MODELING HIGH-RESOLUTION GRIDDED POPULATION SURFACE IN ALACHUA COUNTY, FLORIDA

By

PENG JIA

A THESIS PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE

UNIVERSITY OF FLORIDA

2012

To my mom and dad who always support me unconditionally

ACKNOWLEDGMENTS

It is with a great deal of gratitude that I would like to acknowledge a group of individuals in the Department of Geography for my impressive graduate education experience. First, I want to personally thank Dr. Youliang Qiu, my thesis advisor, for helping me in large and small ways. Through working as his teaching assistant and discussing with him about research, he gave me both criticism and encouragement and looking back, I can see that I gained a lot. I am proud to say he is not only an excellent advisor, but also a reliable friend.

I would also like to thank Dr. Peter R. Waylen and Dr. Liang Mao for serving on my thesis committee. Dr. Waylen was understanding and gave me support in both my academic and personal path. Dr. Mao's logical thinking and challenging questions made me never stop expanding and refining my ideas and research. This thesis would not have been possible without their help and support.

Furthermore, I would like to thank Dr. Richard Rheingans in the Department of Environmental and Global Health. I met him at my turning point and I found my way that was really fit to me through working and discussing with him. Not only did he provide me an excellent working atmosphere to do my research and complete this thesis, he provided me with a flexible working schedule and the freedom to pursue my wide range of interests as well.

I wish to extend my sincere appreciation to Dr. Corene J. Matyas who first encouraged me to come to the University of Florida. Although developing different interests from her afterwards, her kindness, consideration and serious instructions really helped me become accustomed to American ways of thinking and communicating since I arrived at the U.S., which I will forever be grateful for.

Special thanks to Dr. Robert V. Rohli in the Department of Geography & Anthropology at LSU and Dr. Andrew Tatem at UF who gave me generous help and instructions during my tough time of greatest need.

Thank you to Dr. Timothy Fik who introduced me to the statistical world in a unique way. I got a lot of knowledge from his classes and he is a professor I greatly admire.

Thank you also to Dr. Jane Southworth, Ms. Julia Williams, Ms. Desiree Price and Ms. Rhonda Black, who provided me much advice and ease during my graduate study.

Many thanks to my friends for their help and company: to John D. Anderson, my close colleague in the Emerging Pathogens Institute who gave me much help with my work and also gave me a unique opportunity to celebrate a Thanksgiving party; to Mia Jakubisin, my good friend who gave me a great deal of help with my language; to Mario Mighty, who helped me a lot while working as a teaching assistant; to Yilun Sun in the Department of Biostatistics, my great collaborator whose expertise is complementary to mine; and to some of my friends, Zhuojie Huang, Jing Sun, Qiuyin Qi, Yang Yang and Ying Wang, for a lot of fun when we gathered.

Last but not least, I would like to show my biggest appreciation to my parents Xiu Lee and Hongwei Jia for their care and understanding throughout my life. Because of their support, I could always develop my own interests and pursue my dreams since my childhood. Talking with them on the phone always encouraged me and calmed me when I was disappointed or frustrated as they always know just what to say. I appreciate the life they gave me from the bottom of my heart. Although they could not read my thesis easily, they always knew exactly what I was doing.

"It is absolutely an important element for a digital city." Dad said.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AWM	Areal Weighting Method
AHR	Asthma Hospitalization Rate
CEDS	Cadastral-based Expert Dasymetric System
EPA	Environmental Protection Agency
GIS	Geographic Information Systems
GPW	Gridded Population of the World
GRUMP	Global Rural Urban Mapping Project
HGPS	High-resolution Gridded Population Surface
LAH	Limited Access Highways
MAUP	Modifiable Areal Unit Problem
NPL	National Priority List
RS	Remote Sensing
RTS	Regional Transit System
USCG	U.S. Census Grids

Abstract of Thesis Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Master of Science

MODELING HIGH-RESOLUTION GRIDDED POPULATION SURFACE IN ALACHUA COUNTY, FLORIDA

By

Peng Jia

May 2012

Chair: Youliang Qiu
Major: Geography

The demand for small-scale population data is increasing in a variety of fields. Spatial techniques including Geographic Information Systems (GIS) and Remote Sensing (RS) have been more often used for the population study in recent years, which makes it possible to conduct the dasymetric mapping, a vital technique transforming a traditional choropleth map into a continuous statistical surface with values at all locations rather than solely aggregated values over the existing units.

In this study, I adopted the Heuristic Sampling Method and detailed property types of parcels from the Department Property Tax Data Files in 2010 as ancillary data, to disaggregate the population counts from U.S. Census 2010 onto a quadrilateral grid composed of 30m×30m cells covering the Alachua County, Florida, named High-resolution Gridded Population Surface (HGPS).

At last, the HGPS has been applied to two case studies. The first one is to estimate the numbers of potential passengers within the service areas of bus routes in Gainesville, Alachua County. The second one is to estimate the numbers of population potentially exposed to the contamination at Cabot-Koppers Superfund Site in Gainesville. By comparing the results from HGPS and traditional population grids

produced by areal weighting method (AWM), with the actual numbers from relevant records, the HGPS was proven to be more reasonable and valuable.

HGPS is an important step forward towards solving the Modifiable Areal Unit Problem (MAUP) and expected to serve as a more accurate denominator in various research fields, such as public health, crime analysis and so on.

CHAPTER 1
INTRODUCTION

As cities develop dramatically and their inner structures get more complex, the study of population distribution therein is of extreme importance because the location of people, compared to traditional statistical numbers, is more critical to many aspects in our daily life, especially in response to natural and artificial disasters. For instance, emergency management plans evacuation routes and allocates resources speedily based on where people live; environmental health assessment needs to spatially correlate environmental risk factors with the number of people living around and their health outcomes; infrastructure development has a demand for the structure and distribution of people in the target area in order to minimize the inequity caused by unfair inaccessibility; surveying and analyzing the distribution of demographic characteristics need to be done prior to any optimal site selection with the purpose of private profits or public service.

Census data are still the primary data source for a variety of demographic researches, but they are usually released to the public about once in a decade and besides, in an aggregated form in order for confidentiality and administrative purposes (Wu 2006). We call this aggregated form the choropleth map that is composed of a series of vector polygons representing administrative units and values depicting the aggregated attributes over those units.

Martin (2011), in his up-to-date review paper, concluded four major limitations of the conventional choropleth representation from Census data. First of all, the Modifiable Areal Unit Problem (MAUP) is a matter of prime importance, which means that to modify areal boundaries or the level of data aggregation will substantially change the

distribution of population values (Openshaw, 1984). Secondly, the time lag between data collection and publication makes data easily out-of-date for many contemporary applications. Thirdly, the periodic changes of administrative unit boundary hamper the comparison of data in different years. Fourthly, Census data represent the night-time residential population distribution, instead of daytime. In sum, the continuous and dynamic system of the population has been poorly described by discrete aggregated data and static models; furthermore, the choropleth map not only removes the spatial heterogeneity within the given units by replacing a range of values with one aggregated value, but also subjects to abrupt changes at the boundary of artificially defined units that do not necessarily relate to any natural or artificial phenomenon.

Although the finest administrative unit at present, Census block, is usually enough for some large-scale studies, such as estimating the number of people affected by hurricane or earthquake, the problems still remain in small-scale issues, for example, visualizing the population distribution within Census blocks, or counting the people within the area that does not coincide with any Census unit (Wu 2006). In one word, accurate population estimations at small scales, as can be imagined, are being strongly desired to meet the growing demand for immediate and well-informed decision making (Maantay et al., 2007; Krunić, 2011).

Dasymetric mapping is a particular cartographic process by which a traditional choropleth map can be transformed into a continuous statistical surface with values at all locations rather than aggregated values over the existing units (Mennis, 2009; Martin, 2011). In the case of this study where the population counts on Census block level have been disaggregated to a set of finer grids that did not exist before, the

continuous statistical surface is a quadrilateral grid layer with cells representing the population counts within them. Although more accurate, the dasymetric mapping remains unpopular and highly subjective due to lack of good-quality ancillary data, as well as limited computer power. Additionally, the data transferring during the spatial disaggregation process is complex because of the mismatched boundaries between source and target areas and their heterogeneity in terms of density (Li, 2010). Therefore, the dominant spatial models representing demographic characteristics have been still choropleth maps until recently.

Spatial techniques have been more often used for the population study in recent years (Saporito et al. 2007; Langford 2007; Lin et al. 2011), especially Geographic Information Systems (GIS) and Remote Sensing (RS), which make the dasymetric mapping possible and easy to operate. GIS provides a unique platform to perform all necessary spatial operations, and RS provides a large pool of ancillary data to assist disaggregating data. Although Sharkova (2000) proved that GIS was optional in the data-poor research environment as the high cost was a natural concomitant of limited improvements brought by GIS, the results from various methods showed that the GIS method still outperformed others even under the situation of poor data, not to mention that more and more high-resolution spatial data are becoming available.

Most of present population products were produced under the assistance of satellite-derived land cover/use data with a certain degree of misclassification, where the density of different types of housing failed to distinguish from each other, especially in hyper-heterogeneous urban areas. This issue was unavoidable before because of inaccessibility to good-quality data. Based on author's current knowledge, there are few

well-modeled population products which not only base themselves on more accurate data than remote sensing images, but also consider various types of residential and non-residential property. It is fortunate that, over the past 20 years' efforts, the State of Florida has developed a state-wide digital parcel map dataset that contains the boundary of parcels in all 67 counties of Florida with associated tax information from the Florida Department of Revenue's tax database.

Following the brief review of some of the present demographic products and most frequently used dasymetric approaches, I described the study area, data and methodology for producing the High-resolution Gridded Population Surface (HGPS) in Alachua County, Florida, and then applied it to two case studies in order for demonstrating the strengths of the HGPS, as well as exploring the applications of the HGPS in the field of transportation and public health. One case study is to estimate the potential passengers served by Regional Transit System (RTS) in the city of Gainesville, Alachua County; the other is to estimate the numbers of population potentially exposed to the contamination at Cabot-Koppers Superfund Site in Gainesville.

CHAPTER 2
LITERATURE REVIEW

Demographic Products

The major demographic products open to the public at present include U.S. Census Grids (USCG), Gridded Population of the World (GPW), Global Rural Urban Mapping Project (GRUMP), LandScan Global and LandScan USA model. The USCG is a quadrilateral grid layer with a 250m resolution in 50 metropolitan statistical areas with a population of one million or more and a 1km resolution in the remaining regions of the U.S., but this product is simply produced by proportionally allocating the population counts at the Census block level to the grids; for example, if a grid cell contains 40% area of one block and 60% of a second one, the population count for that grid cell would be the sum of 40% population of the first block and 60% of the second. Therefore, the USCG is just simple degradation of the Census block counts without any aid of ancillary data.

The GPW is also simple grid data based on the original Census units with a 5km resolution for 1990-2015 in five year increments. The GRUMP adds urban-rural specification to the GPW by reallocating the urban population on a 1km grid layer with the aid of the night-time lights data.

The LandScan Global model, with a 1km resolution, is the earliest dasymetric product to reflect the diurnal change in population distribution, representing an average ambient population over 24 hours (Dobson et al. 2000). In 2000, the Oak Ridge National Laboratory extended the LandScan Global model and developed the LandScan USA model with a 90m resolution to capture the diurnal population dynamics. The LandScan USA, consisting of two different components, nighttime and daytime population

distribution, is produced based on layers representing land cover/use, transportation networks, various landmarks, elevation, slope, etc.; however, although considering various factors, each grid cell in both LandScan Global and USA model was weighed empirically, so the subjectivity was unavoidable during the process of modeling (Bhaduri et al. 2007).

There is an up-to-date online service named Kingston Automated Geoinformation Service (KAGIS) that allows end users to obtain the dasymetric population map of any area of interest in the United Kingdom based on a set of zone boundaries of a selected Census year by following a few instructive checkbox and button-clicking steps (Shi and Walford 2010). It only takes two minutes to complete the entire process from data transmission to producing a map at the English county level. It is a good Internet service technology to integrate with the data and method in my study.

Dasymetric Methods

All the contemporary techniques for small-scale population estimations can be classified into two categories, according to their goals. One group of methods inputs Census population counts and outputs a finer population surface by interpolation and disaggregation technique, in order to reflect much more spatial details within Census units (Bielecka 2005; Mennis and Hultgren 2006; Maantay et al. 2007; Aubrecht et al. 2009; Li and Corcoran 2010; Ural et al. 2011); in contrast, the other group of methods constructs the relationship between known population counts and other variables first, and then uses that relationship to estimate unknown population, which is often used to estimate intercensal population because Census data are only available every ten years (Wu and Murray 2007; Wu et al. 2008; Tapp 2010; Qiu et al. 2010; Silván-Cárdenas et al. 2010; Dong et al. 2010). Actually, to construct the relationship between population

and at least one additional variable from ancillary data is also necessary for the first group of methods, and the original population counts are all from Census data in both groups. The distinguished distinction between them is that the former redistributes known population counts from up to down in Census years and hence preserves the volume within original units, called downscaling; whereas the latter calculates unknown population counts from down to up in intercensal years, called upscaling. Wu (2005), Maantay (2007) and Mennis (2009), complementing each other, have reviewed nearly all the present approaches in details, so to review them all deviates from the goal of this paper; instead, I focused on how to redistribute known population counts to a finer level in a more accurate way and explored the application of results.

Dasymetric mapping is a vital technique in the first category of methods. The Russian geographer, Semenov-Tyan-Shansky (1827-1914), has often been credited with inventing the dasymetric map (Bielecka 2005) and the American geographer, John Kirtland Wright (1891-1969), may perhaps be the first person to publish a paper on dasymetric mapping in an English-language journal. A dasymetric map is typically described in contrast to a choropleth map where the boundaries of regions are typically, for convenience of enumeration, derived from administrative or jurisdictional divisions rather than the patterns of features the map itself depicts. In this part, I gave a brief review of the most frequently used dasymetric methods that transferred data from one set of geographic zones, commonly Census tract, block group or block, to another set of non-coincident zones, such as school zones or grids.

Areal weighting method works based on the simplest assumption that the population is evenly distributed in the source zone. If the target zone is completely

located within a source zone, the population in the target zone would be the product of the population in the source zone and area ratio of the target zone to the source zone. If the target zone contains or intersects more than one source zone, each intersecting zone within the target zone would get a certain amount of population proportional to the area ratio of itself to the source zone it belongs to and the total population in the target zone would be the sum in all intersecting zones. The U.S. Census Grids is produced based on this approach because of its simple performance. The major weakness of this method is its unrealistic assumption that the population is always evenly distributed across the source zone and population is still assigned to uninhabitable areas in the new target zone using this method, which fails to reflect the strengths of dasymetric mapping. Therefore, to disaggregate the population counts to finer units need the aid of ancillary data.

Filtered areal weighting method is an improvement of the areal weighting method, which employs ancillary data, such as classified remote sensing data or land cover/use polygon, to mask non-residential areas and redistribute the population solely to the remaining areas. For example, Maantay (2007) used the layers of landmark and water body to remove the uninhabitable areas from Census tract and block group maps, as one of three methods involved in his study. This method assumes that all the people live in the classified residential zones and the population density is homogeneous among all residential areas, but the truth is that non-residential areas often have population as well, and the density of different types of residential housing varies a lot. A further step towards accuracy is to use high-resolution satellite imageries or aerial photos to differentiate various residential dwellings, such as single family, multi-family and

apartment (Ural et al. 2011). The major drawback here is the expensive satellite data; moreover, the available pixel resolution is limited and the issue of mixed pixel decomposition in remote sensing images remains unsolved very well, especially in the urban areas (Maantay et al. 2007).

