IMPROVING AN OPEN-SOURCE POPULATION MAPPING METHOD UTILIZING SPACEBORNE, AIRBORNE, AND TERRESTRIAL INSTRUMENTS

by

Nirav Nikunj Patel

A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Earth Systems and Geoinformation Sciences

Committee:

______________________________	Dr. Donglian Sun, Dissertation Director
______________________________	Dr. John J. Qu, Committee Member
______________________________	Dr. Michael Summers, Committee Member
______________________________	Dr. Ruixin Yang, Committee Member
______________________________	Dr. Anthony Stefanidis, Department Chair
______________________________	Dr. Donna M. Fox, Associate Dean, College Office of Student Affairs & Special Programs, College of Science
______________________________	Dr. Peggy Agouris, Dean, College of Science

Date: ______________________________

Summer Semester 2017
George Mason University
Fairfax, VA

Improving an Open-Source Population Mapping Method Utilizing Spaceborne, Airborne, and Terrestrial Instruments

A Dissertation submitted in partial fulfillment of the requirements for the degree of Earth Systems and Geoinformation Science, as Doctor of Philosophy at George Mason University

by

Nirav Nikunj Patel
Master of Science
University of Florida, 2013
Bachelor of Arts
University of Florida, 2011

Director: Donglian Sun, Professor
Department of Geography and Geoinformation Science

Summer Semester 2017
George Mason University
Fairfax, VA

DEDICATION

This dissertation is dedicated to my family: my mother Gaurangi N. Patel, my father Nikunj A. Patel, and my sister Ruchita N. Patel. It is also dedicated to my grandparents and my family's sacrifices in India and around the globe with respect in helping me to even have this opportunity. It is also dedicated to the goals of the WorldPop Project in aims of mapping human population for the benefit of humanity and the Earth. Most of all this is dedicated to all the teachers that have believed in me in all stages of my career. Service to others is truly the highest distinction, and I continue to want to embody that in my research and my life.

ACKNOWLEDGEMENTS

"I hate Indians. They are a beastly people with a beastly religion. The famine was their own fault for breeding like rabbits." – Winston Churchill, 1943

This quotation was from Winston Churchill speaking to his Secretary of State for India, Leopold Amery, in 1943, when 3 million dark-skinned subjects of the British Raj died in the Bengal famine, which was one of history's worst. Oral accounts of survivors show as a part of the Western war effort, food was diverted to starving Indians to well-supplied soldiers of Britain, with stockpiles collected in Britain and elsewhere in Europe. It is telling that Churchill's only response to a telegram from the government in Delhi about people perishing in the famine was to ask why Gandhi had not died yet.[1]

The completion of this dissertation is the result of the sacrifices of many of my family members as well as the continuous and unrelenting support of friends and teachers over the years. It represents a dream realized for my ancestors and a commitment of what needs to be accomplished for the future, for the benefit of humanity and for the benefit of the Earth.

Around February 2014, my mother told me a story about my grandfather after I was admitted to George Mason University. In a small village called Jaspur, my grandfather studied in high school class with a certain Vikram Sarabhai, the eventual founder of the Indian Space Research Organization. Unlike Sarabhai's family, our family had very little wealth at the time, so even though my grandfather had the grades to pursue studying Physics and the Natural Sciences in England, he couldn't afford traveling and living overseas like Sarabhai, and so he instead became a very successful lawyer in India. This PhD is dedicated to him and my grandparents, and I hope that with this path and I have changed that what-if of the past into something actualized, if not for myself, for the spirit of my grandfather and the spirit of my ancestors.

My grandparents had found themselves in a position to pursue any career path and any profession they wanted, despite the challenges of growing up in the British Raj, in a post-Independence India. It is their creativity and perseverance that I acknowledge as the root of this achievement, and it represents the fruits of the sacrifices they have made.

It was a very emotional moment for me to be able start my PhD courses here at George Mason University in August 2014, and be able to pay for everything on my own, knowing that my grandfather didn't have the same opportunity roughly 80 years ago in a rural village just outside of Ahmedabad.

[1] http://content.time.com/time/magazine/article/0,9171,2031992,00.html

It is my grandparents' sacrifice that allowed my parents to come to the United States to take a risk and create a better life for themselves, with my mother becoming a Dentist and Orthodontist and my father owning and operating multiple businesses. They toiled away for years in the Northeast in New Jersey with my older sister before they made the decision to move down to Tampa, Florida in the late 1980s.

When I was born in Tampa in 1990, I was treated to the best childhood I could ask for. I had the privilege of traveling the world, and seeing India when I was a little boy, as well has having some of the best public schooling in retrospect. Having Anant and Shivam Kharod as my childhood (now lifelong) friends helped me truly come out of my shell and foster an unconditional brotherhood that would be the foundation for my toughest times. I cannot forget the influence of their parents, Asha and Manish Kharod in my early years of schooling as well. When I came to my years of middle school, I was fortunate to be at a charter school, called Independent Day School where I met a lifelong teacher, Mr. Tom Bronson, who taught me not to be intimidated of Mathematics and uncovered its beauty to me in understanding the natural world.

At Hillsborough High School, I was fortunate to be influenced by very great teachers, Mr. Zaan Gast, who influenced my interest in research through his passion for European History. Mrs. Katherine Griffin, who pulled me into the world of debate and taught me how to think critically. Mr. Robert "Fred" Digenova, my homeroom teacher and friend, who taught Theory of Knowledge in the IB Curriculum and gave me confidence in my own ability to think and fostered my interest in ontological philosophy. On the debate team, I made two lifelong friends in Hiten Patel and Shashin Chokshi, who are constant guardians of my conscience and keep me honest to staying to my passions. I thank them and love them for their friendship.

When I started at the University of Florida in August 2008, I finally discovered the discipline of Geography. I started my undergraduate career thinking I wanted to be a lawyer, and spent my first two years of school thinking that I would try to work in international development law. I was so thankful that I had the influence of Ted Gonder, a mutual friend of Shashin Chokshi, who showed me that Geography was much more than memorizing states and capitals. Pledging into the brotherhood of Sigma Beta Rho, a national multicultural fraternity, broadened my horizons and gave me access to a nationwide network of brothers who would be at my side through the toughest times. I'm thankful to Trupal Patel introducing me to the brotherhood. My pledgemaster, Joshua Tyler remains a constant supportive force in my life, as are my linebrothers: Jerin George, Michael Vempala, Justin Zacharias, Nicholas Joseph, and Nirav "Neal" Patel. My big brother, Murali Iyyani, also is a constant supporter of everything that I do, and the best big brother I never had. He inspired me to join him in mentorship and join the Big Brothers and Big Sisters program, where I met Miguel Bean, from who I have learned so much. I'm thankful for meeting Eric Morrow, my athletic coach for the past 8 years, who has kept me in shape through these tough times and has been a source of constant positive energy, and helped me be the best Gator Triathlete I could be! I'm thankful to Dr. Joann Mossa, at the Department of Geography, to convince me to switch majors to Geography in the summer of 2010, which led me to graduate with my Bachelors of Arts in Geography and Philosophy in August 2011.

I was so enamored by Geography as a discipline that I entered into a Master's Program at the University of Florida in August 2011, immediately after graduating with my undergraduate degrees. In this time period, I grew a love for medical geography, and was introduced to remote sensing. In my field work in India, I am so thankful that I had the opportunity to work with the organization my grandparents founded in 1980 to assist the most vulnerable in the city where my parents grew up in. In the work I did for my Masters, I'm very thankful for the collaboration of the main doctor I worked with in India, Dr. Dixit Kapadia, and the countless social workers I worked with at the Akhand Jyot Foundation. The guidance of my committee members at the time, Dr. Liang Mao (committee chair), Dr. Timothy Fik and Dr. Peter Waylen was indispensable. They are fantastic mentors to work with, and they continue to guide me in my current work. I am thankful to the family of Dr. Ryan Poehling for awarding me a fellowship for my service to the department of Geography in 2013.

After completing my Masters in May 2013, I had the good fortunate of being recommended by Dr. Jessica Steele as well as other mentors to work at NASA Langley Research Center as an intern for the summer of that same year, which was the seed that inspired me to pursue the PhD after seeing what the educational requirements were of working at NASA. I was so thankful to meet Dr. Bruce Doddridge and Dr. Kenton Ross, who have been instrumental in my NASA research and career thus far.

At NASA, I continued on the work I had started the previous year, with the guidance of Dr. Andy Tatem, Dr. Forrest Stevens and Dr. Andrea Gaughan, who allowed me to work in an assistantship with the World Bank, the last year of my Masters, and concurrent with my NASA internship. It is their influence that led me to ultimately pursue the PhD and become confident in my skills. It with this work on the then AsiaPop team, which is now the WorldPop team, in which I had the opportunity to work at the University of Pavia, in Pavia, Italy, under the guidance of Dr. Paolo Gamba from September 2013 to December 2013, working on the Google Earth Engine, which became the basis of my first paper. Countless colleagues and friends I made in Pavia, Italy as well as the research environment, convinced me that I needed to pursue the PhD.

In making a commitment to apply to both George Mason University and the University of Maryland, in my thought of working full-time and doing the PhD part-time to support myself, I am thankful for the acceptance of George Mason University and the rejection letter I received from University of Maryland for my desire to work full-time while pursuing an education. Dr. Timothy Leslie endured many questions I had and welcomed me to GMU's program, giving me real advice. Mrs. Debbie Hutton provided the best help possible in helping me progress through the paperwork to get to this level. Dr. Ron Mahabir, thank you for all your support and guidance, and being a great friend to me at Mason.

Working full-time at Dito, a Manassas-based company, gave me a new family and support system in Dan McNelis, Jason Taylor, Trish McNelis and Karl von Schilling, as well as a salary to pay for my tuition. My friends and co-workers from NASA Langley who worked at Dito as well, Christopher Ferraro and Kent Sparrow were huge parts of Dito's success, and they will also remain lifelong friends. I'm so thankful to my cousin and his wife, Anshuman and Sejal Patel for being such a foundation for me in Arlington while being at George Mason University. I cannot thank my committee enough for being so flexible and supportive of me, Dr. Sun, you have been incredibly patient with my ambitious goals. Dr. Summers, thank you for giving me a primer on atmospheric dynamics and being so supportive knowing my motivations in joining the USAF as a guardsman. Dr. Yang, thank you for teaching me more scientific data mining skills that I am finding new uses for every single day. Dr. Qu, thank you for giving me a chance in being on my committee, despite your busy schedule! Nothing at George Mason would have been possible without the always excellent office staff as well, in Debbie Hutton and Samantha Cooke, without their assistance, I would have had a much tougher time in getting paperwork done.

Thank you to NASA Ames Research Center for assisting in this research as well as hopefully being a future place of long-term employment for me, and thank you to the United States Air Force's 129th Rescue Wing for being so accommodating to my schedule and encouraging me in my education. To my family in Fremont, Rajesh-uncle, Manisha-auntie, Rohun and Rahul, thank you so much for being there for me in the San Francisco Bay Area. Ruger Dutton of the USAF, I appreciate your help and support at the final stages of this dissertation.

Finally, I thank my immediate family, in my mother, father and sister in their continual support of me in pursuing this degree, and faith that although Earth Science is a not traditionally an "Indian" profession or calling, they have been so proud of what I have been able to accomplish. I'm thankful for all the positive people in my life that supported me wholeheartedly in what I want to accomplish.

This PhD represents another mile marker in my progression to answer these questions:
What's out there? How do we know what is out there? When we attempt to answer these two questions, how can we use this newfound knowledge to reduce suffering?

TABLE OF CONTENTS

Page.

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

Above Sea Level	A.S.L.
Advanced Very High Resolution Radiometer	AVHRR
Applied Remote Sensing Training (NASA)	ARSET
Artificial Neural Networks	ANN
Average Spatial Resolution	ASR
Bluetooth Low Energy	BLE
Center for International Earth Science Information Network	CIESIN
Central Processing Unit	CPU
Classification and Regression Trees	CART
Digital Orthophoto Map	DOM
Distributed Active Archive Centers	DAAC
Earth Observing System (NASA)	EOS
Earth Observing System Data and Information System	EOSDIS
Enhanced Thematic Mapper (Landsat 7)	ETM
File Transfer Protocol	FTP
Food and Agriculture Organization	FAO
Fraction of absorbed Photosynthetically Active Radiation/Leaf Area Index	FPAR/LAI
Geographic Information Systems	GIS
Gigabyte	GB
Gigahertz	GHZ
Global Human Settlement Layer	GHSL
Global Positioning System	GPS
Global Urban Footprint	GUF
Google Earth Engine	GEE
Ground Control Point	GCP
International Union for the Conservation of Nature	IUCN
Kilometer	KM
Land-Use/Land Cover Change	LULCC
Light Detection and Ranging	LiDAR
Linear Spectral Unmixing	LSU
Mean Absolute Error	MAE
Mean Squared Error	MSE
Meter	M
Moderate Resolution Imaging Spectroradiometer	MODIS
Modifiable Areal Unit Problem	MAUP
NASA Earth Exchange	NEX
National Aeronautics and Space Administration	NASA

National Geospatial-Intelligence Agency	NGA
National Land Cover Dataset	NLCD
National Land Use Dataset	NLUD
National Oceanic and Atmospheric Administration	NOAA
National Research Council	NRC
Net Primary Productivity	NPP
Normalized Difference Built-up Index	NDBI
Normalized Difference Vegetation Index	NDVI
Normalized Difference Spectral Vector	NDSV
Normalized Difference Water Index	NDWI
Oak Ridge National Laboratory	ORNL
Object-Based Image Analysis	OBIA
Open Street Map	OSM
Operational Land Imager (Landsat 8)	OLI
Out-Of-Bag	OOB
Persons Per Pixel	PPP
Random-Access Memory	RAM
Random Forest	RF
Red Green Blue	RGB
Regression Tree	RT
Root Mean Square Error	RMSE
Secure Shell protocol	SSH
Shuttle Radar Topography Mission	SRTM
Socioeconomic Data and Application Center	SEDAC
Sub-pixel Inundation Percentage	SIP
Suomi National Polar-orbiting Partnership	Suomi NPP
Support Vector Machines	SVM
Terabyte	TB
Thematic Mapper (Landsat 5)	TM
United Nations	UN
United Nations Environment Programme	UNEP
United States	U.S.
United States Agency for International Development	USAID
United States Geological Survey	USGS
Unmanned Aerial Vehicle	UAV
Vector Map (Level) 0	VMAP0
Very High Resolution	VHR
Visible Infrared Imaging Radiometer Suite (Suomi NPP)	VIIRS
World Food Programme	WFP
World Wildlife Fund US	WWF-US

ABSTRACT

IMPROVING AN OPEN-SOURCE POPULATION MAPPING METHOD UTILIZING SPACEBORNE, AIRBORNE, AND TERRESTRIAL INSTRUMENTS

Nirav Nikunj Patel, Ph.D

George Mason University, 2017

Dissertation Director: Dr. Donglian Sun

With human population growing by over 80 million a year, it has been projected that within the next 50 years, the 10 billion mark will be reached. Most of this growth is expected to be concentrated in primarily urban areas in low income countries. Rapid population growth has been well documented to impact economies, environment and health of nations, which are all expected to undergo significant change. To measure impacts of population growth with high accuracy, high resolution, and contemporary data on human population distributions as well as their compositions are necessary for planning interventions and monitoring changes.

Disease burden estimation, epidemic modeling, resource allocation, disaster management, accessibility modeling, transport and city planning, poverty mapping, and environmental impact assessment have integrated spatial databases of human

population. Low income regions of the world often lack relevant data or the data are of poor quality, whereas in high-income countries, extensive mapping resources and expertise are at their disposal to create such databases. The major obstacles to doing settlement and population mapping across the low income regions of the World include the scarcity of mapping resources, lack of reliable validation data and the difficulty in obtaining high resolution contemporary census statistics.

Focusing in on the open-source WorldPop Project and its associated methods, within the WorldPop Project a range of open geospatial datasets are combined in a flexible regression tree framework to reallocate contemporary aggregated spatial population count data. The resultant maps, backed by statistical assessments, suggest that the resultant maps are consistently more accurate than existing population map products, as well as the simple gridding of census data. The Project's 100m spatial resolution is a finer mapping detail than has even been produced at national extents, and as the data can be integrated with household survey, microdata, satellite and other data sources, this enables the production of more diverse datasets. Population count estimates can now encompass age structures, births, pregnancies, poverty and urban growth.

The aim of this dissertation is to provide a critical eye on how the open-source population mapping pioneered by the WorldPop project can be improve to directly indicate the presence of people within its datasets more accurately. The first section is an experiment with a tool called Google Earth Engine that can rapidly analyze vast amounts of satellite imagery to extract remotely sensed data, in this case applying a Normalized Difference Spectral Vector calculation over Landsat imagery to improve the temporal resolution of the population datasets. The second section is an experiment utilizing volunteered geographic data in the form of Twitter data that is geo-located, providing a layer of information that is human volunteered as a covariate in the population mapping process. The final section is a discussion of future work in mapping population at a continuous 30 meters, as opposed to 100 meters to examine the limitations of the co-variate datasets, as well as exploring the potential future implementation of Unmanned Aerial Vehicles to validate remote sensing classifications.

CHAPTER ONE

This chapter provides an overview of the goals of open-source population mapping in Section 1.1. In Section 1.2, literatures relevant to methodology of population mapping and the specific experiments within this dissertation will be reviewed in this section. Section 1.3 provides the statement of the overall problem. Section 1.4 elucidates the objectives of the study. Section 1.5 described the intended audience for the dissertation. Brief descriptions of the dissertation's main chapters are given in Section 1.6.

1.1 Overview of the goals of open-source population mapping

The global population is projected to increase from 7 billion to over 9 billion over the next four decades, with much of this growth concentrated in low income countries ("World Population Prospects: The 2010 Revision", 2011). The effects of such rapid demographic growth are well documented, with impacts on the economies, environment and health of nations (Bongaarts, 2009). To measure the impact of this population growth, as well as progress towards development goals, there is a need for contemporary, spatially-explicit, high resolution maps that accurately identify population distributions.

High-income countries often have mapping expertise and substantial resources at their disposal to create accurate and contemporary spatial population datasets. However, across the lower income regions of the world, equivalent resources and relevant data can often be either lacking or are of poor quality (Tatem & Linard, 2011). Over the past few decades there has been increasing interest in creating large-area gridded population distribution datasets (Cheriyadat et al., 2007; Balk et al., 2006; Linard et al., 2012) to support applications such as disease burden estimation, climate change and human health adaptive strategies, disaster response, accessibility modelling, transport and city planning, and environmental impact assessment (Balk et al., 2006; Linard et al., 2012; Linard et al., 2010; McMichael et al., 2006; Rasul & Thapa, 2003; Tatem, et al., 2007). Current global gridded population datasets include the Gridded Population of the World (GPW) database (Balk & Yetman, 2004; Tobler et al., 1997) and the Global Rural Urban Mapping Project (GRUMP) (Balk et al., 2005). In addition, there is the LandScan Global Population database (Bhaduri et al., 2007; Dobson et al., 2000), and the United Nations Environment Programme (UNEP) compiled gridded datasets for Latin America, Africa, and Asia (Nelson, 2004; Deichmann, 1996). The WorldPop project provides freely-available gridded population data for Africa, Asia and the Americas (Linard & Tatem, 2012; Linard et al., 2010; Stevens et al., 2015). With the exception of GPW, all of these datasets use spatial covariate datasets on factors related to the way that humans distribute themselves on the landscape to disaggregate areal-unit based census counts to grid squares.

Spatial covariate datasets used in the population disaggregation process tend to include factors known to correlate with population densities, such as satellite-derived maps of human settlements, urban areas, topography, lights at night, and land cover.

Additionally, infrastructure-related variables have been used, including road networks and health facilities (e.g. Stevens et al., 2015). However, all of these covariates are typically static in nature and not direct measures of the presence of people. Recent efforts have shown the potential of ‘big data’ sources, such as mobile phone call data records, to map populations dynamically using the communication patterns of phone users (Deville et al., 2014), but such data are generally difficult to obtain and are highly sensitive, both commercially and for privacy reasons. The rise in data availability of user communications and check-ins through social media presents opportunities however, in terms of a data-source that is freely available, dynamic and without the data sensitivity restrictions of mobile call data records. However it is important to consider that a limitation of utilizing social media is that Internet connected smartphones are often expensive resources in low-income countries and the applicability of these methodologies might be limited (Ramaswamy et al., 2009).

Open-source population mapping thus will be a methodology continuously developed to adapt to the needs of earth sciences and geoinformation sciences. This dissertation aims to suggest and evaluate methods to improve open-source population mapping using the current state of the art instruments available on spaceborne, airborne and terrestrial platforms. The three experiments in this dissertation deal with methodologies being explored to increase the spatial and temporal resolution of population outputs to improve their utility.

1.2 Literature Review

1.2.1 Dasymetric Mapping of Areal Data

A census refers to the methodology employed to acquire and record information about the members of a population over a defined territory with an emphasis on individual enumeration and a certain defined periodicity. Spatially, census mapping refers to the placement of this information in a spatial reference, and distribution of counts in areal units. Humanity had attempted to map its own population distributions in many early civilizations (Edwards, 1969).

Dasymetric maps appeared out of thematic mapping, which utilizes areal symbols to spatial classify certain types of volumetric data. This method was initially pioneered by Benjamin Petrovich Semenov-Tyan-Shansky in 1911 and popularized by others (Petrov, 2012). Historically and currently, dasymetric mapping is favored for mapping population density as it is able to realistically place data over geography. Dasymetric maps often utilize standardized data and places areal symbols by taking actual changing densities within the map area. In order to utilize this method, ancillary information must be acquired in order to inform the statistical data. The method created historically to serve the need of accurately visualizing population data. They are now more common with the ease of production in utilizing geographic information systems (Zakrzewska, 1967; Muehrcke, 1972; Kraus et al., 1974).

Dasymetric mapping represents certain advantages over areal unit representation in the sense that due to its surface-based representation, it allows for population data aggregation to nearly any desired areal unit. This allows it to avoid certain areal-unit derived problems (Bracken, 1993). Surface representations in dasymetric mapping also allow for graphic units of display (like grid cells), that can be uniform in size across a region, and surfaces of populations may offer more accurate cartographic representations of populations that normal chloropleth maps do (Langford & Unwin 1994). In the early days of GIS use on personal computing devices in the early 1990s, raster data was being used to create a surface-based format to convert demographic data into reliable and useful surface based representation of population and its associated attributes from aggregated census data (Langford et al. 1991; Martin & Bracken 1991).

1.2.2 Areal Interpolation and Dasymetric Mapping

When creating a population surface from areal unit data, this is considered to be areal interpolation or transforming a set of geographic data from one set of boundaries to another. This method is typically used when comparing spatial datasets that are stored in incompatible areal units such as congressional districts and census tracts. In this case, raster population surface generation is a special case of areal interpolation, as the desired (target) areal unit (a raster grid cell) is intended to approximate a continuous surface. This means that the grid cell must be much smaller than the size of the original areal unit of data aggregation (Mennis, 2003).

Areal weighting is a technique that can be used whereby each grid cell is assigned a population value based on its percentage area of the given host areal unit. In this method, the requirement of when the summation of the population data to the original set of areal units is preserved in the transformation to a new set of areal units, this meets the requirement of what is considered to be the pycnophylactic property (Tobler, 1979). Similarly intentioned techniques are put forward by (Flowerdew & Green, 1993) and

(Goodchild et al., 1993).

More sophisticated variants of areal interpolation include the inverse distance weighted (IDW) interpolation. In this interpolation, population counts area assigned to summary points what come from the centroids of the original areal units. A moving operation over an “empty” raster grid is then assigns to the window kernel a value that aligns with the population values of the centroids that are contained within the window, with some centroids having more “weight” than other centroids. This methodology assumes that population density must decrease from the centroid when considering distance-decay functions and allows for some areas of the raster surface to have no population (Bracken & Martin, 1989; Martin, 1989).

These techniques advanced as the quality of data became better and more available over the years, and their usage are almost always dependent upon the types of ancillary data that are used (Mennis & Hultgren, 2006).

1.2.3 Evolution of current population mapping methods trends

In the 1990s there was a growing interest in the global mapping of human populations (Deichmann, 1996; Jones, 1990), which lead to the advanced development of methodologies that could spatially downscale human population count data from censuses (sometimes summarized over large and irregular administrative units) to grid squares of 100 meters to 5 kilometer resolution in size (Balk & Yetman, 2004; Tobler et al., 1997; Dobson et al., 2000; Balk et al., 2006; Linard et al., 2012; Gaughan et al., 2013; Azar et al., 2013) . Initially these methods used simple areal weighting (Balk & Yetman, 2004; Deichmann et al., 2001). Others used dasymetric modeling approaches (Balk et al., 2006; Linard et al., 2012, Gaughan et al., 2013), which often utilized ancillary layers in order to redistribute population countries within particular administrative units (Mennis, 2003) .

Methodologies of techniques that spatially downscale population numbers continue to be refined with basic dasymetric models incorporating multiscale remotely sensed and geospatial data, in making improvements to the types of statistical algorithms used in the process (Stevens et al., 2015; Bhaduri et al., 2007; Azar et al., 2010).

It is important to note that no matter how complex and sophisticated these particular methods are, they are largely constrained by the population count data, from censuses and this forms the basis of estimation of population distributions across large areas (Linard et al., 2012; Gaughan et al., 2013).

Global positioning and GIS technologies have allowed for the improvement of census data collection and processing, but they are usually an infrequent and expensive source of detailed population data. Additionally for many low-income countries, the unreliability of estimates, low spatial resolution and lack of contemporary data often represent further limitations. This means that the latest health indicators of populations might be based on outdated or coarse input population data, which is inherently restrictive when contemporary estimates are needed for critical purposes (Tatem et al., 2011; Tatem et al., 2012; Tatem, 2014).

Populations are very dynamic, and moving daily, seasonally and annually, and thus attempts to map these dynamics for high-income countries have been made (Bhaduri et al., 2007; Leung et al., 2010). However, newer methods have opened up the use of mobile phones as a way to measure these densities more accurately in resource-poor regions of the world (Deville et al., 2014).

1.3 Statement of the problem

1.3.1 The modifiable areal unit problem

As the complexity and breadth of data collected for administrative units and census tracts increased over the years, these datasets started to become publically accessible. When areal units are divided, and partitioned, it is typically for arbitrary reasons. The most prominent issue that arises in the modifiable areal unit problem (MAUP). MAUP can best be defined as a situation where modifying the boundaries and/or the scale of data aggregation often significantly affects the results of spatial data analysis (Openshaw, 1983). Given this, it is sometimes unclear whether a census-based analysis actually indicates some reality about individuals living with that particular region or if they are strictly a function of the given areal unit that is used in an analysis. Choropleth maps of population notoriously give the impression that population is often distributed homogenously throughout an areal unit, where it is clear that often times some areas of the region are uninhabited (Dorling, 1993).

The emergence of the surface-based demographic data representation, where data is modeled as a continuous field as opposed to irregular partitioning into arbitrary areal units has often been considered to be the solution. In comparing an areal unit and a surface model of population this can be understood in the context of an object versus the field representations of geographic reality (Goodchild, 1992).

In the object view, population is a set of individual geography entities, where population attributes can be attached. In the field view, population is a continuously varying surface whose value (in this case, population density) can be measured in any given location. In reality, population of course is composed of individual people, and both the object and field representations of population are often abstractions of that particular reality. In the context of GIS, in the object view, this information is displayed using points, lines and polygons in a vector data based model, whereas in the field view this is a tessellation of square grid cells in the raster data model (Mennis, 2003).

