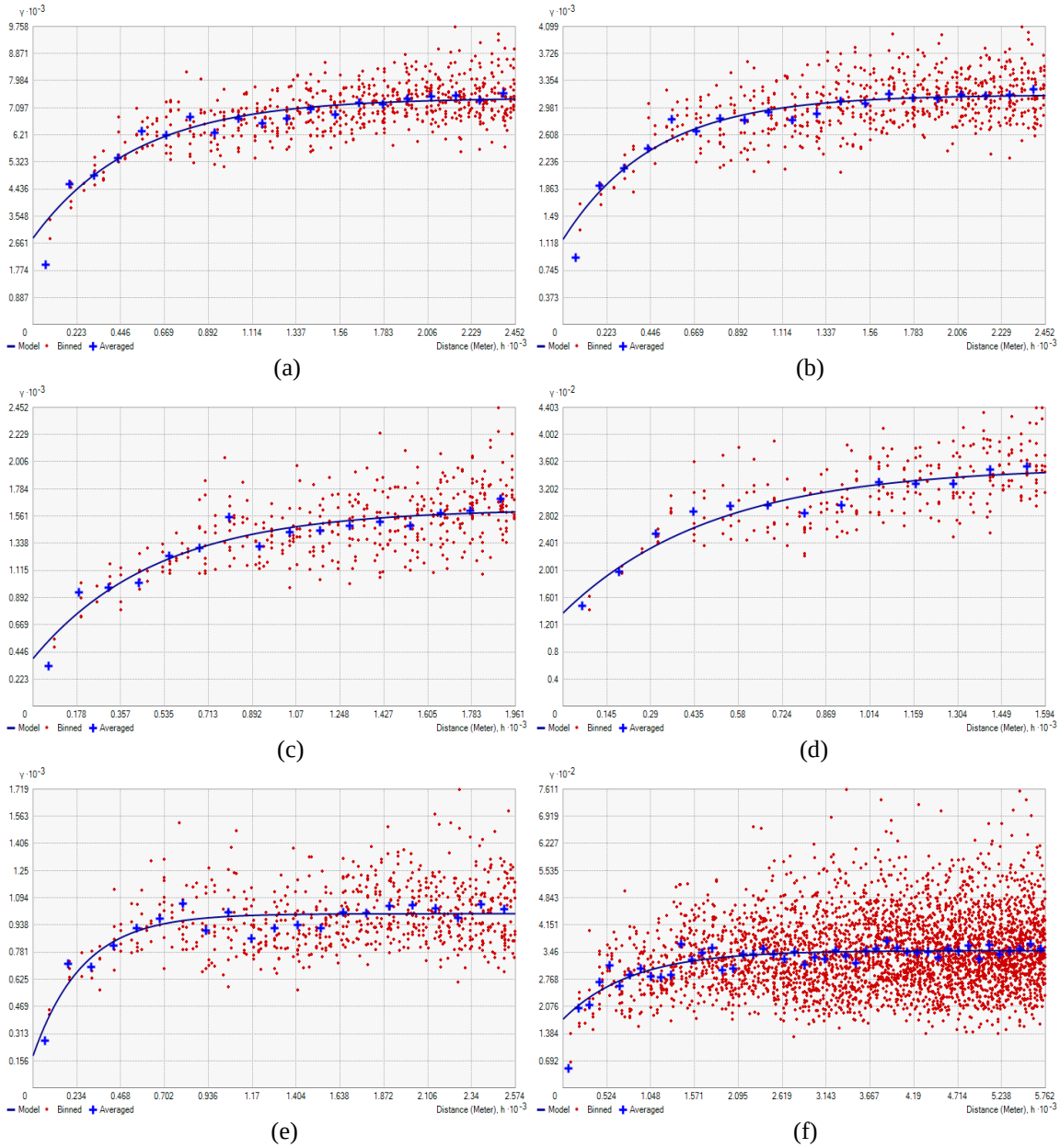


Figure 21. Empirical semivariogram models: (a) Total; (b) White; (c) Black; (d) Other; (e) Age 20 to 34; (f) Age 65 and over.

Thus far, the two important parameters in modeling the empirical semivariogram are determined. By using total- and sub-populations of census block centroids as source inputs and setting the values of lag size and the optimal number of lags accordingly, six different semivariogram models are established as shown in Figure 21. The major ranges of the four models are 1556.977m, 1389.894m, 1507.336m, 1486.206m, 791.14m and 1419.794m, for total, white, black, other population, population of age 20 to 34, and population of age 65 and over, respectively. These values will be used as the value of the "window width" or search radius parameter in the kernel density estimation. Figure 22 shows the kernel density surfaces created from the total- and sub-population values of centroids and their corresponding search radiuses. These kernel density surfaces are in the raster format and have the spatial resolution of 30m×30m. Since the display unit of the map in Figure 1 is meters, by default, the unit of kernel density surfaces is persons per square kilometers. Therefore, each grid cell in the six kernel density surfaces will have the density values of total- and sub-populations.

Sub-populations of each grid cell will be calculated by multiplying the total population in the cell by the ratio of the kernel densities between the sub-population and total population. This can be achieved by using the *raster calculator* tool in ArcGIS 10.1.

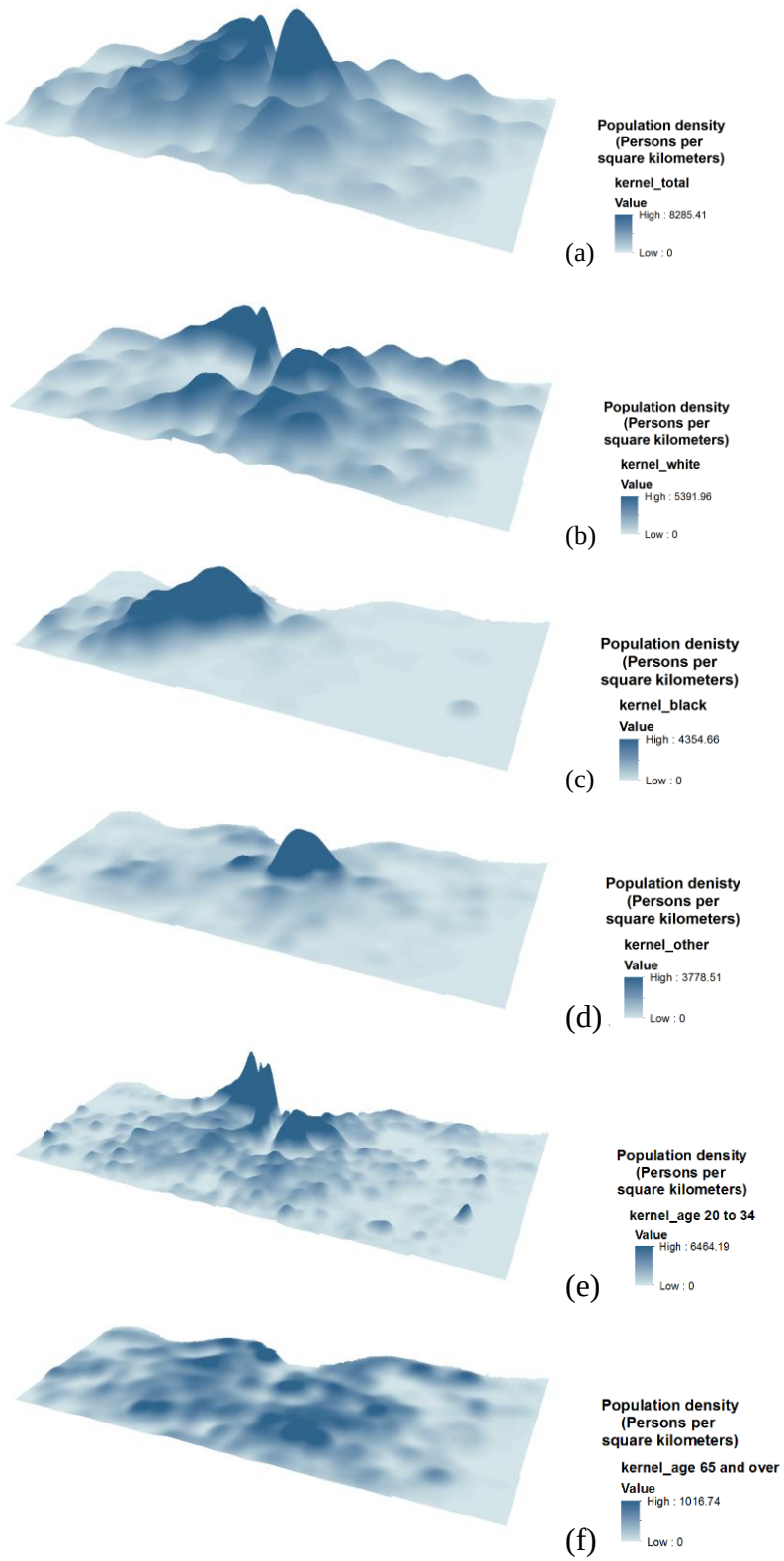


Figure 22. Kernel density surfaces: (a) Total; (b) White; (c) Black; (d) Other; (e) Age 20 to 34; (f) Age 65 and over.

3.2.3. Accuracy assessment

Bhaduri et al. (2007) asserted that it was impractical to validate the LandScan model at grid cell level due to the constraints of research resources and privacy issues. In the meantime, they admitted that human intervention in the modeling process, for instance, the sub-categorizing of each indicator variable and weight assigning to those sub-categories, heavily influenced the model fidelity, which also imposed challenges to result validation. They took following approaches to verify and validate the model. First, to ensure the input data quality, they used census data and residential land use data classified from remote sensing imageries to investigate possible anomalies of population density. Non-typical residential arrangements such as prisons, schools which usually have high population densities are visually verified by high spatial resolution orthophotographs. Second, population disaggregation in the model is controlled at the census block level, which ensures that population grid avoids any irregularities at higher spatial resolutions. Third, LandScan USA data was originally produced at the 30m×30m resolution and aggregated to the 90m×90m resolution as final output. The process of data disaggregation and aggregation may introduce what they called "mensuration errors." The paper used regression analysis between LandScan USA data and block populations to test the model's spatial integrity. Two counties, one urban and the other rural, were chosen as representative enumeration units. Regression results indicated that there was a high correlation between LandScan USA and census data. Finally, the assessment of "location errors" of the LandScan USA model was carried out by geocoding a number of house locations and using previous version of LandScan USA data to validate whether the pixels were assigned inhabited or uninhabited correctly.

In this study, the "mensuration errors" of total- and sub-population grids are introduced not only by the disaggregation process but also by the kernel density estimation. The errors are evaluated by cross-validation with census block populations. The distribution of "mensuration errors" is mapped and census blocks that have highest overestimations and underestimations are especially investigated. The assessment of "location errors" of total population can take the same approach as LandScan USA does, which is randomly sampling pixels and using other data sources such as classified remote sensing images to validate whether the pixels are assigned inhabited or uninhabited correctly. However, the interpolation of total population in this study already takes advantage of land use data. Uninhabited pixels are not allocated with populations, which makes the location error assessment of total population estimates unnecessary. The location errors of sub-populations, however, are impossible since there is no ground information of sub-populations' locations for validation.

4. Results

The final products of this study include six population grids, which are for total-, white-, black-, other-population, population of age 20 to 34 and population of age 65 and over, respectively (see Figure 23-28). These grids are in a 30m×30m raster format and are independent from any zonal system used by the Census Bureau or other government agencies. As shown in the figure, pixels that are in darker red have more estimated population while pixels that are in lighter red have less. Figure 23 shows that total-population values of grid cells have a range of [0, 114.445] and that total-population

values are relatively higher in the areas that are close to downtown Milwaukee while decrease towards the North, South, and West sides. Figure 24 shows that white-population values of grid cells have a range of [0, 90.2018] and that white-population values share similar patterns with total-population except that values are low in the area that is northwest side of downtown. Figure 25 shows that black-population values of grid cells have a range of [0, 60.0791] and that black-population highly concentrates in the northwest side of downtown where white-population values are relatively low. Figure 26 shows that other-population values of grid cells have a range of [0, 22.8377] and that other-population concentrates in the geographic center of the county. Figure 27 shows that population of age 20 to 34 concentrates in the downtown area, which can be explained by the large number of apartment complexes around the University of Wisconsin-Milwaukee and Marquette University. Figure 28 shows the population of age 65 and over is spatially dispersed. An interesting phenomenon is that Milwaukee River vividly separates white and black population and that other population mostly concentrates in the area encompassed by Menomonee River and Kinnickinnic River. These grids give detailed population statistics at a spatial resolution that is much higher than the census blocks.

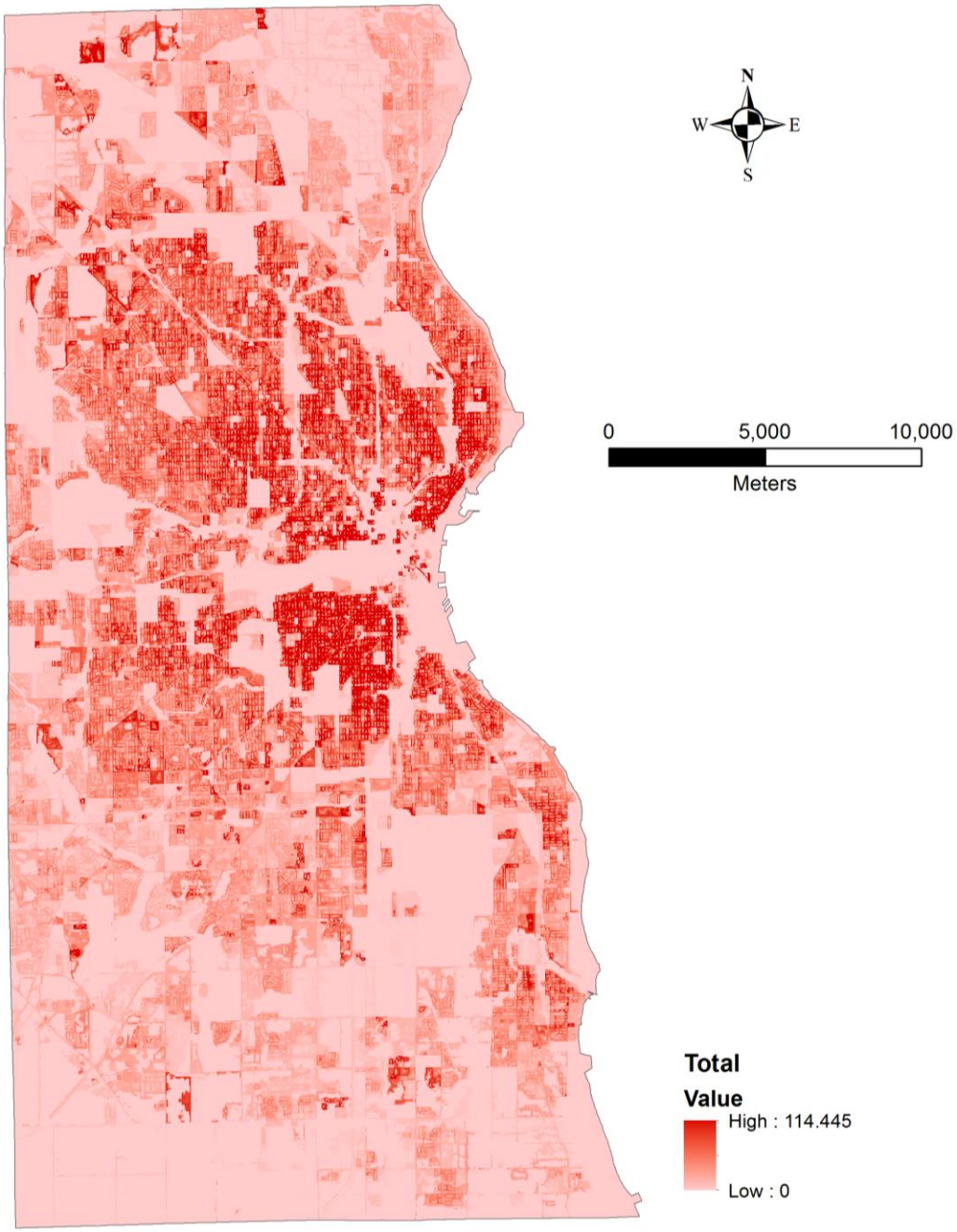


Figure 23. Grid of total population.