Image texture method is to use Census data and high-resolution satellite images, such as ALOS (2.5m for panchromatic and 10m for multi-spectral), IKONOS (1m for panchromatic and 4m for multi-spectral) and QuickBird (0.6m for panchromatic and 2.4m for multi-spectral), to build the correlation between the population density and image texture by extracting Homogeneous Urban Patches (HUP) from images on a basis of the rule of maximizing between-patch differences while minimizing within-patch differences, classifying and matching them with Census data (Liu and Clarke 2002; Maantay et al. 2007). Even the images containing only three visible bands from Google Earth were used in a case study in the Lake Tai basin, eastern China (Yang et al. 2011) due to the high cost of purchasing required remote sensing data. As the technique of object-oriented classification progresses, this method is much easier to realize than before; however, the correlation between the population density and image texture was not high enough due to the misclassification and complexity of spectral texture.

Heuristic Sampling Method uses similar ancillary data involved in other methods, but predefines a series of classes and selects all Census units entirely occupied by a certain class to calculate the density on each class of land type. Mennis (2003) predefined three classes, high-density urban, low-density urban and non-urban areas, and used this approach to produce a population surface in five counties of Pennsylvania, where he calculated the density of each class county by county and the population

counts within block groups were redistributed to each grid cell based on two factors: the relative density of each class and percentage of total area of each class within a block group. However, the ancillary data he used were too coarse, which led to the sparse involved classes and invisible improvements in highly urbanized areas after dasymetric mapping. In addition, this method fails to reveal the intra-class variations in population density.

Regression method is a priori to set up a certain number of classes, classify the land cover/use data and overlay it with Census data in order to build a statistical relationship between the dependent variable, population counts, and independent variables, the area ratio of all predefined classes within Census units. Yuan (1998) built a correlation between the population counts and land cover types in four counties of central Arkansas, where the correlation was assumed to vary county by county and the scaling techniques were applied to adjust the statistical results because the coefficients representing the density of the most sparsely populated areas were negative. The weakness of this method is that the automatic regression procedure hides the intra-class variations in the population density and outliers that would influence the determination of the coefficients, which can be avoided in the manual Heuristic Sampling Method. Another issue is that if using the coefficients from the regression procedure to estimate the population from down to up, rather than redistribute the population from up to down, few aggregated values within Census units would be equal to the original counts.

Cadastral-based Expert Dasymetric System (CEDS), the up-to-date population disaggregation product in New York City, used the highly detailed cadastral data and

two types of proxies, residential area (RA) and number of residential units (RU), to redistribute population from Census units onto each parcel (Maantay et al. 2007). It was assumed that the population was proportional to the residential area; in other word, the population density on all residential types of housing was stable across the New York City where this assumption was applicable due to relatively high density and similar type and area of residential units in that small study area. However, it did not apply to large regions or the regions with a variety of property types, such as Florida.

Most of the methods in earlier studies did not consider detailed property types due to the limitation of the ancillary data that were collected for various purposes, but the Tax Parcel data in my study was specifically collected for the purpose of property tax collection, which has a closer relationship to the population density. High-resolution Gridded Population Surface (HGPS) was produced using a similar disaggregating method as Heuristic Sampling Method (Mennis 2003), but Tax Parcel data was more appropriate than the ancillary data used in his study. The property boundaries and types are clearly recorded in Tax Parcel data and can be directly used without extracting and classifying where the errors often occur.

CHAPTER 3
STUDY AREA AND DATA

Study Area

The State of Florida is in the southeastern United States, located on the nation's Atlantic and Gulf coasts. The study area of dasymetric mapping, Alachua County, is located in the north part of the Florida. Both case studies were conducted in the urban area of Gainesville (Figure 3-1), the county seat and largest city in Alachua County, with a population of 176,096, also home to the University of Florida.

Among all the residential areas in the urban area of Gainesville, the single family homes occupy the largest proportion (23.4%) of the total area, where 38.04% of the populations live (Table 3-1). The second largest percentage of the population (25.12%) live in the multi-family homes with 10 units or more that only take up about 2.9% of the total area, where a large difference between the density of single family and multi-family homes can be seen. There is also a large group of people (28%) living in the vacant residential housing and non-residential areas, according to the Census 2010.

Census Data

U.S. Census data are collected every ten years, which consist of three levels including Census tract, block group and block, and count all the residents within administrative units on each level. In contrast to a de facto population, the census population is a de jure population. A de jure population reports all the residents on a basis of their home address regardless of they were physically present there at the reference date. A de facto population reports all the people physically present in the given area at the reference date, which is less stable and collected harder than a de

jure population, but appropriate to some applications such as emergency response in the daytime (Wu 2006).

Tax Parcel Data

Tax Parcel data is a state-wide digital parcel map dataset developed by the State of Florida, which contains the boundary of parcels in all 67 counties of Florida with associated tax information from the Florida Department of Revenue's tax database. Figure 3-1 shows the percentage of several major property types across the Florida, according to which we can see that single family has the largest proportion (70.3%) among all the property types including both residential and non-residential, followed by vacant residential (15.8%), others (5.4%), commercial/industrial (3.2%), agricultural (2.5%), multi-family (1.8%) and non-agricultural acreage (1.1%). Ideally, people only live on residential parcels, but in reality it is often not the case (Wu 2006). Wu mentioned that, according to U.S. Census 2000, about 2.8% (7.8 million) of the total population in the U.S. lived on non-residential types of parcels that is also defined as group quarters in which unrelated groups of people reside, which was not a small enough number to be ignored. Therefore, I decided to consider the group quarters when matching Census with Tax Parcel data, instead of solely residential types.

Table 3-1. Proportion of area and population of different property types in the urban area of Gainesville, Alachua County, Florida

Property Type	Area (%)	Population (%)
Vacant Residential	3.99	3.23
Single Family	23.4	38.04
Mobile Homes	0.72	0.49
Multi-family (≥10 units)	2.9	25.12
Condominium	0.37	1
Retirement Homes	0.03	0.1
Miscellaneous	0.62	0.3
Multi-family (<10 units)	0.61	4.02
Undefined	2.72	2.93
Non-residential Space	64.64	24.77

Note: The total number of the population in this area is 176,096.

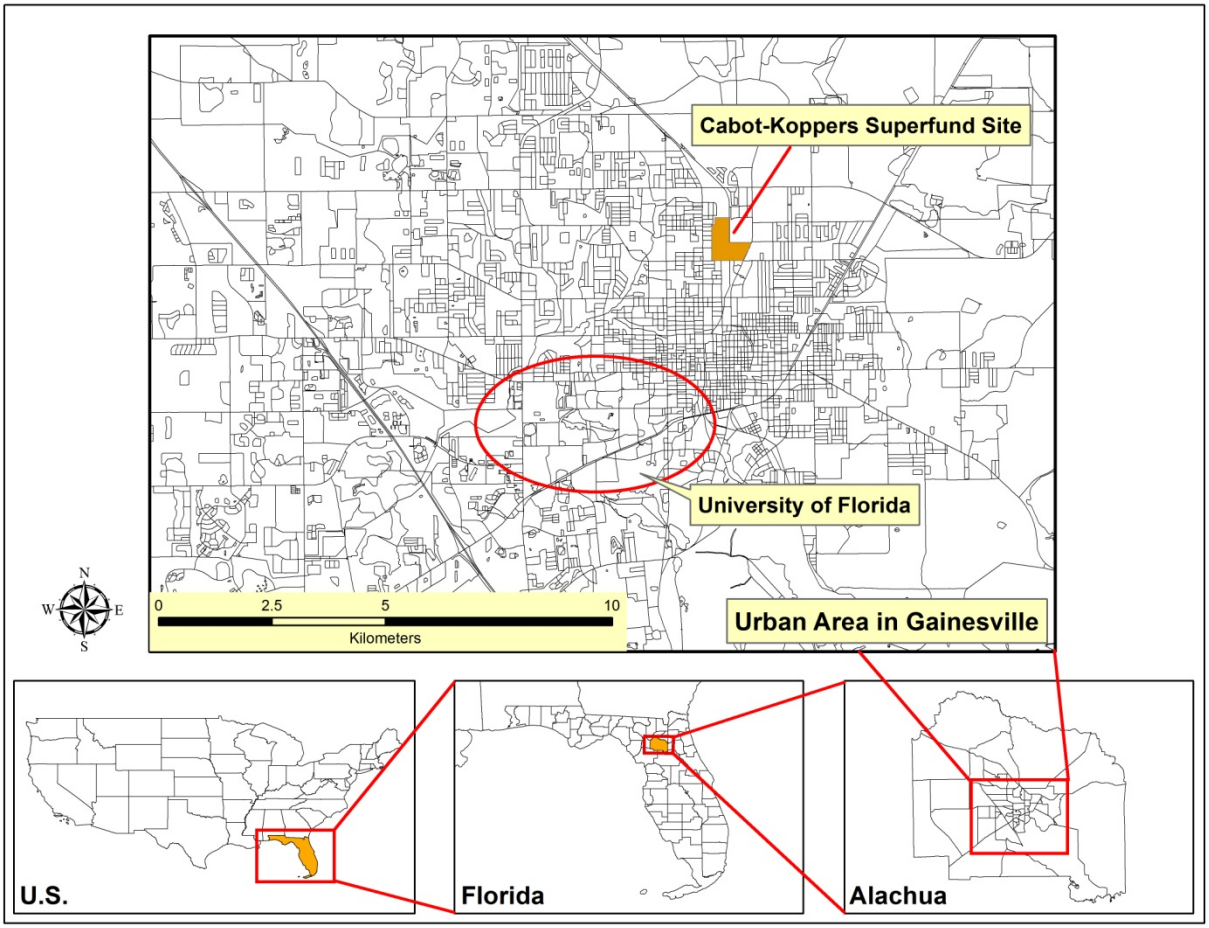


Figure 3-1. Urban area of Gainesville, Alachua County

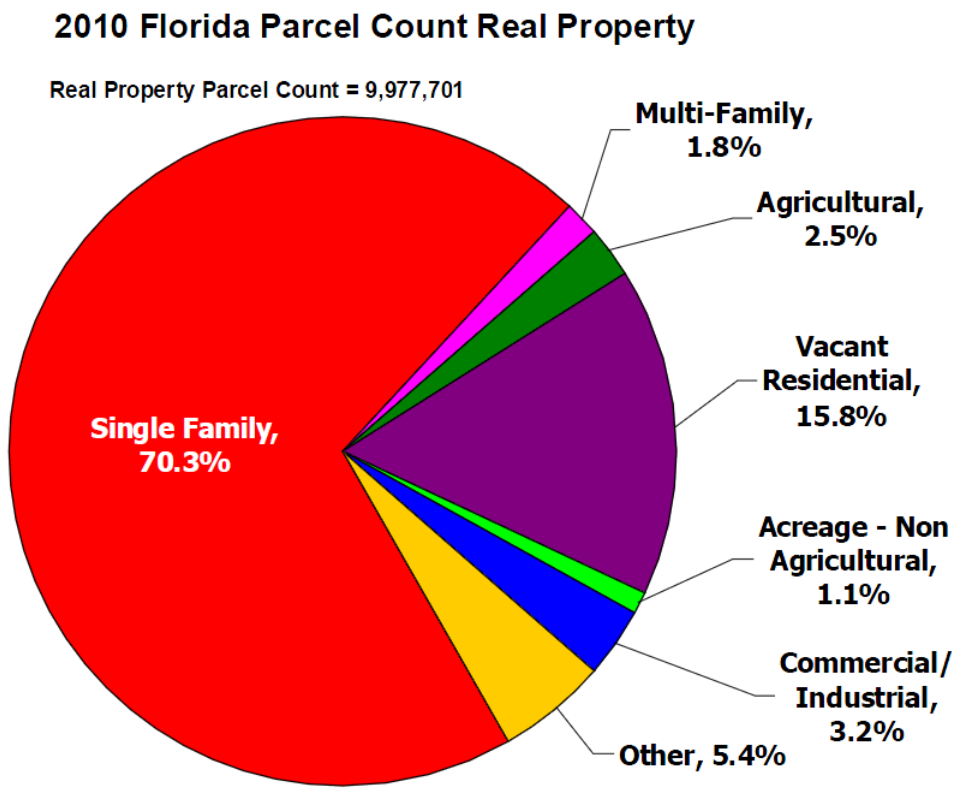


Figure 3-2. Overview of Florida Real Property (Source: Property Tax Oversight, Florida Department of Revenue)

CHAPTER 4
METHODS

For the purpose of preserving pycnophylactic property, meaning that the aggregated estimated values within any unit on the dasymetric map should be equal to the original value of that unit on the choropleth map (Maantay et al. 2007), instead of estimating population from down to up, I intended to redistribute the Census population counts on the block level onto a quadrilateral grid composed of 30m×30m cells by the weights I created. The entire flowchart is shown in Figure 4-1.

In the Tax Parcel data, there is an attribute called CENBLK10 recording which block each parcel belongs to; however, I used the function of spatial intersection in ArcGIS 10.0 (ERSI Inc., Redland CA) to link Census and Tax Parcel data because few parcels matched their corresponding blocks well through my preliminary examination. The parcels with the same type and close to each other were often mapped together as a large parcel, instead of separated appropriately to fit themselves to corresponding blocks, so a large parcel often overlaid more than one block but only one of them was assigned to the parcel, which led to a loss of population on any block that had a non-zero population but was not assigned.

A typical example was shown in Figure 4-2. The shaded community in Figure 4-2A was only divided into two parcels but several blocks, where light red lines and dark black lines represented the boundary of parcels and blocks, respectively. By comparing this vector polygon layer with its high-resolution counterpart on Google Earth (Figure 4-2B and D), the latter clearly showed that the paths and open space within the community were recognized on the Census block map, but hidden on the Tax Parcel map. Wu (2006) also pointed that the Tax Parcel data in the city of Austin, Taxes, did

not depict the city street surface areas. Two blocks to which two parcels were separately assigned were highlighted in dark yellow in Figure 4-2B and the remaining shaded blocks “did not” contain any parcel according to the Tax Parcel records. In Figure 4-2C, all the blocks were colored by population counts and we can see that not all the remaining twelve blocks in Figure 4-2A had zero population; instead, six of them had more than 30 persons, but all of them would be lost and hence the number of the population was underestimated if I used the attribute CENBLK10 to link two sources of data.

Another type of mismatch was shown in Figure 4-3. The dark yellow polygon in Figure 4-3A was a multi-family parcel and black lines depicted block boundaries; green and red lines in Figure 4-3B separately depicted the boundaries of parcels and blocks. By comparing two images, we can see that only about one fourth of the “parcel” is the real parcel and the remaining three fourths were vegetation and water body. In contrast to the Tax Parcel data, Census block data depicted the boundary of parcels more accurately; in other words, the property type was more homogeneous within a block than a “parcel”, but there were no descriptions about the property type in Census block data.

Therefore, I began the dasymetric mapping using the function of spatial intersection in ArcGIS to overlay the block layer on the parcel layer, in order to take advantage of the geographical division of the block data and type information of the parcel data to create a new layer representing all of their intersecting areas with the attribute of property type.

After intersecting, I summed the parcels in each block by property type and calculated the area proportion of each inhabitable type over each block. In addition, I also got the area proportion of each inhabitable type over the total inhabitable area in each block in order for the following weighting process.