In the open-source population mapping modelling process, surface-based population representation is much more preferred over areal unit representation, the objective of this dissertation is to suggest experiments that would improve an open-source population mapping method in addressing the modifiable areal unit problem.

1.4 Objectives of the study

1.4.1 Overall objectives

The aim of this dissertation is to provide a critical eye on how the open-source population mapping pioneered by the WorldPop project can improve to directly indicate the presence of people within its datasets more accurately. The first section will be an experiment with a tool called Google Earth Engine that can rapidly analyse vast amounts of satellite imagery to extract remotely sensed data, in this case applying a Normalized Difference Spectral Vector calculation over Landsat imagery to improve the temporal resolution of the population datasets. The second section will be an experiment utilizing volunteered geographic data in the form of Twitter data that is geo-located, providing a layer of information that is human volunteered as a covariate in the population mapping process. The final section will involve a discussion of proposed work in mapping population at a continuous 30 meters utilizing machine learning methods, as opposed to 100 meters, comparing the current state of the art methods as well as examining a method to validate land cover classifications utilizing UAVs.

1.5 Intended audience

This dissertation is intended for researchers and developers that work in the field of Geoinformatics. The writing assumes that the reader has basic understanding of principles in Geographic Information Systems and Remote Sensing. Therefore, although the topics of disease burden estimation, epidemic modeling, resource allocation, disaster management, accessibility modeling, transport and city planning, poverty mapping, and environmental impact assessment are related to the research within this dissertation, the major caveat is that these topics are examined through the lens of Geoinformatics.

1.6 Organization of Dissertation

Chapter Two will specifically examine how utilizing spaceborne instruments can improve an open-source population mapping method. NASA/USGS Landsat-derived satellite imagery will be analyzed using a tool called Google Earth Engine that can rapidly analyze vast amounts of satellite imagery to extract remotely sensed data, in this case applying a Normalized Difference Spectral Vector calculation over Landsat imagery to improve the temporal resolution of the population datasets.

Chapter Three will examine how an open-source population mapping method can be improved using utilizing terrestrial instruments: In this experiment, volunteered geographic data in the form of Twitter data that is geo-located, provides a layer of information that is human volunteered as a covariate in the population mapping process.

In Chapter Four, spaceborne and airborne instruments will be explored in mapping population at a continuous 30 meters, as opposed to 100 meters, in order to examine the limitations of the co-variate datasets that are being used as well as examining the utility in the varying resolutions for analyzing impacts of climate change on human populations. Inclusion of new airborne instruments (UAVs) will be discussed as a novel way to improve the population mapping method at this resolution, mainly for validating remote sensing classifications. The current state of the art population mapping methods at 30 meter resolution will be also be explained and examined.

Chapter Five includes a discussion on the impact of the dissertation and its place in the dasymetric mapping research, conclusions on the work of Chapters 2 through 4, and an outline of logical and meaningful future work.

CHAPTER TWO: MULTITEMPORAL SETTLEMENT AND POPULATION MAPPING FROM LANDSAT USING GOOGLE EARTH ENGINE

Nota bene: The results and rationale described in this chapter have been published as Patel et al. 2015. Some passages have been quoted verbatim and are allowed to be reprinted with the permission and attribution of the main author, N.N. Patel.

Landsat imagery has proven to be useful in understanding global urbanization trends over different timescales. Satellite-derived data have been integral in understanding trends in urban sprawl and many other dynamics of urbanization (Guindon et al., 2004;Angel et al., 2005; Burchfield et al., 2006; Schneider & Woodcock,2008; Potere et al., 2009; Schneider, 2012; Taubenböck et al., 2012;Sexton et al., 2013).

The Google Earth Engine (GEE) is an online environmental data monitoring platform that incorporates data from the National Aeronautics and Space Administration (NASA) as well as the Landsat Program. After the USGS opened access to its records of Land-sat imagery in 2008, Google saw an opportunity to use its cloud computing resources to allow records of Landsat imagery to be accessed and processed over its online system. This has enabled users to reduce processing times in analyses of Landsat imagery and make global scale Landsat projects more feasible (e.g., Hansen et al., 2013). The 30 m spatial and multispectral resolution is ideal for defining urban areas, and its revisit time is sufficient for monitoring applications (Woodcock et al., 2008). Moreover, because of Landsat’s temporal continuity from 1972 to the present day, it is a popular platform to use for urban change analysis (Alberti et al., 2004; Bagan & Yamagata, 2012; Rawashdeh & Saleh, 2006;Yuan et al., 2005).

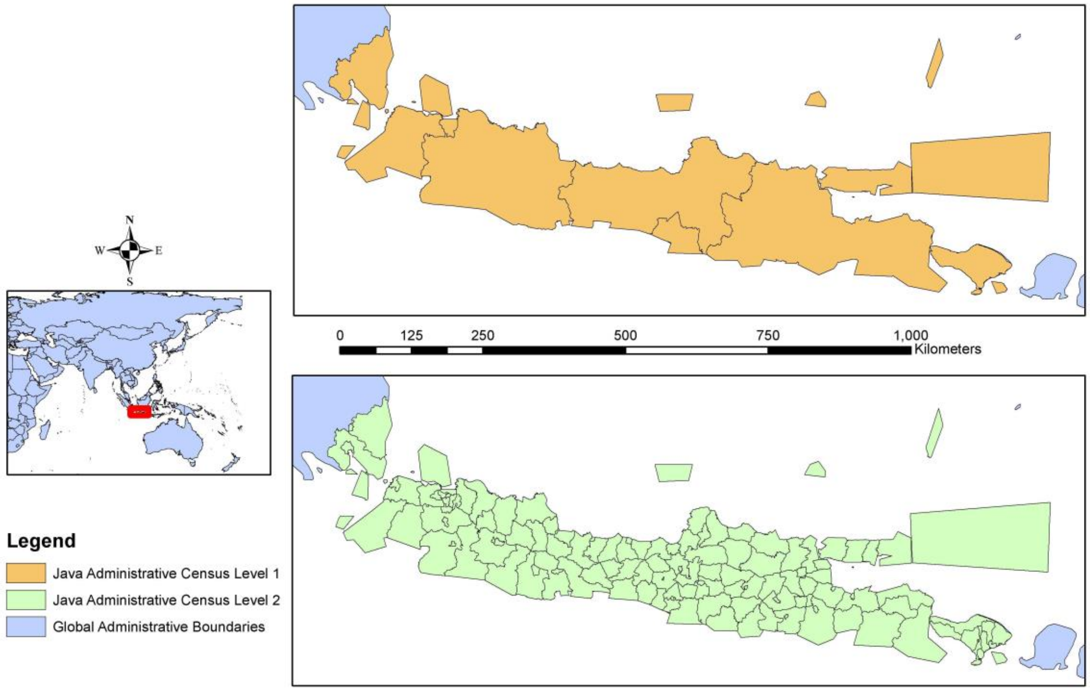
In the past two decades, the Landsat platform has been paired with imagery from the Advanced Very High Resolution Radiometer (AVHRR) (Hansen et al., 1998), the Defense Meteorological Satellite Program’s Operational Linescan System’s nighttime imagery. (Elvidge et al., 1996, 1997, 1999; Sutton, 2003), and NASA’s Moderate Resolution Imaging Spectroradiometer (MODIS) (Schneider et al., 2003, 2009, 2010) to improve the accuracy of urban detection and mapping across large areas. The improvement of methods for detecting urban extents has also driven improvements in population mapping. Satellite imagery has formed the basis of many large area population mapping efforts, such as the Global Rural–Urban Mapping Project (CIESIN, 2004) LandScan (Bhaduri et al., 2007) and WorldPop (Linard et al., 2012; Gaughan et al., 2013; The WorldPop Project, 2014). Satellite-derived urban extents and, more generally, land cover tend to form an important component of accurate population mapping (Linard & Tatem, 2012; Linard et al., 2011), but detailed data can often be costly or time consuming to produce. The GEE presents the possibility for analyzing and classifying satellite data with great speed, so that more relevant and accurate outputs in terms of distributions of population can become a reality (Hansen et al., 2013).

In this chapter an approach for the automated extraction of urban areas from Landsat imagery built into the GEE is presented, and a novel method of validation of this mapping using changes in the statistical performance of a high resolution population mapping method is also explored (Stevens et al., 2015).

2.1 Methods

2.1.1 Study area

The study area for the experiment in this chapter is the Indonesian island of Java, which, along with being the world's most populous island, is also only the fourth largest island in Indonesia but contains more than half of the island nation's population. Jakarta, the capital city is also located on the island and is Indonesia's largest city. The island is 661 miles long from east to west; it ranges in width from about 60 miles in the center to more than 100 miles near each end (Figure 1).


N
W
E
S
0
125
250
500
750
1,000
Kilometers
Legend
🟧 Java Administrative Census Level 1
🟩 Java Administrative Census Level 2
🟦 Global Administrative Boundaries

Sources: Global Administrative Areas (GADM, http://www.gadm.org), Thematic Mapping (http://thematicmapping.org/)

Figure 1: Map of study area and Java administrative boundaries levels 1 and 2.

2.1.2 Urban extent extraction procedure

The urban extraction methodology proposed in Patel et al, 2015, is based on supervised classification of multispectral data. In this work we consider “urban areas” all the portion of a scene with spectrum similar to selected training areas. These training areas include buildings, roads, and other artificial surfaces. Therefore, in the following “urban extents” do not correspond to “built-up extents”. Thus, the definition of urban areas is instead more similar to “impervious surfaces.”

Accordingly, the implemented processing chain is a spectral-based analysis followed by a spatial regularization that is undertaken using the Google Earth Engine cloud computing environment. Processing and implementation in a cloud environment allows for a consistent scaling of the computational efforts when dealing with wide geographical areas. The extraction procedure includes three steps, briefly detailed below: (i) preprocessing and selection of a set of Landsat scenes covering the geographical area and time span of interest, (ii) computation of the Normalized Difference Spectral Vector index (NDSV), a collection of spectral indices that have already been proven (Angiuli & Trianni, 2014) to be an efficient input to urban extent classification algorithms classification and, (iii) spatial-based post-processing.

2.1.2.1 Pre-processing and scene selection

Preprocessing includes orthorectification and coregistration of all the scenes, so that data acquired at multiple dates overlap. This is done internally and seamlessly by the GEE platform at the ingestion of ingestion of the data from the USGS repository. No radiometric intercalibration or atmospheric correction is performed, however. Although all scenes are calibrated according to the sensor parameters, some differences in radiance values due to the illumination and atmospheric conditions still affect overlapping regions among scenes.

Scene selection is instead performed by algorithm developed in Trianni et al., 2015. Specifically, in order to reduce the Landsat data set to the most suitable scenes, a filter on scene parameters is first applied, to consider only those with less than 10% of cloud coverage and the highest radiometric quality.

2.1.2.2 Implementation of the Normalized Difference Spectral Vector index stack into the Google Earth Engine

Unlike threshold-based recognition of human settlement (one index) approaches developed by Pesaresi et al. (2008) and Xu (2008), the main input to the urban extent extraction outlined here is the Normalized Difference Spectral Vector (NDSV), proposed in the technical literature (Angiuli & Trianni, 2014 ; Trianni et al., 2015) as a means to group existing normalized difference indices (such as the Normalized Difference Vegetation Index—NDVI, the Normalized Difference Water Index—NDWI, and the Normalized Difference Built-up Index—NDBI). NDSV includes in one single vector all the possible normalized indexes that can be computed starting from a Landsat 5 or 7 image, considering therefore 6 bands and 15 possible combinations (the dual ones are not considered as their result is the same but with just the opposite sign). NDSV includes in one single vector all the possible normalized indexes that can be computed starting from the 30 m spatial resolution bands a Landsat 5 or 7 image. For each band pair this is computed:

$$NDSV_{ij} = \frac{b_i - b_j}{b_i + b_j}$$

(1)

Hence, using 6 bands and applying Eq. (1) to any possible pair of different bands, a total of 30 indexes are obtained. Due to the symmetry of the definition, 15 of them are only the negative of the other ones, and can be discarded. Each pixel is thus characterized by a set of values, some of which correspond to known indexes (e.g., NDSV43= NDVI, NDSV42= NDWI, NDSV45= NDBI), while other ones have not been explored so far. Each pixel is thus characterized by a set of values that have been at this point "labeled" only partially. Considering a radiometrically and geometrically corrected Landsat scene, the NDSV features characterizing urban areas, compared to other classes, are shown for a few sample pixels in Figure 2. It can be noted that urban areas exhibit a distinct NDSV spectral signature which can be discriminated from other classes by their distinct behavior in this new "multispectral"15-dimensional space. Figure 2 demonstrates NDSV profiles that can be obtained from an image. In summary, instead of relying on threshold-based recognition of human settlements according to a single index (Pesaresi et al., 2008; Xu, 2008), the procedure implemented in this work considers more information as input to a suitable classification chain, aimed at providing a consistent methodology that works in many different environments, and is reasonably robust with respect to the date of acquisition of the image and unaffected by differences in spatial patterns (Trianni et al., 2015). See Appendix II for further detail on how the NDSV calculation was applied in the GEE environment. See Appendix III for further detail on the computing environment utilized to replicate the full experiment detailed in this section as well as other sections in this dissertation.

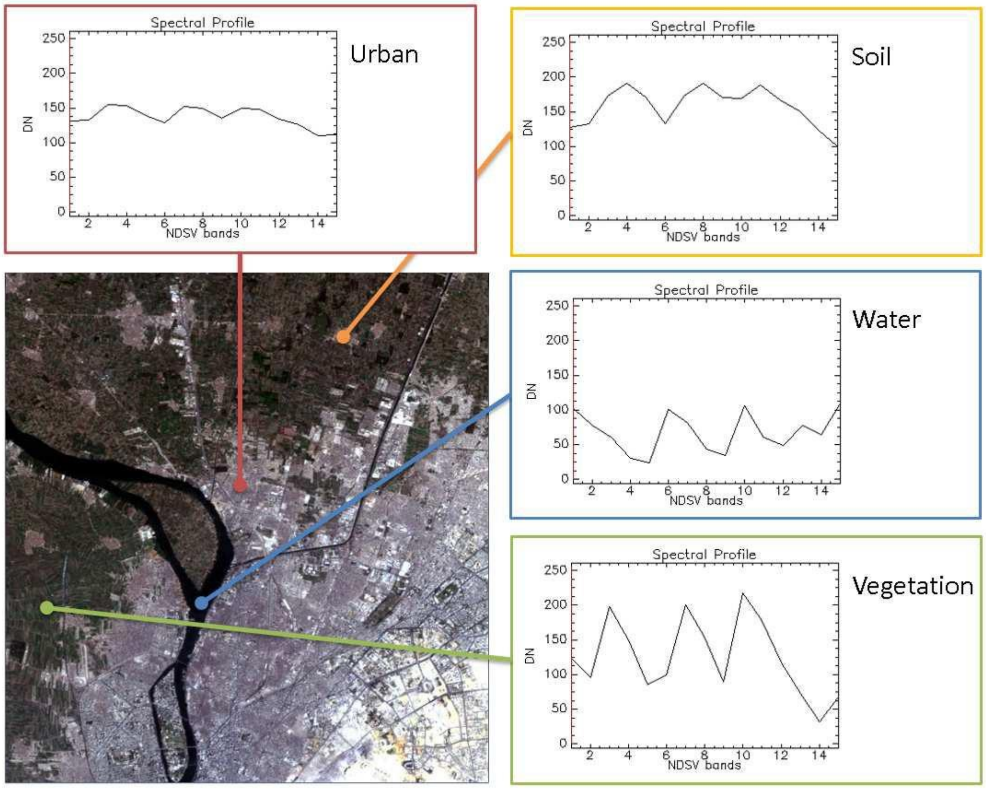


Figure 2: Normalized difference spectral vector (NDSV) profiles for urban areas, vegetation, water, and bare soil.

2.1.2.3 Processing of multitemporal urban extents over Java

Four tests were conducted in order to validate the creation of urban extents using the procedure discussed in the preceding sub-sections. A census-based population disaggregation method was used for validation, a method that rasterizes GIS data and distributes population counts based on the GIS data that is provided. This method was used because it provides the ability to analyze how the urban extents improve the statistical correlations in the disaggregation process. In three out of four tests, the urban extents were considered as one of the inputs to a census-based population disaggregation method (Stevens et al., 2015). In the first test, instead, the same method was run using the original data sets detailed in Table 1 and the land cover map, including urban extents, was taken from the EarthSat Geocover land cover thematic mapper-based dataset(2007, 30 m) by MDA Federal (MDA Federal Inc., 2007). To read further detail on remotely sensed data utilized in the population mapping process that is referred to throughout this dissertation, see Appendix I. Test 1 served as the baseline data for validation because it does not use any Google Earth Engine urban extents and because its classification has been validated by MDA Federal, serving as a useful control test. The qualitative differences of these different land cover-based experiments:

- Test 1: EarthSat GeoCover Landsat Thematic Mapper (TM) derived land cover data from MDA Federal (2007)
- Test 2: GEE urban extents for Java derived using three collections: imagery from 2006, 2007 and 2008 merged with GeoCover.
- Test 3: GEE urban extents for Java derived using three collections: imagery from 2009 T1 (January through April), 2009 T2 (May through August), 2009 T3 (September through December) merged with GeoCover.
- Test 4: GEE urban extents for Java derived using three collections: imagery from 2008, 2009, and 2010 merged with GeoCover.

The GEE urban extractions were obtained using Landsat 5 or Landsat 7 data sets, because both satellites were operative in the years of interest. Specifically, multiple Landsat images in the same area and covering a finite period of time were combined in a so called GEE collection, and each pixel was assigned the median value for all images where it appears. Collections are a powerful way to get rid of many of the cloud-contaminated pixels, because clouds do not appear in the same position in all images. A better approach would be to mask cloud pixels with a dedicated filter, a function which is unavailable in GEE. Although cloud-contaminated pixels may still be present in areas with consistent cloud coverage along the year, this technique was assumed as the best available option. Additionally, it must be noted that collections change the radiometric properties of the data, reducing the effectiveness of the proposed urban extent procedure. To reduce this effect, urban extents for one year were obtained by subdividing the year into thirds. Computing collections for each of these time periods involved extracting urban extents and then combining the resulting maps by majority voting. Similarly, three year collections were subdivided into thirds (one for each year) and then combined by majority voting. To prove the usefulness of the proposed approach for mapping urban extents (and derive population counts) along multiple years, the fourth test repeats the approach of the third one, but using Landsat data collected two years later(2009 versus 2007).

2.1.2.4 Post-processing

Human settlements can be characterized by peculiar spatial patterns, however, it is important to include a post-processing step aimed at reducing issues related to misclassifications at the pixel level. The simplest and most effective approach is to include morphological operators aimed at discarding isolated pixels and at improving the homogeneity of the extracted settlements with respect to their spatial distribution. Additionally, as the classification results may be affected by spectral patterns (and sub-pixel mixing problems) similar to urban ones in water bodies with high turbidity (Carpenter & Carpenter, 1983; Foody, 2000), such as inner reservoirs, coastal areas, and river estuaries, these zones are automatically masked out from the classification in GEE using ancillary GIS data. Similar issues may be caused by clouds, and thus “cloud removal” approaches had to be considered. See Appendix II for further detail on the post-processing steps.

2.1.3 High resolution population mapping method

As mentioned above, the population mapping algorithm in Stevens et al. (2015) serves as the open-source population mapping method being utilized. Thus, its processing steps are briefly described in the following paragraphs.

2.1.3.1 Population data grid

The 130 census polygons for Java (Figure 1) contained population counts from the year 2010. The population mapping algorithm outlined in Stevens et al. (2015) was used, where census counts from the census year are redistributed according to weights, then adjusted up/down based on rural and urban growth rates to a particular year of interest (2007 in this case). This is usually based on the classified urban/rural land cover (built pixels are classified as urban vs. rural using Schneider et al. (2010) urban/rural MODIS-derived classifications), but in this circumstance uses the new GEE-derived urban delineations to identify urban built pixels. The urban/non-urban delineation was integrated into the MDA landcover data as “built” areas (“BLT”). The particular year of interest that was selected was 2007 for all datasets, to pick one year for counts to match and for a point of comparison for the accuracy assessment detailed in section ‘Accuracy assessment’.

The administrative units were used to delineate the areas where the land cover data in continuous raster format and converted vector format are interpolated by means of the random forest method to generate a weighting layer (Stevens et al., 2015). Once this weighting layer is generated, population counts for each census unit are distributed over the weighting layer to provide a map of population counts at a 100 by 100 meter resolution (See Table 1 for detail on all covariate datasets used in the process).

Table 1: Test-specific data sources and variable names used for population density estimation used for dasymetric weights for Chapter Two’s Experiment.

Type	Variable Name(s)*	Description	Indonesia Data
Census		Country-specific census data that is used for disaggregation	2010, Admin-level 2 (GADM , 2014), (GeoHive, 2014)
Land Cover	lan_cls011, lan_dst011	Cultivated terrestrial lands	Landcover utilizing 3-Year Google Earth Engine data & MDA GlobCover with methods from Patel et al. 2015 used.
	lan_cls040, lan_dst040	Woody / Trees	Ibid
	lan_cls130, lan_dst130	Shrubs	Ibid

	lan_cls140, lan_dst140	Herbaceous	Ibid
	lan_cls150, lan_dst150	Other terrestrial vegetation	Ibid
	lan_cls160, lan_dst160	Aquatic vegetation	Ibid
	lan_cls190, lan_dst190	Urban area	Ibid
	lan_cls200, lan_dst200	Bare areas	Ibid
	lan_cls210, lan_dst210	Water bodies	Ibid
	lan_cls230, lan_dst230	No data, cloud/shadow	Ibid
	lan_cls240, lan_dst240	Rural settlement	Ibid
	lan_cls250, lan_dst250	Industrial area	Ibid
	lan_clsBLT, lan_dstBLT	Built, merged urban/rural class	Ibid
Continuous Raster-Format			
	Lig	Lights at night data	Suomi VIIRS-Derived (NOAA, 2012)
	Tem	Mean temperature, 1950-2000	WorldClim/BioClim (Hijmans et al,, 2005)
	Pre	Mean precipitation, 1950-2000	WorldClim/BioClim (Hijmans et al,, 2005)
	Ele	Elevation	HydroSHEDS (Lehner et al., 2006)

	ele_slope	Slope	HydroSHEDS-Derived (Lehner et al., 2006)
Converted Vector-Format			
	roa_dst	Distance to roads	OSM (2013)
	riv_dst	Distance to rivers/streams	OSM (2013)
	pop_cls, pop_dst	Generic Populated Places	VMAP0 merged†
	wat_cls, wat_dst	Water bodies	World Food Programme
	pro_cls, pro_dst	Protected areas	IUCN and UNEP, 2012
	poi_cls, poi_dst	Populated Points	OSM (2013)
	bui_cls, bui_dst	Buildings	OSM (2013)

* The variable names are used in Random Forest model output and throughout the text as reference to the specific data they were derived from. The first three letters are derived from the data type (e.g. "lan" indicates land cover) and the last three letters, if present, indicates what type of data each variable represents (e.g. "_cls" is a binary classification and "_dst" is a calculated Euclidean distance-to variable.

† The default data for populated places is merged from several VMAP0 data sources. There are three VMAP0 data sets used: The point data pop/builtupp and pop/mispopp are buffered to 100 m and merged with the pop/builtupa polygons creating a vector-based built layer. This layer is then converted to binary class and distance-to rasters for use in modeling. (NGA, 2005)

2.1.3.2 Data preparation and the random forest population disaggregation method

The general process used for the data preparation, modeling, and validation for the population mapping is outlined in Figure 3. Full details on these steps are provided in Stevens et al. (2015). In brief, the steps in green represent the data preparation tasks. The aggregated population counts and the raster and vector layers shown in Table 1 are then used to create a random forest model (Breiman, 2001) to predict log population density. Random forest (RF) models are an ensemble, nonparametric modeling approach that grows a “forest” of individual classification or regression trees and improves upon bagging (Breiman, 1996) by using the best of a random selection of predictors at each node in each tree (Breiman, 2001; Liaw & Wiener, 2002).As expected when combining multiple observations that are mostly independent, the best, most unbiased prediction was arrived at by taking the mean of all trees within the forest and back-transforming the log to arrive at an estimate of per-pixel population density. Medians and percentile ranges were also assessed as alternative approaches for prediction; however, the back-transformed mean consistently out-performed the alternative summary methods during validation. The resulting country-wise population density map was then used as a weighting layer for a standard dasymetric mapping approach as described for the AfriPop and AsiaPop (now WorldPop) data sets by (Gaughan et al.,2013; Linard et al., 2012; Linard & Tatem, 2012; Tatem et al.,2007).

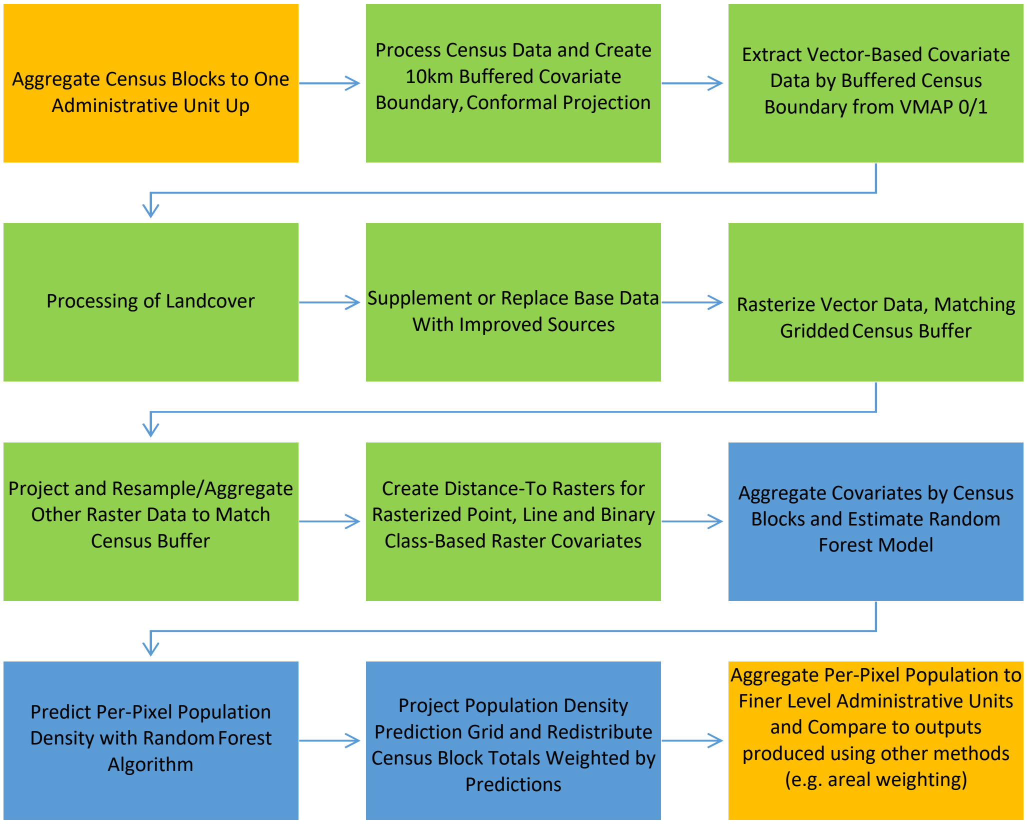

Aggregate Census Blocks to One Administrative Unit Up
Process Census Data and Create 10km Buffered Covariate Boundary, Conformal Projection
Extract Vector-Based Covariate Data by Buffered Census Boundary from VMAP 0/1
Processing of Landcover
Supplement or Replace Base Data With Improved Sources
Rasterize Vector Data, Matching Gridded Census Buffer
Project and Resample/Aggregate Other Raster Data to Match Census Buffer
Create Distance-To Rasters for Rasterized Point, Line and Binary Class-Based Raster Covariates
Aggregate Covariates by Census Blocks and Estimate Random Forest Model
Predict Per-Pixel Population Density with Random Forest Algorithm
Project Population Density Prediction Grid and Redistribute Census Block Totals Weighted by Predictions
Aggregate Per-Pixel Population to Finer Level Administrative Units and Compare to outputs produced using other methods (e.g. areal weighting)

Figure 3: General structure of the data processing and map production procedure used to compare the methodology outlined in Stevens et al. (2015). The orange boxes represent items that are specific to the research presented here and not part of end-user map data product generation. The green boxes represent data preprocessing stages. Items in blue represent random forest model estimation, per-pixel prediction and dasymetric redistribution of census counts.

