CHARACTERIZING THE SOCIOECONOMICS OF METROPOLITAN TRANSPORTATION NETWORK EXPANSION BY MINING A NATIONWIDE ROAD CHANGE DATABASE

by

CHRISTOPHER LOWELL BEAL BROWN

(Under the Direction of Xiaobai Yao)

ABSTRACT

Whereas previous transportation network growth research has focused on long-term historical changes across selected metropolitan areas, short-term comparative explorations—especially those considering all levels of the nationwide road hierarchy—have been neglected. Therefore, a comprehensive road change database of the United States was developed through compilation of U.S. Census Bureau TIGER/Line datasets from 2008 to 2012, while annual and five-year extents of road change were derived using Python geoprocessing scripts. Aggregate percentage road change statistics were presented for each metropolitan statistical area, county, and census tract nationwide, and the data were found to exhibit moderate spatial autocorrelation. Exploratory multiple linear and geographically weighted regression models indicated that the primary mechanisms of change were income and regional differences, while the importance of population change and race increased with finer spatial resolution. Counties surrounding Atlanta, Georgia produced highly statistically significant outliers, suggesting that anomalous expansion processes have uniquely shaped this metropolitan area.

INDEX WORDS: Atlanta, Linear Change Detection, Road Network Expansion, Spatiotemporal Road Database, TIGER/Line, Urban Growth

CHARACTERIZING THE SOCIOECONOMICS OF METROPOLITAN TRANSPORTATION NETWORK EXPANSION BY MINING A NATIONWIDE ROAD CHANGE DATABASE

by

CHRISTOPHER LOWELL BEAL BROWN

B.S., The University of Alabama, 2010

A Thesis Submitted to the Graduate Faculty of The University of Georgia

in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2013

CHARACTERIZING THE SOCIOECONOMICS OF METROPOLITAN TRANSPORTATION NETWORK EXPANSION BY MINING A NATIONWIDE ROAD CHANGE DATABASE

by

CHRISTOPHER LOWELL BEAL BROWN

Major Professor: Xiaobai Yao

Committee: Steven Holloway
Thomas Jordan

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2013

ACKNOWLEDGEMENTS

I would first like to thank Dr. Joe Weber of the University of Alabama for his inspiration during my undergraduate education; without his support and encouragement, I would never have pursued graduate research on transportation networks. I am likewise grateful to Dr. Xiaobai Yao for discussing and nurturing my ideas throughout the development of this document, and also for her generosity of time and interest in all my aspirations—scholastic, personal, and professional. I further appreciate the constructive feedback of my committee members, Dr. Steve Holloway and Dr. Tommy Jordan, both of whom have administered counsel that improved the quality of my logic and the presentation of my results. I especially thank my parents, Jim and Patty Brown, for financing my workstation hardware and providing needed reassurance; and finally, my girlfriend and partner in discovery, Stephanie Wright, for her kindness, patience, and enduring love.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

Overview

The growth of metropolitan road networks in the United States, although often indirectly referenced in academic literature, has been largely neglected as a primary object of scholarship; according to a recent review of transportation growth modeling (Xie and Levinson 2009), only a handful of empirical studies have investigated temporal change in transportation supply. The authors attribute the current lack of statistical analysis and spatiotemporal modeling to a dearth of computer processing power and historical data; however, while these considerations may apply to complex modeling prescriptions and temporal ranges exceeding the maturity of geographic information systems (GIS) technology, they need not preclude future research. As microprocessor densities have increased exponentially according to Moore's law for the past four decades (Moore 1998) and recent historical road network data have been continually archived by both the United States Bureau of the Census and several metropolitan areas, the impediments to scholarship on comparative metropolitan road network growth are steadily being deconstructed.

Enabled particularly by the public-sector advancement of longitudinal GIS data provisioning, this investigation presents a novel analysis of comparative transportation network growth across the Census Bureau metropolitan statistical areas (MSAs) between the years 2008 and 2012. A comprehensive discussion of analytical feasibility addresses several attendant considerations for future transportation growth research, including positional/temporal accuracy, database creation, change detection procedures, and logical consistency. However, the ultimate objectives are to produce descriptive statistics of road change, identify anomalous data regions, and determine the correlative role of socioeconomic conditions with areas of network expansion.

The construction of a nationwide spatiotemporal road network database opens several new avenues of scientific inquiry on the nature and impacts of short-term road network change. As the seminal "ideal-typical sequence" of network expansion theorizes, transportation network growth is powered by discrete external socioeconomic and political factors, which are in turn the key drivers of spatial road diffusion (Taaffe, Morrill, and Gould 1963). In the economic context of the United States, transportation network expansion may be understood as a top-down "spatial fix" for historic overaccumulation crises in the northeastern manufacturing belt, given that the majority of new roads are being constructed in more economically viable "sun belt" regions (Harvey 1985; Harvey 2001). A comparative study of temporal road change across metropolitan areas, enabled by the new road network database, could provide empirical justification for a causal link between metropolitan economic development and new transportation infrastructure; if a temporally lagged structure could be established, it would be possible to "predict" new roads.

Furthermore, given the annual temporal resolution of the database and the data window from 2008 to 2012, the comparative spatial effects of the subprime mortgage crisis could be analyzed over time; because the vast majority of new road mileage in the United States is not composed of major arterials, but local residential subdivision roads on privately-owned land, one could determine whether new roads (as a proxy for economic growth) would strongly correlate with the ability of a metropolis to recover from its stunted spatial growth patterns. As pictured in Figure 1.1, the Atlanta MSA, generally regarded in the urban planning community as a prime example of low-density residential sprawl, has experienced a significant downturn in new road mileage as stakeholders have divested themselves from proposed residential developments. The spatial recovery process could easily be different for each metropolitan area, and socioeconomic trends might also be extracted: for example, would heavily Hispanic counties recover faster?

Therefore, in addition to lending hard evidence for broad-scale, geographically-inclined theories, a scholar could easily utilize descriptive road change statistics to support hypotheses: in order to determine the most relevant predictors for metropolitan road network change, one might perform a regression analysis with areal transportation growth as a dependent variable. The spatiotemporal road network database, then, could be used to develop a new explanatory model with the potential for multi-scalar hierarchical modeling, including variables such as population growth, population density, per capita gross domestic product, income, race, and topographic range, while also ideally incorporating measures of spatial autocorrelation and temporal lag. Although there could be several more applications for the spatiotemporal database, most notably for supporting initiatives against habitat fragmentation and water pollution due to runoff, the key contributions of this thesis are primarily methodological and expository in nature. By expanding an extant hole in the literature and drawing attention to the breadth of possibilities engendered by

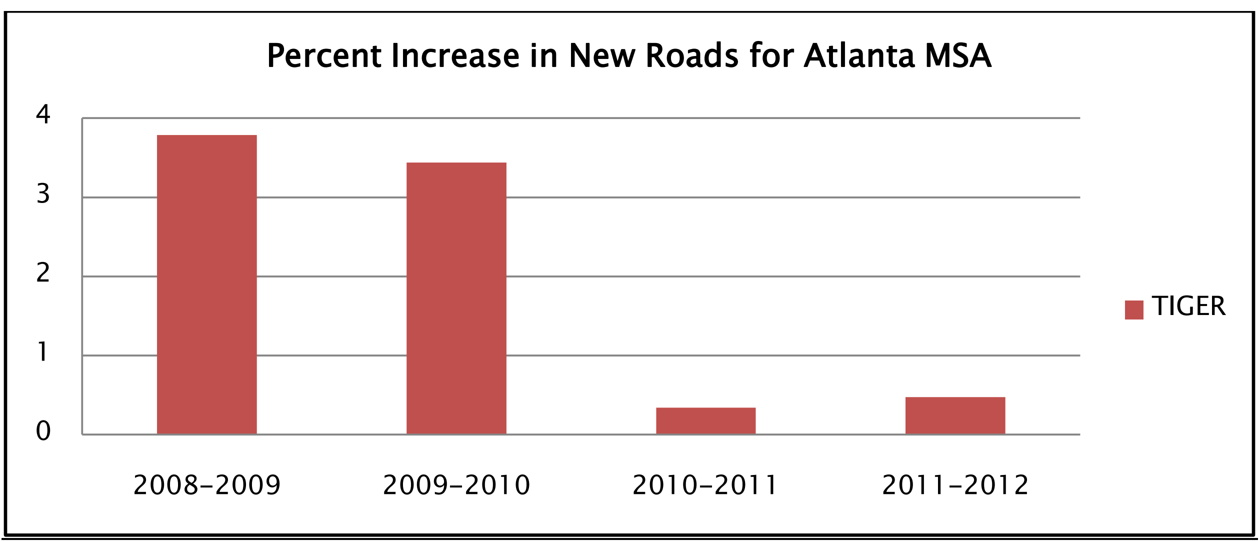


Figure 1.1: Percent Increase in New Road Length for the Atlanta MSA, U.S. Census TIGER

a comprehensive nationwide spatiotemporal GIS road database, interest should be instigated in the academic community for further scholarship on the causative processes and environmental effects of new road construction in the United States.

Research Objectives

The first two objectives of the research are to assess the spatiotemporal accuracy of the database and to develop a programming framework for detecting changes in the road network. As the ultimate value of any created database is dependent upon the characteristics of the input data sources, one must first verify data integrity: therefore, several different geospatial data programs are subjected to a suitability analysis for inclusion in the database. Additionally, the spatial and temporal accuracy of the selected data sources is elucidated, and the entire temporal window is examined for logical consistency and sources of error propagation. Furthermore, the preprocessing steps required to acquire and vet the raw data are furnished, along with a logical presentation of the robust Python geoprocessing framework for identifying new road segments.

The final objective will be to perform exploratory data analyses, including descriptive statistics of road change, spatial autocorrelation statistics, and socioeconomic data associations. The distillation of the voluminous raw GIS data into statistics such as road mileage change or road density change enables outlier detection and investigation, while more functionally complex descriptive statistics, such as the range of percent change across each of the four time intervals, may yield more pertinent information (by proxy) on the comparative spatial curtailing effects of the subprime mortgage crisis across all MSAs. Likewise, autocorrelation statistics such as local Moran's I are utilized to identify spatial clustering of road network changes. Finally, preliminary associative regression analyses seek to determine the suitability of various socioeconomic characteristics and geographical considerations for inclusion in future road change modeling.

CHAPTER 2

LITERATURE

The following conceptualization of comparative transportation network growth differs from previous analyses due to its nationwide scope, metropolitan area scale, and temporal resolution, requiring data assimilation across the entire United States and over a time period of five years—a volume of raw, intermediate, and processed results exceeding one terabyte of data. Although the theoretical construction of a comprehensive spatiotemporal road database may be largely self-evident, the nontrivial effort required for data acquisition and change detection, combined with limited information on accurate, complete, and regularly updated GIS data sources, has appeared to construct a strong barrier of entry to the daunting task. Furthermore, after amassing the data, one must determine how to extract new linear road features given positional or temporal uncertainty; this methodological consideration has likely discouraged many scholars from undertaking the effort to compile, organize, and extract spatiotemporal GIS road data for further examination. However, several types of transportation literature closely approximate different components of the national transportation network growth process, yet the applicability of these analyses are universally limited by scale, scope, or temporal situation. The ideal database, therefore, addresses these structural shortcomings with three characteristics: complete road hierarchy, nationwide coverage area, and temporal network archival (Table 2.1).

The first three rows of the table are unsatisfactory for comprehensive network growth analysis because they only fill one of the three requirements; although many other avenues of research may be substituted for these categories, the listed examples are sufficient to express

Table 2.1: Example Taxonomy of Road Databases in Transportation Literature

	Complete	**Nationwide**	**Temporal**
Space Syntax	X		
Economic Transport		X	
Historic Transport			X
Road Ecology	X	X	
National Planning		X	X
Regional Planning	X		X
Comprehensive Database	**X**	**X**	**X**

their respective deficiencies. Space syntax, an outgrowth of computational architecture that considers the spatial trajectories of network segments and their connectivity, places particular emphasis on accurate fine-scale network representation: however, the drive for accuracy generally comes at the expense of spatial and temporal scope (Kim 2007; Dalton, Peponis, and Conroy-Dalton 2003). Although one comparative study of urban morphology was discovered, articles of this breadth are uncommon (Peponis et al. 2007). The second type of study focuses on comparative nationwide transportation, yet often utilizes national GIS data sources such as the National Transportation Atlas Database, which excludes data on local roads; these papers may incorporate economic statistics to support hypotheses on interstate commerce (Ham, Kim, and Boyce 2005). Some articles may have a temporal component as well, approaching the level of national freight transportation planning, but these usually limit the analysis to a regional scope due to the large volume of data required (Healey and Stamp 2000). The third type of study may describe the evolution of historical transportation networks, but tends to either focus on one metropolitan area or present a general survey of the nation. As much of the data acquired on historic road networks is derived from temporally inconsistent sources such as aerial/satellite imagery or digitized historical maps, historical transportation research has more difficulty in drawing analytical conclusions based on time series, instead preferring a more theoretical or descriptive approach. Furthermore, a considerable portion of historical articles are not explicitly concerned with changes over time; instead, historical road data are often used to observe a particular time period, rather than engage in temporal comparison (Gregory and Healey 2007).

The bottom three rows in Table 1 fulfill two of the three requirements, and thus these categories of transportation research more closely approximate questions of greater scope. The most logical extensions of the first three types are regional and national transportation planning, which simply aggrandize the heretofore opposing concepts of complete, individual metropolitan network coverage and broad, hierarchically-restricted national coverage with temporal density. Although many urban growth models, which are verified according to their ability to reproduce historical conditions at a given time, only consider major roads in a single metropolitan area (Clarke and Gaydos 1998; Levinson and Karamalaputi 2003; Hu and Lo 2007), several cities now provide complete GIS road archives. Therefore, regional planning literature now has the ability to incorporate temporal change, yet the different database schemata, record attributes, and data assimilation methods complicate comparative metropolitan analyses.

These advancements in metropolitan data provisioning should augment research on the effects of urban sprawl (Kenworthy and Laube 1999; Ciscel 2001), as well as inform concerns of systemic racism (Jaret, Adelman, and Reid 2006), inequity of transportation systems and services (Levinson 2002; Bullard 2003; Sanchez, Stolz, and Ma 2003) and job accessibility (Henderson 2004; Weber and Sultana 2007); introducing temporal dimension to these generally static areas of study could produce results detailing if inequitable spatial conditions are being ameliorated. Likewise, national transportation planning has benefited, and will continue to reap insight, from the inclusion of temporal data in economic freight analyses, although unlike air and sea routes, the locations of domestic ground transportation routes remain relatively constant (Feyrer 2009).

The final, likely most specialized field of study, road ecology, assesses the ecological effects of the road network on various species, and may also use landscape ecology metrics to analyze habitat patches outlined by roads and their buffer zones (Forman and Alexander 1998). Although road ecology research has not yet embraced temporal road data, it has presented one static study that includes all roads in the United States (Riitters and Wickham 2003). Unlike any of the aforementioned scholarship, the article compiles all of the county-based road networks in the commercial Tele Atlas/GDT Dynamap/2000 GIS dataset, making this road ecology analysis one of the first to utilize and process such a large linear road representation. Although the information contained in the database was eventually converted into a raster, or cell-based, storage method, the utilization of all available roads in the country, including trails for off-road vehicles, was a novel idea from inception, and could only be achieved through GIS technology.

In the spirit of analyzing all available road data, the following research aims to retain the original linear nature of the road centerlines, while including five years of annual revisions for the nationwide network; additionally, unlike the Riitters and Wickham (2003) paper, the raw data were subjected to spatiotemporal accuracy assessment. Furthermore, the procedures for creating a functional GIS database out of these different temporal layers, and the Python programming routines utilized to extract information about road network change over time, may inform future temporal studies with nationwide scope and/or complete road networks. In order to demonstrate the feasible applications of extending this line of short-term temporal transportation research, three analytical examples will be discussed: descriptive statistics, spatial pattern analysis, and exploratory correlation/regression analyses for future modeling. Both methodological improvements and expository statistics will be necessary to track the human drivers of transportation network expansion; in an era where new roads are being funded before old ones can be maintained, the role of socioeconomic conditions that spur unchecked road network growth deserves further examination in the literature.

CHAPTER 3

METHODOLOGY

Data

Before the advent of GIS technology, individual metropolitan areas did not have a common spatial framework for storing and accessing geographic coordinates of road network representations. Although public-sector GIS data acquisition has matured over the past three decades, differences in accuracy, completeness, boundary definitions, attribute categories, and implementation have discouraged nationwide, comparative studies of metropolitan transportation network expansion; therefore, the minimal interoperability between different data sources has largely resulted in either provincial studies addressing a particular region or nationwide studies of minimal detail. Whereas ideally, a compilation of all the most accurate metropolitan data sources, supplemented by the best available data in rural areas, would produce the most spatially accurate database, data assimilation and integration would prove to be a daunting task across the entire United States, especially considering the increasing number of rural GIS departments. Therefore, when considering a comparative metropolitan area study, the selection of an adequate spatiotemporal GIS data source should be limited to national, standardized databases exhibiting four attributes: logical consistency, positional accuracy, temporal accuracy, and completeness.

Four freely available national data projects were evaluated, including efforts from the United States Geological Survey (USGS), the United States Department of Transportation (USDOT), the United States Bureau of the Census, and the crowdsourced OpenStreetMap project. The USGS Digital Line Graph was the least suitable for the proposed database: temporal updates were completed after a period of around twenty years, and furthermore, revision years were not standardized across the country, obfuscating any possible comparative conclusions. Additionally, the boundaries of the datasets do not coincide with any political boundaries, and the vast majority of the included road features are digitized at a scale of 1:100,000, which would prove to be quite spatially inaccurate for analyzing individual urban or suburban road change. Likewise, the USDOT National Transportation Atlas Database (NTAD) also had scale issues; rather than having problems arise from the data acquisition scale, though, the NTAD dataset purposefully excludes many local roads, choosing to emphasize national, state, and arterial roads instead. Despite its flaws in hierarchical completeness, the USDOT dataset features annual road updates, making it highly suitable for national-scale temporal analyses; however, the vast majority of data in the NTAD is derived from a third-party source: the national census bureau.

Although it would appear that the national and state departments of transportation should have the most complete data at the metropolitan scale, the United States Census Bureau dataset retains the same annual temporal resolution as NTAD, yet is complete and logically consistent. The MAF/TIGER (Master Address File / Topologically Integrated Geographic Encoding and Referencing) road datasets were originally developed to support internal geocoding operations for the decennial census; therefore, all roads are included with standardized attribute information. Another preferred data provider could be OpenStreetMap, a free and open source worldwide road database originating from the United Kingdom; however, while excellent for countries where GIS data are restricted or commercialized, OpenStreetMap data in the United States, like the NTAD datasets, are largely derived from the United States Census Bureau MAF/TIGER roads. Moreover, the datasets do not follow the de facto shapefile standard for vector GIS data, and temporal extraction cannot be readily performed on the raw data, particularly due to the difficulty of creating and then processing a historical "snapshot" of the database features. Finally, the crowdsourced nature of the database, much like Wikipedia, allows anyone to add or edit road features; consequently, this characteristic potentially allows spatiotemporal accuracies to vary widely across the nation. Therefore, due to geographically uncertain accuracy and possible implementation issues, the OpenStreetMap database was rejected in favor of the Census Bureau MAF/TIGER datasets for inclusion in the nationwide spatiotemporal road network database.

Spatiotemporal Accuracy Assessment

The United States Census Bureau MAF/TIGER database has traditionally been eschewed for its low positional accuracy. Originally, for the year 1992, features were digitized from 1:100,000 scale maps, and were assumed (with perfect digitizing) to meet established National Map Accuracy standards of 50 meters. However, in preparation for the decennial Census 2000, the database was updated by several disparate sources, each with inherently different error tolerances; many of these updates used the 1992 data as a reference, ultimately exacerbating the original error tolerances and compromising the 50-meter accuracy specification (see Figure 3.1). Upon realizing the severity of the error propagation in the 2000-vintage data, the U.S. Census began the MAF/TIGER Accuracy Improvement Project (MTAIP) in 2002, which sought to reduce the positional errors plaguing the dataset (Broome and Godwin 2003). By the 2008

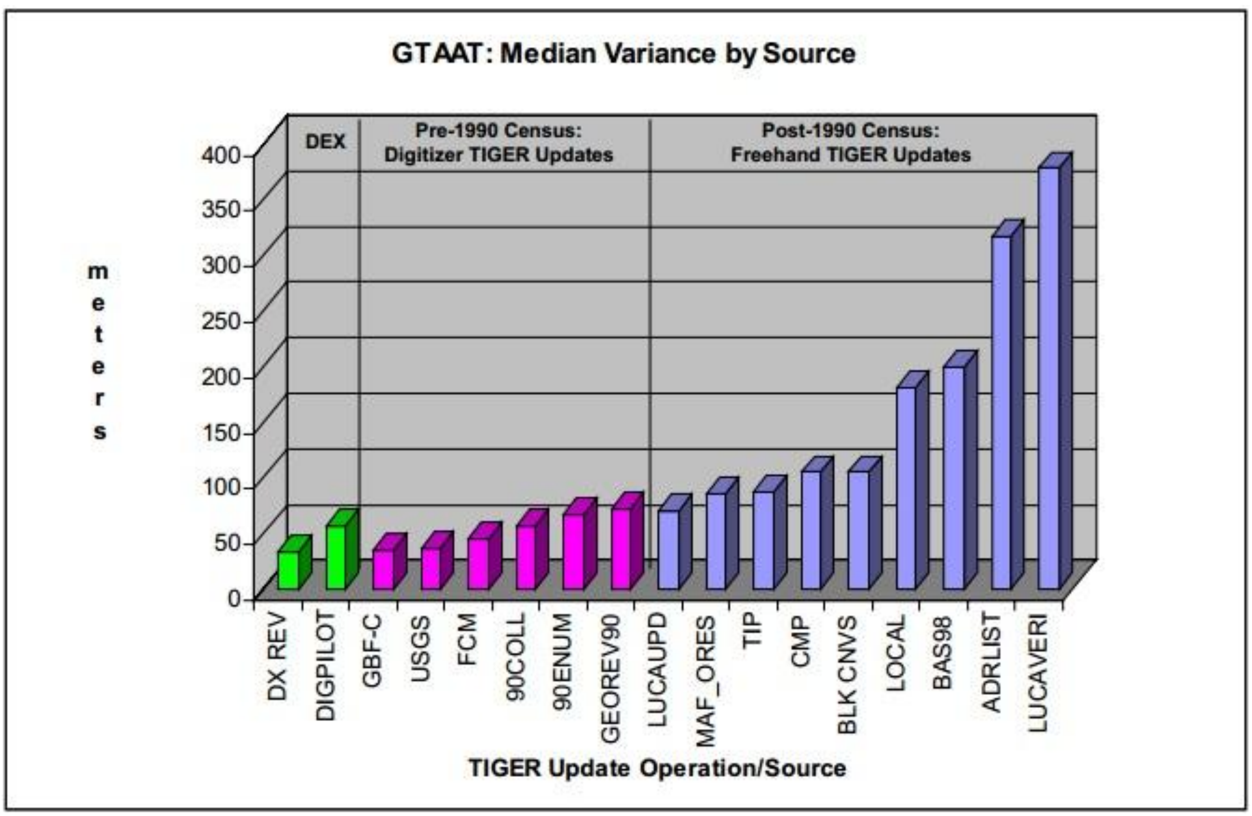


Figure 3.1: Median Variance by TIGER Update Source for 2000 Census (Liadis 2000)

edition of the MAF/TIGER files, the database was declared to meet the National Standard for Spatial Data Accuracy (NSSDA) Circular Error 95 standard of 7.6 meters, meaning that 95% of coordinates were located within a 7.6-meter radius circle around the actual location. However, the methodologies and results of this accuracy assessment were not independently verified.

Therefore, a positional accuracy assessment of the 2009 TIGER/Line roads data was performed by Zandbergen (2011); from a sample of three moderately urbanized counties, the study concluded that none met the specified NSSDA standard, citing that the long tails of the positional error distributions exerted great influence on the final 95th percentile values (Zandbergen 2008). However, as with the in-house accuracy assessment performed by Census Bureau contractors, the ground control coordinates were only located at road intersections, not randomly spread across the linear features of the network, implying different accuracies for locations between the intersection points; furthermore, the improved MAF/TIGER dataset was not tested for any other pertinent measures of accuracy. In order to improve upon the analysis of Zandbergen and to gain a better understanding of the error limitations of MAF/TIGER dataset—specifically the TIGER/Line streets dataset—positional accuracy assessments were performed, including both absolute and relative accuracies, and the datasets were also subjected to testing for completeness and temporal accuracy, using both local GIS data and reference aerial imagery.

Study Area

The Atlanta-Sandy Springs-Marietta Metropolitan Statistical Area (Atlanta MSA) was chosen for quantifying the spatiotemporal accuracies of the TIGER/Line street datasets. The majority of the MSA exhibits a highly curvilinear road network, making GIS road datasets particularly susceptible to linear representational error; therefore, the Atlanta MSA typifies a "worst-case" scenario for minimizing positional accuracies along its linear features (see Figure

3.2). Furthermore, the area has experienced explosive growth throughout the past two decades, with much of its growth attributable to subdivisions of single-family homes; the rapid construction of roads to support these residential developments underscores the importance of temporal accuracy assessment. By providing a challenging study area for the spatiotemporal accuracies of GIS databases, conservative accuracy assessments can be achieved, maximizing the generalization of these findings to the broader MAF/TIGER dataset and not simply limiting the applicability of the quantitative assessments to the case of the Atlanta MSA.

Overview

In order to perform the spatiotemporal accuracy assessment, TIGER/Line street data were downloaded from the U.S. Census Bureau (2012b) via file transfer protocol (FTP) for the counties of the Atlanta MSA from 2008 to 2012. These TIGER/Line files were overlaid on top of a USDA (2010) National Agriculture Imagery Program (NAIP) mosaic for the state of Georgia,

Figure 3.2: Sample Road Network from the Atlanta MSA (USDA 2010)

enabling comparison of the GIS data with the 1-meter resolution aerial imagery. Because road centerlines could easily be resolved with the high-resolution mosaic, it served as the reference data source for accuracy assessment, and was assumed to reflect "ground truth" for the year 2010; although small geo-referencing errors are possible when using ortho-rectified aerial imagery, this insubstantial source of non-systematic error would require extensive ground surveying to ameliorate—a process which must be conceded when working with larger geographic areas. This combination of GIS datasets and imagery enabled planar measurements of absolute positional accuracy, replicating the earlier work of Zandbergen.

To determine the relative positional accuracy of the TIGER/Line streets compared to local data, absolute positional accuracies were also computed from Georgia Department of Transportation (GDOT) GIS data for the Atlanta MSA, retrieved from the Georgia GIS Clearinghouse (2012). The GDOT GIS data were digitized from source USGS Digital Orthophoto Quarter Quadrangles (DOQQs) at a scale of 1:12000, but had no official statement of positional accuracy; furthermore, the data were not updated annually, like the TIGER/Line data, but sporadically for each county in the Atlanta MSA (see Figure 3.3). In keeping with the temporal extent of the study, GDOT road datasets that were last updated before 2008 were excluded from the analysis; although several counties with high growth rates have been excluded, the remaining counties provided a representative cross-section of urbanized areas and county growth rates.

For each temporal period, relative completeness was discovered using a process of buffer thresholding: the buffer around the control lines (the base GDOT network dataset, assumed to be more complete) was expanded until 95 percent of the length of the TIGER/Line dataset fell within the buffer zone. The buffer radius, therefore, represented the actual positional x-displacement error throughout the linear dataset, yielding a much more useful assessment that revealed not just the accuracy of sample points, but the relative accuracy of the entire dataset. Finally, in order to determine relative temporal accuracy, both TIGER and GDOT data sources from different years were compared to the 2010 NAIP imagery, enabling evaluation of temporal mismatch for both datasets.

This combination of data allowed for three questions to be answered: whether the TIGER/Line positional accuracy met the NSSDA standard, whether the local GDOT roads are more accurate or complete than the national TIGER/Line dataset, and finally, whether the temporal accuracy is greater for GDOT or TIGER/Line datasets. In the larger scheme of road change detection, the accuracy of any descriptive road change statistic is largely dependent

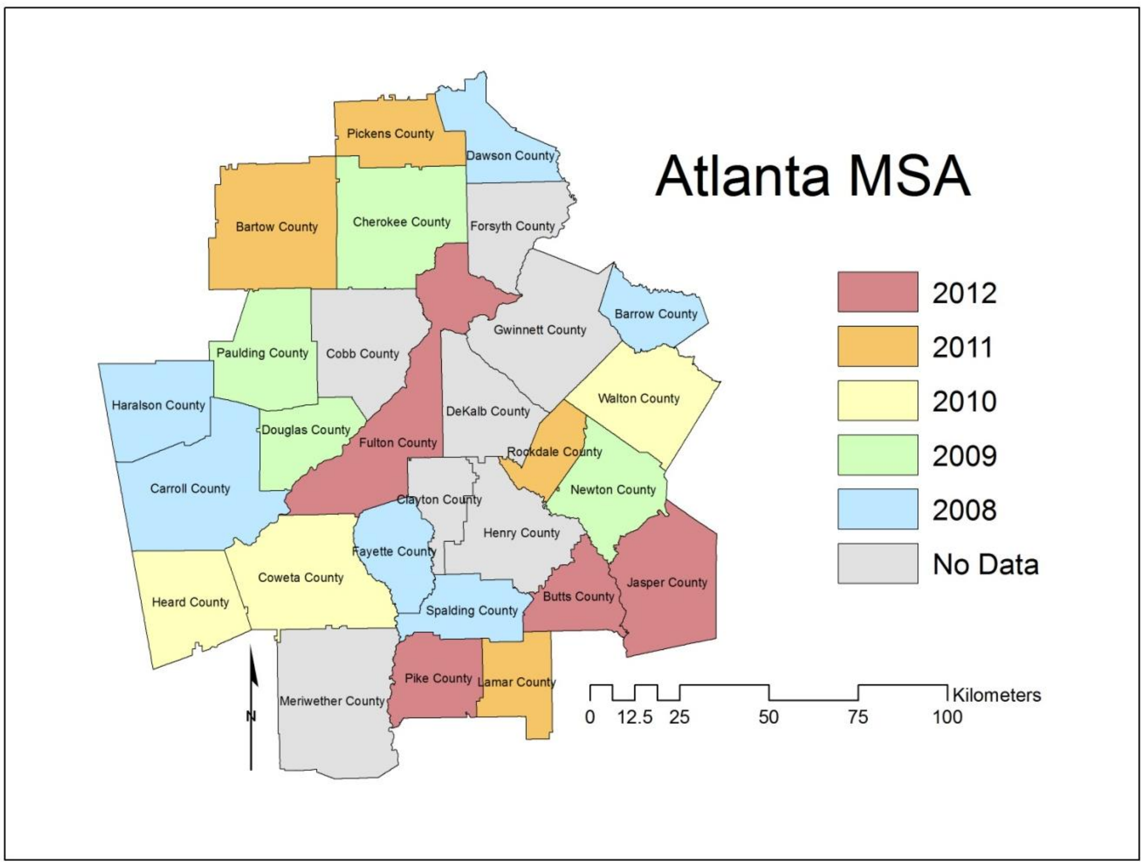


Figure 3.3: GIS Data Availability for Georgia Department of Transportation

upon an accurate conceptualization of the input data accuracy; given that TIGER/Line data have not yet been explicitly used for change detection, it is important to understand its limitations.