Given the proportion of each inhabitable type within blocks, I selected all eligible blocks in a strict way that only the blocks with the proportion of a certain type larger than 50% and the total proportion of the remaining types less than 0.1% were selected for sampling. The little proportion (less than 0.1%) of remaining types in many blocks usually caused by the mismatch between two sets of boundary would not influence the dominance of the mono-type within blocks, so the threshold of 0.1% rather than zero, increased the number of qualified blocks and made the results much more reliable and representative. Next, I found the total population and area of each type and calculated their aggregated population density called the absolute population density, which may be expressed as:

$$\rho_{abs,u} = \sum_{b \in B} P_{ub} \,/\, \sum_{b \in B} A_{ub} \tag{1}$$

where $\rho_{abs,u}$ = absolute population density of the property type u, P_{ub} = population count in block b dominated by type u, A_{ub} = total area of type u in block b, and B represented all the eligible mono-blocks of type u across the Florida.

Then, I merged the types with the similar absolute population density into several classes, normalized all merged absolute densities and got the relative density for each class, also called population density fraction. This step may be expressed as:

$$\rho_{rel,v} = \rho_{abs,v} \,/\, \sum_{x \in V} \rho_{abs,x} \tag{2}$$

where $\rho_{rel,v}$ = relative population density of class v, $\rho_{abs,v}$ = absolute population density of class v, $\rho_{abs,x}$ = absolute population density of class x, and V represented all the classes after merging.

The population counts were transferred from blocks to each inhabitable intersecting zone based on the weight of each parcel, which was calculated in the following way:

$$\omega_{rel,p} = (\rho_{rel,v} \times \alpha_{pb}) / \sum_{x \in P} (\rho_{rel,v} \times \alpha_{xb}) \quad (3)$$

where $\omega_{rel,p}$ = weight of parcel p, $\rho_{rel,v}$ = relative population density of class v that the parcel p belonged to, α_{pb} = the area ratio of parcel p over the total inhabitable area in block b, α_{xb} = the area ratio of parcel x over the total inhabitable area in block b, and P represented all the inhabitable intersecting parcels in block b.

The population count on each block was redistributed to all inhabitable intersecting zones within the block and each zone got a certain amount of population as following:

$$P_p = P_b \times \omega_{rel,p} \quad (4)$$

where P_p = population count on parcel p, P_b = population count in block b in which parcel p was located, and $\omega_{rel,p}$ = weight of parcel p.

In order to transfer population counts from intersecting zones to each grid cell within or intersecting with them, I created a grid of 30m×30m cells covering the entire Florida and intersected it with the layer of intersecting zones. Under the assumption that the population was evenly distributed within each intersecting zone, the population count in each grid was calculated as following:

$$P_g = \sum_{x \in I} (P_x \times \alpha_{gx}) \quad (5)$$

where P_g = population count in grid g, P_x = population count in parcel x, α_{gx} = the area ratio of the overlaid area by grid g and parcel x over the area of parcel x, and I represented all the intersecting zones overlaid with grid g. Till here, the HGPS in Alachua County, Florida, with 30m×30m quadrilateral cells has been created.

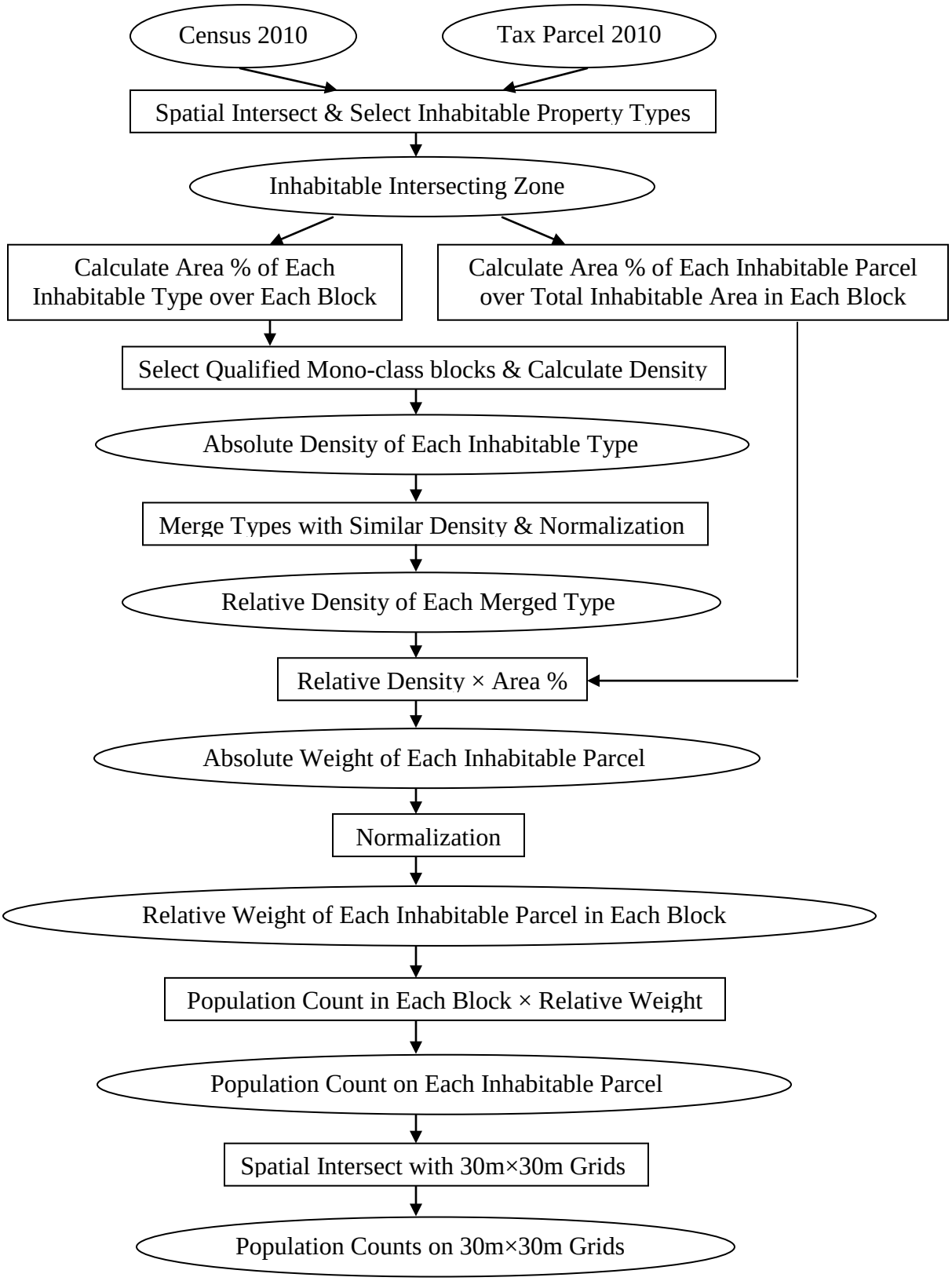


Figure 4-1. Flowchart of dasymetric mapping

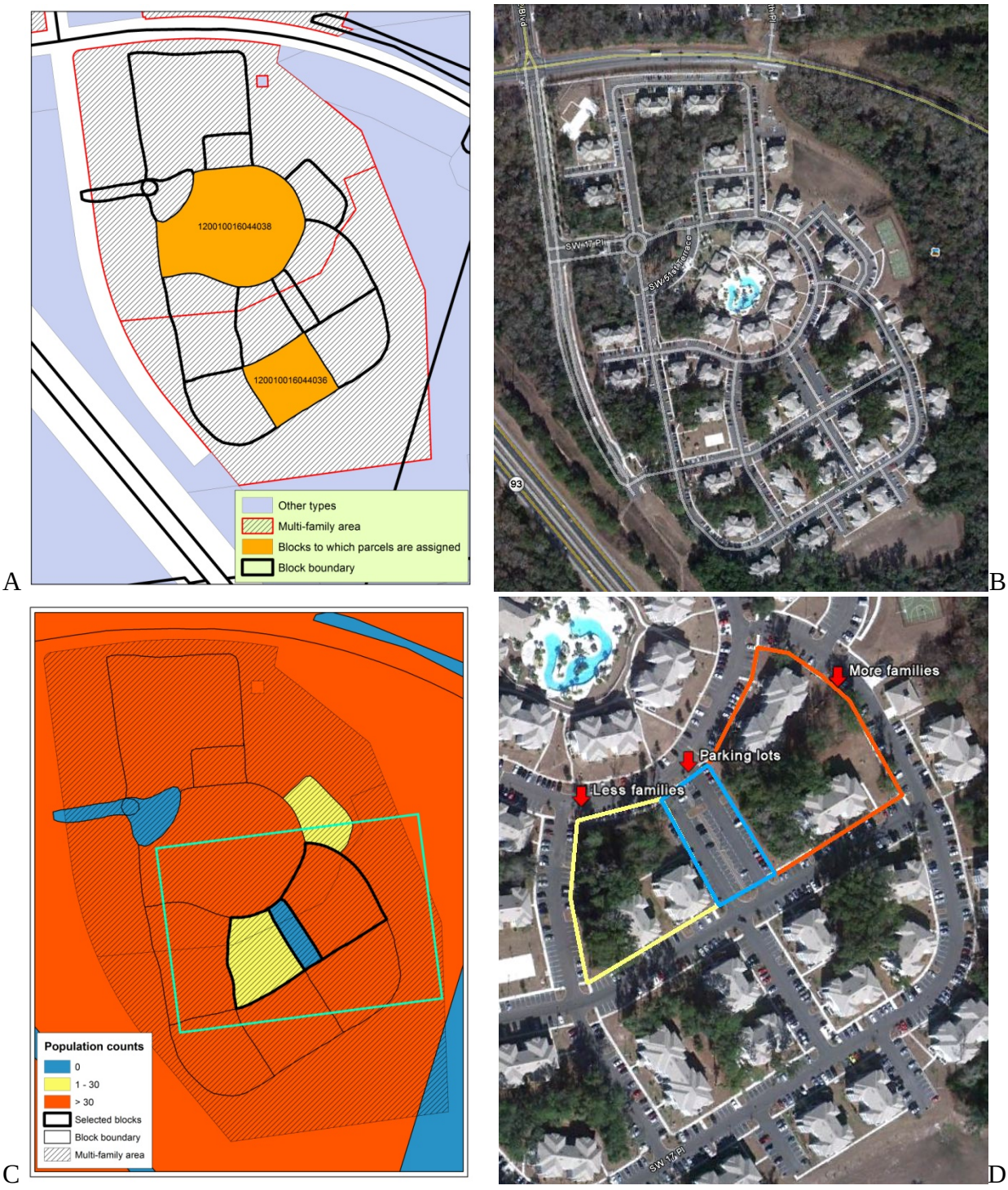
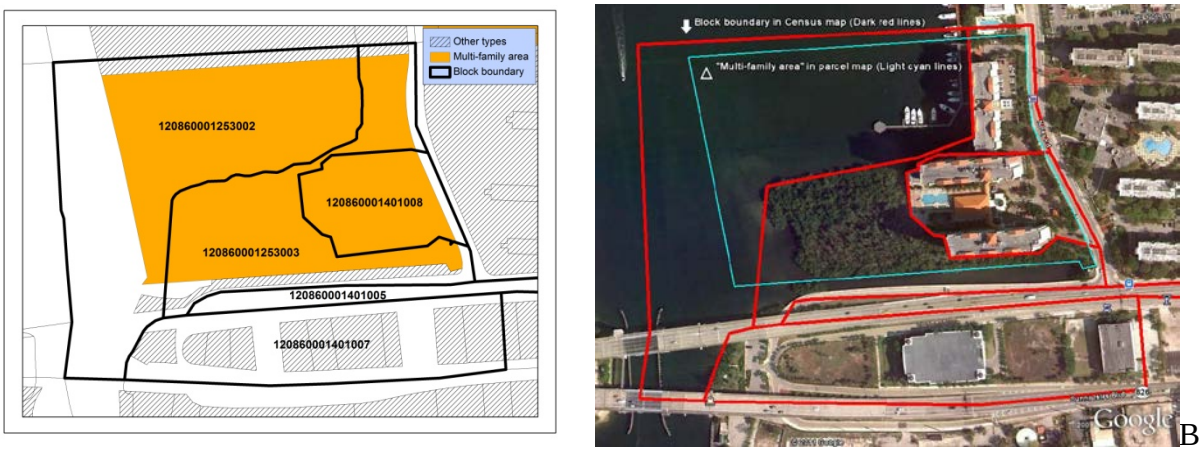


Figure 4-2. Mismatch of boundary between two sources of data



A B

Figure 4-3. Mismatch of boundary between two sources of data

CHAPTER 5
RESULTS

Population Density Fraction

After examining the population density in all qualified mono-type blocks and the definitions of property types in Tax Parcel data, I selected 33 types as the inhabitable parcels in this study (Table 5-1). The classification scheme in Tax Parcel data is very detailed, but has not been taken full advantage of in earlier studies to assist dasymetric mapping. Compared to the three classes, high-density urban, low-density urban and non-urban areas extracted from Landsat TM imagery in the Heuristic Sampling Method (Mennis, 2003) and nine merged classes in Wu's study (2006), including single family, multi-family, commercial, office, industrial, civic, open space, transportation and undeveloped, I involved more independent property types in this study.

Some of the 33 property types have a small number of parcels in Alachua County, which led to insufficient eligible blocks for them. In addition to Alachua County, therefore, I examined the population density on each property type in other four counties with a big city and complex types of parcels there, including Orange County (Orlando), Miami-Dade County (Miami), Duval County (Jacksonville) and Hillsborough County (Tampa).

Although the absolute population density of the same inhabitable type varied a little among five counties, that difference was less significant than the difference of the density among various residential types in any county; moreover, I found that the relative differences among different types in each county were more stable in five counties. Consequently, I applied a set of densities from the entire Florida to Alachua County, which is also easy for future use under the assumption that the accuracy

brought by sampling county by county is further less comparable to the time and cost spent on sampling in all 67 counties of Florida.

I sampled the mono-type blocks across the entire Florida and calculated the absolute aggregated density of each type (Table 5-2). The types with the similar absolute aggregated density were collapsed such that the initial Tax Parcel map was reclassified from 33 original inhabitable categories into 14 classes, the absolute density of which, then, were normalized and transformed into the relative density for each class, called density fraction (Table 5-3).

In Table 5-2, as I expected, the population density in the multi-family (003 and 008) is higher than other residential types, followed in order by the retirement homes (006), condominium (004), single family (001), cooperatives (005), vacant residential lots (000), mobile homes (002), miscellaneous residential lots (007) and undefined residential lots (009). Several non-residential types also have a high density, among which the private schools and colleges (072) have the highest density than any other residential or non-residential type, and the density of privately owned hospitals (073) is higher than condominium but lower than retirement homes; in addition, the homes for the aged (074), office buildings (018), mobile home parks (028) and undefined government lots (080) have the similar density as single family.

Although there is only one eligible block for retirement homes, the density on it is reasonable as a residential type. Actually, there are some other types which have a small amount of mono-type blocks, such as miscellaneous and undefined residential lots (007 and 009), stores (011), office buildings (018), private schools and colleges (072), sanitariums (078), undefined government lots (080), military lots (081), public

schools and hospitals (083 and 085). Due to few amount and small area, any of those types rarely occupy an entire block exclusively, unlike single family; however, I considered the Census as a reliable data source, so the density of any property type would be more than zero only if at least one eligible block with non-zero density exists, and there is no better way to find out a more reliable density but accepting the one from few samples.

