A ZONE-BASED APPROACH TO IDENTIFYING URBAN LAND USES USING NATIONALLY-AVAILABLE DATA

by

James A. Falcone
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Discipline

Committee:

_______________________________ Dr. David Wong, Dissertation Director

_______________________________ Dr. Barry Kronenfeld, Committee Member

_______________________________ Dr. Laurie Schintler, Committee Member

_______________________________ Dr. Nigel Waters, Committee Member

_______________________________ Dr. Peggy Agouris, Department Chairperson

_______________________________ Dr. Richard Diecchio, Associate Dean for Academic and Student Affairs, College of Science

_______________________________ Dr. Vikas Chandhoke, Dean, College of Science

Date: __________________________ Spring Semester 2010
George Mason University
Fairfax, VA

A Zone-Based Approach To Identifying Urban Land Uses
Using Nationally-Available Data

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

James A. Falcone
MS, Remote Sensing, University of New South Wales, 2002
MS, Geographic and Cartographic Sciences, George Mason University, 1999
BA, Environmental Science, University of Virginia, 1979

Director: Dr. David Wong, Professor
Department of Geography and GeoInformation Science

Spring Semester 2010
George Mason University
Fairfax, VA

Dedication

This is dedicated to my mother and father, who put me on a good path, and were always there and always cared.

Acknowledgements

I am very grateful to Dr. Wong for this patient and thoughtful guidance. It has been a real pleasure being his student.

I am also very thankful to Drs. Kronenfeld, Schintler, and Waters, for their advice and generous help.

I would also like to thank Dr. Daren Carlisle at the USGS for his collegial encouragement and introducing me to R, and Drs. Barry Haack and Richard Gomez at GMU, for such a good influence in guiding my earlier studies.

To my wonderful Liz and Katie: there are not enough words in the world to thank you for your unwavering support, love, advice, and listening over these many years: there should be a separate category of sainthood for the family of PhD students. I love you “as much as the world”!!

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ACS – American Community Survey
AVHRR – Advanced Very High Resolution Radiometer
BG – Block Group
CBD – Central Business District
CFCC – Census Feature Class Codes
DEM – Digital Elevation Model
DV – Dependent Variable
ESRI – Environmental Systems Research Institute
GIRAS - Geographic Information Retrieval and Analysis System
GIS – Geographic Information System
GMU GGS – George Mason University Geography and GeoInformation Science
GNIS – Geographic Names Information System
IV – Independent Variable
LU – Land Use
MAUP – Modifiable Areal Unit Problem
MMU – Minimum Mapping Unit
MSA – Metropolitan Statistical Area
NDVI – Normalized Difference Vegetation Index
NLCD01 – National Land Cover Dataset 2001
NLCD92 – National Land Cover Dataset 1992
NRMSE – Normalized Root Mean Squared Error
NTDB – National Topographic DataBase (Canadian)
OOB – Out-of-bag (cases withheld from bootstrap sample in RF)
PCA – Principal Components Analysis
PUMS – Public Use Microdata Sample
RF – Random Forest classification and regression tree tool
RMSE – Root Mean Squared Error
RS – Remote Sensing
SA – Spatial Autocorrelation
SRTM – Shuttle Radar Topography Mission
USGS – U.S. Geological Survey
VIF – Variance Inflation Factor

Abstract

A ZONE-BASED APPROACH TO IDENTIFYING URBAN LAND USES USING NATIONALLY-AVAILABLE DATA

James A. Falcone, Ph.D.

George Mason University, 2010

Dissertation Director: Dr. David Wong

Accurate identification of urban land use is essential for many applications in environmental study, ecological assessment, and urban planning, among other fields. However, because physical surfaces of land cover types are not necessarily related to their use and economic function, differentiating among thematically-detailed urban land uses (single-family residential, multi-family residential, commercial, industrial, etc.) using remotely-sensed imagery is a challenging task, particularly over large areas. Because the process requires an interpretation of tone/color, size, shape, pattern, and neighborhood association elements within a scene, it has traditionally been accomplished via manual interpretation of aerial photography or high-resolution satellite imagery. Although success has been achieved for localized areas using various automated

techniques based on high-spatial or high-spectral resolution data, few detailed (Anderson Level II equivalent or greater) urban land use mapping products have successfully been created via automated means for broad (multi-county or larger) areas, and no such product exists today for the United States.

In this study I argue that by employing a zone-based approach it is feasible to map thematically-detailed urban land use classes over large areas using appropriate combinations of non-image based predictor data which are nationally and publicly available. The approach presented here uses U.S. Census block groups as the basic unit of geography, and predicts the percent of each of ten land use types - nine of them urban - for each block group based on a number of data sources, to include census data, nationally-available point locations of features from the USGS Geographic Names Information System, historical land cover, and metrics which characterize spatial pattern, context (e.g. distance to city centers or other features), and measures of spatial autocorrelation.

The method was demonstrated over a four-county area surrounding the city of Boston. A generalized version of the method (six land use classes) was also developed and cross-validated among additional geographic settings: Atlanta, Los Angeles, and Providence. The results suggest that even with the thematically-detailed ten-class structure, it is feasible to map most urban land uses with reasonable accuracy at the block group scale, and results improve with class aggregation. When classified by predicted majority land use, 79% of block groups correctly matched the actual majority land use with the ten-class models. Six-class models typically performed well for the geographic

area they were developed from, however models had mixed performance when transported to other geographic settings. Contextual variables, which characterized a block group’s spatial relationship to city centers, transportation routes, and other amenities, were consistently strong predictors of most land uses, a result which corresponds to classic urban land use theory. The method and metrics derived here provide a prototype for mapping urban land uses from readily-available data over broader geographic areas than is generally practiced today using current image-based solutions.

1. Introduction

1.1 Rationale for research/background

Mapping urban land is essential to many applications: ecological assessments study the effect of urbanization on biota or stream water quality; urban planning requires understanding of the urban environment at multiple scales; and economic, sociologic, and political decisions are often made to some degree based on knowledge of where people are and how they use the land. The need for such mapping is likely to increase as urban areas grow, and a larger proportion of the world's population inhabits urban environments. Effective solutions to monitoring characteristics of urban landscapes are desirable; cost-effective solutions even more so.

Although numerous variations of land use/land cover classifications exist, most are based on the seminal work of Anderson et al. (1976). The highest level of categorization (Level I) distinguishes among broad land cover types: urban, agricultural, forest, water, wetlands, etc. For urban land, the second level of categorization (Level II) distinguishes among thematically detailed land uses: residential, commercial, industrial, and so on (Table 1-1). While land cover may often be adequately derived from the spectral information in remotely sensed imagery alone, derivation of land use (how humans use the land) requires higher order information that describes the size, shape, pattern, and association among elements in a scene (Estes et al., 1983; Haack et al., 1997;

Graham and Koh, 2002; Aplin, 2003). Land use delineation is primarily an interpretation of the economic function of the land (Campbell, 1996), to which there may be clues in the pattern, shape, extent, and intensity of urban surfaces. However, there is a considerable leap from land cover to land use identification (Lackner and Conway, 2008). Estes et al. (1983) describe a hierarchy of image elements that are fundamental to image interpretation (Figure 1-1), which at the most basic level represent tone/color and at the highest level the site and association of objects.

Table 1-1: Classification scheme of urban land use/land cover, from Anderson (1976)

Level I	**Level II**	**Level III (example for Residential)**
1 Urban or Built-up Land	11 Residential	111 Single-family units
		112 Multi-family units
		113 Group quarters
		114 Residential hotels
		115 Mobile home parks
		116 Transient lodging
		117 Other
	12 Commercial and Services	
	13 Industrial	
	14 Transportation, Communications, and Utilities	
	15 Industrial and Commercial Complexes	
	16 Mixed Urban or Built-up Land	
	17 Other Urban or Built-up Land	

Assimilating the information in these elements, to include the highest levels of interpretation, is often straightforward for a human image interpreter employing high-resolution imagery ("large multi-branched building adjacent to parking lots with school buses, a running track and ball fields: feature is probably a school"). However achieving the same result from digital data, particularly for image processing techniques based on per-pixel extraction of information, is highly challenging (Johnsson, 1994; Jensen et al., 2001; Herold, 2003; Lackner and Conway, 2008).

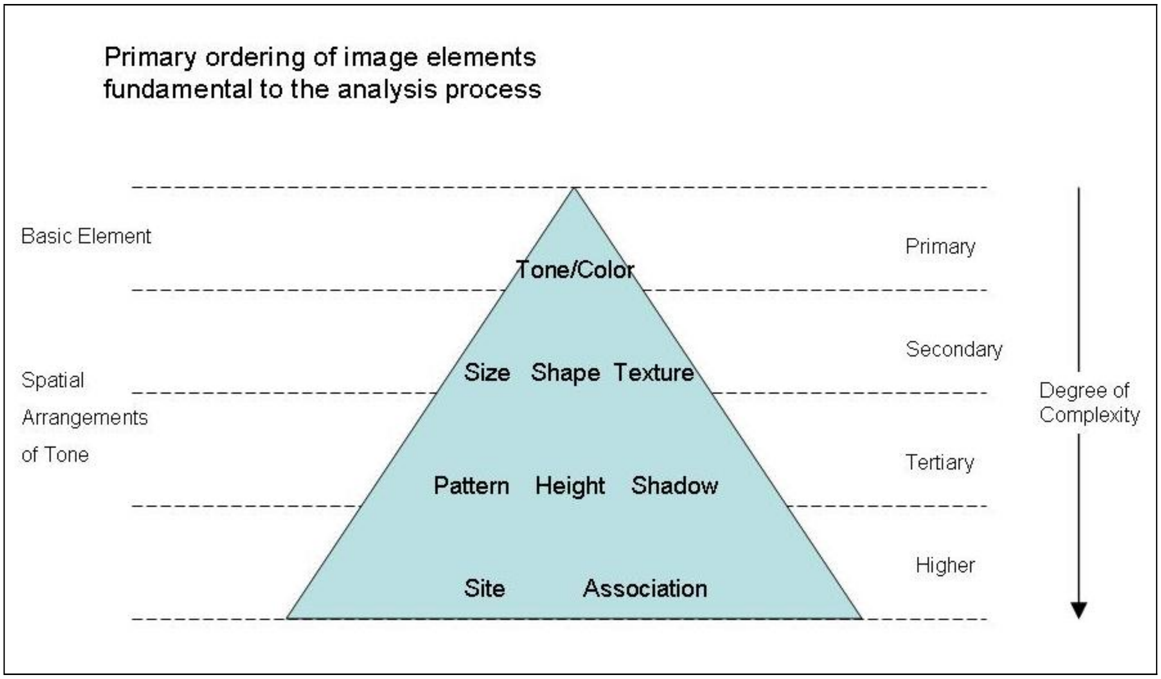


Figure 1-1: The primary ordering or image analysis elements in imagery interpretation (from Estes et al., 1983)

The literature and popular press are often imprecise regarding the terms “land use” and “land cover”. In this paper we will adhere to the differentiation of *land cover* as describing physical surfaces (in the context of urban surfaces as meaning either their physical composition – e.g. asphalt, concrete, metal, etc. – or a specific physical function – e.g. rooftops, roads, parking lots, etc.) vs. *land use* as the broader use/economic function similar to that described by Anderson Level II classes: single-family residential, commercial, industrial, recreation, etc.

It is further recognized that the difficulty in correctly identifying urban land uses increases as the spatial resolution of the image data becomes coarser (Aplin, 2003). Welch (1982) posited that although Anderson Level I urban classification may require only 30 to 80-meter (m) resolution data, Anderson Level II requirements increase dramatically, calling for 0.5 to 10-m spatial resolution, depending on the landscape type. This is echoed by Forster (1985) and Jensen and Cowan (1999), who suggest 5 to 20-m resolution as the minimum resolution feasible for multispectral data in order to adequately classify Anderson Level II categories. Because of the tradeoff between high spatial resolution and swath width/areal coverage, image data of those resolutions is not readily available at anything approaching reasonable cost for broad areas (i.e. multi-county or larger, in the United States). For example, as of this writing, purchase of mosaiced 4-m Ikonos multispectral image data which covers the footprint of a single 30-m Landsat scene would cost more than $800,000 (GeoEye, 2009). Even those studies which have been successful at automated delineation of urban land uses from high-resolution imagery have acknowledged the need to incorporate information outside the spectral domain, either in the form of shape/spatial metrics or ancillary information such as census data (Gong et al., 1992; Mesev, 1998; Chen, 2002; Herold, 2002; Barnsley et al., 2003; Herold et al., 2003; Rocha, 2006; Taubenenboeck, 2006). In short, a traditional multi-spectral image classification solution to mapping Anderson Level II-like land uses at regional or national scales is not currently feasible, and is not likely to be in the foreseeable future using existing methodologies. At present there is no product which maps urban land use across the United States.

The history of urban land use mapping in the United States reflects these realities. Before digital media began to be widely used and available (roughly the 1970s), land use maps were created from manual interpretation of aerial photographs. This method, employed by a skilled interpreter, was (and still is) considered to be the most accurate method for delineating land uses, because (a) it incorporates the knowledge base of a human expert who could interpret the full range of site and association elements of a scene, and (b) it is based on large-scale high-resolution imagery. During the mid-1970s, the U.S. Geological Survey (USGS) began to produce the first nationally consistent maps of land use and land cover for the conterminous United States and Hawaii (Price et al., 2006). Polygons of land use/land cover were delineated manually using aerial photography and mapped following the Anderson classification system, containing the seven urban land use classes noted in Table 1-1. The minimum mapping unit (MMU) was 4 hectares for all urban and water classes (approximately equivalent to 44 30-m pixels in a Landsat scene) and 16 hectares for most other classes. The USGS published these land use and land cover maps at 1:250,000 and 1:100,000 scales for the conterminous United States, Hawaii, and part of Alaska.

The USGS also developed the Geographic Information Retrieval and Analysis System (GIRAS) software (Mitchell et al., 1977) to digitize, edit, and produce cartographic and statistical output from the mapped information (Price et al., 2006). (Although technically the acronym "GIRAS" refers to the software system used to create the digital format of the land use/land cover maps – themselves referred to as the "LULC" – the term GIRAS is often used to refer to the land use/land cover maps and subsequent digital product itself, and will be the term used in this paper). The data are currently distributed by the USGS in the original GIRAS file format (USGS, 2009a), and in a modified format to reflect areas of population density change (USGS, 2009b). Visual examples of the GIRAS and subsequent national land use/land cover products are given in Figures 1-2 through 1-4 with current locations of some landmarks labeled. All three products are publicly-available national datasets.

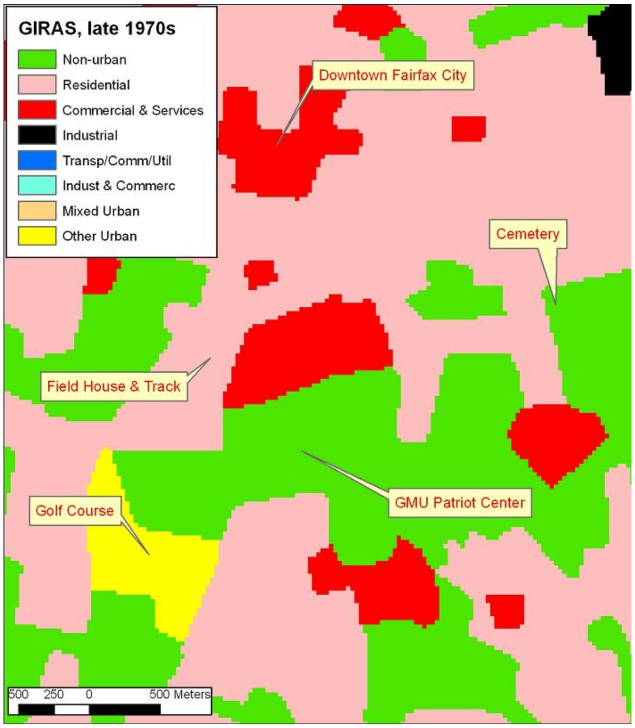


Figure 1-2: Sample of GIRAS for the area of Fairfax City, VA.
The product was derived by manually delineating land use polygons from aerial photography, and contained 7 land use classes (only 4 present in scene shown).

The manual interpretation of imagery that constituted the GIRAS had two drawbacks. First, manual interpretation is recognized to be both a “science and an art” (Graham and Koh, 2002), which draws to some degree on the subjective skill and interpretation of the analyst, and is not necessarily reproducible. Second, it is a highly labor intensive activity, which could require years of work by teams of skilled individuals to map large areas, as was the case with the GIRAS.

The 1970s may also be considered the advent of the era of readily-available satellite imagery, with the launch in 1972 of a series of Landsat satellites, with an initial instrument which imaged five spectral bands at nominal 80-m spatial resolution, then in 1982 with the Thematic Mapper (TM) instrument, with seven bands, six of which had 30-m resolution. The Landsat series of satellites have been imaging the Earth with a repeat cycle of 16 days continually since 1972, and have been by far the most common source of earth information for deriving land use/land cover information at broad scales over the past decades. A 30-m Landsat scene footprint measures approximately 180km x 180km. Although this covers a reasonably broad (multi-county) area, approximately 410 Landsat scenes would be required to image a single season across the conterminous (lower 48) United States.

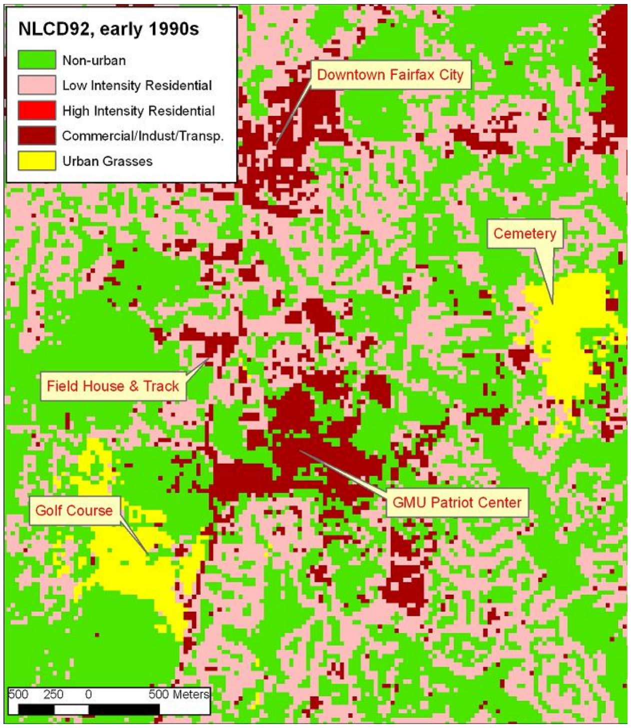


Figure 1-3: Sample of NLCD92 for the area of Fairfax City, VA.
The NLCD92 was a 30-m pixel-based product that had four land use classes.

Because Landsat data had the coverage from which it was possible to map the entire US and reasonably detailed spatial resolution, and recognizing the difficult and labor-intensive nature of manually repeating the GIRAS, the USGS and partner agencies created the National Land Cover Dataset 1992 (NLCD92) dataset, which was a 30-m mapping of 21 land use/land cover classes representing the early 1990s era (Vogelmann et al., 2001). The NLCD92 contained four urban land-use classes (see Figure 1-3), fewer than the GIRAS, but was produced in a more automated fashion at higher spatial resolution. The methodology for producing the NLCD92 was based on an unsupervised classification with subsequent labeling/identification of clusters, and incorporated census data as the basis for distinguishing residential from non-residential urban land. Clusters were in part manually interpreted from aerial photography, particularly for urban areas (James Vogelmann, personal communication, June 19, 2008). The NLCD92 was at least in part a “hybrid” product, produced by both automated and manual means.

Even though considerable effort was put in to creating the NLCD92, and it was widely recognized as a high-quality and valuable product, the difficulty in accurately identifying urban land use on a pixel basis was evident, and the reported accuracy of the individual urban classes generally was less than 50% (Stehman et al., 2003). There was in particular difficulty in separating the high-intensity residential class from the low-intensity residential, and commercial/industrial/transportation class.

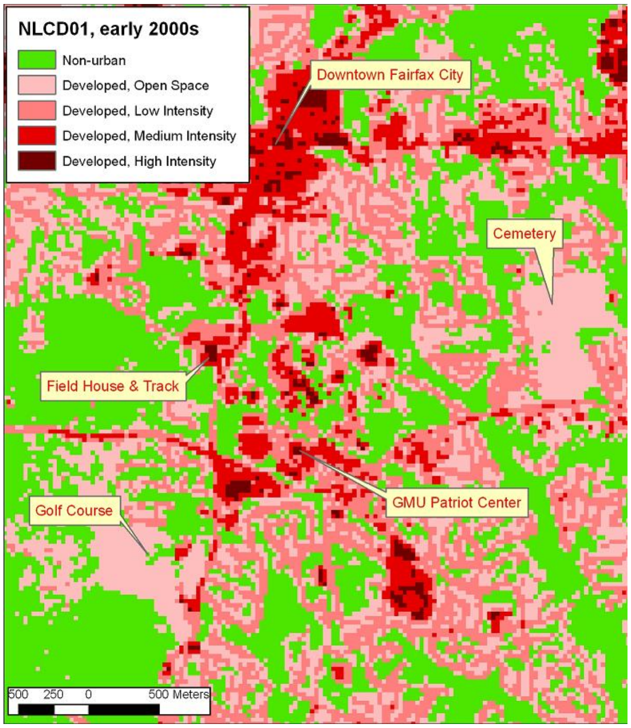


Figure 1-4: Sample of NLCD01 for the area of Fairfax City, VA.
The NLCD01 does not attempt to portray land use, but represents four urban classes as classifications of degrees of imperviousness + urban open space.

In 2006 the USGS completed and released the National Land Cover Database 2001 (NLCD01; Homer et al., 2007). The methodology for the NLCD01, although still based primarily on Landsat data and mapped at 30-m resolution, was a departure from that used for the NLCD92, and was based on a classification and regression tree approach. The NLCD01 contains two products which map urban land, although one is essentially a variation of the other: the first product is a fraction image which maps the percent impervious surfaces (scaled 0-100) for each 30-m pixel, and the second is a categorical version of the fraction image, which is primarily a recoding of the fraction image into four classes (e.g. 80-100% imperviousness was recoded as “Developed, High Intensity”, 50-80% recoded as “Developed, Medium Intensity”, etc.). Although the NLCD01 categorical product does include “Developed, Open Space” as part of the lowest imperviousness class, there is otherwise not an attempt to capture urban land use as part of its depiction of the landscape, and the classes do not impute any economic function to the urban landscape, as the NLCD92 and GIRAS products did.

The United States is not alone in lacking a broad area land use product derived from an automated or semi-automated process. Regions or countries with similar land area to the US (Europe, Canada, Australia) may have national land use/land cover products which map detailed urban land uses, however they were created by manual means. For example, the European CORINE Land Cover dataset (European Environment Agency, 2009) maps nine detailed urban classes in vector polygon format at the 1:100,000 scale, however they were derived from photo-interpretation. The Canadian National Topographic Database (NTDB; Natural Resources Canada, 2009) is a digitized vector version of topographic maps with detailed urban features at the 1:50,000 scale which were originally created from manual interpretation of aerial photos (recent activities by the Canadian government to combine the NTDB with other data to create urban mappings is discussed in Section 2). Products which were derived entirely by automated means from satellite imagery for those countries at the national scale lack detailed urban land use classes.

In short, manual interpretation of aerial photographs is still the norm for mapping urban land use (Nichol et al, 2007). However, because there is widespread recognition in the advantages of finding automated means, there has been great interest in alternative approaches. One of these is processing the landscape in an “object-oriented” (also sometimes referred to as a “segment-based”) approach. In an object-oriented treatment of the landscape two-step processing occurs: first the scene or study area is broken into homogeneous component parts (“objects”) according to some segmentation criteria, then the objects are classified by relating relevant spectral, spatial, or hierarchical properties (Johnsson, 1994; Jensen et al., 2001; Qian et al., 2007; Tiede et al., 2010). It is fundamentally different to traditional pixel-based classifiers in that there is the potential to break the scene into “real-world” objects that are meaningful (e.g. an industrial area), potentially incorporate the kinds of site and association information that a human would recognize, and additionally reduce the noise and heterogeneity that is inherent in a per-pixel approach (De Wit and Clevers, 2004).

Successful applications of object-oriented identification of urban land use exist (and tools to aid the process are available in a number of software packages such as IDRISI or eCognition), however are almost entirely image-based and have employed high-spatial, or high-spectral resolution imagery, often in conjunction with detailed city or county-level supportive data (Segl et al., 2003; Kachouie et al., 2004; Carleer and Wolff, 2005; Taubenboeck et al., 2006; Qian et al., 2007; Sun et al., 2007; Lackner and Conway, 2008). At the national level, even if it were feasible to successfully segment 30-m Landsat data into zones of homogeneous land uses, it would require processing hundreds or even thousands of Landsat images (if multi-season images were included), and essentially reproducing the years of work and many thousands of man-hours that went in to creating the NLCD01, in an attempt to redefine land use from scratch.

Our goal is to find a more accessible, non-image based approach to accurately identifying urban land uses that would be feasible over broad areas, not be constrained by the limited footprints of high or even medium-resolution imagery, and at the same time require a minimum of manual interpretation. We hypothesize that it is feasible to identify thematically detailed urban land uses based entirely on existing public national datasets, and to be able to do so at a scale that is still fine enough to be highly useful for regional or national applications. The greatest problem with the object-oriented approach is successfully breaking the study area into segments of homogeneous land use. Because city or county parcels typically represent a single land use, parcel boundaries present one possible solution to partitioning the landscape. For example, Wu et al. (2007) show the ability to classify parcels according to urban land use for the city of Austin, Texas, using high-resolution data. However, at the national or even regional scale parcel boundaries are inconsistent or unavailable (more detail on the difficulties of parcel data is given in Section 1.3). Nationally-available boundaries – census geographies – present the next best alternative. The primary disadvantage is that census zones do not have homogeneous land use even at the smallest unit (the census block), and therefore are not amenable to being classified categorically. Nonetheless, predicting continuous parameters such as population (Lo, 2003; Wu and Murray, 2007) or imperviousness (Civco et al., 2006) have been well demonstrated using a census ***zone-based*** approach.

We believe that a zone-based approach using census boundaries and predicting percentages of land use as continuous variables (i.e. a regression instead of a classification) is a promising alternative to any of the above-described approaches (i.e. manual interpretation, per-pixel approach, or segmenting/classifying by homogeneous zone), would be readily reproducible, and feasible with national-scale data. A further great advantage of using census zones is that numerous socio-economic parameters which are likely to be predictors of land use (e.g. "number of housing structures with more than 5 units in the structure") are already collected and publicly available at a number of census geography levels.

1.2 Objectives

Given the above background, this dissertation has the following objectives:

- Demonstrate a method for a zone-based identification and mapping of urban land use using readily available national-scale data. The thematic resolution of the classes should be at least equivalent to or better than Anderson Level II classes, and the scale of the zones should be no coarser than census block groups. The result will be decision tree models which map thematically-detailed urban land use over broad areas.
- In so doing, create an inventory of datasets and metrics based on literature and classic urban land use theory which may be logically used as predictor variables for this purpose, to include manipulations or creation of data which may not exist

previously, or spatial pattern, contextual/proximity, or measures of spatial autocorrelation.

- Test the feasibility of applying the method (or some version of the method) to diverse geographic areas that were not part of the model building. The results of this will provide insight into how best to perform an automated mapping of urban land use at the national or regional scales, and whether it is possible to apply models across diverse urban settings.
- Because analysis of this type has not previously been performed at this scale, identify which data sets and metrics are most useful in predicting particular urban land use classes and to investigate the importance of certain classes of predictors. In this way, regardless of the ultimate ability to truly make an accurate "national urban land use map", the result will include a large body of information about the relationship of socio-economic and infrastructure data and types of land use settings at a broad scale, which may aid other urban researchers.

1.3 Feasibility of a national approach to mapping urban land use

The intention of this section is to defend in somewhat more detail the need, usefulness, originality, and feasibility of the project. Because Frequently Asked Questions (FAQs) are a concise way of responding to questions, it is given in that format:

Why does anyone care about urban land use at regional or national scales? Isn't that the kind of detail that only local governments or local studies would be interested in, and they would have access to that kind of information at their scale?

Yes, local governments or local studies would probably have access to some kind of land use information, such as parcel or zoning data from a city/county, or could derive it themselves if needed. However, the format and information content varies between every county, may not be publicly available, and may not even exist in digital format.

There are however, numerous national or regional research bodies and studies for whom a national urban land use map would be extremely valuable. For example the USGS National Water-Quality Assessment Program (NAWQA, U.S. Geological Survey, 2009c) or the USEPA National Wadeable Streams Assessment (U.S. Environmental Protection Agency, 2009) are national monitoring programs which study (among other things) the effects of urbanization on the nation's rivers and streams. NAWQA, for example, has a database of several thousand watersheds across the US which are monitored, and the program desires to have the highest-resolution data possible (thematically and spatially) in order to effectively carry out their mission. The lack of urban land use information at the national scale is a striking data gap.

Isn't there a national parcels database or something similar, that would essentially contain this information? What about private companies that appear to have national parcel data, like zillow.com? What about zoning information?

There has been talk about the creation of a National Parcels Database since at least 1980, however it is still a long way from reality (National Research Council, 2007). While standards have evolved for creating such a product, the National Research Council (NRC) notes there are numerous obstacles. Many cities or counties do not have the resources to participate in a project to update and reformat their data to a national standard, a non-trivial effort. (Fairfax County, for example, has > 350,000 parcels). The NRC estimates that about 30% of county parcel data do not even exist in digital format. Even when/if a national parcels database comes into existence it is not clear that it would necessarily contain the type of information desired (land use of the parcel), or that the information would be nationally consistent even if it did exist.

Some private companies like zillow.com have, at great effort, gathered property information from local sources, primarily for the purpose of pricing. Their databases are generally not for sale, do not necessarily contain land use information as such, and do not necessarily cover the entire United States.

Having said that, this project will test the value of one such proprietary system (whose data can be purchased) – ESRI's Business Analyst data (ESRI, 2009) as a predictor input. It by itself, however, does not contain land use information beyond business and some institutional location information.

Incorporating zoning information entails many of the problems noted above, because while every county or town in the U.S. has publicly-available zoning maps, they have variable class definitions, consistency, and currency. Even for adjacent counties within the same state, such as Massachusetts, zone classifications might have similar names but quite different definitions (MassGIS, 2008). Numerous exceptions may exist for different zoning classes, but the exceptions themselves differ for each county or town. For example, some residential zones may permit churches or recreation or commercial land to exist within the zone in one county, but the same exceptions might not exist for another county, even if the zoning classes are ostensibly identical. Finally, and possibly the greatest difficulty, zoned land use may be considerably different to actual land use.

Why isn't land cover or percent imperviousness, such as that portrayed in the NLCD01, good enough – doesn't it basically have the same information?

No. It's true that the difference between land cover and land use is often ignored in ecological or other studies (Cadenasso et al, 2007; Comber, 2008), however land use differences are important for testing numerous hypotheses, and the land surface's form is different to its function. As noted earlier there is quite a leap from land cover to land use, and knowing the percent imperviousness of a watershed does not have the same information content as knowing how much land is used for industrial or residential or parkland purposes. It is for that reason that many studies have been undertaken over the years (see Literature Review section) to explore better ways to identify land uses.

There was urban land use information in the 1970s GIRAS and in the 1992 NLCD – can't one or both of those simply be updated?

Those products would certainly be helpful in making a current national land use map to the extent that areas that were correctly coded in those products could still have the same land use. The plan here is to test the viability of both of those products as predictors of current land use, and it is likely they will end up participating in the model. So in a sense they will be updated, just at a different scale and resulting in somewhat different classes.

What has the focus been on in the past with automated urban land use mapping?

Many studies have focused on image-based solutions and particularly in demonstrating how well urban land use can be delineated using high-spatial or high-spectral resolution data. Other studies have focused on non-image based solutions, but nevertheless using detailed local-scale data. Those studies show great potential but are extremely difficult or impossible to execute over very broad regions because of the unavailability or expense of those kinds of data. Perhaps someday very high resolution data will be cheap and easily accessible for large areas, however that day is still far away.

And even when that day arrives there are still data processing issues with regard to national mapping. For example, processing 4-m multi-spectral imagery over the entire U.S. for the purpose of land cover delineation would be a stupendous data processing task. Even the proposed method in this project is not without challenge: although the method here uses nationally-available data, from a data processing perspective it would still be a non-trivial amount of work to execute for the entire country (for example, there are about 220,000 block groups in the United States). There are a limited number of organizations who have the mandate or motivation to undertake data-intensive national mapping.

So you think urban land use can be mapped with reasonable accuracy using readily-available national data without having to individually process thousands of unique images. If that's true then doesn't the information essentially already exist?

Yes, the hypothesis is that the information does exist, but needs to be teased out and successfully modeled from the predictor data, which is likely to be non-trivial (and if it were trivial then someone would already have done it). It also likely includes non-linear relationships, i.e. data relationships will exist at some range of values but not at others, which are more difficult to model. Some of the classes to be presented here are definitely not predictable from land cover or population density alone (e.g. "Institutions"), and identifying some classes may require unique combinations of data or metrics which have not been tried before for this purpose or at this scale. It is hypothesized, for example, that spatial pattern metrics might be useful for predicting some classes, but not for others, and the method presented here allows for focus on each individual land use class.

If you intend to use a zone-based approach, what is the appropriate scale? Blocks? Block groups? Tracts? Counties?

This is an important issue, because the resolution of the zones will determine how useful the end product will be. For example, if the resolution of the zones was at the state or county scale they would be of limited use in characterizing watersheds that were (for example) smaller than 50 km^2.

The best result of this project would be accurate estimation of land use at the smallest census geography, possible, i.e. census blocks. However, because the intention is to extract spatial pattern metrics from 30-m data (the spatial resolution of the national land cover datasets), there is an obvious need to match the grain of the data (30-m) to the extent of the zone (Saura, 2002). To that end, the sizes of census geographies were examined in two of the states which will be involved in this study (and likely to be similar to other states) (Table 1-2):

Although there are no definitive guidelines (and the requirement may vary with the type of metric), Herold et al. (2003) suggest that several hundred pixels are an appropriate minimum for executing various spatial pattern metrics. Because at the block geography only a few dozen 30-m pixels would be contained in a median-sized zone (and therefore half would have even fewer than that), the block group geography appears to be the smallest feasible scale for this project. The vast majority of block groups in these two states (> 90%) are made up of at least 200 30-m pixels. This study will therefore be conducted at the block group scale.

Table 1-2: Representative statistics of census geographies for two sample states
(block boundaries from: SILVIS Lab, 2009, other data from: U.S. Census Bureau, 2009a)

	California			**Massachusetts**		
	n	**median size (sq km*)**	**# 30-m pixels in median area**	**n**	**median size (sq km*)**	**# 30-m pixels in median area**
Blocks	561,218	0.03	33	118,171	0.02	22
Block Groups	22,195	0.53	589	5,054	0.86	956
Tracts	7,115	2.03	2,256	1,364	4.25	4,722
Counties	58	4070.81	4,523,122	14	1465.59	1,628,433

* 1 sq km = 100 ha = 247 acres

It is also worth noting that there is a quite a bit of variation in block group sizes. For example, even though the median block group size for Massachusetts above is 0.86 km^2, 25% are > 3.0 km^2 (rural or less developed block groups), and another 25% are < 0.25 km^2 (urban block groups). Census geographies were originally designed to roughly encompass the same number of individuals being enumerated, however both the population and boundaries have changed over time. In any case, it is believed that block groups are a good compromise between the desire for fine-scale mapping, and the reality of doing so over a large area. While block groups may be coarse for some applications, as another perspective, the median values noted above are smaller than a single pixel from a typical regional-scale remote sensing platform such as the AVHRR series of satellites (approx. 1-km resolution) (Figure 1-5).

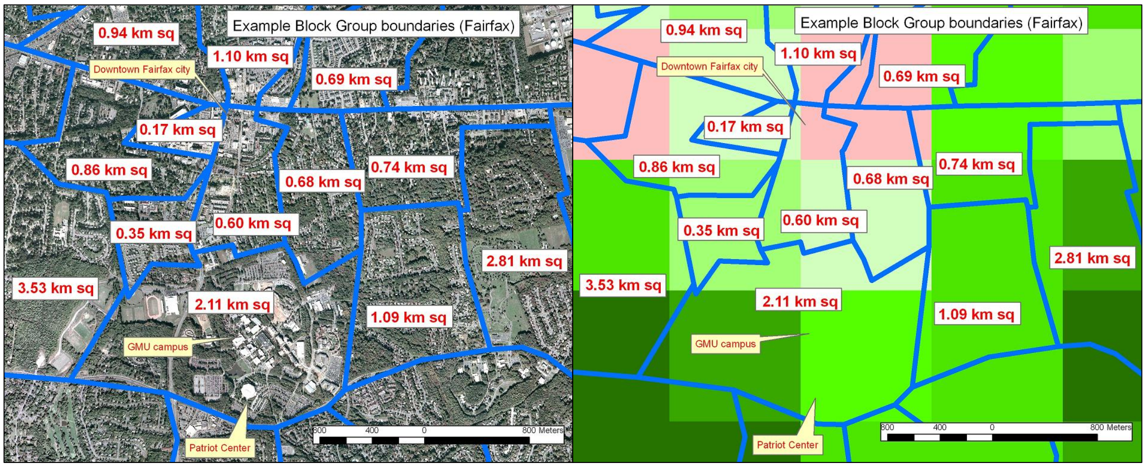


Figure 1-5: Example block group boundaries (blue) over Fairfax, VA.
Area of the block group is shown. Right-hand side: for the same area, 1-km pixels are shown for scale perspective (USGS National Atlas vegetation for year 2000; USGS, 2008).

You said that block groups do not have homogenous land use, and that the method would be demonstrated by predicting land use as a continuous variable. How is that going to work?

The percent of each land use type within a block group will be predicted independently. That is, a block group may be predicted to have 40% single-family residential, 10% multi-family residential, 20% commercial, 20% industrial, and 10% recreational land use. To assess the accuracy those values will be compared to the actual values for withheld validation records. The independent stand-alone models may or may not add up to 100% land use, so they will be integrated and constrained after the fact so that they do in the end represent a “holistic” 100% prediction (although they also have value by themselves as stand-alone predictions). Assignment of categorical classifications (for example, assignment by majority percent) will also be explored to some degree after the fact.

Is the method going to entail using special software or data that will be difficult for someone else to duplicate?

No, and in fact accessibility is one of the key tenets of this project, and it is intended that methods should be reasonably simple to reproduce. Very commonly used GIS software will be employed (ESRI ArcInfo), and other tools, including the regression and prediction software (R platform), are all public domain.

Data to be used will also be public-domain and freely available with the exception (as noted previously) of a test of a sample of the proprietary ESRI Business Analyst data.

2. Literature Review

Literature review for this topic is broken into four categories: determinants of land use, urban land use classification at broad scales, object-oriented or zonal approaches to urban land use identification, and spatial pattern and contextual measures used for urban land use identification. Key terms which provide a conceptual basis for predictors and modeling methods used in this project are **bolded** for emphasis.

2.1 Determinants of land use

The most basic and first formal spatial model of land use theory was proposed by J.H. von Thünen in 1826 pre-industrial Germany (Fellman et al., 1992). Although von Thünen's model was based on agricultural land use, it is noteworthy primarily because it was the first model to incorporate ideas of distance-to-city and transportation costs as drivers of land use and value ("land rent"). In it, an idealized isolated city existed, around which agricultural production occurred, and transport costs were based on distance to the city. The locational rent for any unit area, what could be gained from farming or utilizing it, was therefore a function of distance to the city, and represented as:

$$LR = Y(\text{Price} - \text{ProdCost}) - (Y)(\text{Trans})(d) \qquad (1)$$

where LR = locational rent, Price =maket price per unit of commodity, ProdCost = production cost per unit of the commodity, Trans = transport cost, d = distance to the city, and Y = yield (quantity of production). These market forces resulted in ringed zones of land use (Figure 2-1), in which the innermost land uses were the most intensive, and the outermost land uses were the least intensive, e.g. grazing of livestock. Because locational rent drops as distance from the city increases, the production of some goods which required certain intensities (production per unit area) became unprofitable at some distance. **Distance to the central city** and **transportation costs** as they related to requirements for production intensity were therefore the key determinants of land use. The intensive/extensive nature of urban land use with regard to distance from the city is most prominently evident today for most cities in the form of **residential density**.

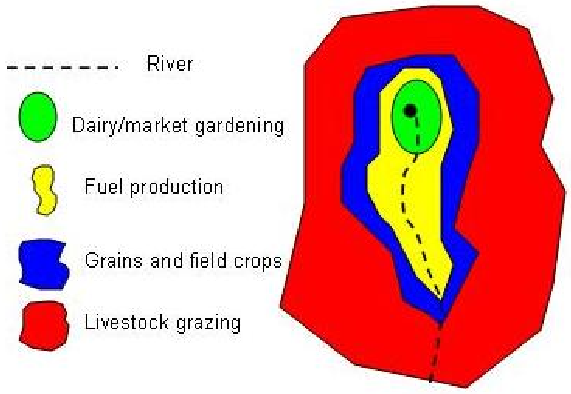


Figure 2-1. Von Thünen land use model. The city (market) is represented by the central black dot and major transportation by the dotted line. The most intensively produced crops are found in the areas closest to the city and transportation routes, while those which have extensive uses of land are located farthest from the city. Figure adapted from Chisholm, 1968.

The idea of **centrality** underlies another seminal work on location theory, that of William Alonso (Alonso, 1964), which extended some of the general von Thünen model to cities and residential location. Alonso's premise was that central sites, which have highest accessibility, are attractive to most land users. However, some land users may value centrality more than others, for example a retail shop may require centrality, a residential user perhaps less so, and agricultural uses may have even lower need for centrality. Every user then has their own "bid-rent curve": a curve that gives the price that a user is willing to pay for sites at various distances from the center, depending on their demand for space and location (Figure 2-2).

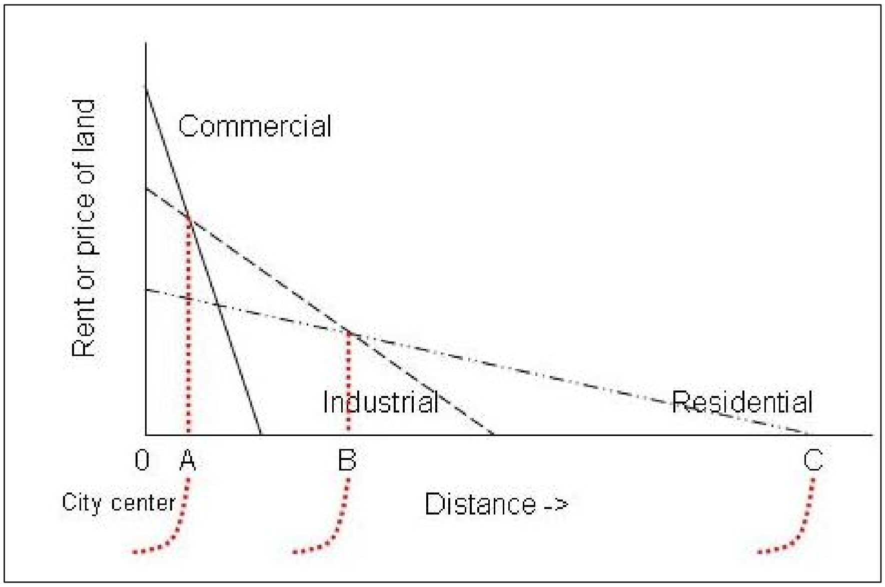


Figure 2-2. Example of bid-rent curves.
Commercial land uses have higher requirement for centrality, and therefore shopkeepers are willing to pay more for land closer to the city center. Industrial land uses have less demand for centrality, and residential uses even less so. If rotated around the central axis the area between 0 and A forms a concentric ring of commercial land, between A and B of industrial land, and between B and C residential land (adapted from Kivell (1993)).

At central locations retail or commercial users will prevail, but at greater distances residential users will be willing to pay a higher price: superimposing the curves allows a pattern of concentric land use to emerge, but in this case within the city itself. As transport or accessibility costs drop the influence of location drops. In this idealized model the Central Business District (CBD) resides at the center of activity. Important works of other urban economists include Richard Muth (1969), Edwin Mills (1980), and Grant Thrall (1980), who expanded and refined the **socio-economic effects** of income and consumption on the monocentric model.

There are three classic urban land use models that describe urban structure, derived from the social and geographic disciplines, but which may be considered as variations of land rent models based on economics. Although urban environments have changed considerably since the models came into being (1920s-1940s) they still broadly describe many aspects of urban places, in a general sense. The models are:

1) Concentric zone model (Burgess, 1925). This model describes a series of six concentric rings around the CBD, which reasonably approximated some American cities at the beginning of the 20th century. The land-rent curves and von Thünen rings described above provide a degree of explanation for this model. At the center the CBD provides commercial services, and surrounded by a second zone of uses which were supportive of that: wholesale, transportation and light manufacturing. The third zone represented older residential areas which had become low income, and abandoned by higher-income residents. The fourth zone represented primarily blue-collar workers who owned their own homes. The fifth zone was a zone of single family homes which were wealthier and who could afford to travel to the central city. The sixth zone represented "satellite cities": low density, upper-middle class residential areas which were the beginnings of

what is today suburbia. There is a distance-decay factor: values of land per unit area generally decrease with distance from the CBD.

The concentric zone model was built on the ecological analog of invasion and succession, i.e. people found their preferred niche, as species do, and represented a kind of continual conversion of land use as cities continued to expand. The model was developed before automobile transportation was commonly available, and therefore did not take into account today's modern transportation modes.

2) Sector model (Hoyt, 1939, as cited in Mather, 1986). This model proposed that **socio-economic status** varied in a sectoral fashion, the **sectors being oriented strongly along transportation axes**. The city is still oriented around the CBD as a commercial district, however other areas take advantage of road and rail systems in their pattern of development. As in the concentric model lower-income residential areas are located closest to the CDB and nearest to industrial areas. Medium-class residential areas are allocated farther out, and a high-income axis exists in which wealthier people with automobiles have access to the central city. As in the concentric model the city center is still assumed to be the center of employment.

3) Multiple nuclei model (Harris and Ullman, 1945, as cited in Mather, 1986). This model acknowledged that **multiple city centers** could exist, and that peripheral growth spread from numerous sources. In this model, an original core of CBD still exists, however now land uses tend to organize themselves in mutually-beneficial ways, aided by the existence of modern transportation. There is

agglomeration of uses, and attractive or repelling forces of land use compatibility play a large role. Locations of specialized functions develop. In this model, there is still separation between highest-class residential areas and low class, as well as from heavy industrial, and land uses may leapfrog non-urban areas. **Income** strongly defines residential zones. The Harris and Ullman model was one of the first to describe the multiple-center suburban patterns of the landscape that were to follow in later decades.

Classic models of land rent and urban structure necessarily make certain assumptions, however, as Mather (1986) notes, there are numerous reasons why patterns of land use in the real world deviate from theoretical patterns. Firstly, land use is affected by **drivers at several scales**. Broadest-scale predictors of land use include natural elements such as climate, topography, geology, soils, water supply, and access to navigable water bodies (Walsh et al., 2003). It is entirely predictable that there are no cities in Antarctica, nor permanent habitation at elevations > 20,000 feet. At some finer scale the presence of anthropogenic features and institutions become important: presence of cities, jobs, transportation and infrastructure facilities, or quality of schools. Other medium-local scale forces include land use regulations, aquatic amenities, or social forces: the attraction of cultural or ethnic enclaves, quality or access to arts or sports, or age of population. At any even finer scale characteristics of specific sites drive demand; for example lakeside or riverside land may be more desirable (or a requirement) to residential or certain industrial users than other users. Technology also plays a part: for example, improvements in telecommuting technologies may reduce the desirability of residential land near employment centers. Not all land is desired or used in an economically rational way, and it is likewise clear that the urban land market functions imperfectly (Kivell, 1993). Economic cycles of boom and bust affect how land is desired – commercial land may be more desirable in certain periods. Political and cultural changes affect land use – for example the construction of public housing may be a sought-after goal of one administration but not another. Inflation, availability of credit, or growing affluence may likewise affect many aspects of urban change – for example, purchase or construction of second homes. Meyer and Turner (1994) categorize the human driving forces of land use change as: (a) population and income, (b) technology, (c) political-economic institutions, and (d) cultural.

McKnight (2001) suggests that the morphology of many North American cities have a general pattern. The CBD is the commercial (and sometimes geographical) center of the city. He also notes the wide variability of commercial and retail forms, to include suburban shopping malls or big-box stores, or strip shopping centers and linear commercial development along road systems. On the margin of the CBD is a transition zone (Mather 1986; McKnight, 2001): a discontinuous area of irregular shape and unpredictable size that has a changing land use pattern: commercial, residential, or industrial. This zone may also function as tenement section of low-income housing. Outside the transition zone are more typically residential areas, which take up the greatest areal extent of the urban form: most generally residential areas are densest closer in to the central city, and less dense in the outer suburbs. Suburban areas have become more complex over the decades and often themselves evolve into nucleated centers of commercial uses. Higher income households are often in suburban locations in the United States. Industrial areas, particularly in older American cities, often organize themselves along major transportation routes: waterfronts, rail, or rivers. Other land uses which have relatively low percentage of land area but are critical to the urban fabric – transportation, recreational, or institutional - have typically a less predictable location and are more scattered. McKnight remarks that the central city in North American has been for decades decreasing in variety and importance in function with suburbanization and sprawl.

In addition to the structure of an individual city, systems of cities evolve. The German geographer Walter Christaller in 1933 was one of the first to specify a *Central Place Theory*, which acknowledges the interaction and complementary effects of urban places. He noted that the system of central places was interdependent (Fellmann et al., 1992), and were one town eliminated the entire system would have to readjust its spatial pattern or change production to provide consumers with needed central place goods and services. He also noted that a regular **hierarchy of central place sizes** exist: for example that the ratio of second order towns to first order towns will be roughly 3:1, and that there is a regular and relatively predictable regular spacing between towns of certain sizes. Christaller’s characterization of central places, while idealized, reinforce the notion of interconnectivity, access, and **influence zones based on proximity to cities of different size**.

Given this background, what are likely to be general indicators that would help us to identify urban land use types? Certainly **socio-economic** factors such as income, home size, family size, or mode of transport to work are likely to be related to a number of different types of land use. **Land cover**, both current and prior (land cover and land use, if available), is an obvious close cousin to land use that has been the basis for identifying uses (e.g. Herold et al., 2003), but it remains a challenge to clearly link urban surfaces to urban functions. **Transportation** routes are key element controlling land use models both theoretically (e.g. Hoyt, 1939; Thrall, 1980), and in practical land use identification studies (e.g. Wu et al., 2007). The presence of **infrastructure landmarks** (such as schools, hospitals, or airports), or **amenities** (such as golf courses) in part drive the desirability of residential areas. The **spatial pattern and agglomeration** of landscape features or socio-economic factors has also been a clear indicator of land use types when based on relatively high-resolution data (Segl et al., 2003; Barr et al., 2004; Mesev, 2005). In this project we distinguish between spatial pattern at a local scale (within a block group), and at a broader scale (e.g. spatial autocorrelation across block groups). Finally, **proximity** and **accessibility** to city centers and other features is perhaps the strongest conceptual predictor of land use from classic land use and land rent models. Proximity to centers is likewise likely to play at a role at multiple scales, e.g. distance to both large and small city centers.

2.2 Urban land use classification at broad scales

As noted earlier, the advent of satellite remote sensing (RS) platforms, particularly Landsat in the early 1970s, made possible some of the earliest efforts at automated broad-area mapping of urban land use/land cover ("broad-area" defined here as multi-county or larger areas, in the U.S.). The history of the national mappings of land use/land cover in the U.S.: the GIRAS (Price et al., 2006), NLCD92 (Vogelmann et al., 2001) and NLCD01 (Homer et al., 2007) has been noted in the Introduction. There is widespread agreement that **land use is not a physically-measurable quantity, but a combination of cultural and economic factors which may only have indirect links to land cover** (Campbell, 1996; Barnsley and Barr, 1997; Mesev 2003). Many studies have long recognized the difficulty of differentiating urban land uses (or detailed urban features) from moderate-resolution imagery such as 30-m Landsat, particularly using spectral based pixel-oriented methods alone (Welch, 1982; Forster 1985; Haack, 1987; Jensen and Cowan, 1999; Jensen et al., 2001; Aplin, 2003; Mesev, 2003; Falcone and Gomez, 2005; Lackner and Conway, 2008).

In addition to the NLCD92, a number of studies have had success in mapping urban land use by incorporating 30-m imagery and data outside the spectral domain: either spatial/textural, contextual, socio-economic, or cadastral-based data. Because Landsat images cover a fairly broad area (~ 180 x 180 km) these 30-m Landsat-based products demonstrate the possibility of mapping land use over a large metropolitan area. The majority of these studies have incorporated fine-scale ancillary data (such as city or county-based data), which generally are unique to that location and may not be available consistently elsewhere. The studies include Moeller-Jensen (1990) who used a knowledge-based (object-oriented) approach incorporating contextual information and textural measures of Landsat-TM imagery to demonstrate discriminating between land uses. Harris and Ventura (1995) used census and local zoning data for post-classification sorting of Landsat-derived urban classes and increased the thematic detail from 5 to 14 urban classes in their final product in doing so. Debeir et al. (2002) likewise incorporated texture and contextual information with Landsat data to map CORINE-like classes (European Environment Agency, 2009). Sun et al. (2007) used an object-oriented approach which incorporated detailed city-level data with Landsat data to simulate time series of urban land use. Smith and Fuller (2001) used parcel-level data to assist Landsat mapping to derive five urban classes. Vogelmann et al. (1998), Chen (2002), and Yu and Wu (2006), all likewise employed ancillary data and Landsat data to good use in identifying some form of urban land use. It is noted in numerous studies that the quality of the ancillary data are very important to the overall result (Vogelmann et al., 1998), and that dasymetric mapping (combining data sources to map a quantity) may improve classification (Schumacher et al., 2000).