Positional Accuracy

In order to perform the absolute accuracy assessments, a set of sample points was randomly distributed along the length of the roads and stratified by GDOT county update year. Each of the five county groupings received thirty random point samples, resulting in one hundred fifty total point samples for determining positional accuracies (see Figure 3.4). Each of these sample points was then compared to the reference NAIP mosaic using the measurement tool in ArcGIS, resulting in an x-displacement value measured perpendicular to the tangent line of the nearest road. From this point of tangency, another measurement was recorded to the nearest

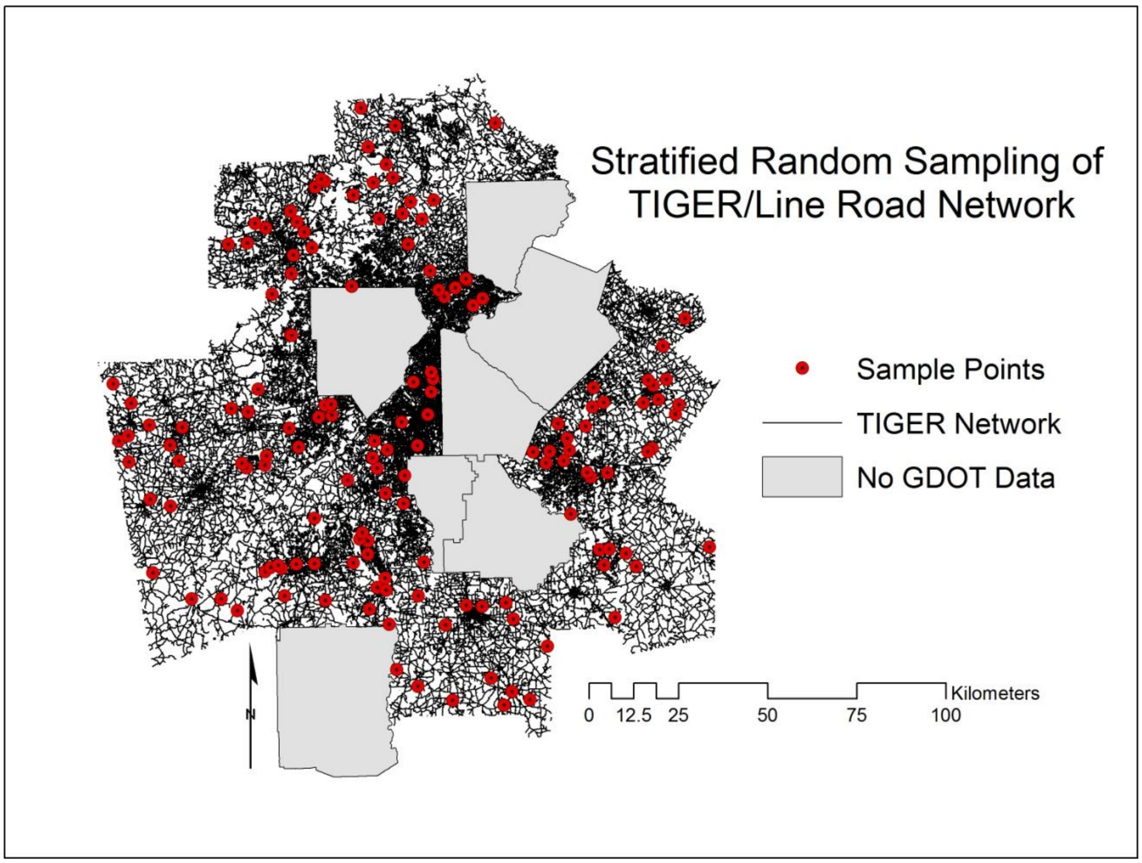


Figure 3.4: Sample Locations for Positional Accuracy Assessment

GDOT road, producing paired x-displacement observations for both TIGER and GDOT streets. Because the sampling was stratified by GDOT update year—the limiting factor in comparative analysis—the TIGER/Line data could be compared directly to GDOT data of the same vintage.

However, due to the relatively low sample size per year, the displacement measurements were combined across the five-year period, resulting in one hundred fifty observations for both datasets. From this list of paired error values for each sample point, a histogram was constructed to demonstrate the relationships between the national TIGER/Line data and the local GDOT data. The graph in Figure 3.5 reflects the expectation of greater accuracy for the local data; while the TIGER/Line distribution is relatively flat, the GDOT distribution has a pronounced peak at a 1m displacement, indicating that more of the locally digitized streets fell within a one meter tolerance of the NAIP road centerline—the same as the resolution of the imagery. Conversely, between four and ten meters, TIGER/Line data have a higher error frequency; additionally, as

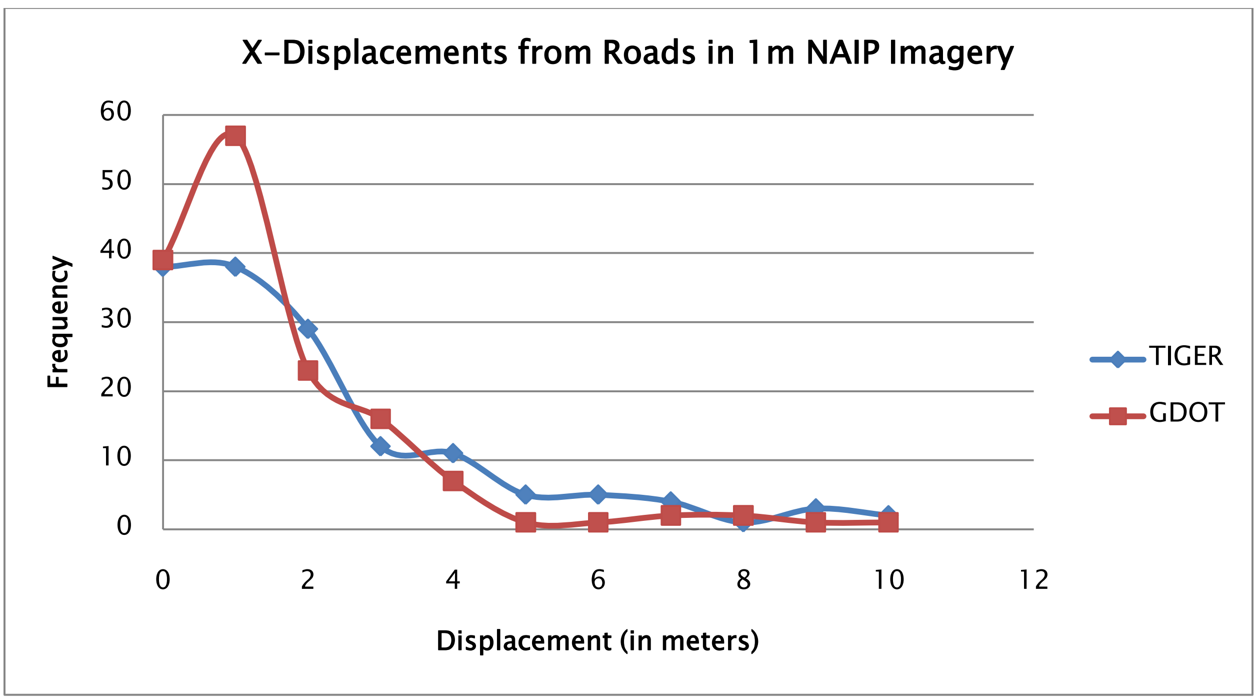


Figure 3.5: Comparative Displacement Error Distributions, Accuracies Truncated at 10m

Zandbergen (2011) notes, the TIGER streets are especially prone to outliers due to their previously erratic and flagrant feature updates. When including the two outliers from the 150-point error sample and assuming a normal error distribution, one can easily see that the positional reliability of the TIGER/Line dataset is highly dependent on which outliers are sampled, especially using metrics such as the NSSDA 95th percentile method (Figure 3.6). However, despite the long tail of the TIGER/Line error distribution, the overall 95th percentile statistic was found to just meet the NSSDA standard of 7.6 meters (Table 3.1). The findings contradict the independent Zandbergen (2011) assessment, yet support the U.S. Census Bureau evaluation of the TIGER/Line positional accuracy; furthermore, the distribution indicates that the Atlanta MSA is a good representative sample of the national dataset as a whole. Therefore, although the sample size may not reveal outlier specificity as well as expected, we may still conclude, at a cursory glance, that the TIGER/Line dataset may be deemed of sufficient positional accuracy for

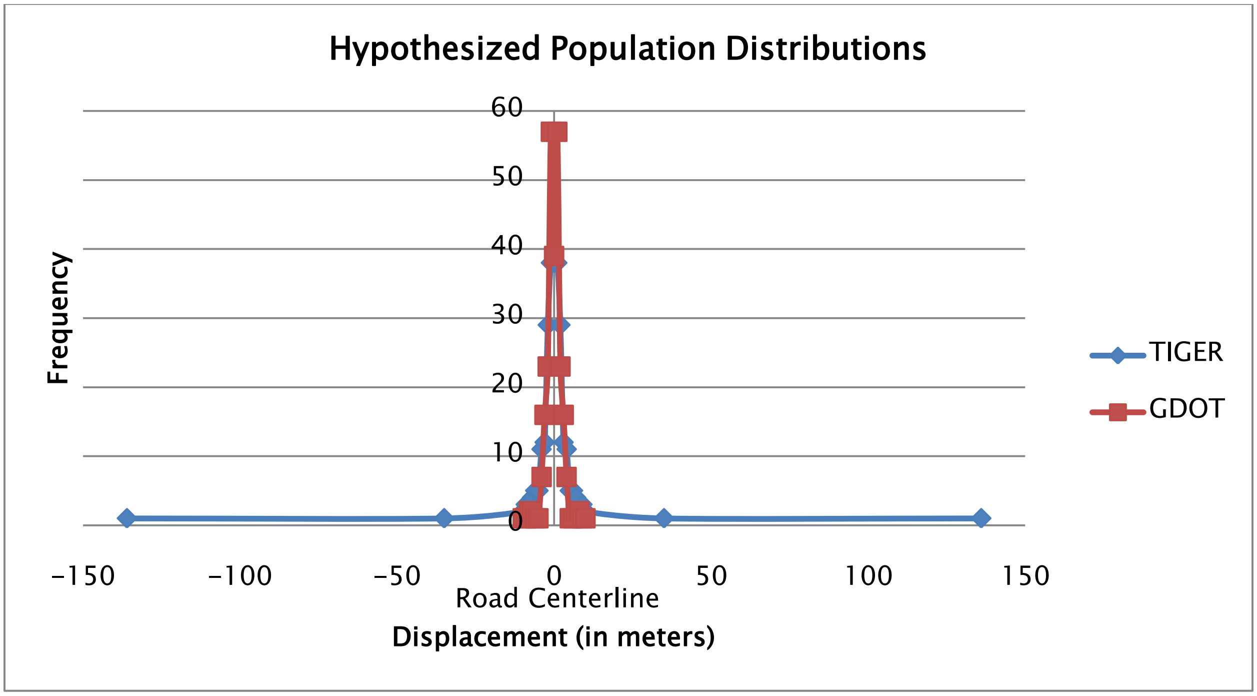


Figure 3.6: Mirrored Displacement Error Distributions

Table 3.1: Circular Accuracy Percentiles for Displacement Error Distributions

Percentile	GDOT	TIGER
99.9th	9.85	120.95
99th	8.51	22.75
95th	4.55	7.55

road change detection. Furthermore, because a liberal sample retained the expected population accuracy, it is geographically stable enough to elicit comparisons in accuracies across the nation.

Completeness Accuracy

The relative completeness accuracy assessment operated upon the assumption that the extent of the local GDOT streets was more accurate than the TIGER/Line streets, given that local knowledge should better inform accurate road locations (Chen, Knoblock, and Shahabi 2006). First, to make the datasets readily comparable, the TIGER streets were preprocessed to remove features that were not strictly roads; because the GDOT would only be concerned about including roads in their dataset, all inapplicable MAF/TIGER Feature Class Codes were automatically removed from the attribute tables (Table 3.2). Once both datasets represented the same informational classes, the TIGER/Line dataset was evaluated for both errors of omission (type II error) and commission (type I error).

To determine errors of omission, a relative completeness accuracy assessment was performed using the concept of epsilon bands (Perkal 1966), in which the TIGER/Line dataset was buffered with its inherent error; in this case, the positional error buffer was set to 25 meters, a value slightly above the 99th percentile of the TIGER/Line error distribution—this offset accounted for the aforementioned errors in rectification. The overall completeness was calculated as the GDOT length intersecting the TIGER/Line buffer divided by the total GDOT length, as shown in Table 3.3 (Goodchild and Hunter 1997). By providing a measure of completeness that

Table 3.2: MAF/TIGER Feature Class Codes Removed for Preprocessing

MTFCC Code	Description
S1500	Vehicular Trail
S1710	Walkway/Pedestrian Trail
S1720	Stairway
S1740	Private Road for Service Vehicles
S1750	Internal U.S. Census Bureau Use
S1780	Parking Lot Road
S1820	Bike Path or Trail
S1830	Bridle Path
S2000	Road Median

operates independently of positional error, the buffer intersection method yielded a much more useful completeness metric for comparing the two road datasets.

To determine errors of commission, the buffer intersection process was reversed; the GDOT dataset was buffered by 10 meters, representing the 100th percentile of the sample error distribution. Next, the total TIGER/Line length was divided by the TIGER/Line length intersecting the GDOT buffer, and 1 was subtracted from the quotient; this yielded a commission error in terms of percent increase in road length, with the GDOT data serving as the reference. However, the errors of commission appeared unacceptable, with more than 20% of the total TIGER/Line length perceived as erroneous, as shown in Table 3.4. Upon investigation of the areas of the TIGER/Line dataset deemed to be superfluous using the 2010 aerial imagery (for 2008-2010) and Google Maps (for 2010-2012), a considerable percentage was incorrectly selected; in order to characterize this problem, it was helpful to distinguish between two different types of commission error: "true" commission error and "false" commission error, or "meta-error." An example of a true commission error would be a road in the GIS dataset where none should exist; a few instances of true error were observed in the TIGER/Line dataset. The remaining false commission errors encompassed attribute, positional, and temporal errors, all

Table 3.3: Errors of Omission by Year, in Kilometers

Year	Intersection	GDOT	Omission Error	TIGER Completeness
2008	7254	7575	4.24%	95.76%
2009	7226	7966	9.29%	90.71%
2010	4575	4801	4.71%	95.29%
2011	4680	4850	3.51%	96.49%
2012	9943	10368	4.10%	95.90%
Total	**33678**	**35560**	**5.29%**	**94.71%**

Table 3.4: Errors of Commission by Year, in Kilometers

Year	Intersection	TIGER	Commission Error
2008	7135	9591	34.42%
2009	7104	8500	19.65%
2010	4406	5397	22.49%
2011	4568	6697	46.61%
2012	9652	11048	14.46%
Total	**32865**	**41233**	**25.46%**

of which may easily masquerade as commission error in a buffer intersection-based GIS analysis. Attribute errors occurred in instances where driveways or service roads were incorrectly characterized as local roads in the original TIGER/Line files; therefore, the automated preprocessing method was not able to remove these roads from the analysis. Because these roads cannot be easily corrected across the entire dataset, they were considered as part of the "true" commission error; however, given that these attribute errors slightly decrease over time, the accuracies for annual percent change would be relatively unaffected. The latter two types of positional and temporal error unequivocally distorted the commission error, although, and were removed from the final commission error estimates by sampling the error distributions from 2008 to 2010, given that temporal accuracies could not be verified for 2011 and 2012 using the 2010 imagery. A stratified random sampling of these three years yielded 90 sample points of the commission errors; each of these samples was characterized by either true or false error, and the percentage of false error was removed from the original commission error estimates (Table 3.5). The average false error was applied to the years 2011 and 2012 to facilitate interpretation of the dataset as a whole; the total commission error value significantly decreased to around 10 percent. Higher values still remain for some years; however, as previously stated, the majority of the modified errors were due to attribute classification errors, not strictly commission errors.

Temporal Accuracy

Due to the significant influence of temporal lag on the errors of commission, the TIGER/Line and GDOT datasets were both tested for temporal accuracy; the inaccuracy of the initial commission estimates was sufficient cause to question the use of the GDOT dataset as an accurate temporal reference for assessing the quality of the TIGER/Line database. For both road representations, there is a distinct possibility that certain errors of omission may not be caught until the next release of the dataset, resulting in a dataset exhibiting temporal lag; it may be extremely difficult for data providers to ensure that their data are spatiotemporally accurate at the moment of distribution, especially for rapidly growing areas such as the Atlanta MSA (Epstein, Payne, and Kramer 2002). Additionally, the phenomenon of spatiotemporal mismatch in national census data has already been documented in the similar Canadian context (Schuurman, Grund, Hayes, and Dragicevic 2006). To address these concerns and determine whether temporal errors are more prevalent in GDOT data, the previous 90-point stratified error sample was re-examined, eliminating positional error from consideration and specifically examining temporal errors. Once the percentage of commission error due to temporal lag was verified, this mismatch value was multiplied by the total commissioned length for the TIGER/Line dataset (Table 3.6). Likewise, the procedure was reversed for the GDOT dataset, testing it against the TIGER/Line dataset with the assumption of greater temporal accuracy (Table 3.7). After multiplying by the total commissioned length, it was evident that the most meta-error in commission occurred for the

Table 3.5: Errors of Commission, Corrected for Positional Error and Temporal Lag

Year	Initial Error	Meta-Error	Modified Error
2008	34.42%	60.00%	13.77%
2009	19.65%	70.00%	5.90%
2010	22.49%	46.67%	12.00%
2011	46.61%	58.89%	19.16%
2012	14.46%	58.89%	5.95%
Total	**25.46%**	**58.89%**	**10.47%**

Table 3.6: Temporal Error in TIGER Commissioned Length, in Kilometers

Year	Mismatch	Commissioned Length	Temporal Error
2008	33.33%	2456	819
2009	50.00%	1396	698
2010	23.33%	991	231
Total		**4843**	**1748**

Table 3.7: Temporal Error in GDOT Commissioned Length, in Kilometers

Year	Mismatch	Commissioned Length	Temporal Error
2008	23.33%	321	75
2009	53.33%	740	395
2010	26.67%	226	60
Total		**1287**	**530**

TIGER/Line dataset, implying that there were many more instances of temporal mismatch in the GDOT dataset. Therefore, although the percentages of mismatch were quite similar, the effects were multiplied by the sheer volume of GDOT data that had not been updated to the improved TIGER/Line temporal accuracy, an unexpected finding since the GDOT data exhibited significantly better measures of positional accuracy.

Conclusion

Despite the lower positional accuracy and high errors of commission endemic to the TIGER/Line dataset, it is still highly suitable for usage in road network change detection. The 95th percentile of positional error lies at a 7.6 meter radial tolerance (less than the width of two interstate lanes) and the 99th percentile rests at a 22.75 meter tolerance (less than a city block). Although the attribute-based commission errors persist through time, later versions of the TIGER/Line dataset have progressively deleted some of the offending driveways and service roads from the dataset, such that inter-annual percentage changes would only slightly be affected by attribute errors. The measures of completeness for the TIGER/Line dataset remain relatively stable, and approach 95 percent when averaged across the five year temporal window, indicating that very little local dataset specificity would be lost when using the data at a national scope. Furthermore, the GDOT dataset appears to be no more temporally accurate than the TIGER/Line dataset, suggesting that no temporal lag would be forgone when using the local GDOT dataset. Therefore, the benefits of continuing with the analysis with a moderately flawed data source far outweigh the costs of not pursuing the idea of a national road network change database.

Database Creation and Change Detection

In order to construct the road change database, TIGER/Line data for the entire United States from 2008 to 2012 were downloaded via FTP from the U.S. Census Bureau, as described in the previous section on spatiotemporal accuracy. However, due to the vast volume of data and hierarchical nature of the FTP storage server, manual acquisition of the data could take several days. Complete road coverage was only available for download at the county level; therefore, one would have to manually right click and download each county individually for each of the five years, executing a total of 15,700 downloads. To assist in this process, an GNU-licensed open-source software called Wget was utilized to batch download all of the necessary files from the U.S. Census Bureau FTP site; the software functions as both a command line-based download manager and a web crawler that can navigate through various internet address hierarchies. Given that Wget allows recursive downloading and the usage of text wildcards for searching, it was ideal for drilling down through multiple folders to extract the TIGER/Line files. Once the files were downloaded in .zip format, 7-Zip software was used to batch process the extraction of the 15,700 zip files and place them in folders corresponding to their creation year.

After saving the county shapefiles (the de facto standard file type for vector GIS data) locally in separate temporal folders, a prototypical "snapshot" spatiotemporal database was produced, in which each geographic area has different data files corresponding to discrete time periods (Langran 1992). However, this form of spatiotemporal representation has two key disadvantages: it does not explicitly demonstrate dynamic changes over time and duplicates elements that have not changed over time, exacerbating data storage problems (Pelekis et al. 2004). The conversion of this extensive static database into a dynamic database requires the accurate identification of changes between each of the annual time "snapshots," ultimately creating a set of four change files (2008-2009, 2009-2010, 2010-2011, and 2011-2012) that capture new additions to the road network (Armenakis 1992). While this process appears to be quite straightforward and simple to implement, the inherent positional error of the TIGER/Line database, along with the fact that TIGER road attributes have become increasingly more accurate, has introduced significant difficulties for the static to dynamic conversion process.

In order to perform the conversion for all counties in the United States, an analytical geoprocessing framework was created with Python scripts to operate in the ESRI ArcMap environment, a common software package for the analysis of spatially embedded processes. Because the scripts interface seamlessly with the Windows operating system and the ArcMap graphical user interface (GUI), the scripts can interactively manage input and output files in the database, automating a tedious workflow that could otherwise take months for large study areas. The Python operating system (os) module was utilized to simplify data entry and automatically organize the outputs of various scripting and processing tasks, while keeping the source data in logical folders for easy retrieval. By default, relevant data are accessed using a predefined file storage hierarchy; the decision to use a static rather than dynamic filing system was made to optimize processing of the raw TIGER/Line data, and ensured that batched outputs of analyses had the same naming conventions and structural relationships as the original U.S. Census Bureau files. Nevertheless, both the scripts and the database are portable and operate on relative pathnames, so that the input paths for the scripts automatically default to the appropriate level of the database in the computer file system. Examples of the database hierarchy, both in its pristine state and after several processes have been executed, can be viewed in Figure 3.7.

The data itself, links to the raw Python scripts, and all the intermediate output files are stored in nested folders within the root folder. The raw TIGER/Line road data for each year may

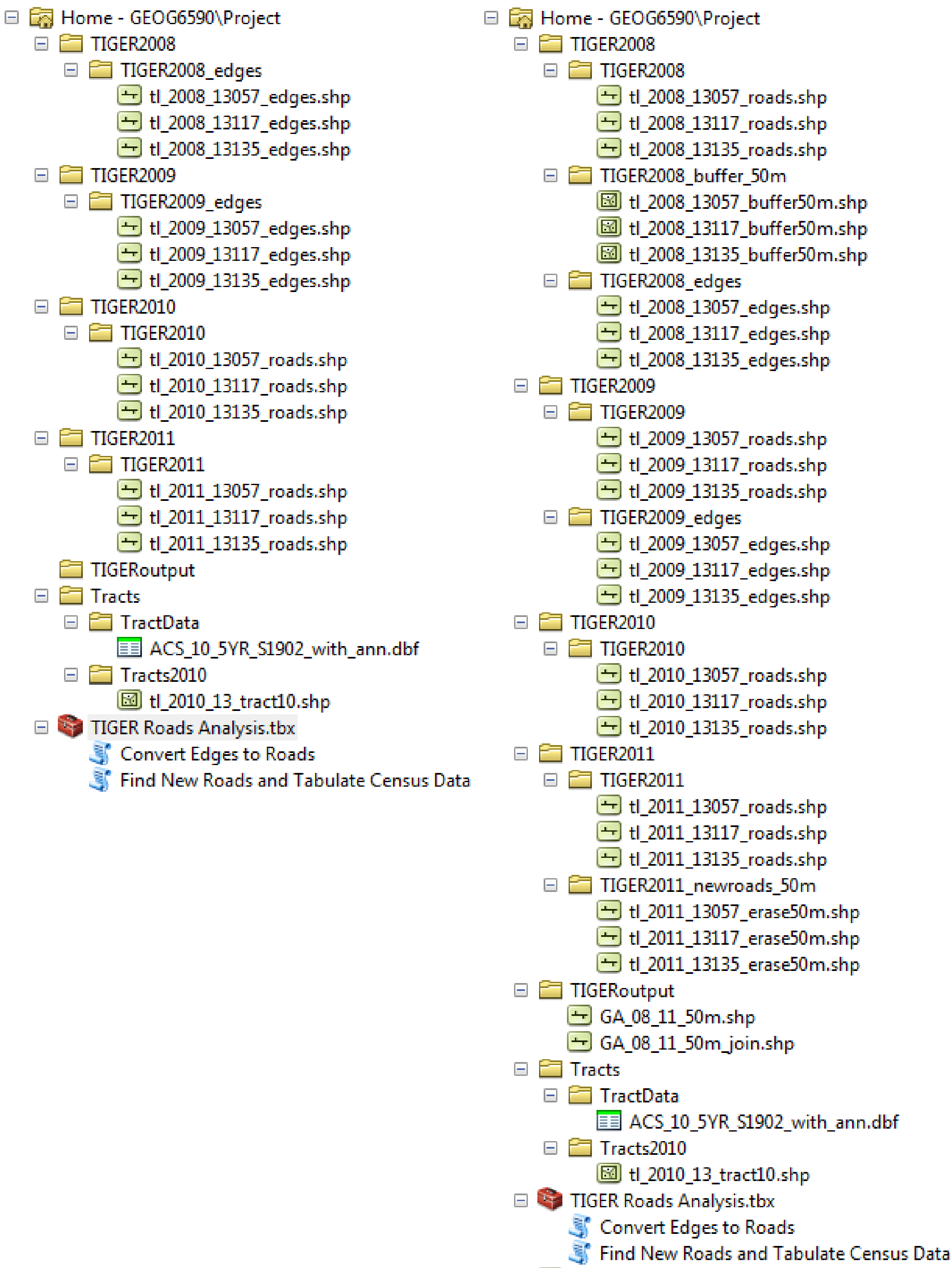


Figure 3.7: Simplified Database Hierarchy Before and After Process Execution

be found in the folder for the respective year. The Tracts folder contains the TIGER/Line tract boundary file for the 2010 census and a raw data table with geographic identifiers (in dBASE format—writable by OpenOffice) for storing socioeconomic variables of interest; note that the Tracts folder is a placeholder that may be expanded to State, County, MSA, etc., as appropriate for the required analyses. An ArcMap "toolbox," which facilitates information exchange between the raw python scripts, the underlying database, and the software graphical user interface, is contained in the root directory of the database, as were all of the scripts used to process the data; the packaging of these three elements within a toolbox enables sharing of the final product with other researchers, and ensures the integrity of the file system for future distribution.

The heart of the programming framework is the toolbox itself (TIGER Roads Analysis.tbx—refer to Figure 3.8), which contains sixteen Python scripts to enable automated geoprocessing (using the arcpy module to interface with ArcGIS) and workspace management (using the os module to interface with the file system). The scripts were separated into toolsets at the levels of individual road segment, tract, and county; the first group handles data preprocessing and change detection, while the following two groupings assist with the aggregation of road network change to the county and tract level, respectively. The separation of tasks into multiple scripts enabled robust performance of the toolbox in the face of a large volume of data (all of the raw U.S. Census TIGER/Line roads from 2008 to 2012 and tract-level socioeconomic data from the American Community Survey, along with several intermediate road datasets). The script output can exhaust storage capacities of between five hundred gigabytes and one terabyte, depending on the complexity and scale of the nationwide road network analysis; therefore, optimization of toolbox performance was not only desirable, but a necessity. Because the bulk of analytical work for this research was performed using these sixteen Python scripts, their structure and function will be reviewed in the following sections, and their contributions to the overall change detection and summarization process will be discussed in detail.

Individual Road Analysis Tools

The first of the Python script "tools" in the All Individual Roads toolset (A1. Convert Edges to Roads) preprocesses legacy TIGER/Line edge files from 2008 and 2009 to remove railroads and other linear features, using a structured query language (SQL) select statement to extract roads (see Figure 3.9 for GUI, Figure A.1 for raw script, and Figure A.2 for outputs). The user inputs the temporary "_edges" workspace, and the script saves the processed data to the same folder naming scheme as the existing data from 2010 through 2012, unifying the raw data.

The second tool (A2. Preprocess Feature Class Codes) removes the MAF/TIGER Feature Class Codes previously mentioned in Table 3.2 to erase service roads, trails, and other pathways

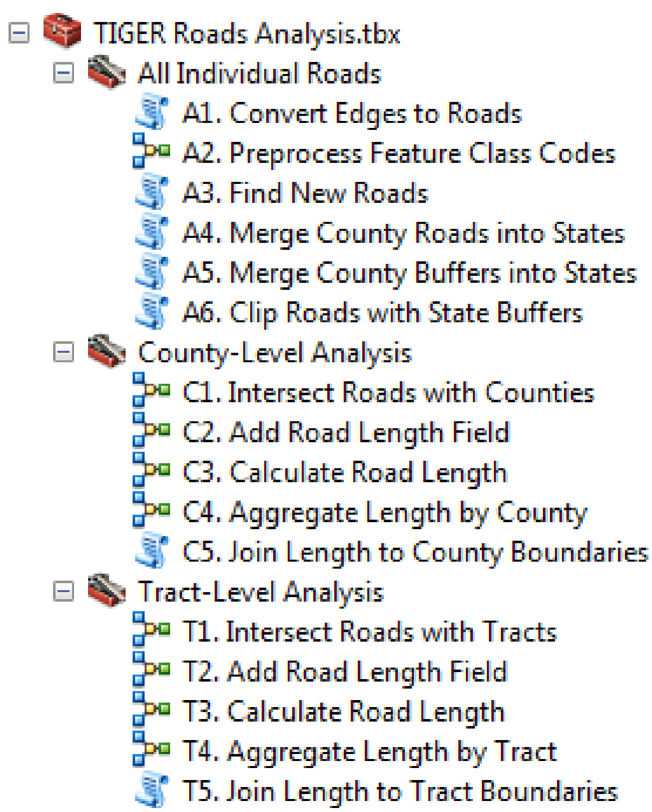


Figure 3.8: ArcGIS Toolbox for Change Detection and Summary Shapefile Creation

from all counties within each year, ensuring that the remaining attribute types are standardized and meaningful. Rather than using a script directly for user interfacing, this tool uses a model created through ArcGIS ModelBuilder, a visual programming language that produces scripts from a series of interconnected geoprocessing modules (see Figure 3.10). Generally speaking, scripts created through ModelBuilder excel at replacing the functions of the Python arcpy module (which accesses geoprocessing scripts bundled with ArcGIS software), yet have limited functionality for manipulating database file names (the role of the Python os module) or handling processing errors without failure of the script. Therefore, for this toolbox, the more robust scripts are used for routines that consistently access and rename files, and for those that process large quantities of data—without error handling, these scripts could silently fail even after several hours or days of runtime!