Dasymetric Mapping

The distribution of population density in the entire Alachua County is shown in both choropleth (Figure 5-1) and dasymetric way (Figure 5-2) for comparison, where a large difference in the spatial patterns of the population distribution can be seen, particularly in two magnified areas. As shown in Figure 5-1, the population was evenly distributed nearly everywhere across the entire county; however, in Figure 5-2, the clustering of inhabitable parcels emerged in the middle area of Alachua County and some were sparsely spread in the surrounding, which explains the dissimilar spatial patterns and might have misled some spatial analysis.

It is worth noting that the significant contrast of spatial patterns in the magnified areas between Figure 5-1 and 5-2 is partly attributed to the fact that they are located in the suburb rather than city. The improvements of dasymetric maps are mainly embodied in two aspects, spatial patterns and intra-block heterogeneity of the population distribution. The changes in spatial patterns are more apparent in the country or suburb, especially visually, than in the city because most areas in the city are already highly populated; however, the geographical heterogeneity in the city would expect to be more obvious, as shown in the following case studies in the urban area of Gainesville.

Case Study 1: Estimation of Bus Passengers

Regional Transit System (RTS) provides high-quality bus transportation services in the city of Gainesville, Florida. From the website of RTS (http://ufl.transloc.com/), I can get the records of how many passengers served per month in 2010 by each route. In addition, the real-time function of locating buses on RTS website allows me to get the courses of all the routes with the background of the city map, so I can exactly digitize all the bus routes.

Although Gainesville RTS maintains 31 city routes and nine campus routes during the weekdays in 2010, 13 out of 31 city routes covering most of main roads in the city of Gainesville were examined in detail: Routes 2, 5, 8, 9, 10, 15, 20, 21, 23, 24, 29, 35 and 36 (Figure 5-3). The similar routes were considered as one route and the numbers of population served by each of them were summed. For example, in Figure 5-4, except the pure blue line curving towards southeast from the northwestern terminal (Oak's Mall) to the first apparent intersection (between SW 62nd Blvd and SW 20th Ave) which is exclusively served by Route 20, the major parts of the routes Route 20 and 21 run overlay from that intersection to the eastern terminal (Rawlings Hall at University of Florida) where the color of blue and orange alternate. Therefore, Route 20 and 21 were merged into one.

One of another two integrated routes was merged by Route 8 and 29 (Figure 5-5), where most of the course Route 29 (dark yellow) runs belong to one part of the course Route 8 (orange) runs; the other integrated route consisted of Route 9 (green), 35 (blue) and 36 (purple) in Figure 5-6, where three routes run the similar course at the north of the west-east major road (Archer Road) and serve a large residential area together at

the south. The routes of Route 2, 5, 10, 15, 23, and 24 are relatively separate from each other, so each of them was taken as an independent route without merging.

The actual numbers of passengers served by each route in February, March, September, October and November downloaded from RTS website were given in Table 5-4. University of Florida is a major part of Gainesville, where many students and employees take the bus to the campus for studying or working. Considering the fact that some local people may probably go out of the town for travelling during vacations, and most of the students in UF may go out for exchange study and internships during summer vacations and go home for spending holidays during winter vacations, I only examined those five months when most of people and students were usually at work or study regularly, and 22 weekdays were counted for one month.

Most of transit planners and researchers universally accept that the 400m (1/4 mile) is the maximum walking distance that most potential passengers are willing to travel to reach their nearest bus stop (Biba et al. 2010). The bus stops are densely placed along the bus routes in Gainesville, so I can easily take the buffer zone of any course as the sum of the buffer zones of all bus stops along that course. For the sake of simplicity, it was assumed that only the people living in multi-family housing and condominium had a demand for bus ride.

Two methods were used to estimate the numbers of population within the 400m buffer zone: the first one was to overlay the buffer zone with the traditional population grids produced by the areal weighting method (AWM), and then sum the values of all the grid cells within the buffer zone, also called the U.S. Census Grids method; the second one was the dasymetric way, that is, to overlay the buffer zone with HGPS and

sum the values of all the grid cells within the buffer zone. The same last step for both methods was to extract the population counts exclusively in the area of multi-family and condominium parcels, that is, the numbers of potential passengers counted in this study. The results before and after clipping from both methods were summarized in Table 5-4.

As expected the overestimation from USCG is quite exaggerative for all routes; however, the results from HGPS shows the same scale of overestimation. After clipping, the estimated population numbers of each route lower dramatically as a whole regardless of the method; in particular, for the two merged routes, Route 20&21 and Route 9&35&36, the estimated numbers by the traditional method even decline below the actual numbers of passengers served in several months, but still fairly overestimated by the dasymetric method. For the other seven routes, while overestimated by both methods, the numbers estimated by the traditional method, rather than dasymetric method, are much closer to the actual numbers of passengers.

For the sake of visual comparison, the choropleth and dasymetric map are shown in Figure 5-7 and 5-8. It can be seen that there were more zero-population areas spread amid blocks in the dasymetric map (white area) and more intra-block variations in population density, which have been completely hidden by constant intra-block values in the traditional map.

Case Study 2: Estimation of Population at Risk to Environmental Pollution

The Superfund is an environmental program created by the Comprehensive Environmental Response, Compensation and Liability Act of 1980 to clean up abandoned hazardous waste sites. A Superfund site is a hazardous waste site placed on the U.S. Environmental Protection Agency's (USEPA) National Priority List (NPL)

after a current or potential health impact assessment. In this study I selected one of the sites on the NPL within my study area to explore the application of the HGPS in environmental and public health.

The Cabot-Koppers Superfund Site is located at the northwestern corner of the intersection between Northwest 23rd Avenue and North Main Street in the City of Gainesville, consisting of two portions (Figure 5-9). The Koppers Site, covering 82 acres on the western side, was polluted with wood treating chemicals in soils and groundwater; and the Cabot Site on the eastern side covers 49 acres and contains groundwater contamination from pinetar, pine oil and charcoal production.

In a study for the benefits and costs of remediation at 150 Superfund sites (Hamilton and Viscusi 1999), it has been found that the cost of averting one cancer case is over $100 million due to the exceedingly small expected number of cancers prevented by remediation. Therefore, it is of importance to estimate the number of population exposed to the major affected media prior to any remediation, which is also imperative to assess the cost effectiveness of the remediation.

Given the numbers of surrounding population within the buffer with a 0.5, 2 and 4 miles radius centered on the Cabot-Koppers Site from the EPA Region 4 Reuse Fact Sheets (Table 5-5), I calculated the total population and solely the population living in residential property within those buffers, separately, by both USCG and HGPS for comparison, which were the same two methods used in the first case study. The results were summarized in Table 5-5.

From the Figure 5-9, we can see that none of residential parcels are located within the Cabot-Koppers Site; hence the population there should be few, which indicates that

the HGPS result is more reasonable than USCG. For the estimated numbers of total population within the buffer with three radii, although faintly, the HGPS results are closer to the numbers in 2007 from the EPA Sheets except within the buffer with a 0.5 mile radius; however, both HGPS and USCG were created based on Census 2010 and the population may possibly increase slightly from 2007 to 2010, so it is quite possible that the slight overestimation of 1.57% is even less than the increasing rate during those three years. Compared to USCG, the HGPS produced not only closer but also more rational and stable results with an overestimation approximately ranging from 3% to 5%.

The Floridan Aquifer under the Cabot-Koppers Site, into which the soil contamination would leach, is the source of drinking water for over 175,000 people in Alachua County. Assuming that only the people living in residential types of housing near the Cabot-Koppers Site are regularly supplied by contaminated groundwater and steadily affected by the soil pollution, that portion of the people probably at risk was separated from the people living in non-residential types of housing and their numbers were recalculated using both methods. The differences between two sets of results are more apparent and the results from HGPS are obviously more sensible than USCG that only depends on the area ratio of the housing over the entire block.

In addition to the distinctions of number, the spatial distribution of the population at risk for contamination within the buffer with a 0.5, 2 and 4 miles radius centered on the Cabot-Koppers Site is separately shown in Figure 5-10, 5-11 and 5-12, where all the spatial units with non-zero values and the finest resolution on their own levels are drawn. It is worth noting that, for the purpose of preserving the volume within original units, all the break values in the legend representing population counts are shown to the

nearest hundredth place because those values are the numbers calculated by multiplying the total population counts with the weight of each sub-unit (sub-block for choropleth maps; grid for dasymetric maps).

The visual comparison between choropleth and dasymetric maps indicates not only somewhat different spatial patterns of the population, but also more detailed intra-block variations of the population on dasymetric maps, which have been completely hidden on the traditional choropleth maps.

In Figure 5-10, for example, due to a large scale, it can be seen that 1) there are lots of people on the choropleth map but few on the dasymetric one within the region of Cabot-Koppers Site; 2) more people on the choropleth map than the dasymetric one live in the northeastern part of the buffer regions; 3) more detailed intra-block variations are observed nearly everywhere on the dasymetric map than the choropleth one.

Table 5-1. Inhabitable property types

Class	Sub-class	Code
Residential Property	Vacant Residential	000
	Single Family	001
	Mobile Homes	002
	Multi-family - 10 units or more	003
	Multi-family - less than 10 units	008
	Condominia	004
	Cooperatives	005
	Retirement Homes	006
	Miscellaneous Residential	007
	Undefined - Reserved for Use by Department of Revenue	009
Commercial Property	Vacant Commercial	010
	Store	011
	Mixed use - store and residential or residential combination	012
	Office buildings	018
	Mobile home parks	028
Industrial Property	Vacant Industrial	040
Agricultural Property	Improved agricultural	050
Institutional Property	Vacant	070
	Churches	071
	Private schools and colleges	072
	Privately owned hospitals	073
	Homes for the aged	074
	Orphanages, other non-profit or charitable services	075
	Sanitariums, convalescent and rest homes	078
Government Property	Undefined - Reserved for future use	080
	Military	081
	Public county schools	083
	Colleges	084
	Public Hospitals	085
	Counties	086
	State	087
	Federal	088
	Municipal	089

Table 5-2. Absolute aggregated density

Code	Type	Block	Pop	Area (km^2)	Density (persons/100m^2)	Mean (persons/100m^2)	Std. (persons/100m^2)
000	Vacant	98	3,902	1.94	0.2	0.54	0.79
001	Single family	11,130	581,343	193.28	0.3	0.33	0.2
002	Mobile home	308	6,469	3.65	0.18	0.24	0.18
003	Multi-fml(≥10)	457	65,369	7.81	0.84	1.06	1.3
008	Multi-fml(<10)	69	4,119	0.6	0.69	0.73	0.61
004	Condominia	144	5,953	1.62	0.37	0.48	0.66
005	Cooperative	29	912	0.32	0.29	0.28	0.2
006	Retirement	1	293	0.06	0.49	--	--
007	Miscellaneous	6	58	0.08	0.07	0.77	1.33
009	Undefined	7	24	0.04	0.06	0.44	0.42
010	Vacant	23	560	0.26	0.21	0.26	0.18
011	Store	2	7	0.04	0.02	0.03	0.02
012	Mixed use	11	200	0.12	0.16	0.2	0.16
018	Office building	4	170	0.07	0.25	0.21	0.35
028	Mobile home	839	25,986	9.5	0.27	0.33	0.38
040	Vacant	0	--	--	--	--	--
050	Agricultural	78	311	11.33	0.003	0.03	0.09
070	Vacant	0	--	--	--	--	--
071	Church	18	150	0.23	0.07	0.07	0.08
072	Private School	7	1,817	0.21	0.87	1.76	0.91
073	Private Hospital	15	2,574	0.6	0.43	0.81	0.72
074	Aged Home	22	1,648	0.56	0.29	0.52	0.64
075	Charitable	0	--	--	--	--	--
078	Sanitarium	1	115	0.06	0.18	--	--
080	Undefined	8	599	0.26	0.23	0.46	0.37
081	Military	2	60	0.28	0.02	0.57	0.55
083	Public school	7	162	0.24	0.07	0.16	0.21
084	College	11	3,171	3.35	0.09	0.77	1.19
085	Public Hospital	2	35	0.06	0.06	0.27	0.26
086	County	39	1,891	1.51	0.13	0.34	0.38
087	State	29	10,028	17.59	0.06	0.86	2.07
088	Federal	79	4,345	18.93	0.02	0.13	0.12
089	Municipal	63	1,989	1.17	0.17	0.35	0.33

Table 5-3. Relative density

Code	Density (persons/100m^2)	Density Fraction
072	0.87	0.18657
003	0.84	0.18014
008	0.69	0.14797
006	0.49	0.10508
073	0.43	0.09222
004	0.37	0.07935
001,005,074	0.3	0.06434
018,028,080	0.25	0.05361
000,002,010,012,078,089	0.2	0.04289
086	0.13	0.02788
007,009,071,083,084,085,087	0.07	0.01501
011,081,088	0.02	0.00429
050	0.003	0.00064
040,070,075	0.00001	0.000002

Table 5-4. Numbers of actual passengers and estimated population within the buffer zone of each route

Route	20&21	9&35&36	5	15	2	8&29	24	23	10
Actual passengers (persons)									
February	6,471	6,869	1,638	882	322	1,273	303	--	349
March	5,786	6,072	1,589	1,001	369	1,261	355	--	353
September	8,498	8,682	1,824	1,179	394	1,673	395	246	492
October	7,780	7,840	1,724	1,126	390	1,474	392	229	456
November	6,705	6,845	1,498	1,008	322	1,318	364	201	402
Average	7,048	7,262	1,654	1,039	359	1,400	362	225	410
Estimated total population (persons)									
AWM	17,132	20,353	13,340	10,201	7,201	22,654	8,961	4,167	14,838
HGPS	17,530	20,060	13,180	11,173	7,335	22,485	9,122	3,955	14,701
Overestimation (%)									
AWM	143.1	180.3	706.5	881.8	1,906	1,518	2,375	1,752	3,519
HGPS	148.7	176.2	696.9	975.4	1,943	1,506	2,420	1,658	3,486
Estimated population living in multi-family and apartments (persons)									
AWM	4,576	7,148	2,958	1,295	933	3,605	690	1,369	2,843
HGPS	7,296	10,789	3,942	2,953	1,368	6,050	1,696	1,922	4,316
Overestimation (%)									
AWM	-35.1	-1.6	78.8	24.6	159.9	157.5	90.6	508.4	593.4
HGPS	3.5	48.6	138.3	184.2	281.1	332.1	368.5	754.2	952.7

Table 5-5. Numbers of actual population and estimated population within the buffer zone with different radii

Radius (miles)	0 (on site)	0.5	2.5	4
Actual population	-	4,274	55,595	97,670
Total population (persons)				
AWM	202	4,341	58,071	101,342
HGPS	20	4,486	57,658	100,745
Difference (persons, HGPS-USCG)		145	-413	-597
Overestimation (%)				
AWM	-	1.57%	4.45%	3.76%
HGPS	-	4.96%	3.76%	3.15%
Population living in residential property (persons)				
AWM	-	2,526	28,625	51,993
HGPS	-	3,638	37,932	70,521
Difference (persons, HGPS-USCG)		1,112	9,307	18,528