There is evidence, therefore, that although 30-m imagery by itself is recognized as not being adequate for thematically-detailed urban land use classification, incorporating supportive GIS ancillary data may greatly improve land use identification from 30-m image data or products. In some studies it is recognized that the GIS data, particularly if fine-scaled, are the primary source of information, and imagery only secondary (Wu et al., 2007).

Several issues exist however with 30-m Landsat-based solutions to urban land use delineation (or other satellite platforms with similar spatial resolution and swath width). The first is that each image is unique based on the distinctive characteristics of the surface materials that were imaged at the moment of sensing, as well as atmospheric conditions, sun angle, cloud cover, sensor calibration, and how the data were processed and stored after the fact. Each image must therefore be processed and classified uniquely, and classifications based on one image cannot generally be applied to another. In a typical land cover classification as was done for the NLCD01, several images for the same image footprint are often processed for better accuracy (i.e. at least a leaf-on and leaf-off image). Even if the goal is to classify a single metropolitan area, the footprint for a single Landsat image may or may not coincidentally cover the area: in the case of Boston, for example, the Landsat path/row break occurs through the middle of the study area. That is, the most basic approach at a Landsat based classification of the Boston metropolitan area would require processing at least four unique images. As another example, a land cover derivation with which the author is intimately familiar (Falcone and Pearson, 2006) for the Dallas-Fort Worth area based on the NLCD01 protocols required 24 unique Landsat images, because of the unlucky way that Landsat footprints fall over that metropolitan area (see Figure 6 of that reference). Image data of higher spatial resolution have correspondingly much smaller area coverage and may be very expensive to acquire over large areas, as has been previously noted. Image processing is additionally less accessible to most users: while many researchers are comfortable with GIS-based processing, a much smaller subset have the specialty image processing tools and requisite skills for handling and classifying multi-spectral imagery. An effective GIS-based method requiring only seamless national-scale data, which are furthermore publicly-available, and requiring only common GIS and public-domain software, has therefore several inherent strong attractions.

Two Canadian efforts have been proposed or exist which are similar in conception to that proposed here. The first of these is proposed by Lemonsu et al. (2008) to create detailed 12-class urban maps for any Canadian or US city based on 15-m ASTER and 30-m Landsat imagery, and incorporating population density data and estimates of building heights using surface DEMS and DEMS from the shuttle radar topography mission (SRTM), which are available continental-wide. The methodology however requires processing unique imagery (ASTER or Landsat) for each city, which in the case of the United States would amount to thousands of unique processing tasks. The second (Leroux et al., 2009) supersedes the Lemonsu idea (they are both co-authors on the other's paper), and is similar, but replaces land cover derived from imagery with land cover/land use derived from the Canadian National Topographic Database (NTDB). **Their replacement of an image-based solution with a GIS-based solution for urban mapping echoes the arguments of this dissertation**. The NTDB is a 102-class mapping of detailed polygon, point and line features, which is **modeled using a decision tree technique** combining census information and building height estimates to create a 5-m 44-class raster land use dataset, of which 34 of the classes are urban. However, because the product, known as UrbanX, is based on the very detailed and consistent Canadian NTDB, it is Canadian-specific. (An equivalent of the NTDB does not exist in the US; the closest thing being USGS 1:24k Digital Line Graphs, for which there would be a number of significant issues to use in a similar manner, primarily consistency). UrbanX is complete for all major Canadian cites, but not publicly available (A. Leroux, personal communication, Jan 20, 2010).

Ridd (1995) suggested that urban landscapes could be mapped to some degree from a linear combination of three bio-physical parameters (vegetation, impervious surface, soil; so-called V-I-S model). V-I-S based classifications have been employed using Landsat data in a number of studies (Wu et al., 2005; Setiawan et al., 2006).

Comber (2008) notes the common confusion between land cover and land use and proposes an approach for separating them using a set of 14 literature-based “data primitives” that allow a mapping of where the NLCD01 primarily depicts land cover (primarily natural land cover types), where land use (agriculture and urban classes), and where the concepts are confused. The primitives, however, are not information external to the NLCD01, but rather, categories which are assigned a score based on the NLCD class descriptions: for example: naturalness, vegetation height, wetness, biomass production, estimated amount of human activity, etc. The categorizations are useful in that they propose a method for separating “land cover” from “land use” classes within the NLCD01, but do not distinguish among urban land uses.

A national mapping effort which is also similar in nature to that proposed here, but focused on population mapping only, is the Oak Ridge National Lab Landscan project (Dobson et al., 2000; Bhaduri et al., 2007). Two products exist, one a global mapping of population, and a second, more spatially detailed mapping of the USA, at 3 arc seconds (about 90-m). The motivation behind the Landscan USA is to map population not only spatially but temporally, including a “daytime” and “nighttime” count. The general methodology allocates population count to individual 90-m cells. Cells are weighted by a probability coefficient (probability of population in the cell) based on 10 variables (in Landscan USA 1.0). Then the population count for the block is apportioned to all the cells in the block according to the probability coefficient weighting for each cell. Weightings are assigned from the input variables based on expert judgment, i.e. how important it is believed those variables are to that particular county or state. The final product is evaluated county by county by a GIS analyst who checks for discrepancies and verifies it against high resolution imagery. The 10 probability variables used are: **land cover, proximity to roads, proximity to rail, slope, landmark polygon feature, parks, schools, prisons, airports, and water bodies**. Most of the data sources other than land cover are based on commercial data, i.e. accessible only at fairly significant cost**.** It is noteworthy that most of these variables are also employed in this dissertation, although their exact form might be somewhat different. There are thus similarities between Landscan and the method presented here in the sense that **multiple national data layers are used to map a characteristic of urbanization** (in Landscan’s case population count), however differ in the end product being produced and in the method details. Landscan’s method benefits from the great advantage of knowing *a priori* what the total population count is in the block so that results may be constrained within the zone, whereas in the case of urban land use no such knowledge exists (except as may exist in disparate and local datasets).

There are at least two additional proposed efforts outside the U.S. which are pertinent to that proposed in this dissertation, and explore similar methods. The first of these is a conceptual framework for a spatial database of Russian urban areas (Perepechko et al., 2005), based on combining multiple source data (e.g. demographic, sociologic, image-based, infrastructure, human-expert categorization). The existence of such a database would allow for detailed urban studies to be enacted at multiple scales. The second proposal is for the construction of a national land use dataset from public domain information for the United Kingdom (Wyatt, 2004). Wyatt notes that enough publicly available data exists in the UK at the land parcel scale to be combined to make a comprehensive land use dataset for any urban area, and proposes a methodology to do so. It is noted here, however, that the same detail of information publicly available in the UK does not exist at the national level in the United States (e.g. parcel-level data descriptions).

In short, while some image-based applications of identifying thematically-detailed urban land use over broad areas have been demonstrated, they have been limited to the area of a single Landsat footprint, and even then only feasible by incorporating ancillary data. A different solution, one which incorporates and models data from numerous sources, to include land cover data already derived from RS data, is recognized as a more viable solution for very broad or national-scale urban applications (e.g. examples are the Canadian UrbanX and U.S. Landscan projects).

2.3 Object-oriented or zonal approaches to urban land use identification

Object-oriented and zonal approaches to identifying land use are conceptually similar to the extent that they are departures from traditional per-pixel classification of the landscape, and that contextual relationships may be incorporated as important elements of land use identification. Humans are very adept at incorporating bits of information into a *knowledge-base* and accurately classifying objects from that. Experienced image interpreters (or even ordinary people) are able to identify features in imagery based on a lifetime of contextual experiences: one recognizes that small-medium sized structures with driveways next to them and arranged in certain patterns are likely to be single-family residential homes; or that very large structures surrounded by acres of parking lot are likely to be commercial or sports complexes, etc. **What is missing from traditional per-pixel classification of imagery is the ability to leverage contextual information about proximity and relationships to other features in the landscape** (for example, agglomeration or spatial autocorrelation of features). There is therefore very good intuitive rationale for not attempting to classify urban land use through traditional per-pixel approaches.

Object-oriented approaches and related techniques such as artificial neural networks, represent a type of artificial intelligence (Jensen, 1996). They are similar to the extent that the system is trained to recognize objects from a knowledge base similar to that of a human expert, which typically includes an understanding or representation of hierarchical, network, or neighborhood relationships. A number of studies have incorporated object-oriented ("knowledge-based") processing in classifying urban land using imagery data. These include Moeller-Jensen, 1990 (30-m Landsat); Johnsson, 1994 (10-m SPOT data); Bauer and Steinnocher, 2001 (4-m IKONOS); Kachouie et al., 2004 (1-m IKONOS); Carleer and Wolff, 2006 (0.6-m QuickBird); Taubenboeck et al., 2006 (4-m IKONOS); Cabral, 2007 (30-m Landsat); Dong and Wu, 2007 (10-m SPOT); Qian et al., 2007 (30-m Landsat); Stow et al., 2007 (0.6-m QuickBird); Sun et al., 2007 (30-m Landsat); Lackner and Conway, 2008 (4-m IKONOS); Su et al., 2008 (0.6-m QuickBird), and Aubrecht et al., 2009 (airborne laser data). Likewise, artificial neural networks have been used to determine urban land from objects (Jensen et al., 2001; De Lira et al., 2006; Rocha et al., 2006)

As noted in the Introduction section, a typical object-oriented treatment of land classification occurs in two steps: a segmentation step and a classification step. How the image is segmented is arguably the more difficult of the two. Wu et al. (2007) suggest it may be done in one of three ways: by manual delineation of boundaries, by segmentation from the image data themselves, or by using pre-existing administrative zonal boundaries. Manual delineation of boundaries for contiguous areas of the scene (perhaps without knowing yet how to label the areas) is advantageous in that a human interpreter is likely to be able to partition the area more effectively than an automated process. This was done by Herold et al. (2003) for the purpose of testing the effectiveness of spatial pattern metrics on identifying land use, and by Dean and Smith (2003) for mapping land cover from high-resolution imagery. It is disadvantageous, however, in that it requires manual intervention by an expert, and is essentially infeasible over very large areas.

The second method is the application of image or scene segmentation techniques, which identify areas of spectral and/or spatial similarities. This also is sometimes called "region-growing" (Qian et al., 2007), and is perhaps the most common technique for creating objects. This has been demonstrated by Segl et al. (2003); Kachouie et al. (2004); Carleer and Wolff (2006); De Lira et al. (2006); Rocha et al. (2006); Taubenboeck et al. (2006); Qian et al. (2007); Sun et al. (2007); Lackner and Conway (2008), among others. Because individual image scenes typically have unique properties which do not apply to other images, this technique is generally practiced from a single image, and therefore is not ideal for regional or national mapping. It has not been demonstrated across regions or multiple metropolitan areas. For non-image data, strategies have been proposed which may "re-partition" Census areas according to user-defined criteria (Openshaw and Rao, 1995; Poulsen, 2002), however, these lack the simplicity and well-recognized nature of already-defined Census boundaries, as well as the potential for loss of accuracy when Census statistics are re-aggregated.

The third method of identifying objects – using pre-defined governmental or administrative boundaries which have a homogeneous land use – has also been used. These are primarily based on parcel boundaries. Wu et al. (2007 and 2009b) used this method based on tax parcel boundaries and building footprint outlines, along with detailed (0.6-m) imagery to classify parcels. Bauer and Steinnocher (2001) also employed parcel boundaries for classification. Smith and Fuller (2001) manually modified detailed vector data to create parcel boundaries of areas of homogeneous land use for classification, with Landsat data. Tiede et al. (2010) likewise used city parcel boundaries as the basis for a land use classification. The advantage of areas as small as tax parcels is that they have homogenous land use and thus are amenable to categorical classification.

An alternative to categorical classification is **estimation of a continuous parameter for a census zone, which is the approach of this project**. Census geographies (blocks, block groups, tracts) have been used as a basis for characterizing various continuous-variable parameters, for example impervious surface (Civco et al., 2006 – by tract), population or housing unit density (Lo, 2003 – by tract; Wu and Murray, 2007 – by block group; Bhaduri et al., 2007 – by block; Hardin et al., 2008 – by block), or "quality of life" (Li and Weng, 2007 – by block group). **However, thematically-detailed urban land use classes have not been processed in this way, even for localized areas.** Of the above studies, Hardin et al. (2008) had perhaps the greatest similarity to this project, except that the target dependent variable was housing unit density, as opposed to our 10 urban land use types. They predicted housing unit density for 1,945 census blocks for the city of Terre Haute, IN. Their final model, based on a multiple regression equation, used seven predictor variables based on the percents of five land cover classes from a classified Landsat image, and two spatial pattern metrics likewise derived from the Landsat classified image. 50% of the study area records were used for training and 50% for validation, and they reported an r^2 of 0.62 if 63 outlier records were removed, and r^2 of 0.37 if outliers were kept.

In summary, while object-oriented approaches, which classify areas of homogeneous land uses, have been demonstrated based on high-resolution imagery or very detailed ancillary data, they are much less feasible over broad areas. The zonal-approach, which estimates a continuous variable for a zone, is a much more promising framework for processing broad areas, however has not been demonstrated to date for urban land use applications.

2.4 Spatial and contextual measures for urban land use identification

Incorporating spatial or contextual measures of the landscape, i.e. information beyond the traditional spectral domain, has been another method which has shown promise over the years in identifying urban land use. Spatial measures encompass a wide variety of possible methods, but may be broadly defined as a technique that characterizes forms and patterns across the landscape, taking into account relationships between elements in some fashion as a human eye might see it (Jensen, 1996) – for example how close buildings are to one another, their shape, whether or not boundaries are sharp or gradual, how much variation exists, etc. The techniques are not new in image processing, and to the extent that they represent a basic form of pattern recognition are not new whatsoever. For example, one of the most common, characterizing the *texture* of a scene has been used successfully in various types of analysis at least back to the 1950s (Haralick et al., 1973; Hsu, 1978). There is no single definition of texture (Debeir et al., 2002), however it is essentially the impression of smoothness or coarseness in an image, and may be measured by the tonal variations within a certain neighborhood (e.g. kernel processing using a 3x3 or 5x5 moving window;). Simple statistics such as range, standard deviation, or variety may be straightforward measures of texture (Gong et al., 1992).

Forster (1993) showed that housing density and size were functions of the coefficient of variation in various moving window sizes from both SPOT and Landsat data. Other texture measures, such as lacunarity, which measures the presence of gaps (Dong, 2000), the grey level co-occurrence matrix (GLCM, Haralick et al., 1973), Moran's I (coefficient of autocorrelation; Bowersox and Brown, 2006; Su et al., 2008), or semi-variograms (Brivio and Zilioli, 2001) are somewhat more complex to compute, but are similar in nature. Many other studies have employed texture as a measure of spatial characterization (Moeller-Jensen, 1990; Zhang et al., 2003; Tso and Olsen., 2004; Moeller-Jensen et al., 2005; Liu et al., 2006; Cabral, 2007). Most texture measures are scale-dependent, i.e. may vary with window size (Dong, 2000).

Other methods exist which are based on the adjacency of classified pixels within a moving window kernel**.** It has been shown that classifying how pixels are clustered within a 3x3 or 5x5 window (for example) are useful for inferring urban land use (Barnsley and Barr, 1996), or other patterns of fragmentation (Riitters et al., 2000).

Another approach might be called a **"patch-based" method**, in which patches of contiguous classified pixels are identified and spatial pattern metrics are calculated based on their association (MacGarigal and Marks, 1995). These have been shown to be advantageous in numerous urban land use studies, in that a wide range of metrics exist which may help to characterize the landscape: connectivity of patches (*cohesion*; Saura, 2004), edge measures (*edge density*; Herold et al., 2003), *fractal dimension* (Bowersox and Brown, 2001), distance of patches to like patches (*Euclidean nearest-neighbor,* Herold et al., 2003), *linearity* (Wang et al., 2008), *perimeter-area ratio* (Salas et al., 2003), among others. In the sense that point locations (e.g. postal addresses) may be considered patches, similar metrics may be executed for points, for example nearest-neighbor distances (e.g. Mesev, 2005). At the national scale, these metrics, based on the NLCD01 categorical land cover data, provide the possibility of aiding land use identification.

Morphological properties of buildings (e.g. size, compactness) have been examined in several studies (Segl et al., 2003; Barr et al., 2004; Wu et al., 2007), however require greater spatial resolution than 30-m to be effective.

Another avenue to be examined in this project is the **"abruptness" of transitions within the landscape, or boundary-based metrics (Jacquez et al., 2000).** While edge-density, above, characterizes the edges of patches, categorical land cover data such as the NLCD01 are not well suited to measuring transitions in urbanization. In the same way that slope may be calculated from a Digital Elevation Model (DEM) and provides unique information about the terrain, the "slope" of urban transitions may provide useful information. At the national scale it is possible to calculate this from the NLCD01 continuous-data impervious surface fraction image. This basic concept has been explored in other studies (Bowersox and Brown, 2001; Zhang and Wang, 2003), in which thresholds are applied to characterize the percent of the landscape that has boundary gradients above or below certain levels. Boundary-based metrics are less well examined than patch-based metrics for the purpose of urban land use identification (Brown et al., 2004), but have potential.

An important point regarding spatial pattern metrics is that nearly all may vary with and are sensitive to scale (Gustafson, 1997; Saura, 2004; Wu, 2004). This is best illustrated with an example (Figure 2-3). The red pixels in Figure 2-3 (“Class1”) as calculated from the single tract area shown have a considerably lower *perimeter-area ratio* (a measure of how dispersed a feature is) than that calculated from its three component block groups because of the overall increase in perimeter (an example of both the scale and zoning effects of the so-called Modifiable Areal Unit Problem (Fotheringham and Wong, 1991)). The implication is that those metrics which may be meaningful in characterizing pattern at one scale may not work well at another scale. How successful some commonly used spatial pattern metrics are at identifying urban land uses using national-scale data is one of the results of this project.

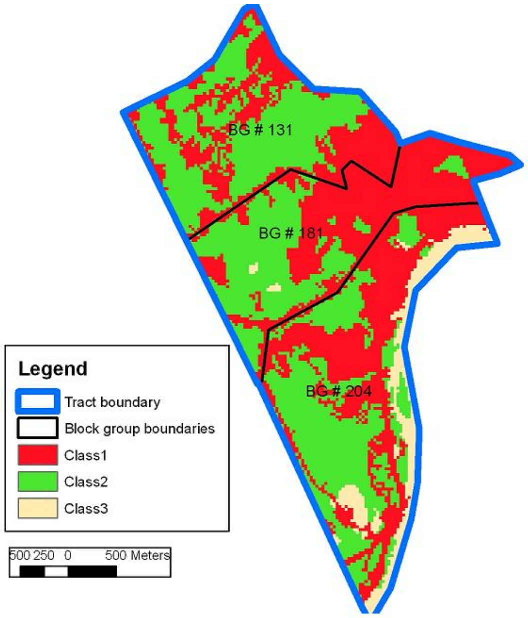


Figure 2-3: Example of how spatial metrics may change with scale.
The perimeter-area ratio for “Class1” calculated from a single tract area may differ considerably from the mean for its three component block groups.

As an aid to urban classification, **contextual measures -** here defined as proximity or spatial association to landmarks, cities, or other known features – have been used in various ways. Moeller-Jensen (1990) incorporated distance to city center as an input variable to predicting land use, and Debeir et al.

(2002) likewise incorporated distance to roads, rail, and other features. As noted above, the Landscan product (Bhaduri, 2007) incorporates proximity to road and rail as predictors of population. Wu et al., (2009a) use nine predictor variables to map land use change, six of which are proximity measures (distances to nearest city center, rail station, major road, city amenity, community service and shopping center). Wu et al. (2007) use the highest road category within 50-m to help predict parcel land use. Proximity of features within certain buffer distances is also frequently used, for example, Walsh et al. (2003) calculated the association of residential land use to aquatic amenities (lakes, wetlands, streams) at various buffer distances and found that residential areas were positively associated with lakes in the upper Midwest (and to roads, within 100-m distance). Distances are typically measured using Euclidean distance, however, it is noted that for some purposes, for example characterizing distributions of urban population, other methods of calculating distance, such as the Minkowskian distance, may be better suited (Griffith and Wong, 2007).

In summary, **spatial pattern metrics** based on imagery or data derived from imagery have been common aids in urban classification studies. We tested some of the most common ones as indicated by the literature, and supplemented those with additional metrics which have a similar basis. **Contextual measures** have been noted throughout the Literature Review section as being key elements which strongly influence the urban landscape, and while they have been employed in a number of studies, we nevertheless feel they have been an under-appreciated tool in previous identifications of land use. Given the potential importance of a land unit's location with regard to city centers, amenities, and other features as a determinant of its use, it is believed that a more thorough examination of contextual/proximity measures should be an outcome of this project.

3. Data and Methods

3.1 Study Area

The main analysis for this study was conducted for a four county area surrounding the city of Boston, Massachusetts (figure 3-1). The study area was chosen primarily because of the availability of high-quality reference data in that region. It is in this area that both 10-class and 6-class land use prediction models were derived and validated.

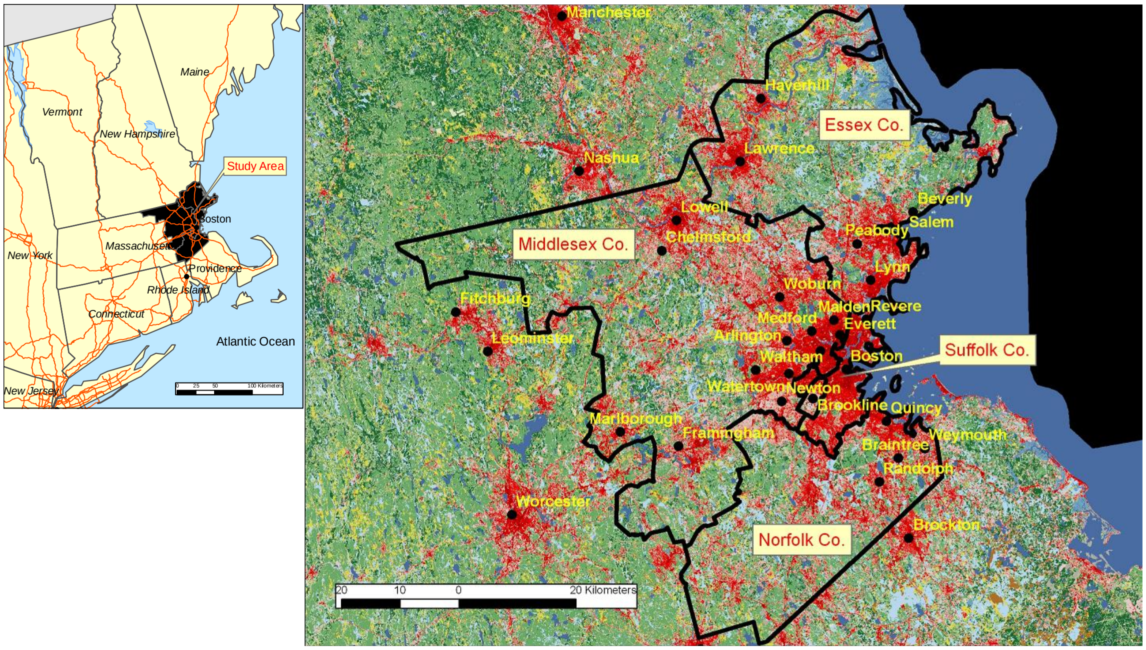


Figure 3-1: Location and 2001 land cover for the four county Boston area (USGS, 2009d). Urban land cover is shown as red and pink tones.

The 6-class model was additionally tested by data samples from three other areas: Atlanta, Los Angeles and Providence. The reference data for all areas and the locations and significance of the three external validation areas are discussed in section 3.2.1.3. The Boston study area includes four counties (Essex, Middlesex, Norfolk, and Suffolk) encompassing 2,764 block groups (BGs), with an area of approximately 4,800 km^2 (1,850 mi^2). A small portion of Norfolk County was omitted because it was discontiguous.

The four-county area includes the city of Boston, as well as outlying suburban areas and smaller cities and towns, and incorporates a total population of approximately 3.5 million people (Census 2000). Boston is essentially a monocentric city (Griffith and Wong, 2007), the center being at the Boston-Cambridge nexus (2000 population roughly 790,000), with Lowell (105,000) being the next largest population center in our study area (and Worcester – 172,000 – exerting possibly a small influence on the western part of our study area). The block groups studied include as wide a range of urban features as are likely to be found in any US metropolitan area, to include extremely dense central-city residential and commercial areas, industry, major airports and harbor facilities, commercial areas of all type, recreation, and areas of lower-density residential of a wide range. The median size of the 2,764 block groups is 0.46 km^2 (= 46 ha = 114 acres). Summary statistics for the four counties are given in Table 3-1. Note that Suffolk County (central Boston) has different land use characteristics to the other three counties, as it is more intensely urbanized (greater density of multi-family and small lot residential, more commercial, institutional, and transportation, less non-urban).

Table 3-1: Summary statistics for four counties and overall study area.
Land use data from MassGIS, 2008 (also described in more detail in next section).
Census data from 2000 Census.

LU class	Essex	Middlesex	Norfolk	Suffolk	Total
Population	723,419	1,466,847	643,047	690,445	3,523,758
Area (sq km)	1385	2191	1045	176	4,797
# block groups	543	1123	467	631	2,764
Actual land use (%):					
Single family residential, > 1/2 acre lot	10.53	13.01	12.25	0.18	11.65
Single family residential, 1/4 - 1/2 acre lot	9.65	13.43	16.50	0.70	12.54
Single family residential, < 1/4 acre lot	5.68	5.51	4.75	21.51	5.98
Multi-family residential	0.94	1.80	1.20	13.98	1.87
Commercial	1.99	2.49	2.08	10.07	2.54
Industrial	1.79	2.82	3.12	3.35	2.61
Institutional	1.15	1.72	1.60	6.03	1.69
Transportation	1.71	1.72	1.83	10.57	2.07
Recreation-Urban Open Space	3.15	3.02	3.59	9.65	3.43
Non-urban	63.42	54.48	53.07	23.96	55.63

3.2 Data

Data for this project fall into two categories: (a) reference data ("ground truth"), used for both training and validation, and (b) predictor data, used as independent variables for model building. The dependent variables to be predicted are the percents of actual land use within a block group. Throughout this document, dependent variables (DVs) are given in all upper case, such as "SFRES_L" (single-family residential, large-lot), while independent variables (IVs) are given in italics, with the first component of the name being the predictor type (described below), such as "*CENS_popden*" (a population density variable derived from census data). The terms predictor variables and independent variables are used synonymously in this dissertation.

3.2.1 Reference Data

3.2.1.1 Reference Data for Massachusetts

The ability to execute this project depends on having high-quality reference data, i.e. an authoritative and detailed mapping of “true” land uses. Approximately 20-25 regional or state-wide land use maps derived by various organizations were examined. These maps had been created from manual digitization of aerial photography. The criteria for potential use in this project were: (1) the time period must be close to the time period for the predictor data to be used in the project (e.g. census data, NLCD01), that is, the period 2000-2001, plus or minus one or two years, (2) the thematic resolution must be at least at Anderson Level II, and preferably at Level III, and (3) the data must include a large metropolitan area. Data available from the Massachusetts State Office of Geographic and Environmental Information (MassGIS, 2008) representing the 1999 time frame were selected as the best candidate. The data are a polygon-based 37-class photo-interpretation of land use for the entire state, with an MMU of 1 acre (0.4 ha) (figure 3-2).

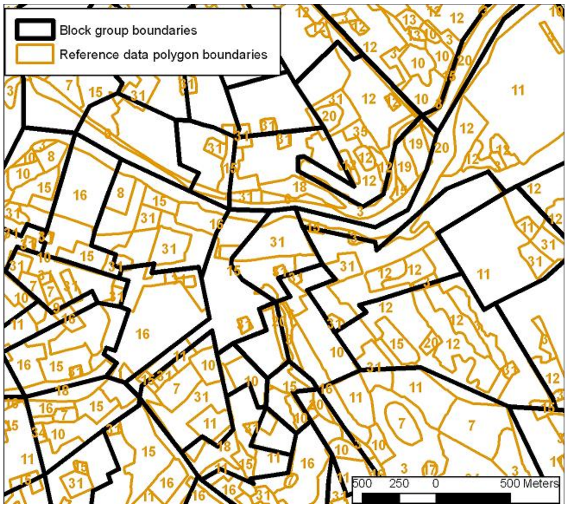


Figure 3-2. Block group boundaries displayed over State of Massachusetts reference data polygon boundaries for a portion of the city of Lowell, MA.
Polygon labels represent the “true” land use class (see Appendix A), as manually derived from aerial photography. For example, class 11 = “Residential, smaller than ¼ acre lots”. Note that block groups are nearly always comprised of multiple land uses.

The data are also noteworthy because (although not used for this project) they include mappings of two prior time periods: 1971 and 1985, which were thought to have potential for future analysis. Class descriptions for the MassGIS mappings are provided in Appendix A.

3.2.1.2 Dependent Variable Definitions and Rationale

Although the Anderson Level II classes are frequently used as a starting point and often referenced in urban land use studies, they are rarely used exactly "as is", and indeed it is difficult to find any two studies which have precisely the same definitions of urban classes at approximately that level. One rationale to deviate from the Anderson classes is that of the seven Anderson II classes two of them are mixed classes ("Industrial and Commercial Complexes" and "Mixed Urban or Built-up Land"). It is preferable to have classes which have a single unambiguous type if possible because results are more interpretable. A second reason to deviate is that there is evidence (Herold et al., 2003; Wu et al., 2007) that some Level III classes may be successfully differentiated along with Level II (e.g. "Single-family residential" and "Multi-family residential" instead of "Residential"). A third reason to deviate from the Anderson classes is that the class structure needs to be matchable to available reference data. That is, if Commercial and Industrial are merged as a single class in the ground truth it is not possible to study them separately. The classes to be examined in this study are given in Table 3-2, and compared to those used by two studies which are similar in nature to this project:

Table 3-2: Land use classes to be used in this project and those used in two similar studies. Study area for Herold was Santa Barbara, CA, and for Wu was Austin, TX.

Herold et al. (2003)	**Wu et al. (2007)**	**Falcone - 10-class**	**Falcone - 6-class**
Single-unit low density residential	Single-family residential	Single-family large lot residential	Residential, low intensity
Single-unit med. density residential	Multi-family residential	Single-family medium lot residential	Residential, high intensity
Single-unit high density residential	Commercial	Single-family small lot residential	Commercial/ industrial/ institutional
Multi-unit residential	Office	Multi-family residential	Transportation
Commercial & industrial	Industrial	Commercial	Recreation & Open Space
Institution	Civic (institution)	Industrial	Non-urban*
Recreation & Open Space	Transportation	Institutional	
Agriculture & rangeland*	Recreation & Open Space	Transportation	
Forest & wetlands*	Undeveloped*	Recreation & Open Space	
		Non-urban*	

* = non-urban classes

Both the Herold and Wu studies showed that there was good separability among their classes based on the attributes studied (spatial pattern metrics of land cover and building height/shape/areas, respectively), and initial tests of prototype data in this project indicated that, even though some classes are closer to each other than others, there was both hypothetical and empirical evidence that they can potentially be distinguished (also see Section 3.3.1).

The 10-class structure was used for identifying land use in a single region only, the Boston area, because reference data at that thematic resolution are not consistently available elsewhere for the correct time period. The 6-class scheme was used for validating and testing the method in both Boston and in Atlanta, Los Angeles, and Providence, because consistent validation data are available for the 6-class level in those regions.

The following are text descriptions and abbreviations of the ten classes used in this project. Figure 3-3 gives examples of the spatial configuration of representative types for the nine urban classes (Non-urban not shown).

- **Single-family residential, large lot (SFRES_L):** Detached residential homes on lots > ½ acre. May include rural residential, including farms. *Generally low population density, large patches of interspersed vegetation.*
- **Single-family residential, medium lot (SFRES_M):** Detached residential homes on lots between ¼ and ½ acre. *Generally medium population density; interspersed vegetation somewhat less; medium road density.*
- **Single-family residential, small lot (SFRES_S):** Detached residential homes on lots of < ¼ acre. *Generally medium-high population density; urban surfaces often have regular patterns and smaller patches of interspersed vegetation; high road density.*
- **Multi-family residential (MFRES):** Residential units which are attached. Includes duplexes, condominiums, low and high-rise apartments, and attached mobile homes. *High population density; medium-high road density; may have sizeable patches of interspersed vegetation.*
- **Commercial (COMMERC):** Office buildings, retail stores, commercial centers, strip shopping centers, malls, hotels, and motels. *Low-medium population density; may have highly concentrated areas of imperviousness or in linear configurations.*

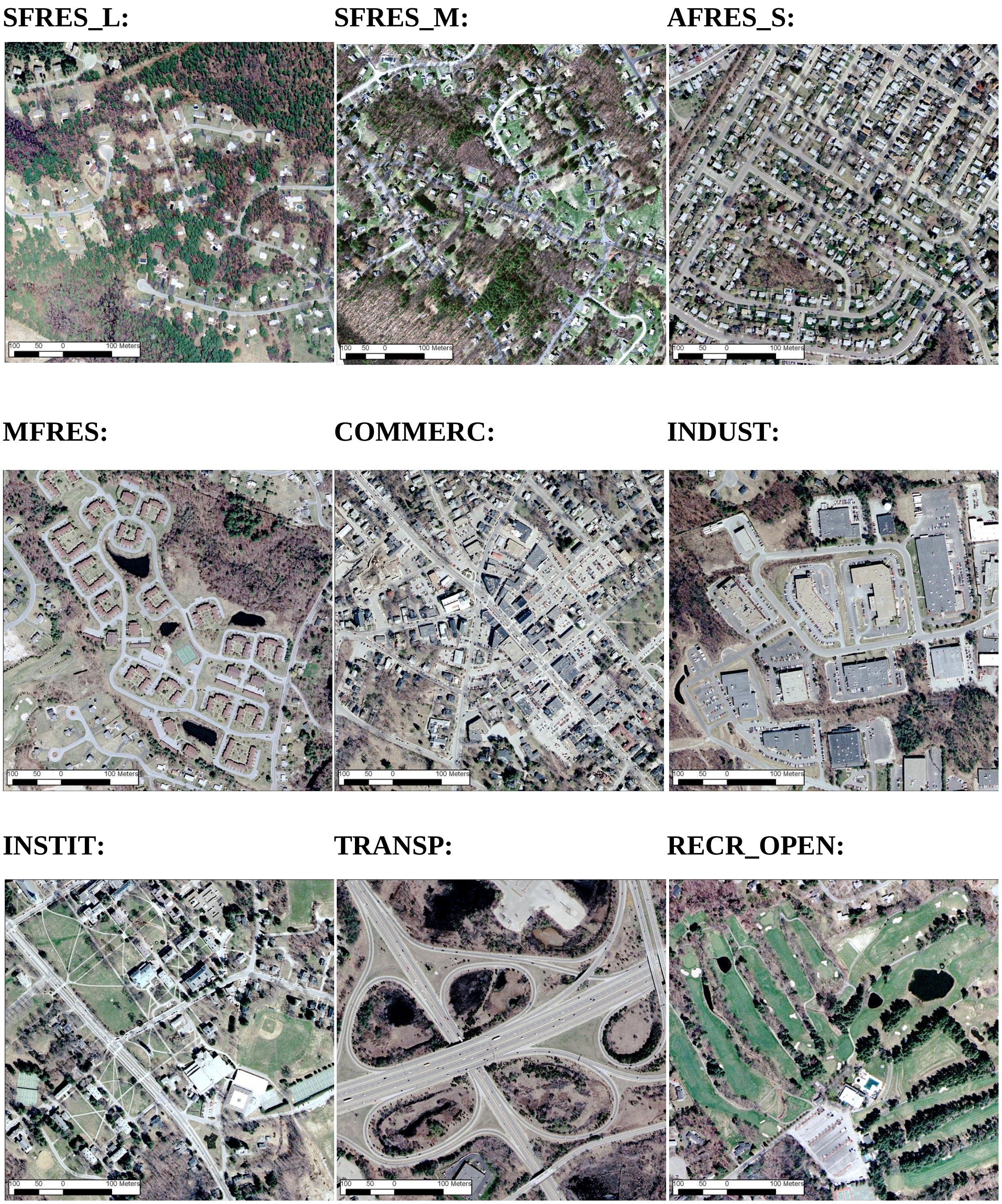
- **Industrial (INDUST):** Light and heavy industry, manufacturing. *Low population density; large structures, often clustered; concentrated imperviousness; low road density.*
- **Institution (INSTIT):** Public facilities, government offices, police and fire stations, hospitals, nursing homes, churches, schools, universities, libraries, prisons, and military bases. *Generally low population density; may have large structures with attendant large patches of vegetation.*
- **Transportation (TRANSP):** Airports, freeways, railways, bus and truck terminals, harbor facilities, major utility and communication facilities. Does not include road systems smaller than freeways, which are considered to be an integral part of other land uses. *Low population density; either very large areas of concentrated imperviousness or linear features.*
- **Recreation and open space (RECR_OPEN):** Parks, golf courses, recreational areas, sport fields, cemeteries, and other open urban space. *Low population density; generally large areas of green vegetation with interspersed roads and structures.*
- **Non-urban (NON_URB)**: All other land not included above. *Very low population density; largely vegetated; very low imperviousness.*

The 6-class categorization is simply a subset of the 10-class, in which 3 classes are merged classes. The following shows the aggregation and abbreviations for the 6-

class names (TRANSP, RECR_OPEN, and NON_URB do not change and are used in both schemes):

SFRES_M + SFRES_L	= RESID_LOW_6CL (low intensity residential)
SFRES_S + MFRES	= RESID_HIGH_6CL (high intensity residential)
COMMERC + INDUST + INSTIT	= COM_IND_INST_6CL (commerc./indust./instit.)
TRANSP	= TRANSP (transportation)
RECR_OPEN	= RECR_OPEN (recreation-open space)
NON_URB	= NON_URB (non-urban)

The above classes (both 10 and 6-class schemes) represent the dependent variables to be used in this project.



SFRES_L: SFRES_M: AFRES_S:

MFRES: COMMERC: INDUST:

INSTIT: TRANSP: RECR_OPEN:

Figure 3-3: Example of spatial configuration of nine urban land use types for the Boston area (“Non-urban” not shown).
Images are 0.5-m orthoimages dated 2001 (USGS Seamless Server).

3.2.1.3 Reference Data for Other Urban Areas

One of the goals of this project is to test how well models created for the Boston area perform if tested against data samples from other urban areas, and vice-versa. During the search for regional land use products it was noted that nearly every one, although loosely based on the Anderson scheme, had different class structures which in some cases made them incompatible, depending on the desired aggregation. For example, some products did not distinguish commercial and industrial land or had difficult-to-compare breakpoints of residential lot size. Others had subtle differences that were apparent only after visual examination, for example in some cases the product "burned in" all roads as part of a transportation class (as opposed to the more common delineation of only major highways). Additionally some were from time periods other than the desired circa 2000 window. Because it was not possible to locate multiple reference datasets external to Massachusetts which could be confidently recoded to the 10-class structure, three datasets were selected which were suitable for the simplified 6-class structure (figure 3-4). These were for Atlanta, GA (data for 13 counties available, Atlanta Regional Commission, 2008), Los Angeles, CA (6 counties, Southern California Association of Governments, 2008), and Providence, RI (5 counties, Rhode Island Geographic Information System,

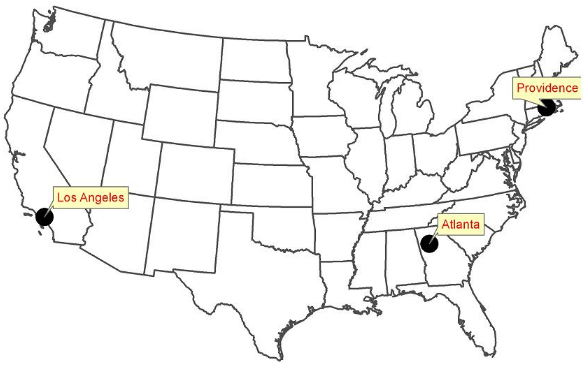


Figure 3-4. Location of three study areas external to Boston where models were tested.

2008). The five counties in Rhode Island encompass the entire state. The product time periods were: Atlanta, 2001; Los Angeles, 2001; and Providence, 2003-4. As with the Boston reference data, the products were all manually interpreted and polygon-based in format. The Los Angeles data were supplemented with county parcel data to more accurately distinguish lot sizes. It was believed these external reference data were as similar as was possible to acquire, given the desired criteria, and represented a variety of urban settings, which included different physical characteristics of background vegetation and terrain.

More detail about the data and methods for the three external areas is provided in Section 3.3.8.

3.2.2 Predictor Data

No previous study has attempted a comprehensive examination of urban land use mapping with national-scale data. The broad aspects of determinants of land use have been discussed in Section 2, and while there is some guidance from literature as to which of these determinants are likely to be important, their specific form (how specific variables should be calculated or derived) is much less clear, particularly from national-scale data. We theorized that sources of commonly-available data (e.g. census data or landmark point locations) or metrics derived from national-scale data (e.g. the rate of change of imperviousness over the landscape, or the degree of spatial autocorrelation of certain features) had a basis to serve as predictors of different types of urban land use. To that end a large number of predictor variables were assembled or derived to predict the 10 different dependent variables. It should be noted here however, that after reducing for multicollinearity not all were necessarily tested as predictors and only a small number were actually used in final models, as is described in sections 3.3.3 and 3.3.4. These independent variables are broken into 10 categories, which broadly match categories of land use determinants and indicators discussed in Section 2: **Census, Landcover, Historical LULC, Transportation, Landmarks, Proximity, Spatial Autocorrelation, Spatial Pattern-categorical data, Spatial Pattern-continuous data,** and **Miscellaneous**.

All data are national, although in some cases are available only for the lower 48 states. The variables for eight of the ten categories (**Census, Landcover, Historical LULC, Transportation, Landmarks, Spatial Pattern-categorical data, Spatial Pattern-continuous data,** and **Miscellaneous**) are calculated based on data contained within a block group. For example, population density (*CENS_popden* in **Census**) is calculated as the number of persons/sq km for the block group. The other two categories (**Proximity** and **Spatial Autocorrelation**) are calculated based on data/relationships that may extend outside the block group (but are summarized for the block group). For example, distance to nearest 250k city (*PROX_city250k_dist* in **Proximity**) is calculated as the distance of the block group centroid to the point location of the nearest city of population > 250,000. The following presents a brief description of the categories and rationale for their use. The names, descriptions, and units for all variables are given in Appendices B1 – B10, however in some cases additional explanation for individual variables is given in the following text, where merited.

1. **Census**. Population density characteristics are clearly related to land use, however several other socio-economic characteristics have intuitive or theoretical relationships which have yet to be examined. For example, household income, median number of rooms in home, or percent of population who take public transport to work may all plausibly be related to land use types (e.g. high percentage of population using public transport positively correlates to multi-family residential housing). One of the advantages of the block-group scale (as opposed to the block scale) is the greater availability of census information. For block groups and coarser geographies, the Census "Summary File 3" data are available, which are detailed information derived from the Census long-form (U.S. Census Bureau, 2009b), then summarized for the geography. Data are publicly available via the Census American Fact Finder (U.S. Census Bureau, 2009c), and include information about population and housing. All information available from the Census for the block-group scale was examined, and a reduced set of 20 variables was created, also based on preliminary testing. This list is given in Appendix B1.
2. **Landcover**: Data from the National Land Cover Data 2001 dataset (NLCD01) are the most current national mapping of US-wide land cover, however, as noted earlier, do not include urban land use classes, but rather gradations of imperviousness. All land cover classes from the NLCD01 (USGS, 2009d), to include some aggregated classes (e.g. "all natural vegetation") were included, as well as derived classes which were hypothesized to have value in predicting

certain classes. For example, low values of the ratio of housing unit density to % imperviousness (*LC_ratio_huden_imperv* in Appendix B2) were theorized to be a good predictor of industrial or commercial land uses, because those LU types typically have low housing density but high levels of imperviousness (Bauer and Steinnocher, 2001). These variables included values for the NLCD01 categorical data as well as the NLCD01 continuous impervious surface data. Variables are listed in Appendix B2.

3. **Historical LULC**: Prior land use information, if available, is likely to be indicative of current land use and has been used in other research to model current land use (Vogelmann et al., 1998). As noted previously, even though the NLCD92 and GIRAS are older datasets of differing format and neither maps completely to the class structure proposed here, they are likely to be valuable to the degree that they were a) originally accurate, and that b) if originally accurate, that land use has not changed since their creation. If both of those conditions are fulfilled it is hypothesized that certain classes of one or both of those historical datasets will be good predictors of current land use: for example the NLCD92 class "Urban/recreational grasses" (*HIST_nlcd92_85* in Appendix B3) is potentially an excellent predictor of the RECR_OPEN dependent variable. A number of variables in this category are indices of land cover over several time periods. For example, *HIST_indust_all_times* is an estimate of industrial land based on the sum of classes which are the most similar to industrial from all three

available time periods (*LC_nlcd01_23*, *LC_nlcd01_24*, *HIST_nlcd92_23*, *HIST_giras12*). Variables are listed in Appendix B3.

4. **Transportation**: This category contains metrics regarding roads, railroads, port facilities, and airports, from the Bureau of Transportation Statistics (BTS, 2009), Census (U.S. Census Bureau, 2009d) and ESRI (ESRI, 2009a). Roads particularly are a defining feature of urban environments (Falcone et al., 2007) that may be useful in distinguishing land use types and have been used in numerous land use studies (Schumacher et al., 2000; Sun et al., 2007; Wu et al., 2007). Several variables in this category are likewise ratios (e.g. *TRANSP_ratio_roadden_imperv*) or an index derived from combining transportation data elements (*TRANSP_alltrans*). Variables are listed in Appendix B4.
5. **Landmarks**: The most comprehensive source for landmark point locations is the U.S. Geological Survey's Geographic Names Information System (GNIS; GNIS, 2009). The USGS maintains this publicly-available registry which consists of approximately two million features considered to be significant enough to have an "official" name (GNIS, 2009). These include locations of hospitals, school, churches, post offices, cemeteries, airports, parks, government buildings, prisons, and more. The GNIS database is continually being updated. Figure 3-5 and Table 3-3 give visual and tabular examples of GNIS point locations in the local Fairfax area.

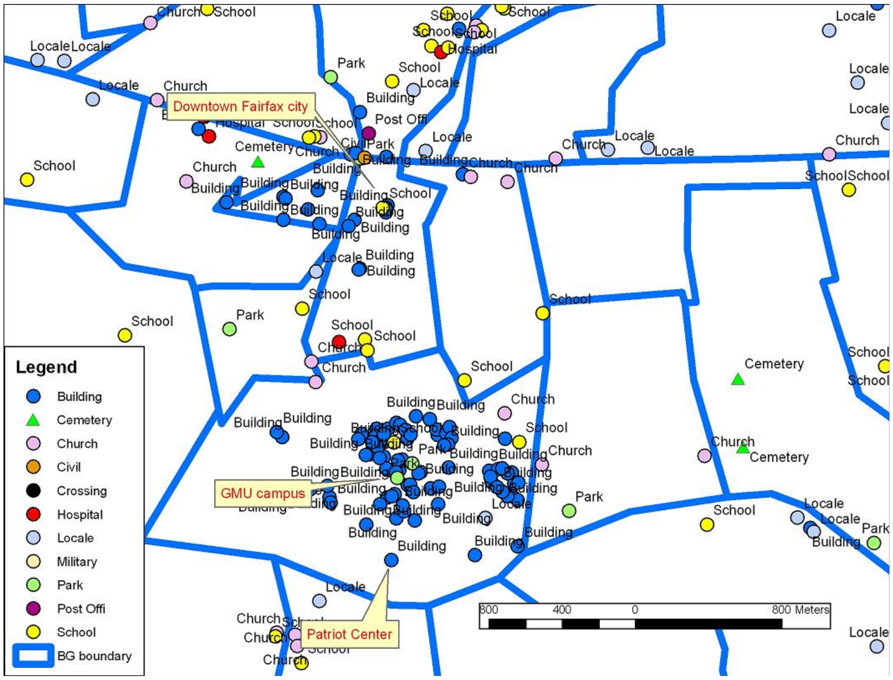


Figure 3-5: GNIS example:
GNIS features for the George Mason University area (same area shown as in Figure 1-5).

Table 3-3: GNIS features as they pertain to this project, example of how many exist in Fairfax County, VA (as reference to a familiar area), and their potential use as predictor

GNIS Feature Class	Number in Fairfax County	Likely useful predictor for:
Airports	9	TRANSP
Buildings	205	Most are "institutional": INSTIT
Cemeteries	39	RECR_OPEN
Churches	410	INSTIT
Civil	13	INSTIT
Crossings	56	Mostly major freeways: TRANSP
Hospitals	19	INSTIT
Locales	293	Variety of features, but mostly shopping centers and industrial parks.
Military	3	INSTIT
Parks	264	RECR_OPEN
Post Offices	34	INSTIT
Resort	20	RECR_OPEN
Schools	396	INSTIT

It was anticipated that the GNIS point locations would primarily be helpful in mapping the INSTIT and RECR_OPEN classes. An uncertainty with the GNIS points is their consistency across the country. Because the database is in part updated by local cooperators they are not necessarily consistent from state to state, and the consistency may additionally vary within specific layers, e.g. point locations for cemeteries may be very current and accurate for specific states or regions, but less so for other areas (Roger Payne, USGS, personal communication, April 24, 2009). However, the data layers that were most useful in this project and were most common – schools, hospitals, public buildings, parks, cemeteries, golf courses, and a few others – seem to be fairly consistent from area to area.

The GNIS points were first sub-categorized to the extent that sub-groupings were discernible from the feature name (the only descriptive information in the GNIS dataset). For example the GNIS points as downloaded contain a single feature class "School" to identify all schools. Recognizing that schools may have tremendous variability in their areal extent (the characteristic being predicted in this project), e.g. high schools vs. elementary schools, several feature classes were broken out into sub-categories, to the degree that that was possible. These derived classes were then examined against known landmarks to determine the most effective way of quantifying their presence for the purpose of estimating their areal extent. A straight density of point locations was used for some predictors (see Appendix B5). Two consolidated predictor variables were constructed from a weighted average for points believed to be good predictors for "recreation" and "institutions", respectively. The weighting was based on a very simple scheme drawn from our visual observations (each value represents the number of point locations of that type):

LANDMRK_gnisconsol_recr_density = ((resorts * 2) + parks + cemeteries)/area

LANDMRK_gnisconsol_inst_density = ((buildings / 2) + churches + correctional + (hospitals * 2) + (military * 2) + post office + school-elementary + school-middle + (school-highschool * 1.5) + (school-univ * 1.5) + school-other)/area

Because it was not know *a priori* if a gridded representation of GNIS points or a calculated point density (above) would be more effective, two other representations for GNIS points as grids were built by "growing" point locations as grid points according to their perceived areal extent (the variables *LANDMRK_gnis_recr_grid* and *LANDMRK_gnis_inst_grid*). Variables are listed in Appendix B5.

6. **Proximity**: Urban land use is formed to a large degree based on the accessibility of land with relation to the central city or other features and amenities, as noted in Section 2. This has been expressed in a classification process by the terms "context" or "proximity" of pixels or land parcels to other known features

(Moeller-Jensen, 1990; Debeir et al, 2002; Bhaduri, 2007). In this project a number of predictor variables were derived which attempted to quantify a block group's relationship to other features likely to be important: roads, city centers, large contiguous urban patches, and GNIS point locations. (Another predictor of a land use type is likely to be other known land uses, however because actual land uses are unknown – they are the dependent variables to be predicted – they are not available as independent variables). The following describes the general derivation of the Proximity category variables (all based on Euclidean distance except for the "cost distance" variables):

- Proximity to roads or airport: mean pixel distance to nearest road (*PROX_mean_dist_road*), and distance of block group centroid to nearest freeway (*PROX_interstate_road_dist*), nearest primary road (*PROX_prim_road_dist*), any major road (*PROX_major_road_dist*), or airport/interstate crossing (*PROC_airport_crossing_dist*).
- Proximity to nearest large (> 2 ha) contiguous urban patch (*PROX_patch_2ha*) (also see Spatial Pattern-categorical data section below for more detail on derivation of large contiguous patches).
- Percent of land in large contiguous patches within 120 m and 240 m of road (*PROX_expand4rds_inters_patchgr2ha* and *PROX_expand8rds_inters_patchgr2ha*). Percent of land in NLCD01 classes 23 and 24 (high intensity urban) within 120 m and 240 m of road (*PROX_expand4rds_inters_2324* and *PROX_expand8rds_inters_2324*).

- Proximity to cities of various sizes: distance of block group centroid to point location of nearest 10k, 20k, 50k, 100k, and 250k city (*PROX_city10k_dist*, *PROX_city20k_dist*, *PROX_city50k_dist, PROX_city100k_dist, PROX_city250k_dist*).
- Proximity to GNIS points: distance of block group centroid to nearest GNIS recreation, institution, and commercial/industrial point locations (*PROX_allrec_gnis*, *PROX_allinst_gnis, PROX_allcomind_gnis*).
- "Cost distance" of any pixel in the landscape to one of six different features. Six cost surface grids were created by calculating every pixel's weighted distance to a) the nearest 10k city, b) the nearest 50k city, c) the nearest 100k city, d) the nearest large urban patch, e) the nearest GNIS institution location (schools, hospitals, etc.), and f) the nearest GNIS recreation location (parks, cemeteries, etc.). The weighting was based on the pathway to the feature via the nearest major road (lowest cost), nearest minor road (second lowest cost) and nearest urban pixel (third lowest cost). It was hypothesized that the mean cost distance for a block group to a particular feature type would be an effective unique metric for measuring a block group's urban "spatial identity": it measures how urban the block group is, how far it is from the feature, and its accessibility to the roads network. Figure 3-6 gives visual examples of the two of the six cost surfaces.