2.1.3.3 Accuracy assessment

The four output population maps produced using administrative level 1 input census data (Figure 1), were then compared to the level 2 census counts to provide one method of assessing mapping accuracies, following Gaughan et al. (2013). The individual cell values of the output population maps represent people per cell, and were then added together for each census unit. These "predicted" sums were then compared with the observed census counts within each unit. Summary statistics were then calculated, including root mean square error (RMSE), the RMSE divided by the mean census unit count (%RMSE) and the mean absolute error (MAE). Together these statistics were used to compare the predictive ability of each methodology.

2.2 Results

2.2.1 Urban extraction results

For the urban extraction results in the test areas, all of them referring to Landsat scenes recorded in 2007, the validation was performed as follows: human settlement extents were manually digitized from Very High Resolution (VHR) Quickbird images available in Google Earth TM, and recoded in 2007, if possible in the same month of the corresponding Landsat scene. The relatively small cities of Manado and Bandung, as well as the big urban agglomeration of Jakarta were considered. This validation was performed in accordance to the methodologies described in Trianni et al. (2015).

The mapping results are shown in Figure 4, while the quantitative validation results for Manado with and without spatial post-processing (see section 'Implementation of the normalized difference spectral vector index stack into the Google earth engine') are reported in Table 2. Visually, the approach shows an accurate extraction of the human settlement extents at the pixel level, with a few misclassifications outside the actual urban area, and missing areas within the boundary of the larger blocks. The quantitative evaluation shows instead a large omission error percentage. After post-processing, however, the overall accuracy improves to 85% and the omission error decreases from 87% to below 19%. Satisfied with the relative accuracy of detecting urban areas using the

NDSV classifier on the GEE system, the process was applied to three collections on the Google Earth Engine, and then integrated with the MDA land cover dataset. This combined land cover dataset, using the GEE-derived built area delineations was then applied to the population mapping process and evaluated statistically for prediction accuracy.

Table 2. Confusion Matrices for Kota Manado without (top) and with (bottom) the Spatial Post-Processing Step (Figure 4):

Overall Accuracy = (2242/4000) 56.05%			
	Ground Truth (Pixels)		
Class	urban	non urban	Total
urban	245	3	248
non urban	1755	1997	3752
Total	2000	2000	4000
Overall Accuracy = (3398/4000) 84.95%			
	Ground Truth (Pixels)		
Class	urban	non urban	Total
urban	1625	227	1852
non urban	375	1773	2148
Total	2000	2000	4000

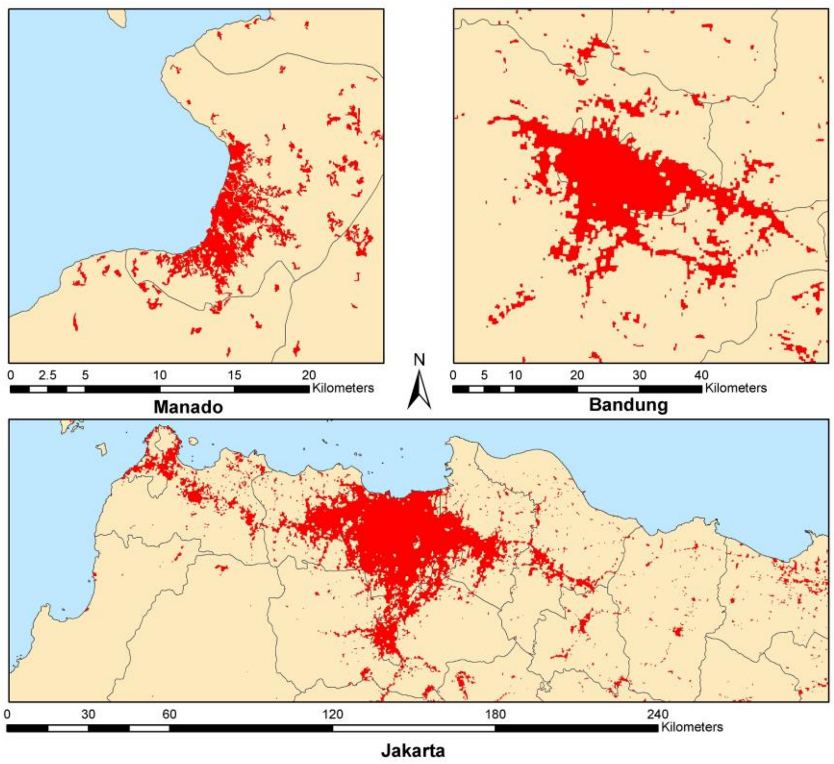


Figure 4: Human settlement extraction results for Manado, Bandung and Jakarta, in Indonesia.

A small sample of the urban extents generated for tests 2, 3, and 4 are shown in Figure 5 for the central part of Jakarta along with the urban extents for the same area in the MDA data set.

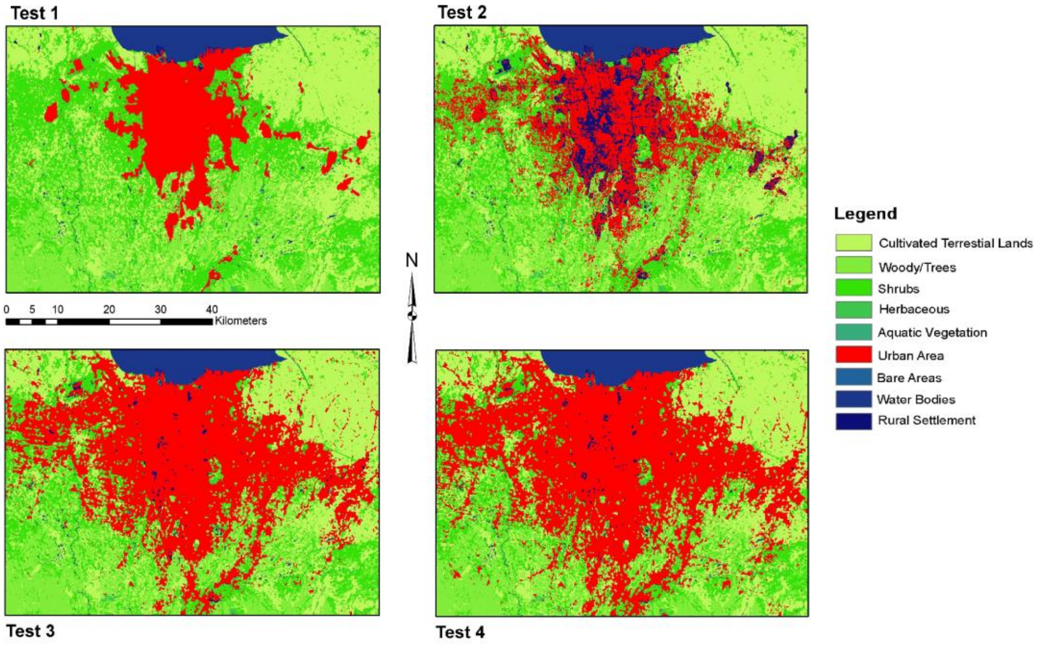


Figure 5: A small sample of the area around central Jakarta used in the tests. Test 1 displays the EarthSat GeoCover Land Cover Thematic Mapper from MDA Federal (reflecting extents from 2007). Tests 2, 3, and 4 represent the Google Earth Engine-derived extents that are merged into Test 1. Test 2 integrates urban areas from 3 collections from 2006, 2007, and 2008 (a collection for each year), test 3 integrates urban areas from 3 collections in 2009, and test 4 integrates urban areas from 2008, 2009, and 2010 (a collection for each year). Classifications reflected in the Legend are all from the MDA Federal dataset other than the "Urban Area" class, which was obtained from the Google Earth Engine derived NDSV extents.

2.2.2 Random forest statistical output

The differences between results are determined by the ancillary datasets used in the population mapping detailed in Table 1.

Referring to the covariate names in Table 1, there are two significant covariates in the random forest mapping process, "BLT" (Built) and "lig" (VIIRS Nightlights). Table 3 provides some insight into the importance of the variables in the mapping process by showing how much mean squared error (MSE) increases when the specified covariate is randomly permuted and predictions re-calculated. The most important variables include the “BLT” covariates, indicating “Built” areas, which include urban and rural settlements. In addition, for all tests, except for test 4 (GEE 2008–2010), the “lig” (VIIRS Nightlights data) have higher importance than other covariates.

Table 3: Top five statistical outputs: percent increase of mean squared error when variable is randomly permuted and total decrease in residual sum of squares when variable is selected for decision tree node.

Percent Increase of Mean Squared Error When Variable Randomly Permuted	
Test 1 (MDA) (total of 23 covariates used), (81% Variance Explained)	**Test 2 (GEE 2006-2008) (total of 22 covariates used), (83% Variance Explained)**
20.2 (Lights)	18.4 (Lights)
12.7 (Landcover Distance to Built Areas)	17.4 (Landcover Distance to Built Areas)
10.8 (Distance to Populated Points)	10.6 (Distance to Populated Points)
9.24 (Distance to Buildings)	7.91 (Distance To Generic Populated Places, VMAP0)
9.08 (Landcover Distance to Cultivated Terrestrial Areas)	7.79 (Landcover Distance to Cultivated Terrestrial Areas)
Test 3 (GEE 2009) (total of 22 covariates used), (83% Variance Explained)	**Test 4 (GEE 2008-2010) (total of 23 covariates used), (84% Variance Explained)**
19.3 (Lights)	19.8 (Landcover Distance to Built Areas)
16.5 (Landcover Distance to Built Areas)	18.8 (Lights)
9.13 (Distance to Populated Points)	8.00 (Distance to Populated Points)
7.13 (Distance to Roads)	7.77 (Landcover Distance to Cultivated Terrestrial Areas)
6.90 (Landcover Distance to Cultivated Terrestrial Areas)	7.23 (Distance to Roads)
Total Decrease in Residual Sum of Squares When Covariate Used	
Test 1 (MDA) (total of 23 covariates used) , (81% Variance Explained)	**Test 2 (GEE 2006-2008) (total of 22 covariates used), (83% Variance Explained)**
53.8 (Lights)	49.3 (Landcover Distance to Built Areas)
31.7 (Landcover Distance to Built Areas)	43.7 (Lights)
19.1 (Distance to Roads)	17.7 (Distance to Populated Points)

17.2 (Distance to Populated Points)	16.9 (Distance to Roads)
16.9 (Distance to Buildings)	13.7(Distance to Buildings)
Test 3 (GEE 2009) (total of 22 covariates used), (83% Variance Explained)	**Test 4 (GEE 2008-2010) (total of 23 covariates used), (84% Variance Explained)**
56.5 (Landcover Distance to Built Areas)	58.4 (Landcover Distance to Built Areas)
46.6 (Lights)	45.3 (Lights)
15.3 (Distance to Roads)	15.3 (Distance to Roads)
13.5(Distance to Populated Points)	12.0 (Distance to Populated Points)
9.76 (Distance to Generic Populated Places, VMAP0)	11.2 (Distance to Buildings)

Table 3 also displays the increase in node purity in each test, which documents reduction in residual sum of squared error for the predictions at the ends of the branches of each tree when the specified variable is used during the random forest mapping process. Again referring to the variables detailed in Table 1, it can be observed that the "BLT" (built) classes with the GEE integrations in tests 2, 3, and 4 are the most important in the random forest process.

Again referring to the variables detailed in Table 1, it can be observed that the "BLT" classes with the GEE integrations in tests 2, 3, and 4 are making the built classes the most important in the random forest process.

2.2.3 Random forest accuracy assessment

The accuracy assessment process detailed in Section 2.1.3.3 shows how much the urban extents improve the output when the census data were aggregated from district to province. The tests in the previous sections detail how well the RF does in predicting population values at the census unit level, but more importantly is whether the population map produced using built land cover data from the three GEE-derived approaches is better at redistributing the population numbers from coarser census units. Two different error assessment methods are presented: root mean square error (RMSE), also expressed as a percentage of the mean population size of the administrative level (% RMSE); and the mean absolute error (MAE).

For both RMSE and MAE, the results in Table 4 indicate that test 4 increased population mapping accuracy the most, with test 3 slightly better than test 1. Notably, the urban extraction from test 2, which used built extents derived from years 2006 to 2008, had the lowest redistribution accuracy. It is notable that the land cover changes allow for test 3 and test 4 to outperform the MDA dataset in reducing error, creating more concurrent built data to correlate better with the other datasets.

Table 4: Accuracy Assessment Results for Four Urban Land Cover Treatments

	RMSE	%RMSE	MAE
Test 1 (MDA)	1450.286	0.129064	787.8362
Test 2 (GEE 2006-2008)	2277.501	0.20268	1352.685
Test 3 (GEE 2009)	1377.889	0.122621	773.6329
Test 4 (GEE 2008-2010)	1346.32	0.119812	759.3168

2.3 Discussion and conclusions

The possibilities that the Google Earth Engine offers in analyzing remotely sensed data on a global scale with the power of Google’s cloud computing are substantial. The inclusion of continuously updated Landsat data along with classification tools and significant processing power will enable newer and more accurate ways to map human settlements across large areas at 30 m spatial resolution, document past changes, and continually update current estimates. The potential of this resource has been recently illustrated for multitemporal forest mapping (Hansen et al., 2013), and in this experiment initial steps for similar efforts in human settlement and population mapping are explored.

The application of the NDSV within the GEE shows significant potential for settlement mapping within the tool. Characterizing human settlements can be considered as a binary problem, but where the “non-urban” class is very heterogeneous. It therefore requires a classifier which is nonparametric, i.e., that does not assume any peculiar statistical distributions of the input values .Moreover, since the NDSV is built through a composition of 15 bands, the classifier has to be able to manage high-dimensional spaces. Therefore, classifiers developed for hyperspectral data are preferable, using, for example, the spectral angle mapper classifier (Angiuli & Trianni, 2014), that captures the differences in multispectral vectors and is robust with respect to difference in illumination. Since this classifier is not available in the GEE environment, support vector machines (SVM) and classification and regression trees (CART) were considered instead (Earthengine-api, 2014), with similarly strong results shown.

Both the SVM and CART are suitable to binary problems, but the tests in Patel, et al. (2015) suggested that CART produced more accurate urban extent maps. The statistical indices explored in the random forest population mapping process in Table 3 highlight to what degree the distance to “built” environments (lan dstBLT) covariate plays a role in reducing error and increasing the quality of the output of the population mapping process. When the focus was on which variable, if removed, would increase the RMSE, the GEE experiments (tests 2, 3, and 4) showed that the distance to “built” covariate was an important one. Table 3 also reflects the same results in increasing node purity in the process. It is important to note that in tests 2 and 4, urban extents extracted in 3 consecutive years are combined, while in test 3 a single year is considered. Test 2 showed the greatest amount of error, utilizing urban extents that were obtained from the GEE for years 2006, 2007, and 2008. It is clear that the modification of the land cover from test 1 for the same time period reflected in test 2, changes the areas within Jakarta significantly. The improved accuracy of test 1 over test 2 could just reflect a better correlation of values instead of informing what is making the data more spatially significant, and in that circumstance, it can be argued that the GEE urban extents can be a critical component in the creation of multitemporal datasets that can modify existing land cover datasets in order to examine trends, in an efficient manner, for different years. Table 4 shows how the integration of GEE extents correlates well in the population mapping process and decreases error, by adding more concurrent built data along with the other covariate datasets.

In using census data from 2010, land cover data closest to this year stands a better chance of being the best proxy for disaggregation if all other factors are equal. In this sense, there is an inherent bias in the tests, but it also highlights the benefits of the GEE approach, that is being able to produce an accurate urban extent map for any time period, with the ability to match up land cover data. Overall, the NDSV is shown here to be a reliable method to detect urban extents, especially when using a powerful tool to analyze the data such as the GEE. Moreover, the GEE represents one of the most powerful tools offered today in remote sensing with its ability to analyze and classify remotely sensed data over different temporal scales. Finally, the use of NDSV derived extents produced in the GEE and integrated in a flexible population mapping method enables testing of the validity of the classifications in improving population distribution mapping, providing an additional novel accuracy assessment approach. As urbanization processes continue to accelerate in many countries around the world, accurate, powerful, and efficient methods for rapid mapping of settlements and their changes, as well as populations within them are a prerequisite for strategic planning and impact assessments. The results in this Chapter point towards the integration of classification and population mapping methods within GEE as a way of meeting this need.

CHAPTER THREE: IMPROVING LARGE AREA POPULATION MAPPING USING GEOTWEET DENSITITES

Nota bene: The results and rationale described in this chapter have been published as Patel et al. 2016. Some passages have been quoted verbatim and are allowed to be reprinted with the permission and attribution of the main author, N.N. Patel.

One of the most popular social media applications over the past decade has been Twitter[2]. Twitter is an online social networking service that allows users to post 140-character messages called "tweets" to a publicly viewable microblog platform, and since its inception in 2006, the service has gained worldwide popularity. Despite the relative lack of tweets with geographic metadata (around 2.02% of tweets are posted with such metadata globally), many useful geographic applications have been derived from tweet data (Takhteyev et al., 2012; Leetaru et al., 2013; Hawelka et al. 2014; Blanford et al., 2015). The maps of geo-located tweets in countries where Twitter is popular show detailed depictions of human activity, with the location of tweets indicative of settlements, transportation networks, and building locations (Leetaru et al., 2013). Such data therefore have the potential to provide a valuable ancillary covariate layer in the population mapping process, and also one that changes dynamically, but its utility has yet to be tested.

[2] https://twitter.com

In this chapter, the potential of geo-located tweets (submitted through terrestrial instruments) to improve population distribution maps is assessed. Tweets are integrated as a covariate layer into a census data disaggregation model. The accuracy of gridded population maps produced with and without geotweet data are compared. The advantages and disadvantages of such social media in improving population mapping accuracies in low and middle income settings are discussed. See Appendix III for further detail on the computing environment utilized to replicate the full experiment detailed in this section as well as other sections in this dissertation.

3.1 Methods

3.1.1 Study Area

Indonesia has one of the highest Twitter user levels in the world (Leetaru et al., 2013) and it also has recent, very high spatial resolution census data. These characteristics combined make the country an ideal case study for the utility of geotweet data for census count disaggregation. The study area is the country of Indonesia, an archipelago made up of thousands of islands, with total land area being around 1.9 million km^2. For the purposes of this study, boundary-matched census data at the Kecamatan administrative level (Level 3, 6,463 units) and Desa/Kelurahan administrative level (Level 4, 79,618 units), were obtained (Figure 6).

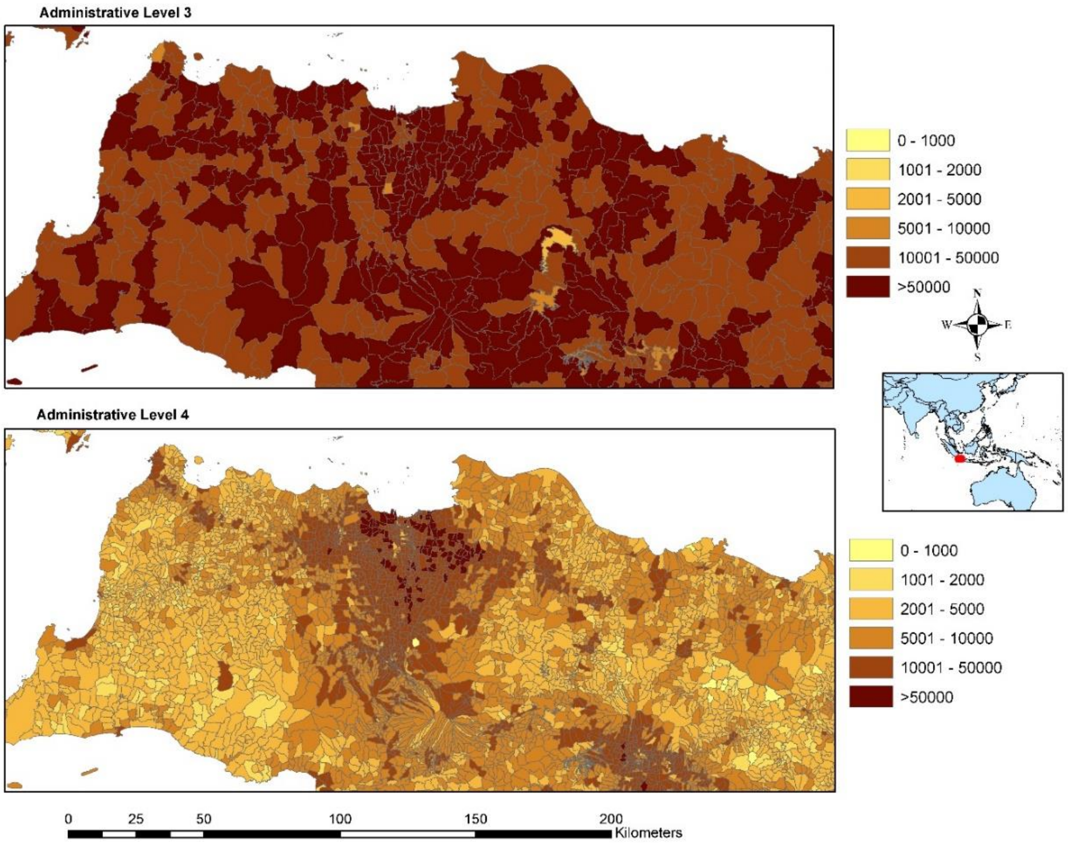


Figure 6: Map of Indonesia administrative boundaries levels 3 and 4, focused around Jakarta, with administrative units shaded to show population counts per administrative unit.

3.1.2 Mapping Geotweets

Two months of geo-located tweets (Jul.8-Aug.8 and Oct.12-Nov.15, 2013) were extracted from the Twitter Streaming API[3] for Indonesia. Morstatter et al. (2013) show the Twitter Streaming API provides around 90.1% coverage of the total available set of geotagged tweets. Note that the geotagged tweets with exact latitude and longitude make up about 1.6% of total number of tweets, which is about 79% of tweets posted with

[3] https://dev.twitter.com/streaming/overview

general geographic metadata (See Twitter Places attributes[4]) (Leetaru et al., 2013). Similar to the process described in Morstatter et al. (2013), the data streaming was performed by utilizing the Tweetpy library[5] on an Amazon Web Service[6] Instance. Further reference on Twitter Streaming API and the most up-to-date usage agreements can be found here[7].

The collected data was automatically uploaded to a storage bucket on Amazon Simple Storage Services (S3)[8] every day during the study period. An Apache Pig [9]process was initiated on the Amazon Elastic MapReduce[10] web service to extract all the tweets within the geographic boundary constraint of Indonesia and to aggregate the raw tweet activity counts into a 0.001 degree by 0.001 degree grid in an unprojected geographic coordinate system. During the aggregation process, the origin latitude and longitude from the geotagged tweets were rounded down to three decimal digits. The list with latitude, longitude and tweet counts were imported into ArcGIS Desktop[11] to produce a raster layer for testing as a covariate in population mapping (Figure 7).

[4] https://dev.twitter.com/overview/api/places#attributes
[5] https://github.com/tweepy/tweepy
[6] http://aws.amazon.com/
[7] https://dev.twitter.com/
[8] http://aws.amazon.com/s3/
[9] https://pig.apache.org/
[10] http://aws.amazon.com/elasticmapreduce/
[11] http://www.esri.com/software/arcgis/arcgis-for-desktop

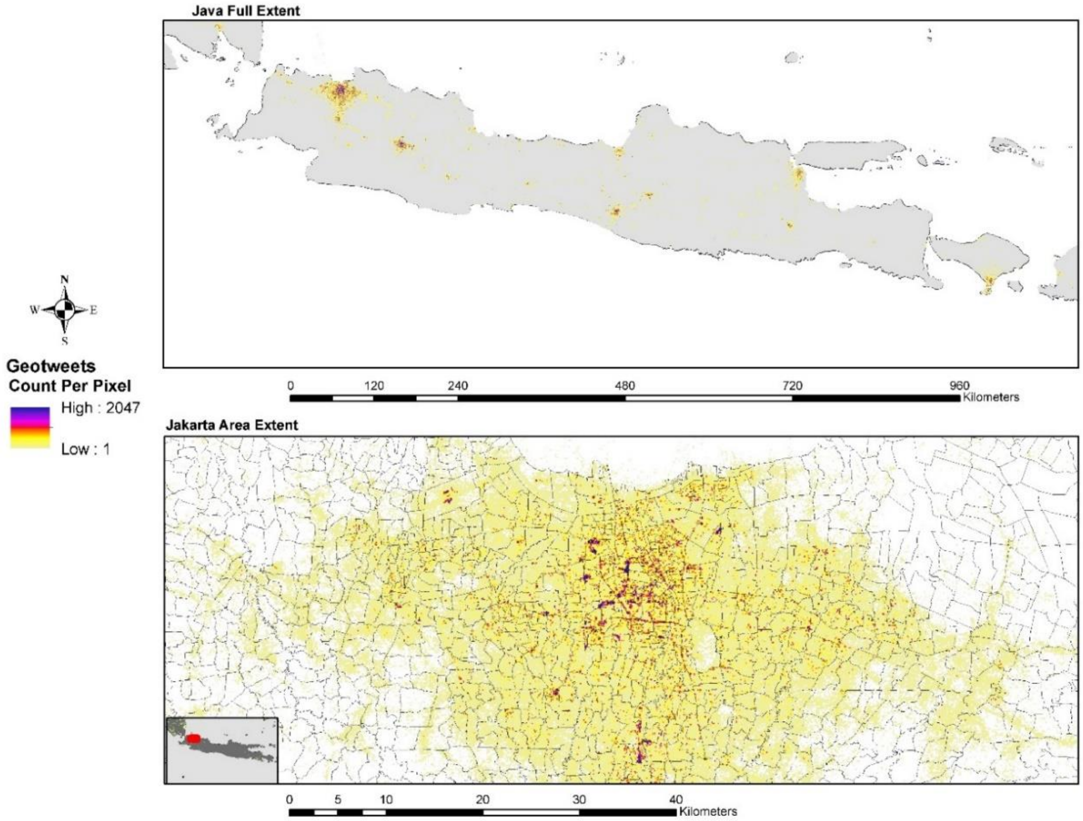


Figure 7: Results of a two-month aggregation of geo-located tweets over the full extent of Java (top) and a view focused on Jakarta (bottom).

The geotweets raster dataset was then integrated into a population mapping process along with other ancillary covariates to disaggregate the administrative unit level 3 census data to a 100 meter by 100 meter grid using a population mapping process detailed in the next section.