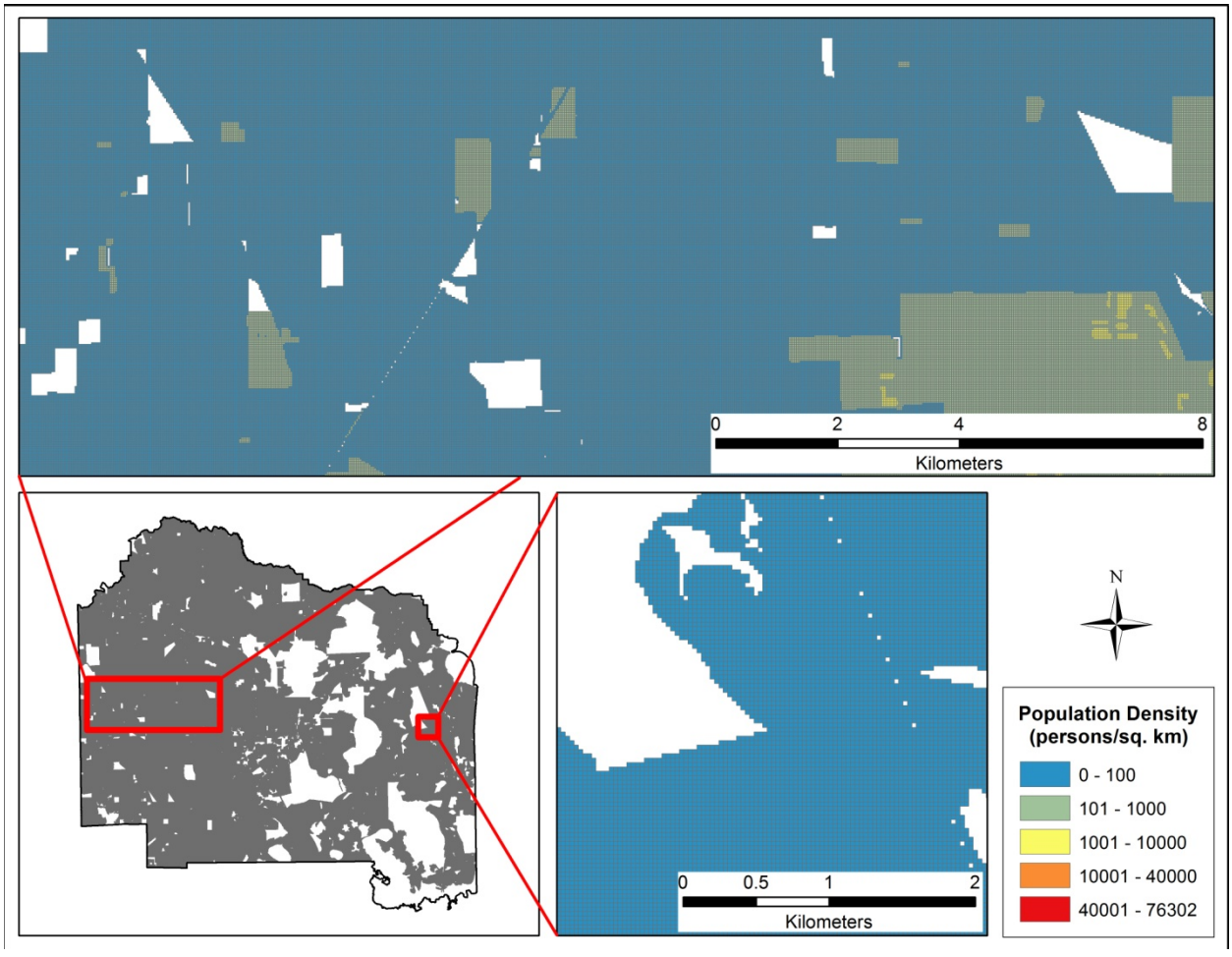


Figure 5-1. Choropleth map of population density in Alachua County

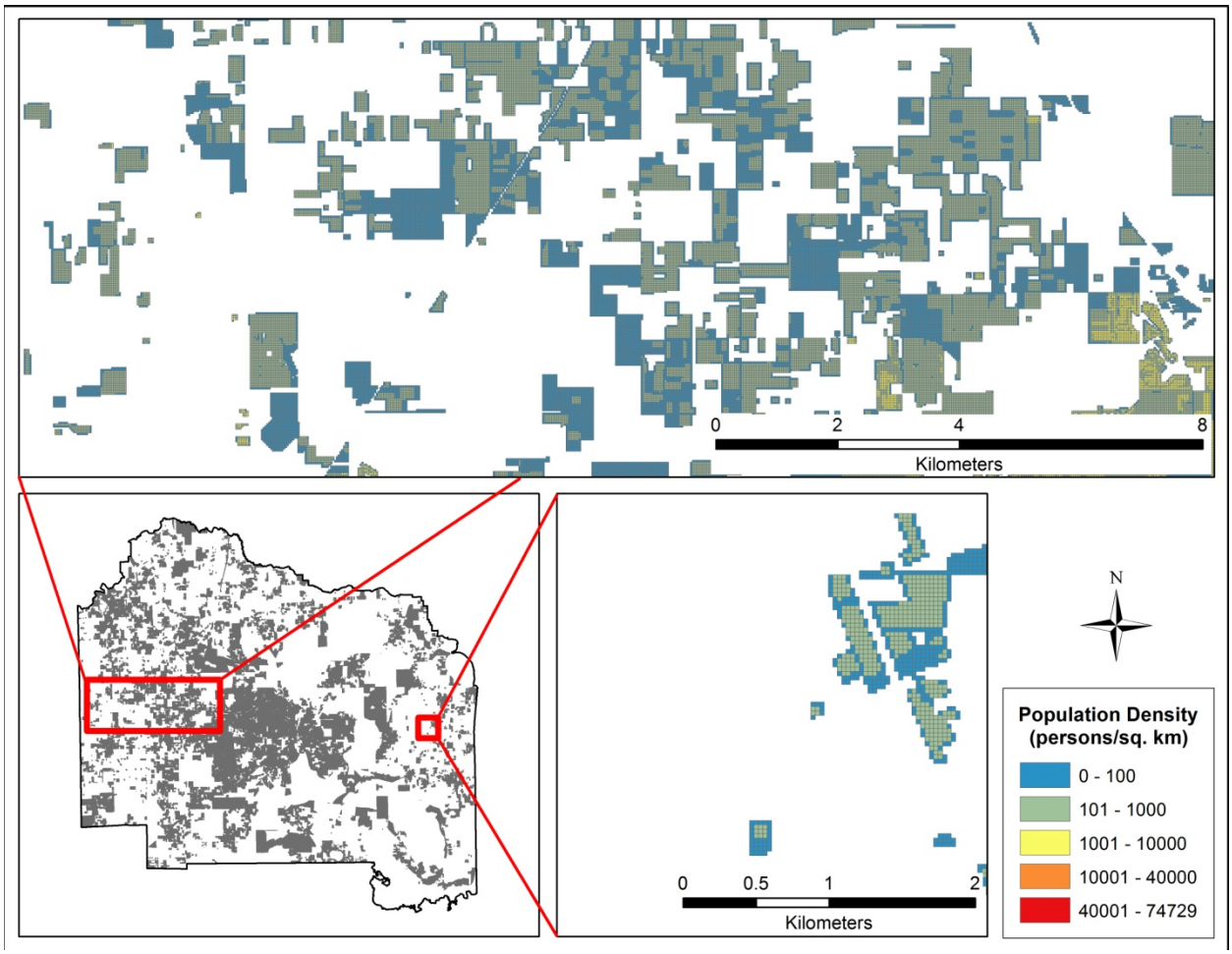


Figure 5-2. Dasymetric map of population density in Alachua County

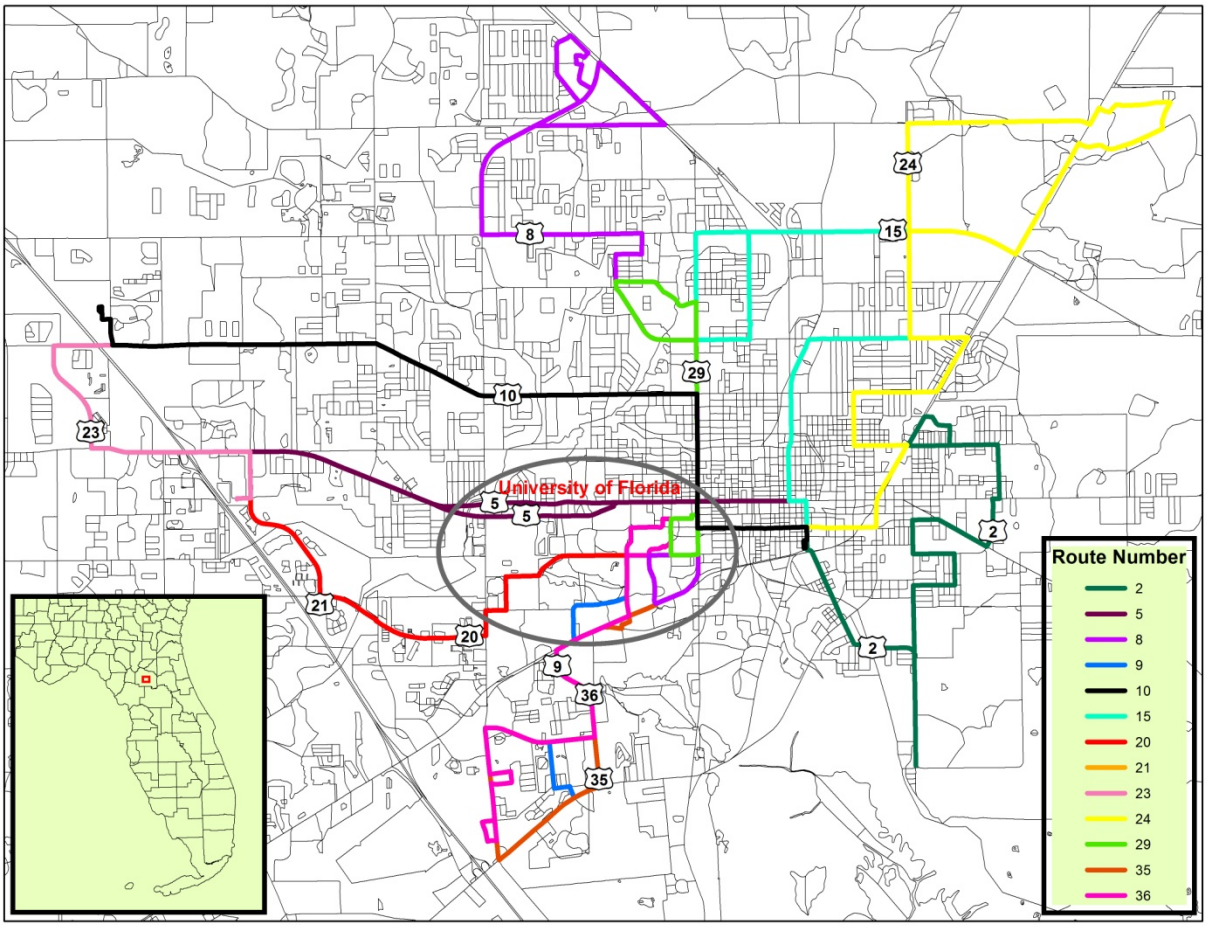


Figure 5-3. Courses of 13 bus routes

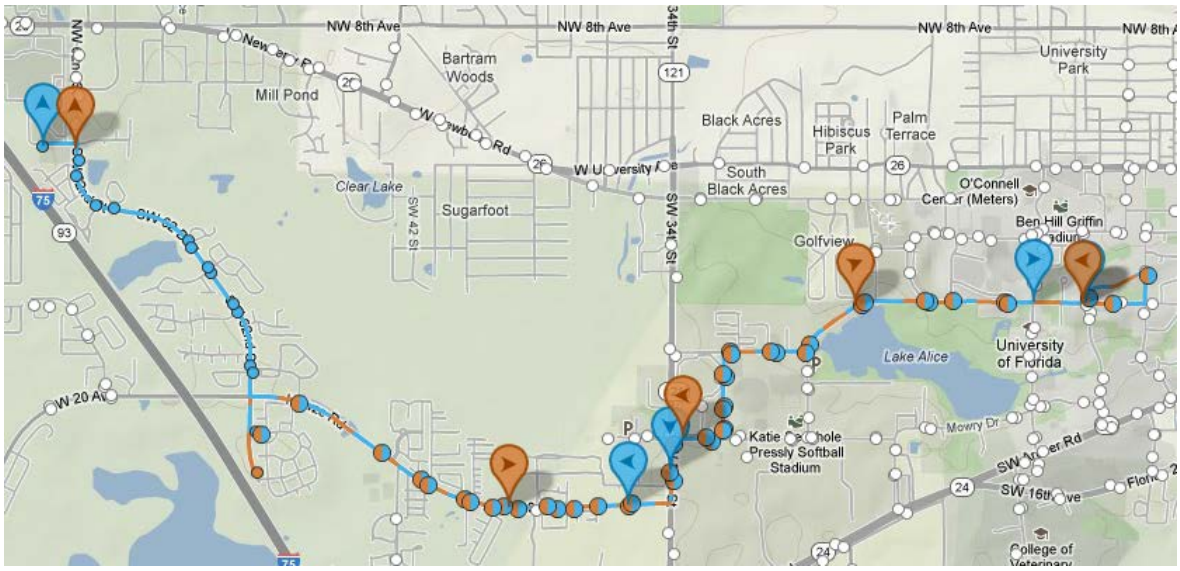
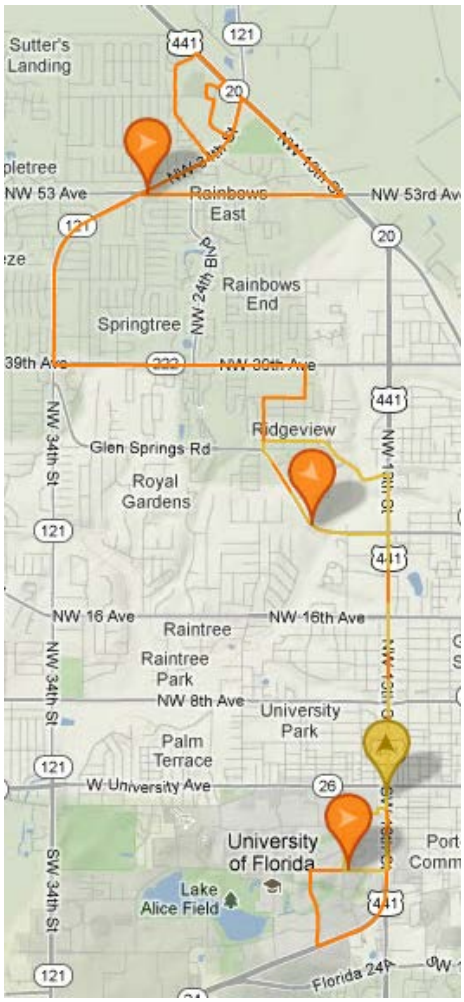


Figure 5-4. Routes 20 (blue) and 21 (orange)