All Proximity variables are listed in Appendix B6.

7. **Spatial Autocorrelation**: A nearly self-evident feature of urban landscapes is that some elements of the landscape may exhibit clustering, and that clustering may be an identifying characteristic of LU types. Griffith and Wong (2007) note that introducing a spatial autoregressive term improves population modeling, and we theorize that predictors based on the same concept would likewise improve decision tree models. While the contiguousness, cohesion, patch density and other spatial pattern metrics of land cover

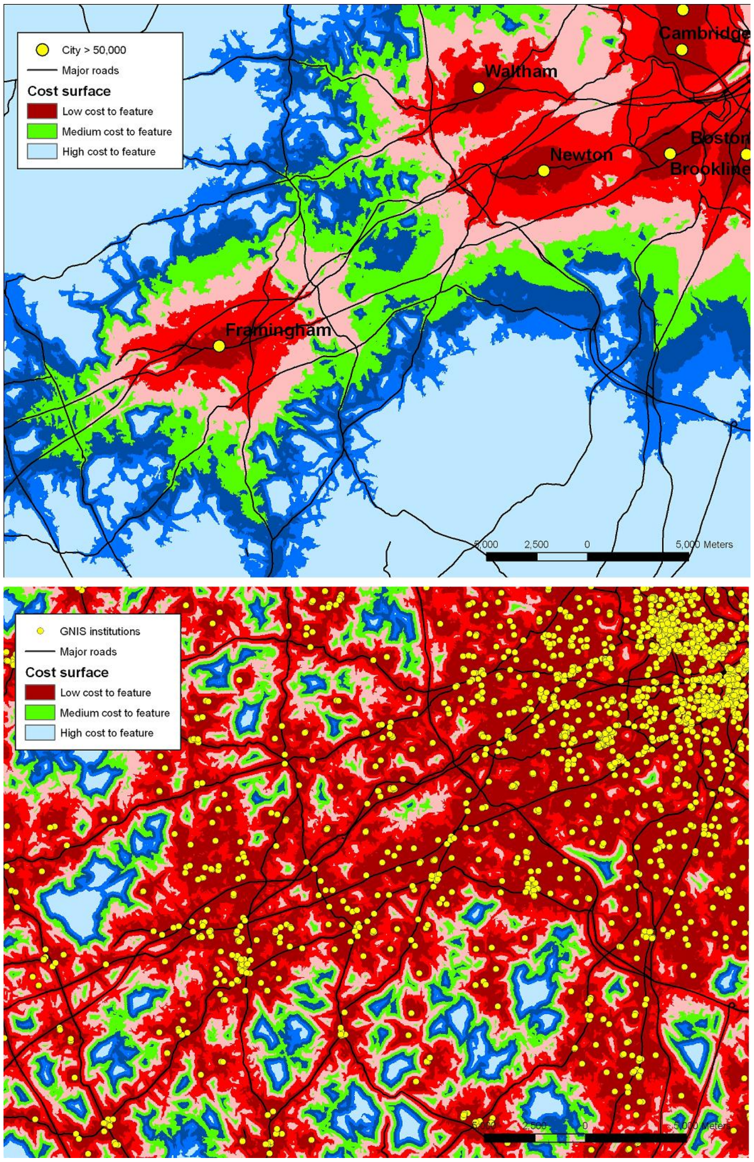


Figure 3-6: Example of two of six cost surfaces:
cost surface to nearest city > 50,000 population (top) and cost surface to nearest GNIS institution point location (bottom), showing the same area in Boston's southwest. The cost surface is a continuous grid which measures a distance to the nearest feature based on a weighted pathway of major roads, minor roads, and urban land cover. Lowest costs therefore tend to be oriented particularly along major roads.

are examined under the category Spatial Pattern-categorical data below, this separate Spatial Autocorrelation (SA) category attempts to capture measures of clustering or similarity that span block group boundaries. These are as follows:

- The Local Moran I statistic was calculated for a number of landscape features whose clustering generally mapped to land uses, based on our visual observation. Using measures of spatial autocorrelation as predictors or input variables to classification is not uncommon in the mapping (e.g. Su et al., 2008) or other sciences (Bell et al., 2007), although typically in land cover/land use mapping they are based on image data. As example, figure 3-7 (left panel) shows a mapping of one of these variables – the census variable median number of rooms per household (*CENS_hu_median_numb_rooms*). There is a clear tendency for households to have fewer rooms in central Boston, and more rooms in the outer suburbs. The right panel on figure 3-7 shows the local Moran z-score (standard deviations), based on inverse distance across the entire landscape. High positive z-scores (if > 1.96, significant at 0.05 level) shown in blue have significant clustering with like values. Low negative z-scores shown in red are significant outliers: they have either high values near low values or vice versa. It is noteworthy that low negative z-scores here generally map to urban transition zones. The z-scores for these variables (e.g. *SA_localMoran_medianrooms*) therefore had the potential to provide additional information about the landscape not evident from simple block group means, and were additionally amenable to the decision tree approach to be used (section 3.3.2). That is, the decision tree would be able to partition high z-scores which were in highly urban areas vs. high z-scores which were in highly non-urban areas, based on splits created by other variables. Detail on decision tree mechanisms are given in 3.3.2.

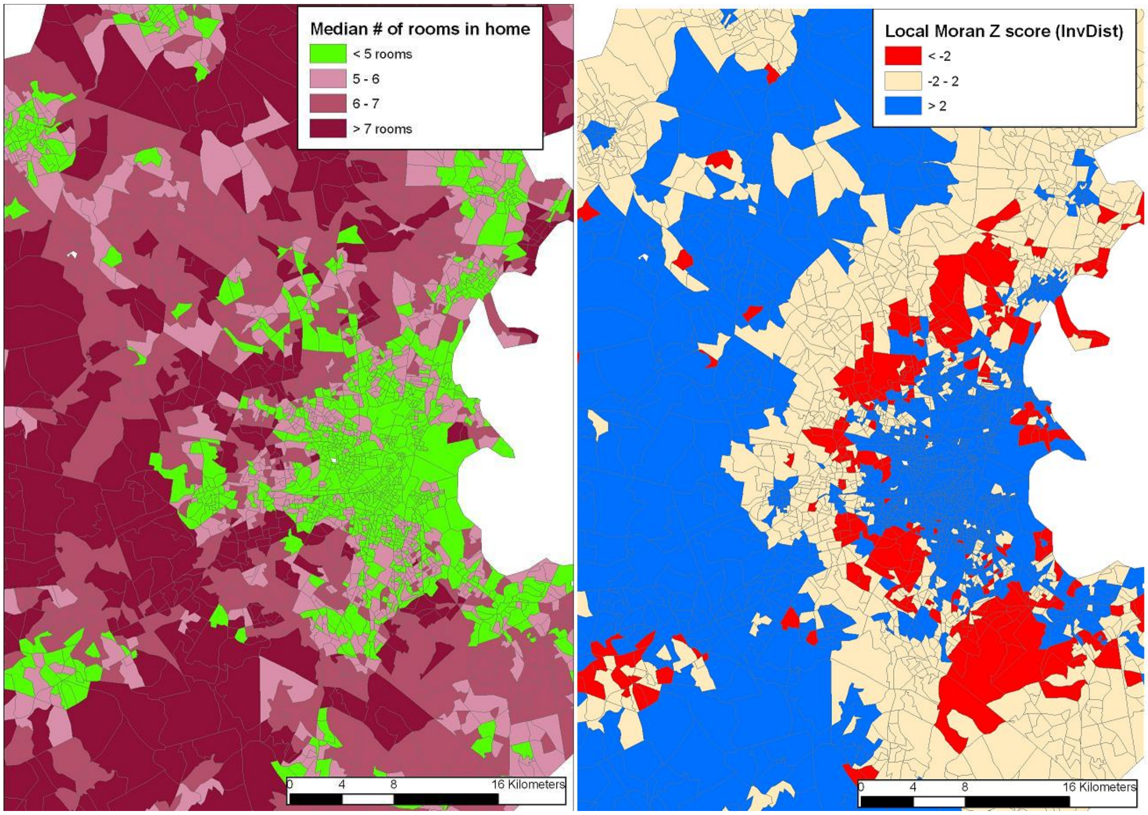


Figure 3-7: Example of calculation of local Moran metrics to be used as predictor: median number of rooms mapped by BG (left) and local Moran z-score (right). Inverse Distance method used. Blue areas exhibit significant clustering of similar values (p = 0.05) and red areas significant dissimilarities. Red areas map well to transition zones.

- The percentage of urban land cover in 400m, 800m, 1200m, and 1600m buffers surrounding the boundary of each BG was calculated, and compared to the percentage in the BG itself. McKnight (2001) notes the common trend for large industrial or commercial areas to be set in undeveloped settings, and this method was an alternative method for estimating how similar urbanization was in the BG to its surrounding areas, and thus to potentially capture those areas. The variables created from this were the difference between the BG's percent urban and each of the buffers and a calculation of the auto-correlation function across the four zones.

Variables are listed in Appendix B7.

8. **Spatial Pattern-categorical data**: As noted earlier, spatial pattern metrics – the shape, size, and configuration of the landscape – have been used as indicators of thematically-detailed land use type in a number of studies, however have been little investigated for that purpose based on relatively coarse national-scale data (e.g. 30-m data). To that end we tested two categories of spatial pattern predictors – one based on the categorical NLCD01 land cover data, and one based on the continuous NLCD01 imperviousness data (next section). The categorical-based predictors are broken out as follows:

    - We used the publicly available and commonly used FRAGSTATS package (McGarigal et al., 2002) to calculate a suite of pattern metrics based on two classes of the NLCD01 data: an aggregated class consisting of the sum of NLCD01 classes 22, 23, and 24 (urban land with imperviousness > 20%), and NLCD01 class 21 (urban land with imperviousness < 20%). Although many hundreds of FRAGSTATS variables are available for calculation, we were guided primarily by Herold et al. (2003) and selected a subset of 17 metrics which we believed would be useful for each of these classes. These included metrics such as patch density, edge density, and perimeter-area ratio, among others (Appendix B8). These metrics are well-established in the literature (Bowersox and Brown, 2001; Saura and Martinex-Millan, 2001; Herold, 2003; Salas et al., 2003; Saura, 2004).

- A grid was created of all contiguous high-imperviousness urban land (classes 23 + 24) which existed in patches greater than 2 ha (about 5 acres). The mean area consisting of these patches was calculated for each BG.
- We used the fragmentation metrics proposed by Riitters et al. (2000) to further characterize fragmentation in the individual urban classes as well as aggregated natural vegetation. The metrics Interior, Transitional, and Edge were calculated for each of those classes using Arc Grid.
- Land cover variety for each block group was calculated based on (a) the Shannon Diversity Index (Entropy) (Odum, 1971), from Anderson Level I classes (e.g. “urban”, “forest, etc.), and (b) the simple variety (# of unique values) from Anderson Level II classes. Land cover variety at both levels was typically inversely related to urbanization in this region at the BG scale.

- Herold et al. (2003) and others have noted that "institutions" (schools, churches, hospitals, etc.) are often large spatially distinct structures surrounded by vegetation. Capabilities in Arc Grid for annulus processing (an annulus defined as the area between two concentric circles) provided one possibility to capture these settings, although it was again unknown if the spatial resolution of the land cover data would be detailed enough for it to be successful. We developed a process to identify all urban pixels (NLCD01 classes 22, 23, 24) which were primarily surrounded by vegetation.
- Measures of shape and compactness have been shown to be strongly related to urban land uses when based on high-resolution data (Barr et al., 2004; Mesev, 2005). A number of shape metrics were calculated: a shape index (area/perim^2) for aggregated pixels in NLCD01 classes 22-24, and for class 21 (higher values indicative of more compact shape), and the ratio of the classes 22-24 ellipse semi-major axis to semi-minor axis (higher values indicative of greater linearity).

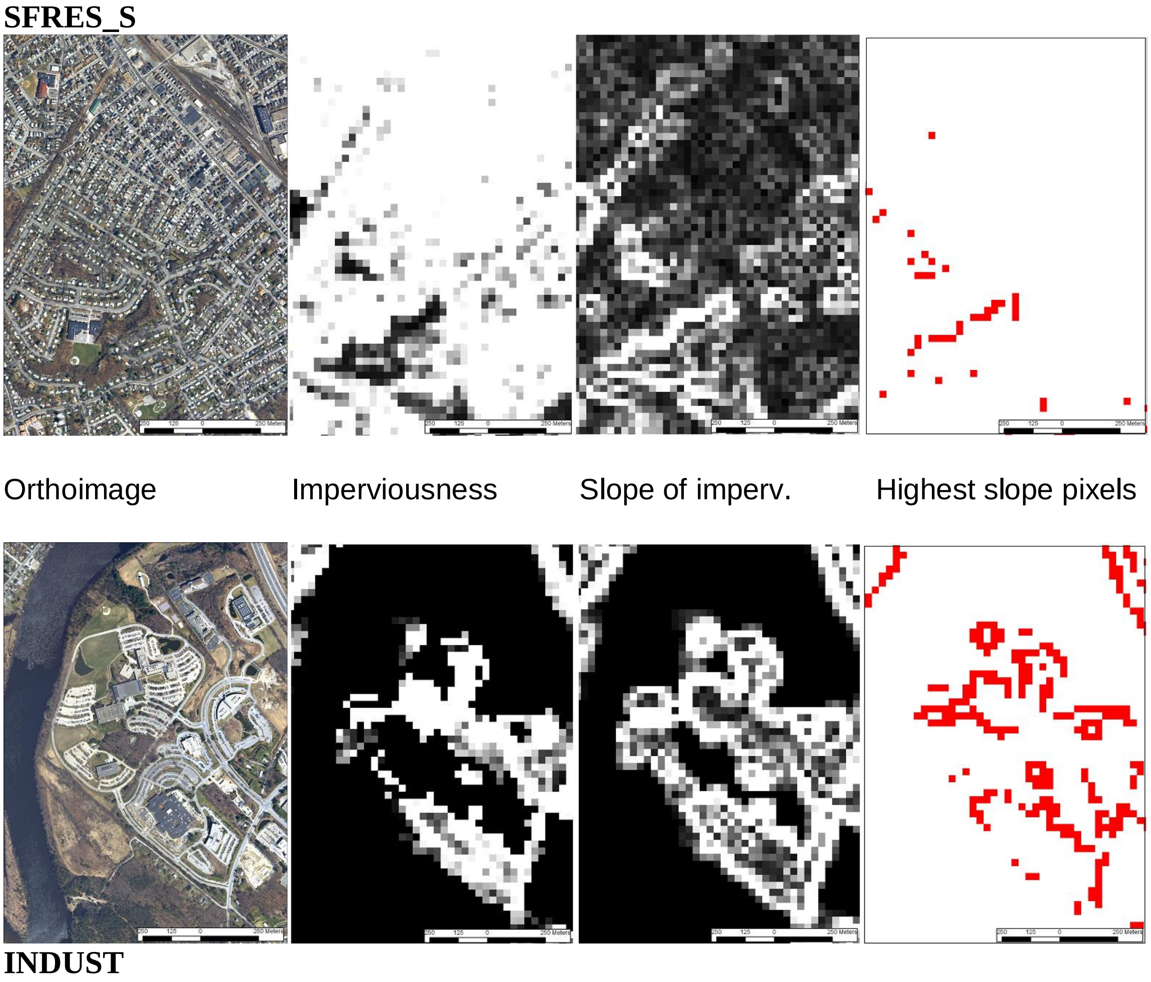
Variables are listed in Appendix B8.

9. **Spatial Pattern-continuous data**: As noted earlier, the literature is rich with examples of the use of spatial pattern metrics in analysis based on categorical data. To some degree this is possibly related to the availability of well-established tools for categorical data (e.g. Fragstats), and that categorical data may be somewhat simpler conceptually. One of the intents of this project is to

also explore the value of using the NLCD01 impervious surface grid as a basis for metrics. The impervious surface grid has integer values scaled from 0-100 (so strictly might not be a “continuous” grid), but as such may be treated as such. Measures of “slope” in the impervious surface grid may be meaningful to the extent that they represent gradual or abrupt changes in the landscape, and that boundary-based metrics based on continuous data may be at least as effective as patch-based metrics based on categorical data for characterizing landscapes in some regards (Jacquez et al., 2000; Bowersox and Brown, 2001; Brown et al., 2004). The use of these metrics in this project is admittedly exploratory, however are based on perceived differences in not only the intensity but also the pattern of imperviousness and rate of change of imperviousness between different land use types (figure 3-8).

Several measures of spatial autocorrelation for the imperviousness grid were calculated using two somewhat different methods: The first was by executing the Grid ‘Moran’ function against all pixels with imperviousness > 50% (because there was greater separation among land uses for high imperviousness than for all imperviousness pixels). The Moran function as implemented in Arc measures only the spatial autocorrelation against immediate pixel neighbors, which limits its usefulness, however it was still a potential measure of spatial pattern of the imperviousness grid. The second method was to calculate the focal mean in a 3x3 window for all pixels with values > 50, then re-executed for a 7x7 window. BGs in which there was little difference between the two values had typically clustered imperviousness.

Variables are listed in Appendix B9.



SFRES_S

Orthoimage Imperviousness Slope of imperv. Highest slope pixels

INDUST

Figure 3-8: Example differences in imperviousness pattern by land use setting.
Top four panels show 4 representations of a high-density single family residential area. Left-most panel is 0.5-m orthoimage (2005). The second panel shows the same area as represented by the NLCD01 imperviousness data (30-m pixels): bright pixels = high imperviousness. The third panel shows imperviousness slope (mean rate of change from each pixel to neighbors): bright pixels = high slope. Right-most panel shows highest slope pixels (slope_class4, described below).
Bottom four panels show same sequence for an industrial park. Note differences in intensity and pattern of imperviousness slope between the two landscape types.

The NLCD01 imperviousness grid was converted to a percent slope grid (Grid 'Slope' function), where each pixel represents the mean rate of change of each pixel to its eight nearest neighbors. Statistics from the slope grid were calculated (e.g. *SPCON_is_slope_max* = maximum imperviousness slope in the BG). The slope grid was visually examined for break points that distinguished land use types (as in Zhang and Wang, 2003) and broken into four categories (slope_class1 – 4), with slope_class1 containing the lowest slopes and slope_class4 the highest slopes. The mean and pattern metrics for each of the four slope classes were then calculated. For consistency, the same suite of Fragstats pattern metrics that were calculated for the categorical land cover classes were calculated for the imperviousness slope classes.

10. **Miscellaneous**: The Miscellaneous category consists of six variables which do not readily fall into the above groups. They were:

    Topography: area in sq km of the block group, mean elevation, and mean (topographic) slope from 30-m elevation data. Measures of topography have been shown to be related to land uses (e.g. Smith and Fuller, 2001).

    Average annual vegetation: the USGS produces a series of 1-km resolution grids based on AVHRR NDVI values which give a measure of the average green vegetation growth for a particular year (USGS, 2008). The grid for the year 2000 was downloaded and the average value per block group calculated. Because most block groups are smaller than 1 km^2 in size, this serves as only a coarse measure of vegetation, but it is useful because it is available for individual years (unlike the NLCD, which has typically been accomplished every 10 years), and therefore could be targeted by individual year in research that might spin off from this project.

Protected areas: Percent "protected lands" (e.g. National or State Parks, Wilderness Areas, etc.) as maintained in a national coverage by the Conservation Biology Institute (2009) was calculated. Protected lands have a clear relation to urban land use in that certain land use types are restricted from those areas.

Maritime: binary value: whether the BG is adjacent to the ocean, or ocean-access via a bay or inlet, or has no direct ocean access (also based on Smith and Fuller, 2001).

Variables are listed in Appendix B10.

3.2.3 Proprietary Data as Predictors

One of the objectives of this project is to demonstrate a method by which urban land use may be mapped given publicly-available national scale data. We are less keen about demonstrating methods which could only be duplicated by purchasing expensive proprietary data or software. Nevertheless, we felt it was informative to at least tangentially examine the effectiveness of other, proprietary data sources, if available. To that end this project examined one of these: ESRI Business Analyst data (although proprietary data will not be built into final models developed). The GMU GGS department has kindly provided a sample from the ESRI Business Analyst package (ESRI, 2009b) for the state of Massachusetts. The ESRI Business Analyst is a software package, which, among other features, contains a national point location database of approximately 12 million U.S. businesses, and the location of some 4,000 major shopping centers/malls. The data have certain re-use restrictions and contain fairly detailed information about each location, to include business name, industry classification code (NAICS code; NAICS Association, 2009), sales, number of employees, and sq footage. The Business Analyst ArcGIS extension is available for purchase as of February 2009 at the commercial rate of $18,000 for the entire U.S. (personal communication, ESRI sales rep, Feb 26, 2009), but less for specific regions or states (e.g. single state purchase = $8,500).

The Business Analyst data are primarily useful for predicting the commercial or industrial classes, and may also be a better source of information for institutions (schools, churches, etc.) than the GNIS data. The Business Analyst data were examined against known locations, and based on that, several predictor variables representing point density, number of employees, and distance to block group centroid were derived as independent variables. These were based primarily on a "cleaned" set of point locations (eliminating probable home-based businesses and other locations not useful for this application) of commercial/industrial and institution locations. These are described in Appendix B11.

3.3 Data Evaluation/Model Building

The following gives an overview of the methods in this project:

- Methods for evaluating actual land use (dependent variables) (section 3.3.1).
- Rationale for use of a decision tree as modeling technique (section 3.3.2)

- Methods to pre-process independent variables and reduce to a manageable set (section 3.3.3).
- Methods to build stand-alone decision tree models: (section 3.3.4)
- Methods for evaluating model performance and variable importance (section 3.3.5).
- Methods for integrating stand-alone model results (section 3.3.6).
- Methods for cross-validating Boston models with data from other urban areas, and building models for those areas as alternatives (section 3.3.7).

3.3.1 Evaluation of Actual Land Use

Regardless of the method ultimately chosen to create predictive models, we felt it was important to systematically evaluate the reference data in order to understand (a) where it is, (b) how it is configured and how the classes relate to each other, and (c) similarities and differences between classes. This part of the project was fundamental to understanding the urban landscape in the study area and provided insight and valuable results even aside from the results of predictive model building. We undertook this using both quantitative and qualitative methods. Some of these results are given in this section, because they provide context to the methods in general.

The first step was to simply map the actual land use in the study area. Figure 3-9 shows the spatial distributions of residential and non-residential classes around Boston (broken out simply for the purpose of making them easier to see on the map).

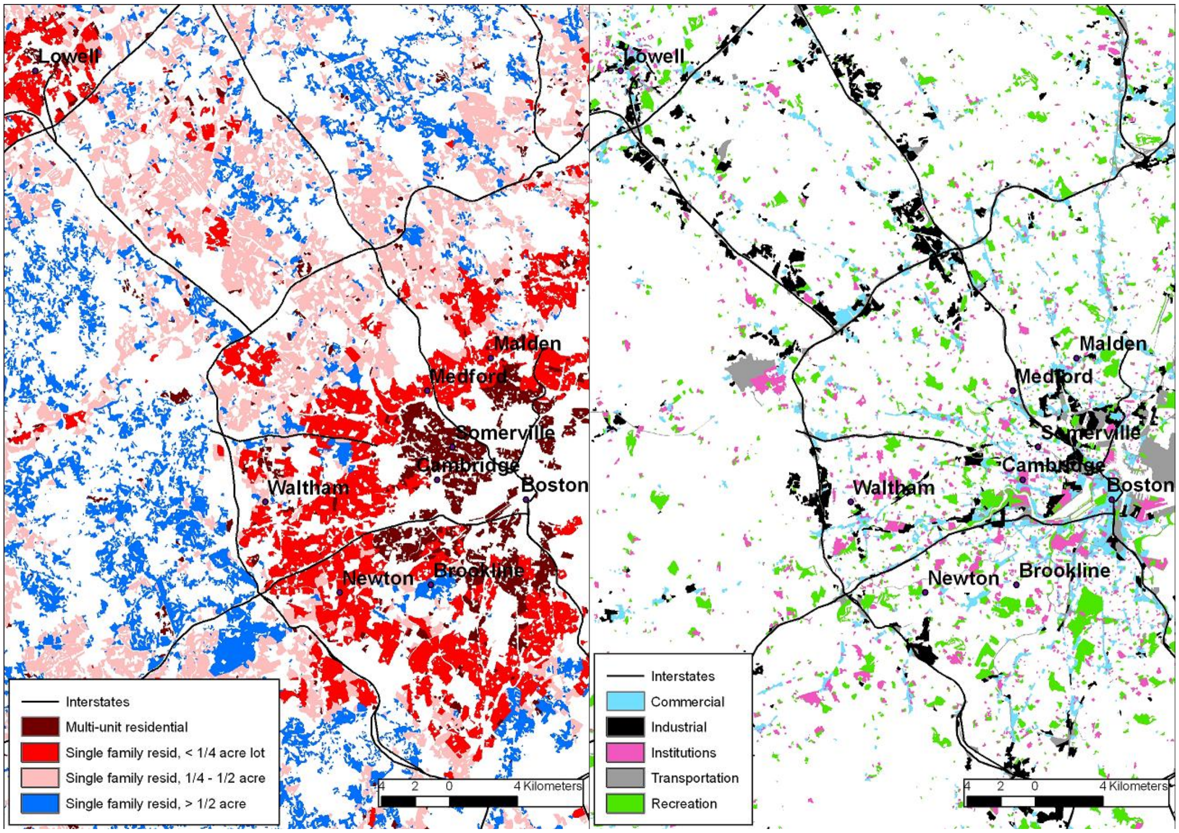


Figure 3-9: Actual land use (MassGIS, 2008).
Left panel shows residential classes. Right panel shows non-residential classes. Non-urban land is not shown for clarity.

Several general hypotheses suggest themselves for urban land use mapping based on these figures:

- There is a general gradient of residential land use intensity around the largest cities: in this region around the largest city, Boston, and to a lesser extent Lowell and Lawrence (Lawrence not shown) (2000 populations 105,000 and 72,000, respectively). With a few exceptions, nearly concentric rings of housing unit density exist around Boston. Distance and access to the central city is certainly a major determinant of land use.

- Non-residential uses dominate in central Boston. Commercial land is prevalent in the CBD and shows linear patterns along road systems. Industrial land is often organized around major transportation routes.
- Institutions and recreational lands are scattered. Predicting their spatial distribution based on location is very challenging with national-scale data.

The global Moran I statistic was calculated for the percent of each type of actual land use across the study area by block group (Figure 3-10). High values of the statistic (normalized here as z-score) indicate greater clustering. This reinforces the visual impression of clustering of the MFRES class, particularly, and more dispersed nature of the INDUST, INSTIT, TRANSP, and RECR_OPEN classes at a regional scale.

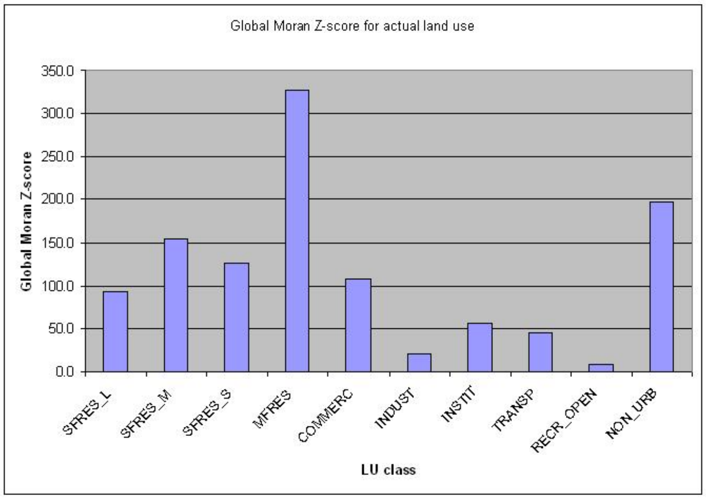


Figure 3-10: Global Moran I z-scores for actual land use distribution.
All values have $p < 0.01$. Highest z-scores indicate greatest spatial clustering.

We explored the pattern and location of actual land use using several quantitative measures. The first of these was to calculate a number of basic statistics for each type, to characterize the classes. Several noteworthy characteristics are given in Table 3-4.

The areal extent of urban lands is dominated by residential land, and not surprisingly, by the two lowest density classes. Mean percent imperviousness (percent of area covered by manmade sealed surfaces) provides a measure of urban intensity, with commercial land having the largest percentage of impervious surfaces. Mean imperviousness, however, may not give information about the spatial configuration (fragmentation) of urban surfaces. For example, the SFRES_S and MFRES classes have similar overall imperviousness (53 and 58%, respectively), but in different configuration: only 24% of SFRES_S land consists of large patches > 2 ha in size, compared to 44% of MFRES land. Other pattern metrics we calculated for the reference classes reinforced that the COMMERC, INDUST, and MFRES classes consisted generally of high levels of concentrated imperviousness, and the INSTIT, TRANSP, and SFRES_S classes of high imperviousness but of a more discontiguous nature. The SFRES_M, RECR_OPEN, and SFRES_L classes have decreasing percent and concentration of impervious surfaces.

Table 3-4: Basic statistics of actual land use in the study

LU class	Percent of urban land in study area	Mean percent imperviousness	Pct of class in large (> 2ha) patches
SFRES_L	26.27	13.0	0.1
SFRES_M	28.26	31.8	1.3
SFRES_S	13.49	53.2	24.0
MFRES	4.20	57.9	44.3
COMMERC	5.71	65.5	52.9
INDUST	5.87	57.4	41.6
INSTIT	3.80	48.5	24.3
TRANSP	4.66	49.8	23.3
RECR_OPEN	7.73	18.6	5.9
NON_URB	-	3.5	0.7

What is the relationship of these classes to the current National Land Cover Dataset? This is a question that may of keen interest to some users. To test this, we calculated the likelihood of a pixel of any NLCD01 class being in a particular LU class, based on the percentage that actually did fall in that class (Table 3-5):

Table 3-5: Percentage of pixels of each NLCD01 class that fell in each land use type in the four-county Boston study area.
Codes are in first column; see Appendix B2, e.g. class 11 is Open Water.

	SFRES_L	SFRES_M	SFRES_S	MFRES	COMMER	INDUST	INSTIT	TRANS	RECR_	NON_UR	
11	1.93	2.12	2.08	0.80	1.10	1.14	2.33	1.53	3.54	83.43	100.00
21	17.14	16.57	7.01	4.29	2.46	2.55	5.51	10.07	28.77	5.63	100.00
22	12.05	19.07	13.38	7.76	4.92	6.12	9.48	13.51	11.35	2.35	100.00
23	2.13	9.60	18.18	16.93	12.00	11.92	12.49	12.51	3.48	0.77	100.00
24	0.06	0.42	6.38	14.31	28.46	22.41	13.24	12.12	2.28	0.31	100.00
31	2.62	1.00	1.89	0.43	6.49	27.64	6.74	8.27	37.52	7.41	100.00
41	27.78	9.84	1.64	3.32	2.61	3.48	5.54	2.52	9.50	33.76	100.00
42	33.19	10.23	1.10	3.75	1.78	2.23	5.16	1.24	8.32	33.01	100.00
43	19.58	6.31	1.04	2.94	2.69	4.98	4.24	2.03	11.85	44.34	100.00
52	20.64	7.30	1.05	4.00	3.76	8.35	6.17	4.56	26.38	17.78	100.00
71	11.49	4.41	0.95	2.04	1.87	6.80	16.63	6.00	39.89	9.92	100.00
81	17.62	5.91	0.80	4.32	5.32	7.60	14.50	4.32	27.13	12.46	100.00
82	3.50	0.87	0.20	2.20	10.41	29.37	14.99	7.96	21.58	8.92	100.00
90	12.45	5.10	0.71	2.43	3.94	6.57	3.84	5.14	8.90	50.93	100.00
95	6.25	3.75	3.91	3.29	4.66	8.22	8.57	10.37	13.61	37.37	100.00

The classes of primary interest in Table 3-5 are the four urban classes (21-24). While there is a very rough correspondence to land use in some cases – for example, the lower intensity classes 21 and 22 do have a somewhat higher likelihood to be in the lower density residential and recreation classes - the thematic detail of this information is very low. For example, a class 23 pixel could just as easily be single-family residential, multi-family residential, commercial, industrial, institutional, or transportation. The NLCD was not designed to provide urban land use information.

Another method we used to quantify relationships amongst the actual land use classes was to calculate the percentage of each land use that was within certain distances from every other land use. We did this by buffering all the pixels of a single land use class by 200m, 400m, and 800m, then calculating the percentage of all other types that fell within those buffers. Results are shown in Table 3-6 for the 400-m buffer results:

Table 3-6: Percentage of type x within 400m of type y, for actual land use.

				type x							
		SFRES_L	SFRES_M	SFRES_S	MFRES	COMMER	INDUST	INSTIT	TRANS	RECR_	NON_URB
	SFRES_L	-	76	30	35	48	51	48	54	62	46
	SFRES_M	50	-	56	46	61	52	56	45	55	23
type y	SFRES_S	7	24	-	44	47	33	40	22	32	6
	MFRES	8	22	37	-	47	30	35	22	24	5
	COMMER	20	41	68	78	-	64	53	52	42	11
	INDUST	11	22	33	43	51	-	29	47	27	8
	INSTIT	17	40	74	69	56	32	-	25	42	8
	TRANS	11	14	25	36	45	51	23	-	22	7
	RECR_	38	54	75	74	68	62	78	50	-	20

The table summarizes spatial proximity of the actual land uses for that buffer distance. For example, looking at the last column: 46% of NON_URB land is within 400-m of a large-lot single family residential area (SFRES_L), but only 5% of NON_URB land is within 400m of an apartment or row house land use (MFRES). Several relationships are noteworthy: for example, that roughly ¾ of high density residential land (MFRES and SFRES_S) is close to COMMERC, INSTIT, and RECR_OPEN.

Table 3-7: Cross-correlation matrix (Spearman's rho) of actual percent land use within block groups.
Values in red have $p < 0.001$. Values greater than |0.30| are bolded.

	SFRES_L	SFRES_M	SFRES_S	MFRES	COMMERC	INDUST	INSTIT	TRANSP	RECR_	NON_URB
SFRES_L	-	**0.59**	**-0.35**	-0.22	-0.24	0.19	-0.15	0.09	0.14	**0.68**
SFRES_M	-	-	-0.29	-0.26	-0.16	0.14	-0.10	0.01	0.14	**0.54**
SFRES_S	-	-	-	**-0.41**	-0.05	-0.13	0.03	-0.17	-0.07	-0.29
MFRES	-	-	-	-	0.15	-0.04	0.07	0.02	-0.03	-0.25
COMMERC	-	-	-	-	-	0.13	0.15	0.09	-0.01	-0.21
INDUST	-	-	-	-	-	-	-0.08	**0.30**	0.15	**0.31**
INSTIT	-	-	-	-	-	-	-	-0.07	0.10	-0.21
TRANSP	-	-	-	-	-	-	-	-	0.11	0.21
RECR_	-	-	-	-	-	-	-	-	-	0.26
NON_URB	-	-	-	-	-	-	-	-	-	-

We additionally characterized spatial distribution with a cross-correlation matrix of percentages of actual land use within each block group (Table 3-7). Block groups are large enough areas to almost always contain several land uses (98% of BGs in this study have multiple land uses). The correlations give only a snapshot of LU within the block group, and do not capture relationships over larger scales, however do give additional information about relationships amongst the classes: for example, the fairly strong negative correlation between MFRES and SFRES_S (-0.41) indicates that they are rarely both present in the same block group, and confirms the visual impression in Figure 3-9 of their general separation into distinct zones. It was hoped these observations would lead to a better understanding of the classes and potentially how they could be predicted and mapped.

A final series of analyses were performed to test class separabilities. That is, to what degree are the land use classes different, from the predictor variables we assembled. To do this we took a random sample of 600 ground truth polygons of each class across the study area (i.e. 6000 polygons total). A random sample was used in order to minimize effects of spatial autocorrelation. We then calculated the mean, standard deviation, and other basic statistics from a wide variety of predictors for those samples and compared the values from each class in order to understand which classes were "most different" and "most similar" and in what regard. The analyses performed included:

- Test for significant differences between any two dependent variable classes for a predictor variable (Wilcoxon signed-rank test). The Wilcoxon signed-rank test is a non-parametric equivalent to a matched-pairs t-test (Burt and Barber, 1996), as many of the independent variables were non-normal in distribution.
- Summarize similarities between classes for multiple variables using a distance measure (Mahalanobis distance).
- Summarize differences among all classes for a specific variable (Kruskal-Wallis test), e.g. does population density or road density more effectively "separate" the LU classes. The Kruskal-Wallis test is a non-parametric equivalent to an ANOVA F-test (Burt and Barber, 1996).
- A hierarchical cluster analysis (dendrogram), as graphical representation of class similarities.

These results are summarized in Section 4.1.

3.3.2 Rationale for Decision Tree as Modeling Technique

We assembled numerous predictor variables for this project because, as national-scale mapping of urban land use by any means is little-researched, the literature guidance on the efficacy of specific variables is not clear. Identifying effective national-scale predictor variables, or classes of variables, is one of the desired outcomes of this project.

We also observed that a number of key relationships between predictor and dependent variables are non-linear in form. For example, high density single-family residential land (SFRES_S) has overall a low correlation to population density by block group, because at very high population densities the land is built up as multi-family residential (MFRES). However, taken over a smaller part of its range where population densities are somewhat lower, there is a much stronger (positive) correlation. The presence of many clearly non-linear data relationships argued against a linear modeling approach in this project.

A number of scientific fields, for example the machine learning, data mining, or medical/genome communities, have for years used tree-based methods to handle similar situations – i.e. where there are high-dimensional data, uncertainty about the form of the independent variables, and potentially complex and non-linear interactions among variables (Cutler et al., 2007; Maindonald and Braun, 2007). Tree-based methods, which produce a set of rules for classification or regression of the data, have a number of advantages in such situations, however are considerably – some would say radically - different than more traditional regression techniques (Venables and Ripley, 1999; Maindonald and Braun, 2007). The strengths and weaknesses of tree-based methods have been detailed in the literature (e.g. Quinlan, 1993; Yohannes and Hoddinott, 1999; Cutler et al., 2007; De'ath, 2007; Maindonald and Braun, 2007), and might be summarized as follows:

Strengths:

- They are non-parametric and make no distributional assumptions of any kind on the dependent or independent variables.
- Relatively complex interactions among large numbers of independent variables may be modeled.
- Independent variables can typically be either categorical or continuous data.
- The model performance is not affected by outliers nor collinearities. Outliers are isolated in a node and do not have any effect on tree splitting.
- Implementations typically have a built-in method for dealing with missing values of a variable for a case.
- They are insensitive to monotonic transformations of the independent variables, i.e. there is no need or effect of transforming explanatory variables to logarithms or square roots.
- They in essence reduce dimensionality: a large number of independent variables may be reduced to a model incorporating only a few important variables.
- The results, for methods resulting in a single tree, generally produce an easily interpretable set of rules.

Weaknesses:

- They are less effective with small datasets or small numbers of explanatory variables, for which parametric or linear regression may be more suitable. That is, straightforward linear relationships may be less obvious.
- Individual trees do not have a probability or confidence interval associated with them.
- Some implementations treat continuous variables, inefficiently, as categories.
- Large trees, or ensembles of trees, make poor intuitive sense and must be treated essentially as black boxes. There are, however, strategies which may ameliorate that (section 3.3.5).

Decision trees in general work on the following principles (Cutler et al., 2007; Maindonald and Braun, 2007): A set of cases are input which have the values of a single dependent variable and multiple independent variables associated with them. The independent variables are examined with regard to the dependent variable and binary splits are identified which partition the data into homogeneous regions ("nodes" of the tree). The nodes are split to maximize the variance between the nodes (between-group residual sum of squares). Each node is then split again recursively until further subdivision achieves some predefined threshold ("when to stop"). That threshold may be a predefined tree depth, a minimum number of cases per node, or a measure of variability such as a Gini index (Cutler et al., 2007). The lower branches of the tree ("terminal nodes") model sample error, so lower branches may additionally be "pruned" to improve performance. The result is a tree (with root at the top), with a set of if-then rules defining the predicted values. This may be done either as a category or continuous value. The result is a non-linear solution which resembles a piecewise linear model, except that rules may overlap (Xian et al., 2002). Decision trees have been used in numerous applications of detection and mapping of land cover and land use (Debeir et al., 2002; Falcone and Pearson, 2006; Homer et al., 2007; Walton, 2008).

Figure 3-11 gives a simple example of a decision tree, based on the R recursive partition tree function rpart() (R project, 2009). The dependent variable to be predicted in this example are values for the INDUST class (i.e. the percent of industrial land in the block group). The root node sets the rule "is the independent variable *TRANSP_ratio_roadden_imperv* >= 0.1962?". If the answer is yes, the branch is followed to the left; if not, to the right, and the cases are split accordingly. On the right

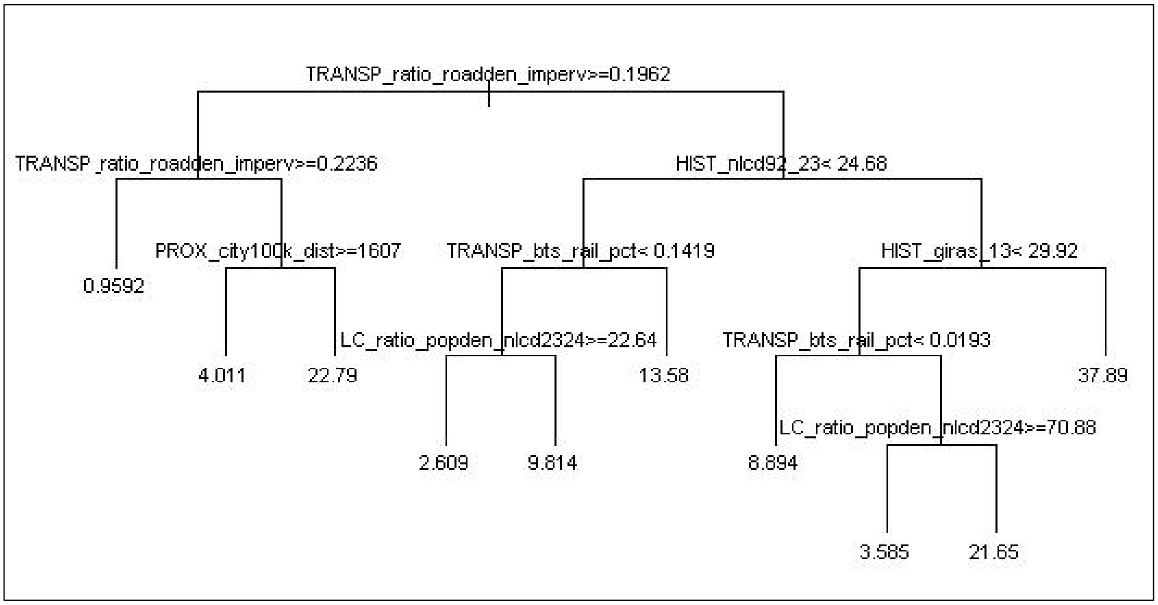

TRANSP_ratio_roadden_imperv>=0.1962
TRANSP_ratio_roadden_imperv>=0.2236
HIST_nlcd92_23< 24.68
0.9592
PROX_city100k_dist>=1607
TRANSP_bts_rail_pct< 0.1419
HIST_giras_13< 29.92
4.011
22.79
LC_ratio_popden_nlcd2324>=22.64
13.58
TRANSP_bts_rail_pct< 0.0193
37.89
2.609
9.814
LC_ratio_popden_nlcd2324>=70.88
8.894
3.585
21.65

Figure 3-11: Example decision tree, based on the R function rpart().

branch, the cases remaining are split on the rule "is *HIST_nlcd92_23* < 24.68?" If yes, then left, if not then right, again. Eventually the terminal nodes are assigned a predicted value based on the dependent variable mean of the remaining cases for that node.

Because it is possible that different training samples from a large dataset may give different trees, a technique known as bagging ("**b**ootstrap **agg**regation") helps reduce instability of the model (Breiman, 1996). In bagging, multiple bootstrap training samples (with replacement) are extracted from the dataset and a tree is created for each. The final result is an aggregation of the results for all the trees, i.e. either a majority vote for categorical classifications or an average for regression. This is sometimes referred to as an example of "ensemble learning" (Liaw and Wiener, 2002). The major drawback of ensemble trees ("forests"), as noted above, is their lack of interpretability: interpreting the averaged results of thousands of trees is obviously more difficult than a single tree.

The specific decision tree tool to be used in this project is the Random Forests (RF) classification and regression method (Breiman, 2001; Liaw and Wiener, 2002; Breiman and Cutler, 2009). We also evaluated another popular (commercial) decision tree package, Cubist (Rulequest, 2008), and found the publicly-available Random Forests performance to be at least as good. Random Forests has been shown to be adept at handling complex non-parametric interactions and highly correlated predictors (Cutler et al., 2007; Strobl et al., 2008). Random Forests runs under the public domain R statistical system (R project, 2009). As in traditional bagging, RF selects a bootstrap sample (5/8 of the training data). The remainder that are left out are referred to as out-of-bag (OOB) observations. A tree is fit to each bootstrap sample, however, RF adds an additional layer of randomness to bagging, in that at each branch of the tree a random subset of the predictor variables is selected, which reduces correlation between trees (Breiman, 2001; Walton, 2008). Trees are grown to their maximum depth (not pruned). This approach has been shown to perform very well compared to other methods, including more traditional ones (e.g. multiple linear regression) or more complex ones (e.g. discriminant analysis, support vector machines, or neural networks) (Liaw and Wiener, 2002, Carlisle et al., 2009). The final prediction for each observation is obtained by averaging the predictions across all the trees. Because of the ensemble nature of the prediction, RF is believed to be relatively robust against overfitting, a primary concern of traditional regression trees (Breiman, 2001; Liaw and Wiener, 2002; Pal, 2005; Carlisle et al., 2009), although there is evidence that a large number of noisy predictors degrades performance (Segal, 2004; Walton, 2008). Random Forests has the facility for classification, regression, or unsupervised learning.

For each tree the fitted model is applied to the out-of-bag observations, and the mean error and coefficient of determination (r^2) (in the case of regression) for those samples are aggregated and reported. These are called the OOB estimate of error, and are the equivalent of what is sometimes referred to as k-fold or v-fold cross-validation (Siroky, 2009). These OOB error rates are believed to be comparable to those that would be reported from completely independent data (Carlisle et al., 2009), have the advantage of not requiring a separate independent dataset for assessing the accuracy of the method, and allowing the model to be built on the maximum number of cases available.

A further desirable feature of RF is that it reports two measures of “variable importance” (Liaw and Wiener, 2002; Maindonald and Braun, 2007; Kuhn et al., 2008). These are:

Mean decrease in accuracy (**%IncMSE**); essentially a “leave-one-out” measure: mean increase in error in the OOB set from the training set when the variable is permuted. Higher values indicate higher variable importance.

Mean decrease in Gini impurity index (**IncNodePurity**): every time a node is split on a variable the Gini impurity index of the two descendent nodes is less than the parent node. This measure sums the Gini index decreases for each variable. A higher IncNodePurity represents a higher value of importance, i.e. nodes are “purer”.

The method of analysis of the importance measures in this project is given in more detail in later sections.

The prediction made by RF for regression is the average of the n tree predictions (e.g. 500 predictions), however it is also possible to obtain a matrix of all of the predictions, for either the training predictions (the ones that went in to making the model), or the out-of-bag validation predictions (those withheld from model building). From this one may calculate a standard deviation, standard error, and confidence intervals for the prediction for each case.

We recognize that there are other possible predictive techniques that might have been employed in the project (e.g. stepwise linear regression, canonical correlation analysis, geographically-weighted regression), however we believed the decision tree approach was most appropriate for the goals of this application, given the complexity of the predictors and the desire to be able to apply some models to withheld data (e.g. other urban areas). Section 3.3.4.2 gives more detail on some simple tests that were performed to compare modeling techniques for this application.

Some researchers identify a dichotomy between explanation and prediction in statistical modeling (e.g. Shmueli and Koppius, 2007). On one hand, if the primary goal is accurate prediction, then models may be more complex/less interpretable. If the primary goal is explaining a process, then some performance may be sacrificed at the benefit of clarity. In this project a middle way is taken to some degree. The primary goal of this project is to demonstrate good accuracy in predicting land use, but with a reasonably parsimonious method involving a small number of predictors.

3.3.3 Predictor Variable Evaluation and Data Reduction

As with the dependent variable evaluation, our first step in examining the independent variables was to simply map them. This step, although qualitative, was nonetheless helpful in understanding the spatial distribution of elements of the landscape, particularly given that an end goal of the predictive modeling was to create land use maps. Figure 3-12 provides an example of visual comparisons: the actual recreation land (dependent variable RECR_OPEN; upper left panel) compared to the land cover class that was most similar to recreation from the three independent variable land cover datasets available (GIRAS, NLCD92, and NLCD01). The GIRAS and NLCD92 classes appeared to have the potential to be reasonably good predictors; the NLCD01 less so.

We created scatterplots and boxplots for many of the IVs against dependent variables, and created a correlation matrix of independent vs. dependent variables (Appendix C). The scatterplots allowed identification of possible blunders or gross errors in the data. The correlation matrix, although it provided only information about linear correlation of IVs and DVs, nonetheless gave insight about which IVs were nonetheless likely to be promising even given the non-linear decision tree approach. In some cases new independent variables were created based on these observations. For example, it was observed that INDUST areas had typically very high imperviousness but low road density, leading to the creation of the variable *TRANSP_ration_roadden_imperv*, which was thought to have the potential to be a better predictor than either of those variables separately.

The data ranges and values for the independent variables were checked and in some cases a small number of values were manually filled. For example, a number of IVs represented ratios (e.g. *LC_ratio_popden_nlcd2324*), where there was a possibility that the denominator could be zero (causing the spreadsheet-calculated value to resolve to "#DIV/0!"), in this example if there was no land cover of either class 23 or 24 in the BG. In this case the denominator was set to a very small but positive number (e.g. 0.1). The number of necessary adjustments of this kind was very small.

Approximately 320 independent variables had been assembled (the variables listed in

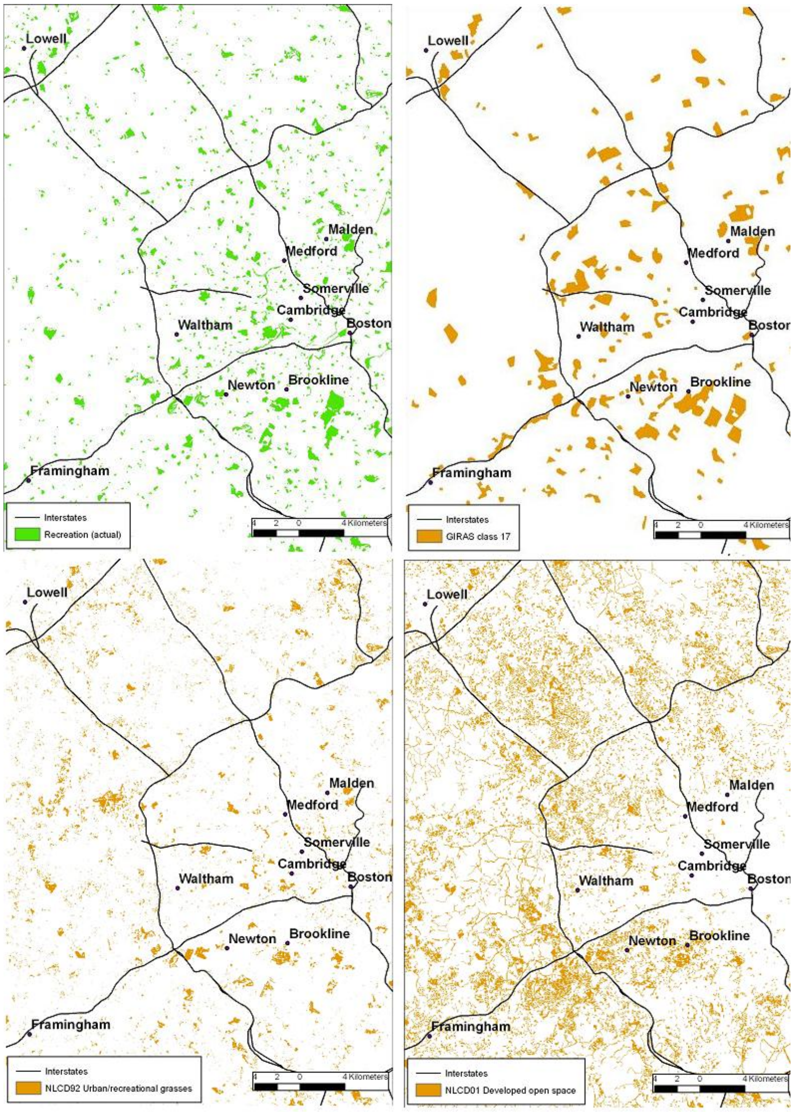


Figure 3-12: Example visual comparison of dependent and independent variables
DV RECR_OPEN (upper left) to the IVs HIST_giras_17 (upper right), HIST_nlcd92_85 (lower left), and LC_nlcd01_21 (lower right).