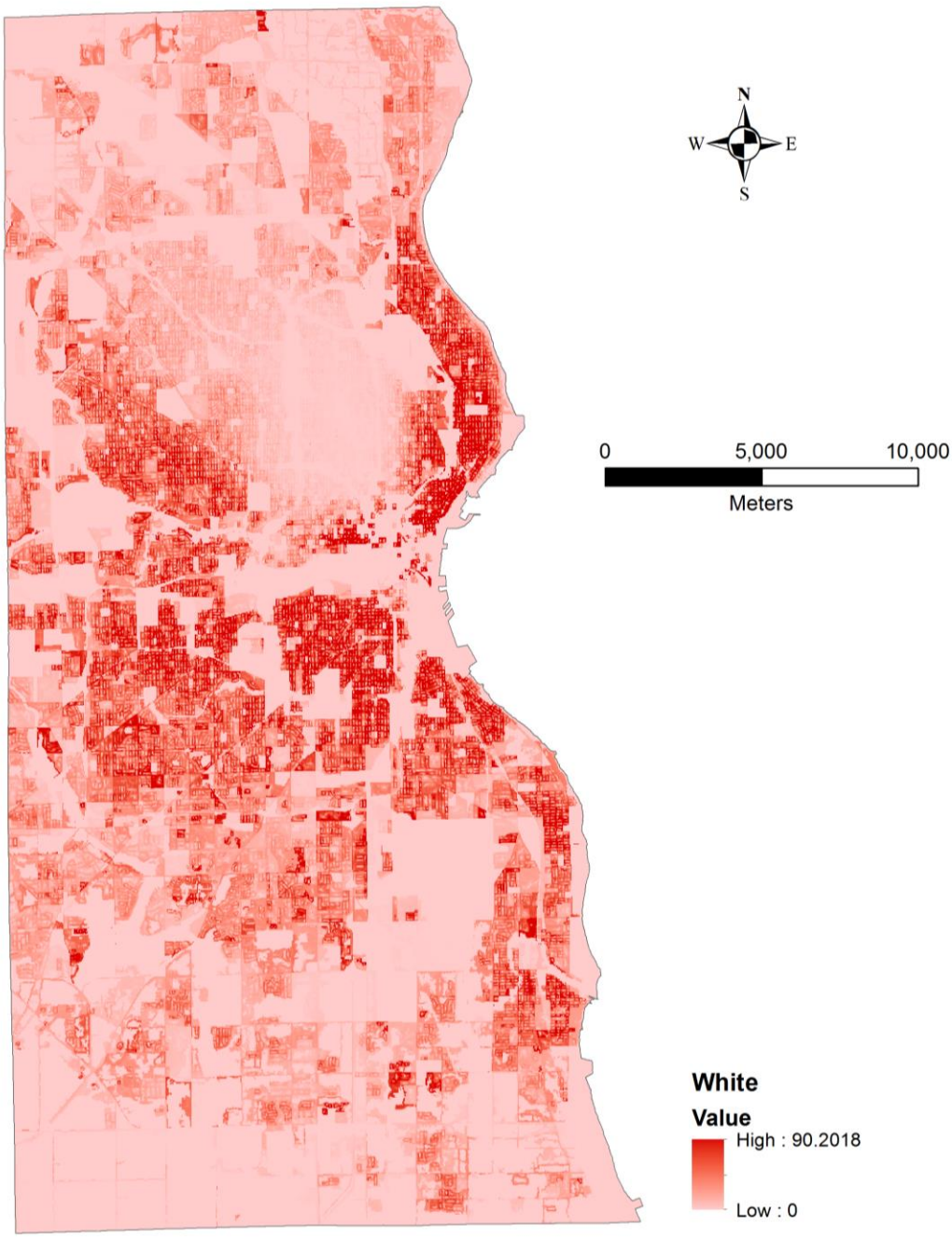


Figure 24. Grid of white population.

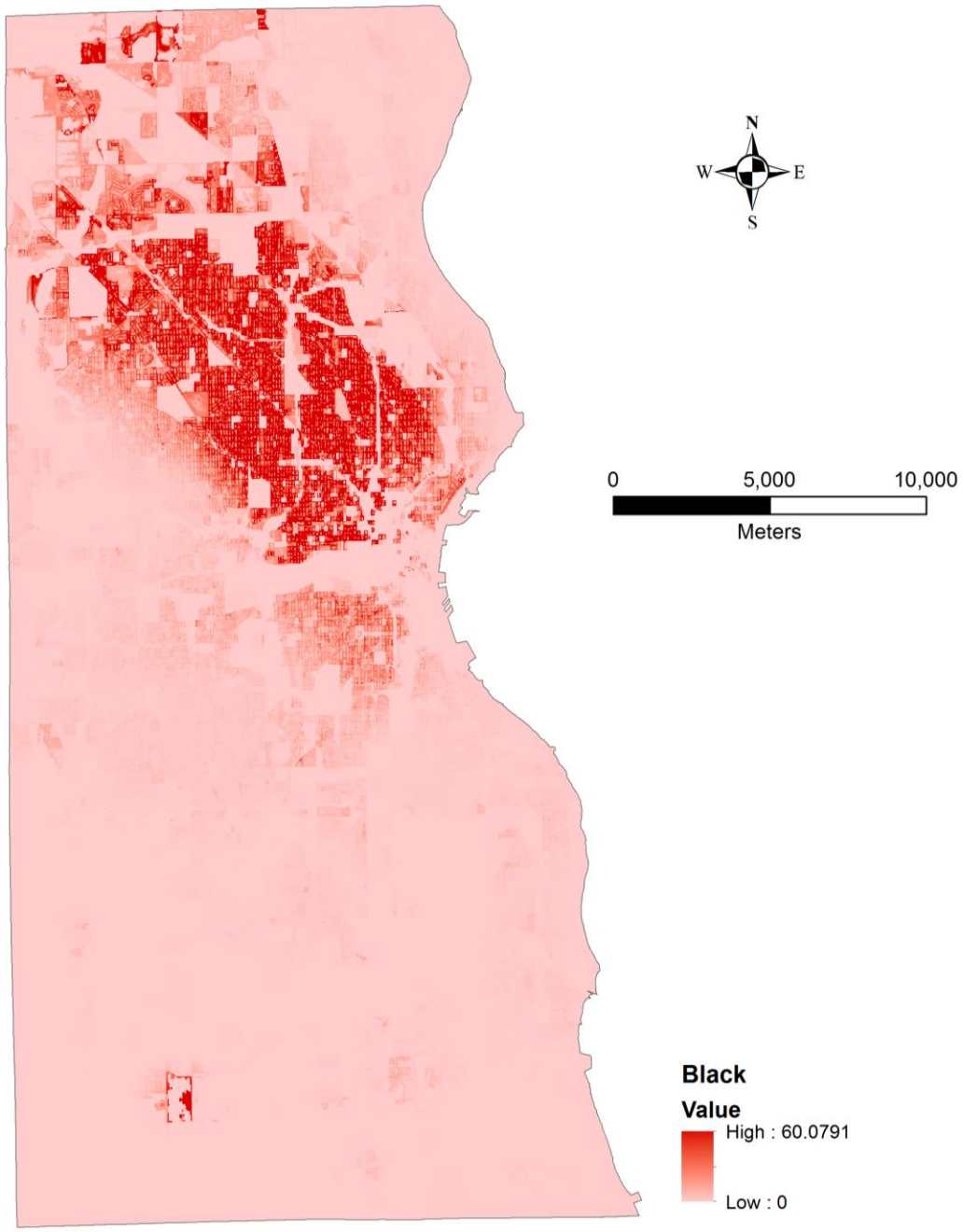


Figure 25. Grid of black population.

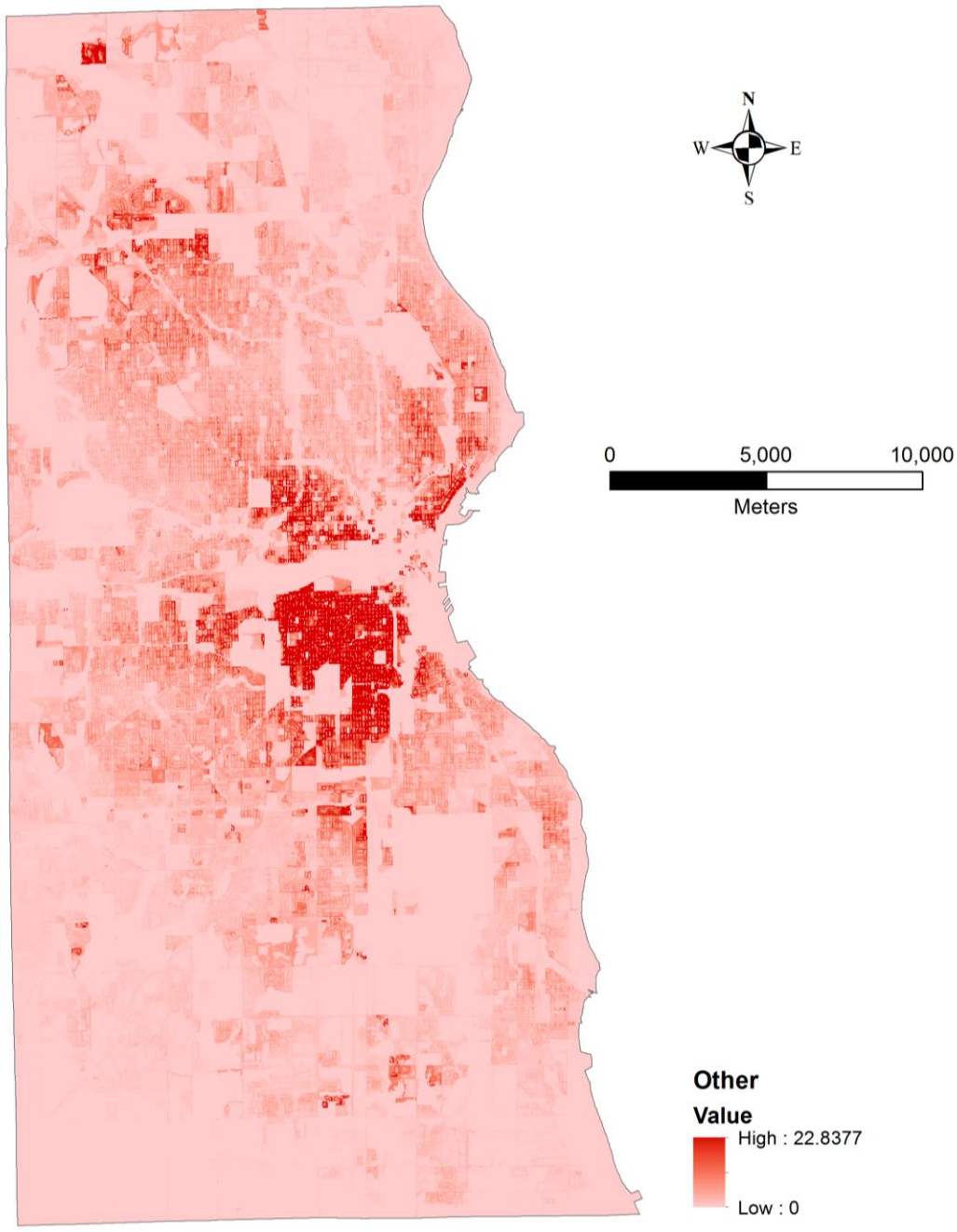


Figure 26. Grid of other population.

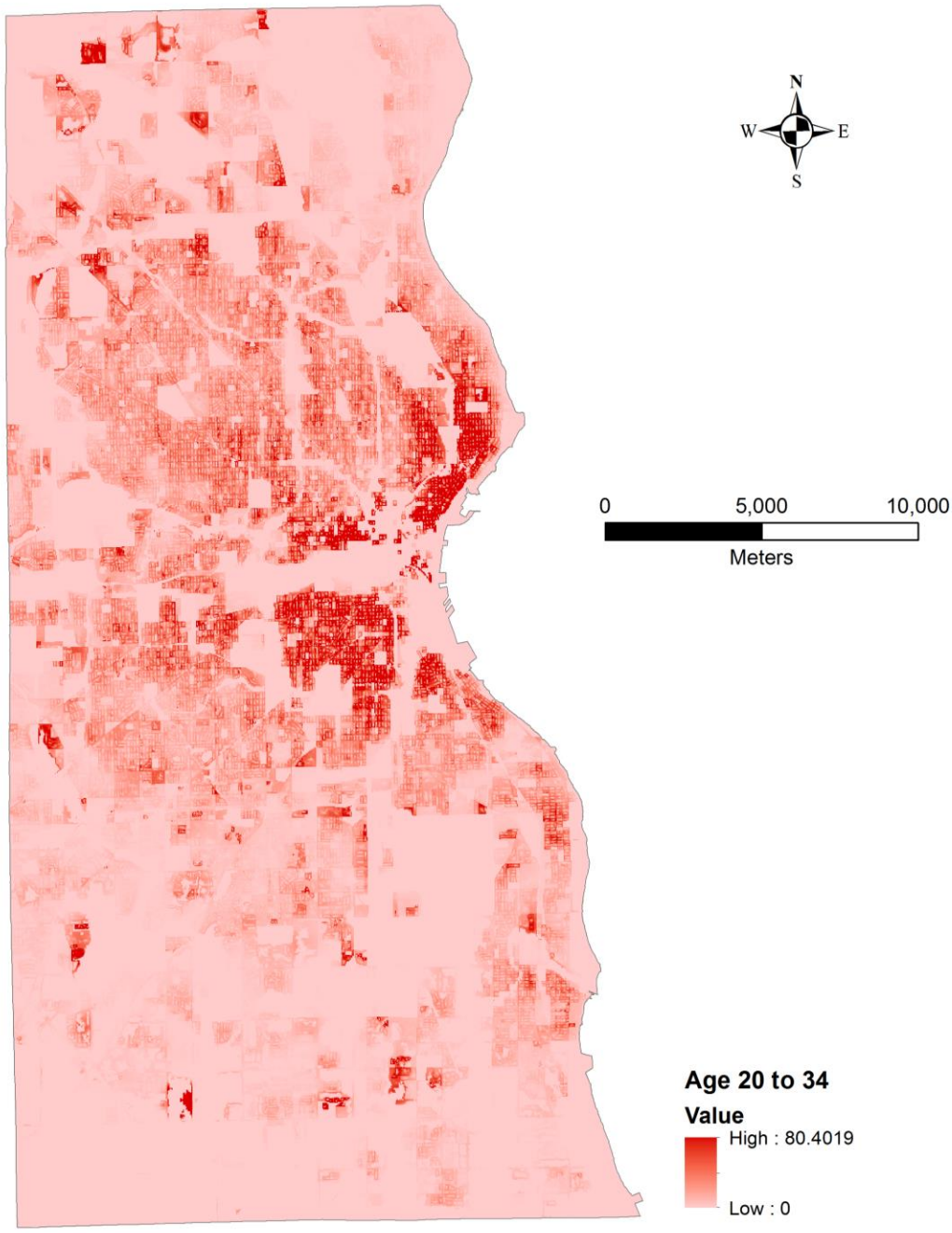


Figure 27. Grid of population of age 20 to 34.