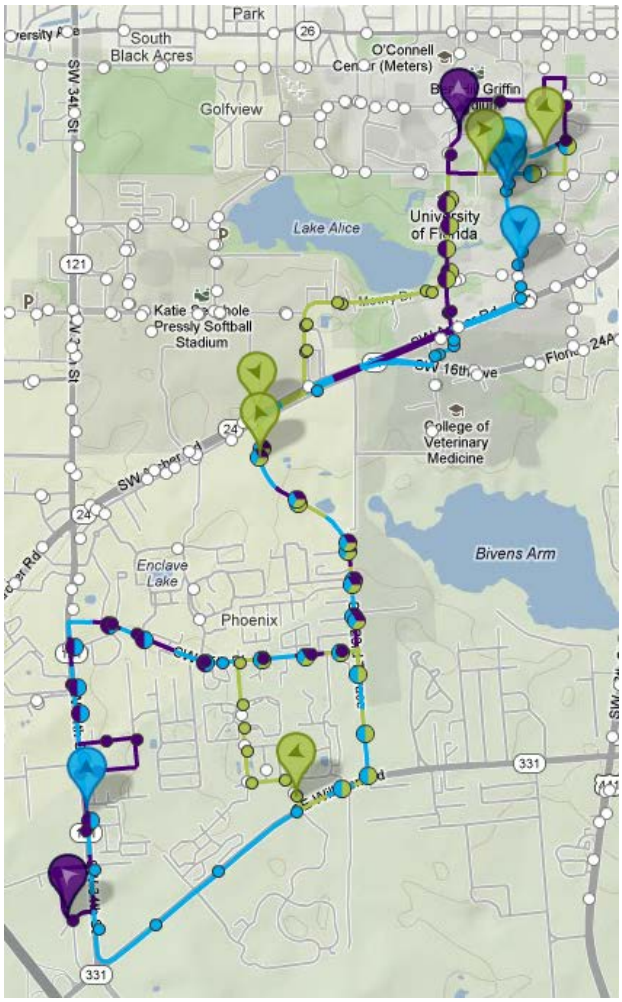


Figure 5-5 (left). Route 8 (orange) and 29 (dark yellow)

Figure 5-6 (right). Route 9 (green), 35 (blue) and 36 (purple)

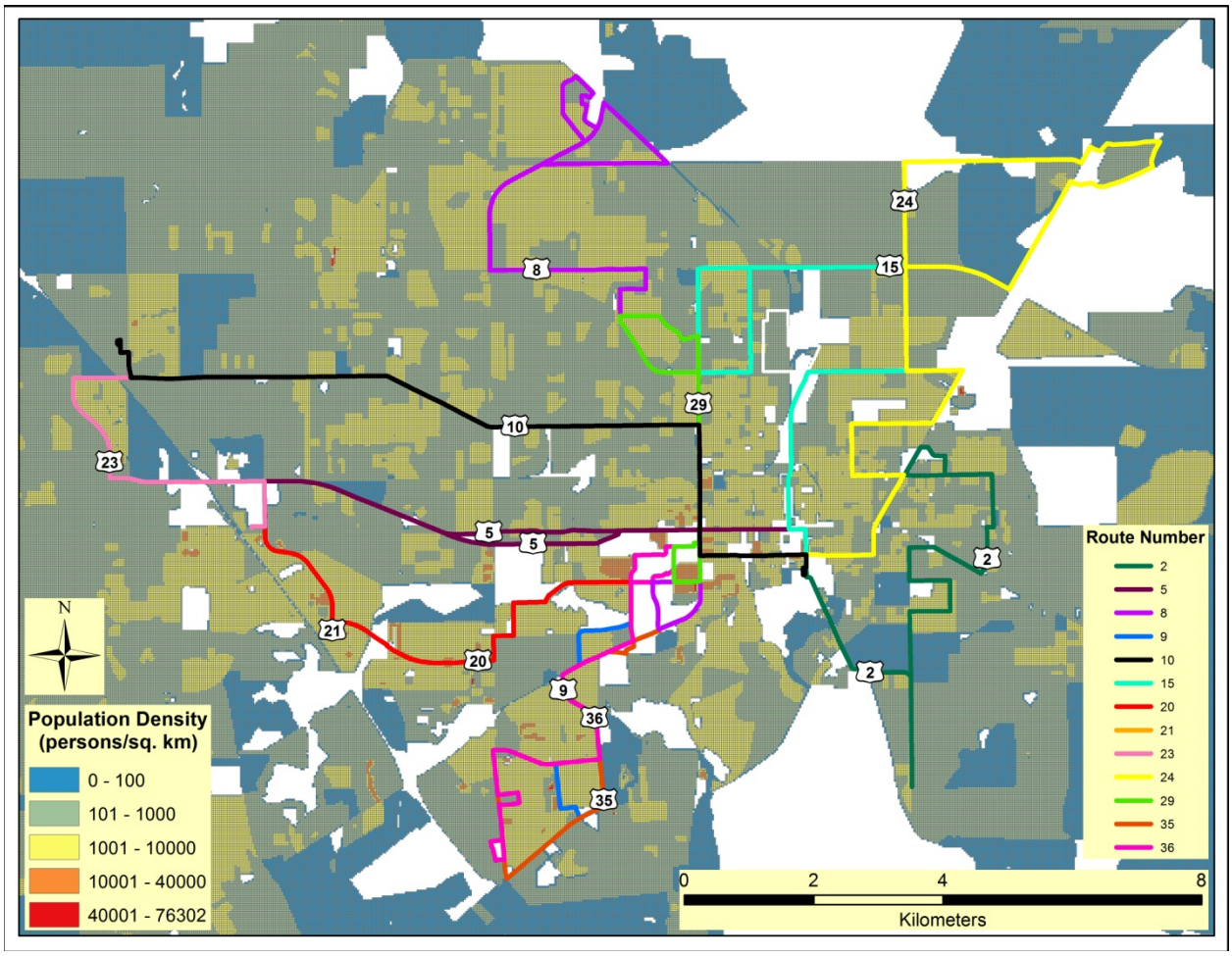


Figure 5-7. Choropleth map of population density in the urban area of Gainesville

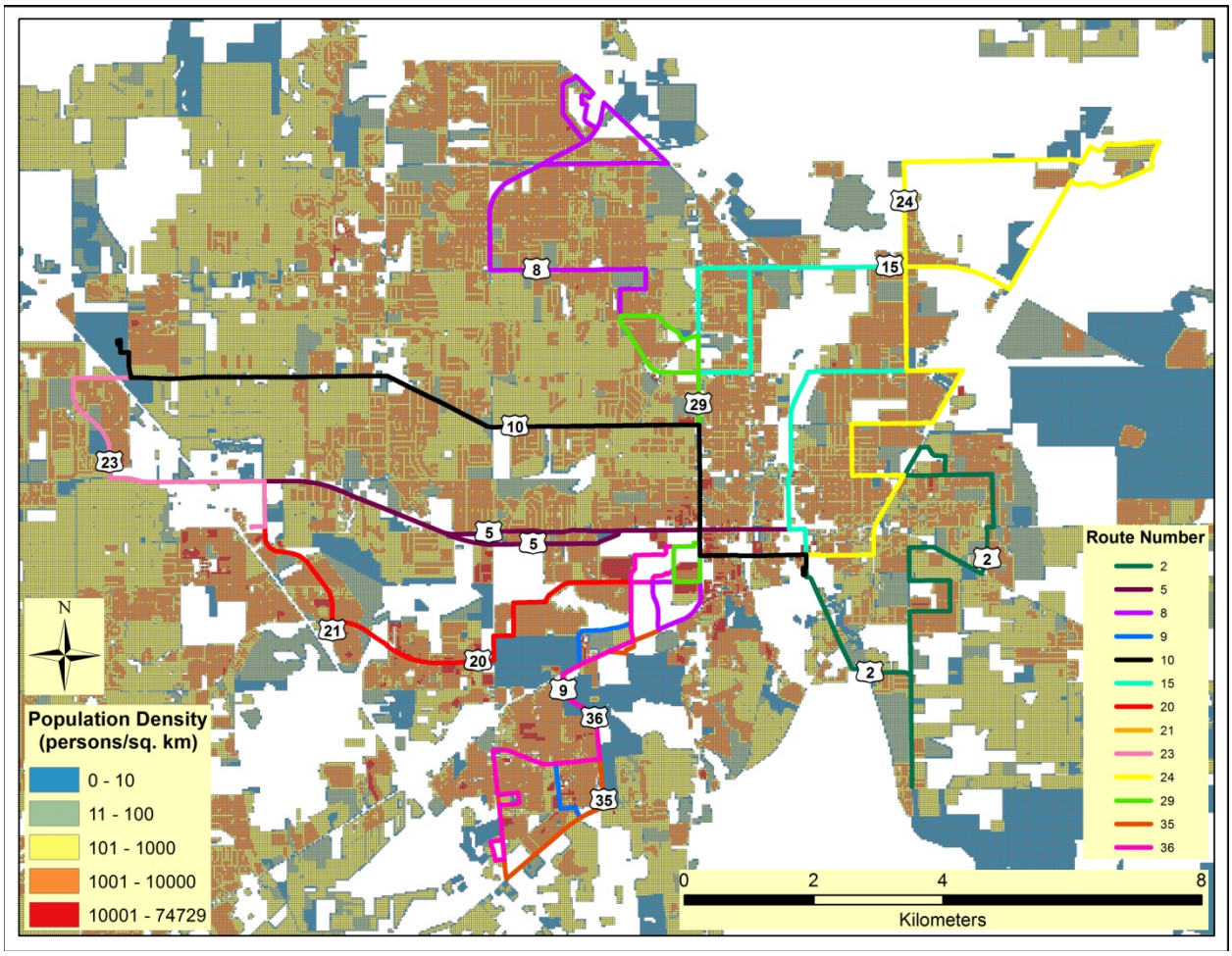


Figure 5-8. Dasymetric map of population density in the urban area of Gainesville

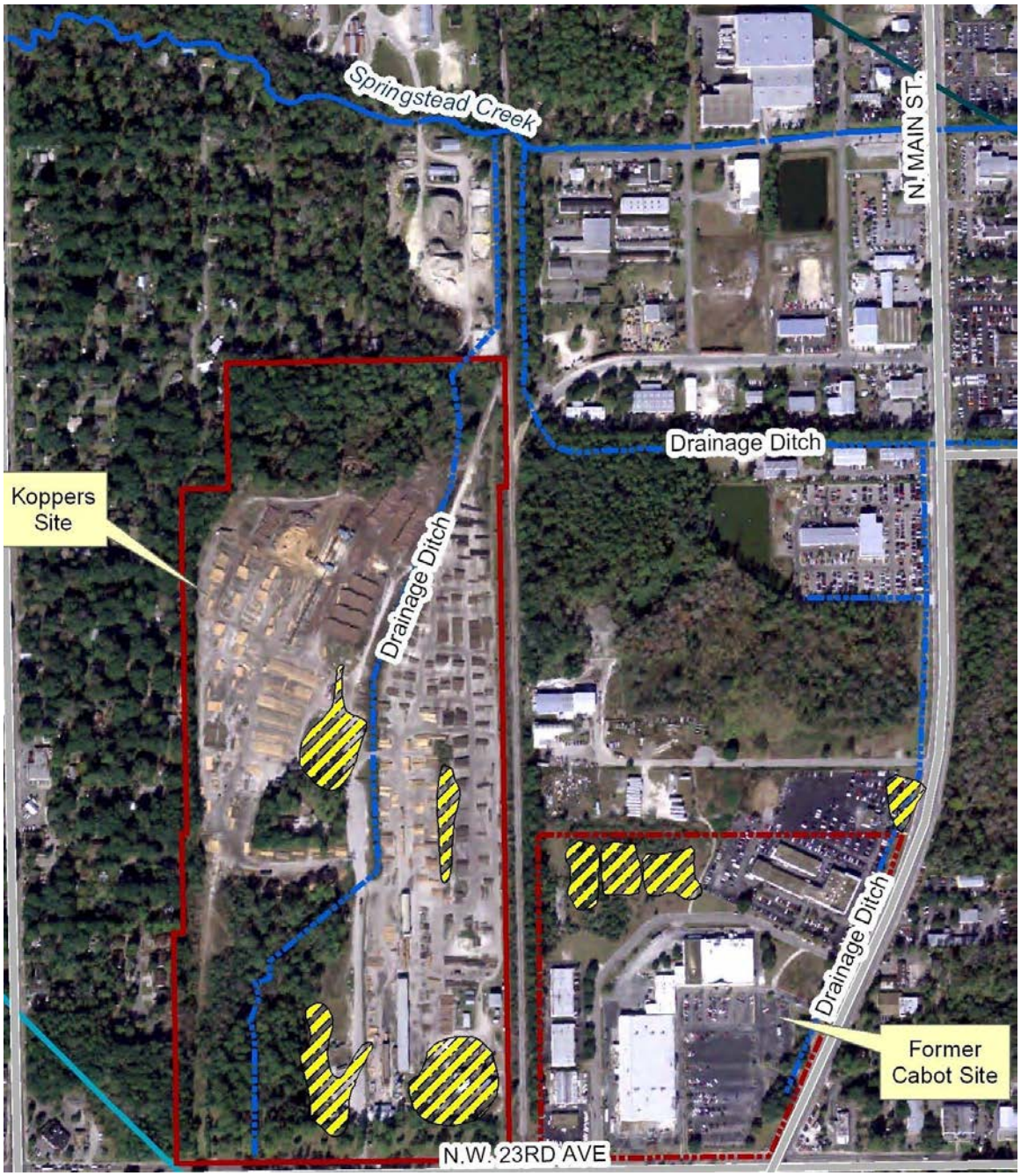


Figure 5-9. Overview of the Cabot-Koppers Superfund Site on Google Earth (Source: http://aquaticpath.phhp.ufl.edu/waterbiology/handouts/CabotKoppers.pdf)

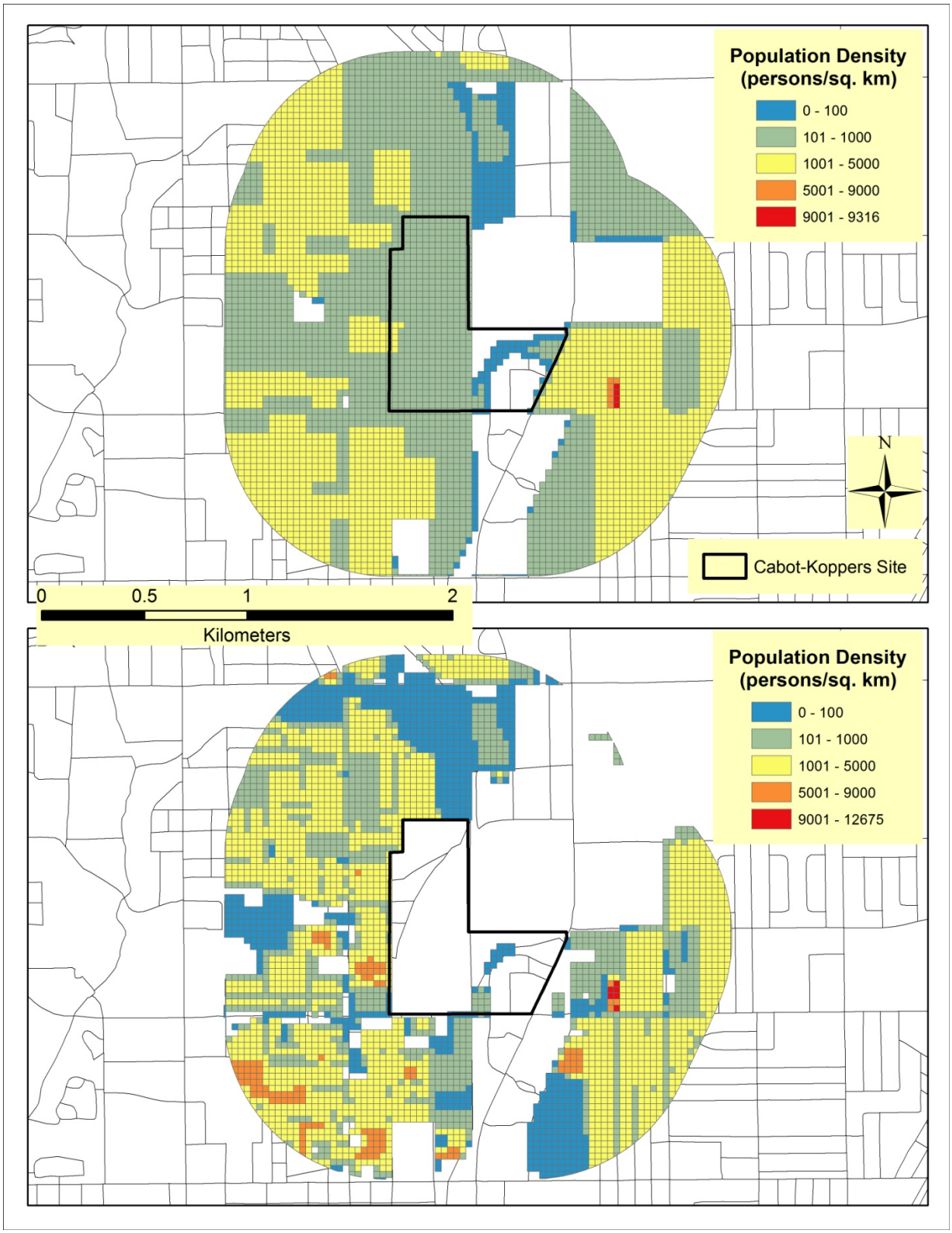


Figure 5-10. Visual comparison of the choropleth (up) and dasymetric map (down) within the buffer with a 0.5 miles radius centered on the Cabot-Koppers Site