3.1.3 High Resolution Population Mapping Method

The population mapping method detailed in Stevens et al. (2015) was utilized to undertake two tests - both mapping the entire Republic of Indonesia, and disaggregating administrative 3 data with and without the extracted geotweets to assess whether the inclusion of the tweets improved mapping accuracies when compared to the administrative level 4 data.

3.1.3.1 Data Processing

Indonesian census counts for 2010 were obtained from the Indonesian Government and matched to GIS-administrative boundaries at administrative level 3 (6,463 units, total spatial area calculation from shapefile = 3,364,560.063 km^2) and administrative level 4 (79,618 units, total spatial area calculation from shapefile = 3,362,579.043 km^2). Both data sets have a total population of 243,530,782 and an average spatial resolution (ASR) of 22.8 and 6.50, for the administrative level 3 and the administrative level 4, respectively. The ASR is calculated as the square root of its surface area (in square kilometers) divided by the number of total administrative units (Balk & Yetman, 2004) and provides a broad measure of mean administrative unit size across the country. When calculated by province (administrative unit level 2), the ASR varies from 1.68 to 68.9 for level 3 units and from 0.919 to 30.8 for level 4 units. The administrative level 3 census data was used in the actual model implementation while the administrative level 4 data was held in reserve for model assessment purposes.

The modeling process uses a suite of continuous and discrete data layers to generate an estimated population density weighting layer. The majority of these data sets are contemporary and freely available (Table 5). The rationale behind using the datasets detailed in Table 1 is to include geospatial data that may correlate with human population presence on the landscape as cited in Stevens et al, 2015.

Table 5: Test-specific data sources and variable names used for population density estimation used for dasymetric weights for Chapter Three's Experiment

Type	Variable Name(s)*	Description	Indonesia Data
Census		Country-specific census data that is used for disaggregation	2010, Admin-level 3 and Admin-level 4 (census datasets received from the Government of Indonesia)
Land Cover	lan_cls011, lan_dst011	Cultivated terrestrial lands	Landcover utilizing 3-Year Google Earth Engine data & MDA GlobCover with methods from Patel et al. 2015 used.
	lan_cls040, lan_dst040	Woody / Trees	Ibid
	lan_cls130, lan_dst130	Shrubs	Ibid
	lan_cls140, lan_dst140	Herbaceous	Ibid
	lan_cls150, lan_dst150	Other terrestrial vegetation	Ibid
	lan_cls160, lan_dst160	Aquatic vegetation	Ibid
	lan_cls190, lan_dst190	Urban area	Ibid
	lan_cls200, lan_dst200	Bare areas	Ibid

	lan_cls210, lan_dst210	Water bodies	Ibid
	lan_cls230, lan_dst230	No data, cloud/shadow	Ibid
	lan_cls240, lan_dst240	Rural settlement	Ibid
	lan_cls250, lan_dst250	Industrial area	Ibid
	lan_clsBLT, lan_dstBLT	Built, merged urban/rural class	Ibid
Continuous Raster-Format			
	Lig	Lights at night data	Suomi VIIRS-Derived (NOAA, 2012)
	Npp	MODIS 17A3 2010 Estimated Net Primary Productivity, 1km	Extraction from MODIS package in R (Running et al., 2004)
	Tem	Mean temperature, 1950-2000	WorldClim/BioClim (Hijmans et al,, 2005)
	Pre	Mean precipitation, 1950-2000	WorldClim/BioClim (Hijmans et al,, 2005)
	Ele	Elevation	HydroSHEDS (Lehner et al., 2006)
	ele_slope	Slope	HydroSHEDS-Derived (Lehner et al., 2006)
	Twe	Tweets	Tweets data obtained from method detailed in Section 2.2
Converted Vector-Format			
	roa_cls, roa_dst	Roads	OSM (2014)

	riv_dst	Distance to rivers/streams	VMAP0 merged†
	pop_cls, pop_dst	Populated Places	OSM (2014)
	wat_cls, wat_dst	Water bodies	VMAP0 merged†
	pro_cls, pro_dst	Protected areas	IUCN and UNEP, 2012
	poi_cls, poi_dst	Populated Points of Interest	OSM (2014)
	bui_cls, bui_dst	Buildings	OSM (2014)
	use_cls, use_dst	Delineated Land Uses	OSM (2014)
	cit_cls, cit_dst	Cities	OSM (2014)
	dwe_cls, dwe_dst	Dwellings	OSM (2014)
	ham_cls, ham_dst	Hamlets	OSM (2014)
	hos_cls, hos_dst	Hospital	OSM (2014)
	loc_cls, loc_dst	Localities	OSM (2014)
	pol_cls, pol_dst	Police	OSM (2014)
	sch_cls, sch_dst	Schools	OSM (2014)
	sub_cls, sub_dst	Suburbs	OSM (2014)
	tow_cls, tow_dst	Towns	OSM (2014)
	vil_cls, vil_dst	Villages	OSM (2014)
	ind_cls, ind_dst	Industrial Land Use	OSM (2014)
	res_cls, res_dst	Residential Land Use	OSM (2014)
	pri_cls, pri_dst	Primary Roads	OSM (2014)
	sec_cls, sec_dst	Secondary Roads	OSM (2014)

	ter_cls, ter_dst	Tertiary Roads	OSM (2014)
	rro_cls, rro_dst	Residential Roads	OSM (2014)
	ser_cls, ser_dst	Service Roads	OSM (2014)
	nei_cls, nei_dst	Neighborhoods	OSM (2014)

* The variable names are used in Random Forest model output and throughout the text as reference to the specific data they were derived from. The first three letters are derived from the data type (e.g. "lan" indicates land cover) and the last three letters, if present, indicates what type of data each variable represents (e.g. "_cls" is a binary classification and "_dst" is a calculated Euclidean distance-to variable.
† The default data for populated places is merged from several VMAP0 data sources. There are three VMAP0 data sets used: The point data pop/builtupp and pop/mispopp are buffered to 100 m and merged with the pop/builtupa polygons creating a vector-based built layer. This layer is then converted to binary class and distance-to rasters for use in modeling. (NGA, 2005)

The MDA-land cover data was modified with the inclusion of a classified urban/rural land cover informed by an urban extent delineation using the Google Earth Engine platform (Patel et al. 2015). The resulting infusion of the binary urban/rural raster layer represents an improved "built" class within the MDA land cover classes.

Each covariate layer contributes to a better understanding of landscape features across Indonesia, both natural and man-made, as each may relate to population densities. In addition, to assess the added-value of including spatially-explicit social media data as a covariate in the model the best available ancillary data were combined with the geotweets. To compare model output and accuracy we create one set of outputs with geotweets included and one set of outputs without geotweets, both using the administrative level 3 census data.

In addition to census data, MDA-derived land cover and additional covariates included those outlined in prior Random Forest-based WorldPop datasets (Stevens, et al.

2015). Raster-based covariates included HydroSheds-based digital elevation data (also converted to slope estimates), Suomi VIIRS-derived lights at night raster layer, MODIS-derived estimates of net primary productivity, WorldClim average temperature and precipitation data, and of course the custom geo-located tweets. Vector-based covariates, which are then processed to raster-based derived products (Stevens, et al. 2015) include Open Street Map derived datasets, NGA VMAP0 data and World Database of Protected Areas boundaries. These are outlined and cited in Table 5.

3.1.3.2 Random Forest population disaggregation method

The general process used for data preparation, modeling and validation for the population mapping is documented in Stevens et al., (2015). In brief, the aggregated population counts and the raster and vector layers shown in Table 1 are used to create a Random Forest-based model (Breiman, 2001), parameterized on census unit population densities to estimate population density using the ancillary covariates. The Random Forest (RF) algorithm, as a nonparametric, ensemble statistical approach, provides flexibility in the modeling process for inclusion of disparate data types (Breiman, 2001). The process involves growing a “forest” by generating individual, unpruned decision trees that are then aggregated to represent a final prediction for each grid cell in the weighting layer (Breiman, 2001; Liaw & Wiener, 2002). The resulting population density map is then used as a weighting layer for a standard dasymetric mapping approach as described for WorldPop population map products (Stevens, et al. 2015; Gaughan et al. 2014; Gaughan et al., 2013; Linard et al., 2012; Linard & Tatem, 2012; Tatem et al., 2007). This process is depicted in Figure 3 of Chapter 2.

The RF model includes an internal cross-validation component that provides additional insight into the prediction error of the model. During the estimation of the random forest, at each node of each tree, one-third of the data is held in reserve from the iterative, bootstrapping process and used to generate an out-of-bag (OOB) error. The OOB error provides an unbiased estimate of prediction error for new, non-reference data points (assuming those points contain covariate data that fall within combinations present in the training data). Another metric that is provided post-hoc from the forest growing algorithm is the variable importance measures for each model and are presented as the mean decrease in residual sum of squares OOB estimates when the variable is included in the tree split.

3.1.3.3 Accuracy Assessment

The population mapping was undertaken using administrative level three census data as input, then the output high resolution map was aggregated at administrative level four and compared to the counts at this level, following Gaughan et al. (2013) and Stevens et al. (2015). Summary statistics were calculated, including root mean square error (RMSE), the RMSE divided by the mean census unit count (%RMSE) and the mean absolute error (MAE). Together these statistics were used to compare the predictive ability of the mapping with and without the geotweets.

3.2 Results

3.2.1 Random Forest statistical outputs

Figure 8 shows the importance of the variables outlined in Table 5 in the mapping process as estimated by the increase in mean squared error (MSE) when the specified covariate is randomly permuted and predictions re-calculated for OOB data. The most important variables include the built land cover covariates, indicating “Built” areas, which include urban and rural settlements that were created in the processes detailed in 3.1.3.2. The built covariates as well as the Suomi NPP Lights-At-Night-derived covariate have been documented in previous literature as strong indicators of population (Patel et al. 2015). Here, the Lights-At-Night and Distance to Villages (derived from Open Street Map data) variables are the most important predictors in the model without geotweets. When the geotweets are included in the modelling process (Figure 9b), differing covariates become key contributors, and the geotweets density variable enters into the top three most important predictors. In comparing the performance of both models, the test without tweets could explain 93% of the variance within the RF model, and test with tweets could explain 94% variance within its RF model.

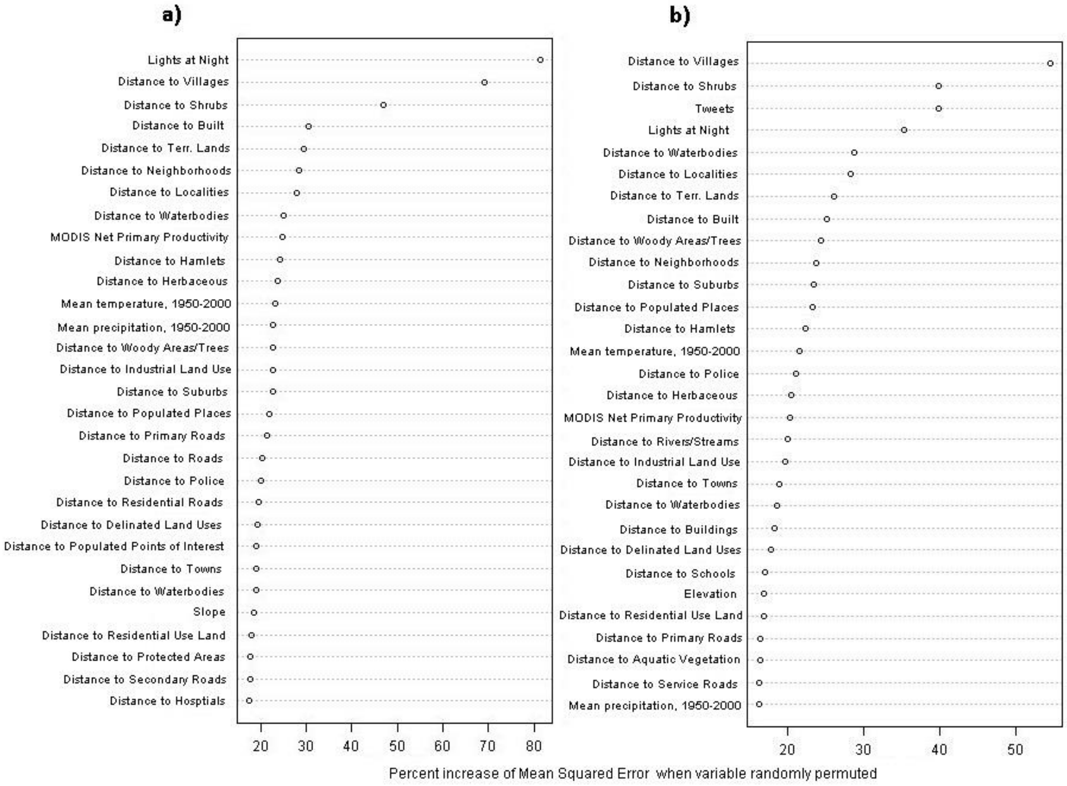


Figure 8: Covariate importance plots for tests a) without geotweets and b) with geotweets.

Figure 9 shows visual examples of the mapping without (Fig 9a) and with (Fig 9b, 9c) the inclusion of the geotweet density covariate. Visually the maps are very different, with the geo-located tweet inclusion resulting in a more constrained and higher density mapping of population density, specifically clustered around settlements and transportation networks.

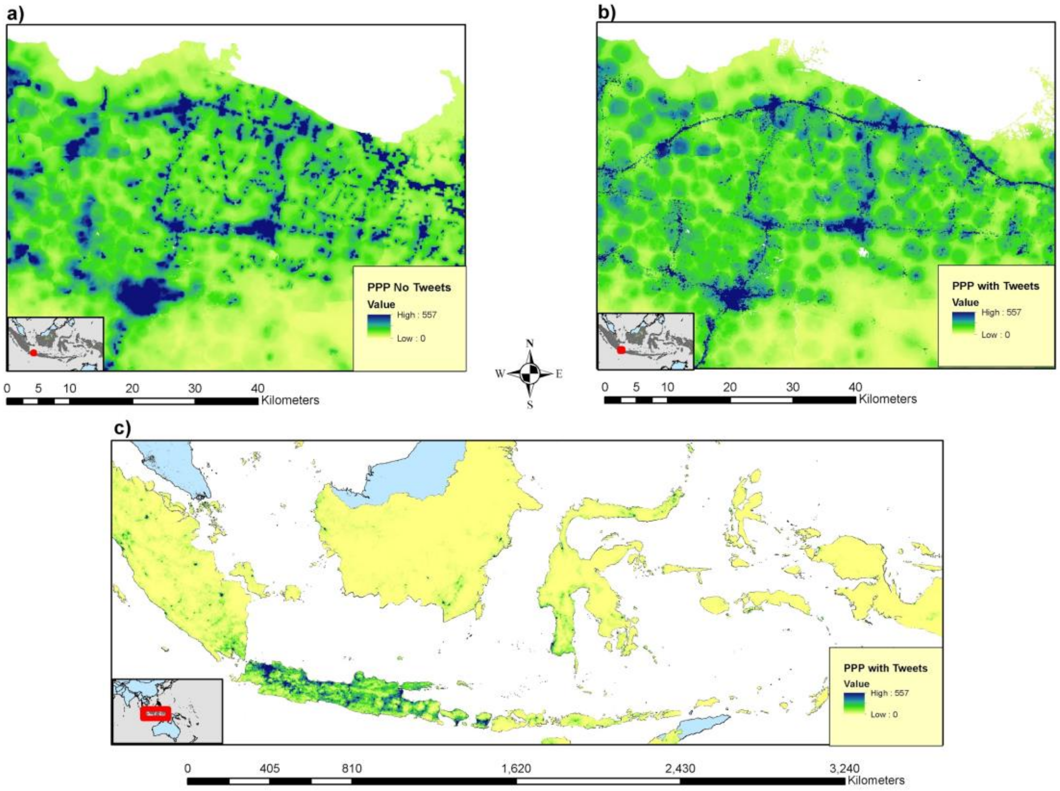


Figure 9: Map of Persons Per Pixel (PPP) produced using high resolution population mapping method (Stevens, et al. 2015), showing the final population maps for a region on the island of Java with a) no geotweet data and, b) geotweet data included, and finally c) the output dataset for the entirety of Indonesia with geotweet data included.

3.2.2 External accuracy assessment

Table 6 presents the results of the comparison of administrative level 3 census data based mapping with and without geotweets against administrative level 4 census counts. The inclusion of the geotweet data as a covariate produced a significant reduction in estimation error for both the root mean square error (RMSE) and mean absolute error (MAE) when considering the sheer difference in this particular comparison on the amount of census units in administrative level 3 versus administrative level 4. The results in Table 6 indicate that the inclusion of geotweets reduced errors relative to the model without the geotweets data.

Table 6: Accuracy Assessment Results for tests

	RMSE (persons)	%RMSE	MAE (persons)
Admin 3 without tweets	2284.14	74.58	1123.44
Admin 3 with tweets	2213.99	72.29	1120.16
Difference (Without - With)	70.15	2.29	3.28

Figure 10 shows the results from comparing the geotweet and non-geotweet population maps constructed using administrative unit level 3 population count data and applying zonal statistics to see how they compare with the finer administrative level 4 counts. It is evident that both datasets produced using administrative level 3 data result in some overestimations and some underestimations of population counts when assessed at administrative level 4. Unsurprisingly, the biggest differences are in Jakarta, where population totals are larger and vary more over shorter distances. Figure 10 does show however, that generally lower levels of over and under-estimation occur using the geotweet model.

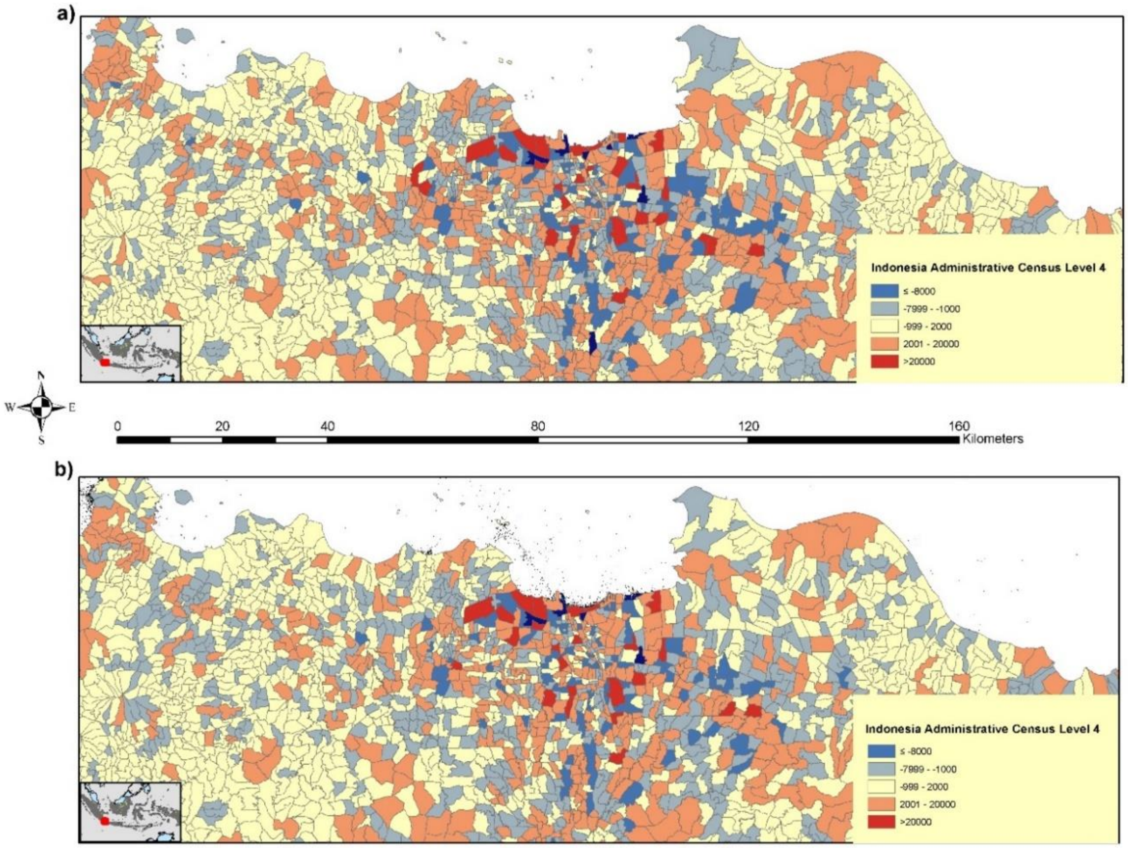


Figure 10 Differenced map produced through comparing the population maps generated with a.) no geotweet data and b.) geotweet data constructed from administrative level 3 census population counts and differencing the zonal sums against administrative level 4 census population count data.

Figure 11 compares the final output population models (with and without geotweets) spatially, subtracting the geotweets model from the non-tweets model to illustrate spatial patterns in differences. In addition to producing a more accurate model (Table 6, Figure 10), the figure highlights how the geotweets model concentrates populations into settlements more tightly, with less spread into more rural areas.

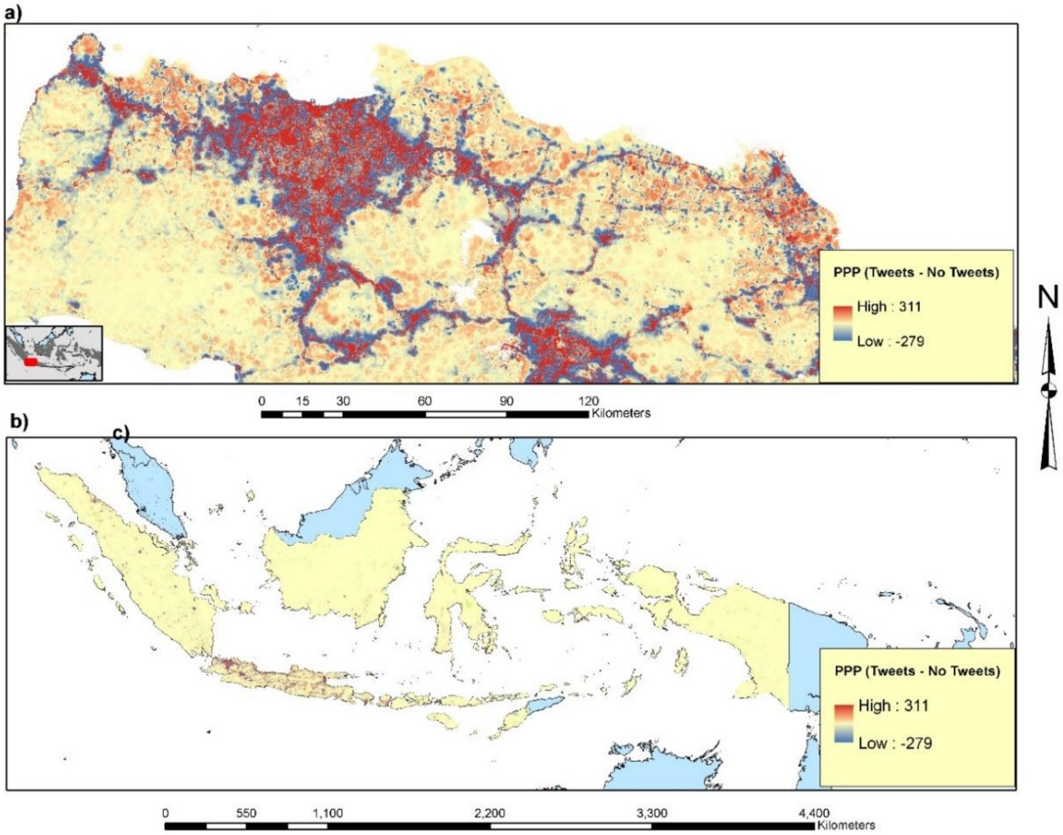


Figure 11 Difference map of persons per pixel (PPP) generated from subtracting the population map generated utilizing no geotweet data from the population map generated utilizing geotweet data. a.) Jakarta and surrounding areas; b.) All of Indonesia.

3.3 Discussion and Conclusions:

Spatially-disaggregated gridded population distribution datasets are becoming widely used, due principally to their flexibility in integration with other spatial datasets and summarization to any chosen level of aggregation. The accuracy with which this disaggregation can be achieved is related to the resolution and age of the input census data, but also to the quality, resolution and relevance of the spatial covariate layers used to statistically aid the disaggregation. The covariate layers typically used are often static in nature and their relationship to population densities can be unclear or interact in nonlinear ways. Despite these difficulties, high mapping accuracies can be achieved with suites of these static covariate layers (e.g. Stevens et al., 2015), but variance often still remains to be explained. In this experiment, it has been shown that data from social media, representative of physical locations of people in space, represents a potential addition to these covariate options that can provide improvements in population mapping accuracies.

The results shown in Tables 6 and Figure 8 demonstrate that the use of geotweet densities result in quantitative improvements in population mapping accuracies. Moreover, Figure 8 emphasizes that the geotweet density covariate was particularly important (third out of 30 covariates retained in the model) in contributing to the variance explained in the Random Forest models. With the number of Twitter users continuing to rise across the world, and the percentage of tweets that are geo-located also rising as smartphones continue to proliferate, the results underline the potential of this data source in contributing to the improvement of population mapping and its dynamic update (Leetaru et al., 2013). Furthermore, other sources of social media data, some country specific like Baidu (China), Instagram, Shutterfly, and others also offer potential when the data is not only geospatially referenced but made available for research such as this.

While the results presented make a strong case for the integration of geotweet densities in improving population mapping accuracies, there are a number of caveats and drawbacks that should be addressed. First, Indonesia has one of the highest Twitter user rates in the World, making it an ideal setting for this test analysis (Leetaru et al., 2013).

However, it remains unclear if similar results would be found elsewhere, particularly in areas such as sub-Saharan Africa, where Twitter usage and smartphone penetration levels are much lower. Within Indonesia, there may also be geographical differences in mapping accuracies. Mapping improvements were only assessed by the analysis at the national level and sub-national assessments may show areas of poorer mapping accuracies where Twitter usage levels are low. Furthermore, population densities in final population maps are highly clustered around transportation networks, and potentially biased in this regard due to people using social media while in transport more frequently than when at home. Moreover, even though Twitter users share their exact location at the time of their tweets, the default spatial granularity of their tweets is set at "Neighborhood" level which is a geographic boundary defined by Twitter[12]. As such, a high level of concentration of geotagged tweets with exact latitude and longitude was observed at aggregated points around city neighborhoods (Wu et al., 2015; Wu et al. 2015). How this aggregation process by neighborhood affects fine scale population mapping still needs further assessment. Additionally, the impacts of demographic biases in Twitter account holders (e.g. they may represent younger segments of the population) remain unclear and warrant further exploration. However, the results overall showed that the geotweets made a positive contribution to mapping accuracies despite these caveats. Another factor that requires further exploration is the timing of the tweets. Here, tweets from all times of day were aggregated, representing a kind of 'ambient' population

[12] https://dev.twitter.com/overview/terms/geo-developer-guidelines

distribution picture, which may not have been as representative of the residential population data from the census against which the outputs were used as a predictor. Further work should examine whether evening or nighttime-only data provides a better representation of residential population, as has been shown previously for nighttime versus daytime mobile phone call densities compared to census counts (Deville et al, 2014). Additionally, including sources of biasing due to age, income, time of day, smartphone availability, cost, and usage habits would make the results more informative and allow for exploration on how twitter data can be used in more diverse scenarios (Ramaswamy et al., 2009). Additionally, novel, open-source social media applications are providing information that can be interpolated with population maps to generate better insights on the basic needs of individuals that exist within the gridded population counts.[13]

Future work will continue to examine the potential of geotweets in combination with other spatial datasets for improving population mapping. In particular, the integration of such data with mobile phone call and cell tower records offers potential for improving dynamic population mapping, and this will be explored for the 15+ low/middle income country call data record datasets being analyzed by the Flow minder Foundation (www.flowminder.org). Further, different methods of measuring and analyzing the geotweets will be undertaken, from varying time periods of capture, to differing spatial windows of aggregation. Integration with upcoming high resolution human settlement datasets will also be explored, including the Global Human Settlement

[13] www.voicelots.com

Layer (“Joint Research Center - Global Human Settlement Layer”, 2015) and the Global Urban Footprint (“Global Urban Footprint”, 2015).