The third step in the All Individual Roads toolset, and the most important script of all, (A3. Find New Roads) generates shapefiles of new roads from "before" and "after" time periods. The graphical user interface may be found in Figure 3.11, where the user picks the starting and ending years of analysis (for example, to find new road additions between 2008 and 2012) and the geographic area of concern (either the entire U.S. or by an individual state). Effectively, for

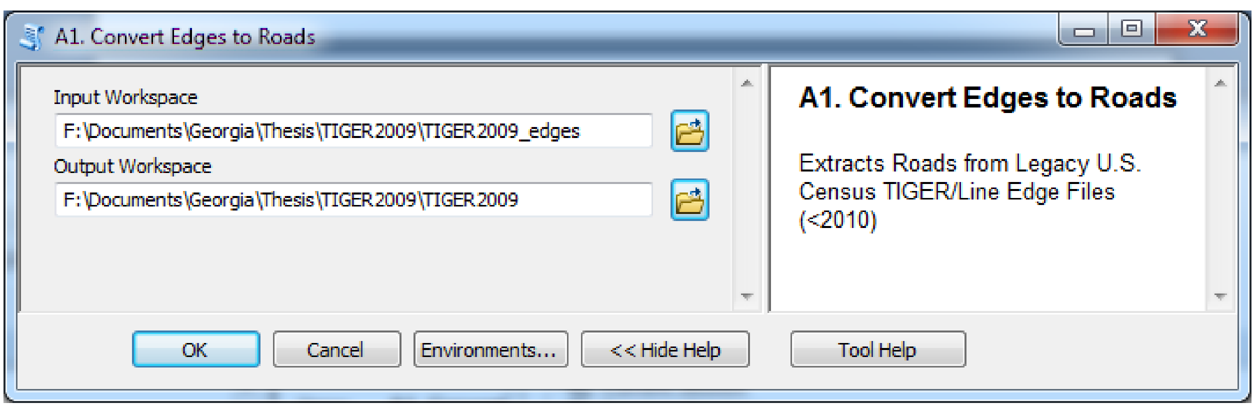


Figure 3.9: Graphical User Interface for Script A1

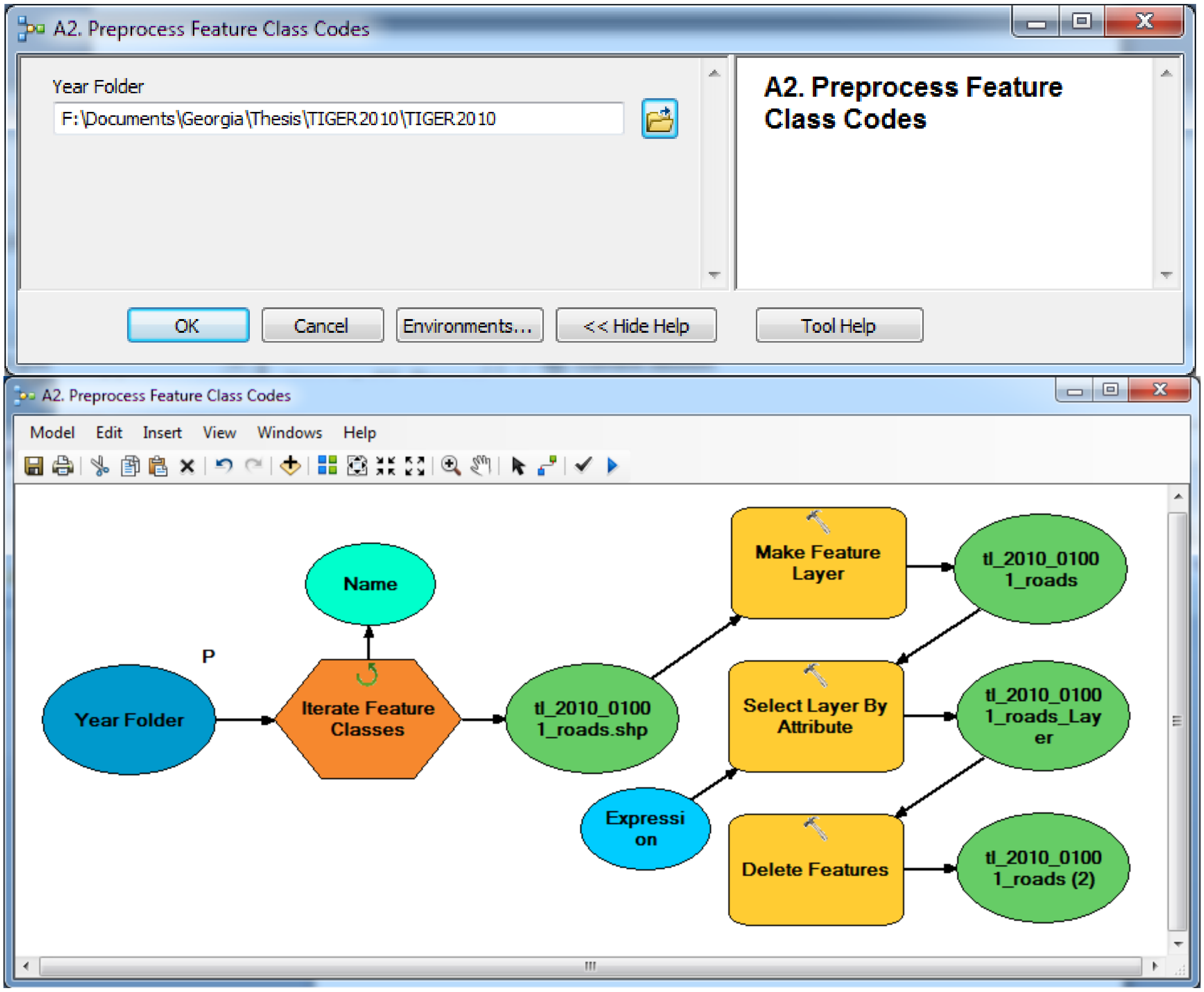


Figure 3.10: Graphical User Interface and Visual Programming for Model A2

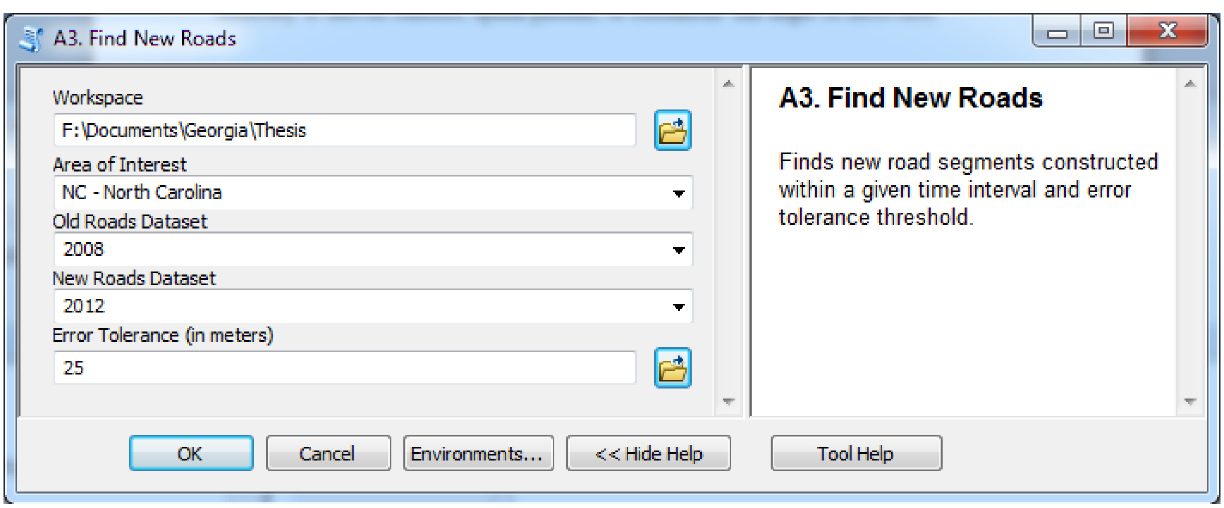


Figure 3.11: Graphical User Interface for Script A3

each county within the area of interest, the first year (2008) is buffered by a specified error tolerance. Here a value of 25 meters was used, in correspondence with the 99th percentile radial error of the TIGER dataset combined with potential errors in control imagery ortho-rectification. Next, the error buffer around these 2008 roads is erased from the last year's (2012) roads, yielding the parts of the road network unique to the 2012 roads dataset (see Figure 3.12). After completing the Find New Roads script, the joined output files are placed in the TIGERoutput folder, each uniquely identified by the state, from and to dates, and the buffer tolerance used.

The Find New Roads analysis would seem to be fairly straightforward, but the datasets are so large that they exceed the memory limitations presented by ArcMap and exacerbate the pre-existing memory leaks in many of the geoprocessing tools. Therefore, an entire state of

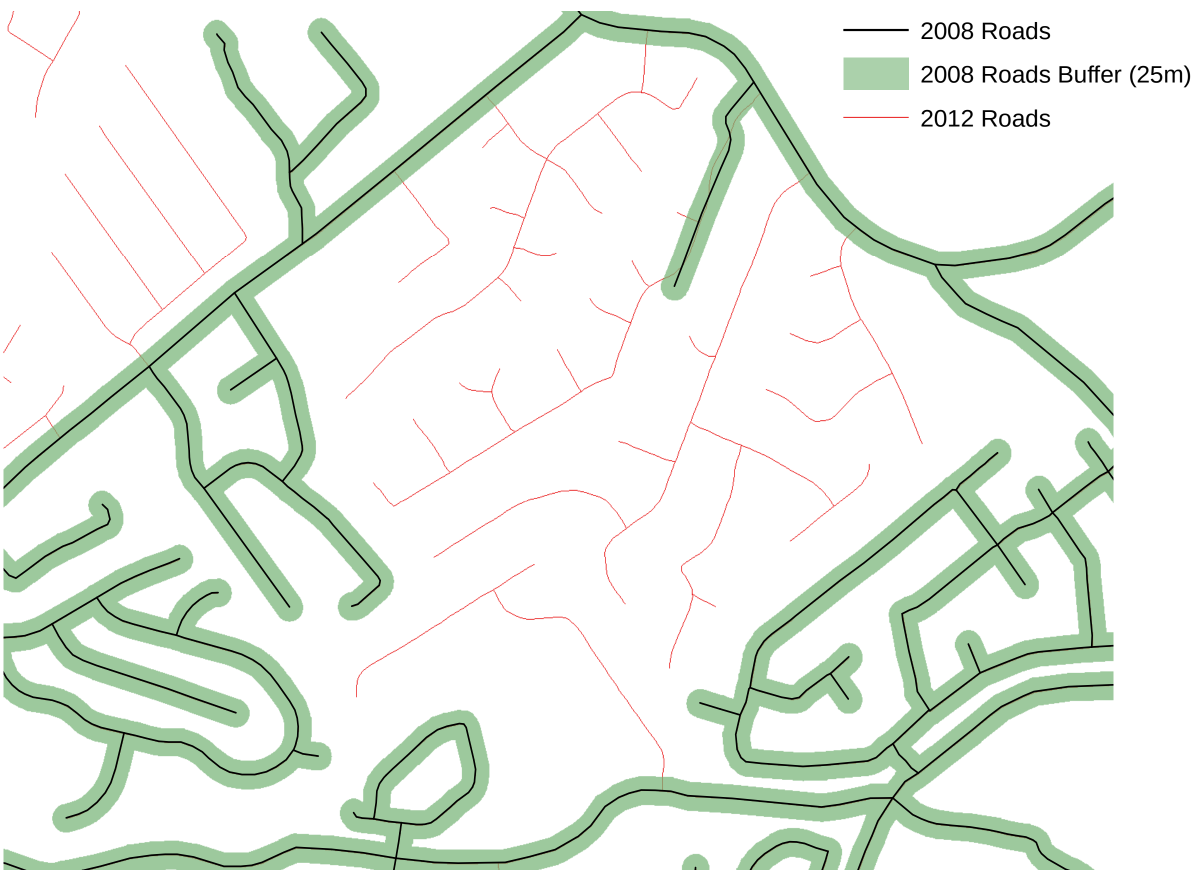


Figure 3.12: Vector GIS Representation of Find New Roads Scripting Logic

roads cannot be batch processed at the same time, and the arcpy Python geoprocessing module generates an "Error 999999", or simply fails with a runtime error while executing. To partially overcome this memory management issue, the Find New Roads tool was split into three different hierarchical scripts, where the highest-level script (for the United States) imports the state-level script, and the state-level script, in turn, imports the county-level script. The hierarchical framework allows the lower-level scripts to terminate once their jobs are finished, thereby clearing up more system memory and enabling the script to run longer without hitting an exception or runtime error. For this implementation, the national-level script translates the state abbreviation into a two-digit FIPS code to dynamically search for the right TIGER/Line files in the right folders (see Figure A.3); the state-level script manages the output folders for all the buffer and erase (see Figure A.4) and merge operations (joining the counties together for a state-level output file—see Figure A.5); and the county-level script performs the buffer and erase geoprocessing operations for each individual county road file, then terminates (see Figure A.6). Although this overall process is structurally complicated—which provides greater incentive for its automation using Python scripting—the line-by-line output of the script has been designed to be as user-friendly as possible, notifying the analyst of the active input and output workspaces for each operation, in real-time, and updating the view with the names of the finished output files as they are created. Because the merging process is much faster than the buffer and erase operations, and only produces one output, the file name was omitted for clarity (see Figure A.7).

The fourth and fifth tools in the All Individual Roads toolset consolidate the results of the county-level Find New Roads tool into a format where they can be more easily analyzed at the state level; although the new roads have been merged into a state-level shapefile, the raw TIGER/Line data and buffers have not yet been reconstituted from county shapefiles to state shapefiles. The fourth tool (A4. Merge County Roads into States—see Figure 3.13) produces a state-level shapefile of the raw road data for the specified area of interest, while the fifth tool (A5. Merge County Buffers into States—see Figure 3.14) compiles the county-level buffers. The roads and buffers are extracted by the specified year and area of interest (and in the case of the buffers, error tolerance), by parsing the informational file names previously created by the Find New Roads script (Figures A.8 and A.9). Reproducing the original buffers at the state level simplifies the process for discovering new roads at the state level; in order to continue from this baseline, one would simply execute the erase command manually in ArcMap between the raw

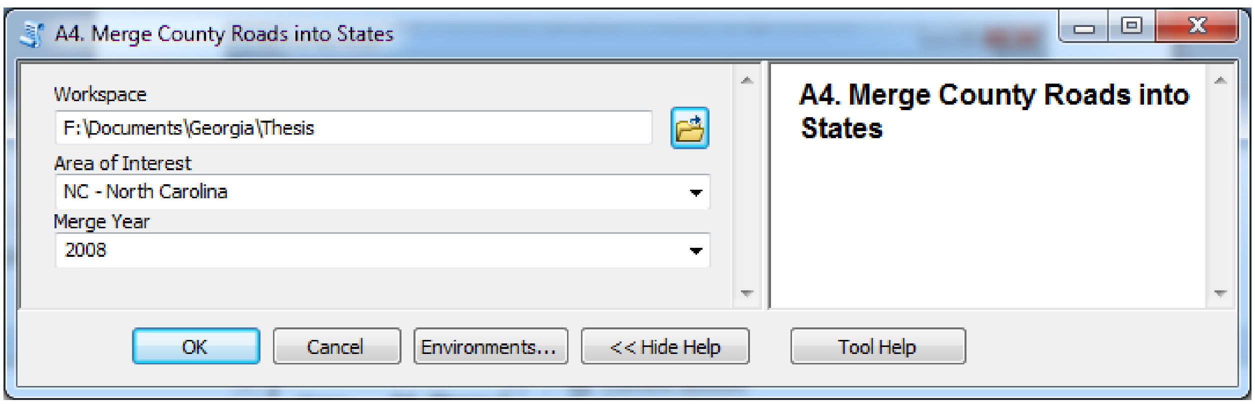


Figure 3.13: Graphical User Interface for Script A4

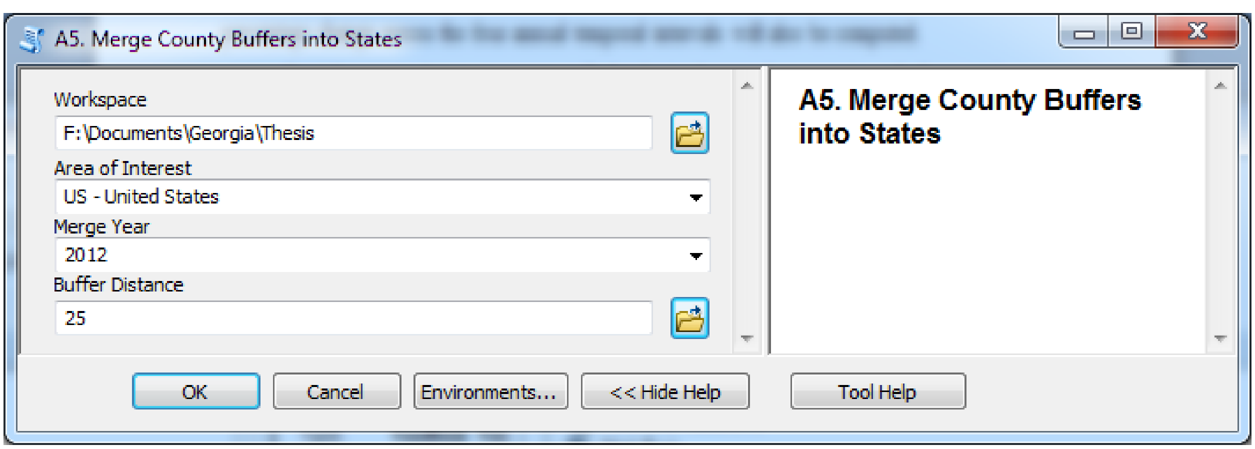


Figure 3.14: Graphical User Interface for Script A5

new roads data and the buffered base year data, yielding new roads by state. Rather than having to identify TIGER/Line files manually by FIPS code, states are easily identified by abbreviation.

The final tool in the toolset for individual road analysis (A6. Clip Roads with State Buffers—see Figure 3.15) was developed to fix aforementioned TIGER/Line attribute errors in which driveways and service roads were misclassified as local roads. Due to these errors, a persistent decrease in small rural roads was observed through time as erroneous "public" roads were progressively being deleted from the annual datasets; although some of these problem road segments were expected to be removed through the Preprocess Feature Class Codes script, the code obviously could not differentiate between correct and incorrect attribute value assignments. This substantial source of error, then, was corrected a second time through this script, which clipped all of the earliest dataset year extents to the error buffer around the latest year's dataset. With the assumptions that human deletions of incorrectly attributed data would be more accurate than eschewing the latest attribute revisions, and that the loss in total road length due to earlier positional errors outside the buffer zone was negligible, it was prudent to systematically obviate this chief source of length inflation via Python scripting (Figure A.10).

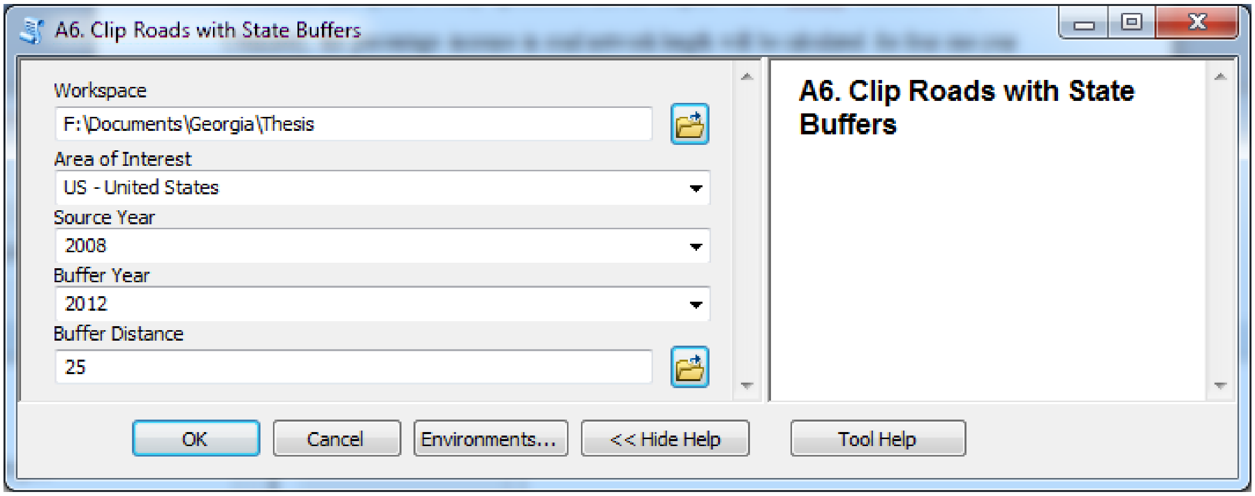


Figure 3.15: Graphical User Interface for Script A6

Aggregate Road Analysis Tools

The remaining two toolsets for county and tract analyses are duplicated given that the same five generalized steps are required to spatially summarize the individual road length data. Due to these workflow similarities, the only differences between the county and tract scripts are the spatial boundary files; therefore, only the county-level scripts will be addressed. The first of these (C1. Intersect Roads with Counties—see Figure 3.16) takes the year as input, and spatially intersects each of the state new road shapefiles with the county boundary file, essentially adding attribute fields for the county name and state on each of the new road segments. The second script (C2. Add Road Length Field—see Figure 3.17) adds another attribute field to store the segment length, while the third script (C3. Calculate Road Length—see Figure 3.18) fills in the length in meters for each road segment, ultimately producing individual length data for every road in the entire nation from 2008 to 2012. The fourth script (C4. Aggregate Length by County—see Figure 3.19) creates a table containing total new road segment length for every county and equivalent jurisdiction in the United States. The final script (C5. Join Length to County Boundaries—see Figure 3.20 for GUI, Figure A.11 for county-level script, and Figure A.12 for tract-level script) attaches this length summary table to the aggregation unit boundaries. The final results, after running all sixteen scripts in the TIGER Roads Analysis toolbox, were national shapefiles (of both counties and tracts) that were attributed with total new road lengths. In order to extract statistics for Metropolitan Statistical Areas, the MSA boundaries were manually intersected with the national new county road length shapefile, yielding urban counties. Then, just as previously indicated, summary statistics were created for total road length by MSA, and the statistics table was joined to the original MSA boundary file. In this manner, new road

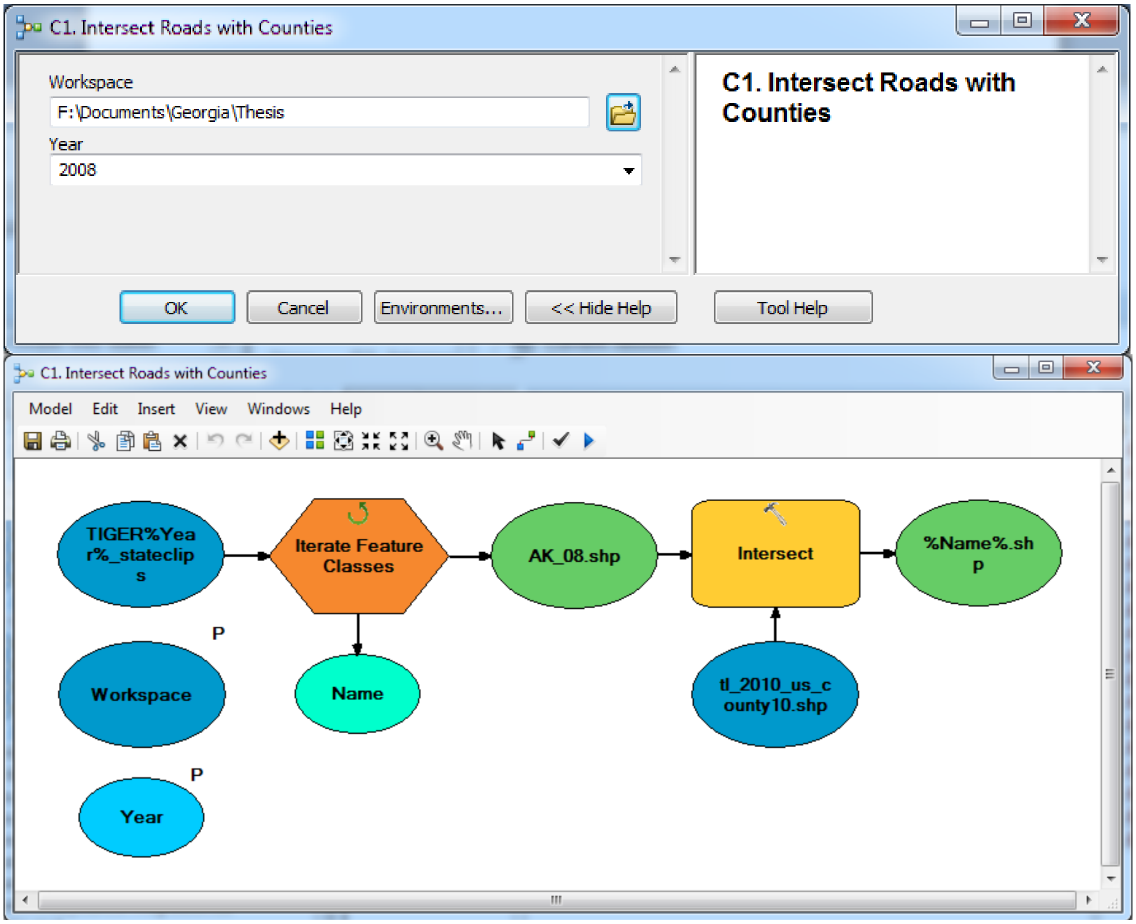


Figure 3.16: Graphical User Interface and Visual Programming for Model C1

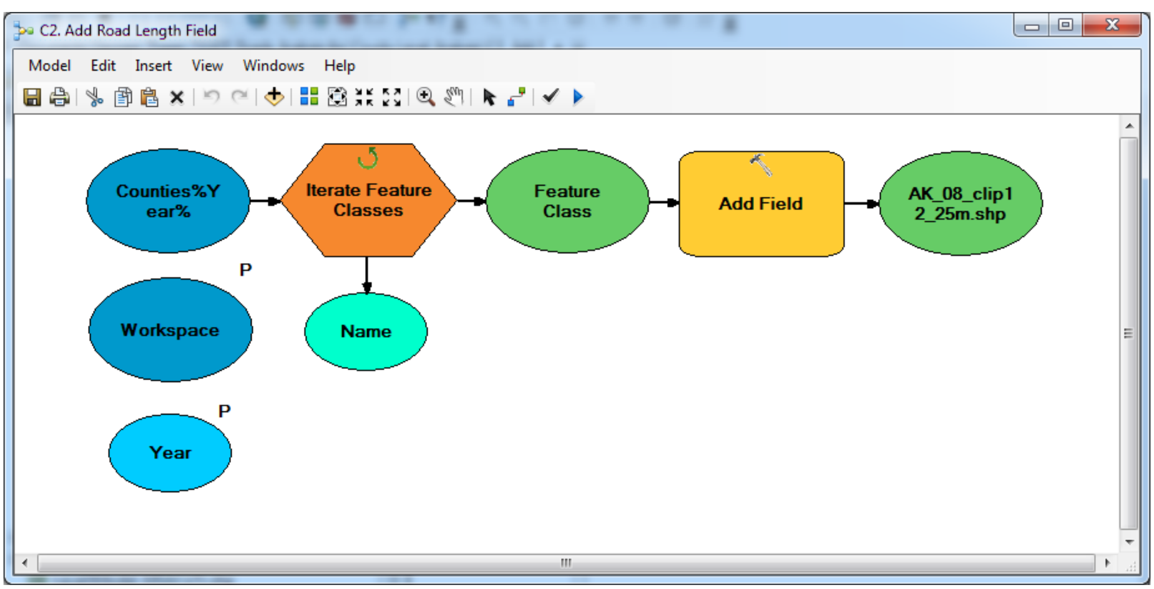


Figure 3.17: Visual Programming for Model C2

length statistics were ultimately produced at the MSA, county, and tract levels of aggregation, which were then extensively analyzed through the following methodologies.