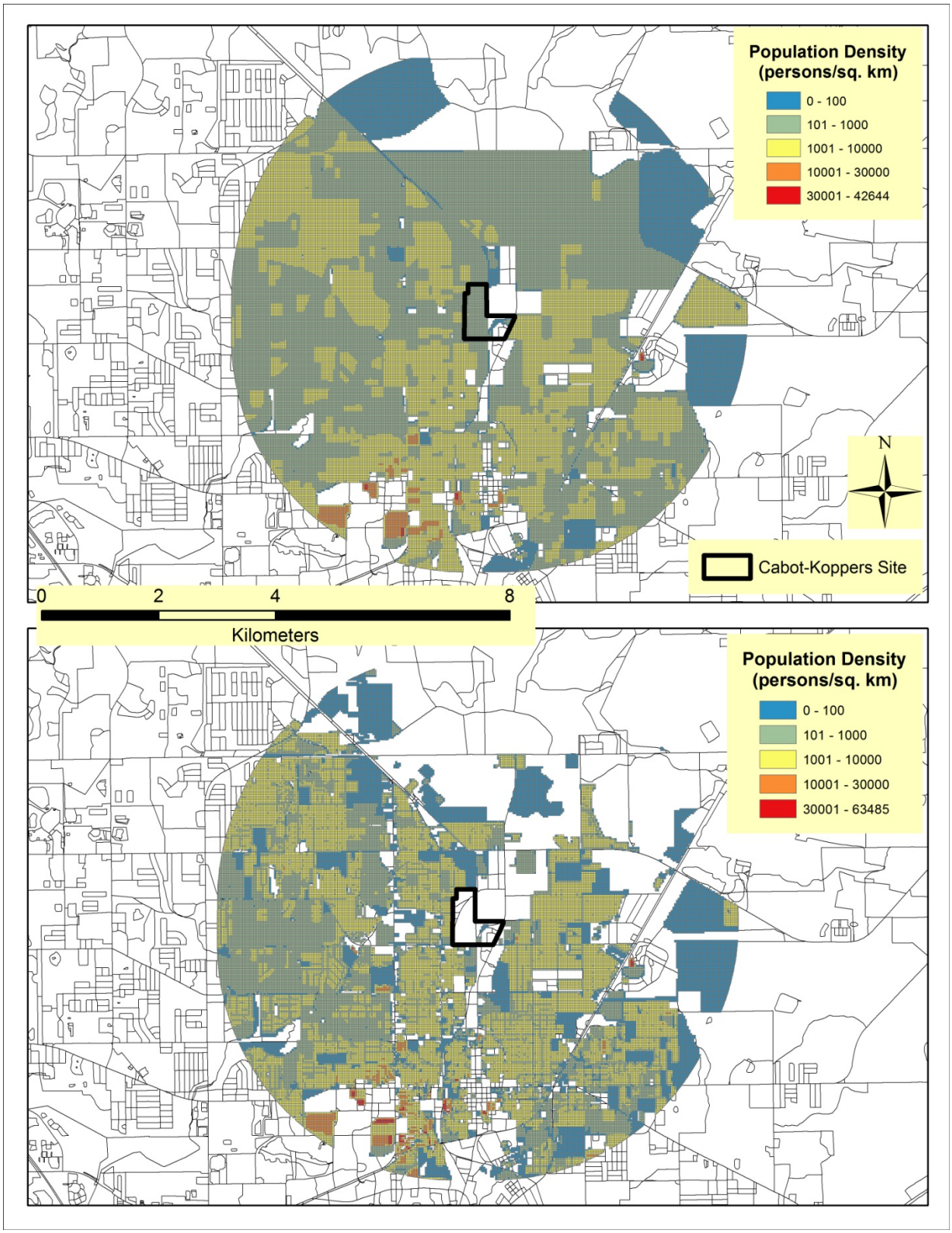


Figure 5-11. Visual comparison of the choropleth (up) and dasymetric map (down) within the buffer with a 2.5 miles radius centered on the Cabot-Koppers Site

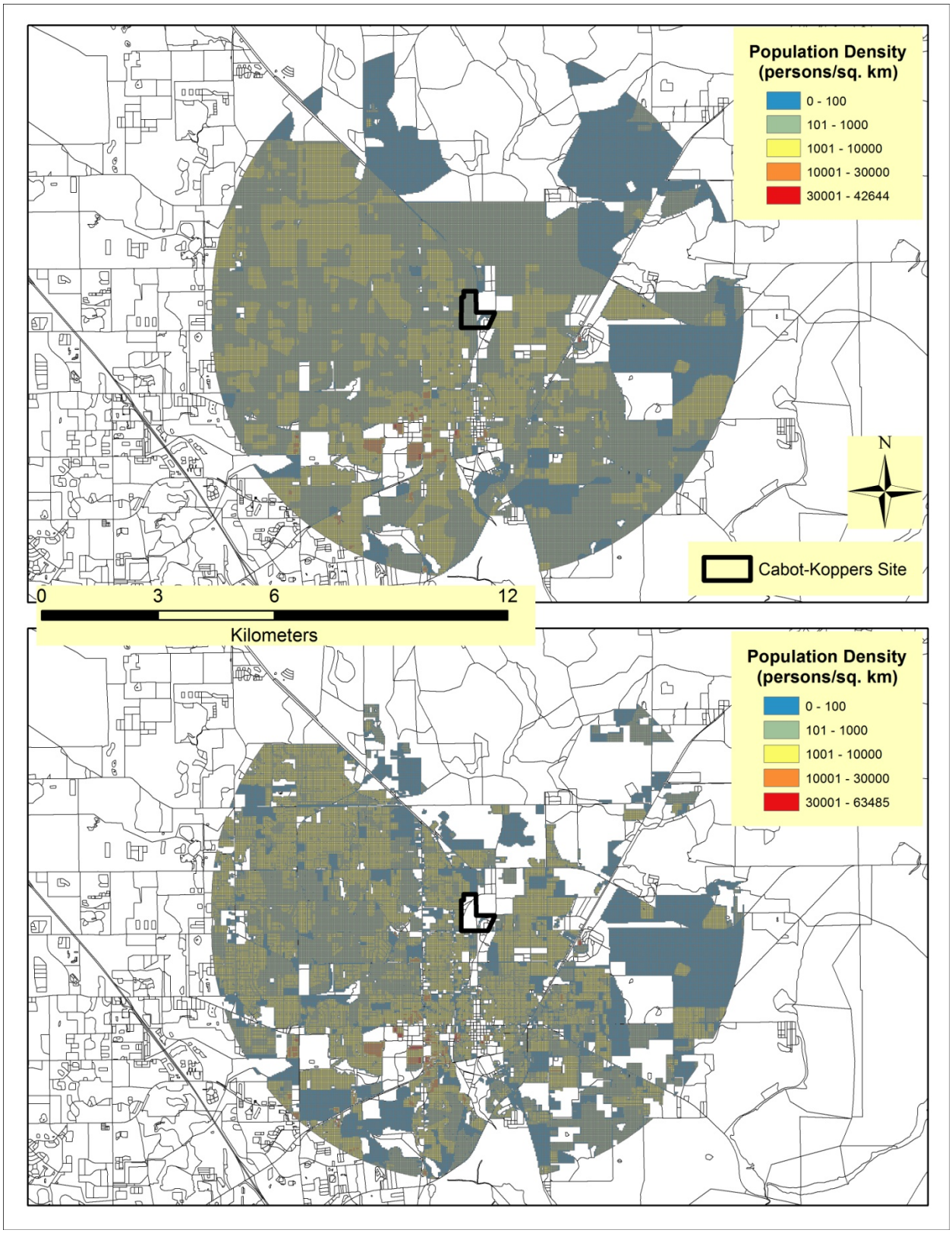


Figure 5-12. Visual comparison of the choropleth (up) and dasymetric map (down) within the buffer with a 4 miles radius centered on the Cabot-Koppers Site

CHAPTER 6
CONCLUSIONS

The purpose of this study is to use the dasymetric mapping method to redistribute aggregated population counts on Census block level to 30m×30m quadrilateral grids in Alachua County, Florida, and apply the HGPS to two case studies in the field of transportation and public health estimate, one of which is to estimate the numbers of potential passengers within the service areas of bus routes, and the other is to estimate the numbers of population potentially exposed to the contamination at a Superfund Site. By comparison of the results from the traditional and dasymetric method, as well as the actual numbers from relevant records, the advantages of the HGPS over the traditional population grids are demonstrated.

The more detailed spatial patterns and intra-block spatial variations of the population distribution coinciding more with real situations but hidden by the traditional aggregated numbers, in preference to the improvement in estimated numbers, are the most significant advantages brought by the HGPS. Furthermore, it allows us to flexibly extract the population counts in any certain group on a basis of various conditions.

The HGPS will definitely advance the GIS application in the field of transportation and public health; additionally, the novel integration of Census and Tax Parcel data also holds promise for improved research in many other areas, such as emergency management and environmental assessment.

This study is an important step forward towards solving the MAUP, where the resolution of 30m provided us with more freedom not only to delineate any types of zone, such as school attendance zone and affected area of pollution sources, or modify present areal boundaries or the level of data aggregation without substantially changing

the distribution of population values, but also to combine with some major spatial data with the same resolution, such as Landsat TM data.

Gridded population surfaces outperform traditionally choropleth representation in every aspect, which has been more and more realized by demographic researchers. It is a necessary step to build a digital city. However, the accuracy of gridded population surfaces needs to be more assessed wherever gridded data are available in the future research, according to which we could get to know which parameters and factors would significantly decide the accuracy of HGPS.

Anyways, we can foresee that a new era of HGPS is on the way.

CHAPTER 7
DISCUSSIONS

To the author's current literature review, this study is the first one to utilize the detailed property types in the Tax Parcel data as ancillary data to disaggregate Census data onto high-resolution grids. Xie (2006) used Tax Parcel data in the City of Boca Raton, Florida, to aid differentiating residential buildings extracted from remote sensing data from other buildings and delineating the housing units; however, the accuracy of the classification was not assessed and there were also two weak assumptions in his study: 1) the number of persons in each housing unit was the same within each block group, calculated by dividing the total population of the block group by the total number of housing units; 2) all the persons lived on residential parcels.

I jumped over the block group level and directly disaggregated population counts from blocks to grids for two reasons. One is that I found some small and irregular blocks that were actually just medians, but had a certain amount of population on them. I assumed that those parts of population did not exist in reality, so to disaggregate population from block groups would count some non-existing population and lead to overestimation. The other reason is that I would like to control the error produced during the disaggregation process within a finer unit; in other words, the error of the redistribution in one block, if existing, would not influence the results in other blocks within the same block group. The weakness of doing this is that no finer data can be used to compare with the disaggregated results in the U.S. Comparing the gridded results with the existing gridded product only appeared once in earlier studies (Martin et al. 2011), where the population counts in Northern Ireland were disaggregated from the

Census output areas onto 100m grids and compared with the true 100m Northern Ireland Grid Square Product.

In the first case study, as the adjacent parcels with the same type were mapped together without any boundary separating them and the information of building footprints were unavailable in Tax Parcel data, there is no way to know the number of households and the position of building entrances on each parcel. Therefore, I did not use the parcel-network method presented by Biba (2010) to calculate the distance of shortest path from the centroid of each parcel to its nearest bus stop. Unless that distance is accurate, otherwise it would be meaningless even though it can be done. I created a 400m buffer zone for each bus stop of each route, merged them all and found that it was the same as the buffer zone of the entire route because of densely and evenly placed stops, so I directly used the 400m buffer zone of each route to represent the service area.

A large difference of estimated population within the 400m buffer zone of each route from two methods was not observed before clipping because, for the purpose of serving more residents, the bus routes have been built in the urban area where the heterogeneity within blocks is less apparent than in the country; moreover, the heterogeneity within blocks is not as obvious as within Census tracts or block groups because blocks are already the finest administrative units. However, when I separated the multi-family and condominium parcels, where the potential passengers in this study lived, from other property types and solely examined the population on those two types, the differences between two groups of numbers appeared. Compared to the dasymetric method, the estimated numbers from the traditional method are much closer to the

actual numbers of passengers for all the routes except Route 20&21. However, for two routes out of nine, the underestimation occurs when using the traditional method, which means that the actual number of passengers overpasses the total number of residents; especially for the Route 20&21, the actual passengers are more than the total residents in each month and compared to the average passengers in five months, the underestimation rate reaches 35%. It is rarely true that all the residents who may have a demand for bus ride truly take the bus but still 35% less than the actual passengers.

It would be too early to say either method is better until I found out the possible reasons behind the contradictory results. From the overlaid layer of bus routes and property types in the urban area of Gainesville (Figure 7-1), we can see that Route 20, 21, 9, 35 and 36 nearly run completely through the campus, multi-family and condominium zone, where, through the field survey, most of the neighborhoods are rental for students at the University of Florida, so they five are appropriate routes used to estimate the potential passengers within their buffer zones because the people living on campus are usually served by campus routes and the ones living in rental neighborhoods may most probably have a demand for bus ride. Among those five routes, Route 20 and 21 run along nearly the same track and serve the same region, like one route, so to merge them is more suitable than merge Route 9, 35 and 36, the service regions of which are a little different at the south of Gainesville. Still merging them is because a large part of the tracks those three routes run overlay, that is, the buffer zone of each route also overlays a lot with each other. If counting the number of population in the buffer zone for each route separately, all the people living in the overlaying zone would be counted twice or three times but in RTS records, they are only

counted once. To merge them is for the better match between the number of population and passengers.