Appendices B1-B10). Some of these were already summaries of a number of data layers (e.g. *TRANSP_alltrans*), or ratios of variables, however there was obvious multicollinearity in the data. Some researchers have suggested that keeping correlated variables is actually beneficial in RF regression because, although some variables may be correlated, they nevertheless may participate in building parts of the tree which add predictive power (Cutler et al., 2007). Nevertheless, for interpretability sake, we believed that reducing some multicollinearity would be beneficial, as models with multiple collinear variables would be more difficult to interpret. In this application there seemed to be several solutions to reducing the effect of redundant variables (Falcone et al., 2009): one was to use a statistical technique such as principal components analysis (PCA) to re-orient the information content of the original variables into new uncorrelated components. A second solution was to use literature-based evidence or judgment to eliminate some variables which showed multicollinearity. We used a conservative approach using the latter solution, for several reasons. Creating predictor variables from PCA component scores, while creating uncorrelated predictors, resulted in variables which had very low interpretability. That is, even when we used PCA rotations which are known to provide more interpretable results (e.g. Varimax rotation), it was very difficult to assign interpretable meaning to most of the resultant variables. Further, our initial tests using PCA-derived variables suggested PCA-based predictors were likely to have lower performance than using original variables. Section 3.3.4 gives more detail on testing with PCA-based predictor variables.

To reduce the number of IVs, we first used results of preliminary testing and judgment to eliminate approximately 70 which we were fairly certain had little predictive power. A cross-correlation matrix was then constructed for the resultant set of 250 variables, which gave the linear correlation (Spearman r value) of each of the 250 against each of the others. This matrix was examined and variable pairs which had an r value > 0.9 were identified. In each of these pairs the variable that was believed to be the least interpretable was eliminated. In a very small number of cases both variables were retained because it was too difficult to judge interpretability and/or we believed there might be small but important differences in the two variables. The primary example of this was the variable pair *CENS_popden* and *CENS_huden* (population and housing unit densities, respectively). The resultant set consisted of 188 variables, which was the final set of independent variables used as the starting point in model building and analysis.

3.3.4 Model Building

As noted earlier, the basic method employed in this project is to create predictive models which estimate the percent of each of 10 different land uses for a block group; therefore 10 predictive models are created (as well as 3 additional ones to create a simpler 6-class scheme to apply to areas outside Boston). The results of those stand-alone models are then integrated (described in section 3.3.6). Individually solving for separate dependent variables has the advantage that each model may be tuned to best suit that target, and the alternative – solving for multiple dependent variables simultaneously – is rarely attempted (Tofallis, 1999), in part because of the highly complex nature of interpreting such a result. We recognize that land uses do not exist independently of one another, however it was not possible to explicitly take other land uses into account in any stand-alone model because the land uses themselves are not known (they are the dependent variables to be predicted). However, it was believed that the predictor data assembled, to include land cover spatial pattern, locations of landmarks, measures of proximity to different feature types, etc. would serve as the best available proxies for incorporating the effect of other land uses into any stand-alone model.

3.3.4.1 Creating 10-class and 6-class stand-alone models

Once the 188 independent variables were finalized, a model was created for each of the dependent variables in Random Forests. Each model was based on the entire dataset of 2,764 block groups, and performance evaluated primarily by the OOB error on withheld records from each tree. RF has a small number of tuneable parameters that a user may control to tweak performance. The primary ones are **ntree**, the number of trees to grow, and **mtry**, the number of variables to be randomly selected at each split. We experimented extensively with both parameters. We found the results to be generally insensitive to increasing ntree greater than 500. Likewise, Liaw and Wiener (2002) suggest varying mtry to ½ the default (default = n / 3), the default, and 2x the default, however note that results do not typically change significantly. That was also our experience. We therefore used 500 trees and the default (n / 3) mtry values throughout for the models of this project. Execution time increased primarily with the number of predictor variables. For example, using a Dell Precision Workstation 670, running RF with 7 predictor variables and 2,764 records building 500 trees took 20-25 seconds, while running the same with 188 predictor variables took 8-9 minutes to complete.

One of the unique features of RF is that, because of the random nature of variable selection at splits, the results are slightly different for each execution (every time a new forest is created), using identical input. Our experience was that, using 500 trees or more there was a fairly small range in results. For example if one execution resulted in an r^2 of .615, others might be .613 or .617. We typically tested models using an average over two-three runs. The precision of the result increased with model performance; i.e. models with very high r^2 varied virtually not at all between executions.

Our primary measure of performance for fine-tuning the models was the root mean square error (RMSE) and r^2 for the out-of-bag withheld data. Those same measures for the training data themselves are less meaningful (and uniformly very low and high, respectively, e.g. $r^2 > 0.9$), because of the ensemble nature of the model (personal communication, Andy Liaw, Dec. 10, 2009). Our guidance from the developers of RF was to focus on the OOB performance measures as the most meaningful measures of model performance.

After each RF execution we ranked the two importance measures (%IncMSE and IncNodeGenerally) that resulted. The importance measures generally, but did not always, agree in their rankings of importance. There is no clear guidance as to which importance measure to prefer (Kuhn et al., 2008). Their agreement is greater with very strong predictor variables. For example, in a typical RF execution with 100 predictor variables, if a predictor was ranked in the top 10 by one measure, it typically was also ranked in the top 10, or close to that, by the other measure. A very weak predictor, on the other hand, might be ranked 75th by one measure, but 50th or 100th, by the other. We interpret this lack of precision in identifying importance for weak predictors as meaning there was essentially no difference between the 50th ranked predictor or the 100th ranked predictor, and that predictors with such low ranks were essentially noise. The importance measures were a general indicator of which variables were likely to most improve performance (RMSE and r^2), and were most useful as a filter for screening variables that had little or no effect on the models.

With that background, we created the models in the following fashion:

A model was created for each dependent variable using all 188 predictor variables as input. The RMSE and r^2 values were noted. The resultant %IncMSE and IncNodePurity were examined for each dependent variable. A cumulative distribution plot was made of those metrics and examined. Figure 3-13 gives an example of the cumulative distribution values of %IncMSE (increase in error if variable is left out) for a typical result: y-axis are importance scores, x-axis represent each of the 188 variables. Higher y-axis values represent variables which are more important in the decision trees.

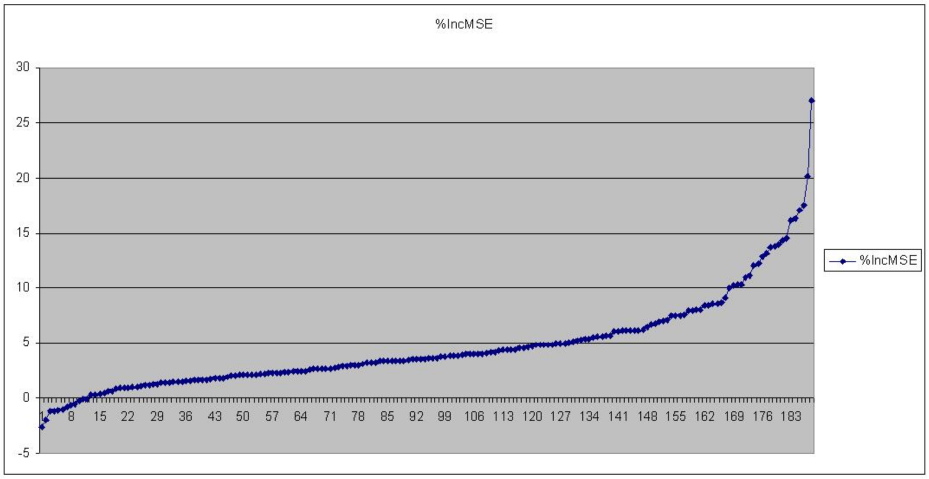


Figure 3-13: Distribution of %IncMSE for an initial RF run of 188 predictors against a dependent variable (in this case MFRES).

Our experience in testing was that a great number of the 188 variables could be dropped with no appreciable loss of performance, and the breakpoint at which variables could be dropped was generally related to the point at which the slope of the distribution lines of the two importance measures started to go up steeply. We qualitatively judged that to include any variable that was in the top 20 importance scores of either %IncMSE or IncNodePurity. There was not always agreement between the two measures, as noted earlier, so this resulted in a reduced set of 20-30 predictor variables for each dependent variable. This set of 20-30 predictor variables for each model represented a "first cut" at creating a manageable and interpretable model from the 188 IVs for each dependent variable. Figure 3-14 shows a flow chart of the progression from 320 variables to final stand-alone models:

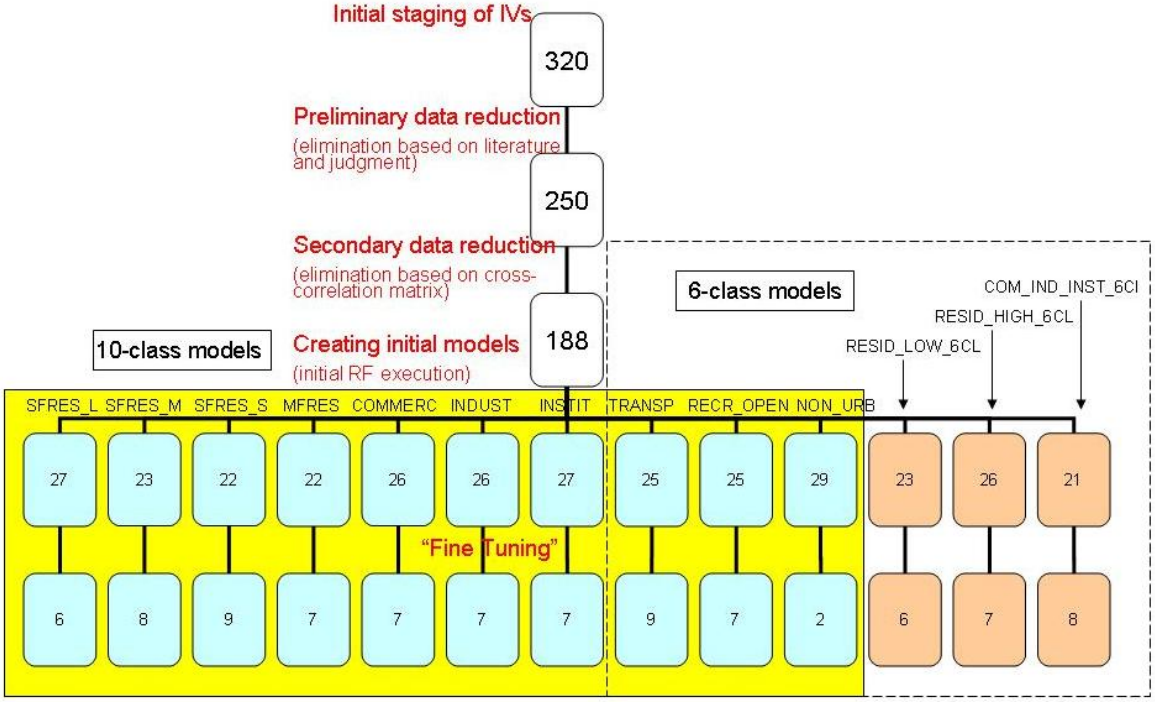


Figure 3-14: Flow chart showing number of independent variables that existed in each step of the process of creating final models.
The TRANSP, RECR_OPEN, and NON_URB classes are shared in 10-class and 6-class schemes. The "fine-tuning" process that occurred in going from 20-30 variables to a final smaller subset is described below.

Our goal was to create a reasonably economical model, with a small number of predictor variables that still had fairly high performance. To that end, the "fine-tuning" stage of model building incorporated a number of different methods to whittle down the set of 20-30 predictor variables remaining. The following general steps were taken for each of the dependent variables:

RF was executed using the new set of 20-30 variables for each dependent variable. The RMSE and r^2 values were noted. In every case the performance of the models was as good or better using the set of 20-30 variables than using the full set of 188. This confirmed results reported by Segal (2004) and Walton (2008) that a large number of noise predictors were likely to degrade performance somewhat. The resultant %IncMSE and IncNodePurity were examined for each variable. As noted earlier, they were useful but not perfect indicators of which variables were likely to make the best models.

Although our experience with creating PCA-based predictors from the set of 188 had not been that fruitful, we nonetheless experimented with doing the same from the set of 20-30. PCA was performed (R princomp() function) using the correlation matrix, then loadings rotated using the Varimax rotation. Component loadings were examined and an attempt to interpret them against original variables was made. The full set of PC scores, as well as subsets of the PC scores (e.g. only the first 5 or 10) were used as predictors and RF re-executed. Typically performance did not improve significantly (and was sometimes worse), and the resultant new variables were also more difficult to understand, and in many cases not interpretable. We eventually abandoned using PCA-based predictors.

Even though the greatest multicollinearity had been previously removed (in reducing to the set of 188), there was still the potential of considerable multicollinearity to be present in the set of 20-30. We examined which variables had similar information content primarily from the Variance Inflation Factor (VIF; R vif() function), which gives a measure of how correlated each variable is to every other variable in a linear regression. High VIF scores indicated variables which were the most collinear with other variables. We attempted to eliminate variables in such a way that the resultant set of variables had low VIF scores, because this would aid in interpreting results.

We also executed a stepwise linear regression for each set of 20-30 variables. While we did not expect a linear regression to give better performance at this point than the decision tree (and it did not), the stepwise process (R step() function) provided information about which variables were most significant in a linear model, which could potentially translate to the tree method. We examined the p-values for the resultant variables as another indicator of which variables were useful.

During this process we also used non-statistical knowledge (judgment) based on our understanding of the physical nature of relationships of the dependent and independent variables. We used the rpart() function to create single trees (e.g. Figure 3-11, previously shown), which helped understanding how the variables were related. We generally favored variables which were more interpretable and easier to understand, all other things being equal.

RF was then re-executed in a “leave-one-out” manner, eliminating variables which were likely to be the least effective in building the model, based on the above criteria. We recorded the resultant RMSE and r^2, and if believed useful, mapped the results and/or residuals to examine them for patterns. This was done in a stepwise manner until eliminating any variable caused the model r^2 to drop by more than 1% of total variance explained (i.e. .01). These were then considered to be the final models of this project for the Boston area. The number of variables that resulted in each model is given in the bottom row of the flowchart in Figure 3-14, above. We calculated the confidence interval for each case for each land use type based on the withheld OOB validation predictions.

Our final step in building models for the Boston area was to examine the effect of adding the ESRI Business Analyst predictor variables (Appendix B11). These were tested only for the INSTIT, COMMERC, and INDUST classes, as they would improbably affect the other dependent variables. The Business Analyst predictor variables were injected into the final models for those three classes, and the results were evaluated.

All results are given in Section 4.

3.3.4.2 Tests of other methods

We wished to convince ourselves that we had not missed some potentially simple method for predicting land use, other than that described above. To that end we compared a version of the method above to two other methods. In order to be able to compare performance between the three methods we randomly separated the 2,764 BGs into two groups: a training set consisting of 1,939 records (2/3) and a validation set consisting of 825 records (1/3) (because not all three methods could use the OOB validation performance measures). The three methods were:

- Executing RF for each dependent variable based on the set of 1,939 records and 188 predictor variables. The resultant models were then applied to the 825 validation dataset and RMSE and r^2 recorded.
- Performing a stepwise linear regression for the same records and variables for each dependent variable. The resultant equation was then likewise applied to the validation dataset.
- Executing principal components analysis for the 188 variables and using the first 90 PC scores used as predictors in an RF model. (Experimentation had shown that performance was slightly better with the first 90 components as opposed to using all 188 or a smaller subset of 30). The resultant models were likewise applied to the validation data set.

Results for these three tests are given in Section 4.

3.3.5 Model Evaluation, Variable Importance, and Training Data Sensitivity

This section describes model evaluation for the stand-alone models described above. Section 3.3.6 describes methods for integrating the stand-alone models.

As noted above, we have used the RMSE and r^2 values of validation datasets as our primary measures of model performance for the stand-alone models described above. R^2 is a measure of association but not necessarily accuracy, however r^2 values may be reasonably compared between models based on the same number of observations (e.g.

comparing r^2 for the SFRES_L model to the MFRES model). RMSE is a better measure of accuracy, however comparing models directly is diffucult, because the mean percentage of land use varies among the classes. The RMSE values in this project are therefore normalized (NRMSE) using the following (Qian and Rasheed, 2004; Karunasinghe and Liong, 2006):

$$\text{NRMSE} = \sqrt{\sum_i (P_i - A_i)^2} \; / \; \sqrt{\sum_i (P_i - \overline{A}\,)^2} \qquad (2)$$

Where P_i = the predicted percent of land use for block group i, A_i = the actual percent of block group i, and $\overline{A}$ = the mean of actual values. This gives a performance measure relative to a prediction of the mean (i.e. a “null model” - Pontius et al., 2004; Wu et al., 2009b): e.g. if the mean percent of SFRES_S land across the study area is 15%, then the null model predicts 15% for every record. If the mean is predicted for every record then NRMSE will be 1. If every prediction is exactly correct then NRMSE will be 0. Values of (1-NRMSE * 100) represent a percent improvement over prediction of the mean (e.g. if NRMSE = 0.6, then it is a 40% improvement over prediction of the mean). Predictions worse than the null model should be considered to be poor.

Stand-alone models were also evaluated by mapping results, residuals, and confidence intervals for each block group and calculating the spatial autocorrelation of residuals for each model. The primary goal of this project is to demonstrate good accuracy in predicting land use with a reasonably parsimonious method, and uses the spatial autocorrelation of predictors as a valuable input, however minimal spatial autocorrelation of residuals is also a desirable outcome.

The interaction of predictor variables in each model was also qualitatively evaluated by creating individual decision tree solutions using rpart(). These individual tree diagrams (as in the example in Figure 3-11) give useful information about general relationships between dependent and independent variable using a decision tree algorithm that is functionally similar to RF. This is a method by which the "black box" nature of RF is circumvented for general analysis.

Variable importance was evaluated by examining the rankings of the two importance measures reported by RF, and more importantly in final models, by calculating the loss of predictive power if a variable is omitted.

Variable importance was also evaluated in two other ways:

- By predictor class. That is, did certain classes of predictors (e.g. Proximity) typically have greater importance than other classes, for a specific land use, or across land uses, and,
- By evaluating specific predictors across all land use classes. That is, are there certain variables or types of variables which are uniformly useful in predicting multiple types of land use.

Finally, we performed sensitivity analysis on training data sample sizes. As noted above, we used all data for the entire study area as input to Random Forests, which then selects random samples of 5/8 (62.5%) of the data for training and 3/8 for validation.

Using all data for the study area maximizes model interpretability, but does not give information about “how little” training data might be needed for comparable performance. To that end, we tested random samples which resulted in 50, 40, 30, 20, 10, and 5% of all study area data being used as training, and validated those against a separate 10% random sample. These results gave an indication of model performance drop-off vs. training sample size.

3.3.6 Integrating Stand-alone Model Results

The stand-alone models described so far make a prediction for a particular land use for each block group, without directly interacting with each other. We believed this was a reasonable but not complete solution to the problem of predicting multiple continuous variables in the same areal unit, and as noted earlier, there are no areal units defined at the national scale which incorporate a single land use (such as parcels). There was also a precedent for the method: two products as part of the USGS NLCD01 had a similar development – the Impervious Surface and Forest Canopy sub-pixel datasets (USGS, 2009d). Each of these two products likewise made an independent prediction of a percent-of-area dependent variable for the same areal unit: the percent imperviousness and percent forest canopy for every 30-m pixel across the US. In their case, as in our case, the results of the models clearly relate to each other in an inverse manner: that is, if a pixel is predicted to have 90% imperviousness, and the prediction is reasonably accurate, then the pixel is likely to have very low forest canopy, and vice versa.

We considered how best to incorporate the results of our models. While we felt that the predictors that were assembled were likely to already account for the effects of other land use forces in the landscape (e.g. distance to cities, spatial correlation of land cover types), we wished to do two things for a more complete product: (1) attempt to leverage information from the other predictions, and (2) constrain the results of all models so that the sum of all predictions equaled 100%. In that way each stand-alone model would be functional as is (i.e. it would be technically possible to make a prediction of recreational land across the US without incorporating any of the other models), but there would also be a prediction of each land use based on the integration of all the other models.

The method was as follows:

- Each model is our best prediction of how much of each land use exists in the block group, and is therefore the best proxy available for the actual land use. The predictions for the other land use classes were therefore used as predictors themselves in a "second pass" of the final model. For example the final SFRES_L model was based on 6 predictor variables: those 6 variables were combined with the predictions for the percent of the other 9 land use classes (which are based on withheld validation data not involved in model building) so that the second pass consisted of 15 predictors. We reasoned this was the most effective way to incorporate as much information as was available about other land uses in each model.

- Theoretically the new predictors would improve the model for that class, leading potentially to a "third pass" (or fourth, etc.) of iteratively including results of the other models. In practice, however, we found that improvements were limited to the second pass and there was no benefit in a third pass.
- Once these integrated models were created they still had the potential to sum to greater than 100% for any block group. The final models were therefore constrained to 100% by dividing each stand-alone prediction by the sum of the stand-alone predictions and multiplying by 100. The result was 10 predictions (for the 10-class scheme) which summed to 100, and 6 predictions (for the 6-class scheme) which also summed to 100 for each block group.

We recognized that there were other possible solutions, such as allowing fuzzy boundaries amongst classes (possibly based on the confidence intervals for each prediction), however we believed that creating a product that was constrained to 100% was the most straightforward and interpretable result.

3.3.7 Cross-validation with Other Urban Areas (6-class models)

Our final major analysis was to compare models across geographic areas. Once the 6 stand-alone models which represented the 6-class scheme were completed (RESID_LOW_6CL, RESID_HIGH_6CL, COM_IND_INST_6CL, TRANSP, RECR_OPEN, and NON_URB – see Figure 3-14 (flowchart)), they were validated and compared against data from the three external urban areas: Providence, RI, Atlanta, GA, and Los Angeles, CA. The intent was not to perform a comprehensive evaluation of land uses from those cities, but to compare how models derived from one urban setting would transport to sample data from other urban areas which had (potentially) different characteristics.   The goal was to gain insight into how a true national mapping of urban land uses could best be achieved, beyond this project. A brief description is given of each of the three external datasets:

The study data for Providence consisted of the entire state of Rhode Island, comprising five counties (Figure 3-15). A handful of block groups on islands were eliminated, so that the validation dataset consisted of 815 BGs. This was equivalent to approximately 1/3 of the BGs in the Boston study area.

We wanted the number of block groups from each of the three external areas to be equivalent, so we limited the number of block groups in Atlanta and Los Angeles to approximately the same number as were available for Providence (although theoretically we could have extended the Rhode Island data by adding block groups from neighboring southern Massachusetts, we felt that the 800+ BGs from RI were adequate as a good sample of data from an area different to Boston). A random block group was selected near central Atlanta and a group of contiguous BGs surrounding that point were selected (Figure 3-16). Contiguous block groups were needed in order to calculate the Spatial Autocorrelation independent variables described above. The final set consisted of 833 BGs.

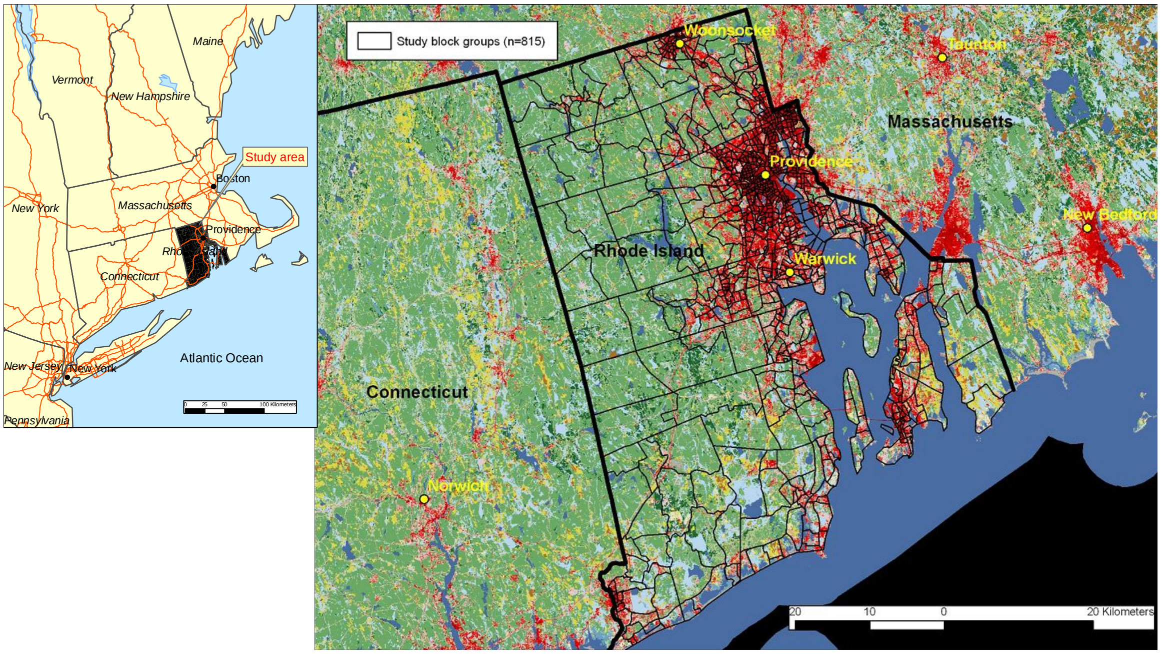


Figure 3-15: Study block groups for Providence, RI.
Background land cover is NLCD01. Red areas are developed. Scale of inset map matches insets for MA, GA, and CA.

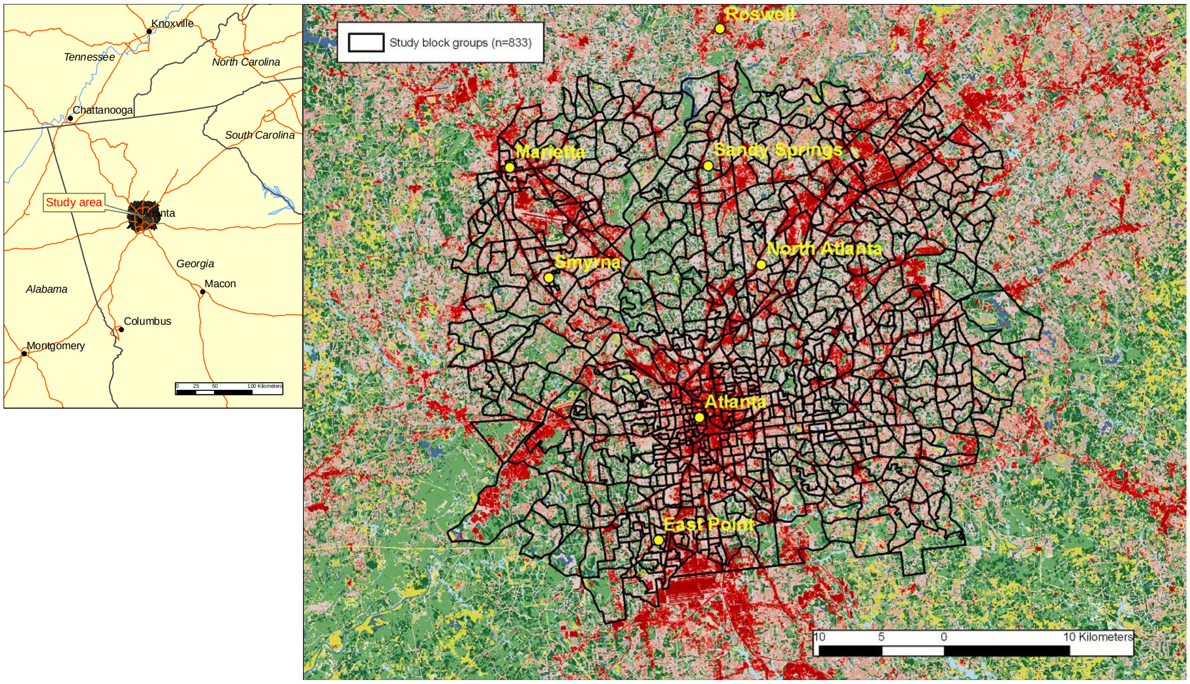


Figure 3-16: Study block groups for Atlanta, GA.

A like number was selected in similar fashion for Los Angeles (figure 3-17).

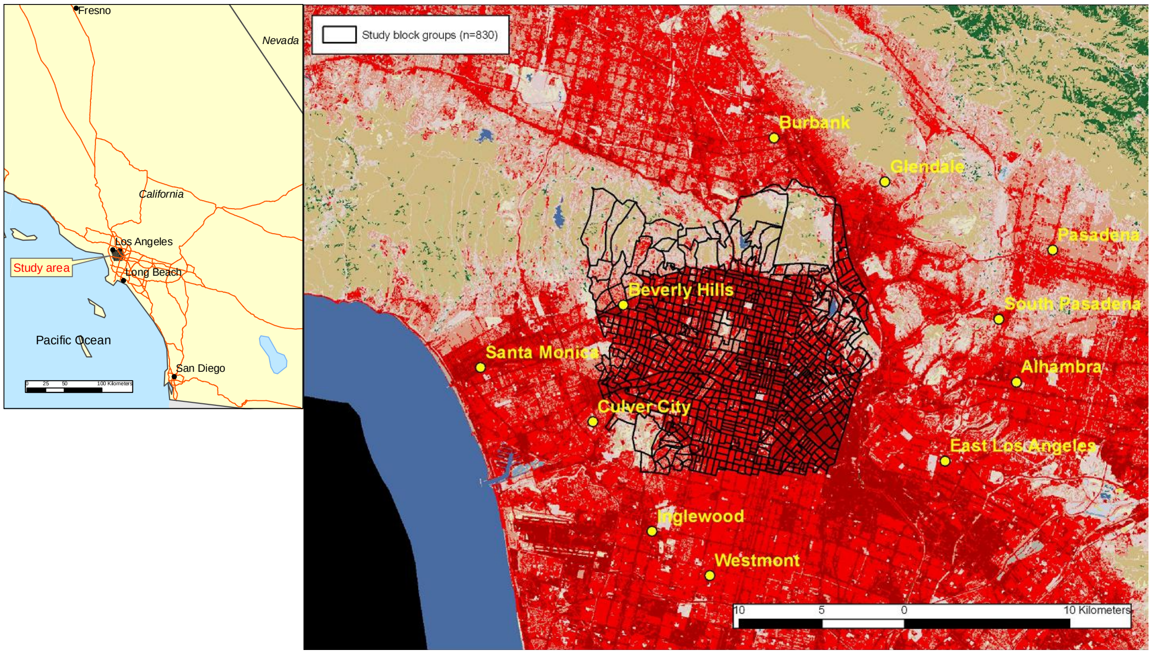


Figure 3-17: Study block groups for Los Angeles, CA.

Some basic statistics for each of the three areas and the Boston validation set are given in Table 3-8. Each area encompassed an area of more than 1 million people.

Table 3-8: General statistics of block group samples from study areas. Census data are from 2000 Census.

	Providence	**Atlanta**	**Los Angeles**	**Boston (validation)**
n (# of BGs)	815	833	830	825
Area (sq km)	2,859	1,535	241	1,531
Population	1,048,000	1,562,057	1,177,000	1,054,000
Median pop. density (#/sq km)	1,565	1,227	6,072	2,306
Number of schools (from GNIS)	781	606	242	598
Median household income ($)	42,100	43,600	28,800	54,400
Pct workers using public transport	1.6	6.2	12.1	9.4
Total population of entire Metropolitan Statistical Area (MSA)*	1,583,000	4,248,000	12,366,000	4,391,000

* MSA: Providence-New Bedford-Fall River, RI-MA

* MSA: Atlanta-Sandy Springs-Marietta, GA

* MSA: Los Angeles-Long Beach-Santa Ana, CA

* MSA; Boston-Cambridge-Quincy, MA-NH

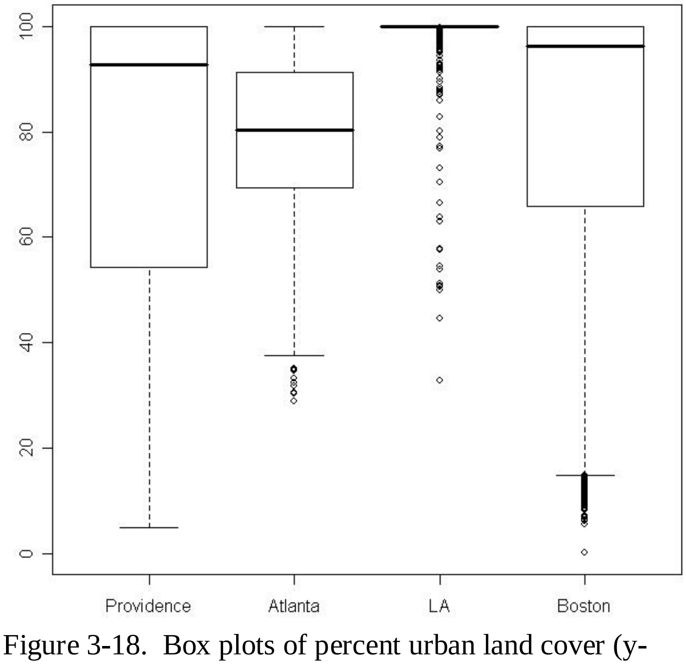


Figure 3-18. Box plots of percent urban land cover (y-axis) from the four datasets in this study.
The median is represented by the center line of boxes and boxes represent inter-quartile range.

The characteristics of the study areas clearly differ. Figure 3-18 shows boxplots of the percent urban land cover (NLCD01) by block group for all four datasets. The data distribution of percent urban land cover is most similar between Providence and Boston. The Los Angeles data represent an area of very high intensity urbanization, and Atlanta and Los Angeles are characterized by Griffith and Wong (2007) as polycentric (multiple major population centers).

The fact that each of these datasets represented different versions of urbanization provided the opportunity to test (a) the hypothesis that models developed for one area would transport best to areas that were most similar, and transport poorly to areas that were most different, and (b) more broadly, to test if it was feasible whatsoever to transport models to areas outside of where the training data originated. After re-calculating the dependent and independent variable values for each of the three areas we compared data characteristics among the four datasets. A clustering technique (R functions dist(), hclust(), and plclust()) was used to create dendrograms based on a number of key indicators of urbanization (population density, percent imperviousness, road density, vegetation index), (Figure 3-19), and socio-economic factors (not shown).

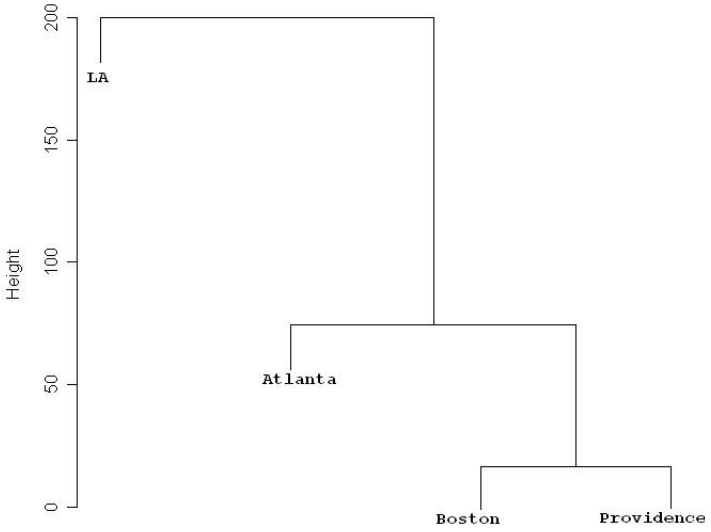


Figure 3-19. Dendrogram based on several key measures of urbanization from the four data samples. Nodes are organized by similarity to each other: Boston and Providence are the most similar to each other by those measures; the Los Angeles block groups are the most different to the other three. Dendrogram based on 19 socio-economic variables yields a nearly identical result.

It should be noted that, although by many measures the Boston and Providence samples are most similar to each other, by some specific measures they may not be. For example, based on the variable “percent of workers using public transport” (Table 3-6) Boston is least similar to Providence and most similar to Los Angeles: a function of the development of mass transit in those particular areas. Similarity may vary depending on the variables used for comparison.

We applied the Boston 6-class models to the data from each of the other areas and recorded the results. The stand-alone models were used because they were more straightforward conceptually. Because the data had already been calculated and the method for creating models had already been established, it was reasonably straightforward to repeat the process for the datasets from the other three areas, as follows.

- 6-class models were created from each of the Providence, Atlanta, and Los Angeles datasets, using the method described above (section 3.3.4.1). The data were validated from the withheld OOB data, as for the Boston 6-class models.
- The models for each city were then validated against the other three cities. The validation dataset used for the Boston area consisted of the 1/3 sample validation dataset described previously (n=825), so that validation for each area was based on a similarly-sized dataset. The result was four sets of tables showing the performance of each set of models against data for its own area, and against the other areas.
- The importance of specific predictor variables and classes of variables was recorded and compared across the four datasets.

4. Results

The results are broken into four sections:

- Evaluation of actual land use classes (section 4.1). This helps to "set the stage" for evaluation of land use generally, and is a continuation of partial results presented in section 3.3.1.
- Results of testing other methods (section 4.2). This was a test to assure ourselves that our modeling method was reasonable, compared to two alternatives.
- Results of 10-class models for Massachusetts study area (section 4.3).
- Results of 6-class models (section 4.4). Transporting models between Massachusetts, Rhode Island, Georgia, and California.

Sections 4.3 and 4.4 represent the key results of this project.

4.1 Land Use Separability

Evaluating the separability of the true land use classes helps to inform the prediction and mapping of land use (even beyond this project), understanding class differences, and the urban landscape in general. As noted earlier, examining the differences of classes with respect to certain variables in some cases led to the creation of new independent variables that in some cases were important predictors. Additionally, evaluating class differences helps to explain why some classes are more difficult to predict than others.

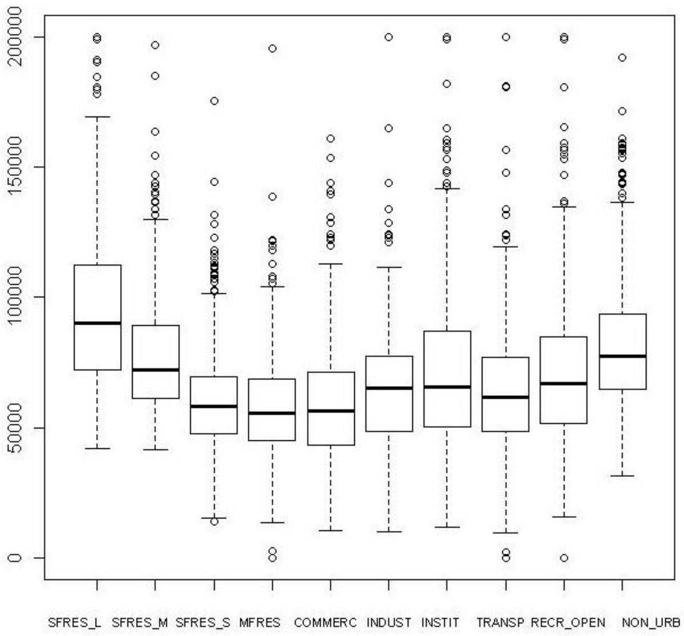


Figure 4-1. Range of values for median household income for 600 reference polygons of each class. Y-axis units are dollars.

As described in the Methods section, the following results were derived from a random sample of 600 polygons of each land use class from the reference data. Differences by class were examined in both graphical and tabular format. For example, Figure 4-1 shows the range of values of block group median household income (*CENS_median_hh_income)* by land use class. There were significant differences between household income for large lot single-family residential (SFRES_L; median = $94,600) and the higher density residential classes (SFRES_S median = $60,400 and MFRES median = $57,600). The income for block groups in which non-residential reference polygons are located are more similar to the high density residential classes. Appendix D provides additional data distributions by LU class for a number of variables.

The statistical difference between any two classes for these and other variables was also assessed with the Wilcoxon signed-rank test. (Rank tests were used here and generally elsewhere because the data for this and many of the variables did not have normal distributions.) For example, differences in the household income data from Figure 4-1 are given in Table 4-1. These confirm the observation above, i.e. that the most significant differences in income are between SFRES_L and the SFRES_S and MFRES classes.

Table 4-1: Pairwise z-value results from Wilcoxon rank-sum tests for block group median household income (*CENS_median_hh_income*).
Higher values indicate greater differences between classes. Values in red have $p < 0.001$.

	SFRES_L	**SFRES_M**	**SFRES_S**	**MFRES**	**COMM**	**INDUS**	**INSTIT**	**TRANS**	**RECR_O**	**NON_UR**
SFRES_L	-	13.94	21.42	21.85	20.98	18.23	14.31	18.76	14.39	7.94
SFRES_M	-	-	11.28	15.08	14.02	9.11	5.87	10.36	5.22	-3.66
SFRES_S	-	-	-	2.37	1.41	-4.16	-5.77	-2.79	-7.29	-16.31
MFRES	-	-	-	-	-0.85	-5.79	-7.41	-4.66	-8.81	-17.12
COMM	-	-	-	-	-	-4.93	-6.49	-3.78	-7.94	-16.19
INDUS	-	-	-	-	-	-	-2.36	1.12	-3.54	-12.19
INSTIT	-	-	-	-	-	-	-	3.36	-0.83	-8.43
TRANS	-	-	-	-	-	-	-	-	-4.50	-13.03
RECR_O	-	-	-	-	-	-	-	-	-	-8.06

Examining pairwise class differences graphically and statistically was useful for ensuring that by individual, or combinations of measures, every class had unique characteristics. This was clearly the case, as even from the one census variable given here as example, there are statistically significant differences between most classes.

We also summarized the differences among all classes for individual variables by using the non-parametric Kruskal-Wallis chi-squared test. Table 4-2 shows results for a number of key indicators (same variables as shown in boxplot figures):

From these results, land use classes have generally greater differences with regard to land cover than socio-economic measures. It is noteworthy, however, that even the strongest variable, imperviousness, does not distinguish between every class (Appendix D), for example SFRES_S and MURES or INSTIT, nor does this table provide information about how well any variable predicts a specific land use class.

Table 4-2: Kruskal-Wallis chi-squared test results.
Higher values indicate there is greater separability among all LU classes for that variable. All p-values < 0.001, and df=9.

Variable	**Chi-sq value**
LC_nlcd01_imperv_mean	3825
LC_sum_nlcd01_allveg	2498
TRANSP_allroads_density	2317
SPCAT_patch_2ha_pct	2141
CENS_huden	2118
CENS_popden	2073
PROX_cost_10k_city	1367
CENS_median_hh_income	1062
CENS_pct_hu_owneroccupied	955
PROX_cost_50k_city	920
PROX_cost_100k_city	626
CENS_pct_nonwhite	546
CENS_pden_change90_00	71

We developed a distance matrix between every land use class using the Mahalanobis distance from eight of the above variables believed to be important (*LC_nlcd01_imperv_mean, LC_sum_nlcd01_allveg, TRANSP_allroads_density, SPCAT_patch_2ha_pct, CENS_huden, CENS_popden, PROX_cost_10k_city,* and *PROX_cost_100k_city).* The Mahalanobis distance is a measure of separation between datasets (Venables and Ripley, 1999), and shows the degree of similarity between classes for those variables (Table 4-3).

Finally, we used the same clustering technique shown earlier to create a dendrogram which organized the LU classes according to those same eight variables (Figure 4-2):

Table 4-3: Mahalanobis distance results.
Values by themselves are not meaningful, but are useful simply for comparison: low values indicate classes are "similar", high values indicate classes are "different" with regard to the measured variables.

	SFRES_L	SFRES_M	SFRES_S	MFRES	COMM	INDUST	INSTIT	TRANS	RECR_	NON_URB
SFRES_L	-	2.46	6.29	7.39	11.78	10.81	6.37	6.12	3.16	4.49
SFRES_M	-	-	1.75	3.88	8.47	8.80	3.95	3.90	6.31	12.68
SFRES_S	-	-	-	2.20	4.70	6.93	3.11	2.75	11.50	17.84
MFRES	-	-	-	-	4.30	5.34	1.57	3.01	9.69	15.58
COMM	-	-	-	-	-	1.37	2.98	1.55	14.69	19.71
INDUST	-	-	-	-	-	-	2.03	1.21	11.75	16.52
INSTIT	-	-	-	-	-	-	-	0.99	7.79	12.89
TRANS	-	-	-	-	-	-	-	-	9.13	13.45
RECR_	-	-	-	-	-	-	-	-	-	5.97
NON_URB	-	-	-	-	-	-	-	-	-	-

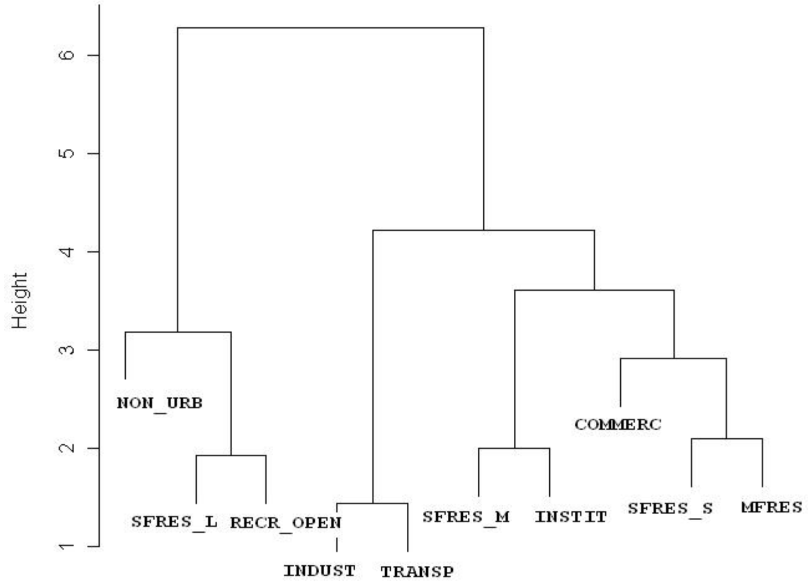


Figure 4-2. Dendrogram based on eight independent variables thought to be important in distinguishing LU types.
Nodes are organized by similarity to each other.

The distance matrix and dendrogram generally confirm prior observations about class similarities and physical proximities. In some cases these results suggest alternative aggregations of classes which could be examined in future work. The 10-class aggregation to 6-classes was based primarily on literature examples (for example our COMMERC, INDUST, and INSTIT classes were combined in part from the example of Gong et al., 1992), as well as perceived interpretability (i.e. combining residential classes seems logical).

As noted previously, these results of class separability tests are provided primarily as background and general information regarding the nature of the land use classes.

4.2 Results of Testing Other Methods

Although we believed the method we followed for predictive modeling was a reasonable one, we were inevitably curious about how some other methods might fare doing the same thing, at least in a general way. It was not the intent of this project to exhaustively explore every possible modeling technique, but the results here are nevertheless informative.

As described earlier, we compared the three following methods:

- Executing RF for each dependent variable based on the set of 2/3 sample of 1,939 records and 188 predictor variables. The resultant models were then applied to the 825-record validation dataset and RMSE and r^2 recorded. This is termed the RF_ORIG method below.

- Performing a stepwise linear regression for the same records and variables for each dependent variable. The resultant equation was then likewise applied to the validation dataset. This is termed the STEP_LIN method below.
- Executing principal components analysis for the 188 variables and using the first 90 PC scores used as predictors in an RF model, likewise applied to the validation dataset. This was a bit of a hybrid method, using Random Forests but with a PCA-derived set of predictors. This is termed the RF_PCA method below.

This was done for the 10-class scheme, with the following results:

Table 4-4: Results of other-methods testing.
R^2 values are given on the left and normalized percent improvement of RMSE from prediction of the mean on the right. Validation based on 825 withheld records.

r^2

LU class	**RF_ORIG**	**STEP_LIN**	**RF_PCA**
SFRES_L	0.481	0.402	0.326
SFRES_M	0.581	0.454	0.515
SFRES_S	0.679	0.427	0.583
MFRES	0.580	0.245	0.567
COMMERC	0.419	0.359	0.287
INDUST	0.442	0.374	0.278
INSTIT	0.382	0.329	0.199
TRANSP	0.576	0.510	0.254
RECR_OPEN	0.593	0.520	0.439
NON_URB	0.957	0.963	0.859

Pct improvement from prediction of means (1 – NRMSE * 100)

LU class	**RF_ORIG**	**STEP_LIN**	**RF_PCA**
SFRES_L	28	21	18
SFRES_M	35	25	29
SFRES_S	42	22	31
MFRES	35	1	34
COMMERC	24	18	16
INDUST	25	21	15
INSTIT	20	13	8
TRANSP	34	29	14
RECR_OPEN	35	30	24
NON_URB	79	81	59

As the RF_ORIG test represented the same (first step of the) method we used in this project we were reasonably satisfied that it would have at least as good, and probably better, performance than either of the other methods tested here, which were simple alternatives.

4.3 Results of 10-class Models

The 10-class models were created for the Massachusetts data only. Separate sections are given for describing stand-alone models (section 4.3.1) and integrated models (sections 4.3.2 and 4.3.3), which integrate the results of all land use models.

4.3.1 Stand-alone Models

The stand-alone models were evaluated from the withheld OOB validation data, as described previously. Table 4-5 summarizes their performance. Each model had a (potentially) different set of predictors, as noted in the **num_vars** column. (Results are discussed in detail in Section 5).

Table 4-5: Results of stand-alone model validation.

LU class	num_vars	r^2	Pct improvement from prediction of means (1 - NRMSE * 100)
SFRES_L	6	0.582	35
SFRES_M	8	0.648	41
SFRES_S	9	0.727	48
MFRES	7	0.706	46
COMMER	7	0.471	27
INDUST	7	0.429	24
INSTIT	7	0.411	23
TRANSP	9	0.568	34
RECR_OP	7	0.535	32
NON_UR	2	0.930	74

Table 4-6 shows the variables that were included in one or more of the final models.

Table 4-6: Variables that were part of final 10-class stand-alone models.
"X" indicates variable was built as part of that model. Variables are given in alphabetical order, which organizes them by category.

Variable	SFRES_L	SFRES_M	SFRES_S	MFRES	COMMERC	INDUST	INSTIT	TRANSP	RECR_OPEN	NON_URB AN
CENS_hu_median_numb_rooms	X		X	X						
CENS_huden							X			
CENS_pct_hu_owneroccupied							X			
CENS_pct_walkbike_to_work							X			
CENS_popden		X	X							
HIST_giras_12					X					
HIST_giras_13						X				
HIST_giras_14								X		
HIST_giras_17									X	
HIST_indust_all_times								X		
HIST_nlcd92_21		X	X							
HIST_nlcd92_22				X						
HIST_nlcd92_23						X				
HIST_nlcd92_85									X	
LANDMRK_gnisconsol_instit_density							X			
LANDMRK_gnisconsol_recr_density									X	
LC_nlcd01_21									X	
LC_nlcd01_24					X					
LC_nlcd01_imperv_mean	X									
LC_ratio_popden_nlcd2324					X	X				
LC_sum_nlcd01_2122		X								
LC_sum_nlcd01_urban										X
MISC_vg2000_mean	X									
PROX_airport_crossing_dist		X	X							
PROX_city100k_dist		X		X	X	X				
PROX_city250k_dist		X	X	X	X	X	X		X	
PROX_cost_10k_city					X					
PROX_cost_50k_city		X	X							
PROX_major_road_dist				X						
PROX_mean_dist_road	X	X	X					X		X
SA_localMoran_allnatveg			X	X						
SA_localMoran_lc_entropy									X	
SA_localMoran_popden			X	X				X		

Variable	SFRES_L	SFRES_M	SFRES_S	MFRES	COMMERC	INDUST	INSTIT	TRANSP	RECR_OPEN	NON_URB AN
SPCAT_shape_index_22_24	X									
SPCON_cohesion_slopeclass2							X			
SPCON_ed_slopeclass1	X									
SPCON_focal77_gt50_is_cv							X			
TRANSP_a11_a17_roads_density								X		
TRANSP_a11_a38_roads_density								X		
TRANSP_alltrans								X		
TRANSP_bts_faf2_pct					X					
TRANSP_bts_rail_pct						X		X		
TRANSP_ratio_roadden_imperv						X		X	X	

Once the final models were created the component variables were re-evaluated using a leave-one-out approach, i.e. the model was successively re-built using all variables but one, and the results recorded. The decrease in performance when a variable was left out was used as a measure for ranking the variables in the final models. Appendix E shows the final variables for each model and their relative importance as judged by that measure.

The above variables were included in our final models, however the importance of all of the 188 predictor variables by class is also of interest. We recorded the two RF importance measures after executing a prediction using all 188 variables, then took the average rank of the two importance measures as a measure of the variable's importance. This is provided in Appendix F, from which the strength of general classes of variables by land use class may be seen. For example, a number of Proximity (PROX_) and Spatial Autocorrelation (SA_) variables were consistently important across many land use classes, whereas Spatial Pattern variables (SPCAT_ and SPCON_) were consistently less important and when so, important to only one or two classes. These results are discussed in detail in Section 5.

What do the results look like? Because of the difficulty of visualizing every class together on one map, Figures 4-3(a)-(j) show a series of panels juxtaposing the actual land use (left side) with the modeled land use (right side) as predicted from the withheld validation records. These series of figures show the same subset of the study area shown in Figure 3-10 (because showing the entire study area makes it too difficult to see details), and uses the same color scheme as Figure 3-9 for the actual land use panels.