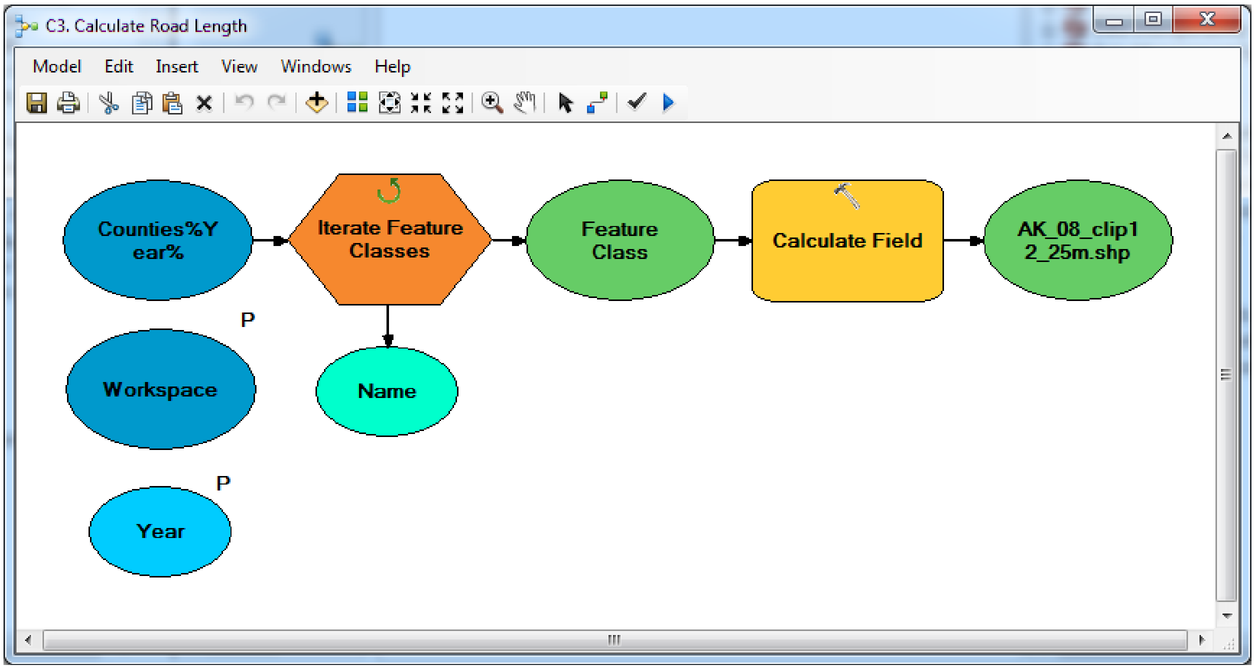


Figure 3.18: Visual Programming for Model C3

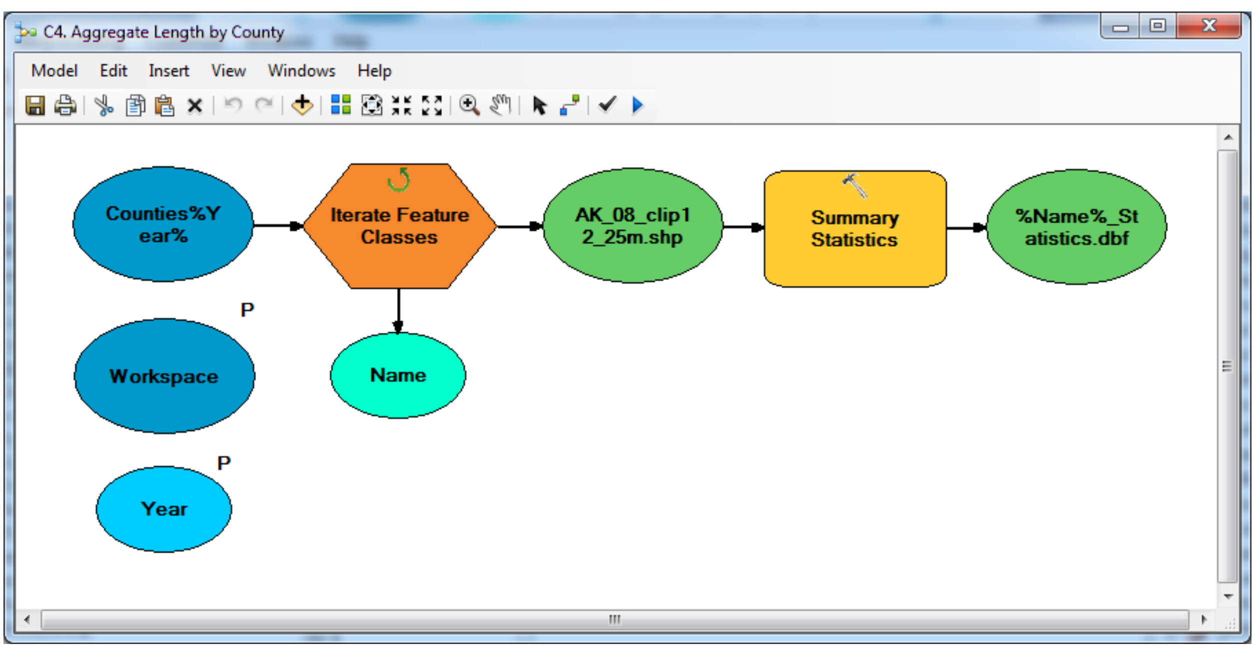


Figure 3.19: Visual Programming for Model C4

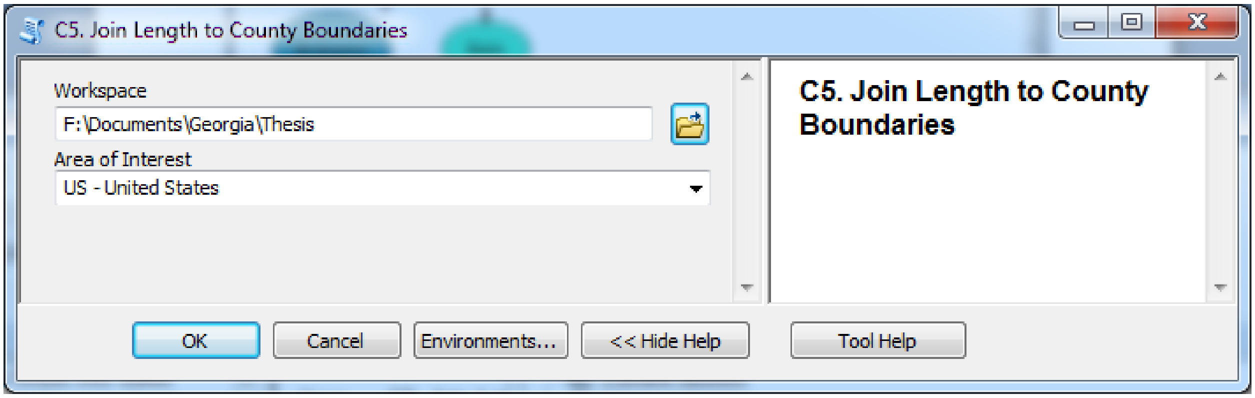


Figure 3.20: Graphical User Interface for Script C5

Exploratory Data Analysis

Once road network change shapefiles were produced for each temporal interval, descriptive statistics were compiled for each static annual dataset as well, enabling normalization of the new road lengths by the existing lengths at the beginning of each change detection period. Percentage increases in road network length were then calculated for the four annual intervals, as well as between 2008 and 2012, enabling identification of significant additions to the network on both annual and five-year bases. In order to determine whether certain MSAs had a wider range of road addition variability (for example, in response to fluctuating low-density housing supply or macroeconomic recession), the range of percentage change across the four annual temporal intervals was also computed. However, the spatial variability of temporal accuracy for the TIGER/Line dataset elicited concern, given that in some cases the Census Bureau relies on third-party information from disparate sources to prevent workload duplication; therefore, annual change statistics were only reported at the MSA level, where differences in update completeness could be averaged out and minimized. These temporal concerns were indistinguishable when temporally aggregating to the entire five-year interval, though, enabling production of spatiotemporally aggregated road network growth totals for the MSA, county, and tract levels.

Spatial Autocorrelation

According to the colloquial Tobler's First Law of Geography, which states that "everything is related to everything else, but near things are more related than distant things," we should expect spatial dependency, or clustering of values, to occur across the nationwide measurements of road network change (Tobler 1970). Therefore, spatial autocorrelation measures were implemented in ArcGIS to examine whether certain counties or tracts of high road network growth tend to be adjacent to other high percentage change values (and vice versa);

in theory, autocorrelation should be pervasive across the country, given that suburban areas in general, and agglomerations in the southern and southwestern United States in particular, would be expected to exhibit larger percentage increases in total road length. However, although the Moran's I statistic for spatial autocorrelation (which calculates spatial autocorrelation using a neighborhood of contiguous polygons) should likely be positive, indicating clustering of similar values (Moran 1950), this global measure of spatial autocorrelation assumes pattern homogeneity: a condition that would likely be unsatisfied due to the organic distribution of growth in a road network with human actors. Therefore, individual clusters were also examined using the local decomposition of the Moran's I statistic, also known as "local indicators of spatial association," or LISA (Anselin 1995); mapping the local Moran's I values by county enabled identification of road network growth clusters and visualizations of their size and magnitude.

Correlation Matrices

The next data analysis initiative was to determine whether selected socioeconomic variables should be included in an areal regression model of road network change. Exploratory correlation analyses were restricted to metropolitan areas to target suburban road growth, and the overall five-year percent change data were used at all three aggregation scales to reduce the impact of temporal reporting inconsistencies on the variable association results. However, because the road change data still needed to be aggregated to a particular zone structure, the correlation coefficients were subject to the modifiable areal unit problem, which states that the level of aggregation, as well as the zonal boundary structure, will affect the results of spatial analyses (Openshaw 1984). To partially address this issue, the different aggregation schemes can each be utilized and compared wherever practical: for example, some variables were only available at the metropolitan area level, while others were available for the census tract level.

The socioeconomic variables hypothesized to be associated with road network growth were mean household income, mean age, population density, percentage population change from 2007 to 2011 (U.S. Census Bureau 2012c), racial/ethnic background, and census divisions (Figure 3.21). Mean household income data were derived from the U.S. Census Bureau (2012a) American Community Survey 2007-2011 five-year estimates, which can be found at MSA, county, and tract aggregation levels. Age and race/ethnicity data were retrieved from the 2010 Census Bureau MSA and county levels. The correlation coefficients for population change and

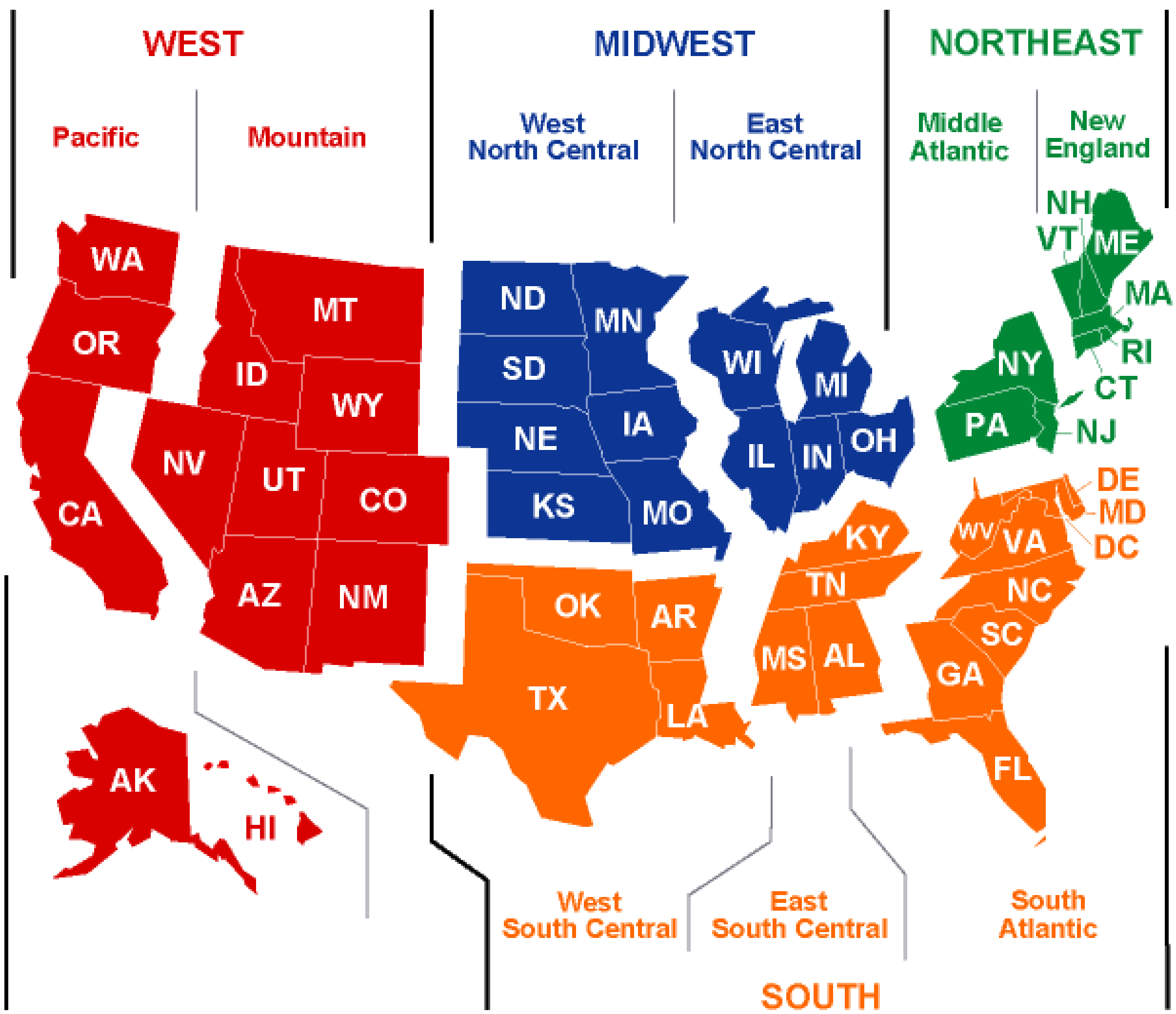


Figure 3.21: U.S. Census Regions and Divisions (United States Department of Energy 2000)

mean household income were expected to be positive, while negative coefficients were hypothesized for age, given that construction of new neighborhoods in economically advantaged neighborhoods tend to cater and market toward younger populations. Population density was not easily predictable due to its varying influence across aggregation levels; in general, larger zones such as MSAs might exhibit as a positive influence, since density could act as a proxy for land value and desirability, while for a small unit such as a census tract, high population density would likely mean that no developable land would be available in the tract. For the categories of race and ethnicity, positive coefficients were expected for Asian/Pacific Islander and White races, along with Hispanic backgrounds, and negative coefficients were predicted for American Indian and Black categories; the composition of the "Other" category was unknown. The influence of the regional divisions was simple to predict for certain regions, yet more difficult for others; in general, due to the protracted migration of economic activity away from the "rust belt" and toward the "sun belt," one might expect positive correlations for the divisions in the greater South region, as well as growth in the Mountain division from its more southerly states. Conversely, the economic stagnation of the state of California might cause the Pacific division to correlate negatively with road network expansion, while the Midwest and Northeast regions are not contemporarily known as being large centers of population growth. The correlation strengths and directions of these twenty potential independent variables, as a whole, should inform future road network models, enabling identification of pertinent explanatory factors for expansion.

Multiple Linear Regression

After producing Pearson product-moment correlation matrices between the twenty potential independent variables and the dependent variable of percentage road network change, coefficient values higher than .8 between independent variables were flagged as potential sources of regression-confounding multicollinearity (see Appendix B). For each of these highly correlated pairs, the independent variable exhibiting the strongest correlation with percentage length change was retained for model inclusion, while the weaker independent variable was dropped from the analysis. This method proved efficacious for eliminating multicollinearity, as each of the independent variables within each aggregation level exhibited variance inflation factors (VIFs) of less than five, a commonly selected threshold for problematic collinearity between predictors. For each of the model aggregation levels (MSA, county, and tract), the smaller Other racial category was dropped due to strong correlation with Hispanic background. Reference variables were also removed for both regional (West North Central) and racial (White) categories: because the racial population percentages added up to nearly one hundred percent, these variables operated much like a complementary set of dummy-coded variables. Model heteroskedasticity was evaluated using Koenker's studentized Breusch-Pagan statistic; significant heteroskedasticity was persistent across each aggregation level (Breusch-Pagan 1979; Koenker 1981). In these situations, ordinary least squares (OLS) regression no longer holds the property of best linear unbiased estimator; therefore, robust standard errors, t-statistics, and probability values were reported for all models. Statistically insignificant variables were retained for the grouped racial/ethnic population percentages and regional dummy variables in order to facilitate comparisons with the reference categories, while other insignificant independent variables were simply dropped from the models. Final models for each level maximized adjusted R^2 values.

Geographically Weighted Regression

Although the expected presence of spatial autocorrelation, largely due to existing clusters of urbanization, was somewhat alleviated by restricting the multiple regression models to MSAs, spatial clustering was not explicitly considered for the previous three models. The technique of geographically weighted regression (GWR) performs an individual linear regression for each data record, with the dependent variables restricted to a particular spatial neighborhood subset. Therefore, GWR accounts for autocorrelation-induced heteroskedasticity in a regression model, producing different regression coefficients, standard errors, and R^2 values in each neighborhood; however, global goodness-of-fit measures are also presented in the form of both global adjusted R^2 and the corrected Akaike Information Criterion (AICc), enabling comparison of GWR results with traditional linear regression models (Sugiura 1978).

In order to assess the efficacy of geographically weighted modeling on road network change, a base model was executed at the county aggregation level for the entire United States, not simply those counties included in an MSA; this distinction was necessary given that GWR cannot be performed adequately with non-contiguous polygons. Because spatial autocorrelation indices were expected to be greater with county-level than with tract-level data, only a county GWR model was utilized. The spatial neighborhood search kernel was defined by an AICc minimization function in ArcGIS that identified an optimal fixed number of neighboring counties, while model specification was more difficult due to the degrees of freedom lost to the GWR modeling process. The increased sensitivity of the model to the number of variables required paring the original model down to three explanatory fields: mean income, percent population change, percent Asian/Pacific Islander. Given that the models were substantively different, their comparative explanatory power was evaluated using measures of R^2 and AICc.

CHAPTER 4

RESULTS

Descriptive Statistics

Although the accuracy assessment, database creation, and Python geoprocessing may be considered results of this investigation, these methodological steps were ultimately preconditions for the expository component of the research: presenting the database and some of its potential applications. While a database itself may be considered a contribution to knowledge, its true value may only be discovered by harnessing and synthesizing the data to create new perspectives; with this frame of reference, even descriptive statistics on temporal road change can be perceived as novel. For example, through summarizing the road change dataset, it was discovered that 289,021 kilometers of new roadway have been constructed in the United States between 2008 and 2012, with each state on average experiencing an increase of 2.7 percent. Four states (North Dakota: 0.43%; Iowa: 0.86%; Kansas: 0.90%; and Nebraska: 0.92%) and the District of Columbia (0.64%) experienced a growth rate of below one percent, while four states in the contiguous United States (Delaware: 5.49%; Connecticut: 6.01%; Maine: 6.18%; and Georgia: 6.70%) experienced a growth rate of above five percent (Figure 4.1). As a whole, road network increases were shown to be highest for Alaska and Hawaii; however, upon further inspection of the 2008 TIGER/Line road networks, positional accuracies were extremely low. Due to the methodologies used in producing the percent road change statistics, positionally inaccurate data (as determined by the most current 2012 vintage shapefiles) were removed from the base 2008 length calculations, thereby inflating the normalized change ratios. Therefore, only road change values from the contiguous United States will be presented in this chapter.

At the MSA level, the Myrtle Beach-North Myrtle Beach-Conway, South Carolina MSA was the fastest-growing metropolitan area in the conterminous United States, followed by Lake Havasu City-Kingman, Arizona and Gainesville, Georgia (Table 4.1). The geographic distributions of metropolitan growth can be observed in Figure 4.2, where both percentage road change and new road density may be compared. Although both statistics serve to normalize the raw length of road change, they each have different strengths: percentage road change favors areas with less existing road mileage (where small additions to length may be drastic when compared to baseline data), while new road density emphasizes areas that may have a great deal of existing roads, yet have still experienced substantial growth in terms of added network length.

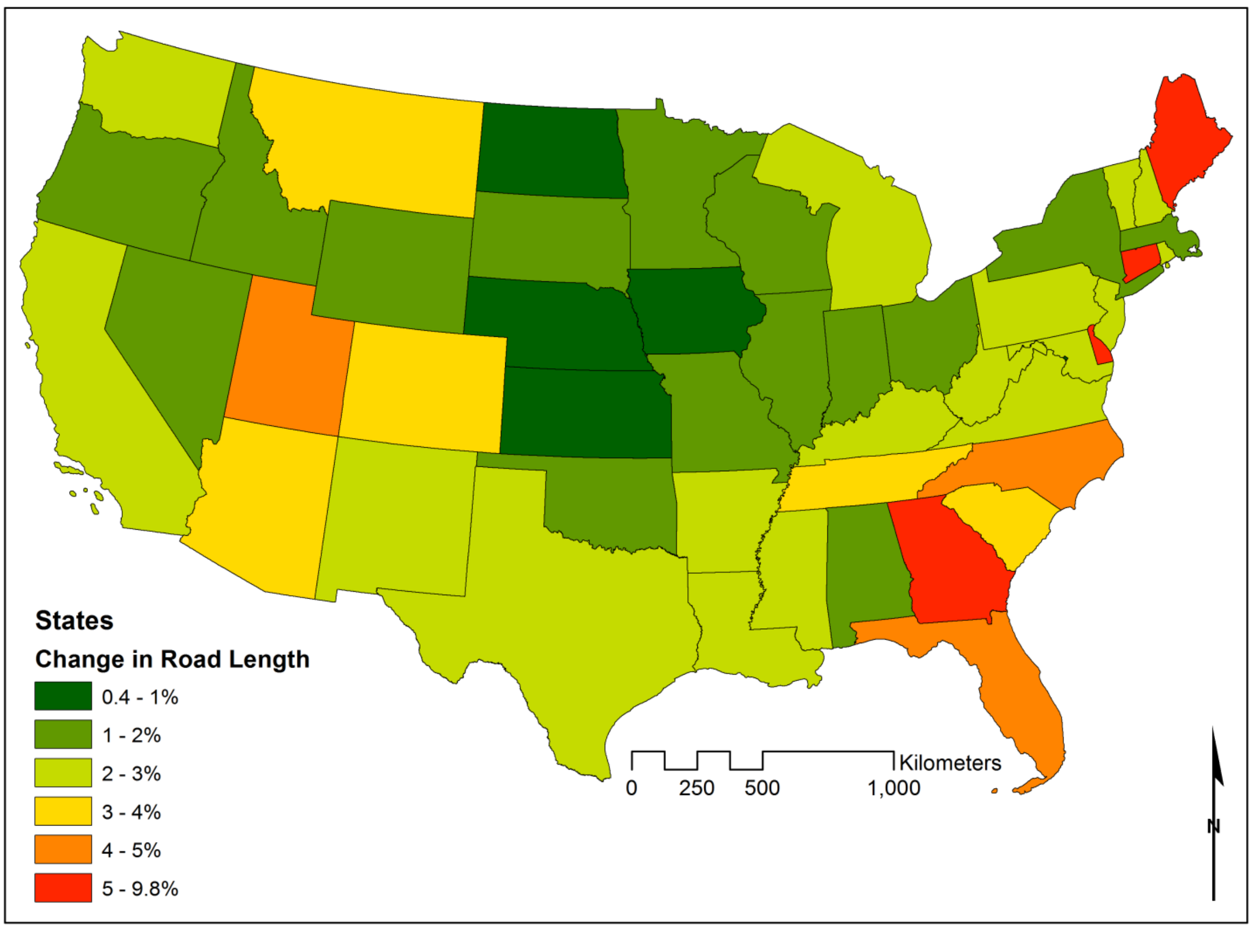


Figure 4.1: Percent Increase in Road Network Length by State

Table 4.1: Top Twenty-Five MSAs for Road Network Growth in the United States

Metropolitan Statistical Area	Percentage Road Increase
Myrtle Beach-North Myrtle Beach-Conway, SC	20.57
Lake Havasu City-Kingman, AZ	16.17
Gainesville, GA	14.34
St. George, UT	13.86
Atlanta-Sandy Springs-Marietta, GA	11.67
Dover, DE	11.19
Orlando-Kissimmee-Sanford, FL	10.65
McAllen-Edinburg-Mission, TX	10.57
Savannah, GA	10.08
Bridgeport-Stamford-Norwalk, CT	9.33
Jacksonville, NC	9.29
Pensacola-Ferry Pass-Brent, FL	8.75
Bangor, ME	8.74
Olympia, WA	8.26
Brunswick, GA	8.24
Athens-Clarke County, GA	8.12
Austin-Round Rock-San Marcos, TX	8.01
Dalton, GA	7.65
Warner Robins, GA	7.59
Houston-Sugar Land-Baytown, TX	7.59
Naples-Marco Island, FL	7.26
San Jose-Sunnyvale-Santa Clara, CA	7.15
Burlington, NC	7.13
Las Vegas-Paradise, NV	7.01
Hartford-West Hartford-East Hartford, CT	6.88

At the county level, three areas in the contiguous United States (Gilmer County, GA: 28.68%; Barrow County, GA: 29.72%; and Forsyth County, GA: 38.65%) exhibited growth rates exceeding twenty-five percent; furthermore, eleven of the top fifteen growth counties by percentage were located in the state of Georgia (Table 4.2). Regardless of whichever method chosen to normalize the added network length, the most rapid road growth counties were located in Georgia (Figure 4.3; Figure 4.4), unequivocally indicating an agglomeration around the city of Atlanta; its strength will be explored in the following section.

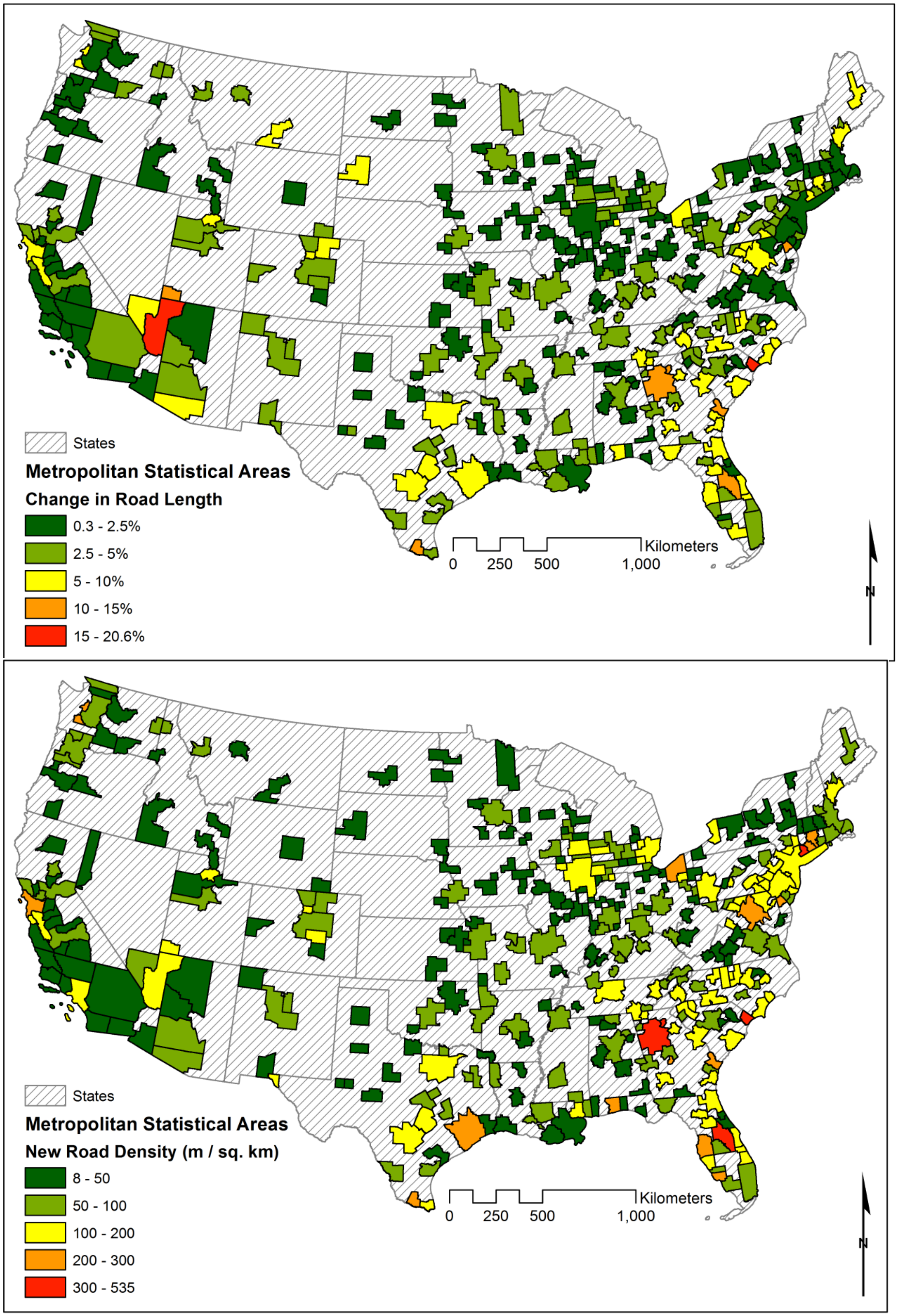


Figure 4.2: Change in Road Network Length and New Road Density by MSA

Table 4.2: Top Twenty-Five Counties for New Roads by Percentage and Density

Rank	County or Equivalent	Percentage	County or Equivalent	Density (m/km^2)
1	**Forsyth County, GA**	38.65	**Forsyth County, GA**	1181.99
2	**Barrow County, GA**	29.72	**Gwinnett County, GA**	1012.61
3	**Gilmer County, GA**	28.68	**Broomfield County, CO**	944.13
4	**Henry County, GA**	24.08	**Barrow County, GA**	757.30
5	**Paulding County, GA**	23.55	**Henry County, GA**	719.02
6	**Jackson County, GA**	22.89	**Manassas Park City, VA**	709.86
7	**Polk County, NC**	21.36	**Clayton County, GA**	662.29
8	**Jefferson County, WV**	21.15	**Tarrant County, TX**	652.74
9	**Horry County, SC**	20.57	**Douglas County, GA**	601.95
10	**Pickens County, GA**	20.40	**St. Louis City, MO**	600.03
11	**Union County, GA**	19.98	**Harris County, TX**	598.72
12	**Douglas County, GA**	19.35	**Rockwall County, TX**	574.56
13	**Gwinnett County, GA**	19.21	**Cuyahoga County, OH**	544.59
14	**Forest County, PA**	18.29	**Fulton County, GA**	541.34
15	**Effingham County, GA**	18.01	**Fairfield County, CT**	535.14
16	**Rockwall County, TX**	17.56	**Paulding County, GA**	524.42
17	**Cherokee County, GA**	17.54	**Contra Costa County, CA**	523.96
18	**Columbia County, GA**	17.40	**Fort Bend County, TX**	521.25
19	**Walton County, GA**	17.07	**Gilmer County, GA**	504.83
20	**Pasco County, FL**	17.06	**Pasco County, FL**	502.59
21	**Oconee County, GA**	16.74	**Alexandria City, VA**	501.74
22	**Osceola County, FL**	16.72	**Jefferson County, WV**	500.04
23	**Lee County, GA**	16.67	**Clarke County, GA**	482.42
24	**Mineral County, MT***	16.37	**Prince William County, VA**	476.06
25	**Fort Bend County, TX**	16.30	**Fairfax County, VA**	475.21

*A majority of new roads were within National Forest boundaries.

However, some abnormalities in data acquisition must be addressed before dismissing the descriptive road change statistics. Along with the previously mentioned positional errors in Alaska and Hawaii, road attribution errors occurred in a few counties where National Forest logging roads or other resource extraction roads were incorrectly labeled as local roads. It should also be noted that percentage change and new road density statistics are most accurate in areas with extensive and dynamic road networks, where static and change sample sizes are maximized.