With the rise of smartphones and social media, the world population is transmitting more data on its presence and activities than ever before. Such data are often highly biased and incomplete, but nevertheless, this work done in this Chapter and in Patel, et al. (2016) has shown its potential in improving our understanding of human population distributions.

CHAPTER FOUR: EXPLORING OPPORTUNITIES AND CHALLENGES IN MAPPING POPULATION AT HIGHER-CONTINOUS RESOLUTIONS

This chapter will explore what opportunities and challenges exist in mapping populations at higher continuous resolutions (sub ~100 meter) on a country-by-country proceeding to global scale. Currently, the state of the art in mapping at 30 meter continuous population grids has been pioneered by the Space Informatics Lab within the Department of Geography at the University of Cincinnati. In the first section, methodologies used in that work will be described and commented on relative to the open-source dasymetric open-source population mapping method utilized thus far in this dissertation as described in Stevens et al. (2015).

In the second section, the open-source high resolution population mapping method described in Stevens et al. (2015), that is used in previous chapters will be proposed to utilize NASA’s high-end supercomputing resources with classified Landsat land cover at 30 meter resolution and mapping gridded population at a 30 meter gridded resolution, which has not been previously attempted before with this population disaggregation method. The perceived impact of the expected results will be discussed within this section.

In the third section, a brief proposed method describing an unmanned aerial vehicle (UAV) as a remote sensing verification tool will be explored, for the purposes of validating remotely sensed data captured and analyzed in close to real time.

In the fourth section, the chapter will be briefly summarized.

4.1 Motivations to map human population at 30 meter continuous resolution

Within the United States, the most authoritative source of population data is the government managed national census, in that U.S. population data is collected every 10 years by the U.S. Census Bureau. The census obtains population data on the resolution of an individual household, but it releases the data as fixed areal units (i.e. census blocks) in order to respect privacy. The census block represents the smallest areal unit with aggregated information. Within urban areas, census blocks can be the size of city block, where as in suburban and rural areas, they can be much larger. Population data aggregated to fixed administrative units does not properly inform population density as described in Chapter One, but more specific examples will be explored here.

In Dmowska & Stepinski (2017), it is aptly noted that aggregating administrative units to fixed administrative units, allows aggregations to suffer from the modifiable areal unit problem (Lloyd, 2014). The spatial data of the aggregated data is also very variable and low, with exception to the most densely populated urban areas. Also, there are spatial inconsistencies between user-desired units (e.g. neighborhoods, tax zones, postal delivery zones, vegetation zones, watersheds, etc.) (Voss et al., 1999). Additionally, the boundaries of census aggregation units, especially in the form of blocks, sometimes changes from one census to another, which makes the analysis of population change at high spatial resolutions very difficult (Holt et al., 2004; Schroeder, 2007; Ruther et al., 2015).

These challenges have made aggregation unit-based data inappropriate for spatial analysis of population related issues, in socio-economic and environmental issues. The population grid has been better suited as a format to deliver such population data. Population grids are constructed from census unit-based data utilizing either areal weighting interpolation (Goodchild & Lam, 1980; Flowerdew & Green, 1992; Goodchild et al., 1993) or dasymetric modeling (Wright, 1936; Langford & Unwin, 1994; Eicher & Brewer, 2001). As defined by Dmowska & Stepinski (2017), a population grid is a geographically referenced lattice of square cells with each cell carrying a population count or the value of the population density at its location. They also describe that population grids have significant advantages, in that all cells must be the same size, the cells must be stable in time, and there cannot be no spatial mismatch problem, as any partition of the study area must be rasterized to be co-registered with a particular population grid. Dasymetric modeling, when utilized for gridded population mapping even offers a superior resolution than that of the aggregated unit-based data.

Dasymetric modeling can be described as the technique of disaggregating unit-based population data into grid cells of a high spatial resolution by using ancillary data that can correlate with the population density but must have a higher spatial resolution. In Petrov (2012), dasymetric modeling has been extensively studied, with focusing in on utilization of different types of ancillary data to increase the accuracy of a given model. Land cover/land use data remain the most original form and most widely used ancillary datasets (Wright, 1936; Mennis, 2003, 2009; Linard et al., 2011). Recent studies have utilized high-resolution satellite images also as ancillary data in order to identify individual buildings (Ural el al., 2011; Lu et al, 2010; Lung et al., 2013). Often, regression analyses are used to link the area or volume of each building to the amount of people within it. Light Detection and Ranging (LiDAR) data was used by Lu et al. (2010) to help establish the volume of a building, for example.

Other dasymetric models sometimes use local infrastructure information such as density of points of interest (Bakillah et al., 2014), street density (Reibel & Bufalino, 2005; Su et al, 2010), tax parcel data (Maantay et al., 2007; Kar & Hodgson, 2012; Mitsova et al., 2012; Jia et al., 2014; Jia & Gaughan, 2016) in order to disaggregate census data. Currently proposed sources of ancillary data include address datasets (Zandbergen, 2011) and light emission data (Briggs et al. 2007; Sridharan & Qiu, 2013).

Population mapping within the United States has had the luxury of being able to utilize high resolution census aggregation data that makes the dasymetric mapping process much more accurate. In the following section, resolution and access to population grids will be discussed globally.

The adoption of demographic data for spatial analysis has been limited in recent times because the majority of potential uses are only able to utilize the ready-to-use product (a population grid) rather than be able to create their own. Recently projects have been released to the public domain in order to increase the adoption of demographic data for spatial analysis. Grids have been developed for all countries in the European Union (Gallego, 2010; Gallego et al., 2011), countries in South and Central America, Asia, and Africa (Gaughan et al., 2013, Linard et al. 2012; Sorichetta et al., 2015), and the United

States through the Socioeconomic Data and Application Center (SEDAC). SEDAC provides demographic grids at 1 kilometer resolution and 250 meters for selected metropolitan areas. However, these grids are only available for the years 1990 and 2000. High resolution (90m) US-wide demographic grids are under-development with the most recent census data by the Oak Ridge National Laboratory (ORNL) (Bhaduri et al., 2007). The product, called LandScan-USA, aims to provide daytime and nighttime (residential) population densities but is not currently available and is not expected be in the public domain once made available.

The work of Dmowska and Stepinski's group at the University of Cincinnati has aimed since 2014 to develop high resolution demographic grids for the entire conterminous US. Their group's main goals aim to develop grids that provide significant improvements over SEDAC grids and make them available for exploration and download through their web-based application called SocScape (Social Landscape) at http://sil.uc.edu. The group's first iteration of grids was called SocScape-90, referring to sharpening SEDAC grids to 90m resolution with using dasymetric modeling with the National Land Cover Dataset (NLCD) as ancillary data (Dmowska & Stepinski, 2014). SocScape was thus released with those methods for 1990 and 2000. Dmowska & Stepinski (2017) note that their original approach had several limitations and drawbacks. The previous approach of their mapping did not use original census data and instead utilized the SEDAC grid, which not only spatially coarser than census blocks in densely populated urban areas, but it also contained a number of errors and inconsistencies (Dmowska & Stepinski, 2014). It was also limited to the years 1990 and 2000, the only years for which SEDAC published these grids.

In their newest work, Dmowska & Stepinski (2017) present the second generation of U.S wide-grids which they call SocScape-30. Their new approach uses dasymetric modeling to disaggregate census blocks directly, rather than disaggregating SEDAC cells and uses two ancillary datasets, the NLCD 2011 and the newly available National Land Use Dataset (Homer et al., 2004; Theobald, 2014). The SocScape-30 grid has a nominal resolution of 30 meters, which is equal to the resolution of both ancillary datasets. An assessment of uncertainty is assessed only if the ground truth is available, and in this particular context the ground truth data would consist of certifiable population counts within aggregation units smaller than those use in the dasymetric model. In the case of the study, they use a similar accuracy assessment method by calculating an additional grid based on the disaggregation of larger units, census block groups, and compare the population of the resultant grid that is aggregated to blocks with the population of the blocks as given by the census. The method is used by the WorldPop Project as featured in Jia et al. (2014) and Stevens et al. (2015).

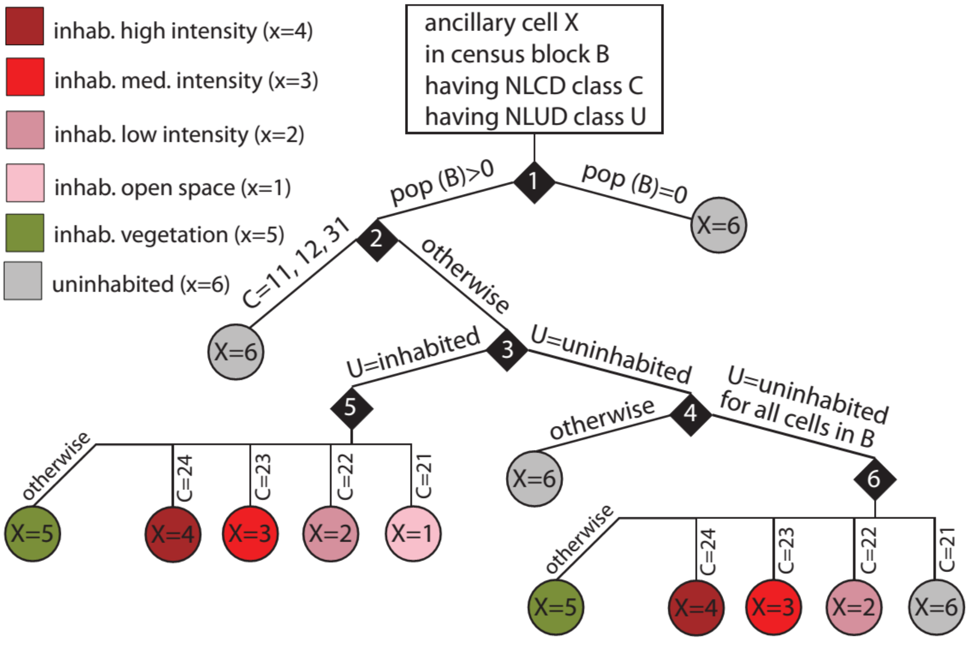


Figure 12: Decision tree utilized by Dmowska and Stepinski 2017 showing the process of assigning a cell's ancillary class on the basis of its NLCD and NLUD classes and the population count in the block to which it belongs.

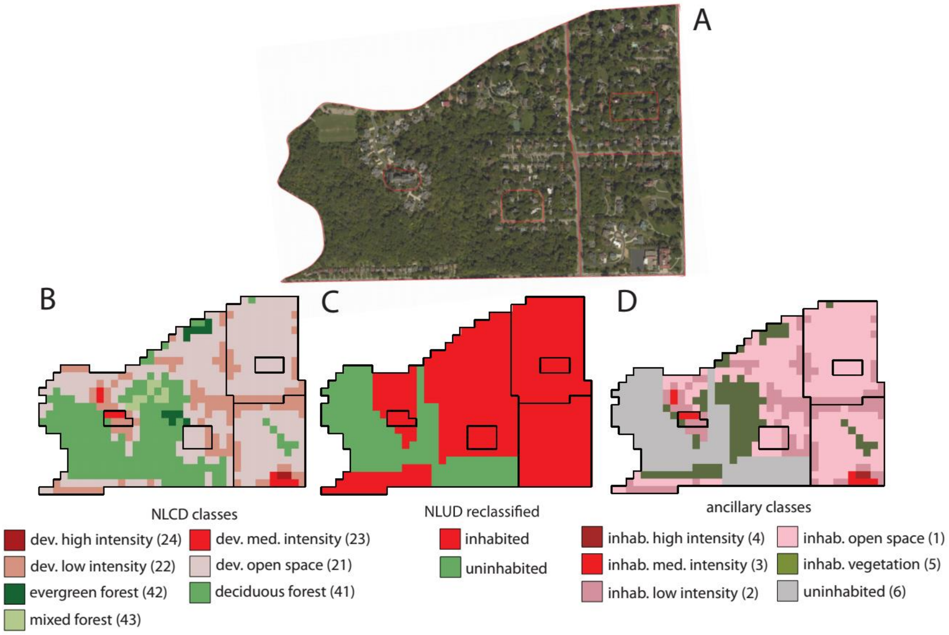


Figure 13: In Dmowska and Stepinski 2017, construction of the ancillary layer using area consisting of six adjacent

blocks in Cincinnati, OH as an example. (A) A satellite image showing the surface masked to the spatial extent of the area; six constituent blocks are indicated by orange lines. (B) Spatial distribution of NLCD classes over the area. (C) Spatial distribution of reclassified NLUD classes over the area. (D) Spatial distribution of final ancillary classes over the area.

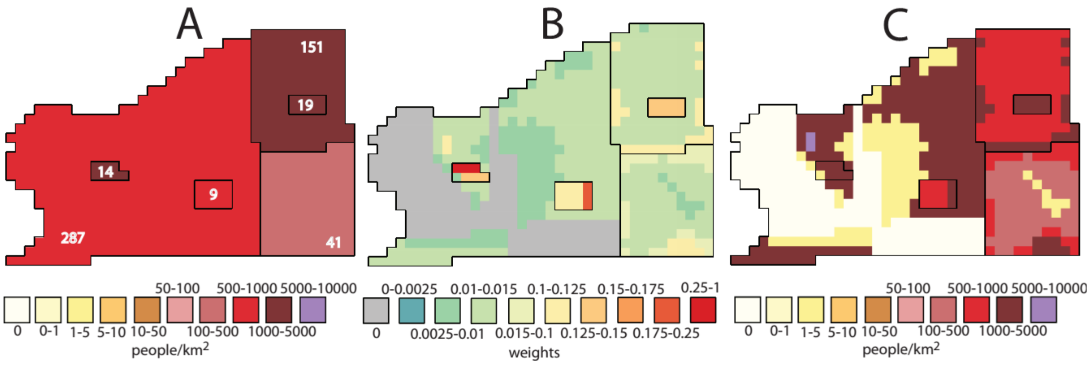


Figure 14: From Dmowska and Stepinski 2017: Demonstration of dasymetric modeling using an area consisting of six adjacent blocks in Cincinnati, OH as an example. (A) Map of population density using only block-level data. Numbers indicate population counts for each block. (B) Spatial distribution of weight values. (C) Map of population density using grid data calculated using a dasymetric model.

Figures 12, 13 and 14 visually explain and show outputs of the population disaggregation methods employed by Dmowska & Stepinski (2017). The authors employed an accuracy assessment strategy similar to what was used in Jia et al., 2014, and compared the nationally produced results versus the results produced by the Alachua County, Florida, results that Jia et al., 2014 produced and enjoyed better accuracy statistics when comparing root mean squared error (RMSE).

This type of dasymetric modeling at this accuracy and resolution is made possible by the sheer amount of funding and resources devoted to projects like the National Land Cover Dataset and the National Land Use Dataset (Homer et al., 2004), as well the luxury of having the lowest average spatial resolution (ASR) census units in the world due to the U.S Census Bureau (Balk & Yetman, 2004). It is important to understand what the current state-of-art methods are being used for simple dasymetric mapping to understand motivations behind applying supervised machine learning at the same resolution. The next section will outline a novel test of applying a supervised machine learning method (Stevens et al., 2015) to perform census disaggregation at 30 meter resolution utilizing high-end computing resources, currently in progress.

4.2 Dasymetric Mapping of Human Populations utilizing supervised machine learning on High-End Computing Capability Resources

In this section, the open-source high resolution population mapping method described in Stevens et al. (2015), that is used in previous chapters will be proposed to utilize NASA's high-end supercomputing resources with classified Landsat land cover at 30 meter resolution and mapping gridded population at a 30 meter gridded resolution, which has not been previously attempted before with this population disaggregation method. The experiment proposed here will be framed around the need to analyze human population distributions with respect to scaling land-use/land cover data. Studying the scales at which land surface is characterized in efforts to estimate carbon flux, investigating scaling effects on measurements, modeling urbanization effects on ecosystems, and developing an interactive modeling environment that can make the datasets easily utilized and downloadable, is increasingly important.

4.2.1 Methods

4.2.1.1 Study Area

The study area is the island of Java, within the country of Indonesia. Section 2.2.1 also has a short description of Java and its importance to the country of Indonesia. For the purposes of this experiment, boundary-matched census data at the Desa/Kelurahan administrative level (Level 4, 26,992 units), were obtained (Figure 15).

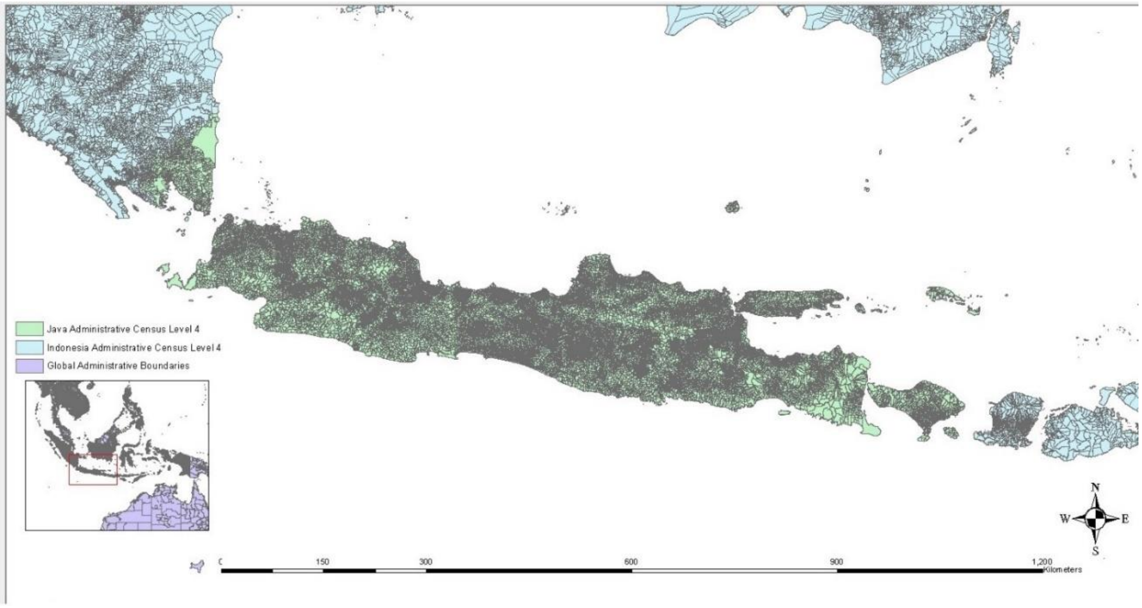

Java Administrative Census Level 4
Indonesia Administrative Census Level 4
Global Administrative Boundaries
N
W
E
S
0
150
300
600
900
1,200
Kilometers

Figure 15: Map of the Indonesian island of Java, at administrative boundaries level 4, focused around Jakarta.

4.2.1.2 Data Management Considerations

For the purposes of this experiment, two tests were conducted in order to examine the differences in the statistical outputs of random forest, both using the same study area, with one mapping population at 100 meter grid cells and the other mapping population at 30 meter grid cells. Both tests utilized the same set of covariate datasets that will be detailed in the next section, 4.2.1.3. However, two different computing environments needed to be set up in order to conduct this experiment, they are detailed further in

Appendix III. A personal computer environment was utilized with higher end components to run the dasymetric mapping at 100 meter resolution. To run the 30 meter dasymetric mapping process, an allocation of 96 cores on a Linux-based environment within NASA's Pleiades super-computer was allocated through the NASA Earth Exchange project, and that work is currently in progress. This collaborative research framework was initiated by the NASA Earth Sciences program to combine state-of-the-art supercomputing, Earth system modelling, remote sensing data from NASA and other agencies, and a scientific social networking platform that provide a complete work environment (Nemani et al., 2011). Further detail is provided in Appendix III.

4.2.1.3 High Resolution Population Mapping Method

The same, open-source population mapping method utilized in Sections 2.1.3 and 3.1.3 was utilized with the exception of the accuracy assessment that is usually done, as the mapping conducted is focused on evaluating the statistical outputs of the Random Forest mapping process with the highest spatial resolution census dataset available at Administrative level Unit 4. Table 7 shows the covariate datasets that are proposed to be utilized in the mapping process, note that the same Open Street Map data from 2014 was utilized and the MDA Federal Inc. EarthSat Geocover land cover dataset was not modified as it was in Chapter Two. The key difference is that if you refer to Figure 3, within Section 2.1.3, the mapping processes in this experiment scale down the covariate datasets even further to 30 meter resolution in the pre-processing steps and the in gridded population mapping steps.

Table 7: Test-specific data sources (for 100m and 30m tests) and variable names used for population density estimation used for dasymetric weights for proposed experiment

Type	Variable Name(s)*	Description	Indonesia Data
Census		Country-specific census data that is used for disaggregation	2010, Admin-level 4 for just the extent of Java (census datasets received from the Government of Indonesia)
Land Cover	lan_cls011, lan_dst011	Cultivated terrestrial lands	MDA Federal Inc., 2007
	lan_cls040, lan_dst040	Woody / Trees	Ibid
	lan_cls130, lan_dst130	Shrubs	Ibid
	lan_cls140, lan_dst140	Herbaceous	Ibid
	lan_cls150, lan_dst150	Other terrestrial vegetation	Ibid
	lan_cls160, lan_dst160	Aquatic vegetation	Ibid
	lan_cls190, lan_dst190	Urban area	Ibid
	lan_cls200, lan_dst200	Bare areas	Ibid
	lan_cls210, lan_dst210	Water bodies	Ibid
	lan_cls230, lan_dst230	No data, cloud/shadow	Ibid
	lan_cls240, lan_dst240	Rural settlement	Ibid

	lan_cls250, lan_dst250	Industrial area	Ibid
	lan_clsBLT, lan_dstBLT	Built, merged urban/rural class	Ibid
Continuous Raster-Format			
	Lig	Lights at night data	Suomi VIIRS-Derived (NOAA, 2012)
	Npp	MODIS 17A3 2010 Estimated Net Primary Productivity, 1km	Extraction from MODIS package in R (Running et al., 2004)
	Tem	Mean temperature, 1950-2000	WorldClim/BioClim (Hijmans et al,, 2005)
	Pre	Mean precipitation, 1950-2000	WorldClim/BioClim (Hijmans et al,, 2005)
	Ele	Elevation	HydroSHEDS (Lehner et al., 2006)
	ele_slope	Slope	HydroSHEDS-Derived (Lehner et al., 2006)
	Twe	Tweets	Tweets data obtained from method detailed in Section 2.2
Converted Vector-Format			
	roa_cls, roa_dst	Roads	OSM (2014)
	riv_dst	Distance to rivers/streams	VMAP0 merged†
	pop_cls, pop_dst	Populated Places	OSM (2014)
	wat_cls, wat_dst	Water bodies	VMAP0 merged†
	pro_cls, pro_dst	Protected areas	IUCN and UNEP, 2012

	poi_cls, poi_dst	Populated Points of Interest	OSM (2014)
	bui_cls, bui_dst	Buildings	OSM (2014)
	use_cls, use_dst	Delineated Land Uses	OSM (2014)
	cit_cls, cit_dst	Cities	OSM (2014)
	dwe_cls, dwe_dst	Dwellings	OSM (2014)
	ham_cls, ham_dst	Hamlets	OSM (2014)
	hos_cls, hos_dst	Hospital	OSM (2014)
	loc_cls, loc_dst	Localities	OSM (2014)
	pol_cls, pol_dst	Police	OSM (2014)
	sch_cls, sch_dst	Schools	OSM (2014)
	sub_cls, sub_dst	Suburbs	OSM (2014)
	tow_cls, tow_dst	Towns	OSM (2014)
	vil_cls, vil_dst	Villages	OSM (2014)
	ind_cls, ind_dst	Industrial Land Use	OSM (2014)
	res_cls, res_dst	Residential Land Use	OSM (2014)
	pri_cls, pri_dst	Primary Roads	OSM (2014)
	sec_cls, sec_dst	Secondary Roads	OSM (2014)
	ter_cls, ter_dst	Tertiary Roads	OSM (2014)
	rro_cls, rro_dst	Residential Roads	OSM (2014)
	ser_cls, ser_dst	Service Roads	OSM (2014)
	nei_cls, nei_dst	Neighborhoods	OSM (2014)

	ghs_dst	Distance to Global Human Settlement Layer polygons	Joint Research Center - Global Human Settlement Layer (2015)

* The variable names are used in Random Forest model output and throughout the text as reference to the specific data they were derived from. The first three letters are derived from the data type (e.g. “lan” indicates land cover) and the last three letters, if present, indicates what type of data each variable represents (e.g. “_cls” is a binary classification and “_dst” is a calculated Euclidean distance-to variable.
† The default data for populated places is merged from several VMAP0 data sources. There are three VMAP0 data sets used: The point data pop/builtupp and pop/mispopp are buffered to 100 m and merged with the pop/builtupa polygons creating a vector-based built layer. This layer is then converted to binary class and distance-to rasters for use in modeling. (NGA, 2005)

4.2.2 Proposed Research

4.2.2.1 Statement of the Problem

The scientific problem that is to be addressed with this initial experiment in 30 meter population mapping is analyzing human population distribution (and associated growth) with respect to the current spatial scales of global land-use/cover change. This research will be complementary to efforts estimating carbon flux, developing land surface characterization parameters for input to hydrologic modeling of surface flows, investigating scaling effects on measurements, modeling urbanization effects on ecosystems, and developing an interactive modeling environment that can make the datasets easily utilized and downloadable. Methods to properly characterize the Earth’s land surface as it relates to global land-use/cover change is essential for global modeling of population distribution and growth.

As detailed in earlier chapters, human population distributions have been utilized extensively for planning interventions and monitoring changes. These datasets are also used for disease burden estimation, epidemic modeling, resource allocation, disaster management, accessibility modeling, transport and city planning, poverty mapping, and environmental impact assessment.

Studying humanity's specific impact as the one of the primary drivers for land surface change compliments NASA's strategic plan and vision for its research and applied science priorities of NASA Earth Science. The specific research end of understanding how human populations at a fine resolution scale impact the Earth system is inherently valuable for NASA's vision to communicate and transfer scientific knowledge, developing and deploying enabling technologies, and inspiring and motivating the nation's students and teachers. NASA's Earth Observing System (EOS) series of satellites offer an unmatched quality of data and modelling human population with the data from EOS would enable better understanding of the Earth System through observations and predictive models, the development and validation of new technologies, and also serve to educate and advance the scientific and technological capabilities of the nation, specifically in relation to evaluating impacts of climate change driven by humans.

The research topic would complement and further the work of contextualizing the work of the NASA EOS missions as well as NASA SERVIR (a joint venture between NASA and the U.S. Agency for International Development), NASA's Socioeconomic Data and Applications Center (hosted by CIESIN at Columbia University), NASA's Applied Remote Sensing Training (ARSET) Program, as well as integrating data from the Earth Observing System Data and Information System (EOSDIS) EarthData program on Human Dimensions.