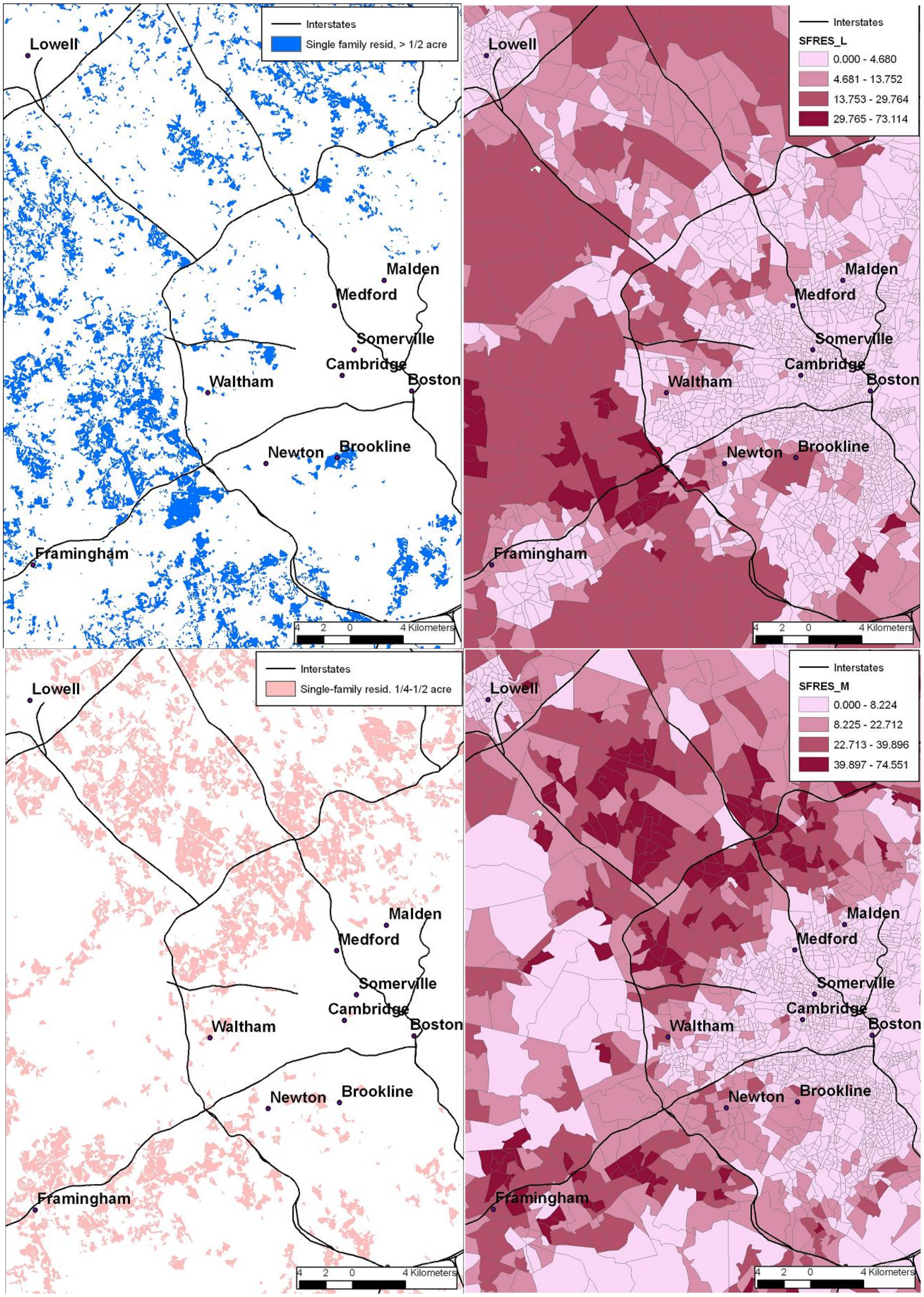


Figure 4-3(a)-(j): Actual (left) vs. predicted (right) land use.
Right hand panels are categorized using Jenks natural breaks breakpoints.

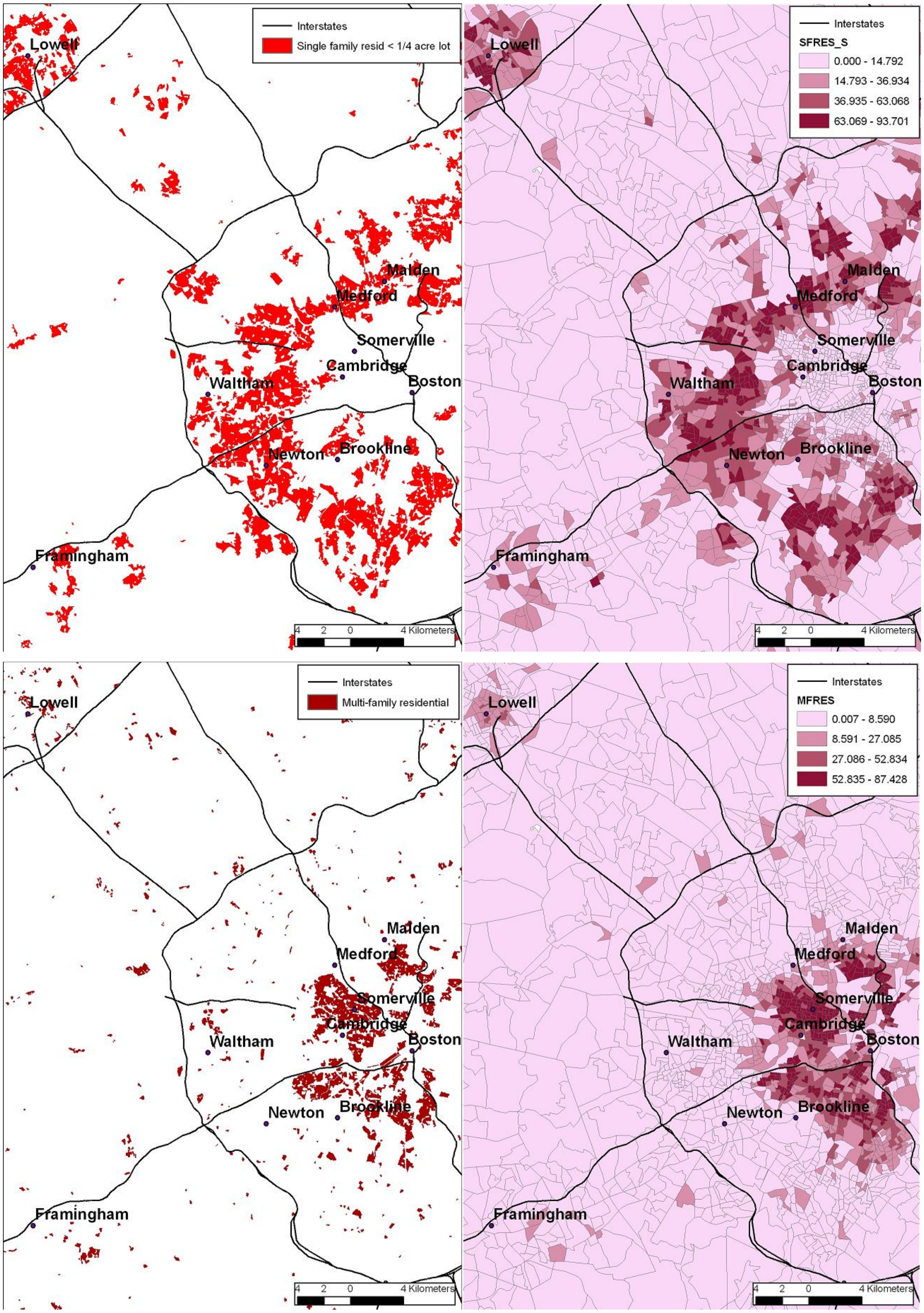


Figure 4-3(a)-(j) continued…

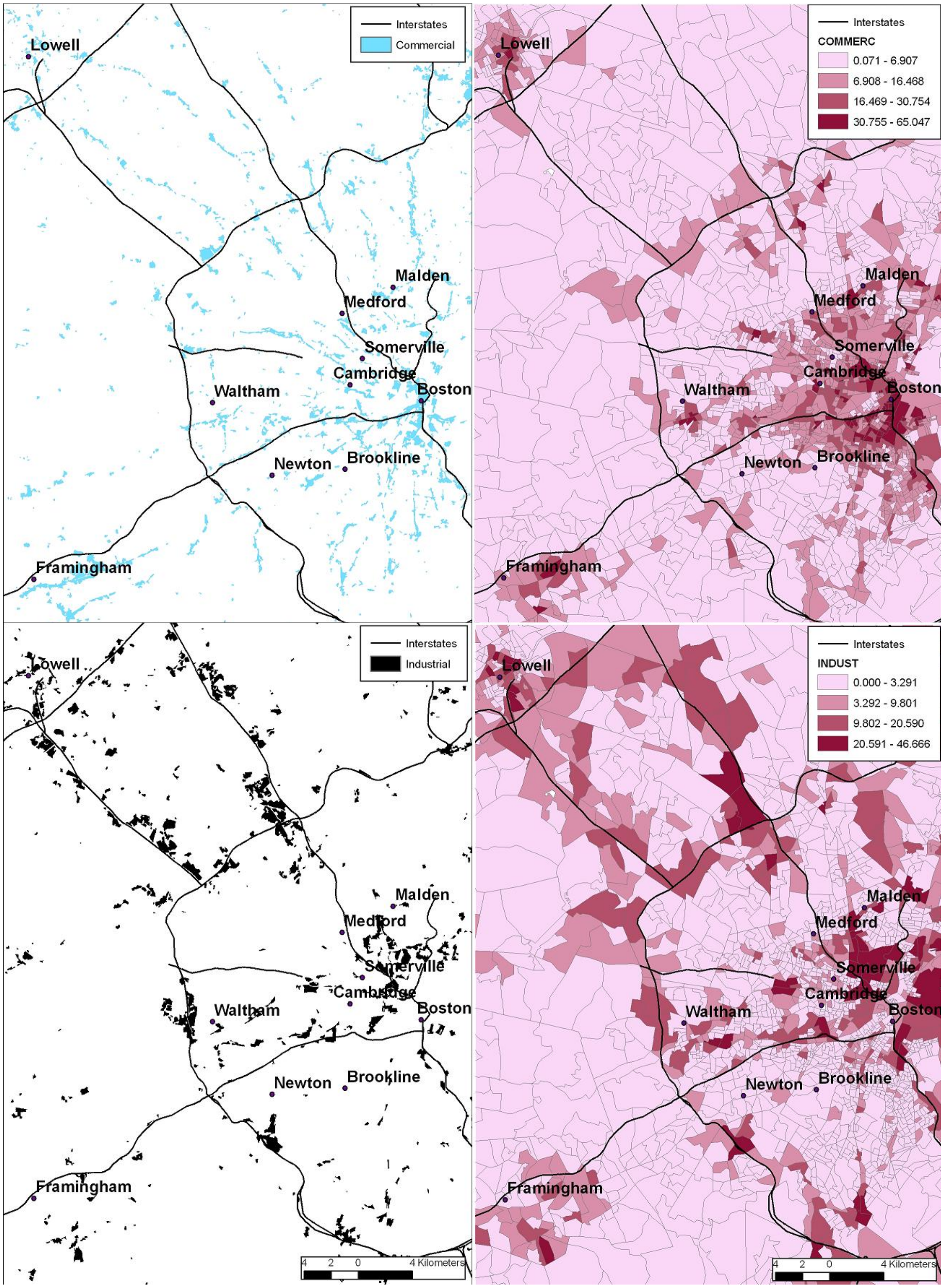


Figure 4-3(a)-(j) continued…

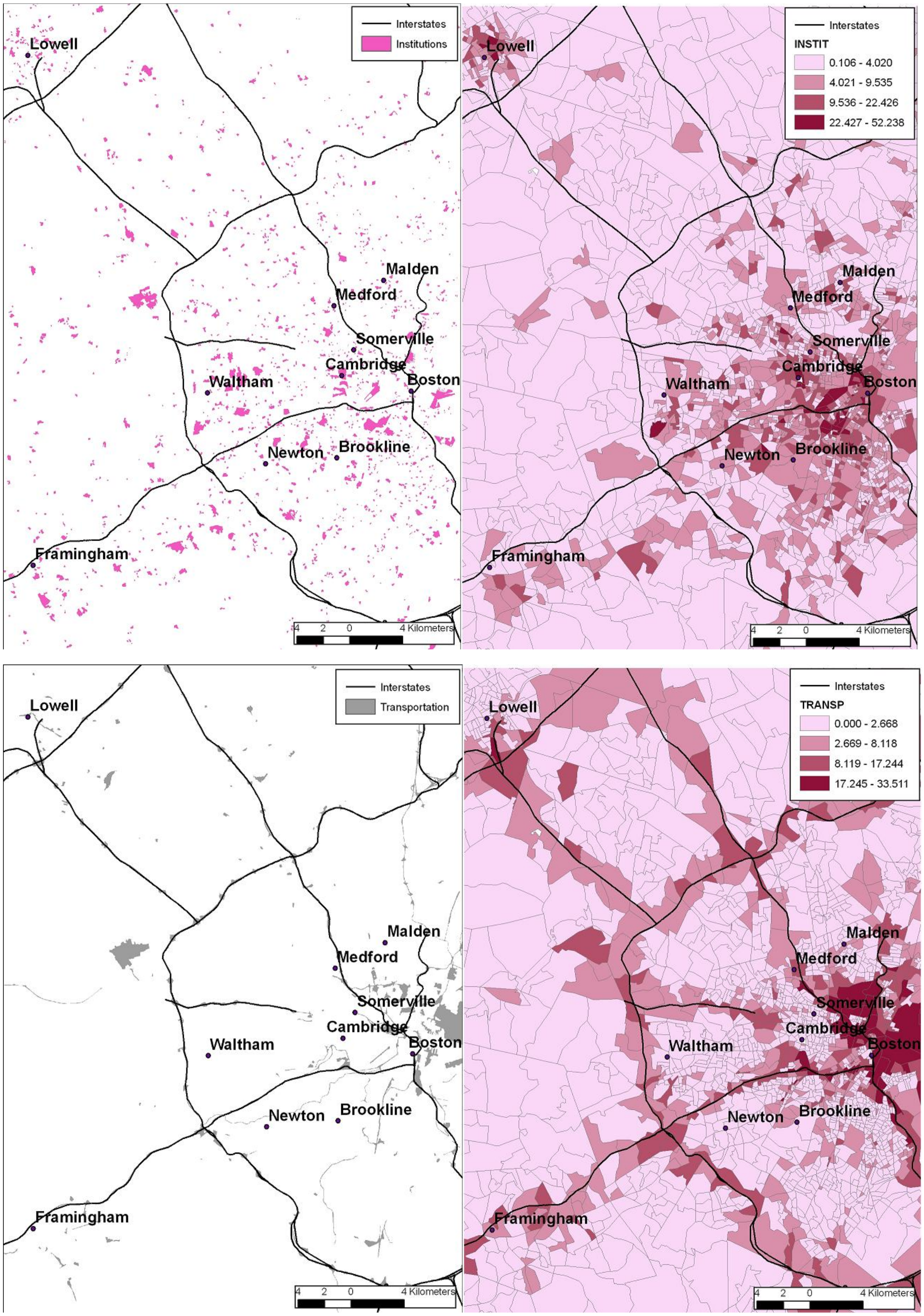


Figure 4-3(a)-(j) continued…

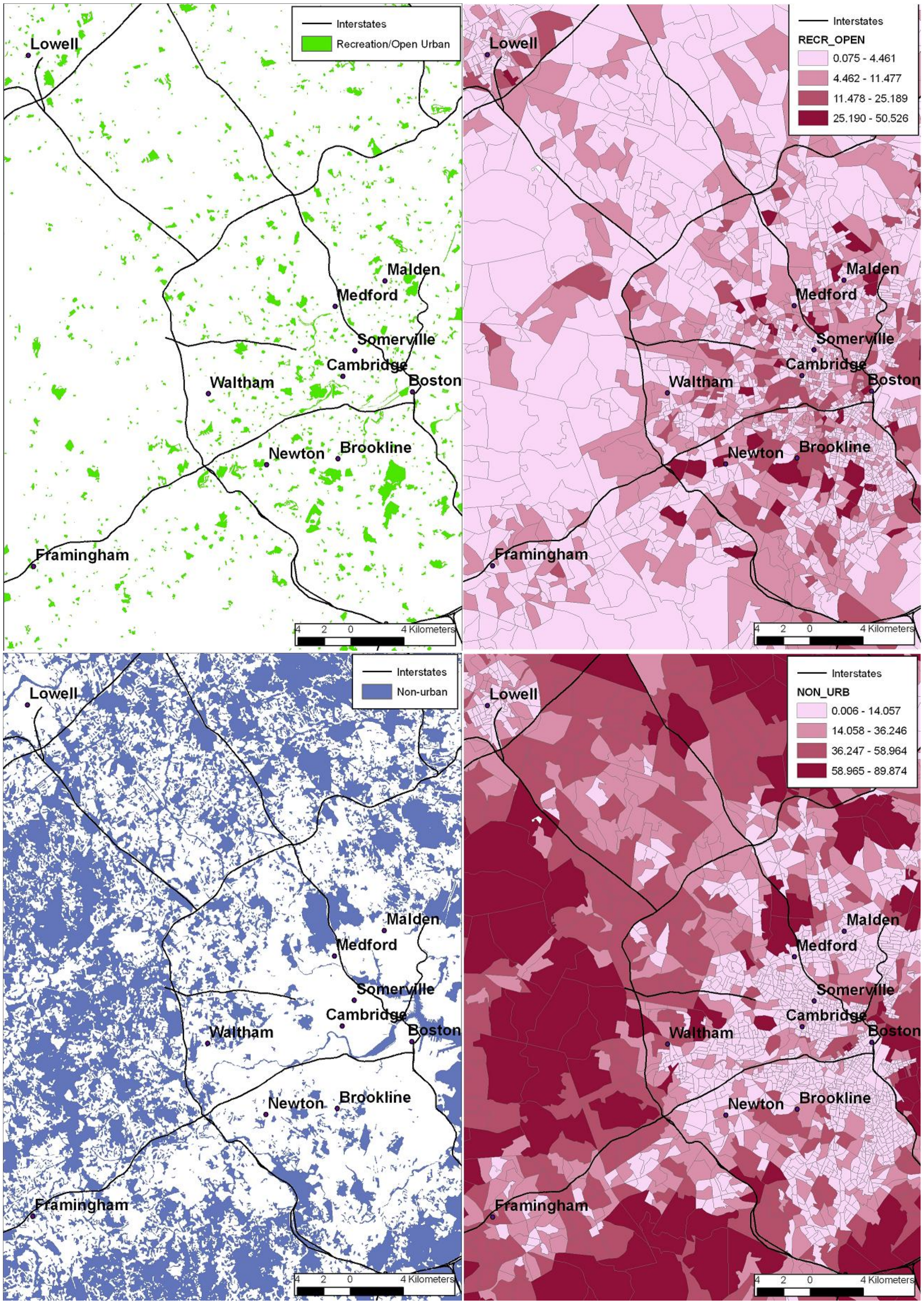


Figure 4-3(a)-(j) – the last

We mapped residuals for each class from both the training and validation data to visually identify patterns of error (Appendix G) and calculated the global Moran I statistic of residuals for both. Although there was statistically significant clustering of residual values for every class ($p < 0.01$), the visual clustering evident from the training residuals was in most cases modest. Additionally, error was not significantly clustered when block groups were evaluated categorically (discussed in section 4.3.2).

Because of the ensemble nature of Random Forests it is not possible to examine a single set of rules for each model. However, as alternative, we took the variables from each final stand-alone model and applied them in the recursive partition tree rpart(), which produces a single set of rules and tree. These tree diagrams give useful information about the general nature of relationships between dependent and independent variables, using a functionality that is similar to Random Forest (albeit with likely lower accuracy – Maindonald and Braun, 2007). The rpart() trees and rules are discussed where pertinent in Section 5.

4.3.2 Integrated (constrained) models

Incorporating predictions from other classes and constraining results to 100% led to small but measurable (above RF variability) improvements in individual models for several classes (SFRES_S and MFRES), but little or no improvement to most others (Figures 4-4(a) and (b)).

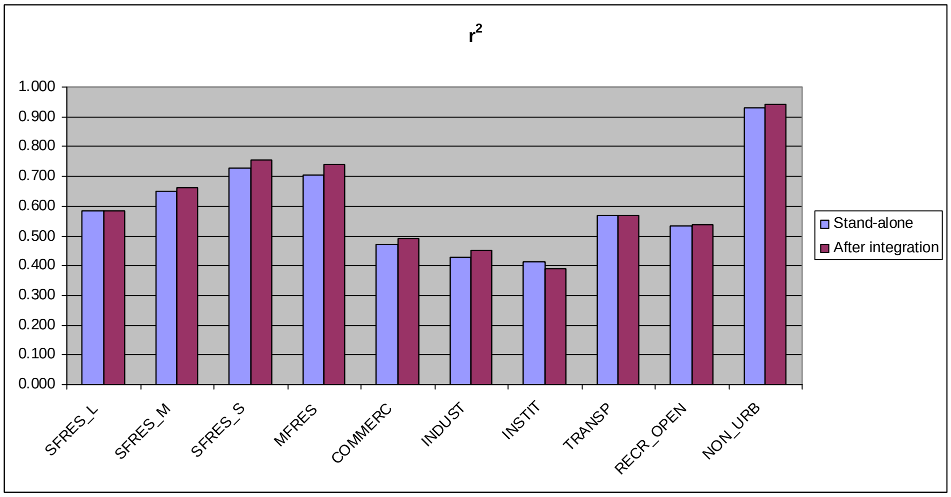


Figure 4-4(a). Change in individual model performance (r^2) after integrating results with other models.

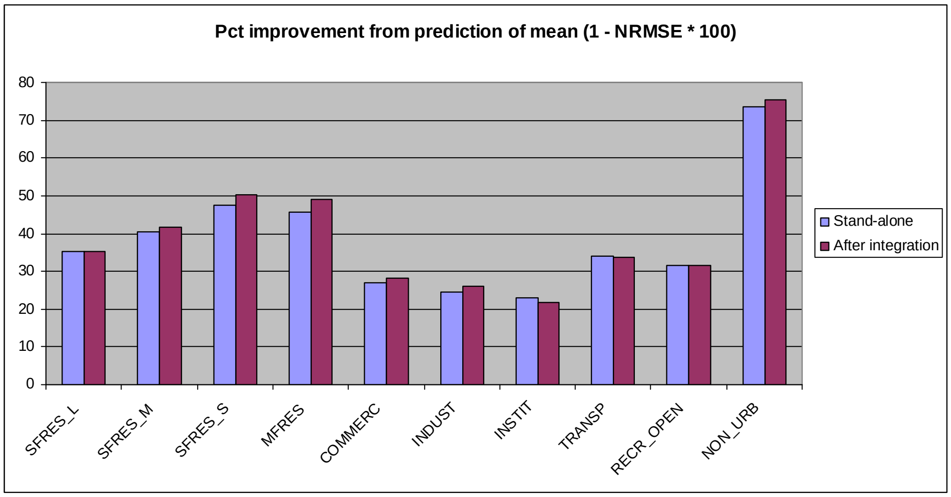


Figure 4-4(b). Change in individual model performance (1 – NRMSE * 100) after integrating results with other models.

We assessed the integrated models, which now summed to 100% land use for each block group, across classes and assigned a majority land use classification (highest percent) to both the actual and predicted land use. A secondary classification (the second highest percent) was also assigned to the predicted land use. Figure 4-5 shows the majority land use in the block group for actual land use (left) and predicted (right).

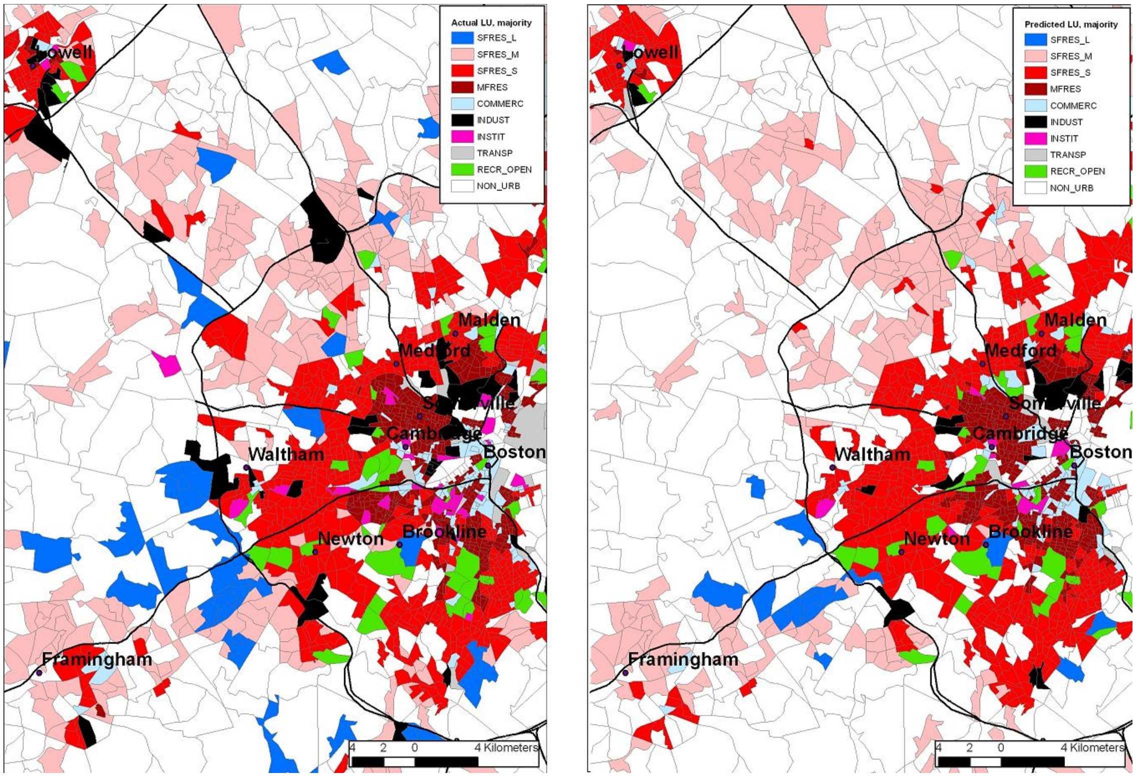


Figure 4-5. Majority actual land use in the block group (left) vs. majority predicted land use. Predicted are based on withheld validation values.

The majority prediction for 79% (2176 / 2764) of all block groups correctly matched the majority actual land use classification. The prediction for 92% of all block groups matched using either the majority or secondary classification (Figure 4-6). There is low spatial autocorrelation of error when assessed in this categorical manner: block

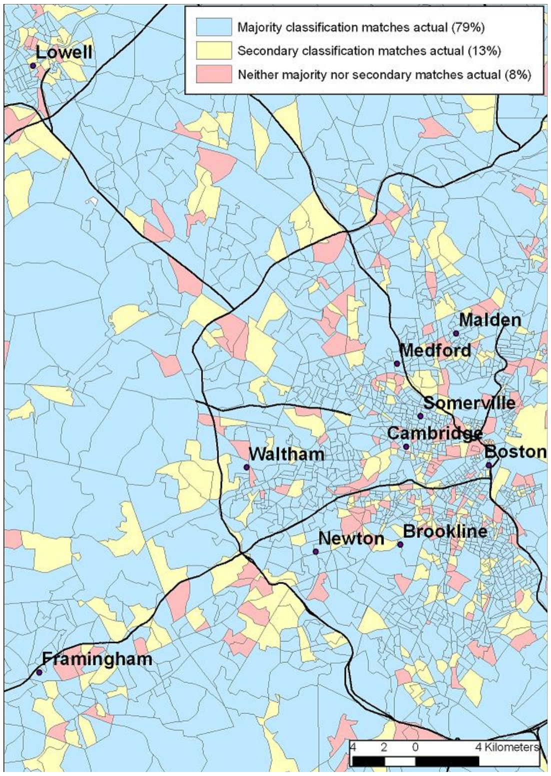

Majority classification matches actual (79%)
Secondary classification matches actual (13%)
Neither majority nor secondary matches actual (8%)
Lowell
Malden
Medford
Somerville
Cambridge
Boston
Waltham
Newton
Brookline
Framingham
4 2 0 4 Kilometers

Figure 4-6. BGs whose predicted majority or secondary classification matched actual land use classification.

groups which did not match the majority land use classification are randomly scattered (Global Moran I = 0.0005, p = 0.17).

We also assessed confidence of predictions by examining the standard deviation of the set of predictions for each class. As noted earlier, we took the matrix of all validation predictions for a class (which for 500 RF trees was 187 predictions), and calculated the standard deviation, standard error and confidence intervals for each record for each class. The sum of the standard errors (Figure 4-7) gives another indication of the overall confidence of predictions: low values indicate block groups where RF made very similar

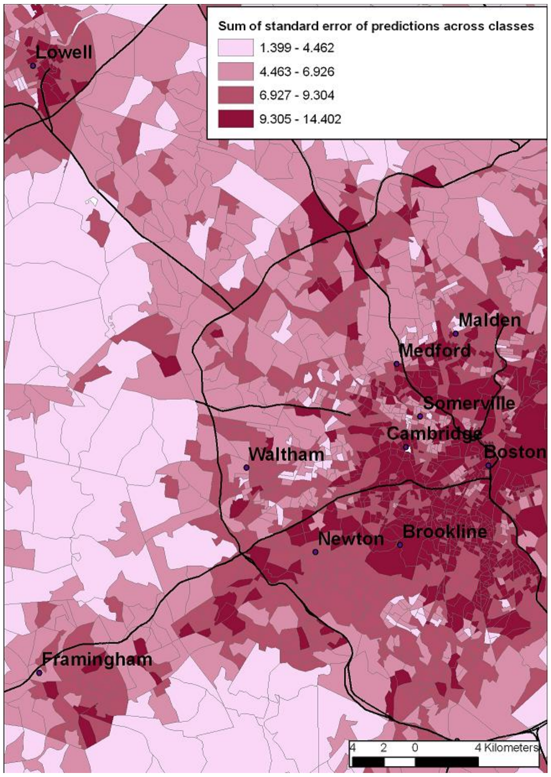


Figure 4-7. Sum of standard error of RF validation predictions across all 10 classes. Categories use Jenks natural breaks breakpoints.

predictions across all trees for most classes, high values indicate the reverse.

4.3.3 Total land area estimation

A key result is how well the method estimates the overall total land area for each class (without regard to error for individual block groups). Comparing overall land area between actual and modeled results is a common practice in land cover/land use product evaluation (Vogelmann et al., 1998). In general one would expect results to be more accurate as aggregated over larger census areas (Wu and Murray, 2005). The predicted percent for each integrated model was multiplied by the actual land area of each block group to give a sum total land area prediction for each class (Table 4-7):

Table 4-7: Comparison of actual land area (sq km) of each class for the entire study area to prediction from integrated models.
Negative differences are underestimation, positive differences are overestimation of land use.

LU class	***actual***	***predicted***	***% difference***
SFRES_L	*559.1*	*556.7*	*-0.4*
SFRES_M	*601.5*	*614.8*	*2.2*
SFRES_S	*286.9*	*305.1*	*6.3*
MFRES	*89.6*	*99.3*	*10.8*
COMMERC	*121.6*	*131.7*	*8.3*
INDUST	*125.1*	*124.4*	*-0.6*
INSTIT	*81.0*	*85.5*	*5.6*
TRANSP	*99.2*	*92.1*	*-7.2*
RECR_OPEN	*164.4*	*171.6*	*4.4*
NON_URB	*2669.0*	*2616.2*	*-2.0*
sum	*4797.4*	*4797.4*	

Figure 4-8 shows this graphically, and additionally shows the calculation of the prediction from the mean for comparison (i.e. if every block group were assumed to have the mean land use percent for that class – our “null model”).

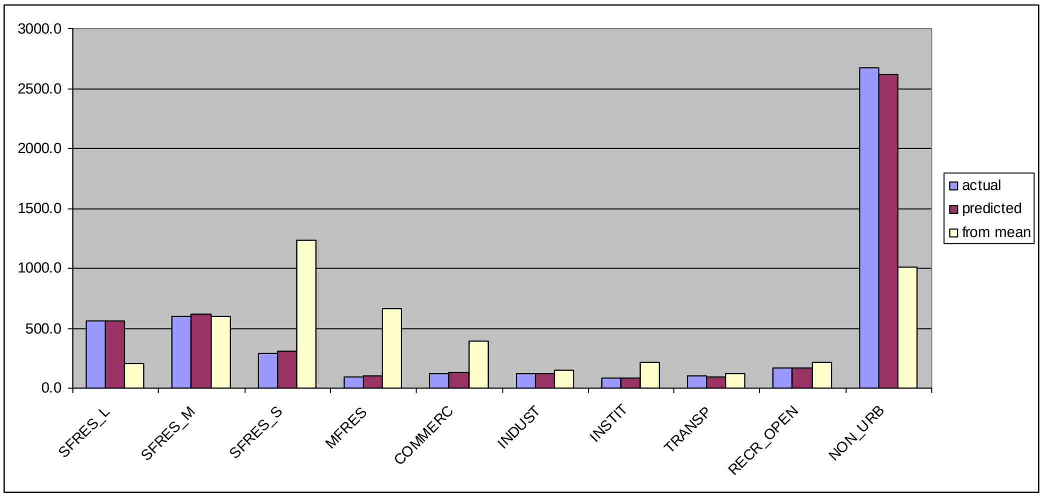


Figure 4-8. Entire study area: comparison of actual land area (sq km) for each class, predicted land area, and prediction from mean (null model).

The land use sums could also be separated out by county (Figures 4-9(a) – (d)): i.e. predictions for individual counties based on models built from the data for the entire study area:

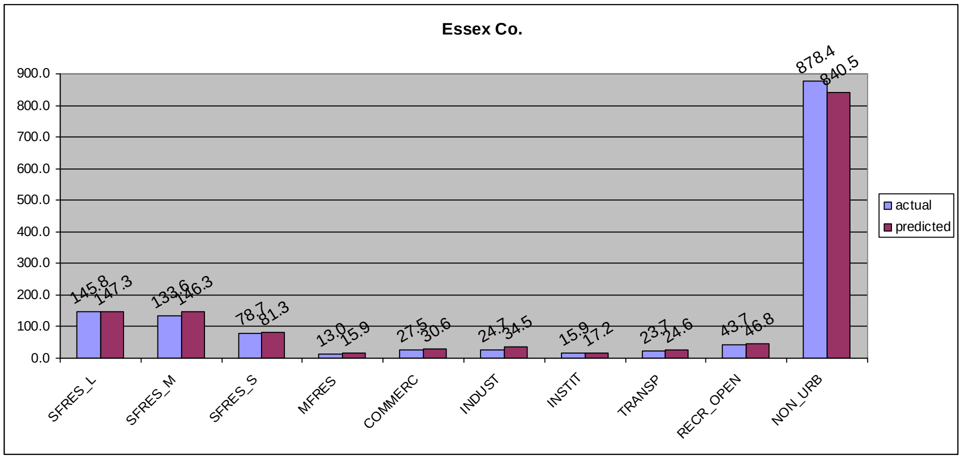


Figure 4-9(a). Essex County: comparison of actual and predicted land area (sq km) for each class.

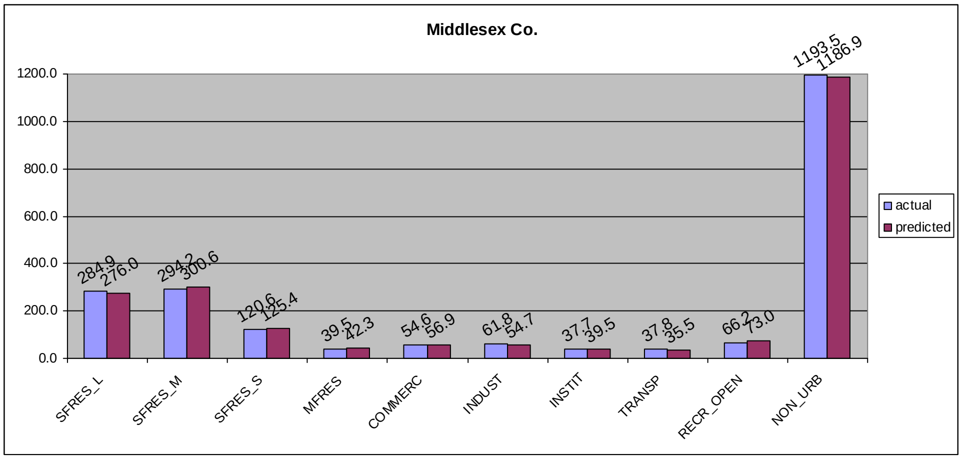


Figure 4-9(b). Middlesex County: comparison of actual and predicted land area (sq km) for each class.

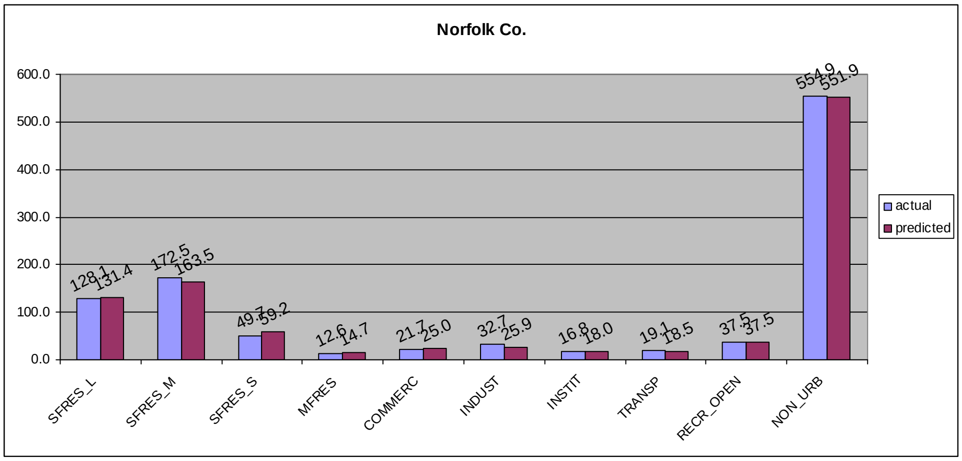


Figure 4-9(c). Norfolk County: comparison of actual and predicted land area (sq km) for each class.

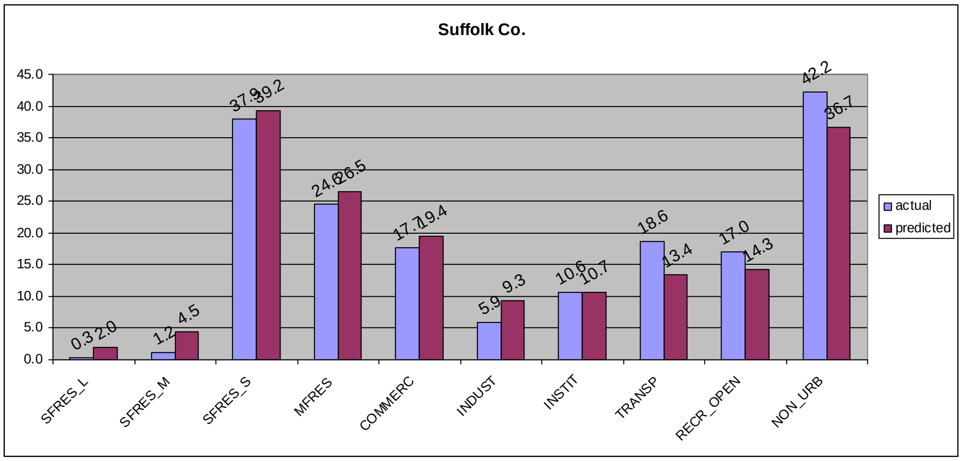


Figure 4-9(d). Suffolk County: comparison of actual and predicted land area (sq km) for each class.

The percent differences for the county comparisons by class are similar to those in Table 4-7, i.e. most differences are within 10%, about half are within 5%, and some are virtually perfect predictions, e.g. RECR_OPEN land area in Norfolk County is an exact prediction of actual land use. It is noteworthy that the predictions for the first three counties have the highest accuracy, but area predictions for Suffolk County (central highly-urbanized Boston) less so. This is not entirely unexpected, as most of the training data comes from the first three counties and they are somewhat dissimilar to Suffolk County. Even given that, the predictions for Suffolk County represent a quite reasonable match to actual data.

4.3.4 Results of Sensitivity Analysis

We tested random samples in which 50, 40, 30, 20, 10, and 5% of all data for the study area were used by RF for model training, and validated these against a separate 10% random sample. Table 4-8 summarizes the results.

Table 4-8: Performance (r^2) vs percent of data used as training.

LU class	**Percent of study area data used as training**					
	50%	**40%**	**30%**	**20%**	**10%**	**5%**
SFRES_L	0.518	0.492	0.490	0.490	0.450	0.416
SFRES_M	0.730	0.711	0.695	0.650	0.614	0.598
SFRES_S	0.811	0.797	0.772	0.750	0.738	0.711
MFRES	0.742	0.726	0.693	0.673	0.648	0.592
COMMERC	0.496	0.494	0.480	0.447	0.416	0.302
INDUST	0.434	0.415	0.416	0.417	0.346	0.269
INSTIT	0.351	0.334	0.329	0.256	0.227	0.178
TRANSP	0.642	0.632	0.627	0.628	0.627	0.592
RECR	0.502	0.490	0.468	0.425	0.418	0.340

These results are discussed in Section 5.

4.3.5 Results of Business Analyst predictors

After including the ESRI Business Analyst predictors in stand-alone models, we evaluated the performance of resulting models for the COMMERC, INDUST, and

INSTIT classes, as noted earlier. The result was a small but measurable increase in performance for those classes (Table 4-9):

Table 4-9: Effect of adding Business Analyst predictors to final models for the COMMERC, INDUST, and INSTIT classes.
As before, all results are from withheld validation records.

	r^2			Pct improvement from prediction of means (1 - NRMSE * 100)	
LU class	**Without BusAnalyst predictors**	**With BusAnalyst predictors**		**Without BusAnalyst predictors**	**With BusAnalyst predictors**
COMMERC	0.471	0.491		27	29
INDUST	0.429	0.463		24	27
INSTIT	0.411	0.434		23	25

4.4 Results of 6-class models

The reason for creating the 6-class models was to be able to test models between urban areas with different physical settings. We had theorized that models built from the Massachusetts dataset should have better performance for areas/datasets that were most similar (e.g. Rhode Island). The feasibility of doing so was unclear prior to testing, as, to our knowledge such a test had not previously been performed with predictive urban land use models, and especially not with national-scale data. The following are the results of validation of the Massachusetts 6-class models.

Table 4-10(a): Results of Boston 6-class model validation.

Column for the area on which the model is built is highlighted. Negative values of pct improvement from prediction of means indicate a fairly poor prediction, i.e. worse than a simple prediction of the mean.

	r^2			
LU class	**MA**	**RI**	**GA**	**CA**
RESID_LOW_6CL	**0.725**	0.304	0.472	0.003
RESID_HIGH_6CL	**0.812**	0.782	0.340	0.325
COM_IND_INST_6CL	**0.630**	0.764	0.751	0.651
TRANSP	**0.568**	0.645	0.573	0.785
RECR_OPEN	**0.535**	0.475	0.303	0.323
NON_URB	**0.930**	0.949	0.501	0.686

Pct improvement from prediction of means (1 - NRMSE * 100)			
MA	**RI**	**GA**	**CA**
47	-35	6	-32
57	36	17	12
39	51	33	40
34	40	34	52
32	18	13	7
74	73	-43	42

With the exception of the RESID_LOW_6CL, the Providence dataset (RI) came closer to matching the performance of the Massachusetts models than the Atlanta (GA) or Los Angeles (CA) datasets. These are discussed in detail in Section 5. The next obvious question was how would models created from those areas perform, both in their own areas, and against the others? These are summarized in Tables 4-10(b) – (d), which are identical in format to Table 4-10(a), but for models created from data from the Providence, Atlanta, and Los Angeles data, respectively.

Table 4-10(b): Results of Providence model validation.

	r^2			
LU class	MA	**RI**	**GA**	**CA**
RESID_LOW_6CL	0.435	**0.594**	0.389	0.003
RESID_HIGH_6CL	0.691	**0.885**	0.232	0.393
COM_IND_INST_6CL	0.490	**0.743**	0.707	0.576
TRANSP	0.471	**0.683**	0.499	0.736
RECR_OPEN	0.469	**0.529**	0.235	0.413
NON_URB	0.926	**0.952**	0.401	0.742

Pct improvement from prediction of means (1 - NRMSE * 100)			
MA	**RI**	**GA**	**CA**
10	**36**	-24	-32
42	**66**	8	11
28	**49**	31	32
27	**44**	28	46
26	**31**	9	19
68	**78**	-91	45

Table 4-10(c): Results of Atlanta model validation.

	r^2				Pct improvement from prediction of means (1 - NRMSE * 100)			
LU class	**MA**	**RI**	**GA**	**CA**	**MA**	**RI**	**GA**	**CA**
RESID_LOW_6CL	0.395	0.331	**0.756**	0.028	20	-87	**52**	-23
RESID_HIGH_6CL	0.612	0.564	**0.639**	0.140	8	-6	**40**	3
COM_IND_INST_6CL	0.302	0.479	**0.819**	0.237	-51	-63	**58**	-21
TRANSP	0.423	0.597	**0.572**	0.770	21	32	**33**	24
RECR_OPEN	0.256	0.289	**0.511**	0.393	14	13	**30**	18
NON_URB	0.880	0.889	**0.735**	0.315	47	37	**48**	14

Table 4-10(d): Results of Los Angeles model validation.

	r^2				Pct improvement from prediction of means (1 - NRMSE * 100)			
LU class	**MA**	**RI**	**GA**	**CA**	**MA**	**RI**	**GA**	**CA**
RESID_LOW_6CL	0.002	0.007	0.109	**0.689**	-30	-78	-18	**44**
RESID_HIGH_6CL	0.575	0.835	0.207	**0.782**	28	38	5	**53**
COM_IND_INST_6CL	0.442	0.579	0.696	**0.826**	20	17	42	**58**
TRANSP	0.128	0.335	0.227	**0.853**	-13	4	-2	**62**
RECR_OPEN	0.182	0.218	0.270	**0.571**	-19	-39	-23	**35**
NON_URB	0.800	0.846	0.395	**0.763**	27	18	-11	**51**

There are obvious difficulties in predicting land use from models of urban areas with different characteristics. For example, while predicting low intensity residential land (RESID_LOW_6CL) in Los Angeles is fairly successful from Los Angeles data (r^2 = 0.689, pct improvement = 44), transporting that model to the other areas is remarkably unsuccessful, and in every case the accuracy is worse than simply predicting from the mean. The characteristics of low intensity residential land are clearly different in the Los Angeles data sample than in the other areas. These results are discussed in more detail in Section 5.

Similar to Table 4-6, the variables that were built in final models for each area are summarized in Tables 4-11(a)-(f). As noted earlier, the process for model building was identical to that described for the 10-class Boston models, i.e. the goal was a model with a small number of variables which had as good a performance as possible for the withheld validation data for the area from which it was built.

Table 4-11(a): RESID_LOW_6CL.
Variables that were part of final models for each area for RESID_LOW_6CL. "X" indicates variable was built as part of that model. Variables are given in alphabetical order, thus organized by predictor class.

Variable	MA	RI	GA	CA
CENS_hu_median_numb_rooms		X		X
CENS_huden				X
CENS_median_hh_income				X
CENS_pct_5_or_more_units_in_structure				X
CENS_pct_hu_owneroccupied			X	
CENS_pden_change90_00		X		
HIST_nlcd92_21	X			
HIST_sum_nlcd92_allveg		X		
LC_nlcd01_21			X	
LC_nlcd01_23			X	
LC_nlcd01_imperv_mean		X	X	
LC_ratio_popden_nlcd2324				X
LC_sum_nlcd01_2122	X	X		
MISC_vg2000_mean	X			X
PROX_city100k_dist		X	X	X
PROX_city250k_dist	X			
PROX_cost_50k_city	X			
PROX_mean_dist_road		X	X	
SA_localMoran_popden	X			
SPCAT_lc_entropy			X	
SPCON_focal33_gt50_is_std		X		
TRANSP_ratio_roadden_imperv			X	

Table 4-11(b): RESID_HIGH_6CL.
Variables that were part of final models for each area for RESID_HIGH_6CL.

Variable	MA	RI	GA	CA
CENS_hu_median_numb_rooms	X			
CENS_huden			X	X
CENS_median_hh_income		X		
CENS_pct_5_or_more_units_in_structure			X	X
CENS_pct_hu_owneroccupied				X
CENS_popden	X			
HIST_giras_11		X		
HIST_highresid_92_and_01		X		
HIST_highresid_all_times		X		X
HIST_nlcd92_21		X		
HIST_nlcd92_22			X	
LC_nlcd01_22			X	
LC_nlcd01_23	X	X		X
LC_ratio_huden_imperv	X			X
PROX_city100k_dist			X	
PROX_city250k_dist	X			
PROX_cost_10k_city				X
PROX_cost_50k_city	X			
PROX_mean_dist_road		X		
SA_localMoran_popden	X		X	
SPCON_focal33_gt50_is_std				X

Table 4-11(c): COM_IND_INST_6CL.
Variables that were part of final models for each area for COM_IND_INST_6CL.

Variable	MA	RI	GA	CA
CENS_hu_median_numb_rooms	X			
CENS_pct_hu_owneroccupied		X		
HIST_commerc_all_times		X	X	
HIST_highresid_all_times				X
HIST_nlcd92_21		X		
HIST_nlcd92_23	X	X		
HIST_sum_giras_comm_ind	X	X		X
LANDMRK_gnisconsol_instit_density	X	X		
LC_nlcd01_23			X	
LC_nlcd01_24	X			X
LC_nlcd01_imperv_mean			X	
LC_nlcd01_imperv_range				X
LC_nlcd01_imperv_stdev			X	
LC_ratio_huden_imperv			X	X
MISC_vg2000_mean				X
SPCON_focal33_gt50_is_std		X	X	

Variable	MA	RI	GA	CA
SPCON_focal77_gt50_is_cv	X			
TRANSP_bts_rail_pct	X			
TRANSP_ratio_roadden_imperv	X		X	

Table 4-11(d): TRANSP.
Variables that were part of final models for each area for TRANSP.

Variable	MA	RI	GA	CA
HIST_giras_14	X	X	X	X
HIST_indust_all_times	X	X	X	
HIST_nlcd92_23		X	X	
LMiZScore_alltrans		X		X
LMiZScore_bgpopden	X			
PROX_city100k_dist				X
PROX_city250k_dist				X
PROX_mean_dist_road	X			
TRANSP_a11_a17_roads_density	X	X	X	X
TRANSP_a11_a38_roads_density	X			X
TRANSP_alltrans	X	X	X	X
TRANSP_bts_faf2_pct			X	
TRANSP_bts_rail_pct	X	X		
TRANSP_ratio_roadden_imperv	X			

Table 4-11(e): RECR_OPEN.
Variables that were part of final models for each area for RECR_OPEN.

Variable	MA	RI	GA	CA
HIST_giras_17	X	X		X
HIST_nlcd92_11				X
HIST_nlcd92_85	X	X	X	X
HIST_recr_all_times		X	X	X
LANDMRK_gnisconsol_recr_density	X	X	X	X
LC_nlcd01_21			X	
LC_nlcd01_22	X			
LC_sum_nlcd01_ag			X	
PROX_city250k_dist	X		X	
PROX_mean_dist_road			X	X
SA_localMoran_allnatveg		X		
SA_localMoran_lc_entropy	X			
SPCAT_lc_entropy		X		
TRANSP_ratio_roadden_imperv	X			

Table 4-11(f): NON_URB.
Variables that were part of final models for each area for NON_URB. Note that in RI the "model" consists of a single variable.

Variable	MA	RI	GA	CA
CENS_median_hh_income			X	
CENS_pct_pop_below_poverty_lev			X	
LC_nlcd01_21			X	
LC_sum_nlcd01_allnatveg			X	
LC_sum_nlcd01_urban	X	X	X	X
MISC_ned30m_slope				X
PROX_mean_dist_road	X		X	
SA_localMoran_allnatveg				X

What do the results look like? Figure 4-10(a)-(d) show mappings of the RESID_HIGH_6CL class for each area as example, from the model for that area. As in the Figure 4-3 series, these also show a series of panels juxtaposing the actual land use (left side) with the modeled land use (right side) as predicted from the withheld validation records. In this series, because the same class is mapped in each figure, the categories on the right-hand side maps use the same breakpoints in each figure.