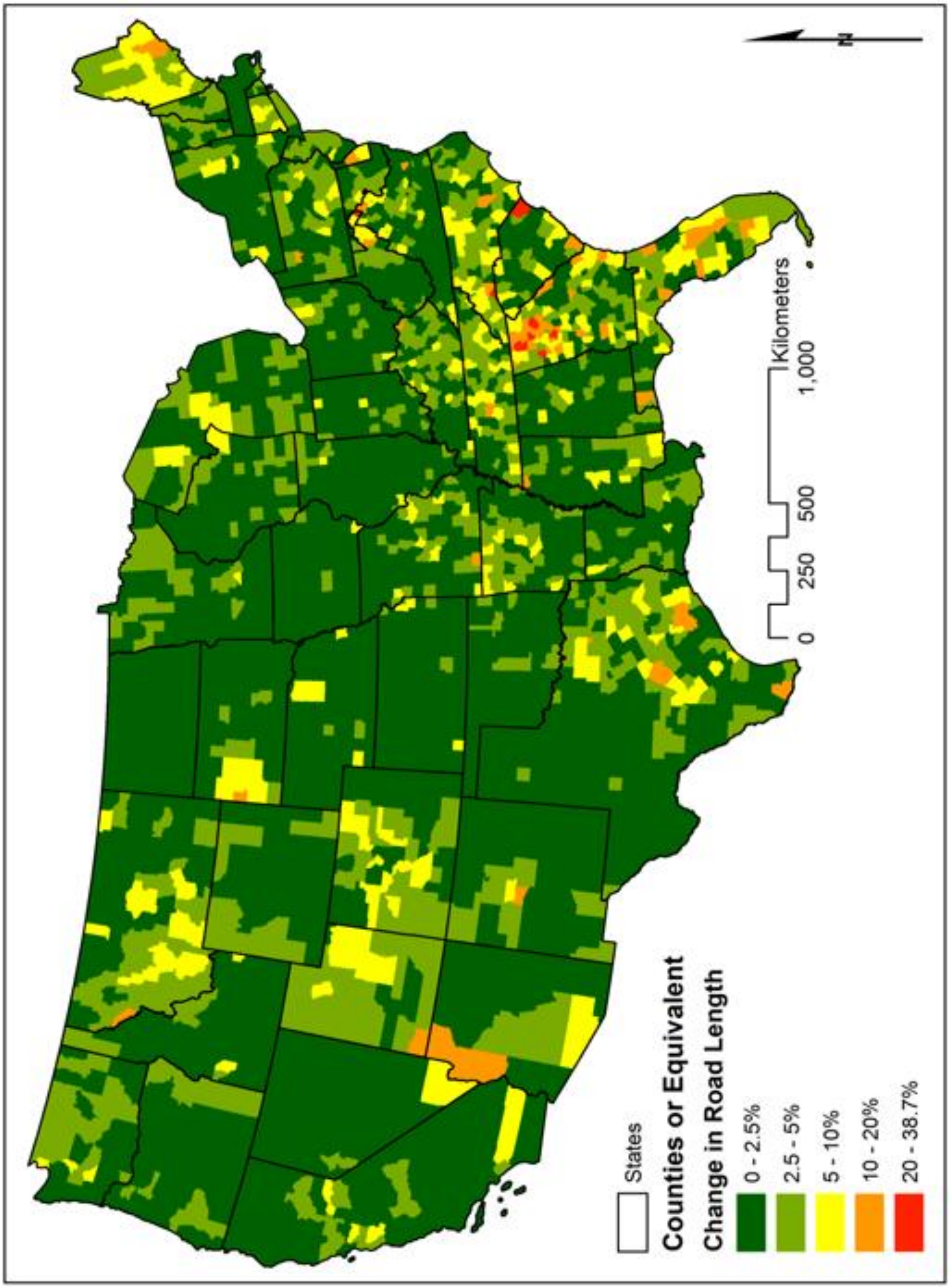


Figure 4.3: Percent Increase in Road Network Length by County or Equivalent

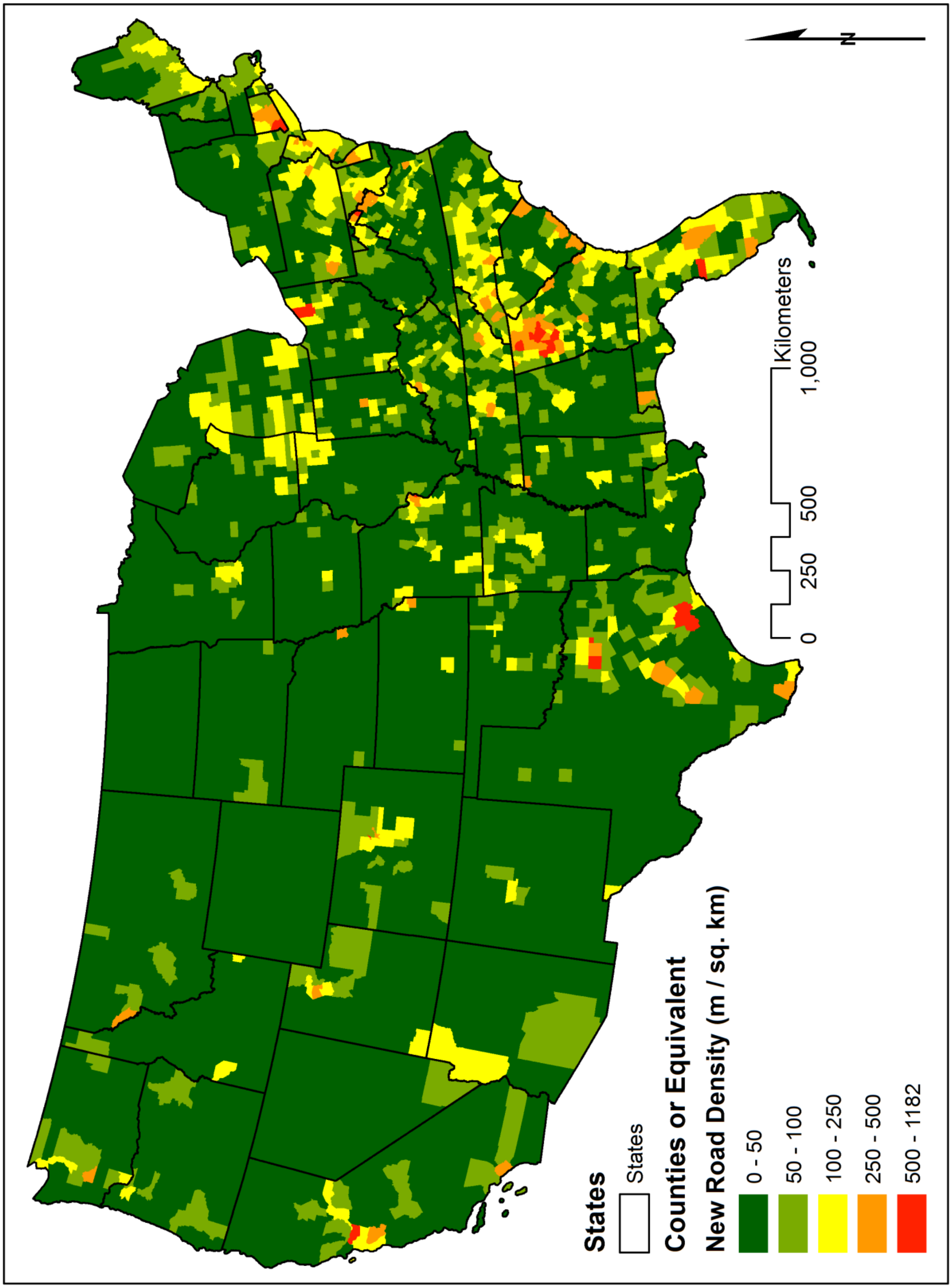


Figure 4.4: Land-Area Density of New Road Network by County or Equivalent

Spatial Autocorrelation

As expected, spatial clustering of road network percent change values was discovered using the graphical output of Figure 4.3; the statistical magnitude and significance of these clusters were examined through global and local versions of the Moran's I statistic. Because the statistic cannot be reliably computed for non-contiguous polygons, MSAs were omitted. Figure 4.5 indicates that the distribution of percent change at the county level across the United States is significantly clustered; therefore, we may reject the null hypothesis that no clustering exists, and state that the distribution deviates from what might be expected by a random number generator. The global Moran's I index values range from negative one to positive one, where a value of -1 describes perfect dispersion and a value of +1 represents perfect clustering. In a nationwide study area with a large number of observations, however, the statistically significant county-level value of .43 describes a moderately strong level of clustering. The tract-level autocorrelation was found to be more statistically significant due to the larger sample size, yet the global Moran's I value turned out to be only half as strong (Figure 4.6). Although both values were positive, the drastic difference between the two index values was attributed to the modifiable areal unit problem; additionally, the construction of new roads, especially for low-density residential developments, tends to be more highly segregated across census tract boundaries. However, though the tract-level analysis may better represent a chief cause of road network expansion, the county-level data were still more reliable and are more easily visualized than the tract-level data.

Therefore, a local analysis of spatial autocorrelation was conducted at the county-level with local Moran's I statistics, which indicate the level of clustering among contiguous polygons. Although the statistic values are not directly comparable with the global version, mapping the local Moran's I results elucidates the clustering dynamics of the road network growth statistics.

Figure 4.7 presents a probability map for each county; those areas shown in orange and red may be considered statistically significant at $\alpha = .05$ and $\alpha = .01$, respectively. However, while this map indicates whether a county has a statistically significant index value, it fails to reveal whether these counties represent areas of clustering or dispersion; therefore, a z-score map has also been included to display the sign and magnitude of the local Moran's I index values (Figure 4.8). Even though a particular county may be surrounded by statistically significant clustering,

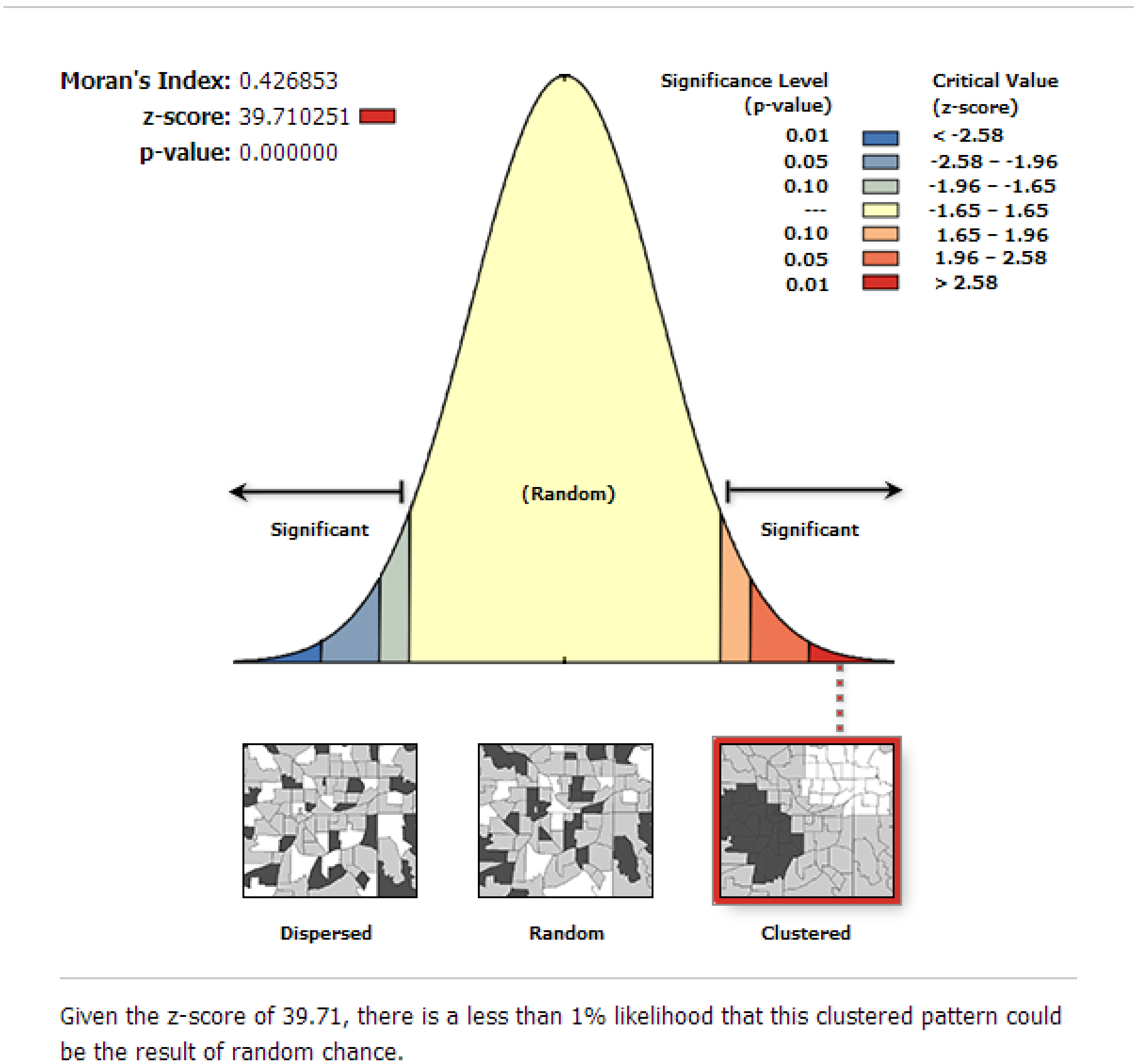


Figure 4.5: County-Level Global Moran's I Output

there is no guarantee that the clustering is of significant magnitude for global importance. The z-score map classified the counties by standard deviation of the calculated index values, yielding a measure of clustering/dispersion magnitude: tier 3 clusters were yellow, tier 2 clusters were orange, and tier 1 clusters were red—further delineating the Atlanta, Georgia metropolitan area.

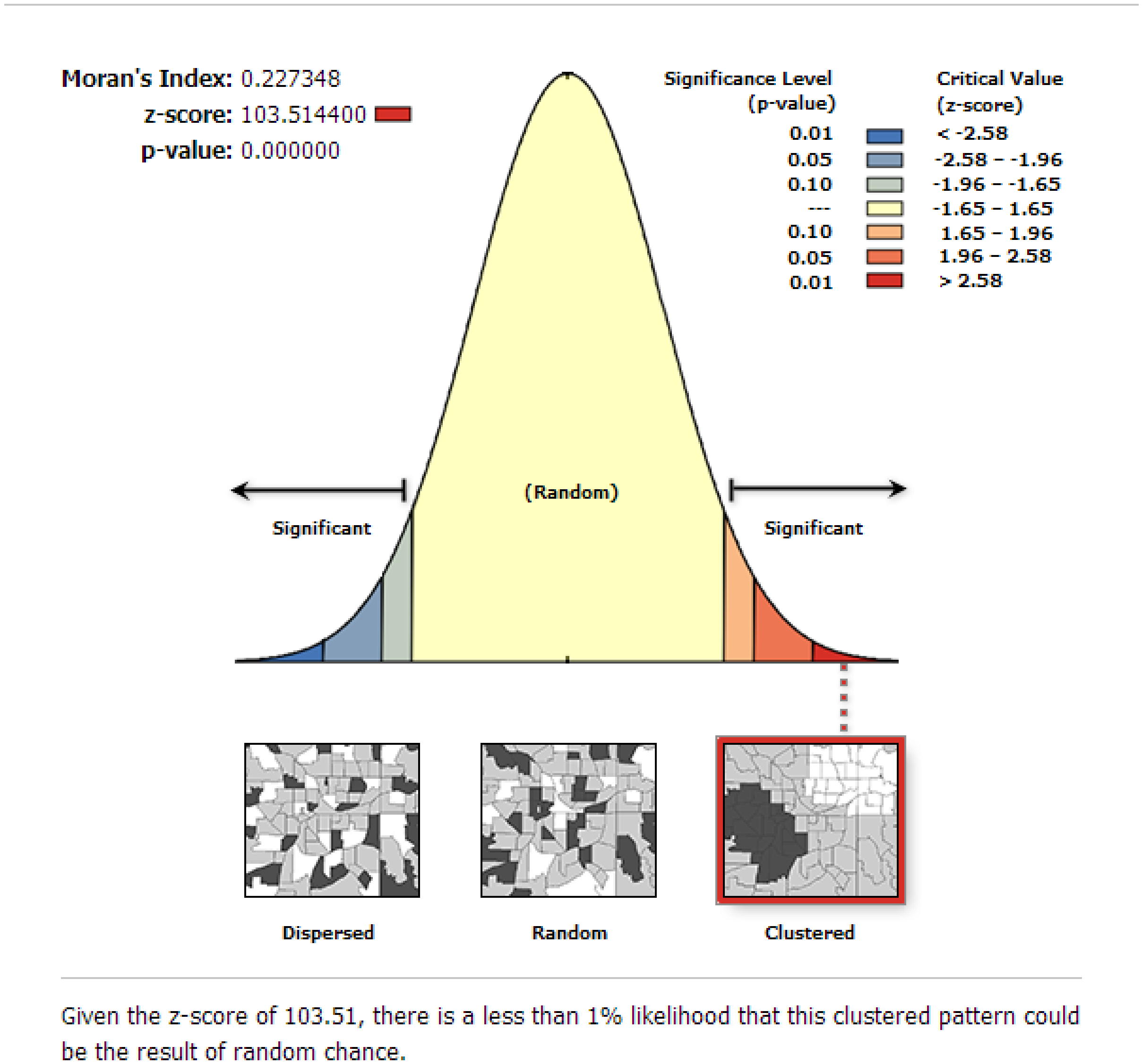


Figure 4.6: Tract-Level Global Moran's I Output

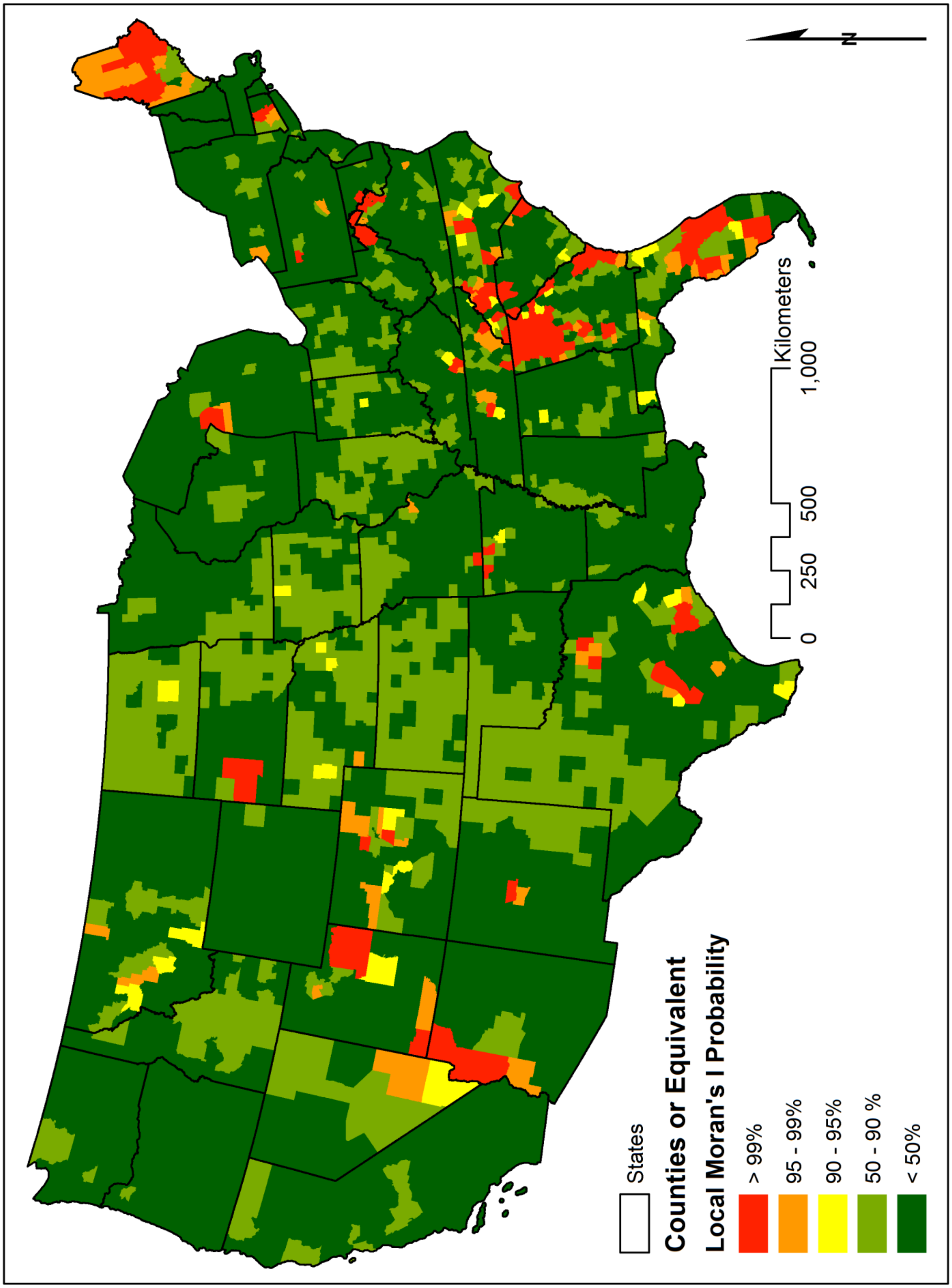


Figure 4.7: County-Level Cluster Probabilities using Local Moran's I

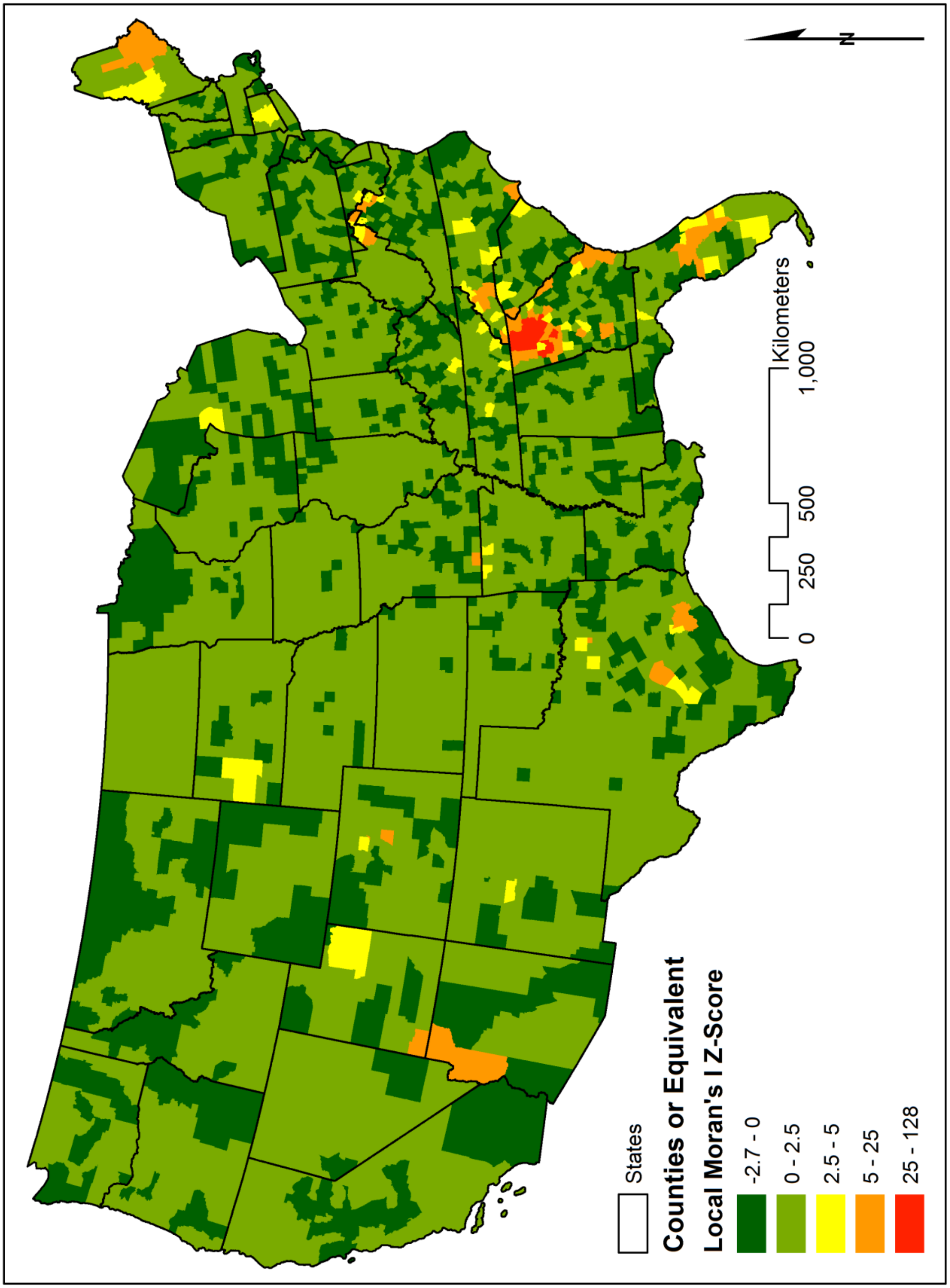


Figure 4.8: County-Level Cluster Z-Scores using Local Moran's I

Multiple Linear Regression

After the descriptive statistics provided information on the central tendencies and upper outliers of the road change database and the Local Moran's I statistic identified spatial clusters of new road growth, multiple linear regressions were utilized in an attempt to discover some driving factors behind the spatial variation of road change across the nation. In order to target metropolitan network expansion rather than rural expansion, data from the county and tract levels of analysis were restricted to the spatial footprints of metropolitan areas. Again, due to the effects of the modifiable areal unit problem, the significance of different explanatory variables changed at different aggregation levels, indicating that road network change is influenced by multi-scalar processes—even the same variables changed in magnitude depending on the scale of analysis; therefore, each of the regression models explained different facets of the phenomenon.

The best-performing model, according to both the coefficient of determination (R^2) and the corrected Akaike Information Criterion (AICc) value, was the MSA model (Table 4.3); this finding was expected due to the smaller sample size and the highly aggregated nature of the dependent variable. The independent variable of mean age was highly insignificant, and therefore dropped from this model, while population density was omitted due to problematic multicollinear interactions with the mean income variable (Table B.1). At this scale, the intercept was not found to be statistically different than zero, and none of the racial categories were deemed significant. Holding all other variables constant, a $10,000 increase in MSA mean income was expected to increase road change percentages by 0.36%, while a population increase of 10% was associated with a 1.39% increase in road change percentage. These regression coefficient results, along with the t-statistic values and probability values, indicate that although income is a relatively significant factor in explaining short-term road network expansion at the

MSA level, concomitant population influxes have more explanatory power for metropolitan areas nationwide. Furthermore, the inclusion of the U.S. Census division dummy variables in the model controlled for geographic differences in MSA racial composition, an important distinction when evaluating racial effects. The East North Central and Pacific divisions were not found to be statistically different from the West North Central reference category, while the South Atlantic, Mountain, and East South Central census divisions were found to be the most statistically influential; metropolitan areas located in the South Atlantic division, on average, experienced 2.78% more road growth than the reference division, Mountain division MSAs gained 1.83% more roads, and East South Central MSAs were found to have 1.09% more growth.

Similar results and explanatory power were discovered at the county level of analysis, where the coefficient of determination was slightly less and the intercept was still not significantly different from zero (Table 4.4). A $10,000 increase in county mean income provided a smaller estimate of a 0.29% road change increase nationwide, while population increases of 10% indicated a strikingly close increase in network expansion of 1.40%, again holding all other variables constant. However, the racial category of percent American Indian became significant at the county level, with road growth slightly decreasing by 0.05% more than the general population for every percentage share in American Indian population; when considering counties within the Mountain and West North Central census divisions with above twenty percent American Indian populations, this effect could translate to a one or more percentage point loss in network growth. However, due to the wider spatial extent of other racial populations, no geographically independent effects were observed for different racial categories. The county census division responses did not vary much from the MSA model, although the Mid-Atlantic region was no longer found to be statistically different from the reference division;

Table 4.3: MSA-Level Linear Regression Output

Regression Statistics	
AICc	1611.4222
R^2	0.2348
Adjusted R^2	0.2038
Standard Error	2.2228
Observations	360

ANOVA

	df	***SS***	***MS***	***F***	***Significance F***
Regression	14	523	37	8	6.89×10^{-14}
Residual	345	1705	5		
Total	359	2228			

	Coefficients	***Std. Error***	***t Stat***	***P-value***
Intercept	-0.791227820	0.9209	-0.8592	0.3908
MeanIncome*	3.5561×10^{-5}	0.0000	2.5142	0.0124
PopulationChange*	0.138878375	0.0541	2.5677	0.0107
%White	N/A	N/A	N/A	N/A
%Black	-0.007353472	0.0132	-0.5570	0.5779
%AmericanIndian	-0.015487906	0.0342	-0.4531	0.6508
%Asian/PacificIslander	0.002284157	0.0135	0.1692	0.8657
%Hispanic	0.009490727	0.0102	0.9277	0.3542
WestNorthCentral	N/A	N/A	N/A	N/A
Pacific	0.097285731	0.4559	0.2134	0.8311
Mountain*	1.834724480	0.6733	2.7248	0.0068
WestSouthCentral*	0.972228055	0.3840	2.5319	0.0118
EastNorthCentral	0.422466401	0.3144	1.3437	0.1799
EastSouthCentral*	1.086333732	0.3868	2.8083	0.0053
SouthAtlantic*	2.780438324	0.5081	5.4727	0.0000
MidAtlantic*	0.757284865	0.3503	2.1619	0.0313
NewEngland	1.284915210	0.7609	1.6886	0.0922

**Variable statistically significant at α=.05*

the influences of South Atlantic and Mountain slightly decreased, while those of East South Central and West South Central increased. Ultimately, the results from the county model did not largely alter the MSA-level perceptions of independent variable magnitude and significance.