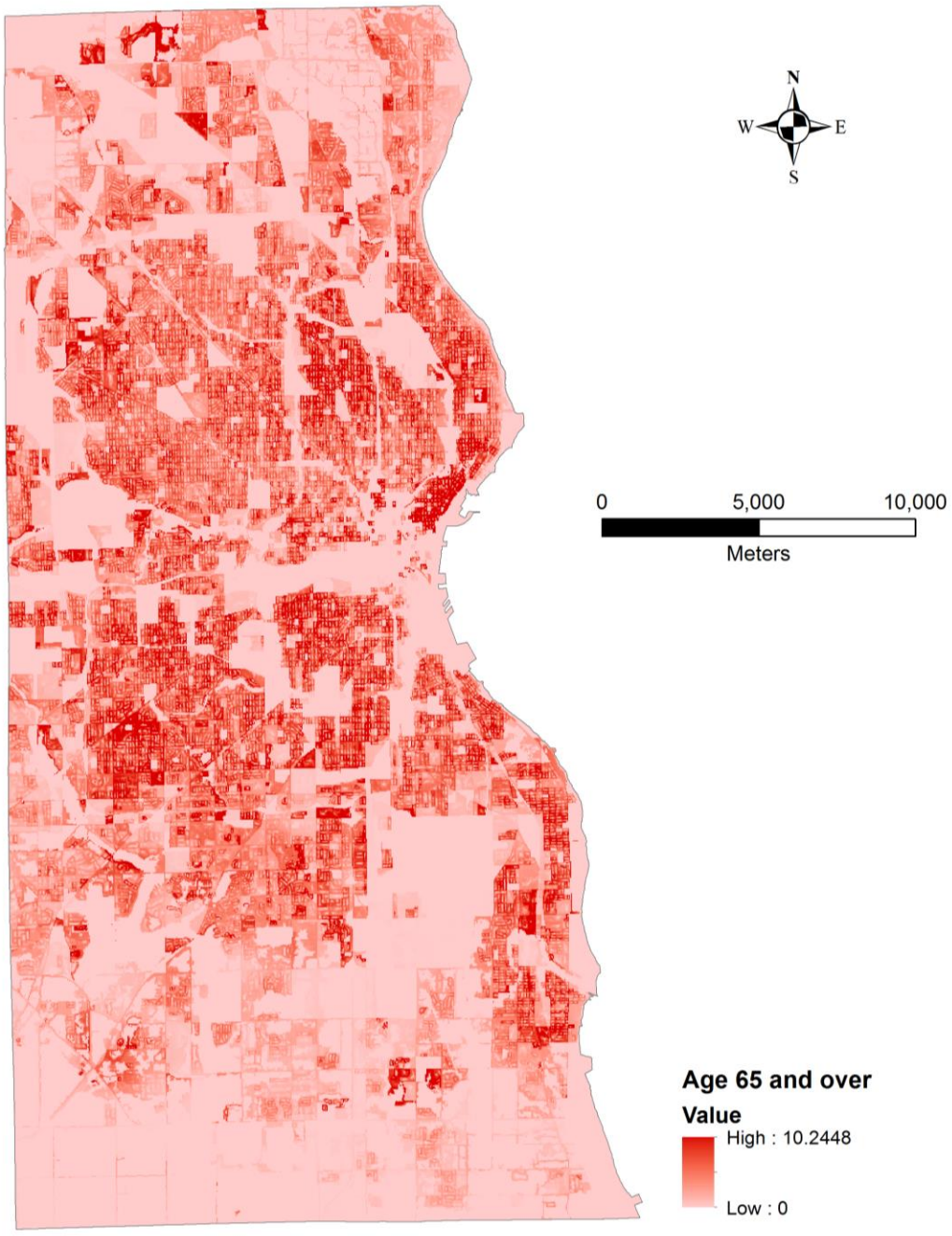


Figure 28. Grid of population of age 65 and over.

5. Discussion

5.1. The assessment of population grids

The accuracy assessment of the population grids cannot be done at the grid cell level as it is impossible, or at least time and labor consuming, to get the *in-situ* population value of every grid cell (Bhaduri et al., 2007). One of the alternatives is to aggregate the values of population grids into a zonal system, and to compare the aggregated values to the true values in the zonal system. In this study, the spatial interpolations of total population are constrained to satisfy the condition that the aggregation of total population of grid cells within each census block is equal to the total population of the block. Thus, the interpolations of total population are absolutely dependent on block boundaries. However, in estimating sub-population in grid cells, the parameters of the kernel density approach are census block sub-populations, the locations of block centroids and the search radiuses. Therefore, after the kernel density to population quantity transformation, the aggregations of grid cell sub-populations in census blocks are not controlled to equal block sub-populations. In light of this understanding, the assessment of sub-population estimation results can still be carried out at the level of census blocks. Using the *zonal statistics* tool in ArcGIS 10.1, one can obtain the sum of grid cell total- and sub-population values within each census block. Then, these aggregated values will be extracted and joined to the blocks in order to assess the estimation accuracy.

5.1.1. Population grids of race

5.1.1.1. Scatterplots and regression statistics

Figure 29 plots the aggregations of total- and sub-population within each census block and the block population, with the linear regression lines plotted through the origins. The X-axis in each plot denotes the census block population while the Y-axis denotes the aggregation of estimated grid cell values. Each red dot represents a census block.

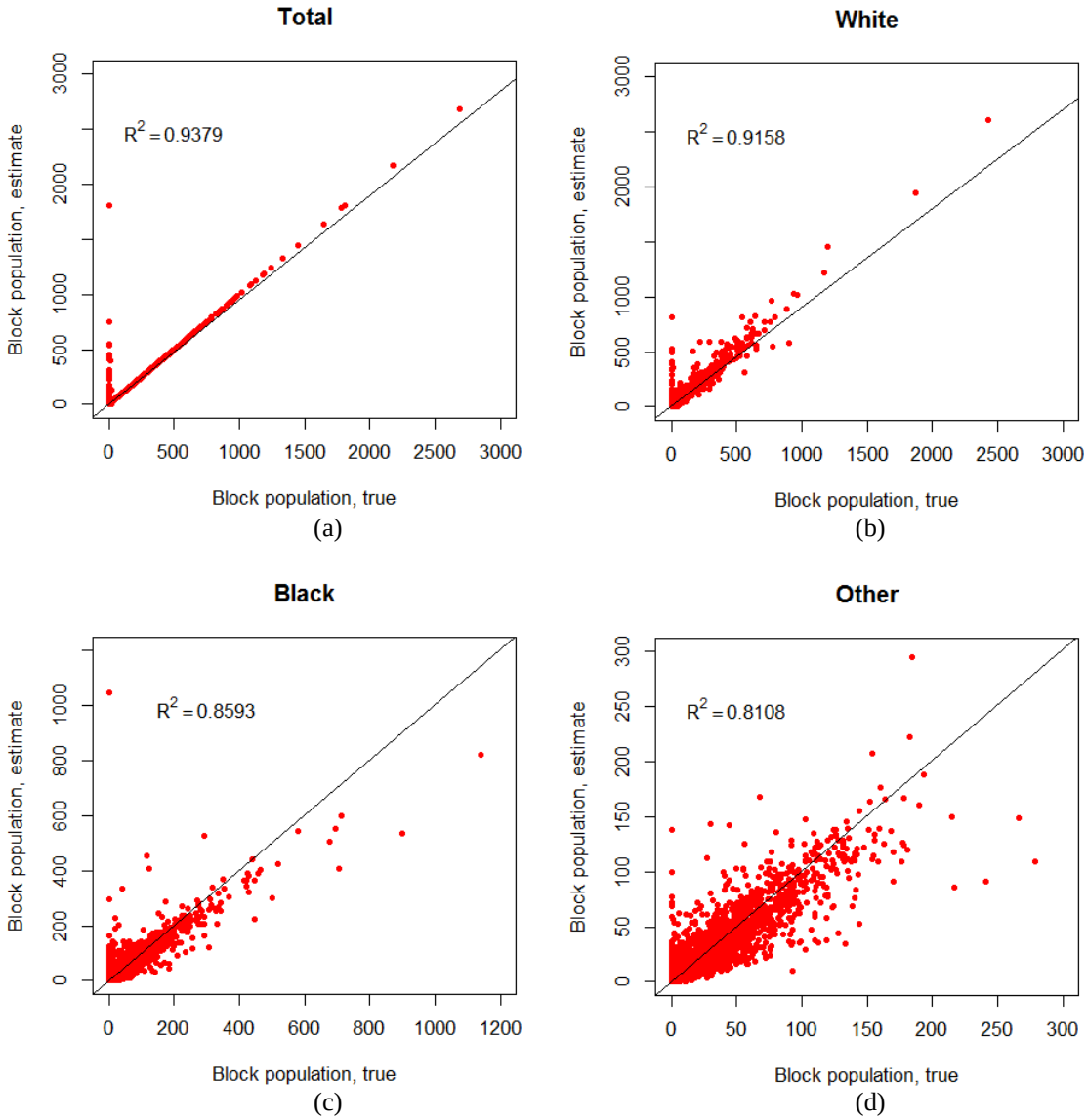


Figure 29. Scatterplots between aggregated block population from grids and the census block population: (a) Total; (b) White; (c) Black; (d) Other.

One can see from the scatterplots above that the combination of the smart interpolation method and the kernel density method generates relatively satisfactory population estimates. The R^2 value, which is the coefficient of determination of a regression model, indicates how well the estimated values fit the true values. The total population estimates have the highest R^2 (0.9379) among all, as shown in Figure 29(a). It is mainly because, in the smart interpolation process, the total population counts of grid cells are allocated within each census block according to their cumulative weights. It is expected that the aggregation of total population values of grid cells within each block should be controlled to equal the block true values. The white population has the second best estimates, which is indicated by the R^2 (0.9158) in Figure 29(b). The black and other populations have the least satisfactory estimation results, with R^2 values of 0.8593 and 0.8108 in Figure 29(c)(d), respectively. The difference among the accuracy of the total and subpopulation estimates may be explained by the intensities of concentration of those populations. The results from the spatial autocorrelation analysis in Table 1indicate that the population of other races is the most spatially clustered among populations of different race, followed by black and white. The total population distribution is the most dispersed. With the very similar search radiuses for the total- and sub-populations of race in the kernel density estimation, the most spatially clustered population will end up being dissolved to the greatest extent, which will consequently result in larger errors.

As one can notice in the scatterplots in Figure 29, especially in Figure 29(a), there are several outlier census blocks that are far away from the regression line. Those census blocks have the true total population value of 0, but the estimated values are larger than 0. The highest estimated population of a block with zero true population is 1809.

Considering the mean total population of all 13,179 census blocks is 71.9, the estimation errors are unacceptable. By examining these "outlier" blocks in the map, it is found that most of the "outlier" census blocks are sliver polygons that are usually street corners, narrow double-line road segments, and so forth. Figure 30 shows six example sliver census blocks. Since the areas of these sliver polygons are so small or the widths of strip slivers are smaller than 30m, there are no whole grid cells or the centers of grid cells falling inside these sliver polygons. In the implementation of the *zonal statistics* tool, the sliver polygons will not be able to recognize the spatial aligned grid cells and summarize their values. By default, the aggregation of a random neighboring census block will be assigned to each of these sliver polygons. In cases that the sliver polygons are surrounded by blocks with high total- and sub-population values, the absolutely errors will be high. These high errors are artifacts of the sliver polygons, which are usually very small blocks.

It is important to note that these errors appear only when the grid cells are aggregated to blocks. It does not necessarily mean that the grid cells with which the sliver polygons overlap have unacceptable estimation errors themselves. In the allocation of census block population values to grid cells, the sliver-overlapping cells are treated as within the neighboring blocks. The principle that a grid cell is considered within a block if its center falls inside the block is kept consistent throughout the study. By pair-wise comparison between the census population and the aggregated population of blocks, 125 blocks are found with the "sliver polygon" problem and excluded from the error analyses. This sorting process is also accompanied by visual validation of the sliver polygons in the map.

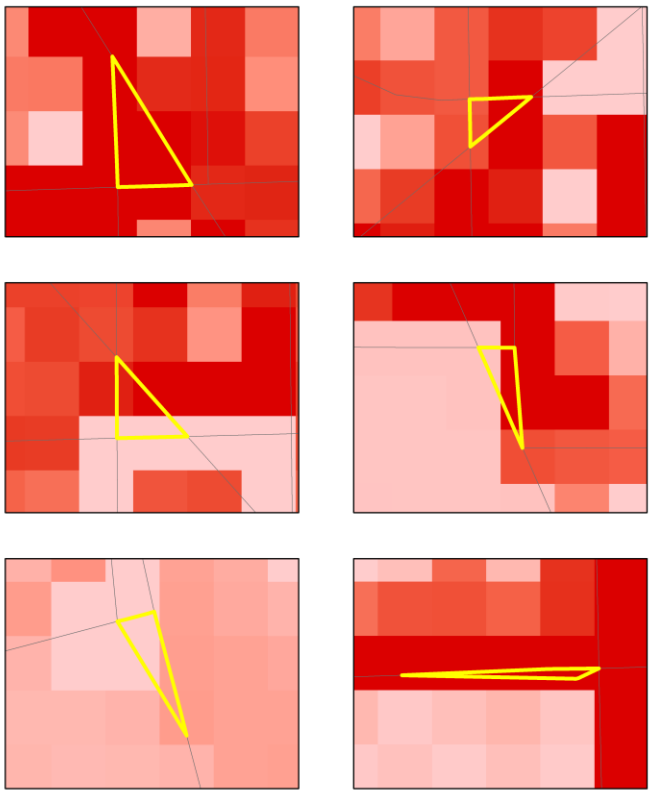


Figure 30. Example sliver census blocks. These blocks are the "outliers" in the scatterplot and have the greatest contribution to the estimation errors.

The exclusion of sliver census blocks will diminish the numeric errors introduced by the grid cell aggregation process and help reveal the real estimation accuracy. Figure 31 shows the scatterplots of total- and sub-populations estimation errors excluding sliver blocks. As one can see, compared to the R^2s in Figure 29, the R^2s of total- and sub-populations plots are significantly improved, with the values of 1, 0.9582, 0.9152 and 0.826 for total, white, black and other populations, respectively. Results from regression analyses in Table 2 also indicate a significant correlation between census block populations and the aggregated populations from grid cells, with coefficients of 1, 0.935, 1.068 and 1.025, for total, white, black and other population, respectively. Standardized coefficients indicate even better regression results.