4.2.2.2 Science Background/Motivation

The idea that human population growth would eventually exceed the capacity of the resources that are required to sustain it was furthered initially by Thomas Malthus, in 1798, in his essay on the principle of population (Malthus, 1798). Malthus's thesis has been explored by scientists with a focus on the impact that exponential population growth would have on the environment. The clear and predictable links between human population dynamics and environmental change have been hard to discern until recently, due to how complex human activities are and how they impacts nature in various ways (Ehrlich, 1970; Meadows et al., 1972; Cohen, 1995; Wilson, 1999; McKee, 2005; Luck, 2007).

As Meyer & Turner explain, human induced global environmental change can be grouped into two overlapping fields of study. The first is the industrial economy concerns the flow of materials and energy through the process of extraction, production, consumption and disposal of modern industrial society. Land-use/land cover change, mainly deals with the alteration of the land surface and its biotic cover (National Research Council, 1990). Environmental changes in either field of study can become the cause of a global change in two ways (Turner et al., 1990), by either affecting a globally fluid system (i.e. the atmosphere, world climate, sea level) or occurring in particular locations summing up to a globally significant total. Land-use change notably contributes to both kinds of global change, such as systemic changes like trace-gas accumulation and patchwork impacts such as biodiversity loss, soil degradation and hydrological changes

(Meyer & Turner, 1992).

Land-cover changes can be characterized as a conversion of one category of land cover to another and the modification of a condition within that particular category. The four main classes have been tracked in most global figures, include forest/woodland, permanent pasture, cultivation, and other lands (Bartholomé & Belward, 2005). The UN Food and Agriculture Organization (FAO) has been collecting this data from member states as a part of their Production Yearbooks since the 1950s, however the FAO class of “other lands” combines several different and distinct forms of land use and land cover.

The climate modelling literature has also been a good source of global land data. Matthews work on presumed pre-agricultural vegetation types and of present day land cover was a foundational study for high resolution mapping of land cover for climate studies (Matthews, 1983). Global carbon flux modelling as well has necessarily involved detailed reconstruction of change in land-cover patterns and evaluating a broad set of data sources such as archival materials and remote sensing. The work of Matthews led to Richards’ continental-scale reconstruction of land-use changes from 1700 to 1980 (with the categories including cultivation, forest/woodland, and grassland/pasture) (Richards, 1990). Historically and currently, issues of data quality and comparability make global-scale assessments of land-use/cover change difficult (Heilig, 1994; Ramankutty & Foley, 1999; Small, 2004; Erb et al., 2007).

Historically in the literature, changes in the major land cover categories has been tracked: cultivated land, forest/tree cover, grassland/pasture, wetlands, and settlement data. Settlement data has historically been valued in population mapping, as it represents areas devoted to human habitation, transportation, and industry (Anderson, 1976).

Land-cover change is associated with many secondary environmental consequences including wetland drainage (impacting biodiversity, trace gas emissions, soil, and hydrological balance), and these are often hard to distinguish from natural variation, with climatic change and water flows being cases in point (Sagan et al., 1979). Five overall classes of impacts that are directly tied to land-cover change have been identified historically in the literature, they include: trace-gas emissions, hydrological change (water quality and water flows), soil impacts, sediment impacts and climatic change.

The human driving forces of global change was considered to be in "intellectual disarray" by Meyer & Turner in 1992 and arguably still are in relative disarray. At the time of their assessment of the literature, either studies were weakly connecting causes to one another or extremely hypothetical (believable arguments that were not supported by case-specific data) (Mortimore, 1989). Ultra-empiricist arguments included those of Newell & Marcus, where they presented a high statistical correlation between world population growth and tropospheric carbon dioxide levels since the 1950s as proof of population's fundamental role (Newell & Marcus, 1987). Ultra-theoretic arguments in the vein of Harvey's work on attack the relevance of neo-Malthusian arguments, on the foundation that they are politically founded, i.e. the primacy of population growth in resource depletion (Harvey, 1979). In ultra-empiricist arguments, it is easy to mistake a correlation for a cause (or vice versa) within an extremely complex area of study, and for the ultra-theoretic arguments this is excessively narrowing the scope of the investigation without any consideration of the data.

Driving forces of population change sometimes vary with the type of change involved. Consider rising interest rates or agricultural prices, which would simultaneously increase deforestation because it would encourage further clearing, but at the same time, provide incentives to adopt soil conservation measures. Additionally, the same type of land-cover change can also have different sources in different areas even within particular world regions. Finally, in understanding the dynamics of underlying causes, there is no agreement existing on the level at which an adequate explanation is achieved. The example that Meyer & Turner give is considering deforestation by an agricultural expansion could be driven by population growth but can also be driven by certain sociopolitical and economic conditions that promote such conditions. For example, this occurring within an economic context where it would make sense for large families to be valuable to subsistence cultivators, while in other cases, emphasizing the role of agricultural expansion creating population growth.

In the literature, comprehensive approach to the question of driving forces is often characterized as the $I = PAT$ which has been used by Ehrlich & Ehrlich, 1990, and by Commoner (1972, 1990), as well as more recent works (Blaikie et al., 2014; Costanza et al., 2014; Moran, 2016). *I* represents environmental impact, which is the product of *P* (population), *A* (affluence), *T* (technology). In this equation, human impact is a product of the number of people, considering a level which they consume, and the character of the material and energy flows in production and consumption. This formula suffers from the relative mismatch of the categories of the driving forces with the exception of population. Arguably, the in more recent works, affluence and technology have more substantial social science theory that back behavior and social structure could be linked with production and consumption.

Because of its ease of quantification and its plausibility for being a driving force for environmental change, human population has been widely used in the literature. The metric incorporates the basic level of resources required per capita for survival and reproduction (biological demand). This role of population is not in dispute, what is controversial is understanding its relative importance with the other forces that generate environmental pressures.

There are various positions of considering population as a driving force of environmental change. There is the neo-Malthusian position, where population growth is considered to have exceeded the capacity of the biosphere. Diametrically opposed to that position is the cornucopian position, which holds that population increases allow for innovations in society and technology that can improve the conditions of life and improve the condition of the environment. The role of population is only lessened in theories where it is held to only cause further degradation stemming from other factors. For example, the Faustian position contends that the use of technology in a careless way is primary to environmental degradation, even though population increases could further increase the severity of the problems that are created. The theory that locates the cause of environmental damage in obstacles to the proper allocation of costs is neopolitical classical economy. The distortion of efficient solutions by government policies or property institutions would be where these obstacles originate from. Neo-marxist political economy emphasizes the roles of the means of production in the global economy of international capitalism, as profit-seeking and capital accumulation require the unsustainable exploitation of natural resources and socioeconomic differentiation causes a situation in which the “haves” place heavy demands on the world’s resources which drives environmental change, leaving the “have-nots” in environments under stress (Meyer & Turner, 1992).

Within any of these positions, the role that population is being given reflects less conflicting evidence than conflicting interpretations of the same evidence. There is no question however, that in the past 300 years, there has been an unparalleled magnitude of human-induced environmental changes, including those of land cover (Ramankutty & Foley, 1999). Correlations between population and land-cover studies have been strong when considering regions possessing similar socioenvironmental characteristics (Schmitz et al., 2014). It is thus not as simple to attribute “overpopulation” to sweeping land cover change like deforestation (Blaikie & Brookfield, 2015), and instead more complex array of policies, institutions, and economic forces must be considered when considering deforestation.

Meyer & Turner noted in 1992, that even though population is an important macro-scale (global) variable in that there is a variable for its direct relationship between total world population and total biological demand for resources. At the time, it was noticed that the connections to land-cover change become weaker at increasingly smaller spatial scales because of the importance of other variables that could affect demand or spatially deflect its particular impacts. It was their recommendation that these other variables must be incorporated to improve the understanding of the human causes of land-use/ land cover change (Meyer & Turner, 1992).

This literature which seeks statistical linkages without a well-developed theory also stress the importance of examining data only within a theoretical framework. The population-agriculture relationship has been widely studied, as associated with neoclassical economics, as a role of population change through the influence on demand as manifested through the market (Setälä et al., 2014). Modern population-pressure theory is sort of corollary to this, attributing agricultural development, and subsistence to market, to the pressures of production that amount from a growing population and operates through a process of intensification (Misra et al., 2014). Sustained population decline has the opposite effect in this theory. Population linkages tend to be spatially congruent with land use in economies with strong subsistence components as well. In supplementing the other candidate forces for the driving forces of land cover, they include arguments furthering technological change, socioeconomic organization, level of economic development and culture (Meyer & Turner, 1992).

Consider sections 1.2 and 4.1 in the literature review of dasymetric mapping processes and how they have relied on the quality of physical data (such as remote sensing) to create more precise and spatially congruent data when it comes to population grids. In order to have better land use characterization for global modelling, it can be contended that analyzing human population distribution data (and associated growth) with respect to the current spatial scales of global land-use/cover change, is an area of research that is necessary given the continual disconnect between physical data that can be gathered from technology such as remote sensing, and descriptions of how the land is actually being used in different regions of the world.

In recent years, substantial growth has been seen in openly available satellite and other geospatial data layers that represent a wide variety of metrics that are relevant to population mapping at fine spatial scales. Recently open-source population mapping projects have been able to harmonize these different data layers in constructing detailed and contemporary spatial datasets, that can accurate describe population distributions (Lloyd et al., 2017).

The work proposed here in mapping population at 30 meters and comparing against 100 meters would further the research in understanding how mapping accuracies shift at different spatial scales, and see how much the process of harmonizing the data would reveal inaccuracies. Additionally, given the access to human population distribution data and data tools to project settlement growth into the future (Nieves et al., 2017), more recent research needs to be conducted to identify how impacts can be assessed with the current spatial scales of land-use/ land cover change as measured by instruments such as NASA’s Earth Observing System. With the current computational resources available and greater access to more temporally relevant data, a more philosophically precise framework can be developed regarding the relationship between human population distribution and actual land-use/land cover.

4.3 Unmanned Aerial Vehicles (UAVs) as remote sensing land cover classification verification tools

The airborne method to improve an open-source population mapping method for the future includes Unmanned Aerial Vehicles, which will initially be used for land cover classification validation. This short section reflects on the work of Xia et al., 2017, which utilizes Landsat-8 OLI data and UAV-based data of a typical wetland region on the Zoige Plateau in China to compare the performance of linear spectral unmixing (LSU), regression tree (RT) and artificial neural networks (ANN) in estimating the wetland sub-pixel inundation percentage (SIP), and its applicability to providing a useful product for the topic of this dissertation.

Unmanned aerial vehicle (UAV) technology over the recent years offers new opportunities and use-cases to explore at the sub-meter image resolution. UAV technology is currently being used and is perceives as a powerful complementary platform to typical remote sensing platforms. UAV technology have been increasingly used in many different applications because of the diminutive size of the platform, lower costs, flexible implementation and the ease-of-use for the typical user. Conventional satellite imaging platforms also cannot ensure the purity of the image pixel in the way that high spatial resolution images from UAV are capable of recognizing the difference between different land cover classes. For example, manned aerial photography platforms or satellite imagery cannot detect small inundations, grass and soil patches (Rango, et al. 2006).

4.3.1 Overview of Methods Utilized for Image Capture from UAV

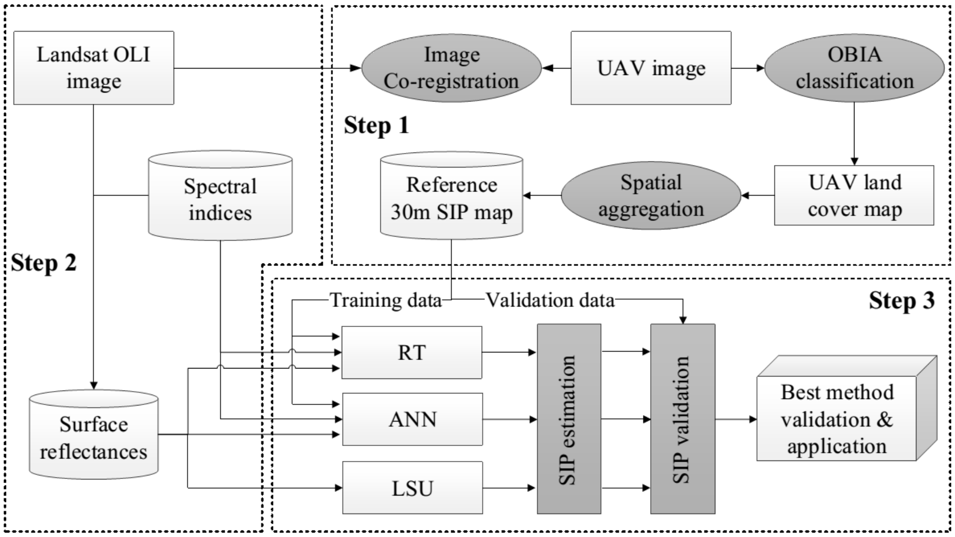


Figure 16: From Xia et al., (2017): Flowchart of the overall approach for mapping the subpixel inundation percentage (SIP), using data on the Landsat-8 Operational Land Imager (OLI) and unmanned aerial vehicle (UAV). The RT, ANN, and LSU are the abbreviations of regression tree, artificial neural networks, and linear spectral unmixing, respectively.

In Figure 16 the overall methodological workflow used by Xia et al., 2017 is shown. In the first step, the UAV image is prepared, preprocessed and classified by an object-based image analysis to derive the UAV land cover classification map. An OLI pixel is aggregated from the UAV land cover map, and from this the reference 30 m SIP map was created. Within the second step, the Landsat-8 OLI image’s available surface reflectance was pre-processed and the spectral indices were calculated for model preparation. The final step involves the UAV-based reference SIP for 2014 (in the case of this study), to train and evaluate three different SIP models (LSU, ANN, RT) to select the optimal model.

Within the Xia et al. (2017) research, a fixed-wing UAV (called Freebird) was utilized. The platform was chosen for its ability to resist wind in comparison to rotary wing based UAVs. The onboard sensor of the instrument is a Canon 5D Mark II, a digital camera that has three optical bands (red, green, blue). For the study, the height of the UAV was set to 800 meters or 4250 meters above sea level (a.s.l.) and was flown at ground level over the transect area in order to image at 0.16m (high spatial resolution). The instrument was set with an 80% forward lap and a 60% side lap so that there was no gap with nearby imaging. With the transect area being 7 km long and 1km wide, a total of 265 pictures were obtained.

The geometric/topographic corrections were implemented after the data collections process to generate a digital orthophoto map (DOM) by using MAP-AT software, automatically processing acquired imagery and altitude data using the UAV image telemetry (ground control points (GCPs), synchronized GPS positions, and roll, pitch and yaw of each image). Figure 17 shows a UAV image of the study area.

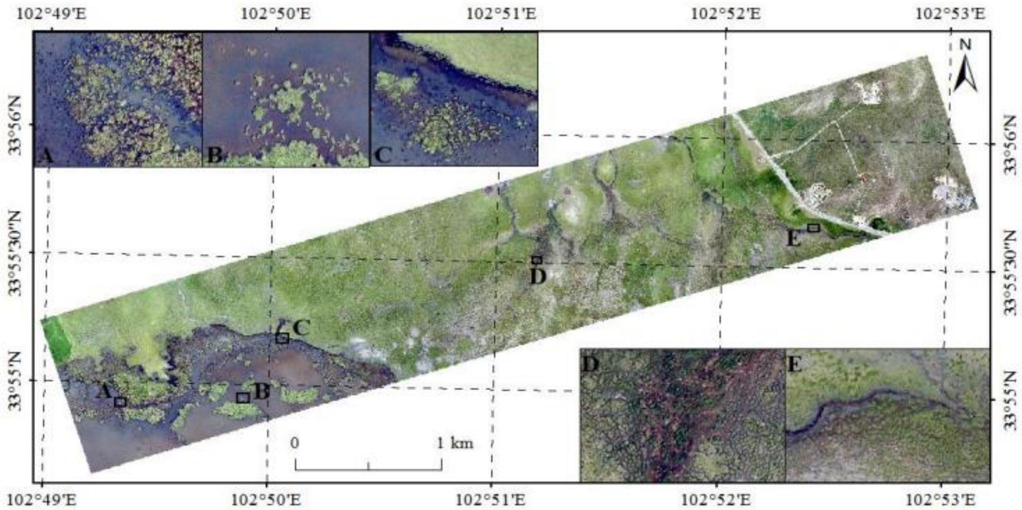


Figure 17: From Xia et al., (2017): The UAV image of the transect in Zoige wetland, acquired in July 2014. A and B are

the small patches of grass in the lake, C is the edge of the lake, and D and E show the area with mixed water and grass.

4.3.2 Object-Based Image Analysis (OBIA) for UAV Image Classification

Due to the high spatial resolution of the captured UAV data, Xia et al. (2017) defined the pixel to be a water surface or a non-water surface. Once this binary classification was complete, the SIP on the Landsat-8 OLI pixel scale was calculated. After these calculations are complete, an OBIA is applied to the UAV image classification, as the geometric and contextual features can be integrated into the classification (Laliberte et al., 2004; Ma et al., 2015). The OBIA approach further segments the UAV image into ecological patches, which is able to improve classification accuracy by combining with a decision tree model at the object level. Equation 2 shows the general form of the spatial aggregation used by Xia et al., 2017:

$$SIP_r = (\sum_{i=1}^{n} s_{w,i})/S_{oli}$$

Equation 2 can be defined as such. The left side of the equation represents the reference SIP for the 30 meter grids of the Landsat image, the right side numerator represents summation of the area of the water pixels for UAV (in meters squared), where n is the number of water pixels in the UAV classifications maps in the 30 meter grids of the Landsat image. The denominator represents the area of 30 meter grids of the Landsat-8 OLI image (in meters squared).

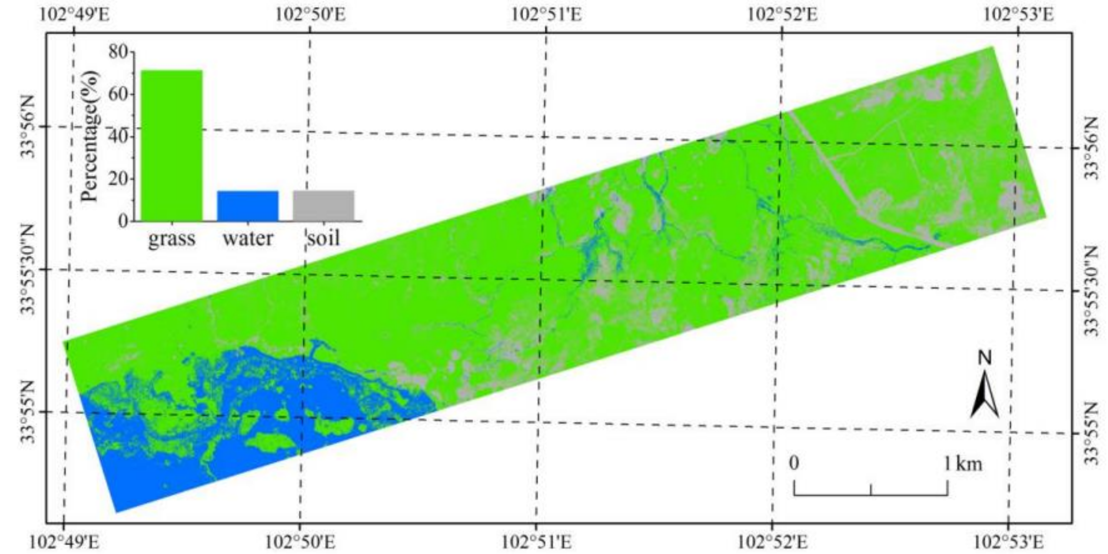
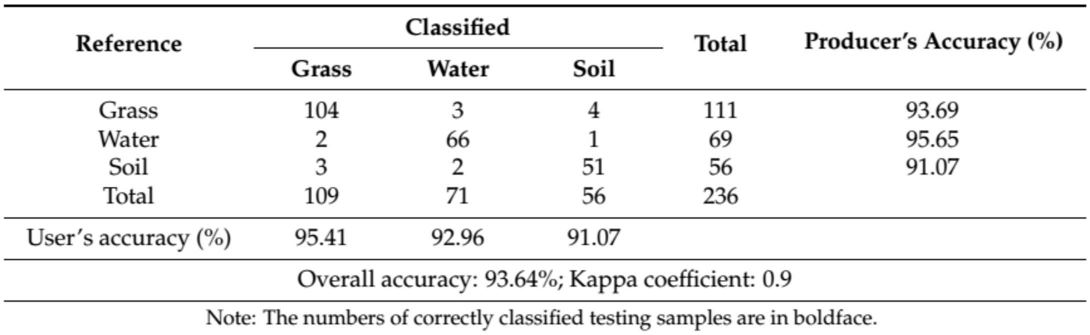


Figure 18: From Xia et al., (2017), Classification Map of the transect based on the UAV image acquired in July 2014.



Reference	Classified			Total	Producer's Accuracy (%)
	Grass	Water	Soil		
Grass	104	3	4	111	93.69
Water	2	66	1	69	95.65
Soil	3	2	51	56	91.07
Total	109	71	56	236	
User's accuracy (%)	95.41	92.96	91.07		
Overall accuracy: 93.64%; Kappa coefficient: 0.9					

Note: The numbers of correctly classified testing samples are in boldface.

Figure 19: From Xia et al., (2017), Confusion Matrix and accuracy estimates for the classified map.

The OBIA method of classification combined with scaling up of pixels using binary classifications as shown in Figure 18 and 19 show significant potential in integrating with the experiment conducted in Chapter 2 of this dissertation. Consider the

NDSV classification on Google Earth Engine and the validation process for urban extents that had to be undertaken using VHR Quickbird imagery (Patel et al., 2015), which is very expensive and less accurate in comparison to the method outlined here.

Generating full classifications would most likely be not optimal just yet for the open-source population mapping method in Stevens et al., (2015), however at least generating binary classifications for essential indicators of human population such as urban land use would be extremely useful when mapping continuous populations at less than 100 meter spatial resolution. This validation method would be especially useful for the high temporal resolution of urban extractions from Landsat data as described in Chapter 2. The swath that UAVs can collect data could potentially be too small, but ideally taking a representative sample and applying to a country-wide scale should be acceptable depending on the type of land-use that is to be observed for informing human population counts (Gaughan et al., 2014). OBIA with current UAV imaging platforms shows strong promise to be the next cost-effective and accurate way to validate and compliment satellite imagery, especially in applications attempting to map urbanized areas for human populations.

4.4 Summary

In the first section of this chapter, the motivations of mapping at 30 meter resolution is described, as this is the highest continuous dasymetric population mapping that has been attempted thus far. The approach in Dmowska and Stepinski (2017) is described and evaluated. In the second section of this chapter, the proposed process and motivation of utilizing the open-source dasymetric population mapping method used in this dissertation (Stevens et al., 2015) to map population at 30 meter resolution (using the test area of Java in Indonesia) is explained. Machine learning processes will be applied on High End Computing resources at NASA Ames Research Center because the processes are too computationally intensive for standard computers. The statistical outputs from utilizing this mapping at 30 meters will be compared to an exact test conducted over the same test area (Java, Indonesia) at 100 meters, to assess the effects of harmonizing data at different spatial scales. In the third section, a UAV based method to categorize reference pixels for sub-pixel inundation mapping was suggested as a possible validation method for the Google Earth Engine extracted images in Chapter 2.

CHAPTER FIVE: DISCUSSION, CONCLUSIONS AND FUTURE WORK

The experiments detailed in this dissertation aimed to improve an open-source population mapping method with the inclusion of spaceborne, terrestrial and airborne instruments. Utilizing novel methods, the inclusion of these instruments were successful in increasing the explanatory value of the population data grids that were created.

In Chapter Two, the spaceborne platform of Landsat was utilized to extract urban extents for different year collections using a tool called Google Earth Engine. Integration of this methodology reduced mapping inaccuracies and error along with improving temporal resolution. In Chapter Three, terrestrial instruments in the forms of mobile phones and computers were utilized as indicators of population utilizing geo-located Twitter data, the extraction of this data and utilization of this data for this purpose was novel, and significantly decreased mapping inaccuracies and error as well as proving to be a significant covariate dataset in the machine learning process. In Chapter Four, state-of-art research at mapping population at higher continuous resolutions was explored, as was proposed research with 30 meter continuous population mapping using NASA Supercomputing resources. An airborne method utilizing UAV image collection for reference classifications for spaceborne image validation was also suggested.

In reflecting on the development of instrumentation as it relates to improvement of tracking populations, on the very high spatial scales, any type of device carried by humans that interface with the Internet or cellular networks will be increasing valuable, with the main constraints surrounding privacy and commercial interests surrounding mobile phone records. However, projects like Flowminder are finding ways to make it very straightforward for phone network operators to work in collaboration with governments to provide planners with accurate population dynamics (Deville et al., 2014). In the assessment of this dissertation, mobile devices will remain the most valuable form of tracking population dynamics for the foreseeable future when considering terrestrial devices. In higher income countries the use of mobile devices will likely manifest itself in more localized tracking utilizing context signals from devices in the form of Bluetooth low energy (BLE). As the technology develops here and as the market forces decrease the cost of the devices, the ease of access to the data should increase as well, especially for humanitarian purposes. However for more temporal accuracy and depth in data, remotely sensed data will still remain an integral component to map populations.

From the remote sensing perspective, as detailed in Chapter Two and Chapter Four, spaceborne and airborne instruments are continually going to be integral sources of valuable datasets for population mapping. The importance of being able to characterize human population distributions in present day is underscored by the fact that the study of global processes has suffered from lack of systematic, quantitative descriptions of land surface characteristics that control the landscape, such as urbanized areas. Spaceborne instruments (like the appropriately vaunted NASA/USGS Landsat series) and airborne instruments (like the low-cost UAV's described in Chapter Four) will need to be used in co-unison in order to underpin accuracy assessments of remotely sensed data.

Developments with methodologies here will lead to the development of appropriate, thorough and scale-consistent geographic descriptions of Earth's land surface. These geographic descriptions will serve as a bridge for *in situ* process knowledge, or data that is gathered directly at the areas where land cover data is being gathered. Research in this area will lead to better integrated parameters of global models, and global population modelling will certainly be one of them.

As detailed in Chapter Four, in order to develop better temporally relevant classifications, the appropriate accuracy assessment data must be obtained as well. And with the relatively low-cost and easy to use UAV-based methods, for large urban areas this should be very scalable. Object-based classifications (appropriate for binary classifications) conducted to identify relevant controlling land surface characteristics, such as urban areas, will be utilized heavily in order to monitor changes in land-cover and hence provide invaluable datasets for evaluating population distributions. The combination of analyses utilizing spaceborne and airborne validation data will be the future of remote sensing, and will improve in pixel classifications, at higher temporal and spatial resolutions.