Table 4.4: County-Level Linear Regression Output

Regression Statistics	
AICc	5647.0910
R^2	0.1951
Adjusted R^2	0.1848
Standard Error	3.1304
Observations	1100

ANOVA

	df	***SS***	***MS***	***F***	***Significance F***
Regression	14	2578	184	19	1.52×10^{-42}
Residual	1085	10632	10		
Total	1099	13210			

	Coefficients	***Std. Error***	***t Stat***	***P-value***
Intercept	-0.289439050	0.5571	-0.5196	0.6035
MeanIncome*	3.0988×10^{-5}	0.0000	3.8231	0.0002
PopulationChange*	0.139845858	0.0260	5.3743	0.0000
%White	N/A	N/A	N/A	N/A
%Black	-0.008977959	0.0086	-1.0466	0.2955
%AmericanIndian*	-0.051883675	0.0203	-2.5599	0.0106
%Asian/PacificIslander	0.022746263	0.0236	0.9655	0.3345
%Hispanic	0.014135549	0.0083	1.7105	0.0875
WestNorthCentral	N/A	N/A	N/A	N/A
Pacific	-0.331405851	0.3746	-0.8848	0.3765
Mountain*	1.579685584	0.4412	3.5805	0.0004
WestSouthCentral*	0.993449838	0.2962	3.3534	0.0008
EastNorthCentral	0.007079975	0.2004	0.0353	0.9718
EastSouthCentral*	1.321117286	0.2895	4.5634	0.0000
SouthAtlantic*	2.695993690	0.3857	6.9890	0.0000
MidAtlantic	-0.006204765	0.2732	-0.0227	0.9819
NewEngland	0.630012297	0.4606	1.3678	0.1717

**Variable statistically significant at α=.05*

However, the tract-level analysis was drastically different than the county and MSA models, exhibiting an extremely low value for the coefficient of determination and a six-digit value for the corrected Akaike Information Criterion, both of which indicated a poorly-fitting and underperforming model (Table 4.5). In contrast to the two previous models, the intercept was statistically significant and quite high, likely due to the lack of availability for population change at this aggregation level and the newly significant independent variables of mean age and population density. Therefore, instead of the default case of zero values for income, population change, and non-White racial percentages within the West North Central census division, we now have assumptions of zero for population density and mean age. The influence of mean income stayed relatively constant at a 0.26% increase in road change for every additional $10,000 of tract-level average income, while an increase in population density by one thousand individuals per square kilometer would cause an estimated 0.17% decrease in network length, consistent with the expected sign of the population density coefficient. For every ten year increase in average age, estimated additional network lengths decreased by 1.30%, reflecting the mobility of younger households and their participation in new low-density residential areas. After controlling for all other variables, but specifically income and geography, all racial variables were found to be statistically significant, while the effects of Hispanic background failed to differ from White race. At the tract level, all racial categories experienced less growth than White and Hispanic except the Asian/Pacific Islander racial category, where for every 10% increase in the share of Asian/Pacific Islander population, 0.45% road growth was expected above and beyond White/Hispanic categories. Likewise, for a 10% increase in the population share of Black race, an 0.18% decrease in road growth was observed, while a 10% increase in the American Indian category was associated with road growth reductions of 0.83%. Compared to the MSA level, the Mid-Atlantic division switched to a negative sign, while the Pacific, East North Central, and New England divisions also switched to statistically negative signs, possibly due to the lack of mediating influence from the population change variable. However, all four of

Table 4.5: Tract-Level Linear Regression Output

Regression Statistics	
AICc	455227.1668
R^2	0.0351
Adjusted R^2	0.0349
Standard Error	11.0212
Observations	59602

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	15	263344	17556	145	0
Residual	59586	7237699	121		
Total	59601	7501043			

	Coefficients	*Std. Error*	*t Stat*	*P-value*
Intercept*	6.113204406	0.3167	19.3000	0.0000
MeanIncome*	2.5902×10^{-5}	0.0000	18.3076	0.0000
MeanAge*	-0.130010573	0.0082	-15.9159	0.0000
PopDensity*	-0.000168585	0.0000	-19.4798	0.0000
%White	N/A	N/A	N/A	N/A
%Black*	-0.017784041	0.0014	-12.8157	0.0000
%AmericanIndian*	-0.082960340	0.0094	-8.8155	0.0000
%Asian/PacificIslander*	0.045153432	0.0051	8.8281	0.0000
%Hispanic	-0.000207763	0.0049	-0.0421	0.9664
WestNorthCentral	N/A	N/A	N/A	N/A
Pacific*	-0.745333651	0.1981	-3.7631	0.0002
Mountain*	2.225035167	0.2793	7.9666	0.0000
WestSouthCentral*	2.359867506	0.2388	9.8836	0.0000
EastNorthCentral*	-0.480000774	0.1281	-3.7466	0.0002
EastSouthCentral*	1.210660654	0.1580	7.6624	0.0000
SouthAtlantic*	2.655173520	0.1940	13.6886	0.0000
MidAtlantic*	-0.488416921	0.1333	-3.6653	0.0003
NewEngland*	-0.619652575	0.1435	-4.3169	0.0000

**Variable statistically significant at α=.05*

the original positive and statistically significant divisions retained their signs; while the South Atlantic and East South Central regions retained similar coefficients, the values for Mountain and West South Central greatly increased in influence. After controlling for the other factors, the

South Atlantic division exhibited a 2.66% increase in road growth values, followed by West South Central (2.36%), Mountain (2.23%), and East South Central (1.21%).

Overall, the relative importance of regional geography decreased as the analysis approached finer aggregation scales; while it was still the most important factor in each of the models, with population change as a close second, other local factors increasingly had greater explanatory power within each progressive model (Table 4.6). Both the mean income and race variables accounted for larger proportions of model goodness-of-fit as the aggregation units decreased in area. While race was not as influential as geography across the regression models, all racial categories gained statistical significance at the tract level; this indicated either that the moderating influence of the absent population change variable caused race to become significant or that the variables only locally exert influence. In order to investigate further, population change was experimentally removed from the county-level model; the racial coefficients did not change much in magnitude or become significant, so it may be concluded that the racial variables made a valid contribution to the tract-level model. Further research on the tract-level drivers of development is desired, as local elements of real estate development are neglected in this model.

Table 4.6: Comparison of Explanatory Power (Adjusted R^2) across Aggregation Schemes

	MSA	**County**	**Tract**
Demographics	0.1026	0.0851	N/A
Economics	0.0136	0.0294	0.0061
Geography	0.1560	0.1006	0.0165
Race/Ethnicity	0.0201	0.0214	0.0115
Total Model	**0.2038**	**0.1848**	**0.0349**

Geographically Weighted Regression

Although around twenty percent of the variation in road network change could be explained by the county-level multiple regression model, spatial autocorrelation was only loosely addressed at the census division level. In actuality, autocorrelation is not simply a regional phenomenon but may be exhibited at multiple scales, magnitudes, and geographic extents, as evidenced by the Local Moran's I clustering maps. Therefore, geographically weighted regression was used in an attempt to improve the explicit modeling of spatial autocorrelation. The integrated AICc minimization function in ArcGIS indicated that the ideal number of neighbors for the GWR was 135, so independent regression models were created for each of the 3143 observations, with each model considering only independent variable data from the closest 135 counties. While the original county model had an R^2 value of 0.1951 and a standard error of 3.1304, the GWR models had an average local R^2 value of 0.2382 and a significantly lower mean standard error of 2.1533. When treating the individual GWR models as a single aggregate model, the global R^2 value jumped up to 0.5106, with an adjusted value of 0.4431 due to the increased degrees of freedom required with global GWR model interpretation.

Furthermore, when the variables of the original county model were reduced to the three variables used in the GWR model (mean income, population change, and Asian/Pacific Islander), the goodness-of-fit dropped to 0.1829, further indicating the superiority of the GWR model. Additionally, the GWR model reduced the AICc value from 15405 to 14400, where a lower value indicates that less information is left unexplained by the model (Table 4.7). The AICc value of a fully-specified OLS model was even higher, at 15423, suggesting that the gains in global and local R^2 from the GWR model were an improvement over the original OLS model.

Table 4.7: County-Level Comparison of OLS and GWR Models

	Ordinary Least Squares	***Geographically Weighted Regression***
Observations	3143	3143
Neighbors	N/A	135
AICc	15405	14400
Global R^2	.1829	.5106
Adjusted R^2	.1790	.4431
Average Local R^2	.1829	.2382

As the AICc and R^2 values are preferred for comparing regression models of different methods and independent variables, the geographically weighted model should be treated as definitive.

Furthermore, the creation of individual models allows for exploration of the variations in model fit across the United States, and can serve as a helpful visualization tool. Figure 4.9 demonstrates how the local decomposition of an OLS regression can provide information on where the model predicts the dependent variable well, and equally important, Figure 4.10 shows how a GWR model can identify counties where the model has difficulty. The clustering of local R^2 values demonstrates that the states of Idaho, Montana, North Dakota, West Virginia, and New York have low explanatory power; however, this does not necessarily mean that the model had difficulty in predicting that area, just that the supplied independent variables were insufficient to capture the variation in road network change. Many of the areas with R^2 values of less than 0.1 are rural, and therefore may not be extremely relevant for metropolitan road expansion; however, the lack of predictive power around Los Angeles, CA; Albuquerque, NM; and southeast Florida perhaps pointed to the fact that Hispanic background was dropped from the GWR model. The pattern of residuals suggested that the most difficult areas to predict are located around Atlanta, Georgia, which is likely due to a large amount of variation within a small geographical area. The neighborhood of 135 counties may have been too large for this area; future research should aim to better explain the variation in this metropolitan area and isolate these unknown local effects.

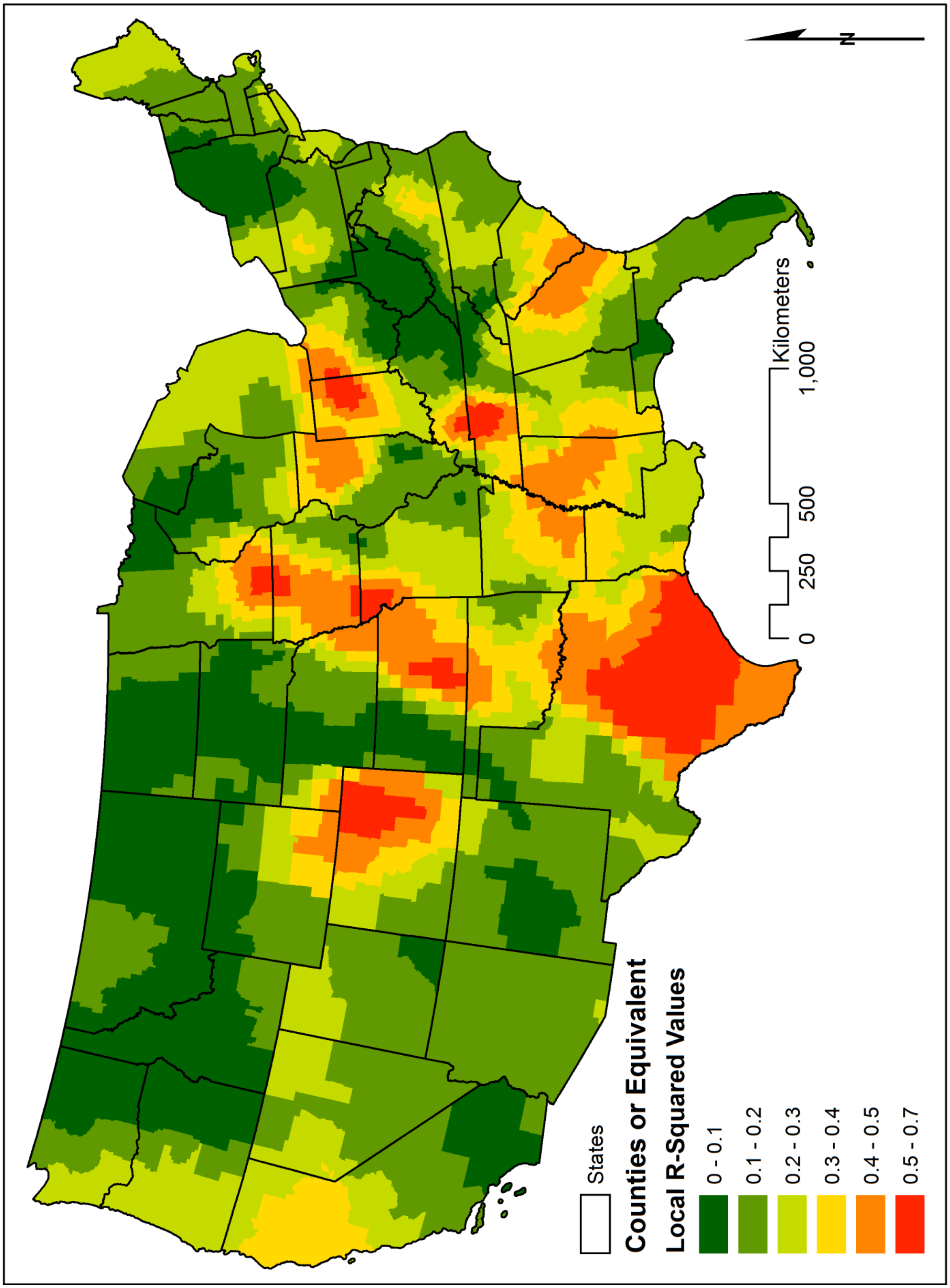


Figure 4.9: Local R^2 Map of County-Level Geographically Weighted Regression

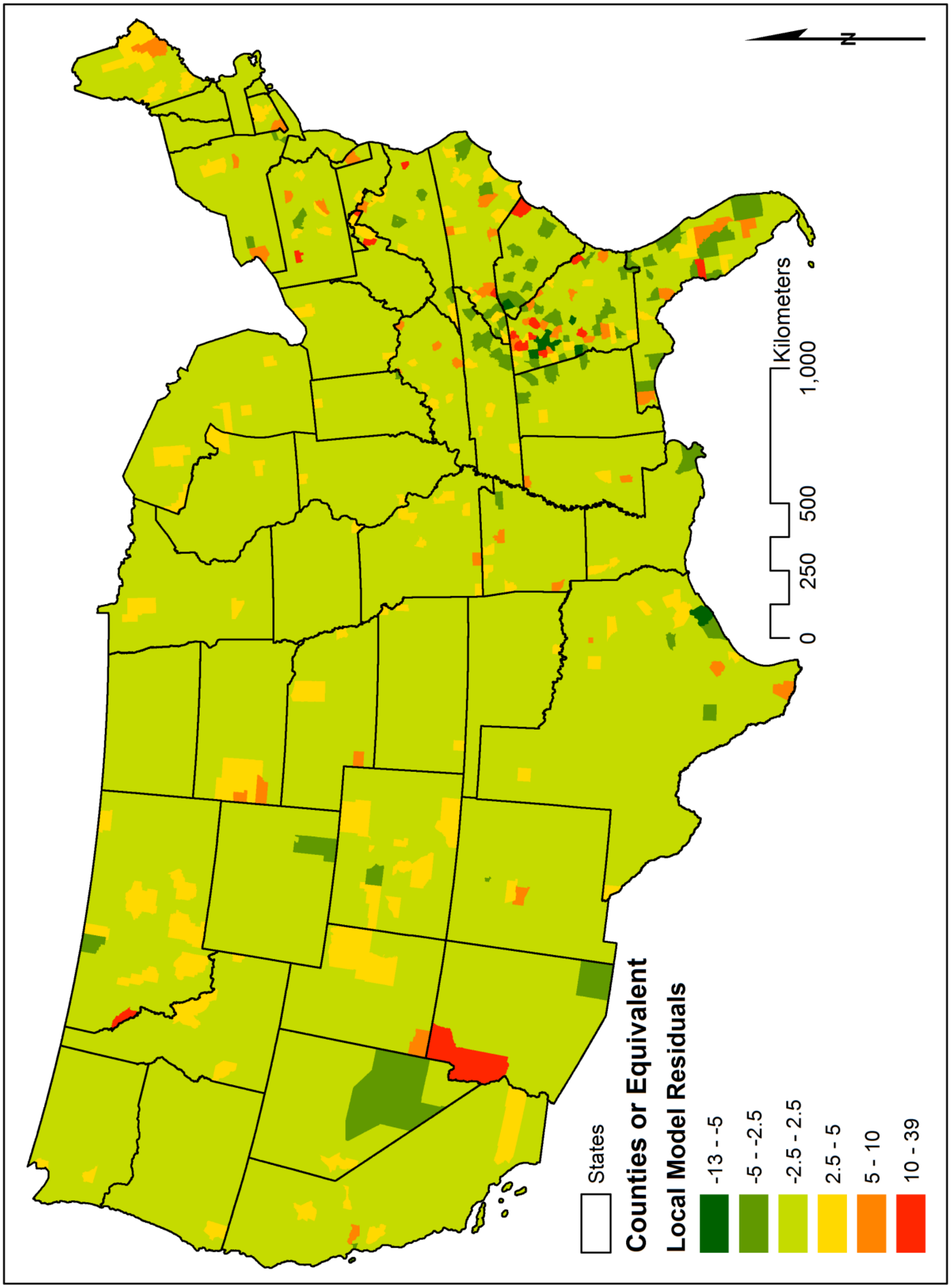


Figure 4.10: Model Residual Map of County-Level Geographically Weighted Regression

CHAPTER 5

CONCLUSION

Summary and Limitations

The construction of a comprehensive spatiotemporal road change database through Python programming enabled aggregate descriptive statistics to be derived for each state, MSA, county, and census tract nationwide; from these statistics, percentage road changes for MSAs, counties, and census tracts served as the dependent variables for three multiple linear regression models. These models indicated that the primary mechanisms of metropolitan growth were due to differences between U.S. Census divisions, while the relative importance of population change and race increased with finer spatial resolution, likely due to an exertion of bounded or targeted effects. Given that the data exhibited moderate spatial autocorrelation, a geographically weighted regression was also executed at the county level, but was found to be inferior to the original model; counties surrounding Atlanta, Georgia produced highly statistically significant outliers, suggesting that a region-specific model would be necessary to better understand local variability.

As each of these results were all from expository analyses, their main functions were to instigate debate and to underscore the importance of geographic variation and socioeconomics in road change scholarship. The presentation of descriptive statistics, spatial autocorrelation indices, linear regression models, and geographically weighted neighborhood regressions, however, would not have been possible without the aid of public data resulting from the TIGER/Line accuracy improvement project. As surface transportation database creators improve their data collection and quality control measures, spatial, temporal, and attribute errors should decrease in magnitude; therefore, in the next decade, we should expect to find longer and more reliable road network time series data disseminated at higher temporal resolutions. In the meantime, though, it is hoped that the methodological framework advanced herein will provide direction on how to proceed with analysis on 2008-2012+ vintage data and deal with their endemic accuracy concerns. Ultimately, the accuracy of the nationwide road change database determined the error profile of the Python change detection scripts, which in turn provided the foundation for the descriptive statistics. Therefore, when considering the summary of results, we must remember that comprehensive quality control and assurance policies were not feasible for a database of this scope and magnitude; thus, spatiotemporally distinct errors may still be present. However, the results of the four types of analyses were dependent upon a set of methodological assumptions aimed at minimizing systematic errors at the aggregate scale; consequently, these considerations will be revisited and their implications will be discussed.

First, an accuracy buffer was identified according to the observed positional displacements of a 150 observation stratified sampling distribution within the Atlanta MSA. While the expected 95th percentile accuracy threshold for the nation was +/- 7.6m, the sample exhibited a spatial accuracy of +/- 7.55m; however, the number of samples could have been increased to better capture the distribution of problematic outliers, and reference positions could have been verified by engineering-grade surveying measurements. In order to account for these potential errors in outlier characterization and reference imagery rectification, the positional accuracy threshold for the road change analyses was extended to the 99th percentile value of 22.75m, and further padded with an extra 2.25 meters to achieve an even 25m error tolerance. The establishment of this error tolerance informed the key Python geoprocessing scripts for delineating new roads, so it indirectly provided the key foundation for all of the calculated statistics. Roads from a later annual dataset that were located more than 25m from the earlier network were considered new roads; whereas a smaller confidence interval would have created a possibility for false positives, a larger buffer distance would have eliminated true road change.

Statistics derived from these new road shapefiles presupposed that the earlier dataset was shown to have the same level of attribute accuracy as the newer dataset. Because the accuracy assessment did not specifically address the temporal differences in attribute accuracy across the entire nation (only for the Atlanta MSA), a substantial number of national forest logging roads, private resource extraction roads, service roads, and driveways in the 2008 and 2009 annual datasets were discovered upon close visual analysis of several states. These attribute errors improved over time, so they were corrected by clipping the older datasets to the positional buffer around the newest dataset (in this case, 2012); therefore, roads that were not located within the new dataset buffer were omitted from the analysis, preventing artificial inflation of length values for the earlier datasets (which in some cases exceeded thirty percent).

Finally, temporal update accuracies were a significant concern for reporting annual road change statistics; a small proportion of smaller counties appeared to update their road extent on an irregular basis, yielding one year with zero percent change and the next compensating for two years worth of road growth. These temporal errors were corrected by only reporting annual change statistics at the MSA level, where the inclusion of multiple counties would reduce the impact of any offending areas. Furthermore, only statistics for the entire five-year period were reported at the county level, in order to average out the potential effects of a single-year omission; therefore, the source of temporal error was mitigated through temporal aggregation. By systematically addressing these three sources of error, the reliability of the descriptive statistics and regression model coefficients for each level of aggregation was maximized.

Future Research

Although the descriptive statistics are relatively immutable, the programming framework used to produce them could be improved for flexibility, efficiency, and ease of use; while researchers may currently utilize and alter the ArcGIS toolbox to develop their own output files, the database structure was hardcoded into the Python scripts, limiting potential applications. Additionally, scholars should further examine the accuracies and limitations of TIGER/Line data for road change analysis, as only a small subset of the data was examined for systematic errors; given that this data source for road change detection has not yet matured, more guidance on its appropriate usage would unequivocally benefit the community of transportation researchers.

Most importantly, continued research should be conducted on the causes and effects of road network expansion, as it is a phenomenon that is still debated in the transportation literature. The presented multiple regression models certainly do not account for all possible independent variables, and a better understanding of the fine-scale interactions between race and road development (especially for low-density, "sprawling" residential areas) would be beneficial for urban social geographers. Specifically, the separation of aggregate road change statistics into arterial and local roads could further elucidate the potential roles of race endemic to either public or private investment regimes. Spatial and temporal lag models, along with added information on human behavior in metropolitan real estate markets, may further increase explanatory power.

Finally, as nearly every expository analysis revealed, road network changes surrounding Atlanta are highly significant and of great magnitude; furthermore, the variations in this region are not sufficiently explained by the nationwide models. Therefore, the last recommendation would be for future scholarship to emphasize this anomalous area, devoting considerable attention to characterizing the undiscovered processes that fuel its rapid road network expansion.

REFERENCES

Anselin, Luc. "Local indicators of spatial association—LISA." *Geographical Analysis* 27, no. 2 (1995): 93-115.

Armenakis, Costas. "Estimation and organization of spatio-temporal data." In *Proceedings of the Canadian Conference on GIS 1992, Ottawa, Canada*. 1992.

Breusch, Trevor S., and Adrian R. Pagan. "A simple test for heteroscedasticity and random coefficient variation." *Econometrica: Journal of the Econometric Society* (1979): 1287-1294.

Broome, Frederick R., and Leslie S. Godwin."Partnering for the people: Improving the US Census Bureau's MAF/TIGER database." *Photogrammetric Engineering and Remote Sensing* 69, no. 10 (2003): 1119-1126.

Bullard, Robert D. "Addressing Urban Transportation Equity in the United States." *Fordham Urban Law Journal* 31 (2003): 1183.

Chen, Ching-Chien, Craig A. Knoblock, and Cyrus Shahabi. "Automatically conflating road vector data with orthoimagery." *GeoInformatica* 10, no. 4 (2006): 495-530.

Ciscel, David H. "The economics of urban sprawl: Inefficiency as a core feature of metropolitan growth." *Journal of Economic issues* (2001): 405-413.

Clarke, Keith C., and Leonard J. Gaydos. "Loose-coupling a cellular automaton model and GIS: long-term urban growth prediction for San Francisco and Washington/Baltimore." *International Journal of Geographical Information Science* 12, no. 7 (1998): 699-714.

Dalton, Nicholas S., John Peponis, and R. Conroy-Dalton. "To tame a TIGER one has to know its nature: extending weighted angular integration analysis to the description of GIS road-centerline data for large scale urban analysis." In *Proceedings of the 4th International Space Syntax Symposium, London, UK* (2003).

Epstein, Jeanne, Karen Payne, and Elizabeth Kramer. "Techniques for mapping suburban sprawl." *Photogrammetric Engineering & Remote Sensing* 63, no. 9 (2002): 913-918.

Feyrer, James. *Trade and Income—Exploiting Time Series in Geography*. No. w14910. National Bureau of Economic Research, 2009.

Forman, Richard T. T., and Lauren E. Alexander." Roads and their major ecological effects." *Annual Review of Ecology and Systematics* (1998).

Georgia GIS Clearinghouse. "Roads & Highways - Base Map." https://data.georgiaspatial.org/ (2012).

Goodchild, Michael F., and Gary J. Hunter. "A simple positional accuracy measure for linear features." *International Journal of Geographical Information Science* 11, no. 3 (1997): 299-306.

Gregory, Ian N., and Richard G. Healey. "Historical GIS: structuring, mapping and analysing geographies of the past." *Progress in Human Geography* 31, no. 5 (2007): 638-653.

Ham, Heejoo, Tschangho John Kim, and David Boyce. "Assessment of economic impacts from unexpected events with an interregional commodity flow and multimodal transportation network model." *Transportation Research Part A: Policy and Practice* 39, no. 10 (2005): 849-860.

Harvey, David. *The urbanization of capital.* Johns Hopkins University Press, 1985.

Harvey, David. "Globalization and the “Spatial fix”." *Geographische Revue* 2 (2001): 23-30.

Healey, Richard G., and Trem R. Stamp. "Historical GIS as a foundation for the analysis of regional economic growth: Theoretical, methodological, and practical issues." *Social Science History* 24, no. 3 (2000): 575-612.

Henderson, Jason. "The politics of mobility and business elites in Atlanta, Georgia." *Urban Geography* 25, no. 3 (2004): 193-216.

Hu, Zhiyong, and C. P. Lo. "Modeling urban growth in Atlanta using logistic regression." *Computers, Environment and Urban Systems* 31, no. 6 (2007): 667-688.

Jaret, Charles, Robert M. Adelman, and Lesley Williams Reid. "Suburban Sprawl, Racial Segregation, and Spatial Mismatch in Metropolitan America." *Sociation Today* 4, no. 2 (2006).

Kenworthy, Jeffrey R., and Felix B. Laube. "Patterns of automobile dependence in cities: an international overview of key physical and economic dimensions with some implications for urban policy." *Transportation Research Part A: Policy and Practice* 33, no. 7 (1999): 691-723.

Kim, Jaecheol. "Testing the Street Connectivity of New Urbanism Projects and their Surroundings in Metro Atlanta Region."In *Proceedings, 6th International Space Syntax Symposium, Istanbul, Turkey* (2007).

Koenker, Roger. "A note on studentizing a test for heteroscedasticity." *Journal of Econometrics* 17, no. 1 (1981): 107-112.

Langran, Gail. *Time in geographic information systems*. Taylor & Francis, 1992.

Liadis, John S. "GPS TIGER accuracy analysis tools (GTAAT) evaluation and test results." *TIGER Operation Branch, Geography Division in the US Census Bureau.* http://www.census.gov/geo/www/tiger/gtaat2000.pdf (2000).

Levinson, David. "Identifying winners and losers in transportation." *Transportation Research Record: Journal of the Transportation Research Board* 1812 (2002): 179-185.

Levinson, David, and Ramachandra Karamalaputi. "Induced supply: a model of highway network expansion at the microscopic level." *Journal of Transport Economics and Policy* 37, no. 3 (2003): 297-318.

Moore, Gordon E. "Cramming more components onto integrated circuits." *Proceedings of the IEEE* 86, no. 1 (1998): 82-85.

Moran, Patrick A. P. "Notes on continuous stochastic phenomena." *Biometrika* (1950): 17-23.

Openshaw, Stan. "Concepts and techniques in modern geography number 38: the modifiable areal unit problem." *Norwick: Geo Books* (1984).

Pelekis, Nikos, Babis Theodoulidis, Ioannis Kopanakis, and Yannis Theodoridis. "Literature review of spatio-temporal database models." *The Knowledge Engineering Review* 19, no. 03 (2004): 235-274.

Peponis, John, Douglas Allen, Dawn Haynie, Martin Scoppa, and Zongyu Zhang."Measuring the configuration of street networks." In *Proceedings, 6th International Space Syntax Symposium, Istanbul, Turkey* (2007).

Perkal, Julian. "On the length of empirical curves." *Michigan Inter-University Community of Mathematical Geographers* 10 (1966): 17-20.

Riitters, Kurt H., and James D. Wickham. "How far to the nearest road?" *Frontiers in Ecology and the Environment* 1, no. 3 (2003): 125-129.

Sanchez, Thomas W., Rich Stolz, and Jacinta S. Ma. *Moving to equity: addressing inequitable effects of transportation policies on minorities*. Civil Rights Project at Harvard University, 2003.

Schuurman, Nadine, Darrin Grund, Michael Hayes, and Suzana Dragicevic. "Spatial/temporal mismatch: a conflation protocol for Canada Census spatial files." *The Canadian Geographer/ Le Géographe canadien* 50, no. 1 (2006): 74-84.

Sugiura, Nariaki. "Further analysis of the data by Akaike's information criterion and the finite corrections." *Communications in Statistics-Theory and Methods* 7, no. 1 (1978): 13-26.

Taaffe, Edward J., Richard L. Morrill, and Peter R. Gould. "Transport expansion in underdeveloped countries: a comparative analysis." *Geographical Review* 53, no. 4 (1963): 503-529.

Tobler, Waldo R. "A computer movie simulating urban growth in the Detroit region." *Economic Geography* 46 (1970): 234-240.

United States Census Bureau (U.S. Census Bureau). "American FactFinder." http://factfinder2.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t (2012a).

United States Census Bureau (U.S. Census Bureau). "Index of geo/tiger." ftp://ftp2.census.gov/geo/tiger/ (2012b).

United States Census Bureau (U.S. Census Bureau). "Population Estimates." http://www.census.gov/popest/data/historical/index.html (2012c).

United States Department of Agriculture (USDA). "ArcGIS Web Service." http://gis.apfo.usda.gov/arcgis/services/NAIP/Georgia_2010_1m_NC (2010).

United States Department of Energy. "Regional Energy Profiles: U.S. Census Regions and Divisions." http://www.eia.gov/emeu/reps/maps/us_census.html (2000).

Weber, Joe, and Selima Sultana. "Journey-to-Work Patterns in the Age of Sprawl: Evidence from Two Midsize Southern Metropolitan Areas." *The Professional Geographer* 59, no. 2 (2007): 193-208.

Xie, Feng, and David Levinson."Modeling the growth of transportation networks: A comprehensive review." *Networks and Spatial Economics* 9, no. 3 (2009): 291-307.

Zandbergen, Paul A. "Positional Accuracy of Spatial Data: Non-Normal Distributions and a Critique of the National Standard for Spatial Data Accuracy." *Transactions in GIS* 12, no. 1 (2008): 103-130.