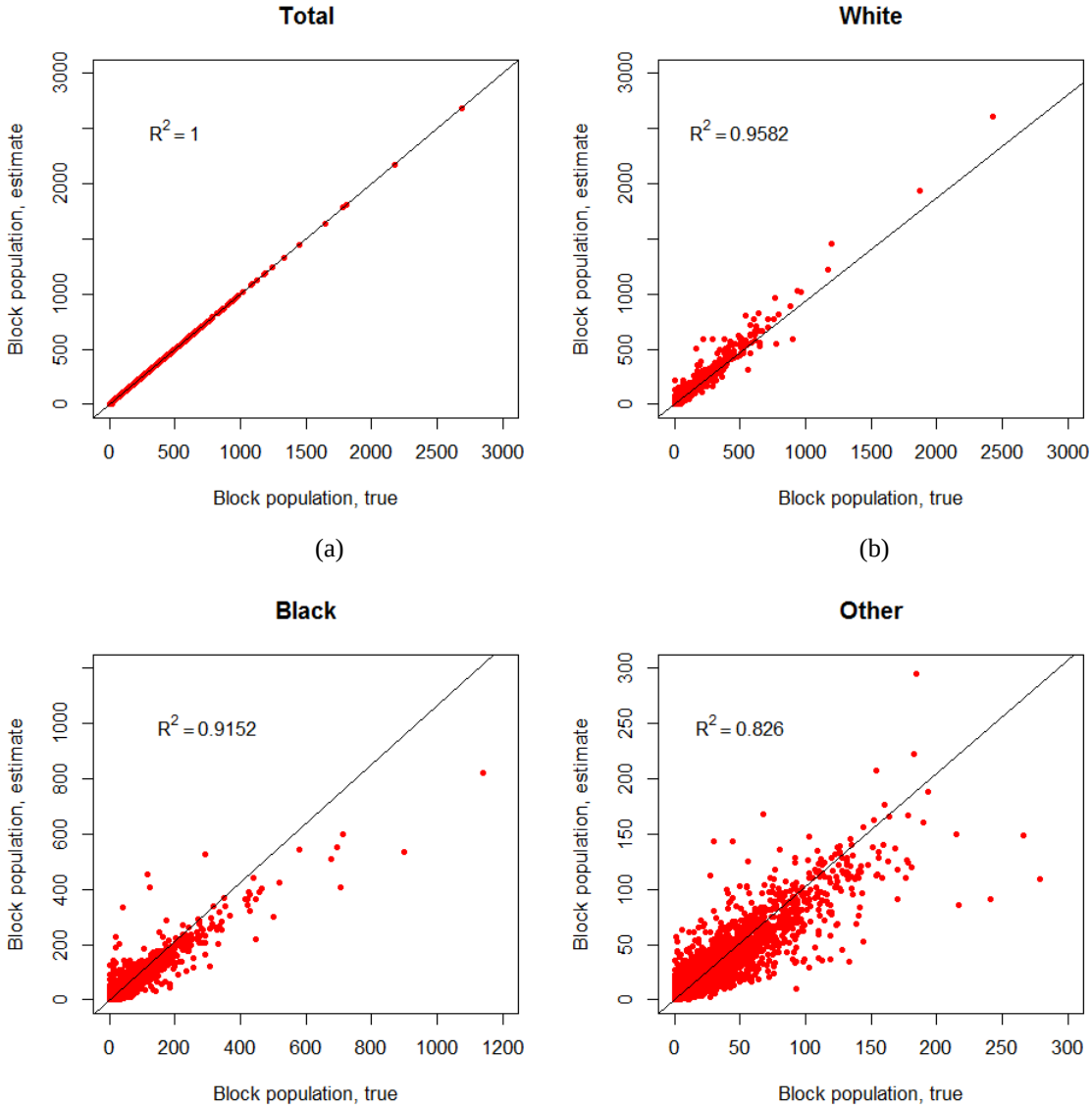


Figure 31. Scatterplots between aggregated census block population from grids and "true" census block population, after exclusion of sliver blocks: (a) Total; (b) White; (c) Black; (d) Other.

Table 2. Regression analysis of total population and sub-populations of race at census block level.

		Coefficients	Standardized Coefficients	R^2	*p*-value
Total	Constant	-0.011	-	1	0.000
	Model	1	1		
White	Constant	2.026	-	0.9582	0.000
	Model	0.935	0.979		
Black	Constant	-1.465	-	0.9152	0.000
	Model	1.068	0.957		
Other	Constant	-0.373	-	0.826	0.000
	Model	1.025	0.909		

5.1.1.2. Overestimation and underestimation

Figure 32 shows the histograms of the estimation errors of total- and sub- populations. The estimation error of white population has the range of [-319, 371], black population has the range of [-364, 339], and other population has the range of [-177, 113]. The histograms show that the mean absolute estimation error of white population is 0.89 and the mean absolute estimation errors of black and other population are 0.14. The errors, which are less than 1 person, indicate satisfactory estimation results. The standard deviations for white, black and other population are 14.877, 13.835 and 8.442, respectively. Considering the ranges of census block population of white, black and other races are [0, 2430], [0, 1288] and [0, 335], the mean absolute estimation errors and the standard deviations suggest statistically accurate estimation results. In fact, 95% of the estimation errors of white, black and other population fall inside the ranges of [-49, 98], [-78, -55], [-53, 39], respectively.

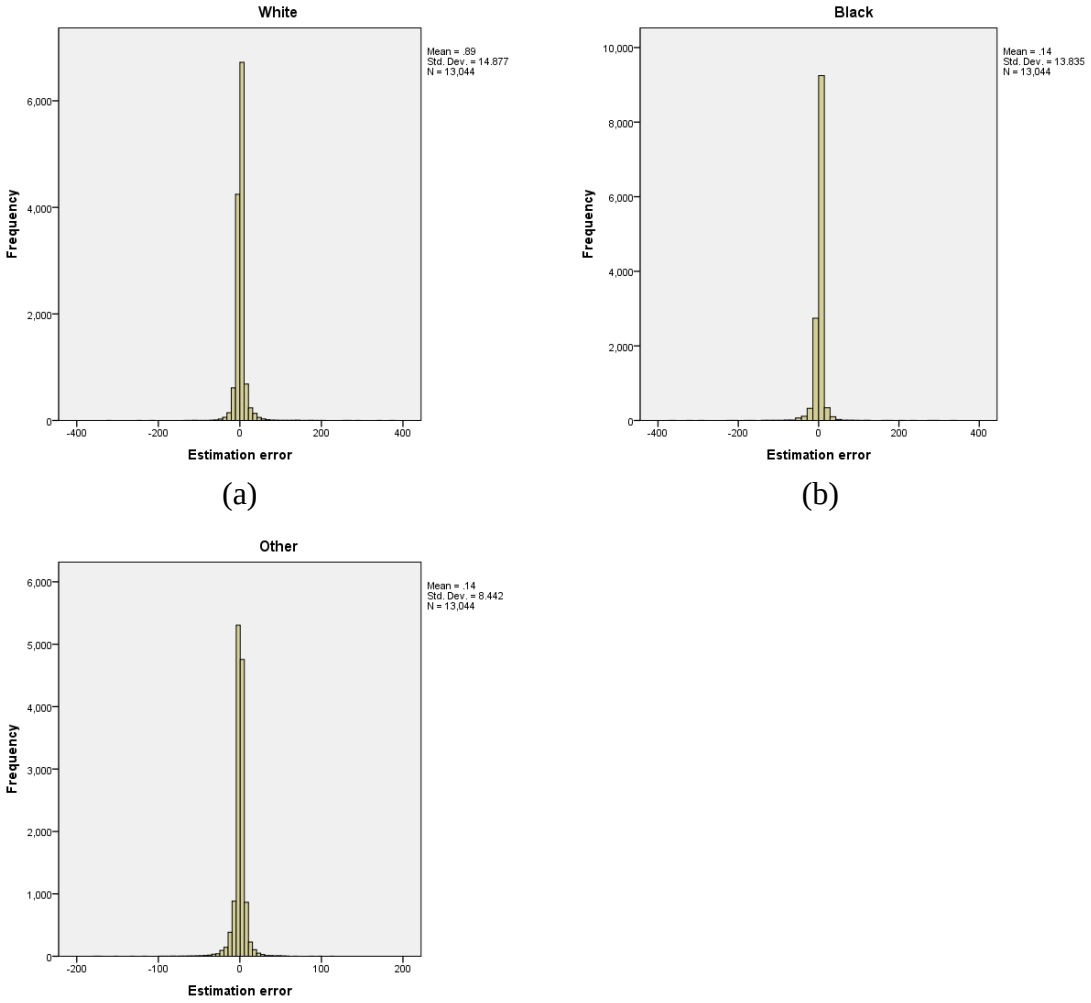


Figure 32. Histograms of estimation errors of sub-populations: (a) White; (b) Black; (c) Other.

To examine the spatial distribution of estimation errors, three choropleth maps of census blocks illustrate the absolute errors (see Figure 33-35). In each error map, blocks that are shaded red are with highest overestimations while blocks that are shaded green have the highest underestimations. As one can see, the estimation errors of sub-populations do not display very recognizable spatial patterns. The analysis of error will focus on the blocks with highest overestimation and underestimation errors.

In Figure 33, the red block (tract 187200, block 2016) that is in the southwest corner of the County is where the county correction facility-south is located. The total-population is 1809. The number of black population incarcerated (1259) in the facility is much higher than white population incarcerated (532). As one can see in Figure 33, White population has one of the highest overestimations (269) in this block; while Figure 34 shows that black population has the highest underestimation (-204) in this block. Another conspicuous block (tract 90100, block 1000) is the red one located in the northwest of downtown, as shown in Figure 34. It has one of the highest black population overestimation (285). The white population underestimation is significant as well (-319). This block has several nursing facilities. In Figure 33, the red block (tract 007400, block 1002) that is north of the Milwaukee downtown area and close to Michigan Lake has very high overestimation of white population (176). The error of black population is not so significant (-20). This block is the location of a high-level university dormitory which has over 2700 residents. The race information of the dormitory residents is not readily available, but the population density of this block is surely very high, considering the block area is small. Another red block (tract 005100, block 2004) with high white population overestimation (118) is the one in northwest of downtown and along Highway 43 in Figure 33. It shows that this block has a very high black population underestimation (-115) as well. In fact, an apartment complex called Fernwood Court Apartment is located in this block. The highest overestimations of other population are in the nearby area of Miller Park, which is the county center. There are several apartment complexes in the block too. By examining these representative blocks with high estimation errors, a possible explanation is that the errors are strongly associated with blocks with population

density abnormalities, which are usually introduced by non-typical residential arrangements such as correctional facilities, nursing homes, high-rise apartment complexes and dormitories in educational institutions.

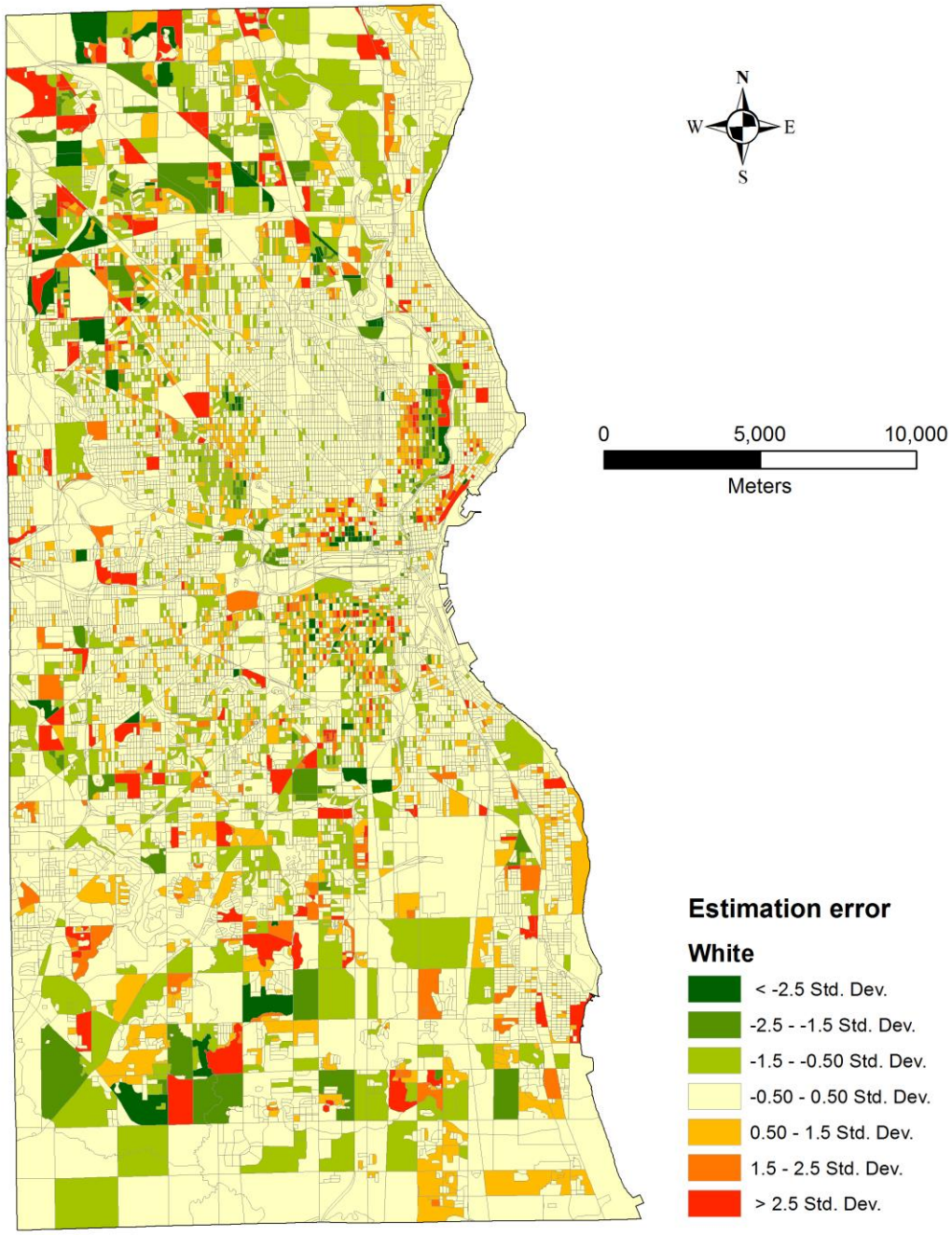


Figure 33. Estimation errors of white population.