Not all the potential passengers in multi-family and condominium truly take the bus every day, some of which also have their own cars, so, compared to the actual passengers, moderate overestimation is more reasonable than a little underestimation. Therefore, the dasymetric method got more realistic results than the traditional method, which can also be explained by the theory of two methods. The former estimates the population by getting a subset of the gridded population surface, on which population has been reasonably redistributed, and calculating the sum of values of all the grids on the subset; the latter just by calculating the area ratio of the sub-block located within the buffer zone over the entire original block and multiplying that area ratio with the population counts on the original block to get the estimated population. Assuming that a sub-block within the buffer zone occupies a certain proportion of the entire block, regardless of how crowded or scarce that sub-block actually is, a constant proportion of the population on the original block the sub-block belongs to would be assigned.

As to why the estimations from the traditional method are closer to the actual passengers for the other eight routes, let us go back to the Figure 7-1 and we can see that, in contrast to five aforementioned routes, Route 5, 15, 2, 8, 29, 24, 23 and 10 are all running amid single family zones where most of people are assumed to rely on their own cars for going out. The single family houses hold the largest part of the population (38.04%) in Gainesville (Table 3-1). There are also some communities of multi-family and condominium there, but few students live there because of being far from the campus. The order of the routes listed in Table 5-4, from the smallest overestimation

rate of the dasymetric result after masking on the left to the largest on the right, also reflects a general trend that the tracks of the routes listed on the left intersect more with the university zone (the white ellipse) than the tracks of the routes on the right.

In sum, for Route 5, 15, 2, 8, 29, 24, 23 and 10, the population estimated by neither USCG nor HGPS is the actual number of people who may have the demand for the bus service. In addition to work or study locations and proximity to the nearest bus stop, the ridership patterns is often influenced by some other demographic characteristics, such as car ownership, age distribution and income level (Biba et al. 2010). The experiences from this case study can also be used to evaluate the cost and efficiency of bus services and optimize them in the future.

In the second case study, knowing the detailed spatial distribution of the population at risk rather than only an estimated number would be greatly helpful to calculate the individual risks, where the extent of individual exposure to contamination usually depends on the distance to the pollution sources, sometimes as well as the direction. For example, the affected media would be the air near a toxic waste dump rather than soil and groundwater at Cabot-Koppers Site. In that case, we need to identify how the population is distributed within the affected area and how far the individuals reside from the dump, as well as which direction the dominant wind goes through that area, in order to assess the individual risks.

In Hamilton's study (1999), the contaminated groundwater plume maps were overlaid with the Census block map to estimate the expected numbers of cancers. A similar procedure might be involved in the process of assessing the extent of exposure

around the Cabot-Koppers Site, so the HGPS would be more appropriate to overcome the inconsistency between two sets of maps.

Maantay (2007) also provided evidence for the advantages of CEDS in the Bronx, New York City, over the traditional filtered areal weighting approach. In his study, the asthma hospitalization rate (AHR) inside the buffer of limited access highways (LAH) calculated by the filtered areal weighting approach is lower than not only the mean AHR for the entire Bronx, but also the AHR outside the buffer of LAH, which rejects the assumption that people living near LAH are more likely to be hospitalized for asthma than those living far from LAH, that is, the correlation between AHR and proximity to major air pollution sources. However, when using CEDS to replace the filtered areal weighting approach, the results support the hypothesis again where the AHR inside the buffer of LAH is higher than both the AHR outside the buffer of LAH and the average level for the entire Bronx.

Some problems remain in my study. The quality issues in the Census data and Tax Parcel data is still unavoidable. Ural (2011) brought up the possible inconsistencies between the number of population and houses in some blocks; for example, 139 people are assigned to five single-family houses in the block 4006 of Census tract 52 in West Lafayette, Indiana, which is most probably false. That is just one of many types of possible inconsistencies among Census data. Other types of inconsistencies found through the data examination in my study include: 1) some blocks with small-size or irregular shapes are actually just medians or parking areas along the street, but considered as blocks and have a certain amount of population on them; 2) the temporal mismatch between Census and Tax Parcel survey may lead to some amount of

population on vacant residential houses, like the 3.23% of the population in Gainesville is distributed on vacant residential lots and another like in Miami, the population density on more than half of vacant residential houses is not zero, especially along the beach; 3) the houses among the agricultural land usually fail to be recognized on the Tax Parcel maps, so the population counts within an agricultural-dominated block can only be evenly distributed across the entire agricultural area, which is not a problem in earlier studies where the study area is limited to the city (Maantay et al. 2007; Ural et al. 2011); 4) some inhabitable parcels have no population on them because of being under construction. For example, the density of the mixed use type is 0.16 persons/100m^2. I randomly selected a mixed use parcel without population and examined it on Google Earth from a vertical and side view (Figure 7-2). The parcel that might have been approved to be used as a combination of stores and residential houses was being under construction at the time of being surveyed.

One of the largest unsolved problems in dasymetric mapping is the variation of intra-class density; for example, the area of a single family house in the city is often smaller than its counterpart in the suburb or in the country due to expensive land prices. National Land Cover Database (NLCD), a national-scope land-cover mapping program, may provide a fit data source to relieve this issue to some extent, which has been utilized in some studies (Langford et al. 1991; Reibel and Agrawal 2007; Zandbergen and Ignizio 2010; Zandbergen 2011). In NLCD dataset, the developed areas are divided into four classes based on the percentage of impervious surfaces: open space (<20%), low intensity (20% to 49%), medium intensity (50% to 79%) and high intensity (80% to 100%). To utilize NLCD data works under the assumption that those four classes

aforementioned are a good proxy for the degree of urbanization, then the density of the single family houses in low intensity and high intensity areas could be separated from each other. The variability of population density on the same type of parcels in different counties, though not severe, still exists due to diverse socioeconomic background. However, to address this issue might be quite time-consuming.

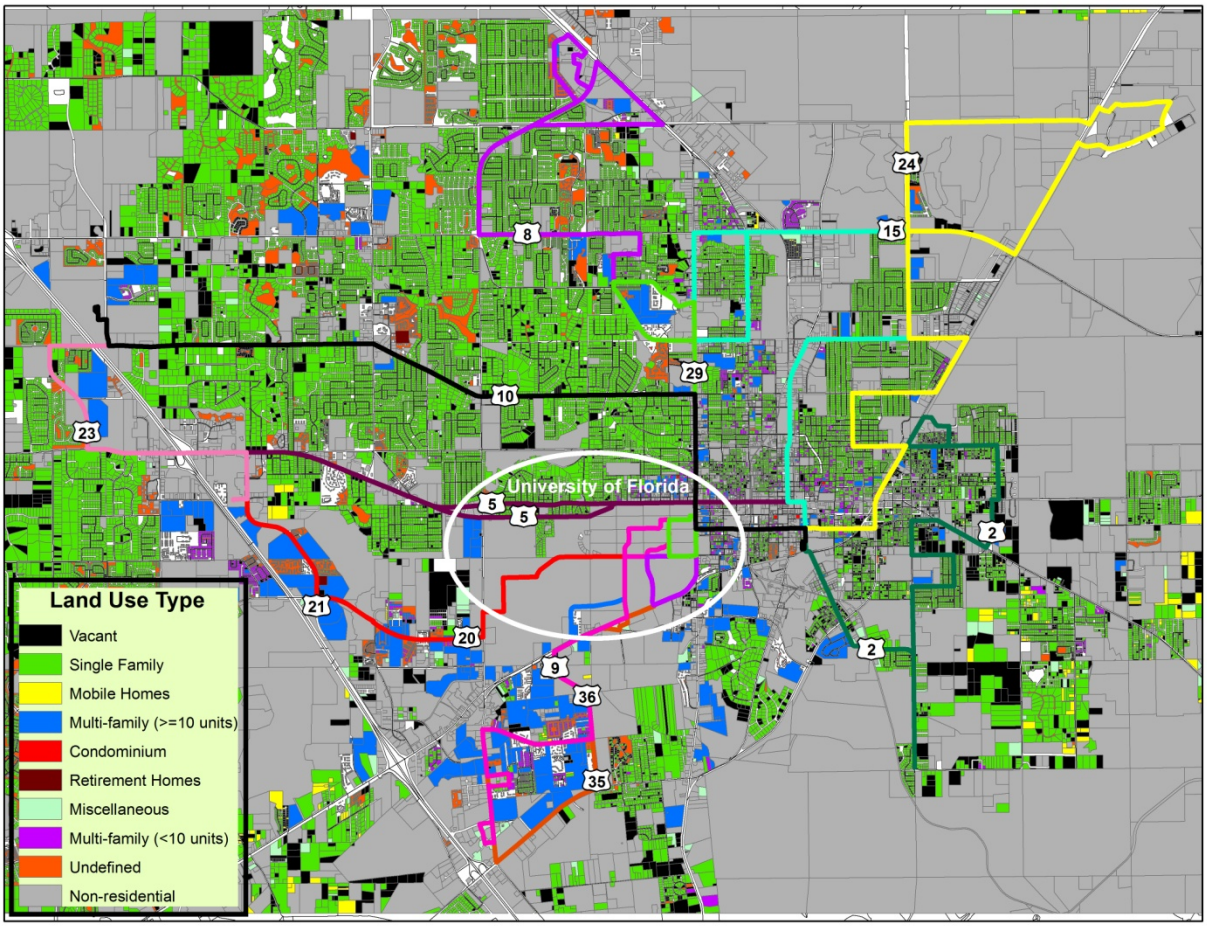


Figure 7-1. Overlay of bus routes and property types

A B

Figure 7-2. The vertical (A) and side (B) view of a mixed use parcel

LIST OF REFERENCES

Aubrecht, C., K. Steinnocher, M. Hollaus, and W. Wagner. Integrating earth observation and GIScience for high resolution spatial and functional modeling of urban land use. *Computers, Environment and Urban Systems* 33(1): 15-25.

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

Biba, S., K.M. Curtin, and G. Manca. 2010. A new method for determining the population with walking access to transit. *International Journal of Geographical Information Science* 24(3): 347-364.

Bielecka, E. 2005. A dasymetric population density map of Poland. In: *Proceedings of the 22nd International Cartographic Conference*, July 9-15. A Coruna, Spain.

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

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

Hamilton, J.T., and W.K. Viscusi. 1999. How costly is "clean"? An analysis of the benefits and costs of Superfund site remediations. *Journal of Policy Analysis and Management* 18(1): 2-27.

Krunić N., B. Bajat, M. Kilibarda, and D. Tošić. 2011. Modelling the spatial distribution of Vojvodina's population by using dasymetric method. *Spatium* 24: 45-50.

Langford, M., D.J. Maguire, and D. Unwin. 1991. The areal interpolation problem: Estimating population using remote sensing in a GIS framework. In: Masser, I., and M. Blakemore (eds), *Handling geographic information: Methodology and potential applications*. London, U.K.: Longman.

Langford, M. 2007. Rapid facilitation of dasymetric-based population interpolation by means of raster pixel maps. *Computers, Environment and Urban Systems* 31: 19-32.

Li, T., and J. Corcoran. 2010. Testing Dasymetric Techniques to Spatially Disaggregate Regional Population Forecasts for South East Queensland. *Urban Research Program Research Paper* 28, Griffith University, Brisbane.

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

Liu, X., and K.C. Clarke. 2002. Estimation of residential population using high resolution satellite imagery. In: Maktav, D., C. Juergens, and F. Sunar-Erbek (eds), *Proceedings of the 3rd Symposium in Remote Sensing of Urban Areas*, June 11-13. Istanbul, Turkey: Istanbul Technical University Press. pp. 153-60.

Maantay, J.A., A. Maroko, and C. Herrmann. 2007. Mapping population distribution in the urban environment: the Cadastral-based Expert Dasymetric System (CEDS). *Cartography and Geographic Information Science* 34(2): 77-102.

Martin, D. 2011. Directions in Population GIS. *Geography Compass* 5(9): 655-665.

Martin, D., Lloyd, C., Shuttleworth, I. 2011. Evaluation of gridded population models using 2001 Northern Ireland Census data. *Environment and Planning A* 43: 1965-1980

Mennis, J. 2003. Generating surface models of population using dasymetric mapping. *The Professional Geographer* 55: 31-42.

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

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

Openshaw, S. 1984. Ecological fallacies and the analysis of areal census data. *Environmental and Planning A* 16: 17-31.

Qiu, F., H. Sridharan, and Y. Chun. 2010. Spatial autoregressive model for population estimation at the census block level using LIDAR-derived building volume information. *Cartography and Geographic Information Science* 37: 239-57.

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

Saporito, S., J.M. Chavers, L.C. Nixon, and M.R. McQuiddy. 2007. From here to there: Methods of allocating data between census geography and socially meaningful areas. *Social Science Research* 36: 897-920.

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

Sharkova, I.V. 2000. With or Without GIS? Evaluating Accuracy, Timeliness and Costs of Population Estimates for User-Defined Areas. In: *Proceedings of the Population Association of America Annual Meeting*, March 23-25. Los Angeles, California.

Shi, S., and N. Walford. 2010. An automated internet geoinformation service for integrating online geoinformation services and generating quasi-realistic spatial population GIS maps. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 38(II): 427-432.

Silván-Cárdenas, J.L., L. Wang, P. Rogerson, C. Wu, T. Feng, and B.D.Kamphaus. 2010, Assessing fine-spatial-resolution remote sensing for small-area population estimation. *International Journal of Remote Sensing* 31: 5605-5634.

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

Wu, C., and A.T. Murray. 2007. Population estimation using Landsat enhanced Thematic Mapper imagery. *Geographical Analysis* 39: 26-43.

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

Wu, S. 2006. Incorporating GIS and remote sensing for Census population disaggregation. *Doctoral dissertation*. Retrieved from ProQuest Dissertations and Theses. (Accession Order No. AAT 3221522)

Wu, S., L. Wang, and X. Qiu. 2008. Incorporating GIS building data and census housing statistics for subblock-level population estimation. *The Professional Geographer* 60(1): 121-35.

Xie, Z.X. 2006. A framework for interpolating the population surface at the residential-housing-unitlevel. *GIScience and Remote Sensing* 43: 233-251.

Yang, X., G. Jiang, X. Luo, and Z. Zheng. 2012. Preliminary mapping of high-resolution rural population distribution based on imagery from Google Earth: A case study in the Lake Tai basin, eastern China. Applied Geography 32(2): 221-227.

Yuan, Y., R. M. Smith, and W. F. Limp. 1997. Remodeling Census Population with Spatial Information from Landsat TM imagery. *Computers, Environment and Urban Systems* 21(3-4): 245-258.

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

Zandbergen, P. A. 2011. Dasymetric Mapping Using High Resolution Address Point Datasets. *Transactions in GIS* 15(Supplement s1): 5-27.

BIOGRAPHICAL SKETCH

Peng Jia was born in Hohhot, Inner Mongolia, China. He earned his B.Eng. in environmental engineering from the Nanjing Agricultural University (NJAU) in 2007 and his M.S. in cartography and geographic information system (GIS) in 2010 from the Institute of Remote Sensing Applications Chinese Academy of Sciences (IRSA, CAS), jointed with the National Museum of China.

He entered the Department of Geography at University of Florida in August 2010 and started working in Emerging Pathogens Institute as a graduate student research assistant since May 2011. His researches focus on the applications of GIS, remote sensing (RS) and other spatial techniques in the study of population, public health, environmental health, health impact assessment and archaeology.

Upon graduating in May 2012 with his M.S. in geography, he will enter the Department of Geography & Anthropology at Louisiana State University to pursue his Ph.D. in geography.