Zandbergen, Paul A., Drew A. Ignizio, and Kathryn E. Lenzer. "Positional accuracy of TIGER 2000 and 2009 road networks." *Transactions in GIS* 15, no. 4 (2011): 495-519.

APPENDIX A

PYTHON GEOPROCESSING SCRIPTS AND OUTPUTS

```python
#Christopher Brown
#TIGER Time-Series Analysis
#Edge to Road Conversion (for TIGER/Line files before 2010)

#import geoprocessing/filesystem modules and enable output overwriting
import arcpy, os
arcpy.env.overwriteOutput = True

### PARAMETERS

workspace = arcpy.GetParameterAsText(0)
edgesyear = arcpy.GetParameterAsText(1)
inWorkspace = workspace + "\\TIGER" + edgesyear + "\\TIGER" + edgesyear + "_edges"
outWorkspace = workspace + "\\TIGER" + edgesyear + "\\TIGER" + edgesyear


### SELECT ROADS FROM EDGE FILES AND EXPORT

arcpy.AddMessage("\nConverting TIGER edges to roads...")
print "\nConverting TIGER edges to roads..."

#set the input workspace
if not os.path.exists(inWorkspace): #check to see if workspace exists
    arcpy.AddMessage("\nInvalid edge input workspace!")
    print "\nInvalid edge input workspace!"
arcpy.env.workspace = inWorkspace

#set the output workspace and coordinate system
if not os.path.exists(outWorkspace): #check to see if workspace exists
    os.mkdir(outWorkspace) #if not, create it

#print input/output directory names
arcpy.AddMessage("\nThe current edge input workspace is: " + "\n\t" + inWorkspace)
arcpy.AddMessage("\nThe output roads will be put in: " + "\n\t" + outWorkspace)
print "\nThe current edge input workspace is: " + "\n\t" + inWorkspace #print current workspace
print "\nThe output roads will be put in: " + "\n\t" + outWorkspace #print output workspace

#create selection set from input folder
fcs = arcpy.ListFeatureClasses("tl_*_edges.shp") #list only TIGER edge files

#execute select function
arcpy.AddMessage("\nSelecting roads from edge files...")
print "\nSelecting roads from edge files..."
for fc in fcs: #for each feature class in the list of feature classes
    outFc = fc[:-9] + "roads.shp" #define the output feature class name using the input
    expression = "ROADFLG = 'Y'" #SQL select all roads from edge records

    #perform calculation
    arcpy.Select_analysis(fc, outWorkspace + "\\" + outFc, expression)
    arcpy.AddMessage("\t" + outFc) #display progress by file
    print "\t" + outFc


### SCRIPT COMPLETED

del fcs
del fc
arcpy.AddMessage("\nConversion from edges to roads completed successfully...\n")
print "\nConversion from edges to roads completed successfully...\n"
```

Figure A.1: Script for Selecting Roads from Edges and Exporting to a New Shapefile

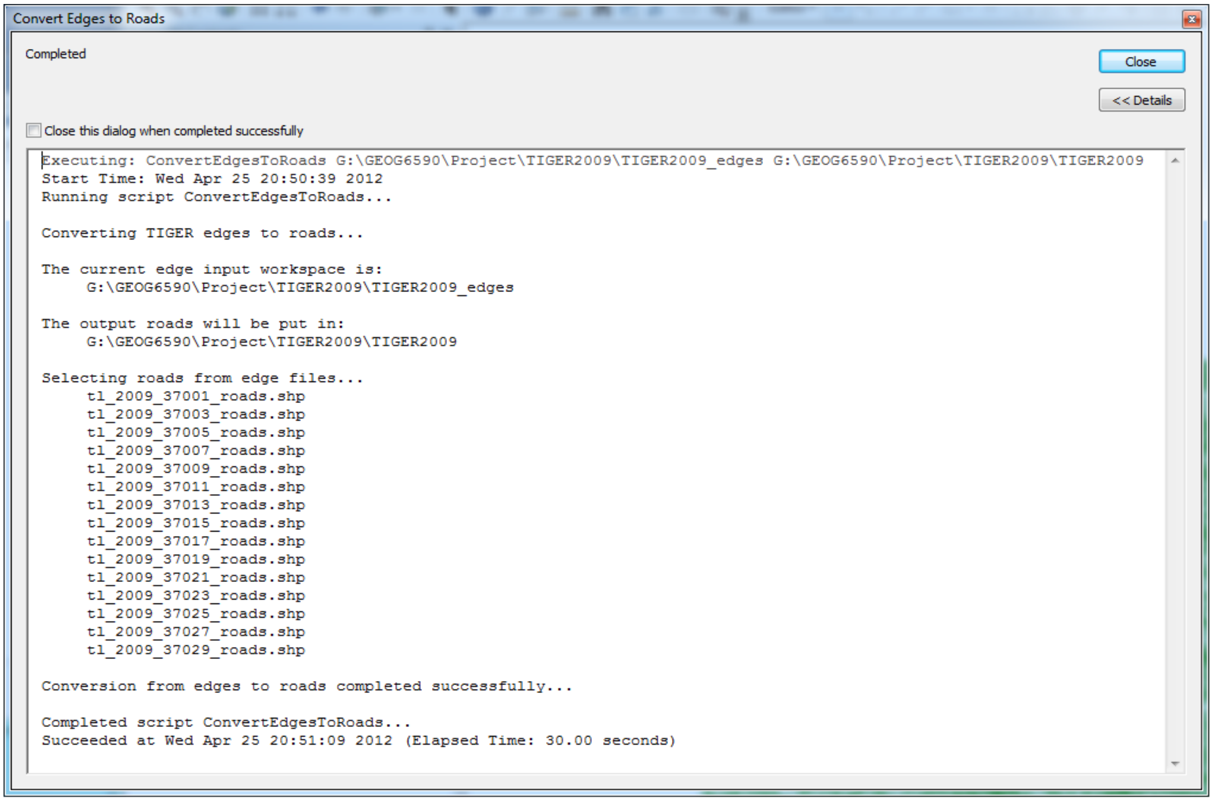


Figure A.2: Output from Convert Edges to Roads Script Tool

```
#Christopher Brown
#TIGER Time-Series Analysis
#United States Master Script File

#import geoprocessing/filesystem modules and roads_state.py
import arcpy, os, roads_state

### PARAMETERS :: WORKSPACE, STATE, YEARS, BUFFER DISTANCE

workspace = arcpy.GetParameterAsText(0)
statestring = arcpy.GetParameterAsText(1)
bufferyear = arcpy.GetParameterAsText(2)
eraseyear = arcpy.GetParameterAsText(3)
bufferdistance = arcpy.GetParameterAsText(4)


#MANUAL PARAMETER ADJUSTMENT
#workspace = os.getcwd()
#statestring = "DE" #use any other abbreviation to process all states
#bufferyear = "2010"
#eraseyear = "2011"
#bufferdistance = "50"


### DICTIONARY TO CONVERT STATE ABBREVIATION TO FIPS CODE

fipsdictionary = {'AL':'01','AK':'02','AZ':'04','AR':'05','CA':'06', \
                  'CO':'08','CT':'09','DE':'10','DC':'11','FL':'12', \
                  'GA':'13','HI':'15','ID':'16','IL':'17','IN':'18', \
                  'IA':'19','KS':'20','KY':'21','LA':'22','ME':'23', \
                  'MD':'24','MA':'25','MI':'26','MN':'27','MS':'28', \
                  'MO':'29','MT':'30','NE':'31','NV':'32','NH':'33', \
                  'NJ':'34','NM':'35','NY':'36','NC':'37','ND':'38', \
                  'OH':'39','OK':'40','OR':'41','PA':'42','RI':'44', \
                  'SC':'45','SD':'46','TN':'47','TX':'48','UT':'49', \
                  'VT':'50','VA':'51','WA':'53','WV':'54','WI':'55', \
                  'WY':'56'}

#return state FIPS code from abbreviation
statefips = str(fipsdictionary.get(statestring[:2]))


### EXECUTE TIME-SERIES ANALYSIS BY STATE

if statefips == "None": #if all states are being processed
    for statestring in fipsdictionary.keys():
        statefips = str(fipsdictionary.get(statestring[:2]))
        roads_state.NewRoadsByState_Function \
            (workspace, statestring, statefips, bufferyear, eraseyear, bufferdistance)
else: #if one state is being processed
    roads_state.NewRoadsByState_Function \
        (workspace, statestring, statefips, bufferyear, eraseyear, bufferdistance)
```

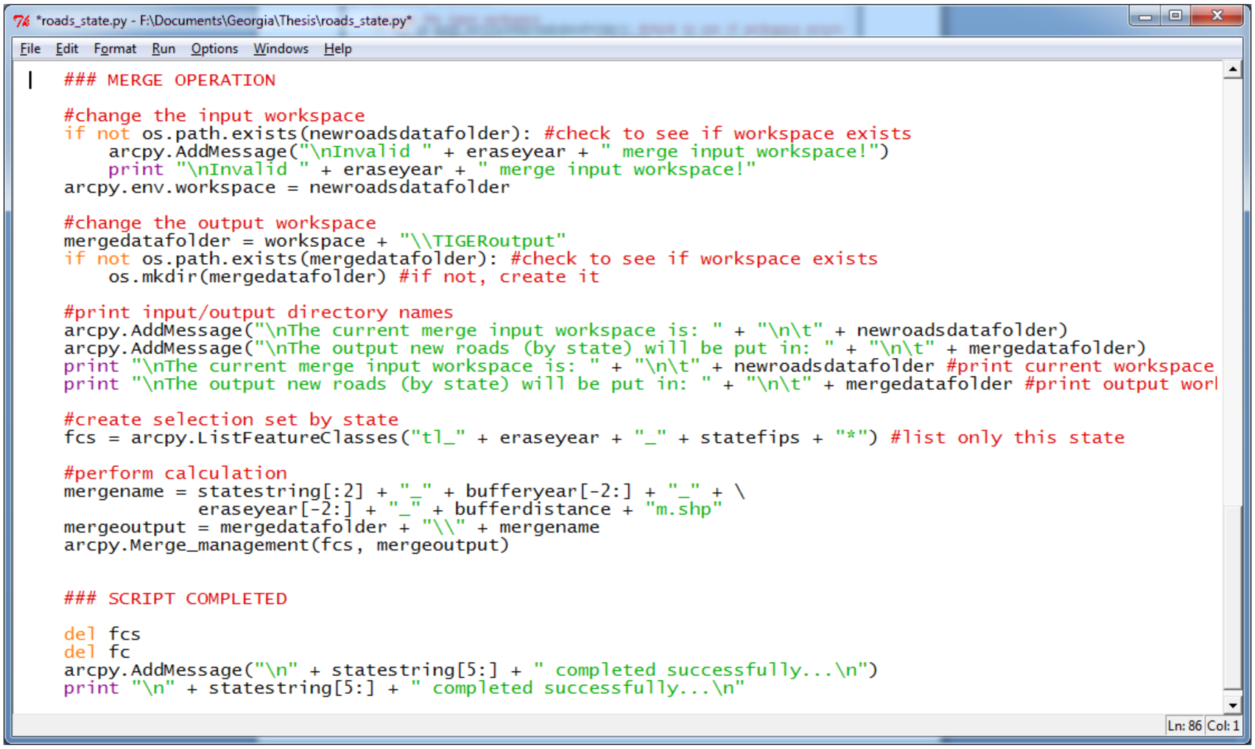
Figure A.3: Script for Retrieving Parameters and Selecting State(s) to Process roads_state.py - E:\Documents\Georgia\GEOG6590\Project\roads_state.py

File Edit Format Run Options Windows Help

```
#Christopher Brown
#TIGER Time-Series Analysis
#State-Level Processing and Folder Manipulation

#import geoprocessing/mapping/filesystem modules, roads_county.py, and enable output overwriting
import arcpy, arcpy.mapping, os, roads_county
arcpy.env.overwriteOutput = True


def NewRoadsByState_Function(workspace, statestring, statefips, bufferyear, eraseyear, bufferdistance):

    ### BUFFER OPERATION

    arcpy.AddMessage("\nFinding new roads in " + statestring[5:] + " from " + bufferyear + " to " + eraseyear + "...")
    print "\nFinding new roads in " + statestring[5:] + " from " + bufferyear + " to " + eraseyear + "..."

    #set the input workspace
    rawdatafolder = workspace + "\\TIGER" + bufferyear + "\\TIGER" + bufferyear
    if not os.path.exists(rawdatafolder): #check to see if workspace exists
        arcpy.AddMessage("\nInvalid " + bufferyear + " buffer input workspace!")
        print "\nInvalid " + bufferyear + " buffer input workspace!"
    arcpy.env.workspace = rawdatafolder

    #set the output workspace and coordinate system
    bufferdatafolder = rawdatafolder + "_buffer_" + bufferdistance + "m"
    if not os.path.exists(bufferdatafolder): #check to see if workspace exists
        os.mkdir(bufferdatafolder) #if not, create it
    arcpy.env.outputCoordinateSystem = "Coordinate Systems\\Projected Coordinate Systems\\Continental\\North America\\USA

    #print input/output directory names
    arcpy.AddMessage("\nThe current buffer input workspace is: " + "\n\t" + rawdatafolder)
    arcpy.AddMessage("\nThe output buffers will be put in: " + "\n\t" + bufferdatafolder)
    print "\nThe current buffer input workspace is: " + "\n\t" + rawdatafolder #print current workspace
    print "\nThe output buffers will be put in: " + "\n\t" + bufferdatafolder #print output workspace

    #create selection set by state
    fcs = arcpy.ListFeatureClasses("tl_" + bufferyear + "_" + statefips + "*") #list only this state

    #execute buffer function
    arcpy.AddMessage("\nBuffering roads in " + statestring[5:] + " for " + bufferyear + " by " + bufferdistance + "m...")
    print "\nBuffering roads in " + statestring + " for " + bufferyear + " by " + bufferdistance + "m..."
    for fc in fcs: #for each feature class in the list of feature classes
        outFc = fc[:-9] + "buffer" + bufferdistance + "m.shp" #define the output feature class name using the input
        distanceargument = bufferdistance + " Meters" #add units to the buffer distance to fit argument syntax

        #perform calculation in separate script to minimize memory leaks
        roads_county.Buffer_Function \
            (fc, bufferdatafolder, outFc, distanceargument) #use data as argument


    ### ERASE OPERATION

    #change the input workspace
    eraserawdatafolder = workspace + "\\TIGER" + eraseyear + "\\TIGER" + eraseyear
    if not os.path.exists(eraserawdatafolder): #check to see if workspace exists
        arcpy.AddMessage("\nInvalid " + eraseyear + " erase input workspace!")
        print "\nInvalid " + eraseyear + " erase input workspace!"
    arcpy.env.workspace = eraserawdatafolder

    #change the output workspace
    newroadsdatafolder = eraserawdatafolder + "_newroads_" + bufferdistance + "m"
    if not os.path.exists(newroadsdatafolder): #check to see if workspace exists
        os.mkdir(newroadsdatafolder) #if not, create it

    #print input/output directory names
    arcpy.AddMessage("\nThe current erase input workspace is: " + "\n\t" + eraserawdatafolder)
    arcpy.AddMessage("\nThe output new roads (by county) will be put in: " + "\n\t" + newroadsdatafolder)
    print "\nThe current erase input workspace is: " + "\n\t" + eraserawdatafolder #print current workspace
    print "\nThe output new roads (by county) will be put in: " + "\n\t" + newroadsdatafolder #print output workspace

    #create selection set by state
    fcs = arcpy.ListFeatureClasses("tl_" + eraseyear + "_" + statefips + "*") #list only this state

    #execute erase function
    arcpy.AddMessage("\nErasing old " + bufferyear + " road buffers from new " + eraseyear + " roads...")
    print "\nErasing old " + bufferyear + " road buffers from new " + eraseyear + " roads..."
    for fc in fcs: #for each feature class in the list of feature classes
        outFc = fc[0:-9] + "erase" + bufferdistance + "m.shp" #define the output feature class name using the input
        eraseFc = bufferdatafolder + "\\" + fc[:5] + bufferyear[-2:] + fc[7:-9] + "buffer" + bufferdistance + "m.shp"

        #perform calculation in separate script to minimize memory leaks
        roads_county.Erase_Function \
            (fc, eraseFc, newroadsdatafolder, outFc)
```

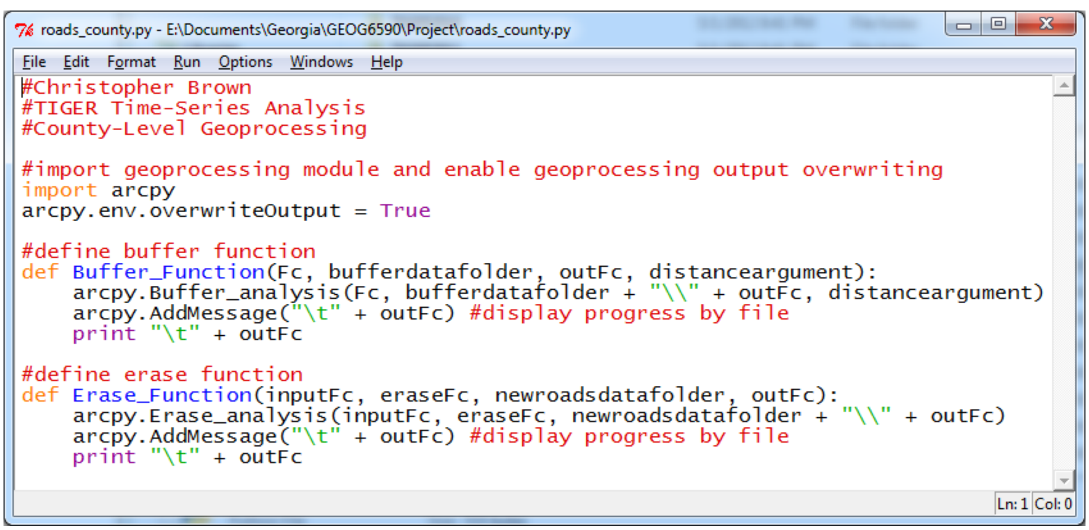
Ln: 1 Col: 0

Figure A.4: Script for Buffering Old Roads and Erasing Buffers from New Roads



```
### MERGE OPERATION

#change the input workspace
if not os.path.exists(newroadsdatafolder): #check to see if workspace exists
    arcpy.AddMessage("\nInvalid " + eraseyear + " merge input workspace!")
    print "\nInvalid " + eraseyear + " merge input workspace!"
arcpy.env.workspace = newroadsdatafolder

#change the output workspace
mergedatafolder = workspace + "\\TIGERoutput"
if not os.path.exists(mergedatafolder): #check to see if workspace exists
    os.mkdir(mergedatafolder) #if not, create it

#print input/output directory names
arcpy.AddMessage("\nThe current merge input workspace is: " + "\n\t" + newroadsdatafolder)
arcpy.AddMessage("\nThe output new roads (by state) will be put in: " + "\n\t" + mergedatafolder)
print "\nThe current merge input workspace is: " + "\n\t" + newroadsdatafolder #print current workspace
print "\nThe output new roads (by state) will be put in: " + "\n\t" + mergedatafolder #print output worl

#create selection set by state
fcs = arcpy.ListFeatureClasses("tl_" + eraseyear + "_" + statefips + "*") #list only this state

#perform calculation
mergename = statestring[:2] + "_" + bufferyear[-2:] + "_" + \
            eraseyear[-2:] + "_" + bufferdistance + "m.shp"
mergeoutput = mergedatafolder + "\\" + mergename
arcpy.Merge_management(fcs, mergeoutput)


### SCRIPT COMPLETED

del fcs
del fc
arcpy.AddMessage("\n" + statestring[5:] + " completed successfully...\n")
print "\n" + statestring[5:] + " completed successfully...\n"
```

Figure A.5: Script for Merging New Roads by State



```
#Christopher Brown
#TIGER Time-Series Analysis
#County-Level Geoprocessing

#import geoprocessing module and enable geoprocessing output overwriting
import arcpy
arcpy.env.overwriteOutput = True

#define buffer function
def Buffer_Function(Fc, bufferdatafolder, outFc, distanceargument):
    arcpy.Buffer_analysis(Fc, bufferdatafolder + "\\" + outFc, distanceargument)
    arcpy.AddMessage("\t" + outFc) #display progress by file
    print "\t" + outFc

#define erase function
def Erase_Function(inputFc, eraseFc, newroadsdatafolder, outFc):
    arcpy.Erase_analysis(inputFc, eraseFc, newroadsdatafolder + "\\" + outFc)
    arcpy.AddMessage("\t" + outFc) #display progress by file
    print "\t" + outFc
```

Figure A.6: Script for Buffering and Erasing Roads at the County Level

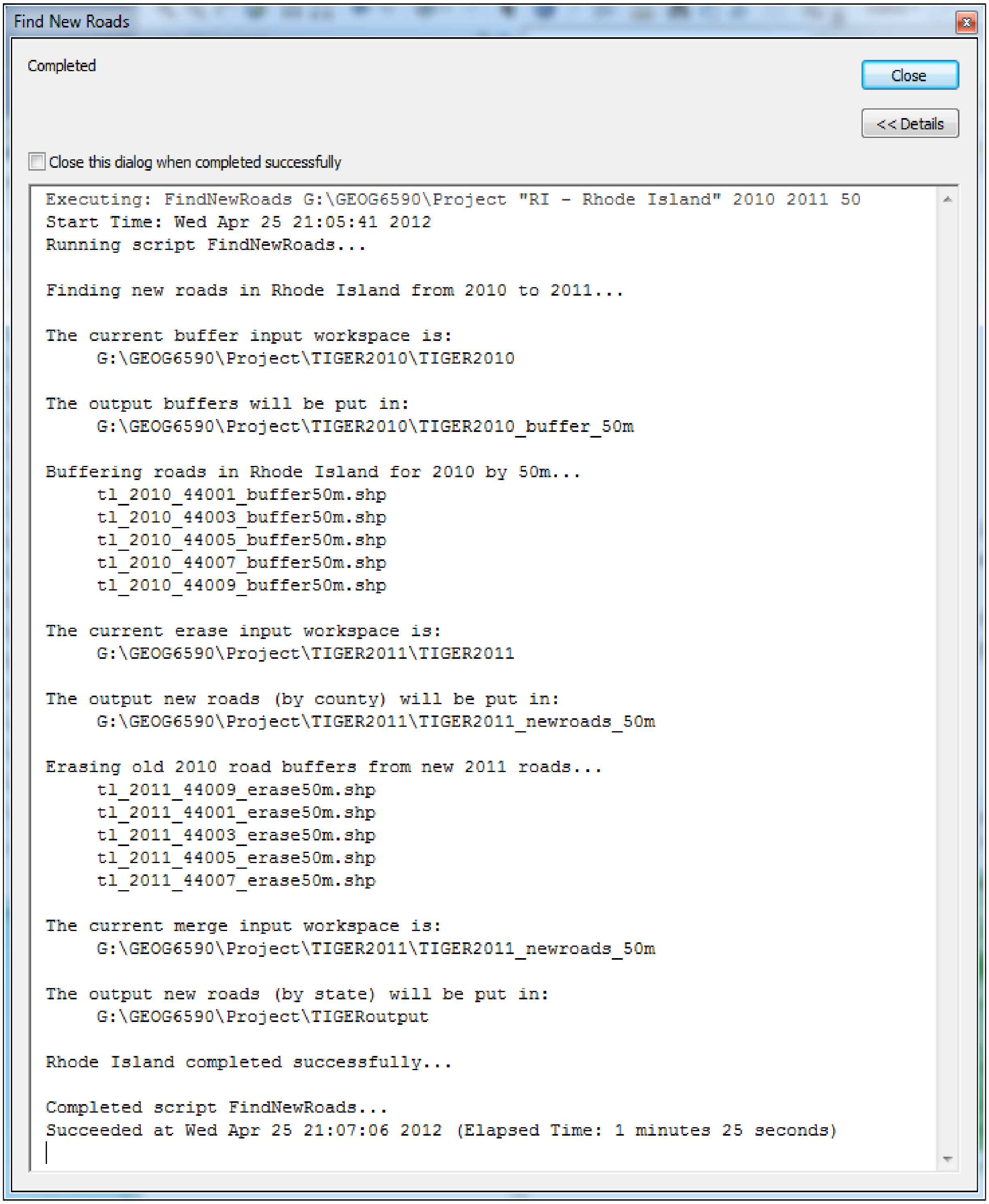


Figure A.7: Output from Find New Roads Script Tool

```
#Christopher Brown
#TIGER County Merging
#United States Master Script File

#import geoprocessing/filesystem modules and enable overwriting
import arcpy, os
arcpy.env.overwriteOutput = True


### MERGE OPERATION

def Merge_Function(workspace, statestring, statefips, mergeyear):

    #change the input workspace
    roadsdatafolder = workspace + "\\TIGER" + mergeyear + "\\TIGER" + mergeyear
    if not os.path.exists(roadsdatafolder): #check to see if workspace exists
        arcpy.AddMessage("\nInvalid " + mergeyear + " merge input workspace!")
        print "\nInvalid " + mergeyear + " merge input workspace!"
    arcpy.env.workspace = roadsdatafolder

    #change the output workspace
    mergedatafolder = workspace + "\\TIGER" + mergeyear + "\\TIGER" + mergeyear + "_states"
    if not os.path.exists(mergedatafolder): #check to see if workspace exists
        os.mkdir(mergedatafolder) #if not, create it

    #print input/output directory names
    arcpy.AddMessage("\nThe current merge input workspace is: " + "\n\t" + roadsdatafolder)
    arcpy.AddMessage("\nThe output roads for " + statestring[5:] + " will be put in: " + "\n\t" + merged
    print "\nThe current merge input workspace is: " + "\n\t" + roadsdatafolder #print current workspace
    print "\nThe output roads for " + statestring[5:] + " will be put in: " + "\n\t" + mergedatafolder #

    #create selection set of county files
    fcs = arcpy.ListFeatureClasses("tl_" + mergeyear + "_" + statefips + "*" + "roads.shp")

    #merge counties into a state file
    mergename = statestring[:2] + "_" + mergeyear[-2:] + ".shp"
    mergeoutput = mergedatafolder + "\\" + mergename
    arcpy.Merge_management(fcs, mergeoutput)

    #clean up
    del fcs
    arcpy.AddMessage("\n" + statestring[5:] + " completed successfully...\n")
    print "\n" + statestring[5:] + " completed successfully...\n"
```

Figure A.8: Script for Merging County Road Shapefiles into State Road Shapefiles

```
### PARAMETERS :: WORKSPACE, STATE, YEARS, BUFFER DISTANCE

workspace = arcpy.GetParameterAsText(0)
statestring = arcpy.GetParameterAsText(1)
mergeyear = arcpy.GetParameterAsText(2)

### MANUAL PARAMETER ADJUSTMENT

#workspace = os.getcwd()
#statestring = "DE - Delaware" #use any other abbreviation to process all states
#mergeyear = "2008"


### DICTIONARY TO CONVERT STATE ABBREVIATION TO FIPS CODE

fipsdictionary = {'AL':'01','AK':'02','AZ':'04','AR':'05','CA':'06', \
                  'CO':'08','CT':'09','DE':'10','DC':'11','FL':'12', \
                  'GA':'13','HI':'15','ID':'16','IL':'17','IN':'18', \
                  'IA':'19','KS':'20','KY':'21','LA':'22','ME':'23', \
                  'MD':'24','MA':'25','MI':'26','MN':'27','MS':'28', \
                  'MO':'29','MT':'30','NE':'31','NV':'32','NH':'33', \
                  'NJ':'34','NM':'35','NY':'36','NC':'37','ND':'38', \
                  'OH':'39','OK':'40','OR':'41','PA':'42','RI':'44', \
                  'SC':'45','SD':'46','TN':'47','TX':'48','UT':'49', \
                  'VT':'50','VA':'51','WA':'53','WV':'54','WI':'55', \
                  'WY':'56'}

#return state FIPS code from abbreviation
statefips = str(fipsdictionary.get(statestring[:2]))


### EXECUTE MERGE ANALYSIS BY STATE

if statefips == "None": #if all states are being processed

    #create state files
    for statestring in fipsdictionary.keys():
        statefips = str(fipsdictionary.get(statestring[:2]))
        Merge_Function(workspace, statestring, statefips, mergeyear)

    #change workspace from county directory to state directory
    arcpy.env.workspace = workspace + "\\TIGER" + mergeyear + \
                          "\\TIGER" + mergeyear + "_states"

    #create selection set of states
    fcs = arcpy.ListFeatureClasses("*" + mergeyear[-2:] + ".shp")

    #merge dissolved states into a national file
    arcpy.Merge_management(fcs, workspace + "\\TIGER" + mergeyear + \
                           "\\" + mergeyear + ".shp")

    #clean up
    del fcs

else: #if one state is being processed
    Merge_Function(workspace, statestring, statefips, mergeyear)
```

Figure A.8 — Continued

merge_buffers.py - F:\Documents\Georgia\Thesis\merge_buffers.py

File Edit Format Run Options Windows Help

```
#Christopher Brown
#TIGER County Buffer Merging
#United States Master Script File

#import geoprocessing/filesystem modules and enable overwriting
import arcpy, os
arcpy.env.overwriteOutput = True


### MERGE OPERATION

def Merge_Function(workspace, statestring, statefips, mergeyear, bufferdistance):

    #change the input workspace
    roadsdatafolder = workspace + "\\TIGER" + mergeyear + "\\TIGER" + mergeyear + \
                      "_buffer_" + bufferdistance + "m"
    if not os.path.exists(roadsdatafolder): #check to see if workspace exists
        arcpy.AddMessage("\nInvalid " + mergeyear + " merge input workspace!")
        print "\nInvalid " + mergeyear + " merge input workspace!"
    arcpy.env.workspace = roadsdatafolder

    #change the output workspace
    mergedatafolder = workspace + "\\TIGER" + mergeyear + "\\TIGER" + mergeyear + \
                      "_statebuffers_" + bufferdistance + "m"
    if not os.path.exists(mergedatafolder): #check to see if workspace exists
        os.mkdir(mergedatafolder) #if not, create it

    #print input/output directory names
    arcpy.AddMessage("\nThe current merge input workspace is: " + "\n\t" + roadsdatafolder)
    arcpy.AddMessage("\nThe output roads for " + statestring[5:] + " will be put in: " + "\n\t" + merged
    print "\nThe current merge input workspace is: " + "\n\t" + roadsdatafolder #print current workspace
    print "\nThe output roads for " + statestring[5:] + " will be put in: " + "\n\t" + mergedatafolder #

    #create selection set of county buffer files
    fcs = arcpy.ListFeatureClasses("tl_" + mergeyear + "_" + statefips + "*" + "buffer" + bufferdistance

    #merge counties into a state file
    mergename = statestring[:2] + "_" + mergeyear[-2:] + "_buffer" + bufferdistance + "m.shp"
    mergeoutput = mergedatafolder + "\\" + mergename
    arcpy.Merge_management(fcs, mergeoutput)

    #clean up
    del fcs
    arcpy.AddMessage("\n" + statestring[5:] + " completed successfully...\n")
    print "\n" + statestring[5:] + " completed successfully...\n"
```