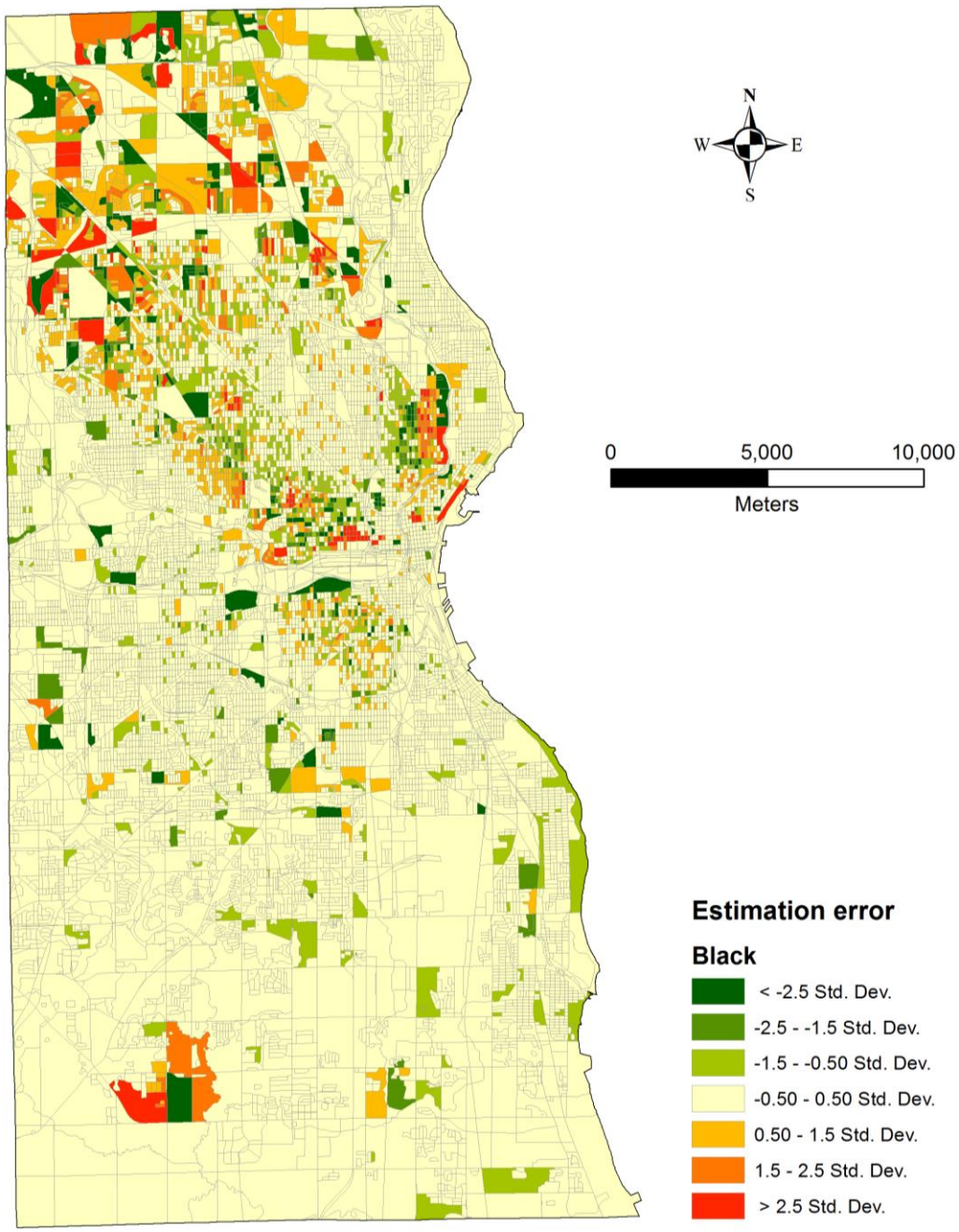


Figure 34. Estimation errors of black population.

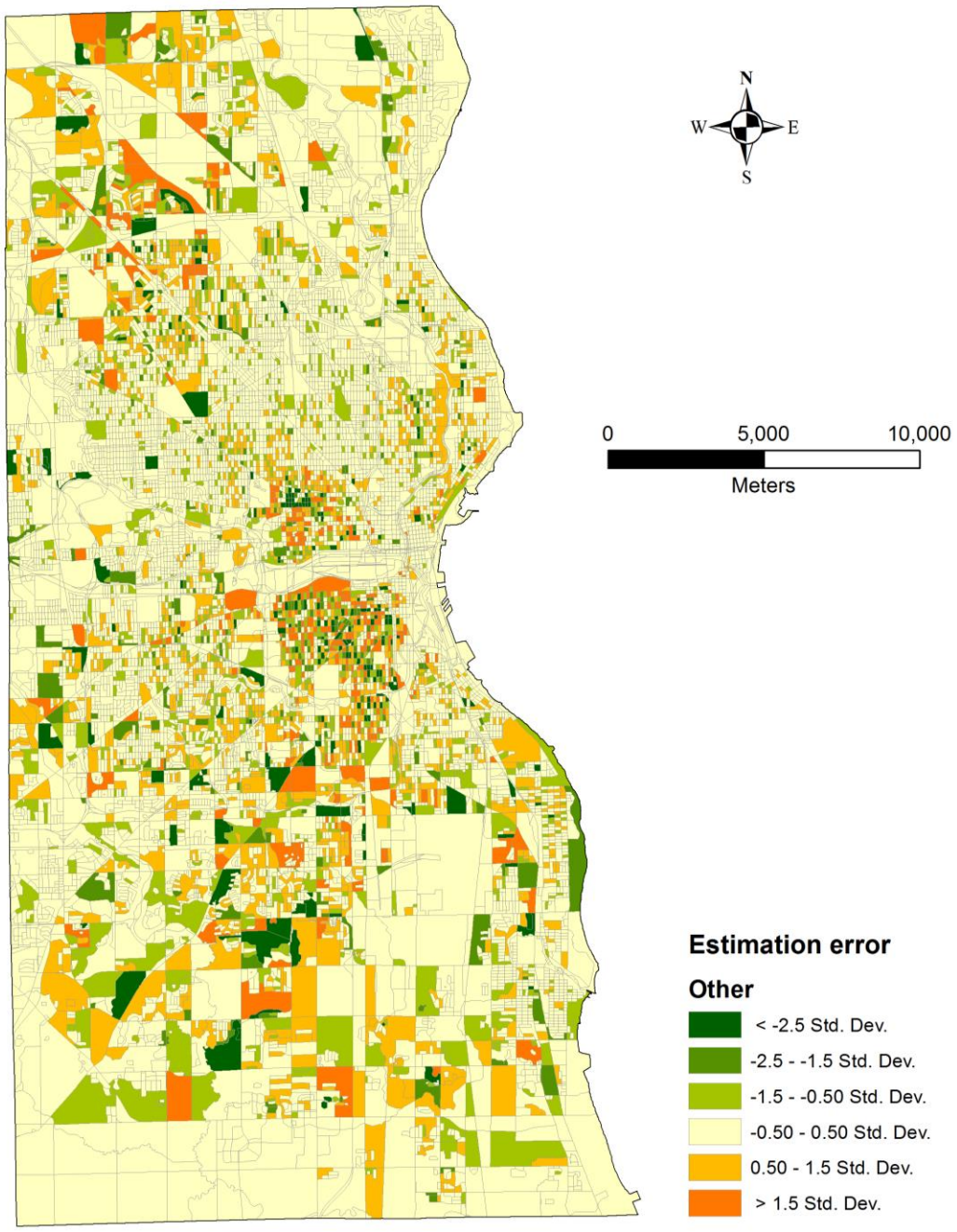


Figure 35. Estimation errors of other population.

5.1.2. Population grids of age groups

5.1.2.1. Scatterplots and regression statistics

The population grids of 20 to 34 and 65 and over age groups are shown in Figure 27-28. As one can see, population of age 20 to 34 has a more clustered spatial pattern than population of age 65 and over. The cluster is mainly located close to downtown area. Figure 36(a)(b) are the scatterplots between block populations aggregated from grids and "true" census block populations for age 20 to 34 and age 65 and over, respectively, before excluding sliver blocks. The R^2 of population of age 20 to 34 is 0.7619; while the R^2 of population of age 65 and over is 0.6037. Figure 36(c)(d) plot the block populations aggregated from age grids and "true" census block populations for age 20 to 34 and age 65 and over, respectively, after excluding sliver blocks. The R^2 of population of age 20 to 34 is 0.8521 while the R^2 of population of age 65 and over is 0.6338. Results from regression analysis in Table 3 show a significant correlation between census block populations and aggregations of grid cells, with standardized coefficients of 0.923 and 0.796, for population of age 20 to 34 and age 65 and over, respectively.

Table 3. Regression analysis of sub-populations of age groups at census block level.

		Coefficients	Standardized Coefficients	R^2	*p*-value
Age 20 to 34	Constant	-1.359	-	0.8521	0.000
	Model	1.265	0.923		
Age 65 and over	Constant	4.828	-	0.6338	0.000
	Model	0.531	0.796		

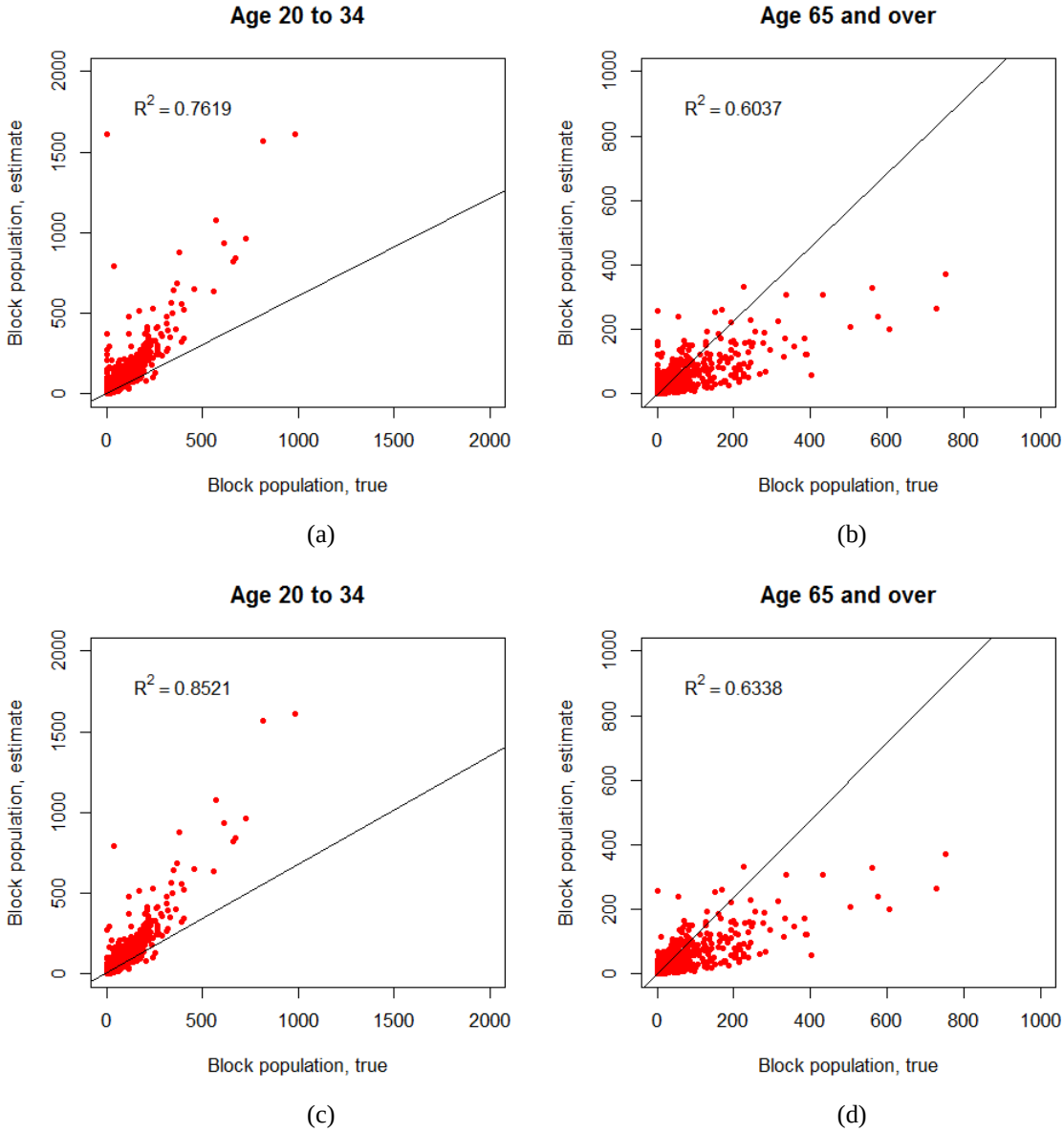
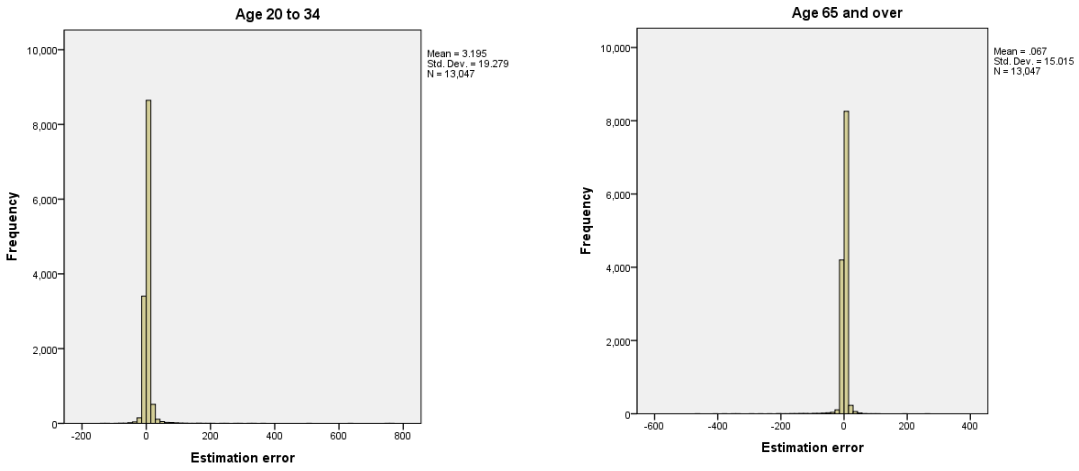


Figure 36. Scatterplots between census block populations aggregated from grids and "true" census block populations: (a) Age 20 to 34; (b) Age 65 and over; (c) Age 20 to 34, after exclusion of sliver blocks; (d) Age 65 and over, after exclusion of sliver blocks.

5.1.2.2. Overestimation and underestimation

Figure 37 shows the histograms of estimation errors of age sub-populations. The estimation errors of population of age 20 to 34 have a range of [-136, 758]; while the estimatio errors of population of age 65 and over have a range of [-461, 257]. The

histograms show that the mean absolute estimation error of population of age 20 to 34 is 3.195 and the mean absolute estimation error of population of age 65 and over is 0.067. The errors, which are less than 1 person, indicate satisfactory estimation results. The standard deviations for population of age 20 to 34 and age 65 and over are 19.279 and 15.015, respectively. Considering the ranges of census block population of age 20 to 34 and age 65 and over are [0, 981] and [0, 752], the mean absolute estimation errors and standard deviations suggest statistically accurate estimation results. In fact, 95% of the estimation errors of population of age 20 to 34 and age 65 and over fall inside the ranges of [-11, 25] and [-12, 14], respectively.



(a) (b)

Figure 37. Histograms of estimation errors of sub-populations: (a) Age 20 to 34; (b) Age 65 and over.

The spatial distribution of estimation error of population of age 20 to 34 and age 65 and over is shown in Figure 38-39**Error! Reference source not found.**. It is obvious in the maps that the population of age 65 and over has much higher overestimations and underestimations than the population of age 20 to 34. The census block that is close to

Veterans Park (tract 186900, block 1002) has one of the highest overestimations (752) for population of age 20 to 34; while for population of age 65 and over, the underestimation is also very high (-382). This block has a number of high-rise high-end condos and apartments for seniors. The block (tract 187200, block 2016), which is the place the county correctional facility-south is located, has high overestimations for both population of age 20 to 34 (633) and age 65 and over (55). The count of population of age 20 to 34 is 981; while the count of population of age 65 and over is 4, which is very small compared to 981. It is predictable that this block will have high overestimation for elderly population since it is surrounded by blocks with high population of age 65 and over. The result that this block also has high overestimation for young population has to do with the area of the block. The search radius for population of age 20 to 34 in kernel density estimation is 791.14m, which is too small to cover the whole block. Therefore, in addition to the young population in this block, part of young population in neighboring blocks are "transferred" into this block, which results in the high overestimation. The block (tract 7400, block 1002) where the University of Wisconsin-Milwaukee residential hall is located has the highest overestimation (258) for population of age 65 and over. This block does not have population of age 65 and over in census, but there are several blocks in surrounding neighborhood that have high population of age 65 and over.