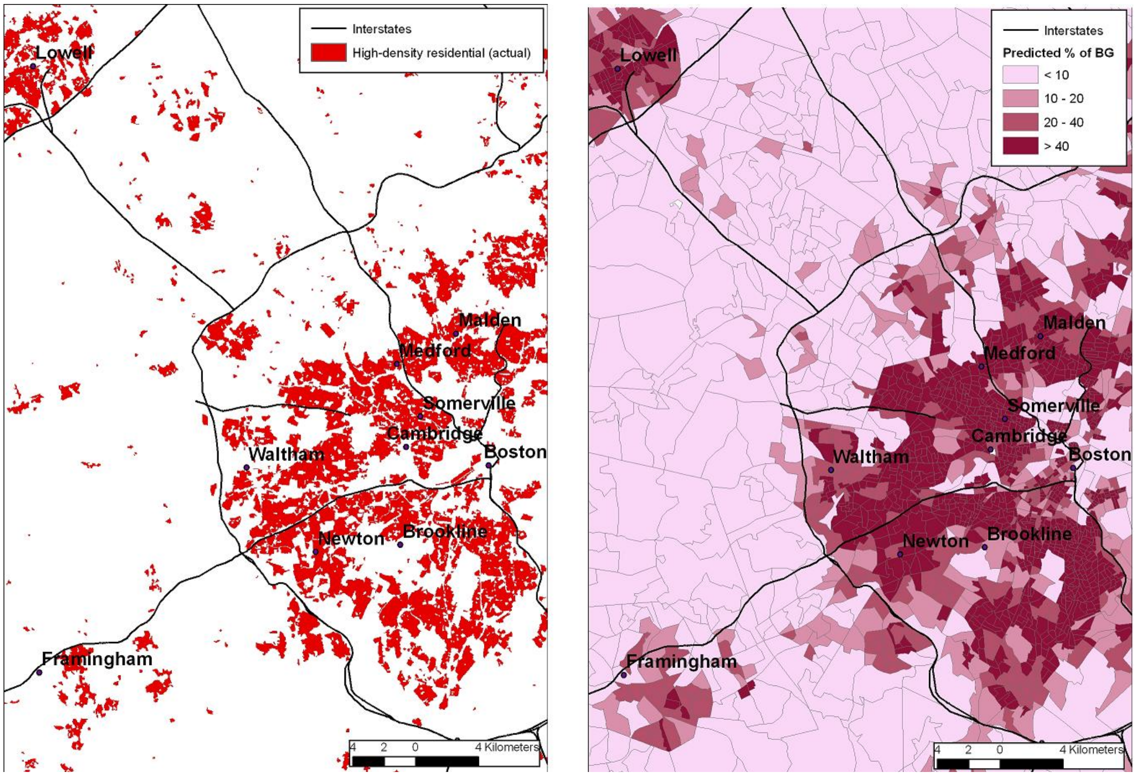


Figure 4-10(a). Boston area. Actual (left) vs. predicted (right) land use for RESID_HIGH_6CL.

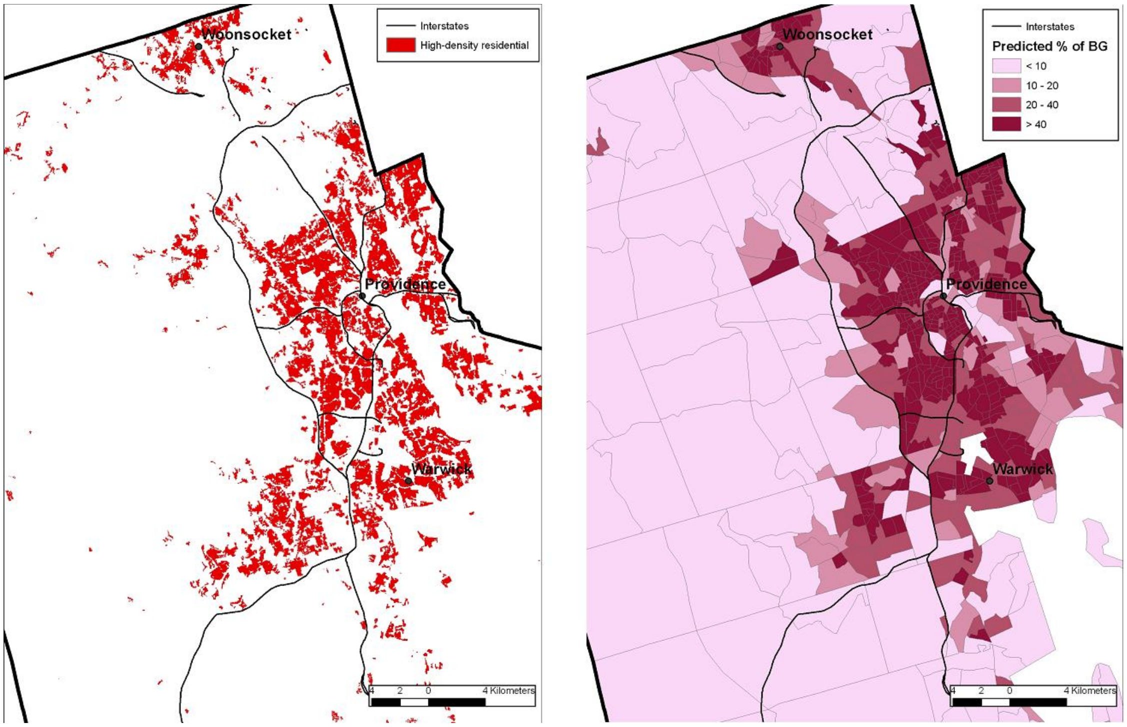


Figure 4-10(b). Providence area. Actual (left) vs. predicted (right) land use for RESID_HIGH_6CL.

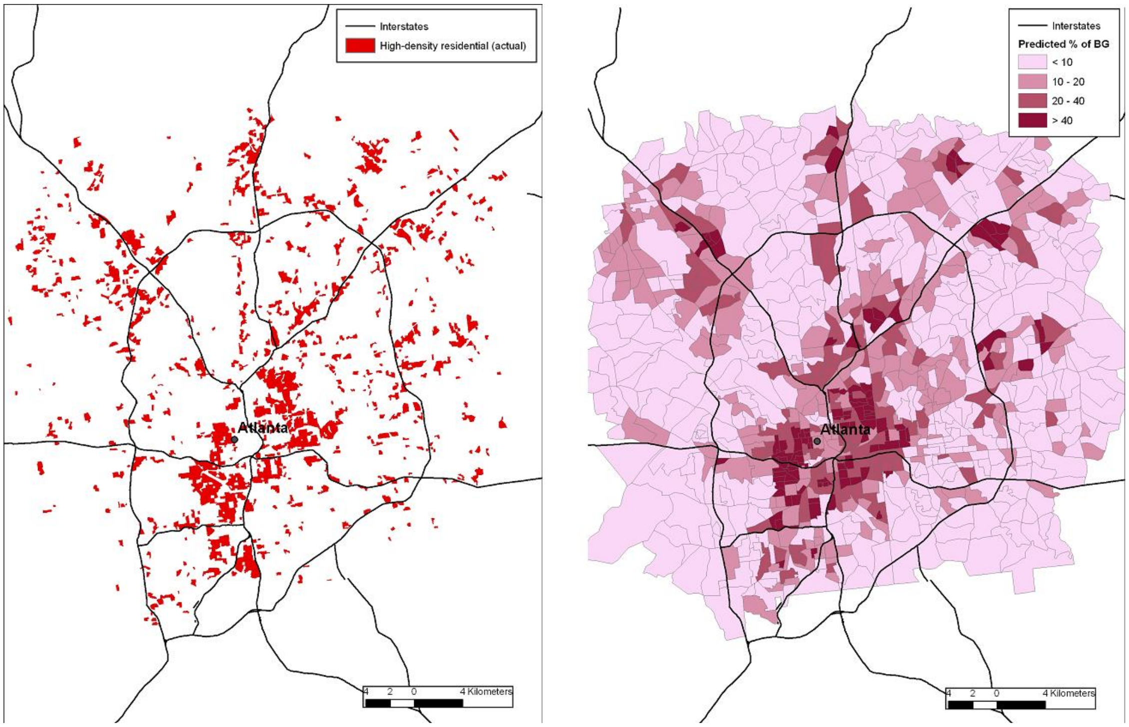


Figure 4-10(c). Atlanta area.
Actual (left) vs. predicted (right) land use for RESID_HIGH_6CL. Actual land use clipped to study block groups' extent for clarity.

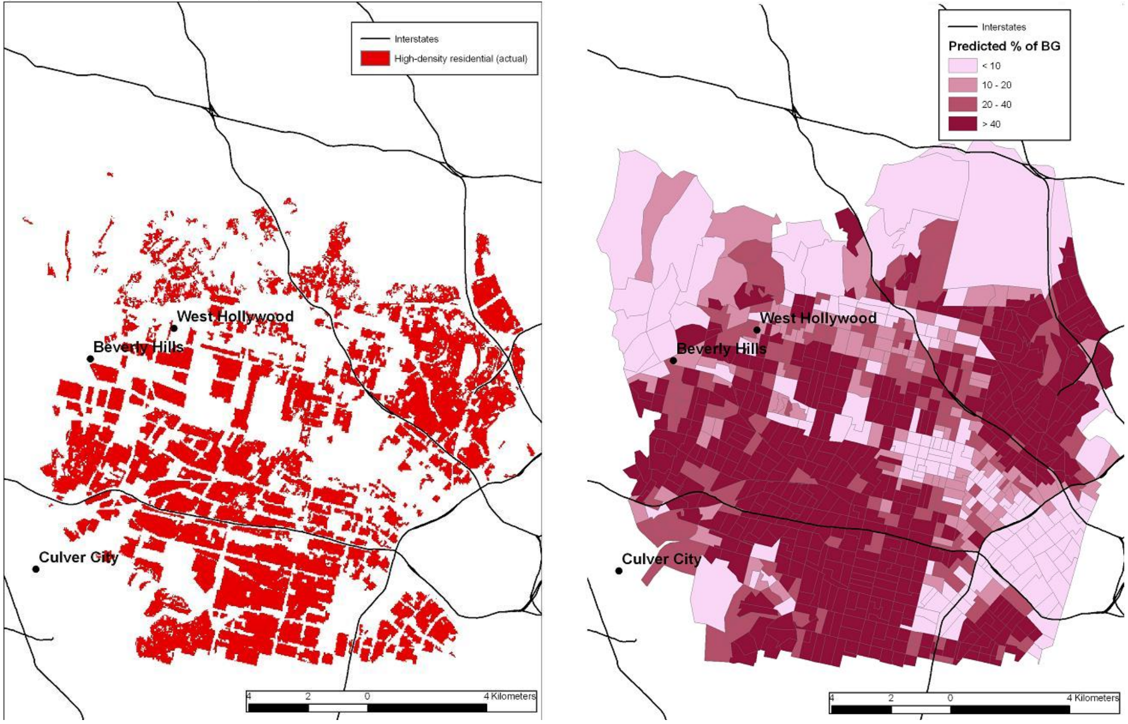


Figure 4-10(d). Los Angeles area. Actual (left) vs. predicted (right) land use for RESID_HIGH_6CL. Actual land use clipped to study block groups' extent for clarity.

5. Discussion

5.1 10-class models

5.1.1 Stand-alone models

Very broad scale drivers of land use: climate, topography, geology or access to water supply, were not major considerations in our analysis of the 10-class Boston models. Except for the Boston metropolis being limited from expansion to the east, we considered the relatively minor variations in those factors to be negligible in our study area. However, classical land use theories suggest that several broad elements of the built environment should be important in driving the pattern of the landscape. These are proximity influences of cities of various sizes, accessibility (transportation), and agglomeration (spatial autocorrelation). Those elements were also key predictors in this study. Examining the importance scores in Appendix F, one readily notes the greater importance of Proximity (PROX_) and Spatial Autocorrelation (SA_) variables across models and higher rankings within models.

From the importance scores of the 188 predictor variables in Appendix F, we summarized the overall importance of variables across classes, assuming equal weight to each of the 10 classes, in two ways. The first of these was to assign a score of 50 to any cell with a dash (i.e. a low score: those were variables we considered to be noise), then averaging the scores across all 10 classes. The top ranked of these are given in Table 5-1 (left side). The second way we measured importance across all classes was to count the number of times (out of 10 classes) that the variable ranked in the top 20 predictors. Those that were in the top ranked predictors in three or more classes are given on the right side of Table 5-1. The prevalence of Proximity and SA variables is clear: from the first method, the top three variables were measures of Proximity to cities or roads, and four of the top 10 were additionally measures of SA. They were also highly placed with the second method. The two most important variables, by far, from both methods were consistent: the most important predictor across classes was distance to nearest 250k city (Boston center), followed by distance to nearest road. Other classes of predictors were useful across a smaller range of land use types, but still regularly or specifically important (CENS_, LC_, TRANSP_, HIST_). Our measures of spatial pattern from 30-m data (SPCAT_ and SPCON_) were much less useful across classes of land use type.

Table 5-1: Summary ranking of predictor variable importance across land use classes, of 188 total variables.

The left hand side shows those variables which had the highest average rank order of importance from all 10 classes (as given in Appendix F). The right hand side shows the number of times (out of 10 classes) that the variable was in the top 20 predictors.

Variable	Overall rank across classes
PROX_city250k_dist	1
PROX_mean_dist_road	2
PROX_city100k_dist	3
TRANSP_ratio_roadden_imperv	4
SA_localMoran_allnatveg	5
MISC_vg2000_mean	6
SA_localMoran_popden	7
SA_localMoran_dist_road	8
CENS_huden	9
SA_localMoran_medianrooms	10
HIST_highresid_all_times	11
LC_ratio_huden_imperv	12
CENS_pct_hu_owneroccupied	13
PROX_interstate_road_dist	14
LC_ratio_popden_nlcd2324	15
LC_nlcd01_imperv_stdev	16
HIST_indust_all_times	17
CENS_pct_5_or_more_units_in_structure	18
PROX_cost_gnis_instit	19
SA_localMoran_imperv	20
HIST_nlcd92_21	21
CENS_hu_median_numb_rooms	22
HIST_nlcd92_23	23
PROX_cost_10k_city	24
SA_localMoran_nlcd2324	25

Variable	Num times in top 20
PROX_city250k_dist	8
PROX_mean_dist_road	7
CENS_hu_median_numb_rooms	4
HIST_indust_all_times	4
LC_nlcd01_imperv_stdev	4
LC_ratio_huden_imperv	4
PROX_city100k_dist	4
SA_localMoran_allnatveg	4
SA_localMoran_popden	4
TRANSP_ratio_roadden_imperv	4
CENS_huden	3
CENS_pct_hu_owneroccupied	3
CENS_popden	3
HIST_commerc_all_times	3
HIST_giras_11	3
HIST_highresid_all_times	3
HIST_nlcd92_23	3
PROX_cost_10k_city	3
PROX_cost_50k_city	3
PROX_cost_gnis_instit	3
PROX_interstate_road_dist	3
SA_localMoran_dist_road	3
SA_localMoran_imperv	3
SA_localMoran_medianrooms	3
SA_localMoran_nlcd2324	3
SPCON_is_slope_max	3
TRANSP_bts_faf2_pct	3
TRANSP_bts_rail_pct	3

There was value in characterizing some elements of the landscape both within a block group and its spatial autocorrelation over a broader area. For example, population density (*CENS_popden*) was a highly ranked predictor for three of the residential classes (see Appendix F: ranked #17 for SFRES_M, #10 for SFRES_S and #3 for MFRES).

However, the local Moran z-score value for population density (*SA_localMoran_popden*), which was calculated over a broader area than the block group (calculation was based on inverse distance for the entire study area), was an even stronger predictor for those same classes (ranked #6, #1, and #1, respectively). This would suggest a landscape-wide effect of population density stronger than a local effect. The overall conclusion from this, however, is the near-universal strength of measures of proximity to cities, roads and other landmarks, and the autocorrelation of some characteristics (vegetation, population density), in predicting urban land use. This strongly reinforces the argument against a per-pixel only attempt to measure land use, and the importance of accounting for distance and clustering effects.

It is highly noteworthy that in predicting these thematically detailed urban land use classes, current land cover was only occasionally one of the strongest predictors, and in no instance were models comprised primarily of land cover variables alone. Four of the 10 final models did not include a LC_ predictor. Another test we did to look at this was to create models based on current land cover variables alone (those in Appendix B2). Table 5-2 show the results for the urban classes, and confirm that predicting most land uses from land cover alone gives poor results. The results of this project overall strongly support the notion that land cover may only give clues to land use, and that the one is not directly mappable from the other.

Table 5-2: Comparison of performance of final models with models built from only current land cover variables.

	r^2 - final model	r^2 - only LC_ predictors
SFRES_L	0.582	0.505
SFRES_M	0.648	0.441
SFRES_S	0.727	0.373
MFRES	0.706	0.323
COMMERC	0.471	0.299
INDUST	0.429	0.255
INSTIT	0.411	0.021
TRANSP	0.568	0.194
RECR_OPEN	0.535	0.241

Examining the results in Table 4-5, the primary measures of performance of the 10-class models, the three classes other than NON_URB which had the best performance were the three highest density residential classes: SFRES_S (r^2 = 0.727), MFRES (r^2 = 0.706) and SFRES_M (r^2 = 0.648). These were also the classes with the highest degree of spatial autocorrelation and clustering into zones (see Figures 3-9 and 3-10). Indeed, the degree of clustering of any class was a reasonably good indicator of how well it could be predicted: those land use classes that were the most dispersed were the most difficult to predict (Figure 5-1).

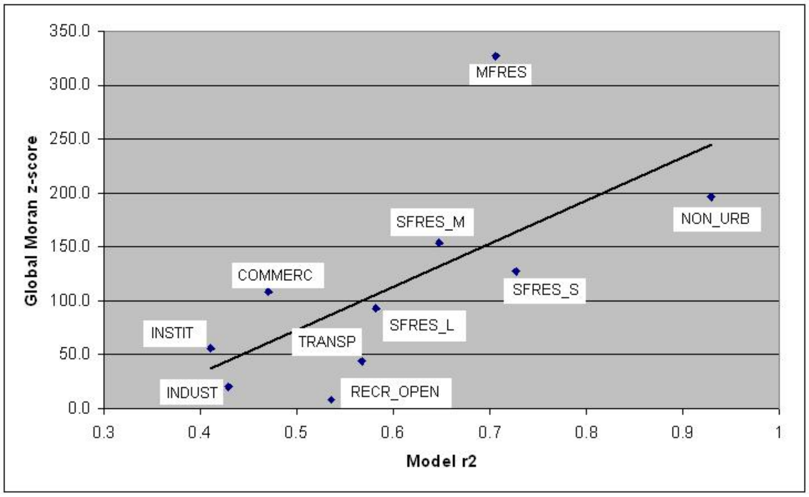


Figure 5-1. Class clustering vs. ability to predict.
Y-axis are Global Moran z-scores (same values as shown in Figure 3-11); X-axis is final model r^2 value. Land use types with higher clustering were easier to predict than dispersed land use types.

Classes which percentagewise were relatively sparse were also more difficult to predict, with the notable exception of MFRES. That is, the classes that were most difficult to predict - INSTIT, INDUST, COMMERC, RECR_OPEN and TRANSP – were also the land uses with the smallest percentages in three of the four study counties (Essex, Middlesex, and Norfolk), although less so in the central Boston county of Suffolk. This, however, corresponded to what we expected – that detecting very small percentages of a land use in a BG would be difficult. The clustered nature of MFRES land made it the exception, i.e. although it made up a small proportion of the entire area, it was centralized in certain areas, and was therefore easier to detect..

Another general factor affecting performance was variability within a class with regard to various predictors. Classes with great variability were more difficult to predict generally. Of the land cover variables, impervious surfaces were very important because they formed the basis for a number of predictor variables. Several classes, notably INSTIT and TRANSP had larger variability of imperviousness within the class (e.g. the large range and interquartile distributions seen in first figure of Appendix D), which made them more difficult to identify.

There were, however, unique aspects of each class and model. A brief evaluation of the prediction of each of the 10 classes is given here:

5.1.1.1 Single-family large lot residential (SFRES_L)

These were the reference polygons with the most scattered urbanization of the urban classes and were recognizable primarily from three measures: a) spatial pattern of urban surfaces, b) low general levels of urbanization, and c) the median number of rooms per home. This was one of only two classes where spatial pattern metrics made it into the final model (the other being INSTIT). The strongest variable in the model (Appendix E) was *SPCON_ed_slopeclass1* – the edge density of the lowest imperviousness slope pixels; i.e. areas where imperviousness was present but of low contrast to its surroundings. The other important spatial pattern metric was *SPCAT_shape_index_22_24*: an index of the shape of high intensity urban pixels: high values indicating more compact shape. The single-family large lot areas represented the opposite of this, and were characterized by very irregular or un-compact shapes of urbanization (secondary roads and scattered or discontiguous housing areas). To interpret the meaning of these variables and help visualize roughly how models were built, the rpart() version of the final model is useful to examine (Figure 5-2): low values of the shape index (branching to the right hand side of tree), high values of the edge density measure, or high values of median number of rooms led to prediction of high values of SFRES_L.

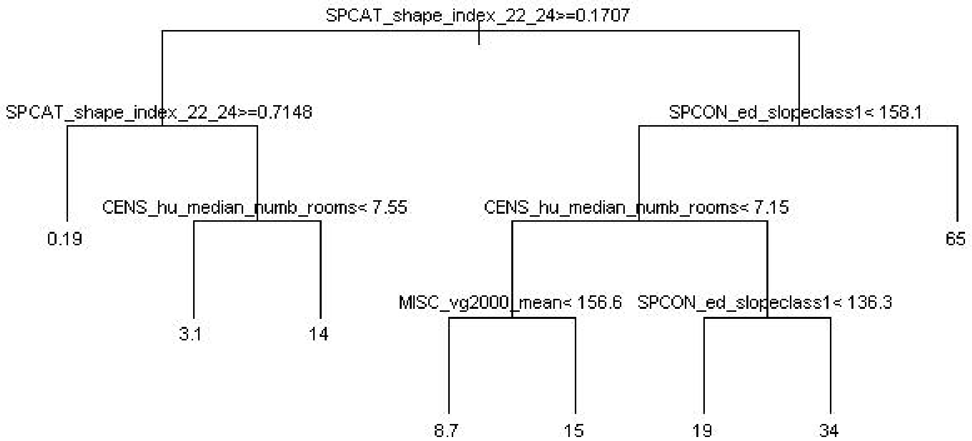

SPCAT_shape_index_22_24>=0.1707
SPCAT_shape_index_22_24>=0.7148
SPCON_ed_slopeclass1< 158.1
0.19
CENS_hu_median_numb_rooms< 7.55
CENS_hu_median_numb_rooms< 7.15
65
MISC_vg2000_mean< 156.6
SPCON_ed_slopeclass1< 136.3
3.1
14
8.7
15
19
34

Figure 5-2. rpart() version of final model for the SFRES_L class.
Highest predicted values are given from the rules leading to the right-hand side of the tree.

These variables, combined with measures of imperviousness or vegetation, and distance to nearest road comprised the final model. The performance of this model was the poorest of the four residential classes (r^2 = 0.582, NRMSE = 0.65), although is better than the non-residential classes and compares well to results from those few studies we can find that are similar. For example, Hardin et al. (2008) use a similar zone-based approach to model block level housing unit density with r^2 validation result of 0.624 with a priori removal of outliers, and 0.370 without removal of outliers (we did not remove outliers). Although it is not advisable to directly compare results of studies which have somewhat different goals and methods (on the one hand, block level prediction is likely to be more difficult than block-group level prediction, but on the other hand the Hardin study has the advantage of using Landsat image data, which we have eschewed as being too unwieldy at the national scale); nonetheless it helps put our results in context.

It is noteworthy that distance to largest city center (*PROX_250k_dist*), although as noted above was almost universally important in other models, was not an especially strong predictor for this class. This was contrary to expectation, however, we attribute this to the existence of numerous block groups which were also distant from the center city but had no or little presence of SFRES_L (i.e. either completely non-urban areas, or small towns/cities which had other types of residential housing), and that other variables such as distance to nearest road better captured the basic concept of lack of centrality for this class.

5.1.1.2 Single-family medium lot residential (SFRES_M)

This was the most common urban land overall in the study area (slightly more so than SFRES_L). Many of the major predictors for this class were measures of proximity or accessibility of the block group to various urban centers (*PROX_city250k_dist*, *PROX_cost_50k_dist*, *PROX_city100k_dist*), or to roads (*PROX_mean_dist_road* and *PROX_airport_crossing_dist*). These variables, combined with measures of low-medium intensity urbanization and population density, comprised the final model. The NLCD01 classes 21 and 22 (low-medium imperviousness) were a very common land cover coding in this class, and therefore a number of metrics derived from those classes were also reasonably strong predictors, even if they did not end up in the final model. The performance of this model was an improvement from the SFRES_L class (r^2 = 0.648, NRMSE = 0.59). The proximity effects from both smaller and larger population centers were evident in this land use.

5.1.1.3 Single-family small lot residential (SFRES_S)

The predictors for this land use class were quite similar to the SFRES_M class, however the spatial extent of the class was more concentrated in a regular proximity to city centers (see Figure 4-3): most of these high-density housing areas are in a ring surrounding central Boston or in close proximity to smaller cities. This allowed predictor variables which characterized proximity, as above, to be effective. In both the SFRES_M and SFRES_S classes the variable *PROX_city250k_dist* was an especially strong predictor (see Appendix E). Two of our measures of spatial autocorrelation (*SA_localMoran_popden* and *SA_localMoran_allnatveg*) were significant contributors to the model. It is also notable that, even though the variable *CENS_hu_median_numb_rooms* had a non-significant linear correlation to this class (see Appendix C), it was a valuable contributor to the final model. This reinforces again the non-linear relationships of some if not most variables. Historical land cover information was also useful in both these classes in the form of the *HIST_nlcd92_21 variable* ("low intensity residential" from 1992). The performance of this model was very successful and was the best of all of the urban classes (r^2 = 0.727, NRMSE = 0.52).

5.1.1.4 Multi-family residential (MFRES)

The predictors for this land use class were likewise similar to the two previous classes, and the spatial extent of the class was even more concentrated than the SFRES_S class (see Figure 4-3): most of these very high-density housing areas are in a tight 8 km ring surrounding central Boston. These lands were by far the most clustered of the land use classes, as noted above and in Figure 3-11. This allowed predictor variables which characterized proximity and spatial autocorrelation, as above, to be effective. The variable which characterized clustering of population density (*SA_localMoran_popden*) was by far the strongest predictor (Appendix E). The census variable *CENS_hu_median_numb_rooms* was also again a contributor to this model. The performance of this model was the second highest of the urban classes (r^2 = 0.706, NRMSE = 0.54).

5.1.1.5 Commercial (COMMERC)

Commercial lands were primarily located either in city centers or strung along road systems, however a large number of exceptions existed. The predictors for this class also incorporated measures of proximity to city centers (*PROX_city250k_dist*, *PROX_cost_10k_city,* and *PROX_city100k_dist*). For the first time in the models discussed so far a measure of current land cover (*LC_nlcd01_24* – highest intensity urban land cover) was a strong predictor, and in fact by far the strongest predictor (Appendix E). Historical 1970s-era commercial land (*HIST_giras_12*) was also a prominent predictor, as well as co-location with major roads (*TRANSP_bts_faf2_pct*) and the ratio of population density to high-intensity urban land cover (*LC_ratio_popden_nlcd2324*). All of these make intuitive sense. The performance of this model was a medium-level result in this project (r^2 = 0.471, NRMSE = 0.73).

5.1.1.6 Industrial (INDUST)

Industrial land is often lumped together with commercial land in a classification scheme because they are both non-residential and both often require large structures which may co-locate in a commercial/industrial park. In the Boston area, however, industrial lands tended to be more distant from city centers, more likely to be near major transportation hubs, but at the same time had low road density themselves. However, as with commercial lands there were very regular exceptions to this. Industrial lands were one of the least clustered land uses (Figure 3-11).

The predictors for this class reflected that. The strongest predictor was *TRANSP_ratio_roadden_imperv,* the ratio of road density to imperviousness. Low values of this indicated few (public) roads and high imperviousness, a typical situation for industrial land. *PROX_city250k_dist* and *PROX_city100k_dist* again were also major predictors, along with measures of historical land use (*HIST_nlcd92_23* and

HIST_giras_13) and co-location with rail lines (*TRANSP_bts_rail_pct*). Low levels of population density to high-intensity urban (*LC_ratio_popden_nlcd2324*) rounded out the model predictors. The performance of this model was also on the low side of our 10 classes, reflecting the difficulty of mapping this class (r^2 = 0.429, NRMSE = 0.76). Nonetheless, a categorical map showing predicted vs. actual as in Figure 4-3 shows a quite reasonable mapped correspondence.

Figure 3-12 was given as a generic example of a decision tree but is in fact the rpart() version of our final model for INDUST (although it omits *PROX_city250k_dist*). The general relationships between variables are discernible: for example the highest prediction of industrial land (37.89, on the right hand side) results from low values of *TRANSP_ratio_roadden_imperv*, high values of *HIST_nlcd92_23* and high values of *HIST_giras_13*. Conversely, if there are somewhat higher values of *TRANSP_ratio_roadden_imperv*, in the range 0.1962 and 0.2236, and *PROX_city100k_dist* is less than about 1600 meters then there is also likely to be fairly high industrial land (22.79). The model is thus able to capture multiple kinds of relationships, even of the same variable.

5.1.1.7 Institutional (INSTIT)

Institutions are scattered. As Fellman et al. (1992) note, they are typically outside the normal rent-bid auction for land because society deems certain functions – schools, hospitals, libraries, public buildings – important without regard to their economic competitiveness. As noted earlier, we anticipated that their scattered and rather sparse nature would make them more difficult to map. Performance of this model was indeed the lowest of our classes (r^2 = 0.411, NRMSE = 0.73).

The strongest predictor, by far, for the class was our aggregation of GNIS institutional point locations (*LANDMRK_gnisconsol_instit_density*). This was also one of the few classes in which spatial pattern predictors played a significant role: SP*CON_cohesion_slopeclass2* and *SPCON_focal77_gt50_is_cv*). Both are in essence measures of the contiguity (or lack thereof) of impervious surfaces from the NLCD01 impervious surface continuous data layer. Three census variables also were important: *CENS_pct_walkbike_to_work*, *CENS_pct_hu_owneroccupied*, and *CENS_huden*. Examining an rpart() tree of this model (not shown), these are interpretable as: high percentages of people who walk or bike to work indicate higher presence of institutions (by itself a somewhat interesting result and suggests perhaps that a greater than average proportion of the people who work in public institutions might be walking or biking there!); low values of homes which are owner-occupied and lower values of housing unit density tend to be predictors of higher values of institutional land.

5.1.1.8 Transportation (TRANSP)

Many of the features in the TRANSP class - major road, rail, airports, terminals, harbors, and utility and communications facilities - are already well mapped from national data sources, e.g. the BTS. This made predicting them more straightforward however, there are enough exceptions to the available data sources – primarily in the form of small-medium sized utility or communications facilities, which are not well mapped at the national scale, to reduce the accuracy of a zonal prediction.

No one predictor variable stands out as extraordinarily strong in this model, but, as might be expected, a number participate from the TRANSP_ predictor class (*TRANSP_bts_rail_pct*, *TRANSP_a11_a17_roads_density*, *TRANSP_alltrans*, *TRANSP_a11_a38_roads_density*, *TRANSP_ratio_roadden_imperv*), and historical predictors (*HIST_indust_all_times*, *HIST_giras_14*), which included previous-era representations of transportation, also participate. Measures of population density clustering and proximity to roads also were in the final model. There is some colinearity among predictors, however enough of the information is unique for all the variables to be useful. The performance of this model was the best of the non-residential urban classes (r^2 = 0.568, NRMSE = 0.66).

5.1.1.9 Recreation and Open Space (RECR_OPEN)

Large features of this class (large parks, golf courses, etc.) were reasonably well mapped from the 1992 NLCD and the 1970s era GIRAS. It might be argued that the class is also well-represented in the NLCD01 in that most of those large features are also correctly coded as NLCD class 21, however many other areas which are also coded as 21 represent areas of low-level imperviousness which are interspersed in every land use (see Figure 3-12). Numerous other small parks, playgrounds, sports fields, etc. may or may not be captured accurately in the NLCD01, and there is regular confusion with agriculture or pasture land in that product. Nonetheless, the combination of those three data sources and the GNIS recreation predictor we created allowed a fairly good prediction. The four strongest variables (Appendix E) were: *HIST_nlcd92_85*, *LANDMRK_gnisconsol_recr_density*, *HIST_giras_17*, and *LC_nlcd01_21*. The performance of this model was reasonably good, and better than the COMMERC, INDUST, and INSTIT classes ($r^2 = 0.535$, NRMSE = 0.68).

5.1.1.10 Non-urban (NON_URB)

This class is of course not an urban land use class at all, and was included essentially for completeness sake, i.e. that the sum of all models could be made to add to 100%. Modeling non-urban land from these data is extraordinarily easy because there is an almost perfect linear correlation to current urban land cover from the variable *LC_sum_nlcd01_urban* (r = -0.96 – see Appendix C).

5.1.2 Integrated (constrained) models

Integrating and constraining the models to 100% gives a more complete solution to mapping urban land use. Although the stand-alone models are entirely functional by themselves – i.e. one might have an interest in mapping Industrial land alone across a large region without regard to any other type of land use – integrating them allows for comparison of percentages of land use within a block group, and for their sum to account for no more and no less than the actual sum. As described in Section 3, the method we selected for integrating model results was to use results from the nine other class models as input to a second pass prediction. The theoretical basis for this is that, if one has a good estimate of other land uses in the block group, then those would aid in prediction if there were relationships between the land uses. For example, there is a fairly strong inverse relationship between the SFRES_S and MFRES land uses (r = -0.41; see Table 3-

7), which would imply that if one could guess that there was a lot of MFRES land, that should argue against SFRES_S land being present.

The results showed only very modest improvements in individual models (Figures 4-4a and b), and then only for the two classes just mentioned: SFRES_S r^2 improved from 0.727 to 0.753 and MFRES from 0.706 to 0.741. In both cases we interpret the improvement as the RF model being able to leverage the information about the other class predictions, in particular the negative relationship between those two classes, to slightly improve those predictions.

The second part of integration was to range-standardize the 10 predictions for each block group to 100%. Although our primary analysis of results in this project has been from regression results (residuals = |pct prediction – pct actual|), one might also analyze the results as categorical data. We did this by taking the majority classification in the block group, as shown in Figure 4-5 and 4-6. Comparing the actual majority to predicted majority, 79% of block groups are classified correctly, and 92% are classified correctly when comparing the actual majority to predicted majority or secondary. Assuming the majority as the block group's classification, the following traditional confusion matrix results (Table 5-3):

Table 5-3: Confusion matrix of block groups from actual majority and predicted majority land use.
Acc = (Consumer's) accuracy.

					predicted							
		SF_L	SF_M	SF_S	MF	COM	IND	INST	TRAN	REC	NON	Totals
	SF_L	**10**	12	1	0	0	0	0	0	0	19	42
	SF_M	5	**248**	65	0	2	0	0	0	1	57	378
	SF_S	0	47	**755**	37	9	0	0	0	6	25	879
actual	MF	0	0	55	**350**	8	1	1	0	2	3	420
	COM	0	5	16	38	**60**	2	3	0	0	5	129
	IND	0	3	8	4	9	**18**	0	3	0	10	55
	INST	0	1	3	10	11	2	**9**	0	1	3	40
	TRAN	0	0	1	1	5	5	0	**4**	1	5	22
	REC	0	1	5	0	1	1	1	0	**28**	11	48
	NON	0	34	16	1	2	1	1	0	2	**694**	751
	Totals:	15	351	925	441	107	30	15	7	41	832	**2764**
	Acc (%):	66.7	70.7	81.6	79.4	56.1	60.0	60.0	57.1	68.3	83.4	78.7

The matrix basically confirms our previous results: the highest density residential classes – MFRES and SFRES_S are predicted very well (79% and 82%, respectively) as majority categories, and the other urban classes follow approximately the same accuracy distribution as given in Table 4-5 for regression results.

5.1.3 Total land area estimation

These results were also very encouraging (Table 4-7 and Figures 4-8 and 4-9(a)-(d)). By multiplying the predicted percentage of land in each class in each block group (from integrated models) by the block group's area, and summing, one obtains an aggregated areal estimate for the entire study area, or for individual counties. For example, the actual SFRES_L land area in the study area was 559.1 km^2, and the predicted was 556.7 km^2 (0.4% difference); the actual INSTIT was 81.0 km^2, and the predicted 85.5 km^2 (5.6% difference). The worst result was a difference of 10.8% (MFRES class).

The fact that these aggregated results are so good is essentially a manifestation of the modifiable areal unit problem (MAUP). In a mapped product it might be that every individual pixel is incorrect, yet over some larger area the product could be very accurate. Because errors of omission or commission average themselves out over a larger area the prediction of some classes (e.g. INSTIT) might be much better than if assessed for 2,764 individual block groups. As noted earlier, preliminary testing and literature results (e.g. Wu and Murray, 2005) led us to believe this was a possible outcome.

The results of breaking out land area numbers by county were also quite good, particularly for Essex, Middlesex, and Norfolk Counties (Figures 4-9(a) – (c)), but less so for Suffolk County (Figure 4-9(d)). Suffolk County is different in nature to the other three (much more densely urbanized overall – central Boston), and it is therefore not surprising that a model built from all four counties might underperform there.

We broke the results out by tract as well, and although not formally reported here, the results were overall better than by block group, but not as good as by county (judged semi-qualitatively by percent difference - we did not calculate r^2 and NRMSE by county, because there were only four of them). This also seemed in line with expectations.

5.1.4 Sensitivity Analysis

While incorporating data for the entire study area provides the most robust possible models for interpretation (e.g. which land uses are best predicted by which variables), it does not provide information about how much training data might be required to achieve a certain performance; information which would be valuable for a practical extension of this project to mapping of other areas, or conceivably the entire U.S. Figure 5-3 gives a graphic portrayal of the data in Table 4-8:

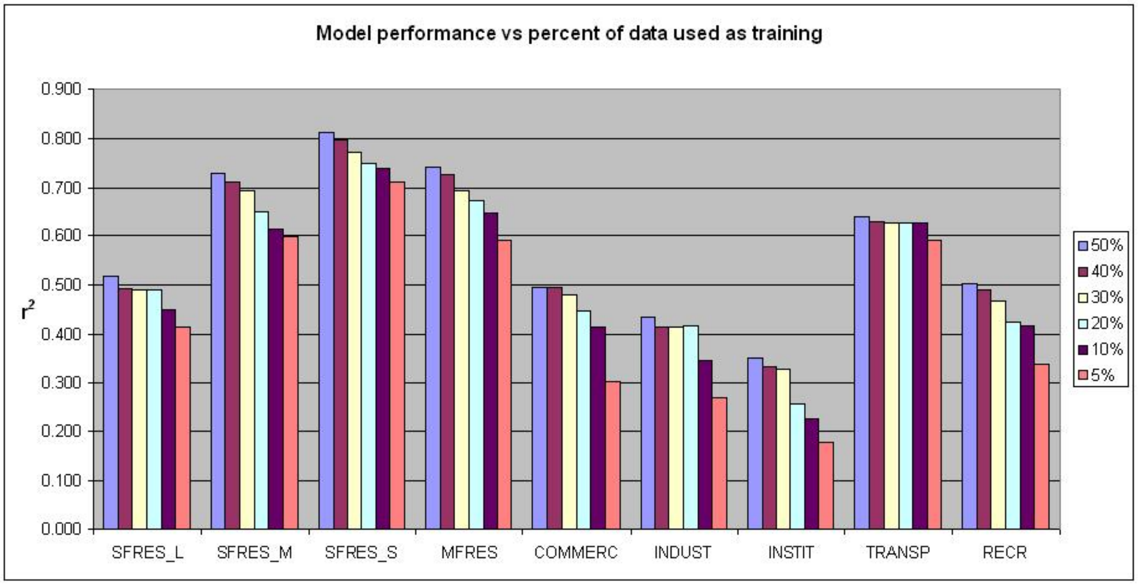


Figure 5-3. Sensitivity of performance to training data sample size

Although in every case performance does decrease with fewer block groups used as training, in some cases that decrease is quite minimal. For example, the TRANSP model is only slightly worse with 5% training data as with 50% training data. The four residential classes are likewise less affected by less training data, and most models maintain reasonable performance with 10% (276 block groups) or even 5% (138 block groups) training data. The classes most affected by less training data are the COMMERC, INDUST, and INSTIT classes, which are the most difficult to predict overall as well. In these cases model performance is reasonably consistent down to 20%

training data, but suffers at lower levels. This suggests that these relatively sparse classes may (a) need to be combined, and/or (b) have more targeted training samples collected, which ensure that enough data are available for robust models.

5.1.5 Incorporating Business Analyst data

Incorporating the ESRI Business Analyst point data, based on the predictors we created from them, led to only modest improvements in the three classes tested (COMMERC, INDUST, and INSTIT; Table 4-8). However, these are useful improvements, because those three classes are the most difficult overall to predict. Although we invested a considerable amount of time in creating these predictors ("cleaning" the dataset of what we believed were residence-based businesses, evaluating NAICS codes, then categorizing locations by commercial/industrial or institution), it is entirely possible that others could find more effective ways to manipulate or use these data as predictors of land use or some other target variable (and that effort by itself could constitute a fairly sizeable project). Our judgment from the experience here is that, for a well-funded national effort to map land uses it might be worthwhile to acquire the data, but that some of the information content is already in other datasets (for institutions particularly, from the GNIS). Of course, for other applications, the dataset might be critically valuable.

5.1.6 Noteworthy variables

Although a number of individual variables have been discussed above in the context of specific models, some discussion about particular variables by predictor class is warranted, either because they were very useful, or because they are of unique interest:

Census: While, as was expected, population and housing unit density (*CENS_popden* and *CENS_huden*) were often major components for predicting land use, the number of rooms per home (*CENS_hu_median_numb_rooms*) ended up being a surprisingly strong predictor of several land uses. Because the variable maps to distance from city centers (more distant homes have more rooms – see Figure 3-7), it incorporates some of the predictive power of the proximity variables. Several other census variables similarly had a generally linear relationship with distance to city centers, and of these *CENS_pct_hu_owneroccupied* (people in outer suburbs more likely to own their homes), and *CENS_median_hh_income* (higher incomes in suburbs) were also useful predictors. Other variables which did not necessarily have a monotonic relationship with distance to city center were also useful for specific land uses (e.g. *CENS_pct_5_or_more_units_in_structure*, *CENS_pct_walkbike_to_work*).

Landcover: Some of the most useful predictors in this category were ratios of land cover with a socio-economic variable. Specifically, the ratio of housing unit density to imperviousness (*LC_ratio_huden_imperv*) and ratio of population density to high-intensity urban (*LC_ratio_popden_nlcd2324*) captured relationships of land cover and socio-economic data that were predictive of land use. The simple mean and variability of imperviousness (*LC_nlcd01_imperv_mean*, *LC_nlcd01_imperv_stdev*) were also regularly useful, as well as specific variables for some classes (e.g. *LC_nlcd01_24* for the COMMERC class).

Historical LULC: One might suppose that representations of land cover/land use from relatively recent periods would be highly predictive of current land use, and the variables from this class were regularly very useful, particularly for the COMMERC, INDUST, TRANSP, and RECR_OPEN classes. However, their usefulness as predictors of current land use is limited by three factors: (1) their original MMU (e.g. many instances of small parks and recreation areas fall below the 1970s GIRAS MMU of 4 hectares), (2) their original accuracy (e.g. the stated accuracy of some NLCD92 urban classes was less than 50%), and (3) their currency.(e.g. one need only look at the GIRAS panel in Figure 1-2 to see that the area around Fairfax, VA has changed quite a bit since the 1970s). Despite that, for specific land uses the correspondence to current land use and helpfulness of a number of these variables is quite strong. A number of composite variables created – *HIST_highresid_all_times*, *HIST_indust_all_times*, *HIST_commerc_all_times*, particularly – our best guesses at those land uses from three time periods – were also helpful.

Transportation: As with the ratios noted above, the variable *TRANSP_ratio_roadden_imperv* (ratio of road density to pct imperviousness) was particularly helpful, and the fourth strongest variable overall across all classes (Table 5-1, above). It was the strongest predictor of the INDUST class, i.e. a class with low road density and high imperviousness. The all-purpose composite of transportation we created – *TRANSP_alltrans* was also useful for the TRANSP class.

Landmarks: The GNIS point locations were among the strongest predictors for two classes (INSTIT and RECR_OPEN), but less useful for other classes. Representations of institutions and recreation areas is quite good in the GNIS, but representation of shopping, industrial, or other land uses is spotty. The consolidated predictors *LANDMRK_gnisconsol_inst_density* and LANDMRK_gnisconsol_recr_density – weighted point density representations were better predictors than gridded versions (*LANDMRK_gnis_inst_grid* and *LANDMRK_gnis_recr_grid*).

Proximity: As noted above, these context variables were overall the strongest predictors of land use as a whole. A fairly global metric – simple Euclidean distance to Boston center (*PROX_city250k_dist*) and mean distance to nearest road (*PROX_mean_dist_road*) – a more local metric - were the two overall strongest predictors of land use generally, and *PROX_city100k_dist* was also a prominently helpful predictor. Our cost-distance representations of context, particularly *PROX_cost_gnis_instit*, *PROX_cost_10k_city*, and *PROX_cost_50k_city* were also regular model participants and helped to characterize relationships to smaller cities and institutions. It is worth pursuing these representations of context in future research.

Spatial Autocorrelation: Using local Moran z-score representations of clustering, particularly of population density, natural vegetation, imperviousness, and median number of rooms, was effective as a prediction method (the variables *SA_localMoran_popden*, *SA_localMoran_allnatveg*, *SA_localMoran_imperv*, and *SA_localMoran_medianrooms*, respectively). The measures capture clustering over larger areas, which are indications of land use patterns, and their use might also benefit from further exploration.

Spatial Pattern-categorical data: As a class these variables were not hugely effective in determining land use, but in specific instances, as noted above several variables were useful. The variable we created to characterize annuluses, *SPCAT_foc_annulus*, was one of the strongest predictors of SFRES_L land, however was fairly collinear with *SPCAT_shape_index_22_24*, and was ultimately not a participant in a final model. Both are essentially measures of the homogeneity/contiguity of urban land. That the spatial pattern variables were very useful for some classes (SFRES_L and INSTIT) is notable, but that they were not useful for most other classes is a reflection of the coarseness of the 30-m data in describing specific land use patterns. One of the goals of this project was to assess the general usefulness of spatial pattern measures derived from 30-m land cover data, and generally one might conclude that, with a few exceptional cases, they have only tangential usefulness.

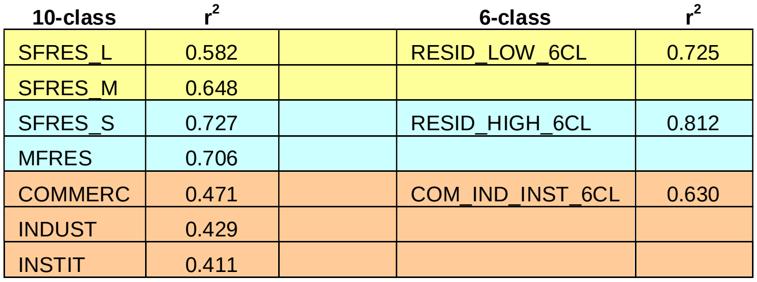
Spatial Pattern-continuous data: The same conclusion may be drawn about the spatial pattern metrics created from the NLCD01 30-m imperviousness data. Although some variables were useful in final models, as noted above, as a class of predictors overall they are useful only in landscape settings with high degree of fragmentation of surfaces, specifically SFRES_L and INSTIT, as above. The slope-of-imperviousness variables we derived were useful for these two classes, but not in others, suggesting transition areas of imperviousness may be more significant for those classes. The results suggest some additional investigation might be warranted.

Miscellaneous: The variable *MISC_vg2000_mean* was a useful predictor in a number of classes. This was of interest, because the variable is based on coarse resolution (1-km pixels) representations of vegetation. If coarse resolution data (e.g. 1-km) provide the same information as higher resolution data (e.g. 30-m) for some characteristics, then that is noteworthy, as the 1-km data are easier and faster to manipulate, require less storage, etc.

5.2 6-class models

The 6-class models created for Boston had good performance as validated from Boston data (Table 4-10(a); red columns). Those classes that were aggregated had better performance than their component classes: for example, in the 10-class models, the classes SFRES_L and SFRES_M had r^2 of 0.582 and 0.648, respectively. When those classes were merged into the single class of RESID_LOW_6CL, the r^2 was 0.725. The same was true of the other aggregations: Table 5-4 summarizes these:

Table 5-4: Comparison of Boston 10-class model performance with their 6-class equivalent, where classes were aggregated.



10-class	r^2		6-class	r^2
SFRES_L	0.582		RESID_LOW_6CL	0.725
SFRES_M	0.648			
SFRES_S	0.727		RESID_HIGH_6CL	0.812
MFRES	0.706			
COMMERC	0.471		COM_IND_INST_6CL	0.630
INDUST	0.429			
INSTIT	0.411			

That thematically-aggregated classes would have better predictions was not too surprising, but the magnitude of the improvement was encouraging. The RESID_HIGH_6CL mapping is shown for Boston in Figure 4-10(a).

When the Boston model was applied to Providence, five of the six models held their performance fairly well (Table 4-10(a)), i.e. r^2 and pct improvement were not much different from Boston. In fact, the Boston COM_IND_INST_6CL model performed considerably better in Providence than with the Boston data (r^2 0.764 vs. 0.630). So that would imply that at least some models are well transportable to another geographic area. However, the class RESID_LOW_6CL had significantly worse performance in Providence, and in fact the metric pct improvement from mean has a negative value (worse prediction than simply using the mean). This implies that there must be sizeable differences in low-intensity residential areas between Boston and Providence (and it might be that there are some), however, it is also noteworthy that the model created from Providence data itself (Table 4-10(b)) has somewhat middling performance even for its own area ($r^2 = 0.594$, pct improvement = 36), and has the poorest performance for that class of any of the four cities. This would suggest either (a) the characteristics of that class might be much more variable in that area, and/or (b) there are inconsistencies in the reference data or (c) there is some other yet-to-be-determined unknown that causes that class to be difficult to predict in that area.

The Boston models as applied to Atlanta and Los Angeles had mixed performance. For example, both the Boston COM_IND_INST_6CL and TRANSP models transport very well to all of the other three cities, and have performance as good or better there as they do in Boston. The other four models, however, transport poorly to Atlanta and Los Angeles.

Examining the performance of the models created from the Providence data (Table 4-10(b)), we find that the models perform fairly well for Providence itself, and lower, but generally reasonable performance for Boston, but in some cases much worse performance for Atlanta, for example, in some cases negative pct improvement from mean.

Examining the performance of the models created from the Atlanta and Los Angeles data (Tables 4-10(c) and (d), respectively), we find that the models for each area perform well for that area (with minor exceptions: the NON_URB models for Atlanta and Los Angeles are not as good as those for Providence and Boston), however generally do not transport well to the other cities. There are some exceptions to this too, for example, the Los Angeles RESID_HIGH_6CL model transports fairly well to Boston and Providence, but on the whole they are not particularly good performers in the other areas, and in some cases quite poor.

Taking a look at which predictors went into the models for each area (Tables 4-11(a) – (d)) gives some insight into why this is so. Firstly, one would not necessarily expect models from different areas to be built from identical variables even if the areas were similar, because there is some colinearity in the predictors, and even very small differences might cause one variable to be preferred over the other. That by itself is not a problem nor unexpected given the modeling technique. There are however obviously very different characteristics in some ways between the four areas and datasets, and the same LU class may have different characteristics in different cities. For example, the variable *SPCAT_lc_entropy* (land cover diversity) is a good predictor of the RESID_LOW_6CL class in Atlanta, because the background vegetation in low density residential areas may consist of deciduous forest, evergreen forest, mixed forest, various types of wetlands, pasture, crops, or other vegetation. However, in Los Angeles land cover diversity in similar areas is very low (primarily shrubland and some evergreen), thus the model predicts poorly from that predictor. In Los Angeles a predictor of RECR_OPEN land is *HIST_nlcd92_11*, which is Open Water land cover: in Los Angeles ponds and lakes are nearly always associated with parkland; in the other areas much less so. And while the relationship of some census variables to land uses may be fairly consistent across areas, for example, population density or housing unit density, other socio-economic variables may differ considerably, for example population density change may not have occurred consistently from area to area within land uses. Numerous other predictors may have different characteristics: for example, presence of rail (*TRANSP_bts_rail_pct*) is a predictor of the COM_IND_INST_6CL class in Boston, but not elsewhere because the characteristics of where rail lines are located differ.

How then to apply models between areas? One solution is to a priori limit the predictor variables to only ones known to be perfectly consistent among the areas. This is a poor solution because (a) it might not be easy to identify consistency, and more importantly (b) there would likely end up being very few predictors, and model performance for all areas would be poor. A second solution is to simply not try to apply models between areas: that is, every urban area should have modeled land use based on training data from that area only. This was in fact the approach of the development of the NLCD01 itself, which was assembled from 66 zones across the US: every zone was modeled based on local training data. That is a perfectly reasonable solution, but requires more data collection. Another solution is to develop a series of more generic models which might be applied to typologies of urban areas which are similar. That is, a set of models might exist for Southwest cities of size greater than 500,000, which might be applicable to El Paso, Albuquerque, Tucson, Phoenix, and Las Vegas. The safer, but more time-consuming option, however, is to develop training data from each area. All of the models we developed, with very minor exception, performed very well for the area from which it was built, albeit using training data available for the entire study area. In addition to the results given above regarding sensitivity of performance to training sample size, it is anticipated that future research will provide additional information about training sample size requirements.

5.3 Limitations, improvements, lessons learned

A limitation of the zonal method shown here is the lack of a mapping at a spatial resolution finer than a block group. For some applications this might cause a problem. For example, if one's goal were to map urban land within 100-m riparian (near-stream) zones for the entire country, data mapped by block group would have insufficient resolution for it to be useful. However, we would counter that block groups are an absolutely useful scale for many, if not most, national or regional-scale applications, and that, as noted previously, the median size of a block group in this study (0.46 km^2) is much smaller than a single pixel from a regional-scale remote sensing platform such as the AVHRR series of satellites (approximately 1-km resolution).

Additionally, a work-around for this, although not demonstrated here, is certainly feasible, which is to perform a dasymetric mapping of the zonal information to 30-m land cover data. That is, if one has estimated that x% of the zone consists of industrial land use, and one has an explicit 30-m pixel mapping of four land-cover categories of increasing imperviousness, one could devise a formula for assigning a new category to define some of those pixels as "industrial". This may be a useful area of future investigation.