Ln: 1 Col: 0

Figure A.9: Script for Merging County Buffer Shapefiles into State Buffer Shapefiles

```
### PARAMETERS :: WORKSPACE, STATE, YEAR, BUFFER DISTANCE

workspace = arcpy.GetParameterAsText(0)
statestring = arcpy.GetParameterAsText(1)
mergeyear = arcpy.GetParameterAsText(2)
bufferdistance = arcpy.GetParameterAsText(3)

### MANUAL PARAMETER ADJUSTMENT

#workspace = os.getcwd()
#statestring = "DE - Delaware" #use any other abbreviation to process all states
#mergeyear = "2008"
#bufferdistance = "25"


### DICTIONARY TO CONVERT STATE ABBREVIATION TO FIPS CODE

fipsdictionary = {'AL':'01','AK':'02','AZ':'04','AR':'05','CA':'06', \
                  'CO':'08','CT':'09','DE':'10','DC':'11','FL':'12', \
                  'GA':'13','HI':'15','ID':'16','IL':'17','IN':'18', \
                  'IA':'19','KS':'20','KY':'21','LA':'22','ME':'23', \
                  'MD':'24','MA':'25','MI':'26','MN':'27','MS':'28', \
                  'MO':'29','MT':'30','NE':'31','NV':'32','NH':'33', \
                  'NJ':'34','NM':'35','NY':'36','NC':'37','ND':'38', \
                  'OH':'39','OK':'40','OR':'41','PA':'42','RI':'44', \
                  'SC':'45','SD':'46','TN':'47','TX':'48','UT':'49', \
                  'VT':'50','VA':'51','WA':'53','WV':'54','WI':'55', \
                  'WY':'56'}

#return state FIPS code from abbreviation
statefips = str(fipsdictionary.get(statestring[:2]))


### EXECUTE MERGE ANALYSIS BY STATE

if statefips == "None": #if all states are being processed

    #create state files
    for statestring in fipsdictionary.keys():
        statefips = str(fipsdictionary.get(statestring[:2]))
        Merge_Function(workspace, statestring, statefips, mergeyear, bufferdistance)

    #change workspace from county directory to state directory
    arcpy.env.workspace = workspace + "\\TIGER" + mergeyear + "\\TIGER" + mergeyear \
                          + "_statebuffers_" + bufferdistance + "m"

    #create selection set of states
    fcs = arcpy.ListFeatureClasses("*" + mergeyear[-2:] + "_buffer" + bufferdistance + "m.shp")

    #merge dissolved states into a national file
    arcpy.Merge_management(fcs, workspace + "\\TIGER" + mergeyear + \
                           "\\" + mergeyear + "_buffer" + bufferdistance + "m.shp")

    #clean up
    del fcs

else: #if one state is being processed
    Merge_Function(workspace, statestring, statefips, mergeyear, bufferdistance)
```

Figure A.9 — Continued

```
#Christopher Brown
#TIGER State Clipping
#United States Master Script File

#import geoprocessing/filesystem modules and enable overwriting
import arcpy, os
arcpy.env.overwriteOutput = True


### CLIP OPERATION

def Clip_Function(workspace, statestring, statefips, sourceyear, clipyear, bufferdistance):

    #specify the input workspace
    roadsdatafolder = workspace + "\\TIGER" + sourceyear + "\\TIGER" + sourceyear + "_states"
    if not os.path.exists(roadsdatafolder): #check to see if workspace exists
        arcpy.AddMessage("\nInvalid " + sourceyear + " source input workspace!")
        print "\nInvalid " + sourceyear + " source input workspace!"

    #specify the clip data folder
    clipdatafolder = workspace + "\\TIGER" + clipyear + "\\TIGER" + clipyear + \
                     "_statebuffers_" + bufferdistance + "m"
    if not os.path.exists(clipdatafolder): #check to see if workspace exists
        arcpy.AddMessage("\nInvalid " + sourceyear + " clip input workspace!")
        print "\nInvalid " + sourceyear + " clip input workspace!"

    #change the output workspace
    outputfolder = workspace + "\\TIGER" + sourceyear + "\\TIGER" + sourceyear + "_stateclips"
    if not os.path.exists(outputfolder): #check to see if workspace exists
        os.mkdir(outputfolder) #if not, create it

    #print input/output directory names
    arcpy.AddMessage("\nThe current roads input workspace is: " + "\n\t" + roadsdatafolder)
    arcpy.AddMessage("\nThe current clip input workspace is: " + "\n\t" + clipdatafolder)
    arcpy.AddMessage("\nThe clipped roads for " + statestring[5:] + " will be put in: " + "\n\t" + outpu
    print "\nThe current roads input workspace is: " + "\n\t" + roadsdatafolder #print current workspace
    print "\nThe current clip input workspace is: " + "\n\t" + clipdatafolder
    print "\nThe clipped roads for " + statestring[5:] + " will be put in: " + "\n\t" + outputfolder #pr

    #create selection set for input roads (one file)
    arcpy.env.workspace = roadsdatafolder
    roads = arcpy.ListFeatureClasses(statestring[:2] + "_" + sourceyear[-2:] + ".shp")
    road1 = roadsdatafolder + "\\" + roads[0]

    #create selection set for clip buffers (one file)
    arcpy.env.workspace = clipdatafolder
    buffers = arcpy.ListFeatureClasses(statestring[:2] + "_" + clipyear[-2:] + "_buffer" + bufferdistanc
    buffer1 = clipdatafolder + "\\" + buffers[0]

    #clip selected state source data using clip year buffers
    clipname = statestring[:2] + "_" + sourceyear[-2:] + "_clip" + clipyear[-2:] + "_" + bufferdistance
    clipoutput = outputfolder + "\\" + clipname
    arcpy.Clip_analysis(road1, buffer1, clipoutput)


    #clean up
    del roads
    del road1
    del buffers
    del buffer1
    arcpy.AddMessage("\n" + statestring[5:] + " completed successfully...\n")
    print "\n" + statestring[5:] + " completed successfully...\n"
```

Figure A.10: Script for Clipping State Road Shapefiles with State Buffer Shapefiles

```
### PARAMETERS :: WORKSPACE, STATE, YEAR, BUFFER DISTANCE

workspace = arcpy.GetParameterAsText(0)
statestring = arcpy.GetParameterAsText(1)
sourceyear = arcpy.GetParameterAsText(2)
clipyear = arcpy.GetParameterAsText(3)
bufferdistance = arcpy.GetParameterAsText(4)

### MANUAL PARAMETER ADJUSTMENT

#workspace = os.getcwd()
#statestring = "DE - Delaware" #use any other abbreviation to process all states
#sourceyear = "2008"
#clipyear = "2012"
#bufferdistance = "25"


### DICTIONARY TO CONVERT STATE ABBREVIATION TO FIPS CODE

fipsdictionary = {'AL':'01','AK':'02','AZ':'04','AR':'05','CA':'06', \
                  'CO':'08','CT':'09','DE':'10','DC':'11','FL':'12', \
                  'GA':'13','HI':'15','ID':'16','IL':'17','IN':'18', \
                  'IA':'19','KS':'20','KY':'21','LA':'22','ME':'23', \
                  'MD':'24','MA':'25','MI':'26','MN':'27','MS':'28', \
                  'MO':'29','MT':'30','NE':'31','NV':'32','NH':'33', \
                  'NJ':'34','NM':'35','NY':'36','NC':'37','ND':'38', \
                  'OH':'39','OK':'40','OR':'41','PA':'42','RI':'44', \
                  'SC':'45','SD':'46','TN':'47','TX':'48','UT':'49', \
                  'VT':'50','VA':'51','WA':'53','WV':'54','WI':'55', \
                  'WY':'56'}

#return state FIPS code from abbreviation
statefips = str(fipsdictionary.get(statestring[:2]))


### EXECUTE CLIP ANALYSIS BY STATE

if statefips == "None": #if all states are being processed

    #create state files
    for statestring in fipsdictionary.keys():
        statefips = str(fipsdictionary.get(statestring[:2]))
        Clip_Function(workspace, statestring, statefips, sourceyear, clipyear, bufferdistance)

    #change workspace to source directory
    arcpy.env.workspace = workspace + "\\TIGER" + sourceyear + "\\TIGER" + sourceyear + "_stateclips"

    #create selection set of states
    fcs = arcpy.ListFeatureClasses("*" + sourceyear[-2:] + "_clip" + clipyear[-2:] + "_" + bufferdistanc

    #merge dissolved states into a national file
    arcpy.Merge_management(fcs, workspace + "\\TIGER" + sourceyear + \
                          "\\" + sourceyear + "_clip" + clipyear [-2:] + "_" + bufferdistance + "m.sl

    #clean up
    del fcs

else: #if one state is being processed
    Clip_Function(workspace, statestring, statefips, sourceyear, clipyear, bufferdistance)
```

Figure A.10 — Continued

join_counties.py - F:\Documents\Georgia\Thesis\join_counties.py

File Edit Format Run Options Windows Help

```
#Christopher Brown
#Joining Length Statistics to County Boundaries
#United States Master Script File

#import geoprocessing/filesystem modules and enable overwriting
import arcpy, os
arcpy.env.overwriteOutput = True


### JOIN OPERATION

def Join_Function(workspace, statestring, statefips):

    #specify the input workspace
    inputdatafolder = workspace + "\\Counties\\CountyBoundaries"
    if not os.path.exists(inputdatafolder): #check to see if workspace exists
        arcpy.AddMessage("\nInvalid input workspace!")
        print "\nInvalid input workspace!"

    #iterate across all data years
    joinyears = ['2008', '2009', '2010', '2011', '2012']
    for joinyear in joinyears:

        #specify the join data folder
        joindatafolder = workspace + "\\Counties\\Counties" + joinyear
        if not os.path.exists(joindatafolder): #check to see if workspace exists
            arcpy.AddMessage("\nInvalid " + joinyear + " join input workspace!")
            print "\nInvalid " + sourceyear + " join input workspace!"

        #print input/output directory names
        arcpy.AddMessage("\nThe current county input workspace is: " + "\n\t" + inputdatafolder)
        arcpy.AddMessage("\nThe current join input workspace is: " + "\n\t" + joindatafolder)
        print "\nThe current county input workspace is: " + "\n\t" + inputdatafolder #print current
        print "\nThe current join input workspace is: " + "\n\t" + joindatafolder

        #create selection set for input county boundaries (one file)
        arcpy.env.workspace = inputdatafolder
        counties = arcpy.ListFeatureClasses("tl_*_" + statefips + "_county*.shp")
        county = inputdatafolder + "\\" + counties[0]

        #create selection set for join tables (one file)
        arcpy.env.workspace = joindatafolder
        tables = arcpy.ListTables(statestring[:2] + "_" + joinyear[-2:] + "*m_Statistics.dbf")
        table = joindatafolder + "\\" + tables[0]
        print table

        #join selected join table to county boundary
        arcpy.JoinField_management(county, "GEOID10", table, "GEOID10", "SUM_Length")


        #clean up
        del counties
        del county
        del tables
        del table

    #state completed
    arcpy.AddMessage("\n" + statestring[5:] + " completed successfully...\n")
    print "\n" + statestring[5:] + " completed successfully...\n"
```

Ln: 1 Col: 0

Figure A.11: Script for Joining County Length Statistics Table to County Boundaries

```
### PARAMETERS :: WORKSPACE, STATE, YEAR, BUFFER DISTANCE

workspace = arcpy.GetParameterAsText(0)
statestring = arcpy.GetParameterAsText(1)

### MANUAL PARAMETER ADJUSTMENT

#workspace = os.getcwd()
#statestring = "US - United States" #use any other abbreviation to process all states


### DICTIONARY TO CONVERT STATE ABBREVIATION TO FIPS CODE

fipsdictionary = {'AL':'01','AK':'02','AZ':'04','AR':'05','CA':'06', \
                  'CO':'08','CT':'09','DE':'10','DC':'11','FL':'12', \
                  'GA':'13','HI':'15','ID':'16','IL':'17','IN':'18', \
                  'IA':'19','KS':'20','KY':'21','LA':'22','ME':'23', \
                  'MD':'24','MA':'25','MI':'26','MN':'27','MS':'28', \
                  'MO':'29','MT':'30','NE':'31','NV':'32','NH':'33', \
                  'NJ':'34','NM':'35','NY':'36','NC':'37','ND':'38', \
                  'OH':'39','OK':'40','OR':'41','PA':'42','RI':'44', \
                  'SC':'45','SD':'46','TN':'47','TX':'48','UT':'49', \
                  'VT':'50','VA':'51','WA':'53','WV':'54','WI':'55', \
                  'WY':'56'}

#return state FIPS code from abbreviation
statefips = str(fipsdictionary.get(statestring[:2]))


### EXECUTE JOIN ANALYSIS BY STATE

if statefips == "None": #if all states are being processed

    #create state files
    for statestring in fipsdictionary.keys():
        statefips = str(fipsdictionary.get(statestring[:2]))
        Join_Function(workspace, statestring, statefips)

    #change workspace to source directory
    arcpy.env.workspace = workspace + "\\Counties\\CountyBoundaries"

    #create selection set of states
    fcs = arcpy.ListFeatureClasses("tl_*_*_county10.shp")

    #merge dissolved states into a national file
    arcpy.Merge_management(fcs, workspace + "\\Counties\\countystatistics.shp")

    #clean up
    del fcs

else: #if one state is being processed
    Join_Function(workspace, statestring, statefips)
```

Figure A.11 — Continued

```
#Christopher Brown
#Joining Length Statistics to Tract Boundaries
#United States Master Script File

#import geoprocessing/filesystem modules and enable overwriting
import arcpy, os
arcpy.env.overwriteOutput = True


### JOIN OPERATION

def Join_Function(workspace, statestring, statefips):

    #specify the input workspace
    inputdatafolder = workspace + "\\Tracts\\TractBoundaries"
    if not os.path.exists(inputdatafolder): #check to see if workspace exists
        arcpy.AddMessage("\nInvalid input workspace!")
        print "\nInvalid input workspace!"

    #iterate across all data years
    joinyears = ['2008', '2009', '2010', '2011', '2012']
    for joinyear in joinyears:

        #specify the join data folder
        joindatafolder = workspace + "\\Tracts\\Tracts" + joinyear
        if not os.path.exists(joindatafolder): #check to see if workspace exists
            arcpy.AddMessage("\nInvalid " + joinyear + " join input workspace!")
            print "\nInvalid " + sourceyear + " join input workspace!"

        #print input/output directory names
        arcpy.AddMessage("\nThe current tract input workspace is: " + "\n\t" + inputdatafolder)
        arcpy.AddMessage("\nThe current join input workspace is: " + "\n\t" + joindatafolder)
        print "\nThe current tract input workspace is: " + "\n\t" + inputdatafolder #print current workspace
        print "\nThe current join input workspace is: " + "\n\t" + joindatafolder

        #create selection set for input tract boundaries (one file)
        arcpy.env.workspace = inputdatafolder
        tracts = arcpy.ListFeatureClasses("tl_*_" + statefips + "_tract*.shp")
        tract = inputdatafolder + "\\" + tracts[0]

        #create selection set for join tables (one file)
        arcpy.env.workspace = joindatafolder
        tables = arcpy.ListTables(statestring[:2] + "_" + joinyear[-2:] + "*m_Statistics.dbf")
        table = joindatafolder + "\\" + tables[0]
        print table

        #join selected join table to tract boundary
        arcpy.JoinField_management(tract, "GEOID10", table, "GEOID10", "SUM_Length")


        #clean up
        del tracts
        del tract
        del tables
        del table

    #state completed
    arcpy.AddMessage("\n" + statestring[5:] + " completed successfully...\n")
    print "\n" + statestring[5:] + " completed successfully...\n"
```

Figure A.12: Script for Joining Tract Length Statistics Table to Tract Boundaries

```
### PARAMETERS :: WORKSPACE, STATE, YEAR, BUFFER DISTANCE

workspace = arcpy.GetParameterAsText(0)
statestring = arcpy.GetParameterAsText(1)

### MANUAL PARAMETER ADJUSTMENT

#workspace = os.getcwd()
#statestring = "US - United States" #use any other abbreviation to process all states


### DICTIONARY TO CONVERT STATE ABBREVIATION TO FIPS CODE

fipsdictionary = {'AL':'01','AK':'02','AZ':'04','AR':'05','CA':'06', \
                  'CO':'08','CT':'09','DE':'10','DC':'11','FL':'12', \
                  'GA':'13','HI':'15','ID':'16','IL':'17','IN':'18', \
                  'IA':'19','KS':'20','KY':'21','LA':'22','ME':'23', \
                  'MD':'24','MA':'25','MI':'26','MN':'27','MS':'28', \
                  'MO':'29','MT':'30','NE':'31','NV':'32','NH':'33', \
                  'NJ':'34','NM':'35','NY':'36','NC':'37','ND':'38', \
                  'OH':'39','OK':'40','OR':'41','PA':'42','RI':'44', \
                  'SC':'45','SD':'46','TN':'47','TX':'48','UT':'49', \
                  'VT':'50','VA':'51','WA':'53','WV':'54','WI':'55', \
                  'WY':'56'}

#return state FIPS code from abbreviation
statefips = str(fipsdictionary.get(statestring[:2]))


### EXECUTE JOIN ANALYSIS BY STATE

if statefips == "None": #if all states are being processed

    #create state files
    for statestring in fipsdictionary.keys():
        statefips = str(fipsdictionary.get(statestring[:2]))
        Join_Function(workspace, statestring, statefips)

    #change workspace to source directory
    arcpy.env.workspace = workspace + "\\Tracts\\TractBoundaries"

    #create selection set of states
    fcs = arcpy.ListFeatureClasses("tl_*_*_tract10.shp")

    #merge dissolved states into a national file
    arcpy.Merge_management(fcs, workspace + "\\Tracts\\tractstatistics.shp")

    #clean up
    del fcs

else: #if one state is being processed
    Join_Function(workspace, statestring, statefips)
```

Figure A.12 — Continued

APPENDIX B

CORRELATION MATRICES

Table B.1: Correlation Matrix for Linear Regression Variables within MSAs

	PerRoadChg	*MeanIncome*	*MeanAge*	*PopDensKm2*	*PopChg0711*
PerRoadChg	1				
MeanIncome	0.127718	1			
MeanAge	-0.03101	0.00045	1		
PopDensKm2	0.105163	0.546452	0.110408	1	
PopChg0711	0.301969	0.037074	-0.47568	-0.15428	1
PerWhite	-0.12596	-0.17808	0.30283	-0.26943	-0.20569
PerBlack	0.104215	-0.10576	-0.07911	0.112044	0.066334
PerAmInd	-0.01684	0.004464	-0.1355	-0.1583	0.061222
PerAsian	0.034107	0.520428	-0.15965	0.459936	0.069264
PerPacIsla	0.022636	0.169224	-0.05072	0.210554	0.087248
PerOther	0.044375	0.218959	-0.31803	0.094505	0.227397
PerHispani	0.103369	0.105763	-0.32509	0.097811	0.297783
Pacific	-0.07744	0.265858	-0.12346	0.059354	0.056597
Mountain	0.126498	0.023738	-0.09486	-0.18151	0.109417
WestNorthC	-0.14266	0.01027	-0.12103	-0.13131	-0.0157
WestSouthC	0.014633	-0.12995	-0.22376	-0.09301	0.265061
EastNorthC	-0.22904	-0.10384	0.082684	0.071293	-0.38397
EastSouthC	-0.03621	-0.1731	5.94E-05	-0.08941	0.084481
SAtlantic	0.328882	-0.11481	0.13708	0.017589	0.185271
MdAtlantic	-0.07238	0.086643	0.188608	0.181192	-0.19698
NewEngland	0.023664	0.265004	0.17691	0.203663	-0.1633

Table B.1 — Continued

	PerWhite	*PerBlack*	*PerAmInd*	*PerAsian*	*PerPacIsla*	*PerOther*
PerWhite	1					
PerBlack	-0.73799	1				
PerAmInd	-0.10583	-0.19222	1			
PerAsian	-0.38351	-0.0817	0.015328	1		
PerPacIsla	-0.22515	-0.07629	0.054707	0.752839	1	
PerOther	-0.37158	-0.22558	0.13477	0.252669	0.054474	1
PerHispani	-0.2462	-0.24452	0.088956	0.158489	0.037396	0.84946
Pacific	-0.19941	-0.26064	0.16404	0.470792	0.272848	0.556691
Mountain	0.129558	-0.27959	0.315316	-0.0576	0.038148	0.15576
WestNorthC	0.224382	-0.15892	0.029457	-0.04098	-0.03518	-0.15367
WestSouthC	-0.14647	0.121146	0.039857	-0.08943	-0.0329	0.118165
EastNorthC	0.214884	-0.05932	-0.12571	-0.08032	-0.07705	-0.22445
EastSouthC	-0.08065	0.222846	-0.08564	-0.11095	-0.03875	-0.16256
SAtlantic	-0.26465	0.433747	-0.14654	-0.0922	-0.05537	-0.15226
MdAtlantic	0.130807	-0.0726	-0.10255	-0.0034	-0.0492	-0.09355
NewEngland	0.132712	-0.10664	-0.05205	0.000117	-0.03005	-0.05206

	PerHispani	*Pacific*	*Mountain*	*WestNorthC*	*WestSouthC*
PerHispani	1				
Pacific	0.366999	1			
Mountain	0.154146	-0.12403	1		
WestNorthC	-0.14433	-0.11188	-0.09714	1	
WestSouthC	0.293929	-0.13736	-0.11926	-0.10757	1
EastNorthC	-0.21697	-0.16903	-0.14676	-0.13237	-0.16253
EastSouthC	-0.16617	-0.11396	-0.09895	-0.08925	-0.10958
SAtlantic	-0.13845	-0.19226	-0.16693	-0.15056	-0.18486
MdAtlantic	-0.09507	-0.11396	-0.09895	-0.08925	-0.10958
NewEngland	-0.06558	-0.07881	-0.06843	-0.06172	-0.07578

	EastNorthC	*EastSouthC*	*SAtlantic*	*MdAtlantic*	*NewEngland*
EastNorthC	1				
EastSouthC	-0.13484	1			
SAtlantic	-0.22748	-0.15337	1		
MdAtlantic	-0.13484	-0.09091	-0.15337	1	
NewEngland	-0.09325	-0.06287	-0.10606	-0.06287	1

Table B.2: Correlation Matrix for Linear Regression Variables within MSA Counties

	PerRoadChg	*MeanIncome*	*PerWhite*	*PerBlack*	*PerAmIndia*	*PerAsian*
PerRoadChg	1					
MeanIncome	0.174124	1				
PerWhite	-0.09449	-0.00874	1			
PerBlack	0.061943	-0.15258	-0.8856	1		
PerAmIndia	-0.03922	-0.0667	-0.06548	-0.10945	1	
PerAsian	0.101456	0.515732	-0.3445	0.015677	0.002964	1
PerPacIsla	0.01819	0.087369	-0.14845	-0.0264	0.065575	0.569104
PerOther	0.069676	0.139974	-0.36244	-0.03423	0.116568	0.34576
PerHispani	0.098548	0.116154	-0.26685	-0.07787	0.091083	0.271724
MeanAge	-0.11502	-0.02254	0.361128	-0.20065	-0.09403	-0.26492
PopChg0711	0.294553	0.180099	-0.06294	-0.04636	0.080018	0.139031
PopDensKm2	-0.05079	0.198848	-0.25237	0.141588	-0.03025	0.309099
Pacific	-0.04132	0.142934	-0.11744	-0.14333	0.140059	0.428894
Mountain	0.060673	0.017219	0.072614	-0.17021	0.21176	-0.02192
WestNorthC	-0.11753	-0.02268	0.225145	-0.18434	0.016656	-0.07036
WestSouthC	0.011995	-0.08547	-0.1057	0.057475	0.196934	-0.07842
EastNorthC	-0.17529	-0.02443	0.22709	-0.15498	-0.11633	-0.08389
EastSouthC	-0.01299	-0.18672	0.001358	0.090099	-0.08689	-0.13957
SAtlantic	0.27785	-0.02906	-0.28808	0.380524	-0.12613	-0.03052
MdAtlantic	-0.08452	0.180331	0.024599	-0.04567	-0.0875	0.114545
NewEngland	-0.0181	0.158723	0.082589	-0.08371	-0.03314	0.030948

	PerPacIsla	*PerOther*	*PerHispani*	*MeanAge*	*PopChg0711*	*PopDensKm2*
PerPacIsla	1					
PerOther	0.111513	1				
PerHispani	0.090881	0.881103	1			
MeanAge	-0.10732	-0.37635	-0.3414	1		
PopChg0711	0.068505	0.221807	0.231102	-0.38118	1	
PopDensKm2	0.028289	0.196019	0.148407	-0.10864	-0.04571	1
Pacific	0.298145	0.460707	0.333903	-0.1223	0.066217	0.002866
Mountain	0.059119	0.183827	0.180163	-0.08188	0.074667	-0.03516
WestNorthC	-0.02455	-0.14755	-0.13983	-0.0673	-0.01638	-0.04656
WestSouthC	-0.02226	0.149775	0.264733	-0.1172	0.154502	-0.05335
EastNorthC	-0.07159	-0.17435	-0.16809	0.067399	-0.18051	-0.03565
EastSouthC	-0.04047	-0.15346	-0.16328	0.003016	0.00662	-0.05404
SAtlantic	-0.04323	-0.10438	-0.11404	0.074094	0.081464	0.004023
MdAtlantic	-0.03701	0.007506	0.002267	0.123872	-0.1266	0.233701
NewEngland	-0.01795	-0.02228	-0.0397	0.098271	-0.10322	0.023753

Table B.2 — Continued

	Pacific	*Mountain*	*WestNorthC*	*WestSouthC*	*EastNorthC*	*EastSouthC*
Pacific	1					
Mountain	-0.06376	1				
WestNorthC	-0.0901	-0.08585	1			
WestSouthC	-0.1	-0.09528	-0.13463	1		
EastNorthC	-0.11214	-0.10684	-0.15098	-0.16756	1	
EastSouthC	-0.08968	-0.08544	-0.12073	-0.134	-0.15026	1
SAtlantic	-0.15516	-0.14783	-0.20889	-0.23184	-0.25998	-0.2079
MdAtlantic	-0.07676	-0.07313	-0.10334	-0.11469	-0.12861	-0.10285
NewEngland	-0.0462	-0.04402	-0.0622	-0.06904	-0.07742	-0.06191

	SAtlantic	*MdAtlantic*	*NewEngland*
SAtlantic	1		
MdAtlantic	-0.17795	1	
NewEngland	-0.10711	-0.05299	1

Table B.3: Correlation Matrix for Linear Regression Variables within MSA Census Tracts

	PerRoadChg	*MeanIncome*	*PerWhite*	*PerAmIndia*	*PerAsian*	*PerPacIsla*
PerRoadChg	1					
MeanIncome	0.078334	1				
PerWhite	0.048556	0.337289	1			
PerAmIndia	-0.01692	-0.12964	-0.0968	1		
PerAsian	0.029809	0.218115	-0.24852	-0.02808	1	
PerPacIsla	-0.00025	-0.00377	-0.11333	0.046048	0.336829	1
PerOther	-0.04412	-0.25953	-0.37604	0.11337	0.071129	0.046162
PerHispani	-0.00036	-0.2389	-0.2682	0.087533	0.037518	0.034042
PerBlack	-0.04491	-0.33811	-0.83417	-0.03911	-0.14762	-0.0382
MeanAge	-0.03163	0.356576	0.409073	-0.12264	-0.04912	-0.05278
PopDensKm2	-0.08461	-0.06413	-0.29569	-0.00521	0.223134	0.034989
Pacific	-0.02523	0.091109	-0.10985	0.0927	0.390407	0.233729
Mountain	0.043559	-0.01335	0.085604	0.168525	-0.05067	0.032254
WestNorthC	-0.01017	-0.0215	0.10169	0.01331	-0.06032	-0.02895
WestSouthC	0.057883	-0.06059	-0.02453	0.082106	-0.0782	-0.03726
EastNorthC	-0.0537	-0.07889	0.040463	-0.07564	-0.11808	-0.06719
EastSouthC	0.004433	-0.07316	-0.00747	-0.04403	-0.09651	-0.03052
SAtlantic	0.076037	-0.00565	-0.04938	-0.08527	-0.10455	-0.05497
MdAtlantic	-0.06678	0.066046	-0.02734	-0.07518	0.04329	-0.05281
NewEngland	-0.02049	0.073891	0.091198	-0.04069	-0.0278	-0.02727

Table B.3 — Continued

	PerOther	*PerHispani*	*PerBlack*	*MeanAge*	*PopDensKm2*	*Pacific*
PerOther	1					
PerHispani	0.888593	1				
PerBlack	-0.04027	-0.09727	1			
MeanAge	-0.44227	-0.40915	-0.24416	1		
PopDensKm2	0.297067	0.252099	0.117685	-0.18355	1	
Pacific	0.350976	0.290225	-0.17922	-0.06155	0.049802	1
Mountain	0.090222	0.094936	-0.13067	-0.03896	-0.05237	-0.12376
WestNorthC	-0.10501	-0.12145	-0.04454	-0.0295	-0.06584	-0.10938
WestSouthC	0.072286	0.18854	0.016432	-0.12379	-0.08429	-0.15606
EastNorthC	-0.14423	-0.17173	0.068439	0.00802	-0.05936	-0.19525
EastSouthC	-0.10984	-0.13108	0.090937	-0.00816	-0.07958	-0.10029
SAtlantic	-0.15509	-0.1025	0.167147	0.100362	-0.10888	-0.218
MdAtlantic	-0.02368	-0.05216	0.029585	0.071892	0.333125	-0.19174
NewEngland	-0.03366	-0.06422	-0.07019	0.054177	-0.01338	-0.10094

	Mountain	*WestNorthC*	*WestSouthC*	*EastNorthC*	*EastSouthC*	*SAtlantic*
Mountain	1					
WestNorthC	-0.06693	1				
WestSouthC	-0.0955	-0.0844	1			
EastNorthC	-0.11948	-0.1056	-0.15066	1		
EastSouthC	-0.06137	-0.05424	-0.07739	-0.09683	1	
SAtlantic	-0.1334	-0.11791	-0.16822	-0.21047	-0.10811	1
MdAtlantic	-0.11733	-0.1037	-0.14795	-0.18511	-0.09509	-0.20668
NewEngland	-0.06177	-0.0546	-0.07789	-0.09746	-0.05006	-0.10881

	MdAtlantic	*NewEngland*
MdAtlantic	1	
NewEngland	-0.0957	1