Recently, technology companies have demonstrated interest in generating population maps for their own interests. Facebook disaggregated population counts from census grids at 5 meter per-pixel resolution for their Internet.org project. They disaggregated census counts over binary pixels that they classified as human-built structures utilizing computer vision techniques on very high resolution Digital Globe Imagery (Gros & Tiecke, 2016). Private companies like Planet Labs are also starting to offer image services from constellations of nano-satellites that are a large number of small, compact satellites that often weigh less than 10 kilograms. These new private companies represent higher accessibility to instruments that can allow for effective management and assessment of the terrestrial system at sub-field scales (between 1 to 10 meter resolutions). Planet Labs currently operates the largest constellation of satellite systems in orbit, with their RGB (red, green, blue) imagery offerings at 3 to 5 meter resolution on a daily scale, utilizing a constellation of satellites (150 to 200 in number). The European Space Agency's Sentinel-2 pair will improve the spatio-temporal frequency of satellite platforms such as Landsat 8, and the synergy of both these platforms will further enhance the temporal resolution of analyses. Other than Planet Labs' offerings, the requirements of very high spatial resolution and near-daily frequency for satellite images can only be acquired through targeted acquisition with companies like WorldView and RapidEye. Planet Labs's satellites have a caveat to their performance in that their sensor designs and utilization of commercial off the shelf components do not compare to the signal-to-noise characteristics, radiometric performance, cross-sensor consistency and spectral enhancements of satellite images that traditional space agencies have been able to generate. The lack of at-sensor radiance conversions and atmospheric correction of the imagery derived from these nano-satellites can cause researchers to question the time and space consistency of any time-series type data, as bands in the visible domain (Red, Green, Blue) are very sensitive to atmospheric correction processes.

Without being able to perform these corrections, the ability to reliably infer actual changes in surface cover conditions is called into question (Houberg & McCabe, 2016).

These new instruments will allow for more spatio-temporally relevant data to be integrated into open-source population mapping processes. (e.g. health, planning, and disaster relief applications). As the resolution of all these datasets increase the computational loads will also proportional increase and this will need equally impressive computer vision analysis and performance optimization in regards to computing. Section 4.2 describes the motivations behind a test for this exact reason and its relevance to testing high-end computing resources. Even classifications of remotely sensed data are being tested using quantum annealing, an experimental and hopefully breakthrough computational technology for handling hard optimization problems (Boyda et al., 2017).

An open-source population mapping methodology was the foundation of the experiments in this dissertation, and significant improvements and insights were observed in each of these experiments in modelling population distribution. The future work of the author after this dissertation will involve the integration of more relevant datasets from different instruments to improve the population mapping method, as well as examining what novel computing methods can be utilized to further optimize the processing. This particular application of Earth Science is essential for the curation of the planet. Tracking our activities as a species on the Earth's surface will allow us to reflect on what parts of our global society and infrastructure need improvement, and it is hoped that by being able to model our footprint, we can create a sustainable future for the Earth while also meeting our needs as a species.

APPENDIX 1

The discipline of remote sensing often encompasses the use of aerial sensor technologies to detect and classify objects on the Earth, referring to objects on the Earth surface as well in the atmosphere and oceans, by the use of propagated signals, usually via electromagnetic radiation. Active remote sensing involves signals being first emitted from aircraft or satellites, and passive remote sensing refers to when information is simply recorded (Liu and Mason, 2014).

The open-source population mapping method tested within this dissertation topic has been detailed in Stevens, et al (2015). This dasymetric modelling approach incorporates various ancillary datasets including vector and raster data, and the raster data sources are often remotely sensed data.

1. Land Cover Remotely Sensed Data:

For all three experiments within this dissertation topic, an MDA EarthSat GeoCover Landsat Thematic Mapper derived land cover raster is utilized in combination with more highly temporally relevant urbanized datasets used in the first experiment utilizing the Google Earth Engine tool (Patel et al., 2015). MDA Federal has a proprietary process in the way it creates this GeoCover land cover raster at 30 meter resolution, self-described as "derived from spectral analysis of consistently orthorectified Landsat Thematic Mapper ™ imagery" and has a standard 13-land cover legend. The version that is used as the base MDA GeoCover dataset is dated to 2007 (MDA Federal Inc., 2007). The first experiment explicitly tests different urban extractions from Google Earth Engine, which also utilizes Landsat Thematic Mapper data from different year sets in order to increase the temporal resolution of the land cover data that is used in the population mapping process. Because the study area is consistent in the second and third experiment, the same best land cover dataset will be used with temporal modifications as necessary.

The remote sensing instrument that is consistent with both of these methods is the NASA/USGS Landsat series. The Landsat program is the longest-running enterprise for the acquisition of satellite imagery of the Earth and has been utilized for a wide variety of applications. The newest satellite Landsat 8 has a total of 11 spectral bands and spatial resolutions ranging from 15 meters to 100 meters, with a temporal resolution being only 16 days (U.S Geological Survey, 2016). The biggest advantage of using this dataset is its availability to the general public as it if freely downloadable as soon as it is available.

The Land Cover datasets used were largely constructed from Thematic Mapper (multispectral scanning radiometer) from Landsat 5 and Enhanced Thematic Mapper (multispectral scanning radiometer) from Landsat 7. The Operational Land Imager on the Landsat 8 satellite might be used for improving the land cover on Experiment 3.

It is important to note that in the methodology the temporal resolution of the Landsat derived land cover is further coarsened to 100 meters due to the needs of the population mapping process, so this is often noted and detailed within each of the experiments.

2. Continuous Raster-Format Remotely Sensed Data:

I. Suomi VIIRS NPP Derived Lights-At-Night data – the Suomi National Polar-orbiting Partnership weather satellite is operated by NOAA and includes a wide array of instruments including VIIRS, the Visible Infrared Imaging Radiometer Suite. This scanning radiometer collects imagery and radiometric measurements of land, atmosphere, cryosphere, and oceans within the visible and infrared bands of the electromagnetic spectrum. There is a product that NOAA distributes from VIIRS called the Day/Night Band sensor that is a 750 meter resolution product that has been useful in measuring radiance and reflectance off the Earth's surface, and orbits the Earth approximately 14 times per day. Night-time radiance data often is useful in mapping the extent of human populations, and hence this data is downloaded from NOAA as they provide it as a Nighttime VIIRS Day/Night Band Composites, frequently updating it on their website at 15 arc-second resolution, which we then resample down to 100 meter resolution for use in our processes (Hillger et al, 2013)

II. MODIS 17A3 2010 Estimated Net Primary Productivity, 1 km data, utilizing the methods of Running et al, 2004. MODIS (MODerate-resolution Imaging Spectroradiometer) was launched into Earth orbit by NASA with the Terra and Aqua satellites, capturing data in 36 different spectral bands, with spatial resolutions varying at 250m (2 bands), 500m (5 bands) and 1km (29 bands) (Salomonson et al., 2002). The estimated Net Primary Productivity product, produces gross primary production of vegetation every day, and sums this to net primary production, which is vegetation growth at the end of the year. This product is computed with daily MODIS landcover and other agriculturally related calculations like FPAR/LAI and GMAO and is accessible via FTP at one of NASA’s DAACs. The product is integrated into our population mapping process as a single variable, mainly to provide further insight on land cover use and inform the spatial distribution of vegetated areas. The 1km gridded data for an entire country is also resampled to a finer 100 meter grid for use in our population mapping process.

III. WorldClim/BioClim Mean Temperature and Mean Precipitation 1950 – 2000 – These continuous raster datasets are created from the methods of Hijmans et al. 2005 specifically using the WorldClim model that are an essentially a set of climate layers that try to represent “current” conditions by interpolating data over a certain set of years (in this case 1950-2000). These are produced at 1 km resolution and the specific datasets that WorldPop’s uses in our population mapping process are BioClim for Mean Temperature and Mean Precipitation as significant variables in our dasymetric weighting process for population mapping. WorldClim uses a major climate databases that collect information from terrestrial sensors, however, the climate data is overlaid with the SRTM (Shuttle Radar Topography Mission) which collected digital elevation models on a near-global scale from 56° S to 60° N, to generate the most complete high-resolution digital topographic database of Earth on the 11 day STS-99 mission of Space Shuttle Endeavour (Van Zyl, 2001).

IV. HydroSHEDS derived Elevation & Slope - HydroSHEDS (Hydrological data and maps based on SHuttle Elevation Derivatives at multiple Scales) provides hydrographic information. The dataset was primarily developed at the Conservation Science Program of the World Wildlife Fund US (WWF-US) in collaboration with the USGS and other agencies. The database was developed to generate key data layers in support of regional and global watershed analyses, hydrological modeling, and freshwater conservation planning that was previously unavailable. HydroSHEDS offers myriad geo-referenced data sets (vector and raster), including stream networks, watershed boundaries, drainage directions, and ancillary data layers such as flow accumulations, distances, and river topology information. Resolutions range from 3 arc-second (~ 90 meter) to 5 minute (~ 10 km) at near-global extent. The HydroSHEDS data based on elevation data of the Shuttle Radar Topography Mission (SRTM) at 3 arc-second (~ 90 meter) resolution. To generate HydroSHEDS, the original SRTM elevation data were hydrologically conditioned in a certain sequence of procedures. Typical data improvement and algorithmic improvements have been applied, including customized gap filling, filtering, and stream burning, and upscaling techniques. The elevation and slope of these datasets are used as variables in the weighting process for the high resolution population mapping method.

APPENDIX 2

“Guide to multi-year script”, executable within Google Earth Engine environment in the Javascript language, developed by Trianni et al., 2015.

--

All the parameters that can be modified are in the first session of the script (PARAMETERS TO BE SET), do not change any line outside this area (there is a couple of exception to this rule, specified in the guide).

Choose the region and the years to analyze:

1. **area, country_name, province_name**, are used to select the region of interest: the first one selects a large area (names of available zones are specified in the script), the second a country, the last a province of a country (available ONLY for Indonesia and China). The parameters are mutually exclusive, and they are checked in order, so if you want to use only a large area the other two parameters must be set to "", otherwise country_name “wins” on area and province_name “wins” on both.

E.g.: with this configuration the script selects the entire North America

```
var area = "North America";
var country_name = "";
var province_name = "";
```

with this one the script selects China

```
var area = "North America";
var country_name = "China";
var province_name = "";
```

IMPORTANT: province_name is considered only if country_name is equal to Indonesia or China, otherwise it must be set to “”, or the script crashes. E.g., with this configuration the script selects province of Bandung

```
var area = "North America";
var country_name = "Indonesia";
var province_name = "Bandung";
```

otherwise, with this one it returns an error

```
var area = "North America";
var country_name = "Brazil";
var province_name = "Bandung";
```

2. **zoom** and **center_on** are used to final visualization, the first one set the level of zoom, the second one specified the city on which center the map, if it is set to "" the map will be placed on the center of the area/country.

3. **year, number_of_years,** and **step** select respectively the first year from which the script starts, the total number of years to analyze, and the step progress between years (e.g. set 1 for 1 year in 1 year, set 2 for 2 years in 2 years, etc.). The script always works backwards, so e.g.

```
var year = 2010;
var number_of_years = 10;
var step = 2;
```

means that the script analyze from 2010 to 2001, 2 year by 2 (in there is no images for one year it skips to the next one).

4. **collection_type, sensor** and **cloud_cover** select the input types (values are specified in the script). All combination are possible but cloud_cover refers only to standard type collection.

IMPORTANT: the script "autoscale" between collection and sensor types in function of the year (e.g. there is not Landsat 7 data before 1995 so it sets automatically Landsat 5), so these parameters are used ONLY for the first year. If you want to change the year range for specific sensors, goes to line 1145 and after.

Set the classifier/s

1. **classifiers** is used to select which classifier/s to use (see the list and the codes in the comments). Generally it must not be changed.
2. **multiclassifier** enable/disable the use of multiclassification: set false to use one classifier for year, true to use three.
3. **compute_area** is used to enable/disable the computation of the area of the classified zone, **area_scale** is a scale factor (generally it must not be changed).
4. **ts_type** is an important parameter to select what kind of training set use as input for the classifier. Different parameters are enable/disable in function of this choice:
    a. If it is set to 0, manual training set are enabled, so the only parameter to check is what fusion table/s use. The list of fusion tables starts from line 235, check what table/s you want to use, and then go to line 1103 and

insert the code(s) separated by commas.

b. If it is set to 1, Globcover random points training set are enabled. In this case you have to specify the area in which will be search the points. In particular Globcover works inside square areas so you must set latitude (ltd) and longitude (lat) of the center of this quad
c. If it is set to 2 Universe of city random points training set are enabled. In this case you can specify in which city search the points, writing the name in **selectedCities**. Alternatively if you want to use ALL the cities in a province set **use_all_cities** to true. **IMPORTANT**: if you set use_all_cities = true, selectedCities are completely ignored. use_all_cities can be set to true ONLY if a province_name is specified, otherwise it must be set to false.

NB: for both the random points training set you must specified the total number of points to use (**num_points**).
IMPORTANT: the parameters used only by specific type are ignored by the others. E.g. if ts_type=0, selectedCities will be ignored, on the contrary if ts_type=2 the chosen fusion tables will be ignored. In this way you can set at the beginning the configuration for all three types without conflicts and then easily switches between them changing the ts_type.

Classification refinements

1. **Waterfilter, NDVI, Elevation, Morphology.** All these refinement “filters” are applied in sequence. Each of them can be enable/disable setting true or false the correspondent flag (**waterfilter, mask_ndvi, mask_slope, morphology_on**). The meaning of each specific parameter is described in the script comments.

2. **class_or_flag** is used to switch between two different correction mode. If false the script applies to each year (except the first one) the logical AND between the current classification and the classification of the previous year. If true the script applies to each year (except the first one) the logical AND between the current classification and a global mask dynamically generated. The global mask is the logical AND between the logical
3. OR of the classifications of all years and the morphological closing of the first year classification.

4. **Ground truth,** test options do not enable.

Examples

Main parameters in some common situations. Refinement parameters such as morphology or NDVI are ignored, because there are not standard values, they can be adapted, enabled or disabled by the users.

1. Kota Bandung, Indonesia from 2010 to 1995, step of 5 years, Landsat 7 Greenest, manual training set

```
// General options
var area = "South Est Asia";   -> this value is irrelevant in this case
var convex_hull = false;   -> this value is irrelevant in this case
var country_name =  "Indonesia”;
var province_name = "Jawa Barat”;

var center_on = "Bandung";
var zoom = 8;

var year = 2010;
var number_of_years = 15;
var step = 5;
var collection_type = "greenest”
var sensor = "L7";
var cloud_cover = 1;

// Classifier parameters
var ts_type = 0;           -> this value specified the use of a manual training set based on
fusion tables
var multiclassifier = false;
var classifiers = new Array(5,9,6);

/* ALL THE PARAMETERS IN YELLOW ARE IGNORED FOR TS_TYPE=0 * /
var num_points = 500;
// Globcover random points parameters
var lng = -48.90564;
var lat = -0.890311;
var radius1 = 500;
var kernelType1 = 'square';

// Universe of cities parameters:
var cities_table = ee.FeatureCollection('ft:1pQ-
PrIEGrYa2Y3v9tsN1xwfYuqRIqOoDPARgpwzS');

// Chose the city or the cities to use
var selectedCities = cities_table.filter(ee.Filter.or(ee.Filter.eq('MAIN_CITY', 'Kunming'),
                                                      ee.Filter.eq('MAIN_CITY', 'Yuxi'),
                                                             ee.Filter.eq('MAIN_CITY',
'Qujing')));
var use_all_cities = true;
```

At line 1103 set

```
var tables_array = new Array(ft2);  -> this is the code of the fusion table  for this area
```

2. Jiangsu Province (P.R. China) from 2008 to 2004, step of 2 years, Landsat 7 Greenest, Globcover training set, multiclassifier

```
// General options
var area = "South Est Asia";   -> this value is irrelevant in this case
var convex_hull = false;   -> this value is irrelevant in this case
var country_name =  "China”;
var province_name = "Jiangsu”;

var center_on = "";
var zoom = 8;

var year = 2008;
var number_of_years = 5;
var step = 2;
var collection_type = "greenest”
var sensor = "L7";
var cloud_cover = 1;

// Classifier parameters
var ts_type = 1;           -> this value specified the use of a random training set based on
Globcover
var multiclassifier = false;
var classifiers = new Array(5,9,6);

var num_points = 500;
// Globcover random points parameters
var lng = 119.43787;
var lat = 31.90554;
var radius1 = 400;
var kernelType1 = 'square';

/* ALL THE PARAMETERS IN YELLOW ARE IGNORED FOR TS_TYPE=1 * /
// Universe of cities parameters:
var cities_table = ee.FeatureCollection('ft:1pQ-
PrIEGrYa2Y3v9tsN1xwfYuqRIqOoDPARgpwzS');

// Chose the city or the cities to use
```

```
var selectedCities = cities_table.filter(ee.Filter.or( ee.Filter.eq('MAIN_CITY', 'Nanjing'),
                                         ee.Filter.eq('MAIN_CITY', 'Suzhou (Jiangsu)'),
                                         ee.Filter.eq('MAIN_CITY', 'Changzhou')) );
var use_all_cities = true;
```

3. Brazil from 2010 to 2000, step of 1 year, Landsat 7 Greenest, Universe of cities training set, single classifier

```
// General options
var area = "South Est Asia";   -> this value is irrelevant in this case
var convex_hull = false;   -> this value is irrelevant in this case
var country_name =  "Brazil”;
var province_name = "”;  -> this value must be “”

var center_on = "";
var zoom = 8;

var year = 2010;
var number_of_years = 11;
var step = 1;
var collection_type = "greenest”
var sensor = "L7";
var cloud_cover = 1;

// Classifier parameters
var ts_type = 2;          -> this value specified the use of a random training set based on
Universe of cities
var multiclassifier = true;
var classifiers = new Array(5,9,6);

var num_points = 500;
/* ALL THE PARAMETERS IN YELLOW ARE IGNORED FOR TS_TYPE=2 * /
// Globcover random points parameters
var lng = 119.43787;
var lat = 31.90554;
var radius1 = 400;
var kernelType1 = 'square';

// Universe of cities parameters:
var cities_table = ee.FeatureCollection('ft:1pQ-
PrIEGrYa2Y3v9tsN1xwfYuqRIqOoDPARgpwzS');

// Chose the city or the cities to use
```

```
var selectedCities = cities_table.filter(ee.Filter.or(ee.Filter.eq('MAIN_CITY', 'Rio de Janeiro'),
                                   ee.Filter.eq('MAIN_CITY', 'Sao Paolo')));

var use_all_cities = false;
```

APPENDIX 3

This Appendix catalogs the computing environments necessary for each of the Chapters.

For Chapter Two's experiment:

- The urban extents for Landsat were processed in the online Google Earth Engine environment with the code detailed in Appendix II (Trianni et al., 2015)
- The WorldPop open-source population mapping method utilized the Python and R code found here: https://github.com/ForrestStevens/WorldPop-RF
- The most current version of ArcGIS was utilized in 2014 as was the most current version of R at the time (3.0.1 – Good Sport)
- The computer specifications to run the Python and R code: Operating System: Windows 10, Processor: Intel ® Core ™ i7-3940XM CPU @ 3.00GHz 3.20GHz Installed Memory (RAM): 16 GB, System Type: 64 bit operating system with x64 based processor, Hard Drive: 393GB internally used, 1 TB and 4 TB externally utilized.

For Chapter Three's experiment:

- The WorldPop open-source population mapping method utilized the Python and R code found here: https://github.com/ForrestStevens/WorldPop-RF

- The most current version of ArcGIS was utilized in 2015 as was the most current version of R in Fall 2015/Spring 2016
- The computer specifications to run the Python and R code: Operating System: Windows 10, Processor: Intel ® Core ™ i7-3940XM CPU @ 3.00GHz 3.20GHz Installed Memory (RAM): 32 GB, System Type: 64 bit operating system with x64 based processor, Hard Drive: 393GB internally used, 2 TB and 4 TB externally utilized.

For Chapter Four's experiment detailing a proposal for Supercomputing resources:

- The WorldPop open-source population mapping method utilized the Python and R code found here: https://github.com/ForrestStevens/WorldPop-RF
- The most current version of ArcGIS was utilized in 2017 as was the most current version of R in 2017
- The computer specifications to run the Python: Operating System: Windows 10, Processor: Intel ® Core ™ i7-3940XM CPU @ 3.00GHz 3.20GHz Installed Memory (RAM): 32 GB, System Type: 64 bit operating system with x64 based processor, Hard Drive: 393GB internally used, 2 TB and 4 TB externally utilized.
- Critically due to needs for massive memory in vector allocations, the R code currently being run on an allocation of 96 cores within NASA's Pleiades Supercomputer, which is managed by NASA's Advanced Supercomputing Division. This work is facilitated by the NASA Earth Exchange (see Nemani et

al., 2011), which also has allocated 1TB of memory and 12GB of RAM in an Linux environment that is accessed through Secure Shell protocol (SSH).

REFERENCES

Alberti, M., Weeks, R., & Coe, S. (2004). Urban Land-Cover Change Analysis in Central Puget Sound. *Photogrammetric Engineering & Remote Sensing, 70*(9), 1043-1052. doi:10.14358/PERS.70.9.1043

Anderson, J. R. (1976). A land use and land cover classification system for use with remote sensor data (Vol. 964). US Government Printing Office.

Angel, S., Sheppard, S. C., Civco, D. L., Buckley, R., Chabaeva, A., Gitlin, L., Kraley, A., Parent, J., Perlin, M. (2005). The dynamics of global urban expansion. Washington, D.C.: World Bank, Transport and Urban Development Department.

Angiuli, E., & Trianni, G. (2014). Urban Mapping in Landsat Images Based on Normalized Difference Spectral Vector. *IEEE Geoscience and Remote Sensing Letters, 11*(3), 661-665. doi:10.1109/LGRS.2013.2274327

Azar, D., Graesser, J., Engstrom, R., Comenetz, J., Leddy Jr, R. M., Schechtman, N. G., & Andrews, T. (2010). Spatial refinement of census population distribution using remotely sensed estimates of impervious surfaces in Haiti. International Journal of Remote Sensing, 31(21), 5635-5655.

Azar, D., Engstrom, R., Graesser, J., & Comenetz, J. (2013). Generation of fine-scale population layers using multi-resolution satellite imagery and geospatial data. Remote Sensing of Environment, 130, 219-232.

Bagan, H., & Yamagata, Y. (2012). Landsat analysis of urban growth: How Tokyo became the world's largest megacity during the last 40 years. *Remote Sensing of Environment, 127*, 210-222. doi: 10.1016/j.rse.2012.09.011

Balk, D. & Yetman, G. (2004) The Global Distribution of Population: Evaluating the gains in resolution refinement. Center for International Earth Science Information Network (CIESIN). Available: http://sedac.ciesin.org/gpw/docs/ gpw3_documentation_final.pdf.

Balk, D., Pozzi, F., Yetman, G., Deichmann, U., & Nelson, A. (2005) The distribution of people and the dimension of place: Methodologies to improve the global estimation of urban extents.; Proceedings of the Urban Remote Sensing Conference; Tempe, AZ. International Society for Photogrammetry and Remote Sensing.

Balk, D., Deichmann, U., Yetman, G., Pozzi, F., Hay, S., & Nelson, A. (2006). Determining Global Population Distribution: Methods, Applications and Data. *Advances in Parasitology Global Mapping of Infectious Diseases: Methods, Examples and Emerging Applications,* 119-156. doi:10.1016/s0065-308x(05)62004-0

Bakillah, M., Liang, S., Mobasheri, A., Jokar Arsanjani, J., & Zipf, A. (2014).

Fine-resolution population mapping using OpenStreetMap points-of-interest. *International Journal of Geographical Information Science*, *28*(9), 1940-1963.

Bartholomé, E., & Belward, A. S. (2005). GLC2000: a new approach to global land cover mapping from Earth observation data. *International Journal of Remote Sensing*, *26*(9), 1959-1977.

Blanford, J. I., Huang, Z., Savelyev, A., & Maceachren, A. M. (2015). Geo-Located Tweets. Enhancing Mobility Maps and Capturing Cross-Border Movement. *PLoS ONE PLOS ONE, 10*(6). doi:10.1371/journal.pone.0129202

Blaikie, P., Cannon, T., Davis, I., & Wisner, B. (2014). *At risk: natural hazards, people's vulnerability and disasters*. Routledge.

Blaikie, P., & Brookfield, H. (Eds.). (2015). *Land degradation and society*. Routledge.

Bhaduri, B., Bright, E., Coleman, P., & 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. doi:10.1007/s10708-007-9105-9

Bongaarts, J. (2009). Human population growth and the demographic transition. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1532), 2985-2990. doi:10.1098/rstb.2009.0137

Boyda, E., Basu, S., Ganguly, S., Michaelis, A., Mukhopadhyay, S., & Nemani, R. R. (2017). Deploying a quantum annealing processor to detect tree cover in aerial imagery of California. PloS one, 12(2), e0172505.

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

Breiman, L., 1996. Bagging Predictors. Mach. Learn. 24 (2), 123–140.

Breiman, L., 2001. Random Forests. Mach. Learn. 45 (1), 5–32, http://dx.doi.org/ 10.1023/A:1010933404324.

Briggs, D. J., Gulliver, J., Fecht, D., & 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.

Burchfield, M., Overman, H. G., Puga, D., & Turner, M. A. (2006). Causes of Sprawl: A Portrait from Space. The Quarterly Journal of Economics, 121(2), 587-633 doi:10.1162/qjec.2006.121.2.587

Carpenter, D., & Carpenter, S. (1983). Modeling inland water quality using Landsat data. *Remote Sensing of Environment*, *13*(4), 345-352. doi: 10.1016/0034-4257(83)90035-4 Center for International Earth Science Information Network (CIESIN), Columbia University, International Food Policy Research Institute (IFPRI), The World Bank and Centro International de Agricultura Tropical (CIAT), 2004. Global Rural–Urban Mapping Project (GRUMP): Urban Extents. Columbia University, Palisades, New York, CIESIN.

Cheriyadat, A., Bright, E., Potere, D., & Bhaduri, B. (2007). Mapping of settlements in high-resolution satellite imagery using high performance computing.

GeoJournal, 69(1-2), 119-129. doi:10.1007/s10708-007-9101-0

Cohen, J. E. (1995). How many people can the earth support?. *The Sciences*, *35*(6), 18-23.

Commoner, B. (1972). *The closing circle: confronting the environmental crisis*. London: Cape.

Commoner, B. (1990). Making peace with the planet.

Costanza, R., Cumberland, J. H., Daly, H., Goodland, R., Norgaard, R. B., Kubiszewski, I., & Franco, C. (2014). *An introduction to ecological economics*. CRC Press.

Deichmann, U. (1996). A review of spatial population database design and modeling. Santa Barbara, CA: University of California, Santa Barbara. National Center for Geographic Information and Analysis (NCGIA). Technical Report 96–3.

Deichmann, U., Balk, D., & Yetman, G. (2001). Transforming population data for interdisciplinary usages: from census to grid. Washington (DC): Center for International Earth Science Information Network, 200(1).

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., Blondela, V. D., Tatem, A. J. (2014). Dynamic population mapping using mobile phone data. *Proc Natl Acad Sci USA Proceedings of the National Academy of Sciences, 111*(45), 15888-15893. doi:10.1073/pnas.1408439111

Dmowska, A., & Stepinski, T. F. (2014). High resolution dasymetric model of US demographics with application to spatial distribution of racial diversity. *Applied Geography*, *53*, 417-426.

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

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

Dorling, D. (1993). Map design for census mapping. The Cartographic Journal, 30(2), 167-183.

Earthengine-api, 2014. Earth Engine Access Library—Google Project Hosting, Retrieved from http://code.google.com/p/earthengine-api/ February 1, 2014.

Edwards, A. C. (1969). Walker's 1870 Statistical Atlas and the Development of American Cartography (Doctoral dissertation, UNIVERSITY OF WISCONSIN).

Ehrlich, P. (1970). The population bomb. *New York Times,* 47.

Ehrlich, P. R., & Ehrlich, A. H. (1990). *The population explosion* (pp. 37-40). New York: Simon and Schuster.