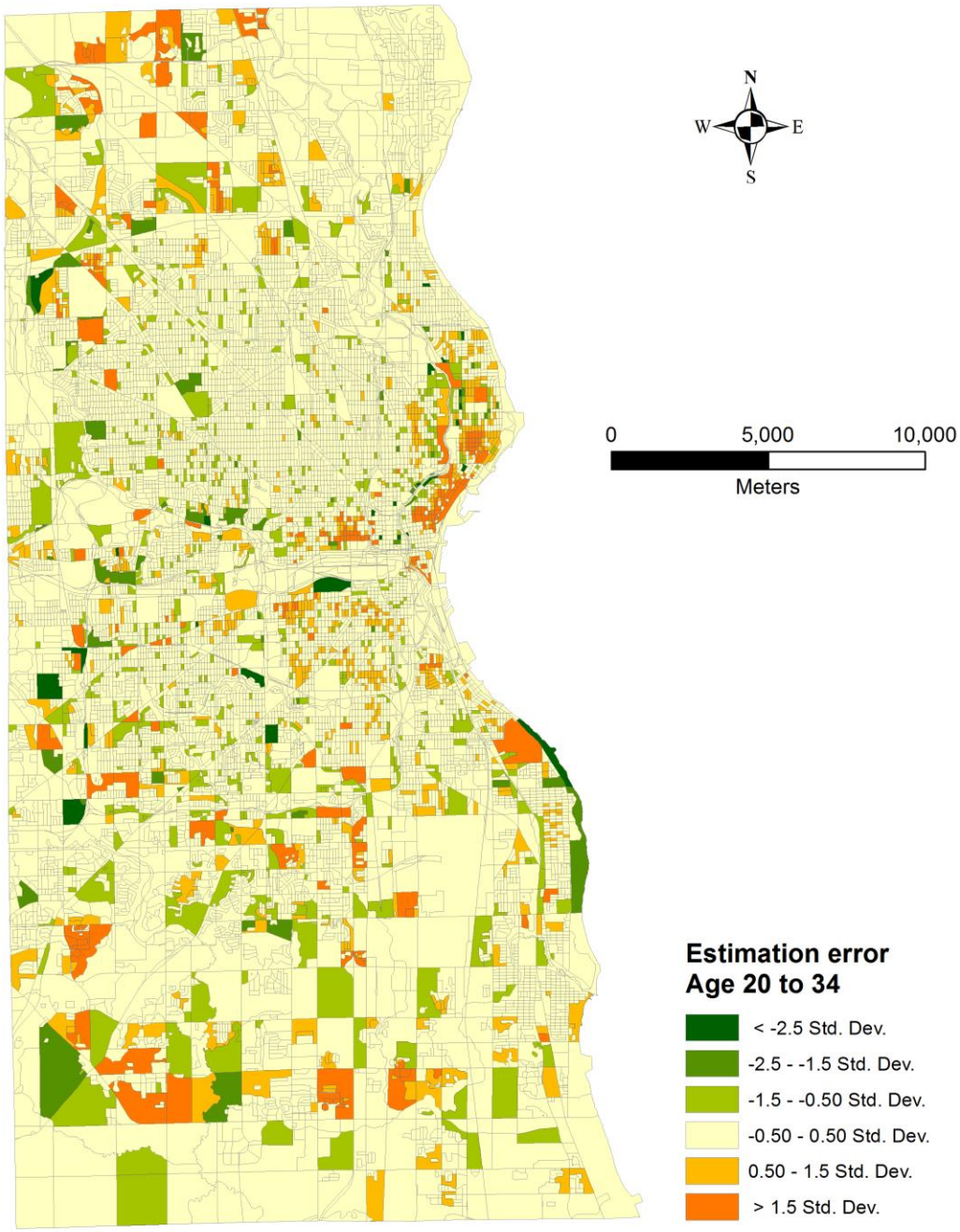


Figure 38. Estimation errors of population of age 20 to 34.

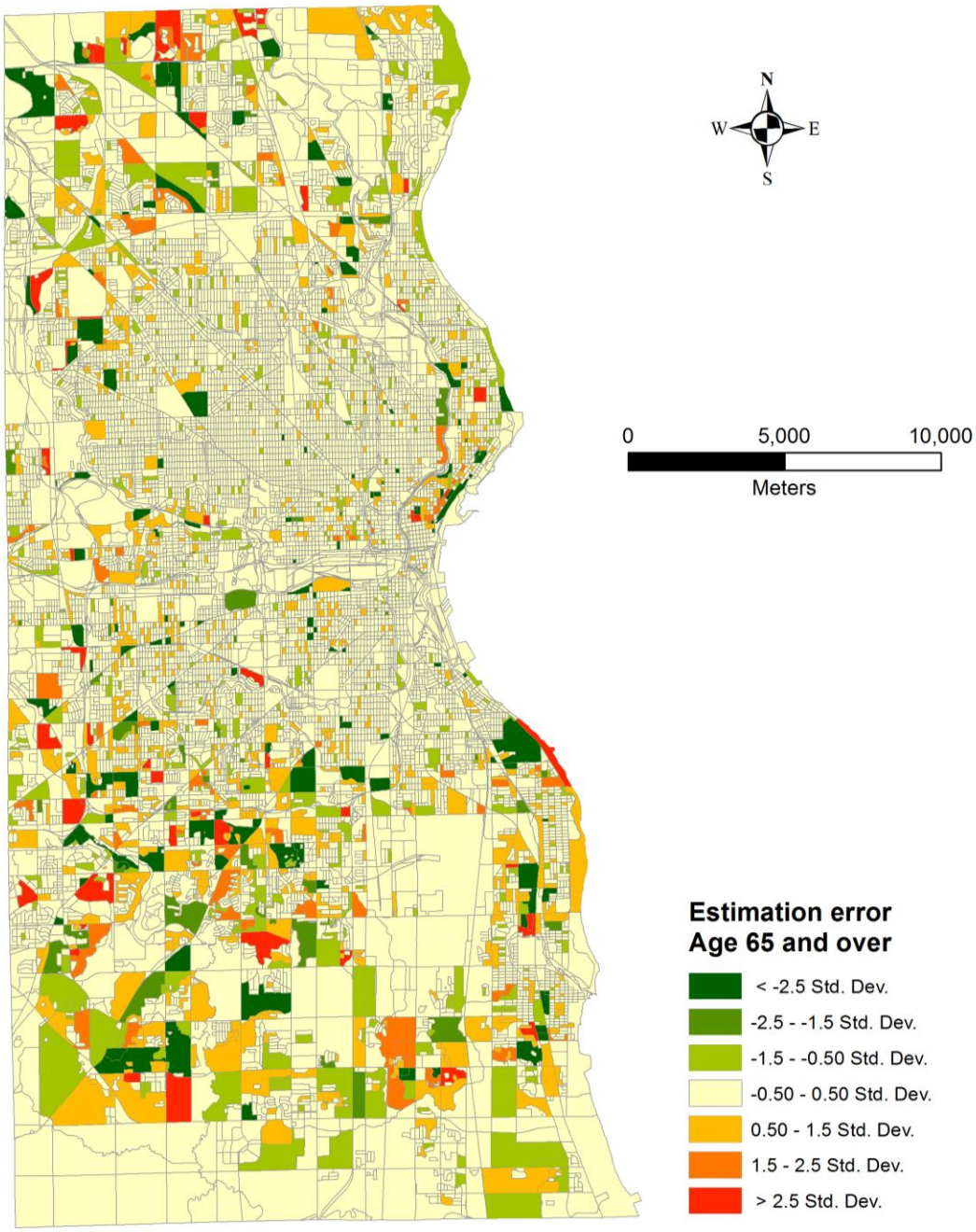


Figure 39. Estimation errors of population of age 65 and over.

5.2. Input data quality

Developing high-resolution population grids requires quality data input in the model. Since block is the finest geography level at which the population data are released in censuses, using census block population data as the original dataset to allocate population values to grid cells is the first option. Turner and Openshaw (2001), in the discussion of the *disaggregative spatial interpolation problem* (DSIP) in spatial interpolation, asserted that estimation accuracy was improved when target zones were relative larger compared with source zones. The problem with using census block population data is that absolute errors, instead of standardized measurements such as absolute percentage errors and RMSE, are used in the estimation result assessment. This is due to the fact that there are 2,377 blocks with zero total population and that sub-populations could be zero even if the total population is not. The situation causes zero as denominator in the percentage error equation, which complicates the percentage error assessment.

Ancillary data that are of finer resolution than census block and that are strongly correlated with population distributions are highly preferred. In the accuracy assessment of NLCD, Wickham et al. (2013, p.303) concluded that the overall quality was very high and that the main confusion came from the "difficulty in distinguishing the context of grass." The variation within the vegetation category does not affect the data quality when the binary dasymetric approach is used in this study. However, NLCD does not have the ability to differentiate residential pixels from other urban pixels, which hinders the spatial accuracy of total population interpolation results. In addition, the latest NLCD was produced in 2006, which is 4 years earlier than the year 2010. Land use and land cover changes may happen during this short period of time due to urbanization, land restoration

projects, etc. This temporal gap may contribute to the estimation errors associated with total-population. By applying novel and reliable land cover classification techniques (e.g. use the elevation differences between the high-points and the low-points of buildings to identify build-up types: high-rise commercial buildings, apartment complexes…) on high-resolution orthophotographs in 2010, it is possible to extract relatively more accurate residential areas. However, interpreting remotely sensed photos requires rigid image classification skills and local knowledge. Since classification errors are likely to be introduced in this process, NLCD that is already validated by USGS may be a better choice. Also, to facilitate urban planning practices, more and more state agencies are producing and releasing parcel-level land use data, which explicitly code different parcel types into residential, commercial, industrial, etc. (Deng and Wu, 2013) These datasets provide great value in helping identify residential areas accurately at sub-block level. However, parcel-level data also have some limitations. First, even though some parcels are assigned as residential type, the actual construction may occur up to several years later (Deng and Wu, 2013). Second, population is not strictly distributed within residential parcels. Other types, such as retail-related or education-related parcels, may be highly populated as well. Therefore, when utilizing parcel data to approximate residential population distribution at a very local level, it is not very reliable to simply use residential parcel boundaries.

Traditional road network method in population interpolation uses the length of road segments in target zones as a weighting factor to disaggregate source zone population. The target zones in this study are 30m ×30m grid cells. Considering such small areas, grid cells that are highly populated may not have any road segment inside.

Thus, the traditional approach does not apply in this case. Proximity to roads, an alternative variable related to road network, is employed under the assumption that people are more likely to reside along roads instead of far away from them. The same approach is applied in the production of LandScan datasets.

Another issue regarding input data quality is the categorization of races. As a socially constructed concept, the identification of race is not as entrenched as other demographic variables such as age and sex. Moreover, the interpretation of race has also been radically evolving. One of the most recent changes regarding the definition of race in census was in the 2000 census. Since the decennial census, people of mixed races are not asked to identify themselves as of only one single race; instead, the census questionnaires provide the option of "two or more races" (Jones et al., 2001). This change in racial category greatly complicates demographic studies. Population of mixed races will need in-depth investigation when highly accurate estimation results are expected. In this study, in order to establish a generic model, population of mixed races is categorized into the "other" population. Future research may need to carefully deal with the issue to improve estimation accuracy.

5.3. Error sources

Estimation errors in this study come from the coaction of several factors such as the errors in original census data, abnormalities in block population density, the centroid locations and search radiuses in kernel density estimation, and the subjectivity of human intervention in the modeling process (e.g. categorization of land use data, grid cell weighting).

First, census block population data have errors themselves. Williams (2009) asserted that, due to atypical living circumstances, language barriers (international migrants), legal status issues, or any combinations of these, racial minorities have undergone recurrent undercounts in censuses. Some sub-populations may have overcounts as well. Even though the U.S. Census Bureau adjusts population statistics before each release, the data can still be different from the true populations.

Second, the kernel density function creates a smooth density surface to fit the block population data. The value of each cell in a kernel density surface is the sum of spatially overlapping kernel values within the search radius of neighboring block centroids. When abnormality of population density appears, such as the HL and LH census blocks in Figure 8-13, the effect of neighboring blocks is intensified. Group quarter populations usually contribute to the abnormality. The assessment of population grids verifies group quarter populations' influence on estimation errors. It poses a question that whether group quarter populations should be treated separately. The issue here is that race and age information of populations in many group quarters such as senior homes and college dormitories are not readily available to be integrated in this study. Prison is an exception because the race and age data are available in Census's Summary File, but the prison populations are still aggregated to census blocks which may contain other residential land.

Third, in kernel density estimation, total- and sub-population density surfaces over the study area are generated from representative points (in this case, census block centroids). This "point to surface" approach works the best when the original point dataset can approximate the true distribution of spatial features or incidents to the most

extent. In this study, the spatial locations of individuals in each census blocks are unknown, due to the protection of privacy in census data releases. Meanwhile, census block is the finest geography level at which detailed population statistics are available. Above all, using census block centroids as representative points is adequate in the kernel density estimation, if not perfect. A possible improvement of the representative points is that they may be restrained to spatially coincide with residential units within each block, given the whole or a part of the block is inhabited. It would be even better for the representative points to coincide with the largest residential area within census blocks if there are several separate residential areas. Moreover, in creating kernel density surfaces, the population each census block is interpolated to a circular area defined by the block centroid and search radius. The fixed search radius for each population type causes a problem because it does not take into accounts the different sizes and shapes of blocks. Populations in small blocks may have high underestimation since the block area is too small compared to the circular area in kernel density estimation; while a large block may have high overestimation because the circular area cannot cover the whole block, and populations in neighboring blocks are transferred to this block. In addition, the circular area in kernel density estimation assumes that spatial autocorrelation is equally strong in all directions. This may not be realistic, especially at the boundaries of areas of different populations such as white and black. Semivariogram model is able to differentiate the directions of spatial autocorrelation. If this information is integrated with the sizes and shapes of census blocks, asymmetric kernel density functions may be constructed to better approximate population distribution.

Finally, human intervention in the modeling process may also contribute to the estimation errors. Subjectivity mainly comes from the selection of weighting factors and the assigning of weights to grid cells. The weighting factors selected in this study are verified in many empirical studies, but there still might be other geophysical, socio-economical variables that have strong correlation with population distribution. These variables need to be further explored. In the grid cell weighting process, the cumulative weights of cells are created by simply multiplying the weights of different variables. This approach does not investigate whether some weighting factors have more influence on population values than others. This is because there is no way to establish a regression between grid cell populations and values of weighting factors since the "true" population values of each grid cell are unknown.

6. Application

Population grids have been used in a wide range of environmental and social studies. Vörösmarty et al. (2000) used population grids to assess the impact of population growth on global water resource. Ewert and Harpel (2004) used LandScan 2001 to estimate population at risk of volcanic eruptions. Hafner et al. (2005) used LandScan 2002 to investigate the influence of local population on atmospheric polycyclic aromatic hydrocarbon (PAH) concentrations. Rowley et al. (2007) used LandScan 2004 to assess the potential risks that changing sea levels are imposing to population and land. Elvidge et al. (2009) used LandScan 2004 as well as satellite-derived nighttime light data to create a global poverty map. Sabesan et al. (2007) developed a set of metrics for geospatial dataset evaluation by using LandScan and GRUMPv1 datasets in disaster response applications. So far, these applications can only estimate how many people are in grid cells. Due to the limit of current population grids on total population, it is still impossible to accurately estimate what kinds of population are in grid cells. With the grids with demographic traits developed in this study, a case study on the population near public recreation facilities is performed to show the potential of these population grids.