An aspect of the method demonstrated here that might require future modification is that the nature of Census data in the U.S. is changing. Previous decadal Census collections (2000, 1990, etc.) have been based primarily on "short-form" (100% count and basic demographic data), and "long-form" data (more detailed questionnaire sent to only a sample of homes, which is then aggregated to block group and coarser geographies). With the 2010 Census, however, the "long-form" data is no longer collected only at decadal intervals but rather is now (since 2003) collected by a rolling set of surveys known as the American Community Survey (ACS). The ACS provides a dataset known as the Public Use Microdata Sample (PUMS), which allows users to tabulate information across specific demographic categories – for example, what are the income characteristics of only unemployed people? This is advantageous, in that the information may be more current for a specific area than a decadal survey, and that information may be more tailorable to users' needs. However, it is somewhat disadvantageous in terms of national mapping efforts in that data for different regions may represent snapshots of somewhat different time frames.

We chose the block group as a the unit of geography partially based on the rationale that it was the finest census geography at which it was defensible to use spatial pattern metrics from 30-m data (as described in Section 1). As it turned out, the spatial pattern metrics we executed were only partially helpful for the task at hand. It would certainly be useful to think about executing the method at the block level (finest possible census scale), however, it would increase the data storage and processing requirements substantially, and there may be a tradeoff of zone resolution with predictive accuracy.

We could have characterized land use at multiple scales, i.e. for both block groups and tracts, but because of the already substantial data processing requirements of the project and volume of results from the block group scale alone, we opted against it. The block group scale, in any case, allows for aggregating up, as was shown in Section 4.3.3.

We have used the term national-scale to refer to data which are available nationally (the data exist in a consistent dataset or format), but also to refer to the feasibility of processing them in a reasonably straightforward and timely way, given today’s technologies. Although technologies (disk storage, compression, computing power) are constantly improving, the feasibility of processing very high resolution data of the kind that would significantly improve some metrics (let’s say 1-m resolution impervious surface data) at the national scale would be a daunting task to even a well-funded and motivated organization, even if those data were available. However, that is in essence one of the points of this project: to show that quite a lot can be done even with the publicly-available data currently existing, given a good method and metrics.

We have used a simple areal estimation based on a 2-dimensional earth, however, of course in reality, land uses may exist in 3-dimensions, i.e. buildings might have multiple land uses in a vertical dimension (even above and below ground). This might be at least partly characterized with data of high enough resolution (Aubrecht et al., 2009), however was not feasible with national data. Having said that, the method proposed by Leroux et al. (2009) shows that building heights may be modeled to some degree using national-scale (continental-scale even) SRTM data, and that exploring this method is intriguing for its potential to add additional predictive information.

It is important to note that the results are only as good as the input data, and that in some cases national-scale data have consistency issues. The GNIS particularly, as noted earlier, have consistency issues with some data layers. These inconsistencies have not been well quantified and are more or less based on our own observations and personal communication with USGS personnel. On the other hand, the data layers that were most useful in this project and were most common – schools, hospitals, public buildings, parks, cemeteries, golf courses, and a few others – seem to be fairly consistent from area to area. Some others, such as interstate crossings, are not. The NLCD01 (and every dataset, frankly) also of course has some data issues. For example, in the RECR_OPEN model for Atlanta, one of the stronger predictors ended up being *LC_sum_nlcd01_ag*: the percentage of agricultural land in the block group. On closer examination it turned out that this was simply because there were a number of park and recreation areas incorrectly coded as agriculture in the NLCD01. These kinds of errors are confounding. Nonetheless, it appears to be a localized error, and we also believe these kinds of errors are relatively rare.

It is also important that the reference data be consistent. The design of this project was based on being able to acquire consistent and high-quality reference data, which we did have for Massachusetts. We believe the reference data from the other three areas were also consistent – or as consistent as is possible to find – however, even minor variations might make some difference – year of collection, subtle differences in class interpretations, etc. We used already-existing reference data because we believed this was more defensible than creating our own ground truth, i.e. that there was little possibility for personal bias to creep in. However, the difficulty of finding consistent and detailed reference data for land use using the same class definitions and from the same year would make it nearly impossible to execute this method over a much larger area or many cities without creating one's own ground truth dataset, which does require some time and labor. On the positive side, being able to control your own ground truth increases its consistency and of course allows you to target any urban area. And most importantly, it would allow finer thematic detail. The only reason we modified the testing from 10 to 6 classes was because it was so difficult to find consistent detailed classes in multiple datasets.

One approach that might be taken to improve the models, at the risk of complicating or over-training, is to segment them by some criteria, notably, either by population density or by a "more urban/less urban" segmentation. Lo (2003) found that dividing census tracts into two groups of "periphery" and "centre" improved population modeling. The danger in our case might be of a proliferation of models even for a single metropolitan area, and of a paucity of certain land uses within certain areas.

While we did measure proximity and access in several ways and to different kinds of centers and features (population centers, major and minor roads, large urban patches, institutions, recreation), there might have been additional ways. Retrospectively, one of these that we note is distance to employment centers (McMillen, 2004), which are feasible to derive from U.S. Department of Transportation data. Employment centers, to the extent that they are different from (nighttime) population centers are additional landscape forces of land use and should be included in future research.

5.4 Next steps

5.4.1 Applying methods to systematic mapping of US or regions

Because the method presented here is based entirely on nationally-available predictor data, it is feasible for it to be applied to large metropolitan areas across the continental US (some predictor data for Alaska and Hawaii may not be as readily available). If reference data are derived for the project, then the full thematically-detailed 10-class structure, or a similarly-detailed variation thereof, could be used. Although we invested a large block of time deriving predictor variables and testing their usefulness, the number that were truly useful and necessary for final models was rather small (Table 4-6), and could fairly readily be calculated for any area in the continental US. We believe this is promising for anyone interested in urban land use at broad scales.

As noted above, if one wished to map urban land use for (as example) the 20 largest US cities, it would likely be necessary to create training and validation data by visually interpreting imagery samples for each area, because consistently-derived reference data do not exist over that kind of scale, and transporting models from one area to another may be problematic. However, the great advantage, as noted earlier, is that creating one's own reference data allows for much greater flexibility in where to map, which classes to define (10-class or similarly detailed), and having consistent methods. As noted above, we anticipate that we will conduct future study of training data sample size requirements.

Over what area(s) should or could the method presented here be applied? We have intentionally not adopted a strict definition of a city in this project because the proposed method is potentially modifiable or applicable to different levels of urbanization. For the 2000 Census the U.S. Census Bureau defined an urban area as "core census block groups or blocks with population density of at least 1,000 people per square mile (386/sq km) and surrounding blocks that have an overall density of at least 500 people per square mile" (U.S. Census Bureau, 2009e). Metropolitan Statistical Areas (MSAs) – counties surrounding core areas of at least 50,000 people – are an alternate definition. Numerous studies have used MSAs as a basis for large-city characterization (e.g. Griffith and Wong, 2007), and they would be a perfectly reasonable approach to executing this method over a larger area, for example, the top 10, or 20, or 50 largest MSAs. Such a product would be a valuable complement to existing and future land cover datasets.

5.4.2 Related research questions/future work

In the course of this dissertation, a number of potential avenues of future research have been noted. These are summarized here:

- Investigating how models may more effectively be transported between cities, to include investigating whether models for city "typologies" is feasible. That is, could the largest U.S. metropolitan areas be partitioned into city types, and a single model be effective for all the cities in each group, and if so, using what class structure. Even though the results here suggested transportability was generally low, even our two most similar cities, Boston and Providence, are still fairly different (the Boston MSA is 3x more populous than the Providence MSA). If reference data were derived specifically for the project then any city or cities could be studied. We limited ourselves here to only places where fairly compatible reference data already existed, which limited the choice of study areas.
- Identifying more specific requirements for training data sample size. Can a metropolitan area be modeled well with training data from 200 block groups? What about 100? Our initial results indicate different classes likely have different sample size requirements.
- An alternative spatially-explicit mapping of cities is shown by Leroux et al. (2009), which, however, is possible because of the existence of very detailed feature mapping in Canadian data (the Canadian NTDB). It would, however, be possible to apply some of the same data and methods to create a similar, if

less detailed, product here. Essentially, one could create a categorical urban mapping of US cities which integrated available national data: - roads and other transportation features, 30-m landcover, population density, building height estimates from SRTM/DEM, GNIS point locations, and possibly other data - into a single thematic layer. That layer might or might not have the same kind of thematic land use information as proposed in this dissertation (for example classes might be something like “High intensity urban, > 40m structures”). At present these datasets that describe urbanization exist only as separate entities and it is up to individual users to integrate them, if done at all. The availability of a single dataset that had more detailed thematic information would be helpful for many purposes.

- Extending the method to a dasymetric mapping by pixel. That is, take the broader knowledge of percent land use in a block group and apply that to individual 30-m pixels.
- Explore other possible data sources that might improve the process, such as publicly available volunteer-collected data available from openstreetmap.org.
- Are there better ways to relate land cover – portrayed in the NLCD01 as levels of imperviousness – to economic-function land use areas? We showed the relationship between the NLCD01 classes and Boston land use in this project (Table 3-5), which indicated only a very rough correspondence between land cover and land use, and then only for some classes. However providing

additional information and analysis as to how one might be able to relate land cover and land use would be of great value to some national-scale data users.

6. Summary

Why should we care about characterizing urban land? Because increasingly the landscape of the United States is dominated by humans. The late film director Anthony Minghella, in describing why the 2003 film “Cold Mountain”, set in 19th century North Carolina, had been filmed in Romania, stated that when scouting locations in the eastern US it was nearly impossible to find a broad landscape view that did not have visible presence of 20th century man – roads, cars, houses, power lines, mechanized agriculture, cell phone towers, etc. (Chicago Sun-Times, 2004). In the United States population nearly doubled between 1950 and 2000, and land change has even outstripped that: for decades the extent of urbanized land area in the United States has exceeded the percentage of population growth (Theobald, 2005). “Urban sprawl”, generally considered the undesirable growth pattern of a small population consuming a disproportionate amount of natural land, is now a defining aspect of the US urban landscape. It is clear that the issues of the third millennium will be urban (Weber, 2001).

The distribution of people on the landscape has important implications to the environment (U.S. EPA, 2010), in addition to many other aspects of modern life. Improved methods and information about the urban environment, and especially about how the land is used, are therefore noteworthy. As noted earlier, the lack of consistent urban land use information at the national scale in the U.S. is a striking data gap. This gap makes it difficult or impossible to perform certain types of study of the environment, of climate, of landscape trends, and much else, at broad scales in the U.S.

The fundamental motivation for this project is to demonstrate that a zone-based mapping of urban land use using block groups as the unit of measurement can be effectively done with national data, and public data at that. Block groups (of which there are about 220,000 in the U.S.) provide a grain size that is sufficiently detailed for most national and regional applications, and the method avoids some of the problems with other methods of identifying urban land use (manual interpretation, per-pixel mapping, delineating and classifying areas of contiguous land use).

These are the major outcomes of this project:

1. A zone-based approach to mapping thematically detailed (10-class) urban land use with publicly-available national data was demonstrated for a four-county area around Boston, using block groups as the zonal unit. Performance was very good for some classes, particularly the highest-density residential classes (e.g. integrated model validation r^2 of 0.753 for single-family small lot residential, and 0.741 for multi-family residential), but more difficult for other classes, namely institutions (r^2 = 0.411), industrial (0.429), and commercial (0.471). We discussed in detail some of the issues with predicting and mapping these classes at broad scales, and explored some alternative methods. We demonstrated a decision tree as a good modeling method for this application.
2. An aggregation of results to less thematically detailed classes (10 classes to 6 classes) resulted in improved results. For example, an aggregation of the single-

family small lot and multi-family residential classes to “high density residential” resulted in r^2 of 0.812.

3. An aggregation of results to the county or study-unit level likewise resulted in high accuracy of predicting land use, sometimes within 1% of a land use type. For example, actual industrial land use in the four counties was 125.1 km^2, and predicted, based on withheld validation data, was 124.4 km^2 (0.6% difference). The median difference by class between actual and predicted was about 5%.
4. Stand-alone models were built for each class, then integrated so that predicted land use summed to 100%. Integrating the models slightly improved prediction of some land use classes. If classifying by majority land use, 79% of our predicted majorities by block group matched the actual majority land use, and 92% if taking the secondary land use.
5. A large suite of predictor variables were created and tested from nationally-available sources, some of them novel approaches to characterizing measures of spatial pattern, proximity, or spatial autocorrelation. By general category, measures of proximity to city centers and roads, and measures of spatial autocorrelation were some of the consistently strongest predictors of land use. We note that land cover by itself was rarely one of the strongest predictors. We discussed the strengths and weaknesses of many of the variables and provide a listing of their relative usefulness in predicting various land uses. The results support classical land use theory of the importance of distance to city centers and access routes as determinants of land use.

6. Although spatial pattern metrics derived from either 30-m categorical or continuous data were useful for two classes where fragmentation of the landscape is fairly prevalent (single-family large lot residential and institutions), overall, spatial pattern metrics from 30-m data appear to have limited predictive power for thematically detailed land use. They are likely to be simply too coarse to give good detailed information about land use.
7. Models based on 6 land use classes, when applied to a completely different geographic area (Providence, Atlanta, Los Angeles) had mixed performance, and models built from data of those cities likewise generally did not transport particularly well, with the exception of fair transportability between Boston and Providence. The results imply that, with the exception of possible transportability between areas which are similar, the surest way to apply the methods here to many cities is for models to be built from training data for that city. This is perfectly reasonable, given that there may be considerable differences in urban characteristics, vegetation, etc. between urban areas. The 6-class models built for each city all had good performance for their own area.
8. Training data requirements were examined for the 10-class models. The residential and transportation classes maintained fairly good performance with only 5 or 10% of data used for training, however the commercial, industrial, and institutional classes had poorer performance at those levels. These results may help to guide how classes should be aggregated and/or how training data are collected for future efforts.

9. The separability and unique characteristics of each of the 10 detailed land use classes were examined extensively, to include some inter-class relationships (e.g. land use type proximities), and their linear correlation to each of the predictor variables is given. The relationship of each of the land use classes to the current NLCD01 is also given.
10. A brief examination of the usefulness of the commercial ESRI Business Analyst product for land use prediction is given.

All data or detail about specific methods or calculations are available from the author on request.

Appendix A

Land use codes and descriptions for Massachusetts reference data (MassGIS, 2008). For historical reasons classes 23-37 may in some cases overlap with classes 1-22.

Code	Description	Recoded to
1	Intensive Agriculture	NON_URB
2	Extensive Agriculture	NON_URB
3	Forest	NON_URB
4	Nonforested freshwater wetland	NON_URB
5	Mining - sand, gravel, rock	NON_URB
6	Open land - abandoned agriculture, power lines, areas of no vegetation	NON_URB
7	Participation recreation - golf, tennis, playgrounds, skiing	RECR_OPEN
8	Spectator recreation - stadiums, racetracks, fairgrounds, drive-ins	RECR_OPEN
9	Water-based recreation - beaches, marinas, swimming pools	RECR_OPEN
10	Multi-family residential	MFRES
11	Single-family residential - smaller than 1/4 acre lots	SFRES_S
12	Single-family residential - 1/4 to 1/2 acre lots	SFRES_M
13	Single-family residential - larger than 1/2 acre lots	SFRES_L
14	Salt marsh	NON_URB
15	Commercial	COMMERC
16	Industrial	INDUST
17	Urban open - parks, cemeteries, public greenspace, vacant land	RECR_OPEN
18	Transportation - airports, docks, divided highway, freight, railroads	TRANSP
19	Waste disposal - landfills, sewage lagoons	RECR_OPEN
20	Water	NON_URB
21	Woody perennial - orchards, nursery, cranberry bog	NON_URB
22	(used for MassGIS QA/QC)	NON_URB
23	Cranberry bog	NON_URB
24	Power lines	NON_URB
25	Saltwater sandy beaches (part of #9, no longer used)	NON_URB
26	Golf	RECR_OPEN
27	Tidal salt marshes (part of #14, no longer used)	NON_URB
28	Irregularly flooded salt marshes (part of #14, no longer used)	NON_URB
29	Marina	NON_URB
30	New ocean (areas of accretion)	NON_URB
31	Urban public	INSTIT
32	Transportation facilities	TRANSP
33	Heath	NON_URB
34	Cemeteries	RECR_OPEN
35	Orchard	NON_URB
36	Nursery	NON_URB
37	Forested wetlands (part of #3, no longer used)	NON_URB

Appendix B1

Census-based predictor variables (U.S. Census Bureau, 2009c).

Variable name	**Description**
CENS_hu_median_numb_rooms	Median number of rooms per household
CENS_hu_median_year_struct_built	Median year structure built
CENS_hu_pct_bottledgas	Percent housing units using bottled gas as home heating source
CENS_hu_pct_lacking_complete_plumbing	Percent housing units lacking complete plumbing facilities
CENS_huden	Housing unit density (units/sq km)
CENS_median_hh_income	Median household income
CENS_pct_2_or_more_units_in_structure	Percent housing units with 2 or more units in structure
CENS_pct_5_or_more_units_in_structure	Percent housing units with 5 or more units in structure
CENS_pct_foreignborn	Percent population foreign-born
CENS_pct_households_with_ss_income	Percent households with social security income
CENS_pct_hu_5ormore_person_household	Percent housing units with 5 or more person households
CENS_pct_hu_occupied	Percent housing units occupied
CENS_pct_hu_one_person_household	Percent housing units with one person household
CENS_pct_hu_owneroccupied	Percent housing units owner-occupied
CENS_pct_nonwhite	Percent population non-white
CENS_pct_pop_below_poverty_lev	Percent population below poverty level
CENS_pct_publictransport_to_work	Percent population commute via public transport
CENS_pct_walkbike_to_work	Percent population walk or bike to work
CENS_pden_change90_00	Change in population density, 1990-2000 (persons/sq km)
CENS_popden	Population density (persons/sq km)

Appendix B2

National Land Cover Data (NLCD01) predictor variables (USGS, 2009d).

Variable name	Description
LC_nlcd01_11	NLCD01 percent Open Water
LC_nlcd01_21	NLCD01 percent Developed, Open Space
LC_nlcd01_22	NLCD01 percent Developed, Low Intensity
LC_nlcd01_23	NLCD01 percent Developed, Medium Intensity
LC_nlcd01_24	NLCD01 percent Developed, High Intensity
LC_nlcd01_31	NLCD01 percent Natural Barren
LC_nlcd01_41	NLCD01 percent Deciduous Forest
LC_nlcd01_42	NLCD01 percent Evergreen Forest
LC_nlcd01_43	NLCD01 percent Mixed Forest
LC_nlcd01_52	NLCD01 percent Shrubland
LC_nlcd01_71	NLCD01 percent Herbaceous
LC_nlcd01_81	NLCD01 percent Pasture/Hay
LC_nlcd01_82	NLCD01 percent Cultivated Crops
LC_nlcd01_90	NLCD01 percent Woody Wetlands
LC_nlcd01_95	NLCD01 percent Emergent Herbaceous Wetlands
LC_nlcd01_imperv_mean	NLCD01 mean percent impervious surfaces
LC_nlcd01_imperv_range	NLCD01 range of impervious surface values
LC_nlcd01_imperv_stdev	NLCD01 std dev of impervious surface values
LC_ratio_huden_imperv	Ratio of housing unit density to pct imperviousness
LC_ratio_popden_nlcd2324	Ratio of population density to high intensity urban lc
LC_sum_nlcd01_2122	NLCD01 percent all lower intensity urban classes (21+22)
LC_sum_nlcd01_ag	NLCD01 percent agriculture classes (81+82)
LC_sum_nlcd01_allnatveg	NLCD01 percent all natural veg. classes (41+42+43+52+71+90+95)
LC_sum_nlcd01_allveg	NLCD01 percent all veg. classes (41+42+43+52+71+81+82+90+95)
LC_sum_nlcd01_forest	NLCD01 percent forest classes (41+42+43)
LC_sum_nlcd01_urban	NLCD01 percent all urban classes (21+22+23+24)

Appendix B3

Historical predictor variables: NLCD92 (USGS, 2009d), GIRAS (USGS, 2009a),

Variable name	Description
HIST_nlcd92_11	NLCD92 percent Open Water
HIST_nlcd92_21	NLCD92 percent Low Intensity Residential
HIST_nlcd92_22	NLCD92 percent High Intensity Residential
HIST_nlcd92_23	NLCD92 percent Commercial/Indust/Transp.
HIST_nlcd92_31	NLCD92 percent Bare Rock/Sand/Clay
HIST_nlcd92_32	NLCD92 percent Quarries/Strip Mines/Gravel Pits
HIST_nlcd92_33	NLCD92 percent Transitional
HIST_nlcd92_41	NLCD92 percent Deciduous Forest
HIST_nlcd92_42	NLCD92 percent Evergreen Forest
HIST_nlcd92_43	NLCD92 percent Mixed Forest
HIST_nlcd92_51	NLCD92 percent Shrubland
HIST_nlcd92_61	NLCD92 percent Orchards
HIST_nlcd92_81	NLCD92 percent Pasture Hay
HIST_nlcd92_82	NLCD92 percent Row Crops
HIST_nlcd92_85	NLCD92 percent Urban/Recreational Grasses
HIST_nlcd92_91	NLCD92 percent Woody Wetlands
HIST_nlcd92_92	NLCD92 percent Emergent Herbaceous Wetlands
HIST_sum_nlcd92_urban	NLCD92 percent all urban classes (21+22+23)
HIST_sum_nlcd92_forest	NLCD92 percent forest classes (41+42+43)
HIST_sum_nlcd92_allveg	NLCD92 percent all veg. classes (41+42+43+51+61+81+82+85+91+92)
HIST_sum_nlcd92_ag	NLCD92 percent agriculture classes (81+82)
HIST_sum_nlcd92_allnatveg	NLCD92 percent all natural veg. classes (41+42+43+51+71+91+92)
HIST_giras_11	GIRAS percent residential
HIST_giras_12	GIRAS percent commercial and services
HIST_giras_13	GIRAS percent industrial
HIST_giras_14	GIRAS percent transportation, communication and utilities
HIST_giras_15	GIRAS percent industrial and commercial complexes
HIST_giras_16	GIRAS percent mixed urban or built-up land
HIST_giras_17	GIRAS percent other urban or built-up land
HIST_giras_21	GIRAS percent cropland and pasture
HIST_giras_22	GIRAS percent orchards, groves, vineyards, nurseries
HIST_giras_24	GIRAS percent other agricultural land
HIST_giras_41	GIRAS percent deciduous forest
HIST_giras_42	GIRAS percent evergreen forest
HIST_giras_43	GIRAS percent mixed forest
HIST_giras_51	GIRAS percent streams and canals
HIST_giras_52	GIRAS percent lakes

Variable name	**Description**
HIST_giras_53	GIRAS percent reservoirs
HIST_giras_54	GIRAS percent bays and estuaries
HIST_giras_61	GIRAS percent forested wetlands
HIST_giras_62	GIRAS percent nonforested wetlands
HIST_giras_72	GIRAS percent dry salt flats
HIST_giras_75	GIRAS percent bare exposed rock
HIST_giras_76	GIRAS percent transitional areas
HIST_sum_giras_comm_ind	GIRAS sum classes commercial-industrial (12,13,15,16)
HIST_sum_giras_urban	GIRAS sum urban classes (11-17)
HIST_sum_giras_forest	GIRAS sum forest classes (41-43)
HIST_sum_giras_allveg	GIRAS sum all veg classes (21-43,61,62)
HIST_sum_giras_ag	GIRAS sum all veg classes (21-24)
HIST_sum_giras_allnatveg	GIRAS sum all veg classes (41-43,61,62)
HIST_delta_natveg_1970_2001	NLCD01_sum_allnatveg minus HIST_sum_giras_allnatveg
HIST_recr_all_times	Index of classes most similar to recreation in all 3 time periods
HIST_lowresid_all_times	Index of classes most similar to low intensity residential in all 3 time periods
HIST_highresid_all_times	Index of classes most similar to high intensity residential in all 3 time periods
HIST_commerc_all_times	Index of classes most similar to commercial in all 3 time periods
HIST_indust_all_times	Index of classes most similar to industrial in all 3 time periods
HIST_highresid_92_and_01	Percent pixels which were residential in 92 (21 or 22) AND high intensity urban in 2001 (23 or 24)
HIST_veg1970_urban01	Percent pixels which were vegetation in GIRAS and urban in NLCD01

Appendix B4

Transportation predictor variables (Geolytics, 2001; BTS, 2009, ESRI, 2009a; U.S. Census Bureau, 2009d).

Variable name	Description
TRANSP_bts_faf2_pct	Bureau of Transportation Statistics (BTS) Freight Analysis Network freeways gridded at 200m, percent
TRANSP_bts_portfac_pct	Bureau of Transportation Statistics (BTS) Port Facilities gridded at 400m, percent
TRANSP_bts_rail_pct	Bureau of Transportation Statistics (BTS) Rail lines gridded at 100m, percent
TRANSP_allroads_density	Census 2000 TIGER roads, all roads density, km/sq km
TRANSP_ratio_roadden_imperv	Ratio of road density to pct imperviousness
TRANSP_culdesac_density	Census TIGER shapefiles, cul-de-sac point locations, number/sq km.
TRANSP_a11_a17_roads_density	Census 2000 TIGER roads, A11-A17, + A63 density (interstates + cloverleafs), km/sq km
TRANSP_a21_a28_roads_density	Census 2000 TIGER roads, A21-A28 density (primary roads), km/sq km
TRANSP_a11_a38_roads_density	Census 2000 TIGER roads, A11-A38 density (interstates + primary + secondary roads), km/sq km
TRANSP_alltrans	Index of all major transportation: grid of polygon airports, and expanded interstate, rail and port facilities. Percent.

Appendix B5

Landmark predictor variables (GNIS, 2009).

Variable name	**Description**
LANDMRK_gnis_indust_density	Density of GNIS "locale" points believed to be industrial in nature (based on name), number/sq km
LANDMRK_gnis_shopping_density	Density of GNIS "locale" points believed to be shopping centers (based on name), number/sq km
LANDMRK_gnis_station_density	Density of GNIS "station" points (primarily rail or subway), number/sq km
LANDMRK_gnisconsol_recr_density	Density of consolidated GNIS points which are recreational in nature (see text), number/sq km
LANDMRK_gnisconsol_instit_density	Density of consolidated GNIS points which are institutional in nature (see text), number/sq km
LANDMRK_gnis_comind_density	Density of consolidated GNIS points which are commercial/industrial/post offices, number/sq km
LANDMRK_gnis_recr_grid	Gridded representation of recreational points, point locations expanded according to type, percent
LANDMRK_gnis_inst_grid	Gridded representation of institutional points, point locations expanded according to type, percent

Appendix B6

Proximity predictor variables. Citations as noted.

Variable name	**Description**
PROX_mean_dist_road	Mean distance of any pixel to nearest road in meters (m), based on national grids obtained from Watts et al. (2007)
PROX_city10k_dist	Linear distance of block group centroid to point location of nearest city of population > 10,000 (m). City point locations from ESRI (2009a).
PROX_city20k_dist	Linear distance of block group centroid to point location of nearest city of population > 20,000 (m). City point locations from ESRI (2009a).
PROX_city50k_dist	Linear distance of block group centroid to point location of nearest city of population > 50,000 (m). City point locations from ESRI (2009a).
PROX_city100k_dist	Linear distance of block group centroid to point location of nearest city of population > 100,000 (m). City point locations from ESRI (2009a).
PROX_city250k_dist	Linear distance of block group centroid to point location of nearest city of population > 250,000 (m). City point locations from ESRI (2009a).
PROX_interstate_road_dist	Linear distance of block group centroid to nearest interstate road line (CFCC categories a11-a17) (m)
PROX_prim_road_dist	Linear distance of block group centroid to nearest primary road line (CFCC categories a21-a28) (m)
PROX_major_road_dist	Linear distance of block group centroid to nearest major road line (CFCC categories a11-a38) (m)
PROX_allrec_gnis	Linear distance of block group centroid to nearest GNIS "recreation" point (m)
PROX_allinst_gnis	Linear distance of block group centroid to nearest GNIS "institution" point (m)
PROX_patch_2ha	Linear distance of block group centroid to nearest large (> 2ha) urban patch (m)
PROX_expand4rds_inters_patchgr2ha	Percent of land in BG consisting of a large contiguous patch (> 2ha) and within 120 m of road.
PROX_expand4rds_inters_2324	Percent of land in BG consisting of NLCD01 classes 23 and 24 and within 120 m of road.
PROX_expand8rds_inters_patchgr2ha	Percent of land in BG consisting of a large contiguous patch (> 2ha) and within 240 m of road.
PROX_expand8rds_inters_2324	Percent of land in BG consisting of NLCD01 classes 23 and 24 and within 240 m of road.
PROX_allcomind_gnis	Linear distance of block group centroid to nearest GNIS

Variable name	**Description**
	"commercial/industrial" point (m)
PROX_airport_crossing_dist	Linear distance of block group centroid to nearest airport or interstate crossing (m)
PROX_cost_10k_city	Mean cost in BG to nearest 10k city. Cost as calculated here represents a weighted distance (m) to the feature taking into account connectivity to major and minor roads and intervening urban pixels.
PROX_cost_50k_city	Mean cost in BG to nearest 50k city. Cost as calculated here represents a weighted distance (m) to the feature taking into account connectivity to major and minor roads and intervening urban pixels.
PROX_cost_100k_city	Mean cost in BG to nearest 100k city. Cost as calculated here represents a weighted distance (m) to the feature taking into account connectivity to major and minor roads and intervening urban pixels.
PROX_cost_patch_2ha	Mean cost in BG to nearest large urban patch. Cost as calculated here represents a weighted distance (m) to the feature taking into account connectivity to major and minor roads and intervening urban pixels.
PROX_cost_gnis_instit	Mean cost in BG to nearest GNIS institution location. Cost as calculated here represents a weighted distance (m) to the feature taking into account connectivity to major and minor roads and intervening urban pixels.
PROX_cost_gnis_recr	Mean cost in BG to nearest GNIS recreation location. Cost as calculated here represents a weighted distance (m) to the feature taking into account connectivity to major and minor roads and intervening urban pixels.

Appendix B7

Spatial Autocorrelation predictor variables.

Variable name	**Description**
SA_localMoran_popden	Local Moran Z-score (standard deviations) using Inverse Distance method for CENS_popden
SA_localMoran_medianrooms	Local Moran Z-score (standard deviations) using Inverse Distance method for CENS_hu_median_numb_rooms
SA_localMoran_medianyear	Local Moran Z-score (standard deviations) using Inverse Distance method for CENS_hu_median_year_struct_built
SA_localMoran_imperv	Local Moran Z-score (standard deviations) using Inverse Distance method for LC_nlcd01_imperv_mean
SA_localMoran_lc_entropy	Local Moran Z-score (standard deviations) using Inverse Distance method for SPCAT_lc_entropy
SA_localMoran_dist_road	Local Moran Z-score (standard deviations) using Inverse Distance method for PROX_mean_dist_road
SA_localMoran_nlcd2122	Local Moran Z-score (standard deviations) using Inverse Distance method for sum NLCD01 classes 21+22
SA_localMoran_nlcd2324	Local Moran Z-score (standard deviations) using Inverse Distance method for sum NLCD01 classes 23+24
SA_localMoran_allnatveg	Local Moran Z-score (standard deviations) using Inverse Distance method for LC_sum_nlcd01_allnatveg
SA_localMoran_gnis_recr	Local Moran Z-score (standard deviations) using Inverse Distance method for LANDMRK_gnis_recr_grid
SA_localMoran_gnis_inst	Local Moran Z-score (standard deviations) using Inverse Distance method for LANDMRK_gnis_inst_grid
SA_localMoran_alltransp	Local Moran Z-score (standard deviations) using Inverse Distance method for TRANSP_alltrans
SA_diff_urbanbuf400m	Percentage of urban land in 400m buffered area surrounding BG minus percent in BG
SA_diff_urbanbuf800m	Percentage of urban land in 800m buffered area surrounding BG minus percent in BG
SA_diff_urbanbuf1200m	Percentage of urban land in 1200m buffered area surrounding BG minus percent in BG
SA_diff_urbanbuf1600m	Percentage of urban land in 1600m buffered area surrounding BG minus percent in BG
SA_acf1_400_1600m_bufs	Autocorrelation function over the 400, 800, 1200, 1600m buffered differences
SA_acf2_400_1600m_bufs	Autocorrelation function over the 400, 800, 1200, 1600m buffered percents

Appendix B8

Spatial pattern predictor variables derived from NLCD01 categorical land cover data.

Variable name	**Description**
SPCAT_np_22_24	Fragstats NP variable: number of patches of aggregated NLCD01 classes 22-24
SPCAT_pd_22_24	Fragstats PD variable: patch density of aggregated NLCD01 classes 22-24
SPCAT_lpi_22_24	Fragstats LPI variable: largest patch index of aggregated NLCD01 classes 22-24
SPCAT_ed_22_24	Fragstats ED variable: edge density of aggregated NLCD01 classes 22-24
SPCAT_area_mn_22_24	Fragstats AREA_MN variable: mean area of patches of aggregated NLCD01 classes 22-24
SPCAT_area_sd_22_24	Fragstats AREA_SD variable: std dev of area of patches of aggregated NLCD01 classes 22-24
SPCAT_area_cv_22_24	Fragstats AREA_CV variable: cv of area of patches of aggregated NLCD01 classes 22-24
SPCAT_frac_mn_22_24	Fragstats FRAC_MN variable: fractal dimension mean of patches of aggregated NLCD01 classes 22-24
SPCAT_frac_sd_22_24	Fragstats FRAC_SD variable: fractal dimension std dev of patches of aggregated NLCD01 classes 22-24
SPCAT_frac_cv_22_24	Fragstats FRAC_CV variable: fractal dimension cv of patches of aggregated NLCD01 classes 22-24
SPCAT_para_mn_22_24	Fragstats PARA_MN variable: perimeter-area ratio of patches of aggregated NLCD01 classes 22-24
SPCAT_para_sd_22_24	Fragstats PARA_SD variable: perimeter-area ratio std dev of patches of aggregated NLCD01 classes 22-24
SPCAT_para_cv_22_24	Fragstats PARA_CV variable: perimeter-area ratio cv of patches of aggregated NLCD01 classes 22-24
SPCAT_circle_mn_22_24	Fragstats CIRCLE_MN variable: smallest circumscribing circle mean of patches of aggregated NLCD01 classes 22-24
SPCAT_circle_sd_22_24	Fragstats CIRCLE_SD variable: smallest circumscribing circle std dev of patches of aggregated NLCD01 classes 22-24
SPCAT_circle_cv_22_24	Fragstats CIRCLE_CV variable: smallest circumscribing circle cv of patches of aggregated NLCD01 classes 22-24
SPCAT_cohesion_22_24	Fragstats COHESION variable: cohesion index of patches of aggregated NLCD01 classes 22-24
SPCAT_np_21	Fragstats NP variable: number of patches of NLCD01 class 21
SPCAT_pd_21	Fragstats PD variable: patch density of NLCD01 class 21
SPCAT_lpi_21	Fragstats LPI variable: largest patch index of NLCD01 class 21
SPCAT_ed_21	Fragstats ED variable: edge density of NLCD01 class 21
SPCAT_area_mn_21	Fragstats AREA_MN variable: mean area of patches of NLCD01 class 21

Variable name	**Description**
SPCAT_area_sd_21	Fragstats AREA_SD variable: std dev of area of patches of NLCD01 class 21
SPCAT_area_cv_21	Fragstats AREA_CV variable: cv of area of patches of NLCD01 class 21
SPCAT_frac_mn_21	Fragstats FRAC_MN variable: fractal dimension mean of patches of NLCD01 class 21
SPCAT_frac_sd_21	Fragstats FRAC_SD variable: fractal dimension std dev of patches of NLCD01 class 21
SPCAT_frac_cv_21	Fragstats FRAC_CV variable: fractal dimension cv of patches of NLCD01 class 21
SPCAT_para_mn_21	Fragstats PARA_MN variable: perimeter-area ratio of patches of NLCD01 class 21
SPCAT_para_sd_21	Fragstats PARA_SD variable: perimeter-area ratio std dev of patches of NLCD01 class 21
SPCAT_para_cv_21	Fragstats PARA_CV variable: perimeter-area ratio cv of patches of NLCD01 class
SPCAT_circle_mn_21	Fragstats CIRCLE_MN variable: smallest circumscribing circle mean of patches of NLCD01 class
SPCAT_circle_sd_21	Fragstats CIRCLE_SD variable: smallest circumscribing circle std dev of patches of NLCD01 class 21
SPCAT_circle_cv_21	Fragstats CIRCLE_CV variable: smallest circumscribing circle cv of patches of NLCD01 class 21
SPCAT_cohesion_21	Fragstats COHESION variable: cohesion index of patches of NLCD01 class 21
SPCAT_patch_2ha_pct	Percent of land consisting of urban patches > 2 ha in size
SPCAT_riit21_interior	Riitters et al. (2000): :"Interior" pixels of class 21, percent
SPCAT_riit21_transitional	Riitters et al. (2000): :"Transitional" pixels of class 21, percent
SPCAT_riit21_edge	Riitters et al. (2000): :"Edge" pixels of class 21, percent
SPCAT_riit22_interior	Riitters et al. (2000): :"Interior" pixels of class 22, percent
SPCAT_riit22_transitional	Riitters et al. (2000): :"Transitional" pixels of class 22, percent
SPCAT_riit22_edge	Riitters et al. (2000): :"Edge" pixels of class 22, percent
SPCAT_riit23_interior	Riitters et al. (2000): :"Interior" pixels of class 23, percent
SPCAT_riit23_transitional	Riitters et al. (2000): :"Transitional" pixels of class 23, percent
SPCAT_riit23_edge	Riitters et al. (2000): :"Edge" pixels of class 23, percent
SPCAT_riit24_interior	Riitters et al. (2000): :"Interior" pixels of class 24, percent
SPCAT_riit24_transitional	Riitters et al. (2000): :"Transitional" pixels of class 24, percent
SPCAT_riit24_edge	Riitters et al. (2000): :"Edge" pixels of class 24, percent
SPCAT_riit2324_interior	Riitters et al. (2000): :"Interior" pixels of aggregated classes 23+24, percent
SPCAT_riit2324_transitional	Riitters et al. (2000): :"Transitional" pixels of aggregated classes 23+24, percent
SPCAT_riit2324_edge	Riitters et al. (2000): :"Edge" pixels of aggregated classes 23+24, percent
SPCAT_riitnatveg_interior	Riitters et al. (2000): :"Interior" pixels of aggregated class for all natural vegetation, percent

Variable name	**Description**
SPCAT_riitnatveg_transitional	Riitters et al. (2000): :"Transitional" pixels of aggregated class for all natural vegetation, percent
SPCAT_expan_zone_diff_21	Percent diff between random sample class 21 pixels and percent in expanded (by 3) zone
SPCAT_expan_zone_diff_22	Percent diff between random sample class 22 pixels and percent in expanded (by 3) zone
SPCAT_expan_zone_diff_23	Percent diff between random sample class 23 pixels and percent in expanded (by 3) zone
SPCAT_expan_zone_diff_24	Percent diff between random sample class 24 pixels and percent in expanded (by 3) zone
SPCAT_riitnatveg_edge	Riitters et al. (2000): :"Edge" pixels of aggregated class for all natural vegetation, percent
SPCAT_lc_entropy	Land cover variety (entropy; Odum, 1971) based on Anderson Level I classes,unitless
SPCAT_lc_zonalvariety	Zonal variety (# of unique values) of Anderson Level II lc
SPCAT_foc_annulus	Mean percent pixels in class 22-24 which intersect with focal annulus (3-6 pixels) of vegetation (incl. class 21), and at least 80% vegetation in focal window
SPCAT_flattening_22_24	Aggregated classes 22-24 mean semi-major divided by semi-minor axis (from Grid Zonalgeometry)
SPCAT_shape_index_22_24	Aggregated classes 22-24 area divided by perimeter squared, unitless, higher values = more compact
SPCAT_shape_index_21	Class 21 area divided by perimeter squared, unitless, higher values = more compact

Appendix B9

Spatial pattern predictor variables derived from NLCD01 impervious surface continuous data layer.

Variable name	Description
SPCON_is_slope_max	Maximum slope of imperviousness in BG, percent. Slope is the mean rate of change between each pixel and its 8 nearest neighbors.
SPCON_is_slope_mean	Mean slope of imperviousness in BG, percent
SPCON_is_slope_std	Std dev of slope of imperviousness in BG, percent
SPCON_slopeclass1_mean	Mean slope_class1 imperviousness (see text) in BG, percent
SPCON_slopeclass2_mean	Mean slope_class2 imperviousness (see text) in BG, percent
SPCON_slopeclass3_mean	Mean slope_class3 imperviousness (see text) in BG, percent
SPCON_slopeclass4_mean	Mean slope_class4 imperviousness (see text) in BG, percent
SPCON_np_slopeclass1	Fragstats NP variable: number of patches of slope_class1
SPCON_pd_slopeclass1	Fragstats PD variable: patch density of slope_class1
SPCON_lpi_slopeclass1	Fragstats LPI variable: largest patch index of slope_class1
SPCON_ed_slopeclass1	Fragstats ED variable: edge density of slope_class1
SPCON_area_mn_slopeclass1	Fragstats AREA_MN variable: mean area of patches of slope_class1
SPCON_area_sd_slopeclass1	Fragstats AREA_SD variable: std dev of area of patches of slope_class1
SPCON_area_cv_slopeclass1	Fragstats AREA_CV variable: cv of area of patdches of slope_class1
SPCON_frac_mn_slopeclass1	Fragstats FRAC_MN variable: fractal dimension mean of patches of slope_class1
SPCON_frac_sd_slopeclass1	Fragstats FRAC_SD variable: fractal dimension std dev of patches of slope_class1
SPCON_frac_cv_slopeclass1	Fragstats FRAC_CV variable: fractal dimension cv of patches of slope_class1
SPCON_para_mn_slopeclass1	Fragstats PARA_MN variable: perimeter-area ratio of patches of slope_class1
SPCON_para_sd_slopeclass1	Fragstats PARA_SD variable: perimeter-area ratio std dev of patches of slope_class1
SPCON_para_cv_slopeclass1	Fragstats PARA_CV variable: perimeter-area ratio cv of patches of slope_class1
SPCON_circle_mn_slopeclass1	Fragstats CIRCLE_MN variable: smallest circumscribing circle mean of patches of slope_class1
SPCON_circle_sd_slopeclass1	Fragstats CIRCLE_SD variable: smallest circumscribing circle std dev of patches of slope_class1
SPCON_circle_cv_slopeclass1	Fragstats CIRCLE_CV variable: smallest circumscribing circle cv of patches of slope_class1
SPCON_cohesion_slopeclass1	Fragstats COHESION variable: cohesion index of patches of slope_class1
SPCON_np_slopeclass2	Fragstats NP variable: number of patches of slope_class2
SPCON_pd_slopeclass2	Fragstats PD variable: patch density of slope_class2
SPCON_lpi_slopeclass2	Fragstats LPI variable: largest patch index of slope_class2
SPCON_ed_slopeclass2	Fragstats ED variable: edge density of slope_class2
SPCON_area_mn_slopeclass2	Fragstats AREA_MN variable: mean area of patches of slope_class2
SPCON_area_sd_slopeclass2	Fragstats AREA_SD variable: std dev of area of patches of slope_class2
SPCON_area_cv_slopeclass2	Fragstats AREA_CV variable: cv of area of patdches of slope_class2
SPCON_frac_mn_slopeclass2	Fragstats FRAC_MN variable: fractal dimension mean of patches of slope_class2
SPCON_frac_sd_slopeclass2	Fragstats FRAC_SD variable: fractal dimension std dev of patches of slope_class2
SPCON_frac_cv_slopeclass2	Fragstats FRAC_CV variable: fractal dimension cv of patches of slope_class2

Variable name	Description
SPCON_para_mn_slopeclass2	Fragstats PARA_MN variable: perimeter-area ratio of patches of slope_class2
SPCON_para_sd_slopeclass2	Fragstats PARA_SD variable: perimeter-area ratio std dev of patches of slope_class2
SPCON_para_cv_slopeclass2	Fragstats PARA_CV variable: perimeter-area ratio cv of patches of slope_class2
SPCON_circle_mn_slopeclass2	Fragstats CIRCLE_MN variable: smallest circumscribing circle mean of patches of slope_class2
SPCON_circle_sd_slopeclass2	Fragstats CIRCLE_SD variable: smallest circumscribing circle std dev of patches of slope_class2
SPCON_circle_cv_slopeclass2	Fragstats CIRCLE_CV variable: smallest circumscribing circle cv of patches of slope_class2
SPCON_cohesion_slopeclass2	Fragstats COHESION variable: cohesion index of patches of slope_class2
SPCON_np_slopeclass3	Fragstats NP variable: number of patches of slope_class3
SPCON_pd_slopeclass3	Fragstats PD variable: patch density of slope_class3
SPCON_lpi_slopeclass3	Fragstats LPI variable: largest patch index of slope_class3
SPCON_ed_slopeclass3	Fragstats ED variable: edge density of slope_class3
SPCON_area_mn_slopeclass3	Fragstats AREA_MN variable: mean area of patches of slope_class3
SPCON_area_sd_slopeclass3	Fragstats AREA_SD variable: std dev of area of patches of slope_class3
SPCON_area_cv_slopeclass3	Fragstats AREA_CV variable: cv of area of patdches of slope_class3
SPCON_frac_mn_slopeclass3	Fragstats FRAC_MN variable: fractal dimension mean of patches of slope_class3
SPCON_frac_sd_slopeclass3	Fragstats FRAC_SD variable: fractal dimension std dev of patches of slope_class3
SPCON_frac_cv_slopeclass3	Fragstats FRAC_CV variable: fractal dimension cv of patches of slope_class3
SPCON_para_mn_slopeclass3	Fragstats PARA_MN variable: perimeter-area ratio of patches of slope_class3
SPCON_para_sd_slopeclass3	Fragstats PARA_SD variable: perimeter-area ratio std dev of patches of slope_class3
SPCON_para_cv_slopeclass3	Fragstats PARA_CV variable: perimeter-area ratio cv of patches of slope_class3
SPCON_circle_mn_slopeclass3	Fragstats CIRCLE_MN variable: smallest circumscribing circle mean of patches of slope_class3
SPCON_circle_sd_slopeclass3	Fragstats CIRCLE_SD variable: smallest circumscribing circle std dev of patches of slope_class3
SPCON_circle_cv_slopeclass3	Fragstats CIRCLE_CV variable: smallest circumscribing circle cv of patches of slope_class3
SPCON_cohesion_slopeclass3	Fragstats COHESION variable: cohesion index of patches of slope_class3
SPCON_np_slopeclass4	Fragstats NP variable: number of patches of slope_class4
SPCON_pd_slopeclass4	Fragstats PD variable: patch density of slope_class4
SPCON_lpi_slopeclass4	Fragstats LPI variable: largest patch index of slope_class4
SPCON_ed_slopeclass4	Fragstats ED variable: edge density of slope_class4
SPCON_area_mn_slopeclass4	Fragstats AREA_MN variable: mean area of patches of slope_class4
SPCON_area_sd_slopeclass4	Fragstats AREA_SD variable: std dev of area of patches of slope_class4
SPCON_area_cv_slopeclass4	Fragstats AREA_CV variable: cv of area of patdches of slope_class4
SPCON_frac_mn_slopeclass4	Fragstats FRAC_MN variable: fractal dimension mean of patches of slope_class4
SPCON_frac_sd_slopeclass4	Fragstats FRAC_SD variable: fractal dimension std dev of patches of slope_class4
SPCON_frac_cv_slopeclass4	Fragstats FRAC_CV variable: fractal dimension cv of patches of slope_class4
SPCON_para_mn_slopeclass4	Fragstats PARA_MN variable: perimeter-area ratio of patches of slope_class4
SPCON_para_sd_slopeclass4	Fragstats PARA_SD variable: perimeter-area ratio std dev of patches of slope_class4
SPCON_para_cv_slopeclass4	Fragstats PARA_CV variable: perimeter-area ratio cv of patches of slope_class4
SPCON_circle_mn_slopeclass4	Fragstats CIRCLE_MN variable: smallest circumscribing circle mean of patches of slope_class4
SPCON_circle_sd_slopeclass4	Fragstats CIRCLE_SD variable: smallest circumscribing circle std dev of patches

Variable name	**Description**
	of slope_class4
SPCON_circle_cv_slopeclass4	Fragstats CIRCLE_CV variable: smallest circumscribing circle cv of patches of slope_class4
SPCON_cohesion_slopeclass4	Fragstats COHESION variable: cohesion index of patches of slope_class4
SPCON_moran_gt50_is_30m	Grid 'Moran' function mean for all imperviousness pixels with value > 50
SPCON_moran_gt50_is_adjusted	Grid 'Moran' function mean for all imperviousness pixels with value > 50 adusted by percent class23-24 pixels
SPCON_moran_gt50_is_60m	Difference between executing Moran at 30m and for data resampled to 60m
SPCON_focal33_gt50_is_mean	Focal mean of imperviousness for pixels > 50% in 3x3 window
SPCON_focal33_gt50_is_std	Focal std dev of imperviousness for pixels > 50% in 3x3 window
SPCON_focal33_gt50_is_cv	Focal cv of imperviousness for pixels > 50% in 3x3 window
SPCON_focal77_gt50_is_mean	Focal mean of imperviousness for pixels > 50% in 7x7 window
SPCON_focal77_gt50_is_std	Focal std dev of imperviousness for pixels > 50% in 7x7 window
SPCON_focal77_gt50_is_cv	Focal cv of imperviousness for pixels > 50% in 7x7 window
SPCON_focal_diff_33_77	Difference between SPCON_focal33_gt50_is_mean and SPCON_focal77_gt50_is_mean, i.e. change in imperviousness over broader area where high imperv pixels exist
SPCON_is_variety	Variety (# of unique values) of imperviousness

Appendix B10

Miscellaneous predictor variables (USGS, 2008; Conservation Biology Institute; 2009; USGS, 2009a)

Variable name	**Description**
MISC_area_km2	Area in square km of the BG
MISC_ned30m_elev	Mean elevation in the BG, from 30m National Elevation Data, meters
MISC_ned30m_slope	Mean slope in the BG, from 30m National Elevation Data, percent
MISC_vg2000_mean	Mean annual green vegetation index, 1-km pixels, from USGS National Atlas, unitless
MISC_padcat1_2	Percent land in one of the first 2 protected categories (e.g. National Parks) from the Conservation Biology Institute
MISC_maritime	Whether the block group is adjacent to the ocean or ocean-access bay or inlet, binary value (1/0)

Appendix B11

ESRI Business Analyst predictor variables (ESRI, 2009b, NAICS Association, 2009).

Variable name	Description
ESRI_nai_indco_pt_density	Point density of NAICS codes believed to represent non-home based commercial/industrial physical ("bricks and mortar") locations, #/sq km
ESRI_numemp_nai_indco_cvr	Total number of employees in non-home based commercial/industrial locations in the block group
ESRI_nai_indco_dist	Linear distance of block group centroid to point location of nearest commercial/industrial location, meters
ESRI_nai_inst_pt_density	Point density of NAICS codes believed to represent non-home based institution physical locations, #/sq km
ESRI_numemp_nai_inst_cvr	Total number of employees in non-home based institution locations in the block group
ESRI_nai_inst_dist	Linear distance of block group centroid to point location of nearest institution location, meters

Appendix C

Correlation matrix (r values) of dependent (columns) and independent (rows) variables for the final 188 independent variables used in testing, for 10-class prediction. All values > |0.07| have p-values < 0.001. Positive correlations >= 0.50 are highlighted in red, negative correlations <= -0.50 are highlighted in blue. Variables are given in alphabetical order, which organizes them by category.