Eicher, C. L., & Brewer, C. A. (2001). Dasymetric mapping and areal interpolation: Implementation and evaluation. *Cartography and Geographic Information Science*, *28*(2), 125-138.

Elvidge, C., Baugh, K., Kihn, E.A., Davis, E.R., 1996. Mapping city lights with nighttime data from the DMSP operational linescan system. Photogrammetric Engineering and Remote Sensing. 63, 727–734

Elvidge, C. D., Baugh, K. E., Dietz, J. B., Bland, T., Sutton, P. C., & Kroehl,

H. W. (1999). Radiance calibration of DMSP-OLS low-light imaging data of human settlements. Remote Sensing of Environment, 68, 77 – 88

Elvidge, C., Baugh, K. E., Hobson, V. R., Kihn, E. A., Kroehl, H. W., Davis, E. R., et al. (1997). Satellite inventory of human settlements using nocturnal radiation emissions: A contribution to the global toolchest. Global Change Biology, 3, 387 – 395.

Erb, K. H., Gaube, V., Krausmann, F., Plutzar, C., Bondeau, A., & Haberl, H. (2007). A comprehensive global 5 min resolution land-use data set for the year 2000 consistent with national census data. *Journal of land use science*, *2*(3), 191-224.

Flowerdew, R., & 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.

Foody, G. M. (2000). Estimation of sub-pixel land cover composition in the presence of untrained classes. Computers & Geosciences, 26, 469 – 478.

Gallego, F. J., Batista, F., Rocha, C., & Mubareka, S. (2011). Disaggregating population density of the European Union with CORINE land cover. *International Journal of Geographical Information Science*, *25*(12), 2051-2069.

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

Gaughan, A. E., Stevens, F. R., Linard, C., Jia, P., & Tatem, A. J. (2013). High Resolution Population Distribution Maps for Southeast Asia in 2010 and 2015. *PLoS ONE*, *8*(2). doi:10.1371/journal.pone.0055882

Gaughan, A., Stevens, F., Linard, C., Patel, N., & Tatem, A. (2014). Exploring nationally and regionally defined models for large area population mapping. International Journal of Digital Earth, 1-18. doi:10.1080/17538947.2014.965761

GeoHive. (2014). *Population Statistics*. Retrieved October 1st, 2013, from http://www.geohive.com/

Global Administrative Areas. (2014). *Global Administrative Areas*. Retrieved October 1st, 2013, from http://www.gadm.org/

Global Urban Footprint. (2015). Retrieved from http://dlr.de/eoc/en/desktopdefault.aspx/tabid-9628/16557_read-40454/

Goodchild, M. F. (1992). Geographical data modeling. Computers & Geosciences, 18(4), 401-408.

Goodchild, M. F., Anselin, L., & 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., & 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.

Gros, A., & Tiecke, T. (2016, February 22). Connecting the world with better maps. Retrieved April 16, 2017, from https://code.facebook.com/posts/1676452492623525/connecting-the-world-with-better-maps/

Guindon, B., Zhang, Y., & Dillabaugh, C. (2004). Landsat urban mapping based on a combined spectral–spatial methodology. Remote Sensing of Environment 92(2), 218-232. doi:10.1016/j.rse.2004.06.015

Hansen, M., DeFries, R., Townshend, J.R.G., Sohlberg, R., 1998. 1km Land Cover Classification Derived from AVHRR. The Global Land Cover Facility, College Park, Maryland.

Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D., Stehman, S.V., Goetz, S.J., Loveland, T.R., Kommareddy, A., Egorov, A., Chini, L., Justice, C.O., Townshend, J.R.G. (2013). High-resolution global maps of 21st-century forest cover change. Science 342 (6160), 850–853, http://dx.doi.org/10.1126/science.1244693

Harvey, D. (1979). Population, resources, and the ideology of science. In Philosophy in geography (pp. 155-185). Springer Netherlands.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. (2014). Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, *41*(3), 260-271. doi:10.1080/15230406.2014.890072

Heilig, G. K. (1994). Neglected dimensions of global land-use change: reflections and data. *Population and Development Review*, 831-859.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol. International Journal of Climatology, 25*(15), 1965-1978. doi:10.1002/joc.1276

Hillger, D., Kopp, T., Lee, T., Lindsey, D., Seaman, C., Miller, S., Solbrig, J., Kidder, S., Bachmeier, S., Jasmin, T., Rink, T. (2013). First-light imagery from Suomi NPP VIIRS. Bulletin of the American Meteorological Society. http://dx.doi.org/10.1175/BAMS-D-12-00097.1

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

Homer, C., Huang, C., Yang, L., Wylie, B., & Coan, M. (2004). Development of a 2001 national land-cover database for the United States. *Photogrammetric Engineering & Remote Sensing*, *70*(7), 829-840.

Houborg, R., & McCabe, M. F. (2016). High-resolution NDVI from Planet's constellation of earth observing nano-satellites: a new data source for precision agriculture. *Remote Sensing*, *8*(9), 768.

IUCN and UNEP, 2012. The World Database on Protected Areas (WDPA). UNEP-WCMC, Cambridge, UK, Retrieved from http://www.protectedplanet.net (October 15th, 2013).

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

Jia, P., Qiu, Y., & Gaughan, A. E. (2014). A fine-scale spatial population distribution on the high-resolution gridded population surface and application in Alachua County, Florida. *Applied Geography*, *50*, 99-107.

Joint Research Center - Global Human Settlement Layer (2015). Retrieved from http://ghslsys.jrc.ec.europa.eu/

Jones, H. R. (1990). Population geography. Guilford Press.

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

Laliberte, A. S., Rango, A., Havstad, K. M., Paris, J. F., Beck, R. F., McNeely, R., & Gonzalez, A. L. (2004). Object-oriented image analysis for mapping shrub encroachment from 1937 to 2003 in southern New Mexico. *Remote Sensing of Environment*, *93*(1), 198-210.

Langford, M., Maguire, D. J., & 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., & Unwin, D. J. (1994). Generating and mapping population density surfaces within a geographical information system. The Cartographic Journal. 31:21–26

Leetaru, K., Wang, S., Cao, G., Padmanabhan, A., & Shook, E. (2013). Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday,18*(5).

Lehner, B., Verdin, K., Jarvis, A., Fund, W.W., 2006. HydroSHEDS Technical Documentation. World Wildlife Fund, pp. 27.

Leung, S., Martin, D., & Cockings, S. (2010, September). Linking UK public geospatial data to build 24/7 space-time specific population surface models. In GIScience 2010: Sixth International Conference on Geographic Information Science.

Liaw, A., Wiener, M., 2002. Classification and Regression by random forest. R News 2 (3), 18–22.

Linard, C., & Tatem, A. J. (2012). Large-scale spatial population databases in infectious disease research. International Journal of Health Geographics Int J Health Geogr, 11(1), 7. doi:10.1186/1476-072x-11-7

Linard, C., Alegana, V. A., Noor, A. M., Snow, R. W., & Tatem, A. J. (2010). A high resolution spatial population database of Somalia for disease risk mapping. *International Journal of Health Geographics Int J Health Geogr, 9*(1), 45. doi:10.1186/1476-072x-9-45

Linard, C., Gilbert, M., & Tatem, A. J. (2010). Assessing the use of global land cover data for guiding large area population distribution modelling. *GeoJournal, 76*(5), 525-538. doi:10.1007/s10708-010-9364-8

Linard, C., Gilbert, M., Snow, R. W., Noor, A. M., & Tatem, A. J. (2012). Population Distribution, Settlement Patterns and Accessibility across Africa in 2010. PLoS ONE, 7(2). doi:10.1371/journal.pone.0031743

Lloyd, C. D. (2014). *The Modifiable Areal Unit Problem, in Exploring spatial scale in geography*. John Wiley & Sons

Lloyd, C. T., Sorichetta, A., & Tatem, A. J. (2017). High resolution global gridded data for use in population studies. *Scientific Data*, *4*.

Liu, J. G., & Mason, P. J. (2013). *Essential image processing and GIS for remote sensing*. John Wiley & Sons.

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

Luck, G. W. (2007). A review of the relationships between human population density and biodiversity. *Biological Reviews*, *82*(4), 607-645.

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

Ma, L., Cheng, L., Li, M., Liu, Y., & Ma, X. (2015). Training set size, scale, and features in geographic object-based image analysis of very high resolution unmanned aerial vehicle imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, *102*, 14-27.

Maantay, J. A., Maroko, A. R., & 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.

Malthus, T. (1798). AN ESSAY ON THE PRINCIPLE OF POPULATION, AS IT AFFECTS THE FUTURE IMPROVEMENT OF SOCIETY WITH REMARKS ON THE SPECULATIONS OF MR. GODWIN, M. CONDORCET, AND OTHER WRITERS. LONDON, PRINTED FOR J. JOHNSON, IN ST. PAUL'S CHURCH-YARD, 1798. *St. Paul’s Church-yard, London*.

Martin, D. (1989). Mapping population data from zone centroid locations. Transactions of the Institute of British Geographers, 90-97.

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

Matthews, E. (1983). Global vegetation and land use: New high-resolution data bases for climate studies. *Journal of climate and applied Meteorology*, *22*(3), 474-487.

McKee, J. K. (2005). *Sparing nature: the conflict between human population growth and earth's biodiversity*. Rutgers University Press.

McMichael, A. J., Woodruff, R. E., & Hales, S. (2006). Climate change and human health: Present and future risks. *The Lancet, 367*(9513), 859-869. doi:10.1016/s0140-6736(06)68079-3

MDA Federal Inc. (2007). EarthSat GeoCover LC Overview. Retrieved October 15, 2013, from http://www.mdafederal.com/geocover/geocoverlc/gclcoverview

Meadows, D. H., Meadows, D. L., Randers, J., & Behrens III, W. W. (1972). *The limits to growth: a report to the club of Rome (1972)*. Universe Books, New York.

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

Meyer, W. B., & Turner, B. L. (1992). Human population growth and global land-use/cover change. *Annual review of ecology and systematics*, *23*(1), 39-61.

Misra, A. K., Lata, K., & Shukla, J. B. (2014). Effects of population and population pressure on forest resources and their conservation: a modeling study. *Environment, development and sustainability*, *16*(2), 361-374.

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

Moran, E. F. (2016). *People and nature: an introduction to human ecological relations*. John Wiley & Sons.

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? comparing data from Twitter's streaming api with Twitter's firehose. arXiv preprint arXiv:1306.5204.

Mortimore, M. (1989). *Adapting to drought: Farmers, famines and desertification in West Africa*. Cambridge University Press.

Muehrcke, P. (1972). Maps in geography. Cartographica: The International Journal for Geographic Information and Geovisualization, 18(2), 1-41.

National Research Council. (1990). Research strategies for the US global change research program. National Academies Press.

Nelson, A. (2004). African population database documentation, Retrived from http://na.unep.net/siouxfalls/globalpop/africa/Africa_index.html.

Nemani, R., Votava, P., Michaelis, A., Melton, F., & Milesi, C. (2011). Collaborative supercomputing for global change science. *Eos, Transactions American Geophysical Union*, *92*(13), 109-110.

Newell, N. D., & Marcus, L. (1987). Carbon dioxide and people. *Palaios*, 101-103.

NGA, 2005. Vector Map;1; (VMap) Level 0. National Geospatial-Intelligence Agency (NGA), Retrieved from http://geoengine.nga.mil/geospatial/SW_TOOLS/NIMAMUSE/webinter/rast_roam.html

Nieves, J. J., Tatem, A. J., Sorichetta, A., Bird, T., Stevens, F. R., Linard, C., & Gaughan, A. (2017, April 27). Globally Mapping Past and Future Settlements – Initial Outputs and Considerations [Scholarly project].

NOAA., 2012. VIIRS Nighttime Lights—2012. In: Earth Observation Group, National Geophysical Data Center. National Oceanic and Atmospheric Administration (NOAA), Retrieved from http://www.ngdc.noaa.gov/dmsp/data/viirs_fire/viirs_html/viirs_ntl.html.

OSM, (2014). OpenStreetMap Base Data. OpenStreetMap.org., Retrieved from http://www.openstreetmap.org/ (November 08th, 2014).

OSM., 2013. OpenStreetMap Base Data. OpenStreetMap.org., Retrieved from http://www.openstreetmap.org/ (October 15th, 2013).

Openshaw, S. (1983). The modifiable areal unit problem. Concepts and Techniques in Modern Geography, vol. 38. Norwich: Geobooks

Patel, N. N., Angiuli, E., Gamba, P., Gaughan, A**.,** Lisini, G., Stevens, F. R., Tatem, A. J., & Trianni, G. (2015). Multitemporal settlement and population mapping from Landsat using Google Earth Engine. *International Journal of Applied Earth Observation and Geoinformation*, *35*, 199-208. doi:10.1016/j.jag.2014.09.005

Patel, N. N., Stevens, F. R., Huang, Z., Gaughan, A. E., Elyazar, I., & Tatem, A. J. (2016). Improving Large Area Population Mapping Using Geotweet Densities. Transactions in GIS.

Pesaresi, M., Gerhardinger, A., & Kayitakire, F. (2008). A Robust Built-Up Area Presence Index by Anisotropic Rotation-Invariant Textural Measure. IEEE Journal of

Selected Topics in Applied Earth Observations and Remote Sensing, 1(3), 180-192. doi: 10.1109/JSTARS.2008.2002869

Petrov, A. (2012). One hundred years of dasymetric mapping: back to the origin. The Cartographic Journal, 49(3), 256-264.

Potere, D., Schneider, A., Angel, S., & Civco, D. (2009). Mapping urban areas on a global scale: Which of the eight maps now available is more accurate? International Journal of Remote Sensing, 30(24), 6531-6558. doi: 10.1080/01431160903121134

Ramankutty, N., & Foley, J. A. (1999). Estimating historical changes in global land cover: Croplands from 1700 to 1992. *Global biogeochemical cycles*, *13*(4), 997-1027.

Ramaswamy, L., P., D., Polavarapu, R., Gunasekera, K., Garg, D., Visweswariah, K., & Kalyanaraman, S. (2009). CAESAR: A Context-Aware, Social Recommender System for Low-End Mobile Devices. 2009 Tenth International Conference on Mobile Data Management: Systems, Services and Middleware. doi:10.1109/mdm.2009.66

Rango, A., Laliberte, A., Steele, C., Herrick, J. E., Bestelmeyer, B., Schmugge, T., Roanhorse, A. & Jenkins, V. (2006). Using unmanned aerial vehicles for rangelands: current applications and future potentials. *Environmental Practice*, *8*(03), 159-168.

Rasul, G., & Thapa, G. B. (2003). Shifting cultivation in the mountains of South and Southeast Asia: Regional patterns and factors influencing the change. *Land Degrad. Dev. Land Degradation & Development,14*(5), 495-508. doi:10.1002/ldr.570

Rawashdeh, S. A., & Saleh, B. (2006). Satellite Monitoring of Urban Spatial Growth in the Amman Area, Jordan. Journal of Urban Planning and Development, 132(4), 211. http://dx.doi.org/10.1061/(ASCE)0733-9488(2006)132:4(211).

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

Richards, J. F. (1990). Land transformation. *The earth as transformed by human action,* 163-178.

Running, S. W., Nemani, R. R., Heinsch, F. A., Zhao, M., Reeves, M., & Hashimoto, H. (2004). A Continuous Satellite-Derived Measure of Global Terrestrial Primary Production. *BioScience, 54*(6), 547. doi:10.1641/0006-3568(2004)054[0547:acsmog]2.0.co;2

Ruther, M., Leyk, S., & Buttenfield, B. P. (2015). Comparing the effects of an NLCD-derived dasymetric refinement on estimation accuracies for multiple areal interpolation methods. GIScience & Remote Sensing, 52(2), 158-178.

Sagan, C., Toon, O. B., & Pollack, J. B. (1979). Anthropogenic albedo changes and the earth's climate. *Science*, *206*(4425), 1363-1368.

Schneider, A., Woodcock, C.E., (2008). Compact, dispersed, fragmented, extensive? A comparison of urban growth in twenty-five global cities using remotely sensed data, pattern metrics and census information. Urban Studies. 45 (3), 659–692,http://dx.doi.org/10.1177/0042098007087340.

Schneider, A. (2012). Monitoring land cover change in urban and peri-urban areas using dense time stacks of Landsat satellite data and a data mining approach.

Remote Sensing of Environment, 124, 689-704. doi: 10.1016/j.rse.2012.06.006

Schneider, A., Friedl, M. A., & Potere, D. (2009). A new map of global urban extent from MODIS satellite data. *Environmental Research Letters, 4*(4), 044003. doi: 10.1088/1748-9326/4/4/044003

Schneider, A., Friedl, M. A., & Potere, D. (2010). Mapping global urban areas using MODIS 500-m data: New methods and datasets based on 'urban ecoregions'. *Remote Sensing of Environment, 114*(8), 1733-1746. doi: 10.1016/j.rse.2010.03.003

Schneider, A., Friedl, M. A., McIver, D. K., & Woodcock, C. E. (2003). Mapping urban areas by fusing multiple sources of coarse resolution remotely sensed data. Photogrammetric Engineering and Remote Sensing, 69, 1377 – 1386

Schroeder, J. P. (2007). Target-Density Weighting Interpolation and Uncertainty Evaluation for Temporal Analysis of Census Data. *Geographical Analysis*, *39*(3), 311-335.

Setälä, H., Bardgett, R. D., Birkhofer, K., Brady, M., Byrne, L., De Ruiter, P. C., De Vries, F.T., Gardi, C., Hedlund, K., Hemerik, L & Hotes, S. (2014). Urban and agricultural soils: conflicts and trade-offs in the optimization of ecosystem services. Urban Ecosystems, 17(1), 239-253.

Sexton, J. O., Song, X., Huang, C., Channan, S., Baker, M. E., & Townshend, J. R. (2013). Urban growth of the Washington, D.C.–Baltimore, MD metropolitan region from 1984 to 2010 by annual, Landsat-based estimates of impervious cover. Remote Sensing of Environment, 129, 42-53. doi: 10.1016/j.rse.2012.10.025

Salomonson, V. V., Barnes, W., Xiong, J., Kempler, S., & Masuoka, E. (2002, June). An overview of the Earth Observing System MODIS instrument and associated data systems performance. In *Geoscience and Remote Sensing Symposium, 2002. IGARSS'02. 2002 IEEE International* (Vol. 2, pp. 1174-1176). IEEE.

Schmitz, C., van Meijl, H., Kyle, P., Nelson, G. C., Fujimori, S., Gurgel, A., Havlik P., Heyhoe E., d'Croz D.M., Popp A. & Sands, R. (2014). Land-use change trajectories up to 2050: insights from a global agro-economic model comparison. *Agricultural Economics*, *45*(1), 69-84.

Small, C. (2004). Global population distribution and urban land use in geophysical parameter space. *Earth Interactions*, *8*(8), 1-18.

Sorichetta, A., Hornby, G. M., Stevens, F. R., Gaughan, A. E., Linard, C., & Tatem, A. J. (2015). High-resolution gridded population datasets for Latin America and the Caribbean in 2010, 2015, and 2020. *Scientific data*, *2*, 150045.

Sridharan, H., & Qiu, F. (2013). A Spatially Disaggregated Areal Interpolation Model Using Light Detection and Ranging-Derived Building Volumes. *Geographical Analysis*, *45*(3), 238-258.

Stevens, F. R., Gaughan, A. E., Linard, C., & Tatem, A. J. (2015). Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data. *PLoS ONE PLOS ONE, 10*(2). doi:10.1371/journal.pone.0107042

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

Sutton, P. (2003). A scale-adjusted measure of urban sprawl using nighttime satellite imagery. Remote Sensing of Environment, 86, 353 – 369.

Takhteyev, Y., Gruzd, A., & Wellman, B. (2012). Geography of Twitter networks. *Social networks*, *34*(1), 73-81.

Tatem, A. J. (2014). Mapping the denominator: spatial demography in the measurement of progress. International health, ihu057.

Tatem, A. J., Adamo, S., Bharti, N., Burgert, C. R., Castro, M., Dorelien, A., Fink, G., Linard, C., John, M., Montana, L., & Montgomery, M. R. (2012). Mapping populations at risk: improving spatial demographic data for infectious disease modeling and metric derivation. Population health metrics, 10(1), 1.

Tatem, A. J., Campiz, N., Gething, P. W., Snow, R. W., & Linard, C. (2011). The effects of spatial population dataset choice on estimates of population at risk of disease. Population Health Metrics, 9(1), 1.

Tatem, A., & Linard, C. (2011). Population mapping of poor countries. *Nature*, *474*(7349), 36-36. doi:10.1038/474036d

Tatem, A. J., Noor, A. M., Hagen, C. V., Gregorio, A. D., & Hay, S. I. (2007). High Resolution Population Maps for Low Income Nations: Combining Land Cover and Census in East Africa. PLoS ONE, 2(12). doi:10.1371/journal.pone.0001298

Taubenböck, H., Esch, T., Felbier, A., Wiesner, M., Roth, A., & Dech, S. (2012). Monitoring urbanization in mega cities from space. Remote Sensing of Environment, 117, 162-176. doi:10.1016/j.rse.2011.09.015

The WorldPop Project. (2014). Retrieved from http://www.worldpop.org.uk/

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

Tobler, W., Deichmann, U., Gottsegen, J., & Maloy, K. (1997) World population in a grid of spherical quadrilaterals. International Journal of Population Geography 3: 203–225.

Trianni, G., Lisini, G., Angiuli, E., Moreno, E. A., Dondi, P., Gaggia, A., & Gamba, P. (2015). Scaling up to national/regional urban extent mapping using Landsat data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *8*(7), 3710-3719.

Turner, B. L., Kasperson, R. E., Meyer, W. B., Dow, K. M., Golding, D., Kasperson, J. X., Mitchell, R. C., & Ratick, S. J. (1990). Two types of global environmental change: Definitional and spatial-scale issues in their human dimensions. *Global Environmental Change*, *1*(1), 14-22.

U.S. Geological Survey (2016). Landsat—Earth observation satellites (ver. 1.1, August 2016): U.S. Geological Survey Fact Sheet 2015–3081, 4 p., http://dx.doi.org/10.3133/fs20153081.

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

Van Zyl, J. J. (2001). The Shuttle Radar Topography Mission (SRTM): a breakthrough in remote sensing of topography. *Acta Astronautica*, *48*(5-12), 559-565.

Voss, P. R., Long, D. D., & 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.

Woodcock, C.E., Allen, R., Anderson, M., Belward, A., Bindschadler, R., Cohen, W., Gao, F., Goward, S.N., Helder, D., Helmer, E., Nemani, R., Oreopoulos, L., Schott, J., Thenkabail, P.S., Vermote, E.F., Vogelmann, J., Wulder, M.A., Wynne, R. (2008). Free Access to Landsat Imagery. *Science, 320*(5879), 1011a-1011a. doi: 10.1126/science.320.5879.1011a

Wilson, E. O. (1999). The diversity of life. WW Norton & Company.

World Population Prospects: The 2010 revision. (2011). New York: United Nations.

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

Wu, F., Li, Z., Lee, W. C., Wang, H., & Huang, Z. (2015). Semantic Annotation of Mobility Data using Social Media. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 1253-1263). International World Wide Web Conferences Steering Committee.

Wu, F., Wang, H., Li, Z., Lee, W. C., & Huang, Z. (2015). SemMobi: A Semantic Annotation System for Mobility Data. In *Proceedings of the 24th International Conference on World Wide Web Companion* (pp. 255-258). International World Wide Web Conferences Steering Committee.

Xia, H., Zhao, W., Li, A., Bian, J., & Zhang, Z. (2017). Subpixel Inundation Mapping Using Landsat-8 OLI and UAV Data for a Wetland Region on the Zoige Plateau, China. *Remote Sensing*, *9*(1), 31.

Xu, H. (2008). A new index for delineating built-up land features in satellite imagery. *International Journal of Remote Sensing, 29*(14), 4269-4276. doi: 10.1080/01431160802039957

Yuan, F., Sawaya, K.E., Loeffelholz, B.C., & Bauer, M.E. (2005). Land cover classification and change analysis of the twin cities (Minnesota) metropolitan area by multitemporal Landsat remote sensing. Remote Sensing of Environment. 98 (2–3), 317–328, http://dx.doi.org/10.1016/j.rse.2005.08.006

Zakrzewska, B. (1967). Trends and methods in land form geography. *Annals of the Association of American Geographers*, *57*(1), 128-165.

Zandbergen, P. A. (2011). Dasymetric mapping using high resolution address point datasets. *Transactions in GIS*, *15*(s1), 5-27.

BIOGRAPHY

Nirav Nikunj Patel was born in Tampa, Florida in the United States of America on June 8th, 1990. After graduating from Hillsborough High School's International Baccalaureate Program in May 2008, he earned his Bachelors of Arts in Geography and Philosophy in August 2011 at the University of Florida. In the same month of his undergraduate graduation he started the Master of Science in Geography at the same university. In his first year of his Masters he worked part-time at Nationwide Insurance as a Commercial Technical Clerk to support his education, and to save for his field research trip in the summer of 2012. With the mentorship of his then committee chair, Dr. Liang Mao, Nirav designed and collected data for his Master's thesis, and had the good fortune and honor to utilize the resources at his grandparents' non-governmental organization, the Akhand Jyot Foundation in Ahmedabad, India. During the summer of 2012, Nirav designed a method using GIS to measure spatial accessibility to treatments for HIV/AIDS and Tuberculosis in the Ahmedabad metropolitan area. While in India, a department e-mail circulated for an assistantship with funding from the World Bank through Dr. Andrew (Andy) Tatem. Nirav applied to it and received a source of funding for his second and last year of his Masters. Upon returning to the University of Florida in August 2012 after completing his field research, he started mapping urban change for 27 countries in the World Bank's East Asia and Pacific Division while completing his Master's thesis. He was awarded the Dr. Ryan Poehling Fellowship, awarded to students for their service to the department. With the guidance of Dr. Liang Mao, Dr. Timothy Fik and Dr. Peter Waylen, he successfully defended his Master's thesis on March 15th, 2013 and graduated in May. One month later, he started an internship at NASA's Langley Research Center in Hampton, Virginia, working with the NASA DEVELOP Program on two Earth science projects for the Republic of Rwanda, one focusing on mapping solar insolation and the other mapping soil degradation. At the end of this internship in August 2013, Nirav was informed that Dr. Andy Tatem and Dr. Paolo Gamba wanted to collaborate on a project using Google Earth Engine, and needed a representative from the WorldPop Project to work with Dr. Gamba at the University of Pavia in Italy. He took the opportunity and it gave him motivation to apply for the PhD at two programs in the DC metro area, after truly being immersed in the research lifestyle at NASA and in Italy. In spring 2014, Nirav applied to George Mason University and the University of Maryland, College Park. Without funding, his plan was to work full-time and do a PhD program part-time, and after applying to multiple jobs, he received and accepted an offer at

Dito as a GIS Engineer, while also accepting his admission to Mason, and filing away his rejection letter from Maryland (who were not keen on students that worked full time while doing their program). In his first year at Mason (starting Fall 2014), Nirav successfully built and directed a Google Maps business line at Dito while balancing school. In the summer of 2015, an opportunity arose for Nirav to move to Silicon Valley to further build Dito and start working as a contractor for NASA Ames Research Center. While completing classes and requirements at a distance, Nirav published two experiments featured in this dissertation and plans to publish the final experiment as a NASA Postdoctoral Fellow later on this year. He hopes to be the fastest part-time student to complete the Earth Systems and GeoInformation Science PhD at Mason in 2 years, 8 months and 25 days.