Public recreation facilities are provided and maintained by administrations to improve the quality of local residents' physical living environment. Urban parks, as a major aspect of the public recreation system, have been playing a vital role in accommodating people's increasing desire for a higher quality of life (Nicholls, 2001). However, empirical researches have shown that the efficiency of certain public services is imbalanced among different racial groups, usually minority population are underprivileged in the accessibility to those services. This application uses the total- and

sub-population grids to evaluate the racial inequity in the walking accessibility to urban parks.

The service areas of urban parks are determined using a linear 400m buffer, which is considered as a reasonable walking distance. Figure 40 shows the urban parks and their service areas. The service areas include the urban parks themselves. Then, using the *zonal statistics* tool, these service areas are delineated as zone features to aggregate the total- and sub-population grids. The sums of grid cell values are the service total- and sub-populations. To test the population results against those generated using polygon-based spatial interpolation methods, the area weighting method is chosen. Intersections between census blocks and service areas are created by spatially overlaying the two features. The total and sub-populations of the intersections are calculated by splitting blocks total and sub-populations according to the area ratios between intersections and census blocks. In the end, the total and sub-populations of intersections are dissolved to get the total and sub-populations in the service areas.

Table 4 shows the population statistics. Compared to the county population ratios (white: 60.63%, black: 26.78%, other: 12.59%) in the description of study area, the white population ratios calculated by using both the area weighting method (62.54%) and aggregation of population grids (63.73%) suggest that the service areas have the higher than average white population. The black population ratios calculated by using both the area weighting method (25.69%) and aggregation of population grids (25.76%), however, indicate that the service areas have worse accessibility to urban parks than average black population. The statistics of other population show the same disparity as black population. Also, the difference between white and black population ratios in the area

weighting method (36.85%) is smaller than the difference between white and black population ratios in the population grids method (37.97%). The difference between white and other population ratios in the area weighting method (50.78%) is also smaller than the difference between white and other population ratios in the population grids method (51.52%). Compared to the population grids, using the area weighting method to investigate the racial difference of accessibility to urban park services between white and minority populations can weaken the intensity of racial inequality.

In terms of populations of age groups, the populations of age 65 and over in the service area do not show much difference between the two methods, but the area weighting method underestimates the population of age 20 to 34 by 16,749 persons, compared to aggregated population from grid. This huge gap may contribute to inefficient decision-making when administrators try to improve young population's accessibility to urban parks.

Table 4. Comparison of service populations and ratios using area weighting method and aggregation of population grids, using a 400m linear buffer of urban parks as service areas.

	Total	White	Black	Other	Age 20 to 34	Age 65 and over
Area weighting	370714	231862 (62.54%)	95260 (25.69%)	43591 (11.76%)	93484 (25.22%)	54716 (14.76%)
Population grids	369197	234180 (63.73%)	95114 (25.76%)	45084 (12.21%)	110233 (29.86%)	53216 (14.41%)

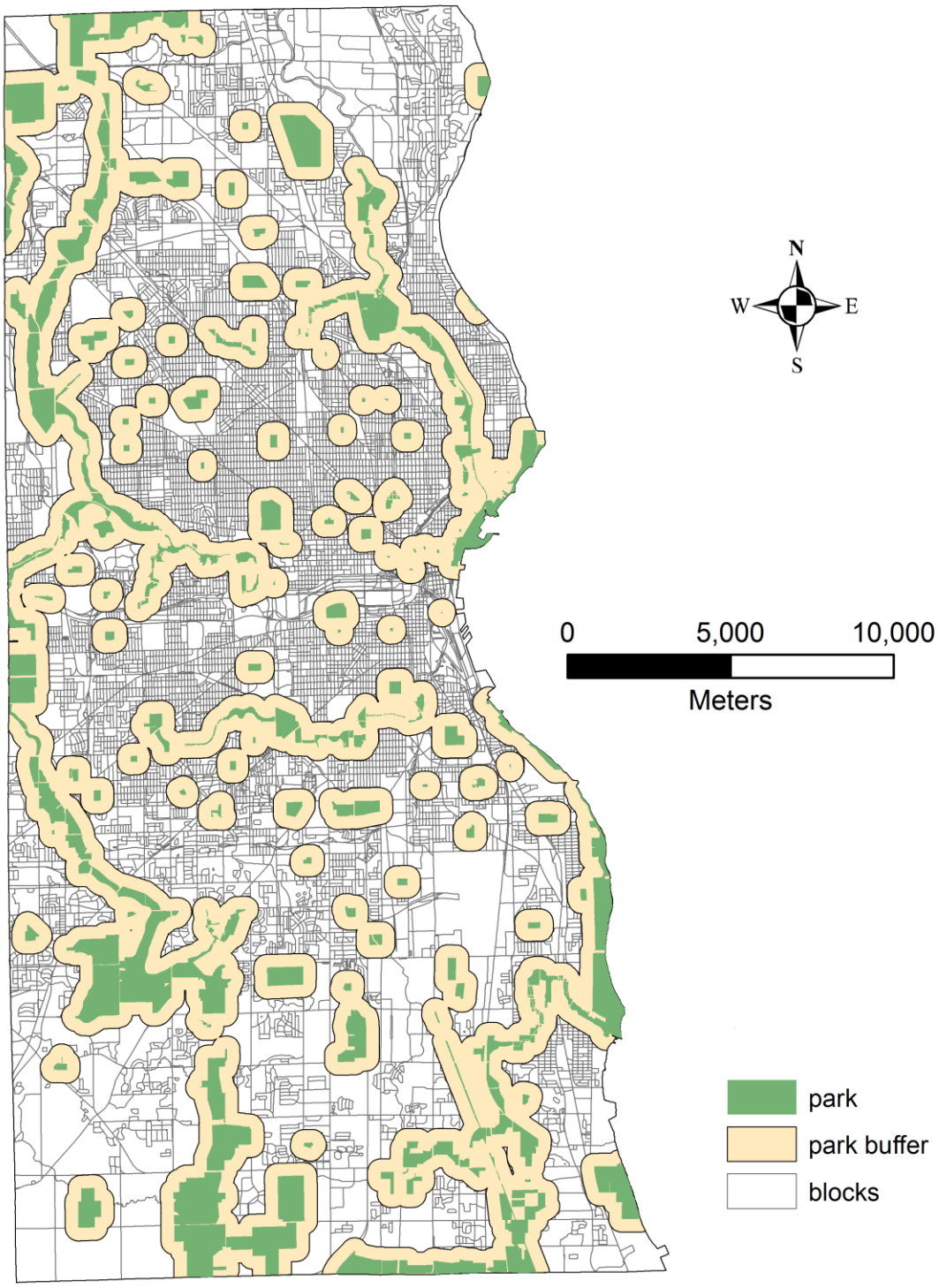


Figure 40. Service areas of urban parks in Milwaukee County.

Another advantage of population grids is that the difference among population distributions in a zone is readily visible at a very high spatial resolution. Figure 41 zooms in closer on the total- and sub-population grids to a noise zone defined by a 400m linear buffer outward a section of the State Highway 100. One can see that the distribution patterns of total- and sub-populations are quite different. Black population tends to be more concentrated at the west end and the south side of the noise zone than white population. Other population is relatively small and mostly located at the west end. In addition, the zone has more population of age 20 to 34 than population of age 65 and over. This is especially true in the west half of the noise zone.

At last, the implementation of the aggregation of population grids is much easier than spatial interpolation methods since the *zonal statistics* tool is readily usable in ArcGIS while polygon-based spatial interpolation tools are not. There are some areal weighting tools developed by ESRI users, but those tools often need to be customized for specific research projects.

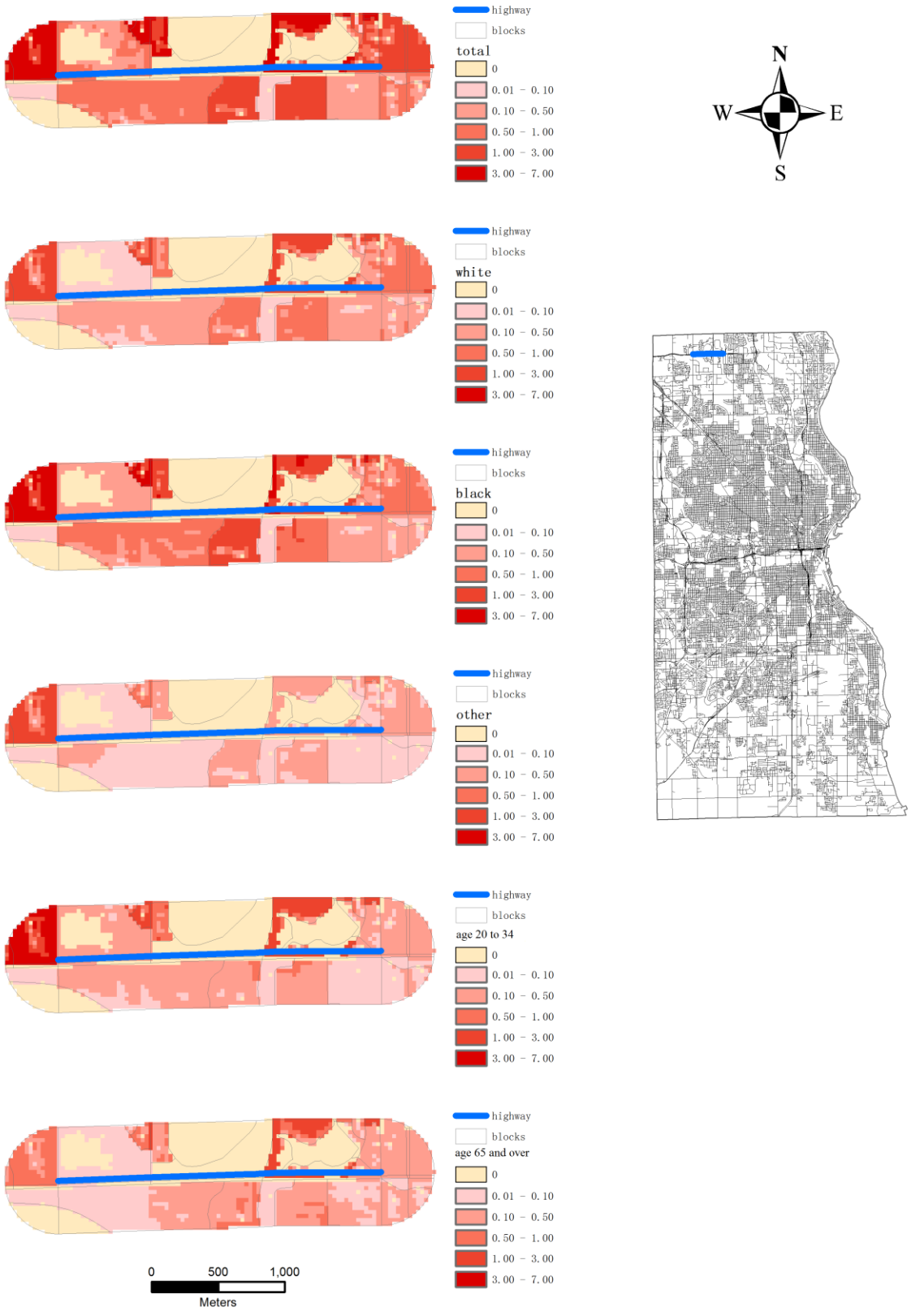


Figure 41. Total- and sub-population grids in the noise zone along a state highway section.

7. Conclusions

This study generates high spatial resolution, race and age-specific population grids for Milwaukee County, Wisconsin. It intends to address the issues in population studies that census population aggregate data are not adequate to portray population distribution variation at geography levels that are lower than census zonal system, and that most of previous population grids only deal with the estimation of total population, neglecting the distribution variation among sub-populations. Formerly produced population grids at different levels and many spatial interpolation methods are reviewed in this study. The creation of high-resolution population grids with demographic trait (race and age specifically in this case) consists of two stages: the first one uses selected areal interpolation method to create a total-population grid; the second employs a point-based kernel density function to create density surfaces for sub-populations from block centroid. In kernel density estimation, spatial autocorrelation is investigated to determine the optimal values of the "search radius" parameter. The sub-population density surfaces are then used as relative weights to calculate sub-populations in each grid cell. To validate the estimation accuracy of the total- and sub-population grids, grid values are aggregated to census blocks in order to compare the estimated populations to true block populations. Scatterplots and regression analyses show that total and sub-population grids using the integration of "smart" interpolation and kernel density function generates satisfactory population estimates. The overestimation and underestimation of sub-populations do not show significant spatial patterns but the relationship between estimation errors and block population characteristics is briefly discussed. Non-typical and highly populated residential arrangements such as prisons, dormitories, nursing

homes and high-rise apartments show large contributions to estimation errors in the corresponding blocks.

The importance of the quality of input data in developing population grids is also discussed. Census blocks are chosen as source zones to disaggregate population values to grid cells because they are the finest geography unit at which census data are reported. Compared to other census zonal systems such as block groups and tracts, the relatively smaller difference in spatial resolution between blocks and grid cells helps better approximate the true population distribution. However, block population data limit the possibility of using percentage errors in estimation error assessment, which may present different error patterns to absolute errors. The discussion about the quality of ancillary data in total population estimation majorly focuses on land use data, road network, and categorization of race. To develop better population grid models, more non-traditional indicator variables that are correlated with population distribution may be needed.

To testify the advantage of population grids over population aggregate data in census zonal units in population-related spatial analysis, accessibility to urban parks of racial and age groups in the study area are calculated using both area weighting interpolation and aggregation of population grids. Populations and population ratios in urban park service areas of the two approaches are compared. Results indicate that using area units with homogenous population distribution in area weighting method distorts the intensity of racial and age inequality compared to aggregation of population grids. The application is a simple example of how population grids with demographic trait can benefit population-related spatial analysis.

Population grids have the great potential to portray the spatial heterogeneity of population distribution, and to facilitate micro-level population-related studies. In the meantime, they are very flexible in terms of the ability to aggregate according to different geography levels. Population grids endowed with demographic details, such as race, age and sex, expand the dimensions of total population grids and should be of great need in planning, emergency management, disaster relief practices of public agencies and private sectors.

REFERENCES

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

Bentley, G. C., Cromley, R. G., and Atkinson‐Plombo, C. (2013). The Network Interpolation of Population for Flow Modeling Using Dasymetric Mapping. *Geographical Analysis*, *45*(3), 307-323.

Bhaduri, B., Bright, E., Coleman, P., and Dobson, J. (2002). LandScan. *Geoinformatics*, *5*(2), 34-37.

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

Bracken, I. (1991). A surface model approach to small area population estimation. *Town Planning Review*, *62*(2), 225.

Bracken, I. (1993). An extensive surface model database for population-related information: concept and application. *Environment and Planning B: Planning and Design*, *20*(1), 13-27.

Bracken, I. (1994). A surface model approach to the representation of population-related social indicators. *Spatial analysis and GIS*, 247-259.

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

Bracken, I., and Martin, D. (1995). Linkage of the 1981 and 1991 UK Censuses using surface modelling concepts. *Environment and Planning A*, *27*(3), 379-390.

Briggs, D. J., Gulliver, J., Fecht, D., and Vienneau, D. M. (2007). Dasymetric modelling of small-area population distribution using land cover and light emissions data. *Remote sensing of Environment*, *108*(4), 451-466.

Cai, Q. (2007). New techniques in small area population estimates by demographic characteristics. *Population Research and Policy Review*, *26*(2), 203-218.

Cai, Q., Rushton, G., Bhaduri, B., Bright, E., and Coleman, P. (2006). Estimating Small‐Area Populations by Age and Sex Using Spatial Interpolation and Statistical Inference Methods. *Transactions in GIS*, *10*(4), 577-598.

Center for International Earth Science Information Network (CIESIN)., Columbia University., and Centro Internacional de Agricultura Tropical (CIAT), (2004). Gridded Population of the World (GPW), Version 3. *Palisades, NY: Columbia*

University. Retrieved from http://beta.sedac.ciesin.columbia.edu/data/collection/gpw-v3, on 11/15/2013.

Center for International Earth Science Information Network (CIESIN). (2004). Global Rural-Urban Mapping Project (GRUMP), v1. *Palisades, NY: Columbia University*. Retrieved from http://sedac.ciesin.columbia.edu/data/collection/grump-v1, on 11/15/2013.

Center for International Earth Science Information Network (CIESIN). (2006). U.S. Census Grids. *Palisades, NY: Columbia University*. Retrieved from http://sedac.ciesin.columbia.edu/data/collection/usgrid, on 11/15/2013.

Citro, C. F. (1998). Window on Washington: Model-Based Small-Area Estimates: The Next Major Advance for the Federal Statistical System for the 21st Century. *Chance*, *11*(3), 40-50.

Coombes, J. (2002). Handy hints for variography. *Snowden Associates Ltd*.

Deichmann, U. (1996). *A review of spatial population database design and modeling*. National Center for Geographic Information and Analysis.

Deichmann, U., Balk, D., and Yetman, G. (2001). Transforming population data for interdisciplinary usages: from census to grid. *Washington (DC): Center for International Earth Science Information Network*.

Deichmann, U., National Center for Geographic Information and Analysis (NCGIA), and University of California, Santa Barbara (UCSB). (1996a). *Corresponding (E. Asia) Population Density Map, Metadata.* Retrieved from https://na.unep.net/metadata/unep/GRID/EASPOPD.html, on 11/20/2013.

Deichmann, U., National Center for Geographic Information and Analysis (NCGIA), and University of California, Santa Barbara (UCSB). (1996b). *Corresponding (E. Asia) Population Density Map, Metadata.* Retrieved from https://na.unep.net/metadata/unep/GRID/WASPOPD.html on 11/20/2013.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal statistical Society*, *39*(1), 1-38.

Deng, C., and Wu, C. (2013). Improving small-area population estimation: an integrated geographic and demographic approach. *Annals of the Association of American Geographers, 103*(5), 1123-1141.

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.

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.

Elvidge, C. D., Sutton, P. C., Ghosh, T., Tuttle, B. T., Baugh, K. E., Bhaduri, B., & Bright, E. (2009). A global poverty map derived from satellite data. *Computers & Geosciences*, *35*(8), 1652-1660.

European Forum for GeoStatistics (EFGS). (2012). *ESSnet project GEOSTAT 1A-Representing Census data in a European population grid-Final Report.*

ESRI. (2012a). *ArcGIS Desktop Help 10.1 - How Kernel Density works*. Retrieved from http://resources.arcgis.com/en/help/main/10.1/index.html#//009z00000011000000 , on 2/15/2014.

ESRI. (2012b). *ArcGIS Desktop Help 10.1 - Semivariogram and covariance functions*. Retrieved from http://resources.arcgis.com/en/help/main/10.1/index.html#//003100000036000000 , on 2/20/2014.

ESRI. (2012c). *ArcGIS Desktop Help 10.1 - Modeling semivariogram and covariance functions: Changing the lag size and number of lags*. Retrieved from http://resources.arcgis.com/en/help/main/10.1/index.html#/Changing_the_lag_size_and_number_of_lags/0031000000mn000000/, on 2/20/2014.

ESRI. (2013). *ArcGIS Desktop Help 10.1 - Cluster and Outlier Analysis (Anselin Local Moran's I) (Spatial Statistics)*. Retrieved from http://resources.arcgis.com/en/help/main/10.1/index.html#/Cluster_and_Outlier_Analysis_Anselin_Local_Moran_s_I/005p0000000z000000/ , on 3/10/2014.

Ewert, J. W., and Harpel, C. J. (2004). In harm's way: population and volcanic risk. *Geotimes*, *49*(4), 14-17.

Fisher, P. F., 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.

Fisher, P. F., and Langford, M. (1996). Modeling Sensitivity to Accuracy in Classified Imagery: A Study of Areal Interpolation by Dasymetric Mapping∗. *The Professional Geographer*, *48*(3), 299-309.

Flowerdew, R. (1988). Statistical methods for areal interpolation: predicting count data from a binary variable. *Northern Regional Research Laboratory, Universities of Newcastle and Lancaster*.

Flowerdew, R., and Green, M. (1989). Statistical methods for inference between incompatible zonal systems. *Accuracy of spatial databases*, 239-247.

Flowerdew, R., and Green, M. (1990). *Inference between incompatible zonal systems using the EM algorithm*. University of Lancaster, North West Regional Research Laboratory.

Flowerdew, R., and Green, M. (1993). Developments in areal interpolation methods and GIS. In *Geographic Information Systems, Spatial Modelling and Policy Evaluation* (pp. 73-84). Springer Berlin Heidelberg.

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

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

Fry, J., Xian, G., Jin, S., Dewitz, J., Homer, C., Yang, L., Barnes, C., Herold, N., and Wickham, J. (2011). Completion of the 2006 National Land Cover Database for the Conterminous United States. *PEandRS*, Vol. 77(9), 858-864.

Gallego, F. J. (2010). A population density grid of the European Union. *Population and Environment*, *31*(6), 460-473.

Ghosh, M., and Rao, J. N. K. (1994). Small area estimation: an appraisal. *Statistical science*, 55-76.

Goerlich, F. J., and Cantarino, I. (2013). A population density grid for Spain. *International Journal of Geographical Information Science*, (in press), 1-17.

Goodchild, M. F. (1989). Modeling error in objects and fields. *Accuracy of spatial databases*, 107-114.

Goodchild, M. F., Anselin, L., and Deichmann, U. (1993). A framework for the areal interpolation of socioeconomic data. *Environment and Planning A*, *25*(3), 383-397.

Goodchild, M. F., Lam, N. S. N., and University of Western Ontario. Dept. of Geography. (1980). *Areal interpolation: a variant of the traditional spatial problem*. London, Ont.: Department of Geography, University of Western Ontario.

Gotway, C. A., Ferguson, R. B., Hergert, G. W., & Peterson, T. A. (1996). Comparison of kriging and inverse-distance methods for mapping soil parameters. *Soil Science Society of America Journal*, *60*(4), 1237-1247.

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.

Gregory, I. N., and Ell, P. S. (2006). Error‐sensitive historical GIS: Identifying areal interpolation errors in time‐series data. *International Journal of Geographical Information Science*, *20*(2), 135-152.

Griffith, D. A. (2013). Estimating Missing Data Values for Georeferenced Poisson Counts. *Geographical Analysis*, *45*(3), 259-284.

Hafner, W. D., Carlson, D. L., & Hites, R. A. (2005). Influence of local human population on atmospheric polycyclic aromatic hydrocarbon concentrations. *Environmental science & technology*, *39*(19), 7374-7379.

Handcock, M. S., and Stein, M. L. (1993). A Bayesian analysis of kriging. *Technometrics*, *35*(4), 403-410.

Hawley, K., and Moellering, H. (2005). A comparative analysis of areal interpolation methods. *Cartography and Geographic Information Science*, *32*(4), 411-423.

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.

Honeycutt, D., and Wojcik, J. (1990). Development of a population density surface for the conterminous United States. In *Proceedings GIS/LIS* (Vol. 90, No. 1, pp. 484-496).

Jones, N. A., Smith, A. S., Black, A. A., Hawaiian, N., and Indian, A. (2001). *The two or more races population, 2000* (Vol. 8, No. 2). US Department of Commerce, Economics and Statistics Administration, US Census Bureau.

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

Lam, N. S. N. (1982). An evaluation of areal interpolation methods. In *5 th Int. Symp. Comp. Assisted Cartography and Int. Soc. Photogrammetry and Remote Sensing Commission IV*.

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., Maguire, D. J., and Unwin, D. J. (1991). The areal interpolation problem: estimating population using remote sensing in a GIS framework. *Handling geographical information: Methodology and potential applications*, 55-77.

Langford, M., and Unwin, D. J. (1994). Generating and mapping population density surfaces within a geographical information system. *Cartographic Journal, The*, *31*(1), 21-26.

Lo, C. P. (2003). Zone-based estimation of population and housing units from satellite-generated land use/land cover maps. *Remotely sensed cities*, 157-180.

Nicholls, S. (2001). Measuring the accessibility and equity of public parks: a case study using GIS. *Managing Leisure*, *6*(4), 201-219.

Maantay, J. A., 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.

Maantay, J. A., Maroko, A. R., and Porter-Morgan, H. (2008). Research note—A new method for mapping population and understanding the spatial dynamics of disease in urban areas: Asthma in the Bronx, New York. *Urban Geography*, *29*(7), 724-738.

Markoff, J., and Shapiro, G. (1973). The linkage of data describing overlapping geographical units. *Historical Methods Newsletter*, *7*(1), 34-46.

Martin, D., and Bracken, I. (1991). Techniques for modelling population-related raster databases. *Environment and Planning A*, *23*(7), 1069-1075.

Martin, J. H., and Serow, W. J. (1978). Estimating demographic characteristics using the ratio-correlation method. *Demography*, *15*(2), 223-233.

McCleary, G. F. (1969). *The dasymetric method in thematic cartography* (Doctoral dissertation, University of Wisconsin--Madison).

Mennis, J. (2002). Using geographic information systems to create and analyze statistical surfaces of population and risk for environmental justice analysis. *Social science quarterly*, *83*(1), 281-297.

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., and Hultgren, T. (2006). Intelligent dasymetric mapping and its application to areal interpolation. *Cartography and Geographic Information Science*, *33*(3), 179-194.

Moon, Z. K., and Farmer, F. L. (2001). Population density surface: A new approach to an old problem. *Society and Natural Resources*, *14*(1), 39-51.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, *37*(1-2), 17-23.

Mugglin, A. S., Carlin, B. P., Zhu, L., and Conlon, E. (1999). Bayesian areal interpolation, estimation, and smoothing: an inferential approach for geographic information systems. *Environment and Planning A*, *31*(8), 1337-1352.

Mugglin, A. S., Carlin, B. P., and Gelfand, A. E. (2000). Fully model-based approaches for spatially misaligned data. *Journal of the American Statistical Association*, *95*(451), 877-887.

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.

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. (1983). *The modifiable areal unit problem* (Vol. 38). Norwich: Geo books.

Openshaw, S., and Taylor, P. J. (1979). A million or so correlation coefficients: three experiments on the modifiable areal unit problem. *Statistical applications in the spatial sciences*, *21*, 127-144.

Plane, D. A., and Rogerson, P. A. (1994). The Geographical Analysis Of Population: With Applications To Planning And Business Author: David A. Plane, Peter A. Roger.

Qiu, F., and Cromley, R. (2013). Areal Interpolation and Dasymetric Modeling. *Geographical Analysis*, *45*(3), 213-215.

Qiu, F., Zhang, C., and Zhou, Y. (2012). The Development of an Areal Interpolation ArcGIS Extension and a Comparative Study. *GIScience and Remote Sensing*, *49*(5), 644-663.

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.

Rowley, R. J., Kostelnick, J. C., Braaten, D., Li, X., and Meisel, J. (2007). Risk of rising sea level to population and land area. *Eos, Transactions American Geophysical Union*, *88*(9), 105-107.

Sabesan, A., Abercrombie, K., Ganguly, A. R., Bhaduri, B., Bright, E. A., & Coleman, P. R. (2007). Metrics for the comparative analysis of geospatial datasets with applications to high-resolution grid-based population data. *GeoJournal*, *69*(1-2), 81-91.

Sadahiro, Y. (2000). Accuracy of Count Data Estimated by the Point‐in‐Polygon Method. *Geographical Analysis*, *32*(1), 64-89.

Schroeder, J. P., and Van Riper, D. C. (2013). Because Muncie's Densities Are Not Manhattan's: Using Geographical Weighting in the Expectation–Maximization Algorithm for Areal Interpolation. *Geographical Analysis*, *45*(3), 216-237.

Silverman, Bernard W. *Density estimation for statistics and data analysis*. Vol. 26. CRC press, 1986.

Smith, S. K., and Cody, S. (2004). An evaluation of population estimates in Florida: April 1, 2000. *Population Research and Policy Review*, *23*(1), 1-24.

Smith, S. K., and Lewis, B. B. (1980). Some new techniques for applying the housing unit method of local population estimation. *Demography*, *17*(3), 323-339.

Smith, S. K., and Mandell, M. (1984). A comparison of population estimation methods: Housing unit versus component II, ratio correlation, and administrative records. *Journal of the American Statistical Association*, *79*(386), 282-289.

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.

Starsinic, D. E. (1974). Development of population estimates for revenue sharing areas. *Census Tract Papers, Series GE*, *40*.

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.

Tober, W., Deichmann, U., Gottsegen, J., and Maloy, K. (1995). *The global demography project*. Technical Report 95-6. Santa Barbara, CA: National Center for Geographic Information and Analysis.

Turner, A., and Openshaw, S. (2001). Disaggregative spatial interpolation. *GISRUK, University of Wales, April*.

U.S. Census Bureau. (1966). Methods of population estimation: Part I, Illustrative procedure of the Bureau's component method II. *Current Population Reports, series P-25, No. 339*.

Ver Hoef, J. M., Krivoruchko, K., & Lucas, N. (2001). *Using ArcGIS geostatistical analyst* (Vol. 380). Redlands: Esri.

Voss, P. R., Long, D. D., and Hammer, R. B. (1999). *When census geography doesn't work: Using ancillary information to improve the spatial interpolation of demographic data*. Center for Demography and Ecology, University of Madison-Wisconsin.

Vörösmarty, C. J., Green, P., Salisbury, J., & Lammers, R. B. (2000). Global water resources: vulnerability from climate change and population growth. *science, 289*(5477), 284.

White, D. (1978). A design for polygon overlay. *Harvard Papers on Geographic Information Systems, 6*.

Wickham, J. D., Stehman, S. V., Gass, L., Dewitz, J., Fry, J. A., & Wade, T. G. (2013). Accuracy assessment of NLCD 2006 land cover and impervious surface. *Remote Sensing of Environment, 130*, 294-304.

Williams, J. D. (2009). The 2010 Decennial Census: Background and Issues. Library of Congress Washington DC Congressional Research Service.

Willmott, C. J., and Matsuura, K. (1995). Smart interpolation of annually averaged air temperature in the United States. *Journal of Applied Meteorology, 34*(12), 2577-2586.

Wong, D. W. (2004). The modifiable areal unit problem (MAUP). In *Worldminds: Geographical Perspectives on 100 Problems* (pp. 571-575). Springer Netherlands.

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

Wu, J. (2004). Effects of changing scale on landscape pattern analysis: scaling relations. *Landscape Ecology, 19*(2), 125-138.

Wu, S. S., Qiu, X., and Wang, L. (2005). Population estimation methods in GIS and remote sensing: a review. *GIScience and Remote Sensing*, *42*(1), 80-96.

Wu, S. 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, Z. (2006). A framework for interpolating the population surface at the residential-housing-unit level. *GIScience and Remote Sensing*, *43*(3), 233-251.

Yang, T. C. (2005). Modifiable areal unit problem. *GIS Resource Document*.

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., and Ignizio, D. A. (2010). Comparison of Dasymetric Mapping Techniques for Small-Area Population Estimates. *Cartography and Geographic Information Science*, *37*(3), 199-214.

Zucchini, W., Berzel, A., and Nenadic, O. (2003). Applied smoothing techniques.