Variable	SFRES_L	SFRES_M	SFRES_S	MFRES	COMMERC	INDUST	INSTIT	TRANSP	RECR_OPEN	NON_URB AN
CENS_hu_median_numb_rooms	0.48	0.35	0.03	-0.39	-0.45	-0.13	-0.27	-0.19	-0.09	0.39
CENS_hu_median_year_struct_built	0.09	0.06	-0.06	-0.03	0.00	0.02	-0.02	-0.01	-0.07	0.06
CENS_hu_pct_bottledgas	-0.07	-0.18	-0.06	0.22	0.12	0.00	0.05	0.02	-0.05	-0.06
CENS_hu_pct_lacking_complete_plumbing	-0.12	-0.15	0.03	0.14	0.13	0.09	0.06	0.08	0.01	-0.16
CENS_huden	-0.26	-0.30	0.00	**0.60**	0.31	-0.11	0.14	0.00	-0.09	-0.46
CENS_median_hh_income	**0.52**	0.28	-0.09	-0.30	-0.30	-0.13	-0.19	-0.13	-0.06	0.35
CENS_pct_5_or_more_units_in_structure	-0.23	-0.24	-0.21	0.29	0.43	0.09	0.27	0.19	0.10	-0.21
CENS_pct_foreignborn	-0.28	-0.35	0.07	0.41	0.32	0.11	0.16	0.09	0.00	-0.45
CENS_pct_households_with_ss_income	-0.01	0.18	0.11	-0.27	-0.13	0.01	-0.07	-0.02	0.07	0.08
CENS_pct_hu_5ormore_person_household	0.13	0.03	0.08	-0.04	-0.13	0.00	-0.08	-0.08	-0.10	0.01
CENS_pct_hu_occupied	0.09	0.16	0.05	-0.13	-0.22	0.00	-0.05	-0.09	-0.03	0.08
CENS_pct_hu_one_person_household	-0.29	-0.23	-0.04	0.20	0.35	0.07	0.26	0.15	0.11	-0.25
CENS_pct_hu_owneroccupied	0.41	0.44	0.00	-0.48	-0.46	-0.11	-0.31	-0.15	-0.07	0.48
CENS_pct_nonwhite	-0.25	-0.31	0.10	0.34	0.26	0.05	0.19	0.02	0.01	-0.39
CENS_pct_pop_below_poverty_lev	-0.26	-0.32	-0.05	0.37	0.31	0.07	0.28	0.14	0.06	-0.32
CENS_pct_publictransport_to_work	-0.29	-0.37	0.01	**0.53**	0.33	-0.05	0.18	0.15	0.04	-0.47
CENS_pct_walkbike_to_work	-0.17	-0.22	-0.17	0.30	0.34	0.02	0.39	0.14	0.06	-0.24
CENS_pden_change90_00	-0.03	-0.06	-0.04	0.10	0.08	0.01	0.06	0.03	-0.03	-0.07
CENS_popden	-0.30	-0.34	0.06	**0.63**	0.29	-0.12	0.20	-0.02	-0.11	**-0.53**
HIST_commerc_all_times	-0.47	-0.35	0.15	0.32	**0.53**	0.33	0.27	0.27	-0.01	**-0.65**
HIST_delta_natveg_1970_2001	-0.07	-0.11	0.09	0.07	0.06	-0.03	0.09	0.05	0.14	-0.19
HIST_giras_11	-0.25	0.01	0.47	0.28	-0.05	-0.19	0.06	-0.18	-0.12	**-0.68**
HIST_giras_12	-0.17	-0.12	-0.08	0.04	0.42	0.21	0.21	0.15	0.08	-0.23
HIST_giras_13	-0.06	-0.07	-0.07	-0.01	0.05	0.37	0.00	0.16	0.02	-0.02
HIST_giras_14	-0.06	-0.07	-0.08	-0.01	0.16	0.02	0.04	0.39	-0.03	-0.01
HIST_giras_17	-0.01	0.00	-0.04	-0.07	-0.05	0.00	-0.01	-0.03	0.45	0.02
HIST_highresid_92_and_01	-0.48	-0.34	0.46	0.49	0.23	-0.08	0.19	-0.11	-0.12	**-0.78**
HIST_highresid_all_times	-0.41	-0.12	**0.49**	0.41	0.09	-0.14	0.14	-0.14	-0.09	**-0.83**

Variable	SFRES_L	SFRES_M	SFRES_S	MFRES	COMMERC	INDUST	INSTIT	TRANSP	RECR_OPEN	NON_URB AN
HIST_indust_all_times	**-0.51**	-0.38	0.23	0.39	0.48	0.25	0.24	0.30	-0.06	**-0.72**
HIST_nlcd92_11	-0.03	-0.09	-0.14	-0.13	-0.08	0.00	-0.07	0.12	0.04	0.43
HIST_nlcd92_21	-0.08	0.30	0.46	-0.14	-0.20	-0.20	-0.03	-0.24	-0.02	-0.40
HIST_nlcd92_22	-0.31	-0.38	0.09	**0.54**	0.33	0.01	0.21	0.01	-0.09	**-0.51**
HIST_nlcd92_23	-0.16	-0.14	-0.20	0.01	0.42	0.45	0.11	0.48	0.04	-0.12
HIST_nlcd92_31	0.03	-0.02	-0.03	-0.05	-0.04	-0.02	-0.05	0.02	0.02	0.13
HIST_nlcd92_32	0.06	-0.01	-0.08	-0.06	-0.04	0.08	-0.05	0.01	-0.02	0.16
HIST_nlcd92_33	0.09	0.03	-0.14	-0.13	-0.08	0.11	-0.09	0.06	0.02	0.30
HIST_nlcd92_85	0.07	-0.03	-0.18	-0.10	-0.04	0.11	-0.02	0.15	**0.55**	0.12
HIST_nlcd92_91	0.24	0.10	-0.26	-0.22	-0.14	0.08	-0.13	0.02	-0.03	**0.51**
HIST_nlcd92_92	0.08	-0.01	-0.18	-0.14	-0.07	0.07	-0.08	0.06	0.03	0.38
HIST_recr_all_times	0.26	0.26	-0.19	-0.24	-0.18	-0.02	-0.08	-0.03	**0.49**	0.18
HIST_sum_giras_ag	0.23	0.01	-0.17	-0.12	-0.11	-0.03	-0.07	-0.04	-0.02	0.39
HIST_sum_giras_allveg	0.44	0.13	-0.39	-0.29	-0.26	-0.04	-0.20	-0.09	0.07	**0.79**
HIST_sum_giras_comm_ind	-0.18	-0.13	-0.09	0.03	0.42	0.32	0.20	0.19	0.08	-0.23
HIST_sum_giras_urban	-0.42	-0.10	0.41	0.30	0.27	0.03	0.22	0.03	0.08	**-0.90**
HIST_sum_nlcd92_ag	0.27	-0.02	-0.21	-0.14	-0.13	-0.01	-0.10	-0.04	-0.07	**0.50**
HIST_sum_nlcd92_allveg	**0.50**	0.18	-0.42	-0.35	-0.29	-0.02	-0.20	-0.04	0.08	**0.83**
HIST_veg1970_urban01	0.27	0.22	-0.31	-0.24	-0.19	0.06	-0.15	0.04	-0.04	**0.53**
LANDMRK_gnis_comind_density	-0.06	-0.05	0.03	-0.02	0.13	0.06	0.06	0.10	0.03	-0.10
LANDMRK_gnis_indust_density	0.00	0.00	-0.04	-0.03	-0.02	0.17	-0.03	0.06	0.01	0.02
LANDMRK_gnis_inst_grid	-0.03	-0.01	-0.09	-0.02	0.12	0.02	0.08	0.02	0.30	-0.04
LANDMRK_gnis_recr_grid	-0.11	-0.04	-0.08	0.12	0.22	-0.03	0.49	-0.02	0.04	-0.23
LANDMRK_gnis_shopping_density	-0.04	-0.01	-0.03	-0.04	0.17	0.11	0.03	0.03	0.03	-0.05
LANDMRK_gnisconsol_instit_density	-0.15	-0.16	-0.06	0.23	0.24	-0.06	**0.50**	-0.01	0.01	-0.27
LANDMRK_gnisconsol_recr_density	-0.09	-0.10	-0.06	0.11	0.24	0.00	0.12	0.00	0.16	-0.16
LC_nlcd01_11	-0.03	-0.07	-0.13	-0.14	-0.09	-0.01	-0.08	0.09	0.02	0.43
LC_nlcd01_21	0.48	**0.52**	-0.22	-0.34	-0.27	-0.10	-0.13	-0.11	0.16	0.27
LC_nlcd01_22	0.16	**0.50**	0.07	-0.32	-0.25	-0.13	-0.09	-0.12	0.16	-0.01
LC_nlcd01_23	-0.48	-0.18	**0.54**	0.28	0.07	-0.05	0.11	-0.08	-0.07	**-0.71**
LC_nlcd01_24	-0.34	-0.40	0.01	0.42	**0.51**	0.23	0.24	0.24	-0.06	**-0.51**
LC_nlcd01_82	0.16	0.01	-0.17	-0.11	-0.08	0.02	-0.06	-0.01	-0.01	0.34
LC_nlcd01_90	0.36	0.13	-0.32	-0.22	-0.19	-0.02	-0.16	-0.05	-0.07	**0.62**
LC_nlcd01_95	0.03	-0.05	-0.13	-0.12	-0.08	0.02	-0.07	0.07	0.03	0.37
LC_nlcd01_imperv_mean	**-0.53**	-0.35	0.34	0.47	0.43	0.14	0.25	0.13	-0.06	**-0.86**
LC_nlcd01_imperv_range	0.18	0.27	-0.11	**-0.52**	-0.21	0.13	-0.07	0.00	0.22	0.45
LC_nlcd01_imperv_stdev	-0.03	0.28	-0.17	-0.36	-0.12	0.20	-0.04	0.14	0.24	0.31
LC_ratio_huden_imperv	-0.22	-0.28	-0.02	**0.56**	0.25	-0.15	0.10	-0.03	-0.07	-0.37
LC_ratio_popden_nlcd2324	0.11	-0.11	-0.05	0.25	0.03	-0.13	0.06	-0.07	-0.07	-0.11

Variable	SFRES_L	SFRES_M	SFRES_S	MFRES	COMMERC	INDUST	INSTIT	TRANSP	RECR_OPEN	NON_URB AN
LC_sum_nlcd01_2122	0.33	**0.57**	-0.06	-0.37	-0.29	-0.13	-0.12	-0.13	0.18	0.12
LC_sum_nlcd01_ag	0.32	0.05	-0.28	-0.19	-0.17	-0.03	-0.10	-0.07	-0.01	**0.59**
LC_sum_nlcd01_allnatveg	**0.52**	0.12	-0.41	-0.30	-0.27	-0.06	-0.21	-0.08	-0.04	**0.86**
LC_sum_nlcd01_urban	-0.45	-0.08	0.43	0.33	0.29	0.06	0.21	0.03	0.03	**-0.96**
MISC_area_km2	0.38	0.00	-0.30	-0.22	-0.20	-0.03	-0.16	-0.03	-0.06	**0.69**
MISC_maritime	-0.03	-0.08	-0.01	-0.05	-0.06	-0.04	-0.04	0.03	0.01	0.20
MISC_ned30m_elev	0.36	0.33	-0.14	-0.28	-0.27	-0.09	-0.14	-0.17	-0.10	0.39
MISC_ned30m_slope	0.17	0.12	0.07	-0.10	-0.24	-0.16	-0.09	-0.15	-0.05	0.12
MISC_padcat1_2	0.04	-0.04	-0.15	-0.02	-0.06	-0.02	-0.04	0.05	0.06	0.26
MISC_vg2000_mean	0.29	0.28	-0.06	-0.22	-0.17	-0.03	-0.08	-0.15	-0.01	0.16
PROX_airport_crossing_dist	0.05	-0.19	0.09	0.10	-0.05	-0.13	0.03	-0.15	0.01	0.01
PROX_allcomind_gnis	0.38	0.12	-0.20	-0.22	-0.27	-0.11	-0.16	-0.10	-0.07	**0.55**
PROX_allinst_gnis	0.38	0.12	-0.27	-0.28	-0.24	0.02	-0.28	-0.02	-0.06	**0.67**
PROX_allrec_gnis	0.33	0.16	-0.19	-0.24	-0.23	-0.03	-0.18	-0.04	-0.19	**0.53**
PROX_city100k_dist	0.31	0.30	-0.16	-0.41	-0.23	-0.03	-0.20	-0.10	-0.06	**0.55**
PROX_city20k_dist	0.36	0.09	-0.26	-0.22	-0.17	-0.05	-0.17	-0.05	-0.06	**0.59**
PROX_city250k_dist	0.27	0.23	-0.11	-0.39	-0.23	0.04	-0.18	-0.10	-0.07	**0.52**
PROX_cost_10k_city	0.39	0.08	-0.30	-0.24	-0.22	-0.08	-0.19	-0.06	-0.09	**0.70**
PROX_cost_50k_city	0.39	0.25	-0.30	-0.32	-0.24	-0.05	-0.21	-0.06	-0.06	**0.65**
PROX_cost_gnis_instit	0.38	0.08	-0.33	-0.29	-0.27	-0.01	-0.25	-0.02	-0.04	**0.81**
PROX_cost_gnis_recr	0.30	0.04	-0.25	-0.26	-0.26	-0.01	-0.19	-0.01	-0.13	**0.73**
PROX_expand8rds_inters_2324	-0.28	-0.25	0.07	0.26	0.38	0.09	0.18	0.26	-0.03	-0.43
PROX_interstate_road_dist	0.24	0.12	-0.07	-0.23	-0.19	-0.11	-0.11	-0.25	-0.07	0.41
PROX_major_road_dist	0.19	0.13	-0.05	-0.19	-0.21	-0.08	-0.12	-0.16	-0.05	0.34
PROX_mean_dist_road	0.27	0.00	-0.36	-0.29	-0.24	0.02	-0.16	0.05	0.05	**0.84**
PROX_patch_2ha	-0.38	-0.47	0.16	**0.50**	0.41	0.13	0.23	0.13	-0.03	**-0.61**
PROX_prim_road_dist	0.17	0.08	-0.07	-0.16	-0.15	-0.06	-0.12	-0.04	-0.03	0.31
SA_acf1_400_1600m_bufs	0.01	0.01	0.01	-0.03	-0.01	0.01	-0.02	0.01	0.01	0.01
SA_acf2_400_1600m_bufs	-0.01	-0.01	-0.01	0.01	0.07	-0.01	-0.01	0.02	-0.01	-0.02
SA_diff_urbanbuf800m	0.05	-0.20	-0.19	-0.02	-0.06	0.00	-0.02	0.01	0.08	0.43
SA_localMoran_allnatveg	0.19	-0.31	-0.13	0.20	0.06	-0.06	0.02	-0.01	-0.16	0.17
SA_localMoran_alltransp	-0.06	-0.08	-0.08	-0.02	0.16	0.01	0.08	0.36	-0.01	0.00
SA_localMoran_dist_road	-0.03	-0.27	-0.01	0.27	0.10	-0.13	0.03	-0.08	-0.18	0.01
SA_localMoran_gnis_inst	-0.06	-0.08	-0.12	0.15	0.16	-0.04	0.33	0.00	0.01	-0.11
SA_localMoran_gnis_recr	-0.04	-0.06	-0.07	0.07	0.22	-0.01	0.09	0.00	0.04	-0.08
SA_localMoran_imperv	-0.04	-0.30	-0.18	**0.51**	0.34	0.06	0.15	0.12	-0.18	-0.29
SA_localMoran_lc_entropy	-0.09	-0.27	-0.10	**0.49**	0.27	-0.01	0.09	0.07	-0.27	-0.26
SA_localMoran_medianrooms	-0.01	-0.17	-0.27	0.28	0.34	-0.02	0.25	0.23	0.03	-0.15
SA_localMoran_medianyear	0.03	-0.01	-0.03	0.00	-0.04	-0.02	-0.04	-0.02	0.01	0.09

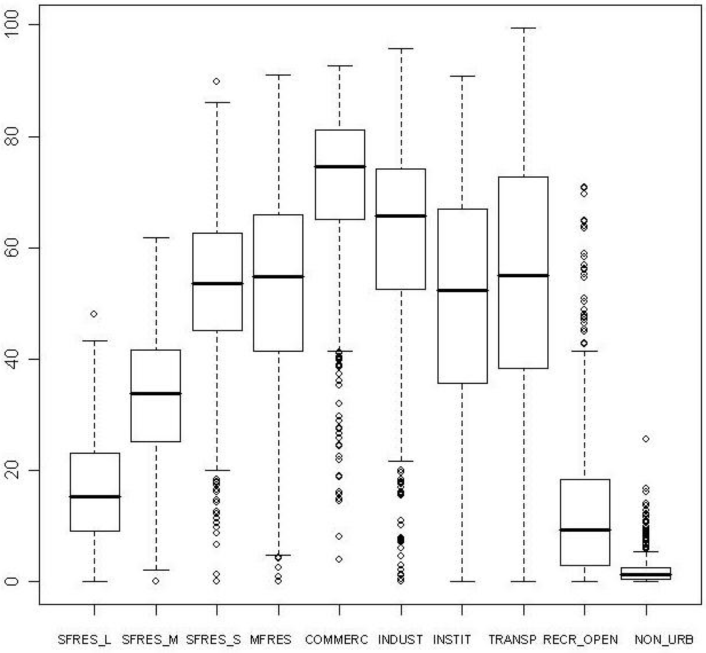
Variable	**SFRES_L**	**SFRES_M**	**SFRES_S**	**MFRES**	**COMMERC**	**INDUST**	**INSTIT**	**TRANSP**	**RECR_OPEN**	**NON_URB AN**
SA_localMoran_nlcd2122	0.05	0.01	-0.06	0.24	0.12	-0.01	0.07	0.11	-0.12	-0.29
SA_localMoran_nlcd2324	-0.02	-0.29	-0.12	**0.53**	0.27	0.02	0.14	0.07	-0.22	-0.33
SA_localMoran_popden	-0.10	-0.15	-0.18	0.48	0.22	-0.10	0.15	-0.01	-0.10	-0.22
SPCAT_area_mn_22_24	-0.22	0.15	0.09	-0.18	0.04	0.32	0.01	0.21	0.08	-0.17
SPCAT_area_sd_21	**0.54**	0.19	-0.24	-0.22	-0.20	-0.07	-0.12	-0.06	0.11	0.35
SPCAT_circle_mn_21	0.33	0.40	-0.20	**-0.50**	-0.31	0.01	-0.16	-0.05	0.17	**0.54**
SPCAT_circle_mn_22_24	0.13	0.09	-0.13	-0.15	-0.04	0.01	-0.10	0.06	0.01	0.25
SPCAT_circle_sd_21	0.42	0.45	-0.31	-0.44	-0.32	-0.02	-0.20	-0.07	0.05	**0.63**
SPCAT_ed_22_24	-0.21	-0.01	0.12	0.43	0.21	-0.15	0.14	-0.06	-0.01	**-0.64**
SPCAT_expan_zone_diff_21	0.22	0.37	-0.16	-0.44	-0.28	0.02	-0.13	-0.05	0.17	0.46
SPCAT_expan_zone_diff_22	0.13	0.03	-0.03	-0.29	-0.11	0.06	-0.07	0.01	0.13	0.31
SPCAT_expan_zone_diff_23	0.42	0.25	-0.49	-0.33	-0.10	0.08	-0.13	0.09	0.08	**0.67**
SPCAT_expan_zone_diff_24	0.05	0.11	-0.07	-0.21	-0.17	0.00	-0.06	-0.04	0.13	0.29
SPCAT_flattening_22_24	0.00	-0.02	-0.04	0.00	0.04	0.03	-0.04	0.04	-0.04	0.06
SPCAT_foc_annulus	**0.69**	0.17	-0.36	-0.26	-0.24	-0.09	-0.17	-0.09	-0.03	**0.62**
SPCAT_frac_mn_22_24	0.15	0.37	-0.14	-0.33	-0.18	0.00	-0.12	-0.01	0.14	0.29
SPCAT_frac_sd_21	**0.51**	0.44	-0.35	-0.44	-0.33	-0.04	-0.20	-0.09	0.04	**0.66**
SPCAT_lc_entropy	0.44	0.34	-0.44	-0.43	-0.31	0.03	-0.19	-0.04	0.12	**0.81**
SPCAT_lpi_21	0.35	0.29	-0.08	-0.22	-0.17	-0.08	-0.06	-0.08	0.23	0.06
SPCAT_np_21	**0.51**	0.22	-0.38	-0.30	-0.26	-0.04	-0.20	-0.06	-0.05	**0.71**
SPCAT_np_22_24	**0.56**	0.03	-0.33	-0.23	-0.22	-0.08	-0.17	-0.09	-0.05	**0.70**
SPCAT_para_mn_22_24	0.46	0.32	-0.41	-0.31	-0.28	-0.06	-0.19	-0.05	0.03	**0.69**
SPCAT_pd_21	0.21	**0.56**	-0.11	-0.35	-0.24	-0.07	-0.11	-0.09	0.10	0.18
SPCAT_pd_22_24	-0.01	-0.23	-0.06	**0.50**	0.18	-0.18	0.10	-0.05	-0.08	-0.31
SPCAT_riit21_interior	0.37	0.19	-0.14	-0.16	-0.15	-0.06	-0.07	-0.05	0.24	0.13
SPCAT_riit22_interior	0.07	0.25	0.04	-0.18	-0.15	-0.08	-0.06	-0.03	0.23	-0.01
SPCAT_riit23_edge	-0.39	-0.04	0.47	0.09	0.00	0.01	0.09	-0.09	-0.06	**-0.53**
SPCAT_riit23_interior	-0.25	-0.10	0.45	0.04	-0.07	-0.06	-0.01	-0.09	-0.07	-0.35
SPCAT_riit23_transitional	-0.48	-0.12	0.44	0.29	0.10	0.01	0.14	-0.04	-0.02	**-0.71**
SPCAT_riit2324_edge	-0.24	0.20	0.25	-0.17	-0.05	0.05	0.04	-0.01	0.06	-0.23
SPCAT_riit2324_transitional	-0.46	-0.10	0.32	0.38	0.23	-0.08	0.15	-0.03	-0.05	**-0.71**
SPCAT_riit24_interior	-0.16	-0.19	-0.11	0.08	0.41	0.33	0.11	0.36	-0.01	-0.20
SPCAT_shape_index_21	-0.17	-0.12	0.23	-0.06	0.05	0.02	0.08	0.04	0.11	-0.19
SPCAT_shape_index_22_24	-0.43	-0.46	0.34	**0.51**	0.32	-0.01	0.21	0.02	-0.12	**-0.68**
SPCON_area_cv_slopeclass1	0.37	0.09	-0.30	-0.28	-0.15	0.08	-0.15	0.09	0.10	**0.54**
SPCON_area_cv_slopeclass2	0.35	0.35	-0.38	-0.41	-0.23	0.14	-0.17	0.08	0.07	**0.64**
SPCON_area_cv_slopeclass3	0.30	0.41	-0.23	-0.47	-0.26	0.06	-0.13	-0.02	0.09	**0.52**
SPCON_area_sd_slopeclass2	0.29	0.41	-0.10	-0.36	-0.18	0.07	-0.10	0.00	0.14	0.15
SPCON_circle_cv_slopeclass3	0.19	0.26	-0.06	-0.47	-0.21	0.09	-0.08	0.01	0.15	0.37

Variable	SFRES_L	SFRES_M	SFRES_S	MFRES	COMMERC	INDUST	INSTIT	TRANSP	RECR_OPEN	NON_URB AN
SPCON_circle_mn_slopeclass2	-0.10	-0.11	0.12	0.17	0.08	-0.05	0.03	-0.01	-0.10	-0.21
SPCON_circle_sd_slopeclass3	0.22	0.32	-0.11	**-0.49**	-0.22	0.08	-0.09	0.00	0.17	0.42
SPCON_cohesion_slopeclass2	-0.08	0.10	0.27	0.03	0.00	0.03	0.08	-0.06	0.00	-0.47
SPCON_cohesion_slopeclass3	0.24	0.37	-0.14	**-0.49**	-0.24	0.06	-0.10	-0.04	0.18	0.43
SPCON_ed_slopeclass1	-0.28	-0.31	0.26	0.44	0.30	0.03	0.13	0.06	-0.08	**-0.68**
SPCON_ed_slopeclass2	-0.39	-0.08	0.35	0.39	0.27	-0.01	0.19	-0.01	-0.01	**-0.89**
SPCON_ed_slopeclass3	0.00	0.48	-0.02	-0.30	-0.11	0.03	-0.01	-0.04	0.16	-0.03
SPCON_focal_diff_33_77	-0.16	0.28	-0.04	-0.15	0.03	0.10	0.09	0.07	0.19	-0.12
SPCON_focal33_gt50_is_std	**-0.57**	-0.39	0.28	0.36	0.35	0.19	0.27	0.14	0.02	**-0.63**
SPCON_focal77_gt50_is_cv	**-0.54**	-0.16	0.29	0.20	0.21	0.15	0.23	0.09	0.10	**-0.56**
SPCON_frac_mn_slopeclass2	-0.19	-0.22	0.33	0.26	0.12	-0.10	0.09	-0.06	-0.11	-0.48
SPCON_frac_mn_slopeclass3	0.10	0.15	0.04	-0.34	-0.14	0.06	-0.05	-0.01	0.11	0.20
SPCON_frac_sd_slopeclass2	0.15	0.26	-0.07	-0.34	-0.13	0.06	-0.06	-0.01	0.17	0.22
SPCON_frac_sd_slopeclass3	0.24	0.39	-0.15	**-0.51**	-0.25	0.06	-0.10	-0.03	0.17	0.45
SPCON_is_slope_max	0.24	0.26	-0.28	**-0.49**	-0.22	0.18	-0.12	0.08	0.18	**0.62**
SPCON_is_slope_mean	-0.22	0.34	0.15	-0.12	0.02	0.10	0.10	0.03	0.16	-0.40
SPCON_is_slope_std	0.12	0.27	-0.26	-0.41	-0.17	0.17	-0.09	0.12	0.19	**0.49**
SPCON_is_variety	0.28	0.43	-0.20	**-0.58**	-0.30	0.11	-0.14	-0.02	0.18	**0.56**
SPCON_lpi_slopeclass1	-0.17	-0.27	0.03	0.33	0.30	0.05	0.06	0.14	-0.02	-0.35
SPCON_lpi_slopeclass2	-0.38	-0.17	0.46	0.35	0.21	-0.04	0.21	-0.04	-0.05	**-0.86**
SPCON_lpi_slopeclass3	-0.20	0.13	0.14	0.04	0.04	-0.06	0.09	-0.06	0.10	-0.32
SPCON_lpi_slopeclass4	-0.16	-0.07	-0.02	-0.05	0.04	0.13	0.05	0.16	0.13	0.03
SPCON_moran_gt50_is_30m	-0.20	-0.15	-0.12	0.04	0.28	0.29	0.11	0.22	0.05	-0.06
SPCON_moran_gt50_is_60m	-0.23	-0.19	0.04	0.20	0.18	0.06	0.16	0.09	-0.02	-0.22
SPCON_moran_gt50_is_adjusted	-0.47	-0.37	0.19	0.38	0.46	0.25	0.24	0.20	-0.06	**-0.66**
SPCON_para_mn_slopeclass2	0.20	0.28	-0.28	-0.37	-0.16	0.11	-0.12	0.07	0.13	**0.50**
SPCON_para_mn_slopeclass3	0.02	0.02	0.11	-0.21	-0.07	0.07	-0.02	0.01	0.04	0.07
SPCON_para_mn_slopeclass4	0.19	0.30	-0.10	-0.44	-0.20	0.06	-0.08	-0.03	0.17	0.36
SPCON_para_sd_slopeclass2	0.13	0.16	-0.08	-0.28	-0.10	0.05	-0.05	0.00	0.14	0.25
SPCON_para_sd_slopeclass3	0.21	0.29	-0.08	-0.48	-0.21	0.10	-0.08	-0.01	0.15	0.39
SPCON_pd_slopeclass1	-0.33	-0.16	0.35	0.39	0.19	-0.06	0.16	-0.05	-0.08	**-0.73**
SPCON_pd_slopeclass2	-0.15	-0.08	-0.16	0.20	0.16	0.03	0.04	0.10	0.01	-0.03
SPCON_pd_slopeclass3	-0.17	0.13	0.19	-0.05	0.03	0.03	0.08	-0.01	0.10	-0.32
SPCON_pd_slopeclass4	-0.06	0.33	-0.07	-0.24	-0.06	0.08	0.02	0.01	0.17	0.06
TRANSP_a11_a17_roads_density	-0.03	-0.04	-0.09	-0.03	0.10	0.08	-0.02	**0.52**	-0.02	-0.01
TRANSP_a11_a38_roads_density	-0.10	-0.11	-0.08	0.09	0.24	0.02	0.12	0.38	0.04	-0.16
TRANSP_a21_a28_roads_density	-0.10	-0.11	-0.05	0.11	0.18	-0.01	0.14	0.24	0.07	-0.17
TRANSP_alltrans	-0.04	-0.10	-0.12	-0.02	0.15	0.11	0.05	**0.55**	0.01	-0.02
TRANSP_bts_faf2_pct	-0.21	-0.22	-0.05	0.22	0.43	0.06	0.19	0.32	0.03	-0.35

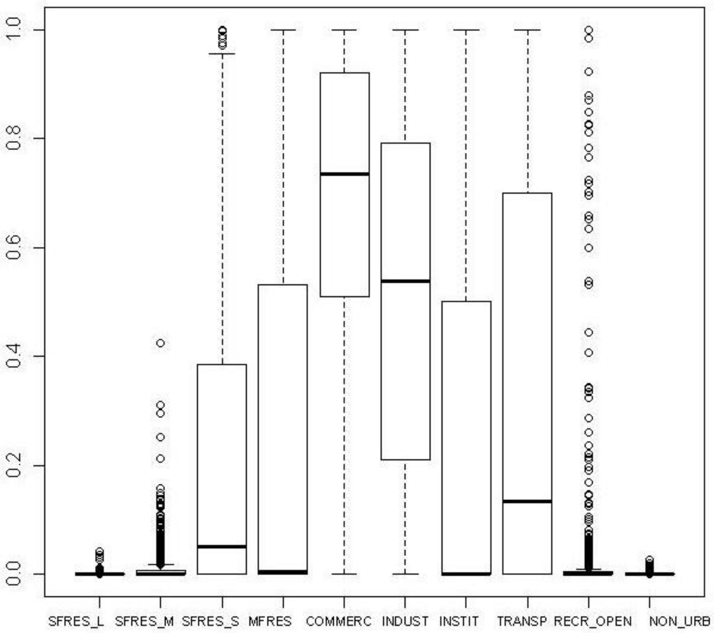
Variable	SFRES_L	SFRES_M	SFRES_S	MFRES	COMMERC	INDUST	INSTIT	TRANSP	RECR_OPEN	NON_URB AN
TRANSP_bts_portfac_pct	-0.05	-0.07	-0.08	0.01	0.06	0.08	0.04	0.21	0.01	0.03
TRANSP_bts_rail_pct	-0.14	-0.12	-0.03	0.08	0.23	0.25	0.03	0.32	0.00	-0.20
TRANSP_culdesac_density	0.03	0.06	-0.03	0.01	-0.02	-0.02	-0.05	-0.01	-0.02	0.01
TRANSP_ratio_roadden_imperv	0.38	0.02	-0.09	-0.08	-0.17	-0.26	-0.14	-0.16	-0.15	0.36

Appendix D

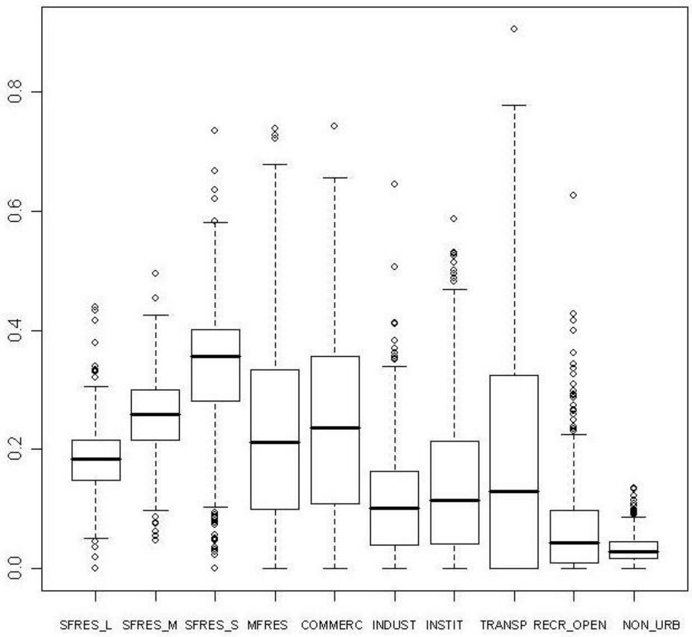
Data distributions of independent variables of interest for 600 randomly-selected reference polygons of each land use class.



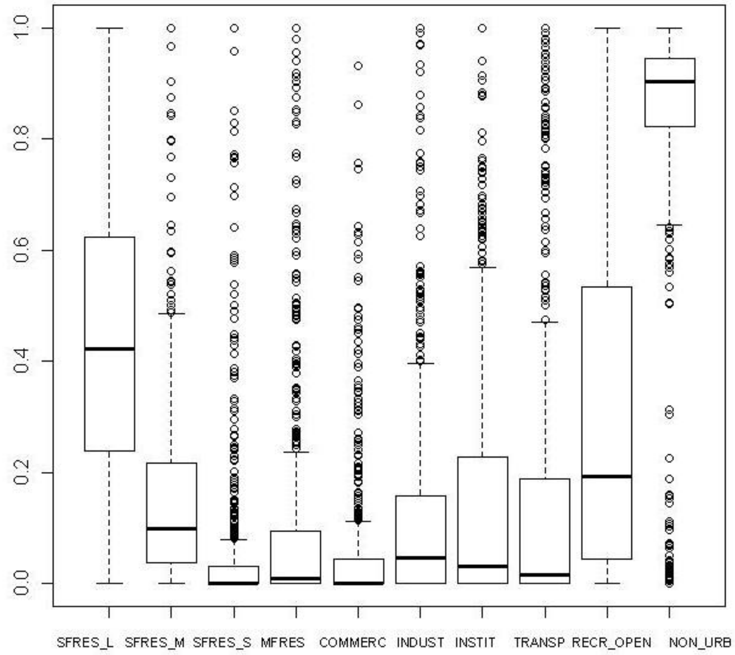
Percent impervious surfaces. (*LC_nlcd01_imperv_mean*).



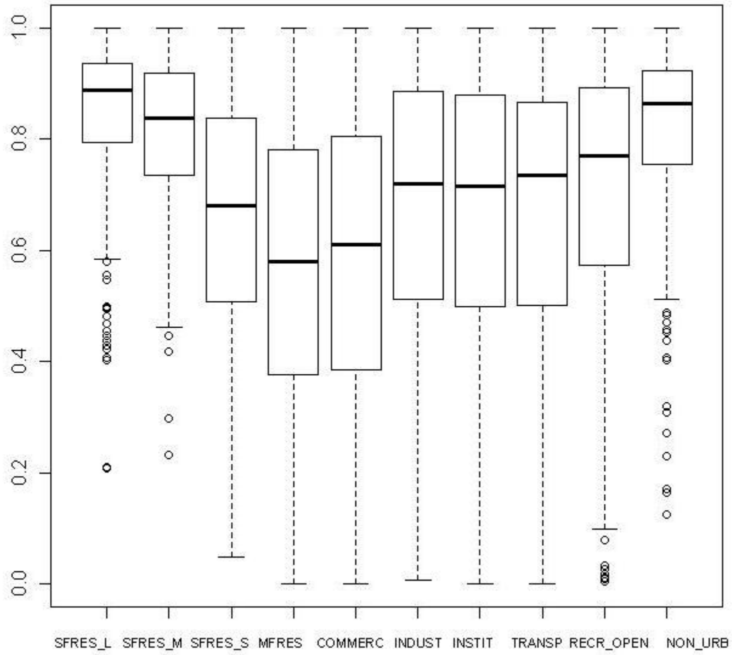
Proportion of land consisting of urban patches > 2 ha in size. (*SPCAT_patch_2ha_pct*).



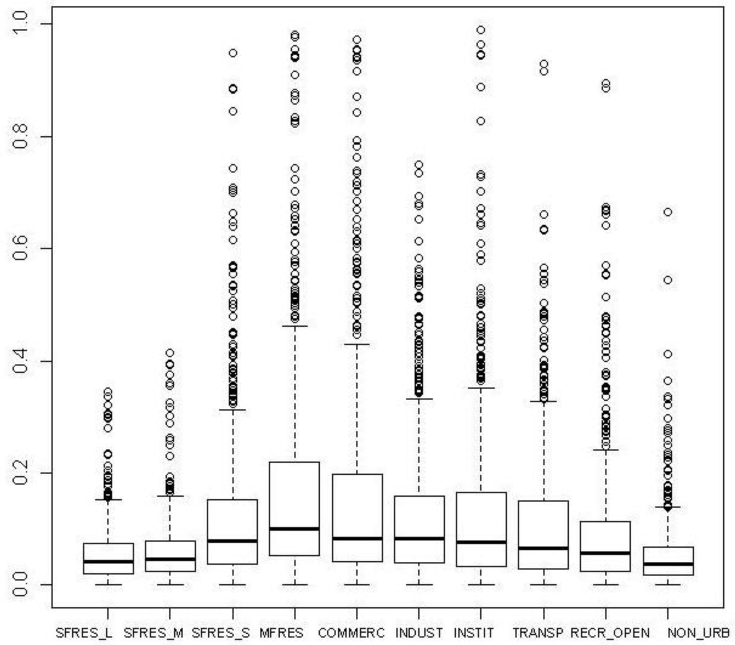
Road density, all roads, (proportion of land from gridded version (*TRANSP_allroads_density*).



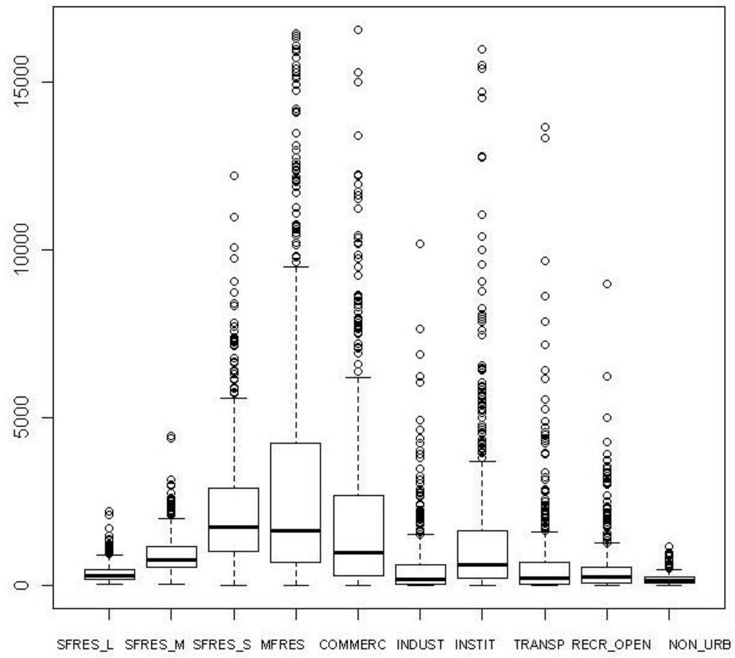
Percent vegetated land. (*LC_sum_nlcd01_allveg*).



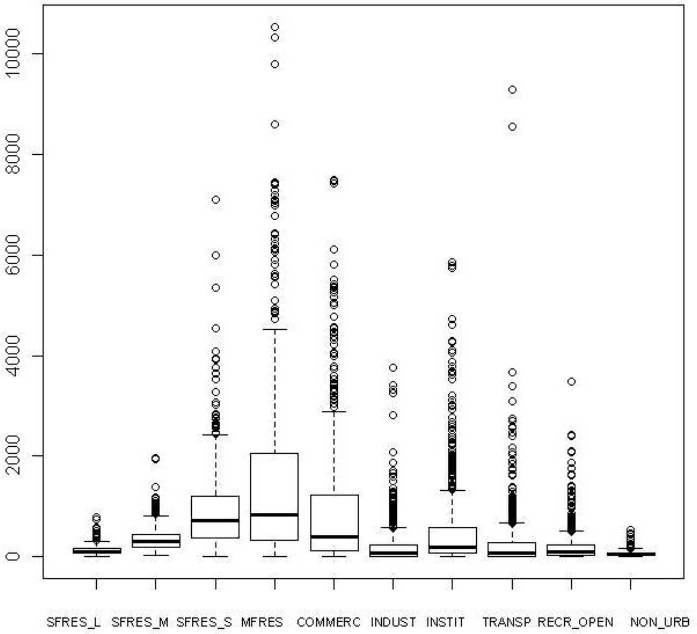
Percent housing units which are owner occupied (*CENS_pct_hu_owneroccupied*).



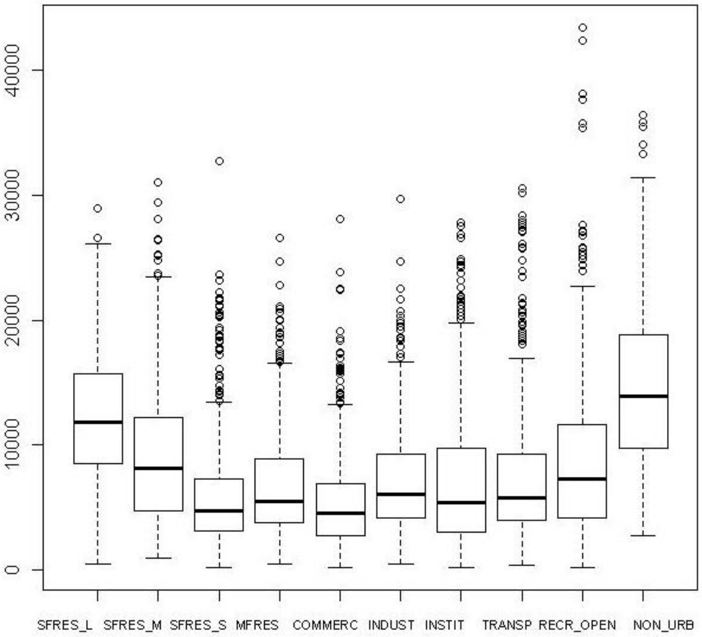
Percent population non-white (*CENS_pct_nonwhite*).



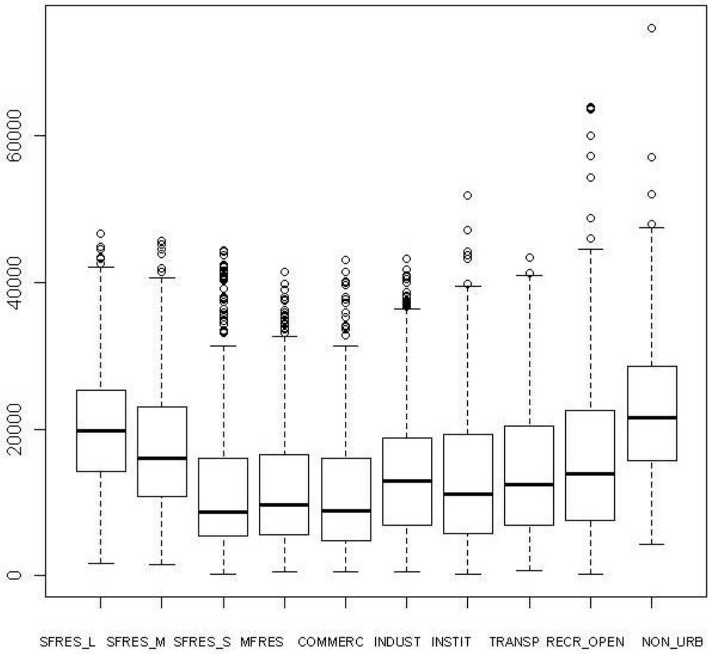
Population density (#/sq km) (*CENS_popden*).



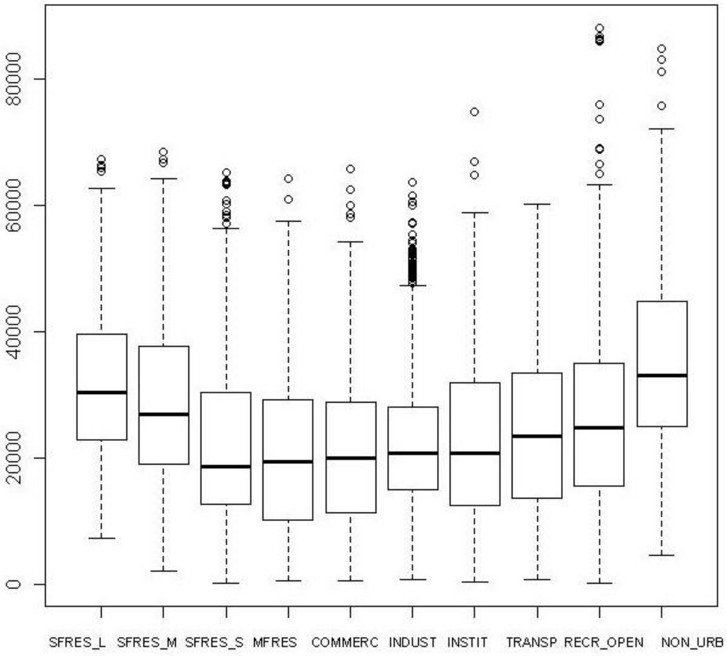
Housing unit density (#/sq km) (*CENS_huden*).



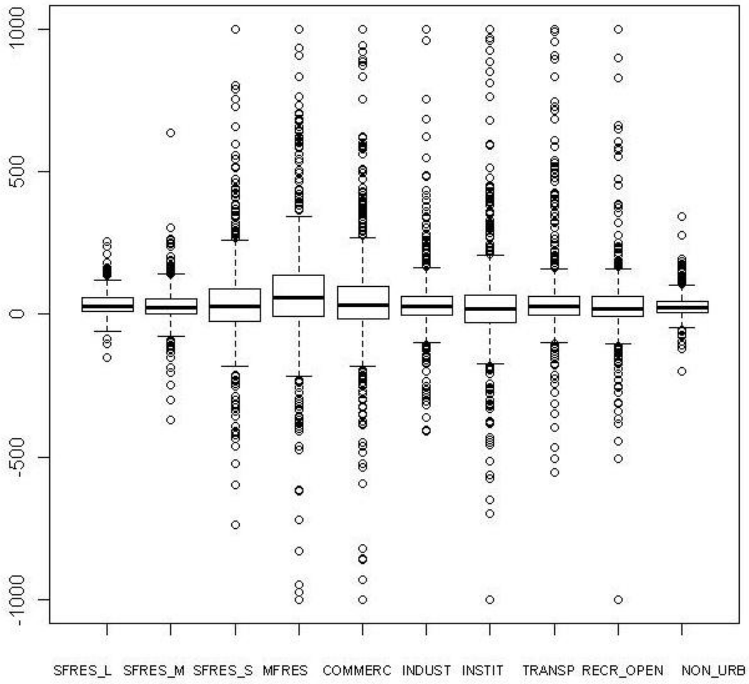
Mean cost distance to nearest city of population > 10,000 (*PROX_cost_10k_city*).



Mean cost distance to nearest city of population > 50,000 (*PROX_cost_50k_city*).



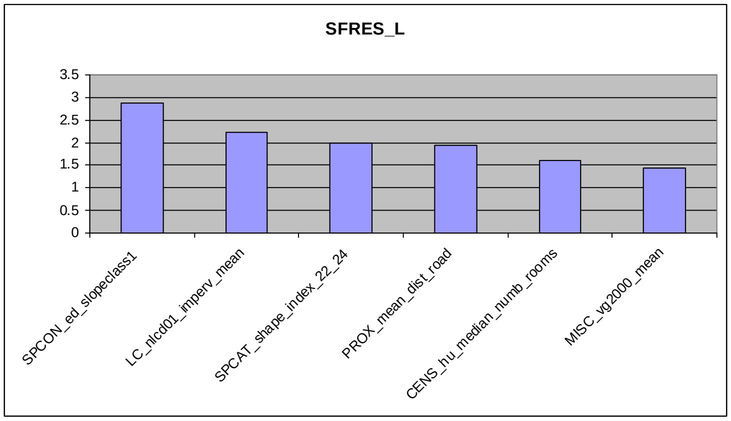
Mean cost distance to nearest city of population > 100,000 (*PROX_cost_100k_city*).

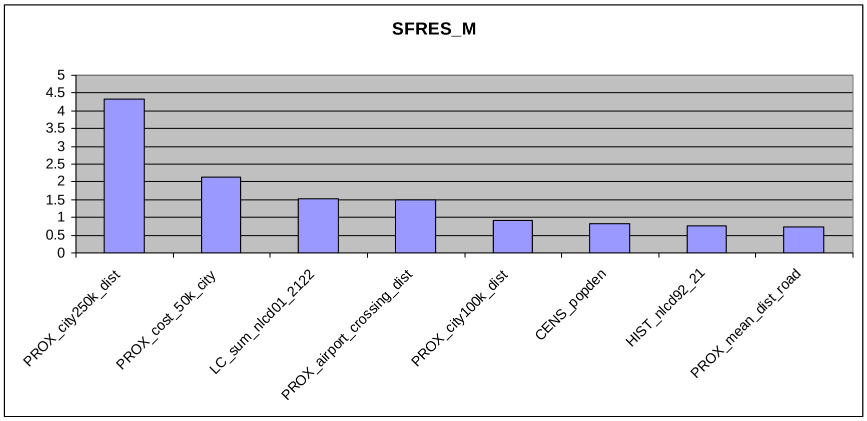


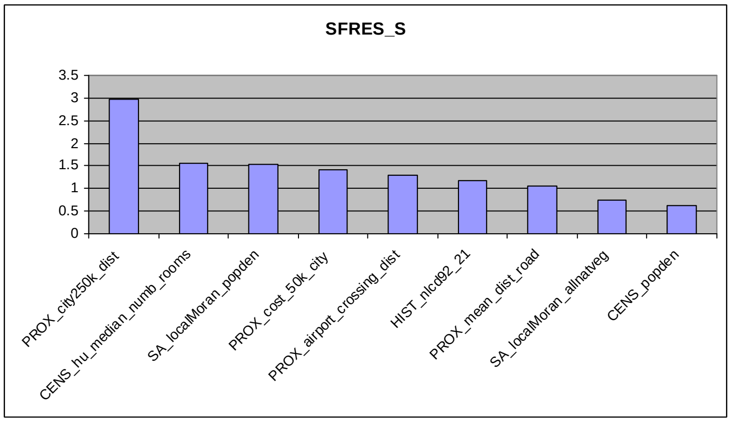
Population density change, 1990-2000 (#/sq/km) (*CENS_pden_change90_00*).

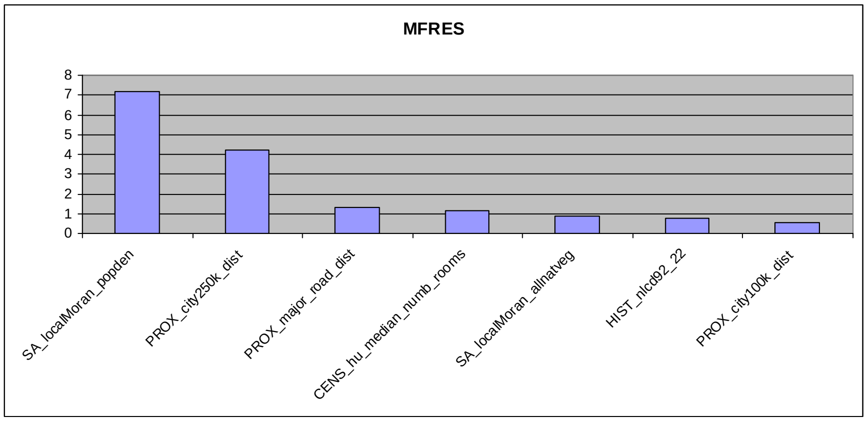
Appendix E

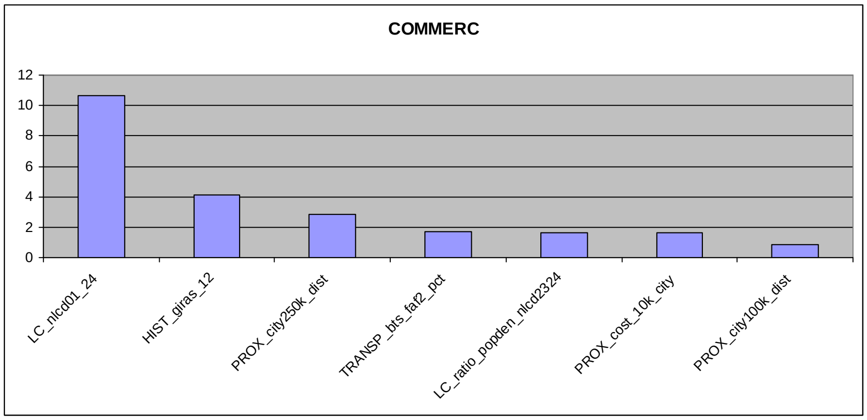
Component variables in final models for 10-class stand-alone models. The y-axis is the drop in % variance explained (decrease in r^2) if variable is left out, i.e. variables with higher values are more important to the model.

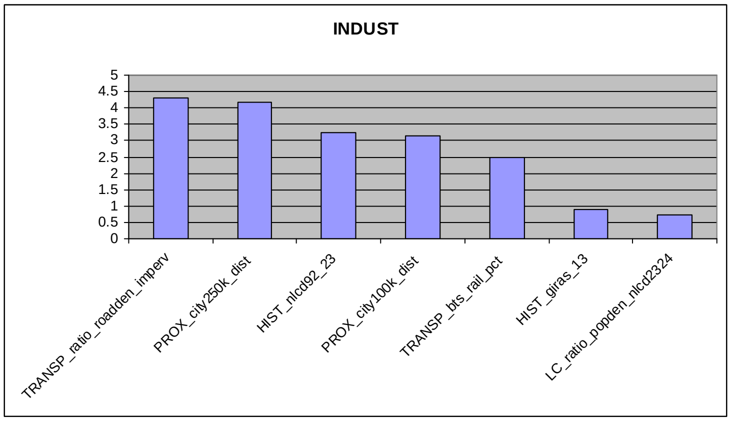


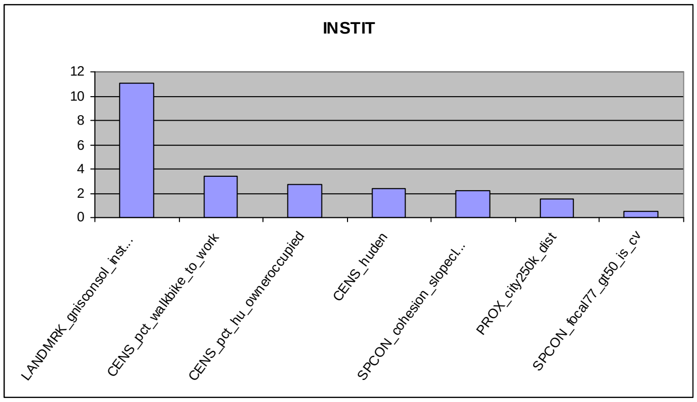


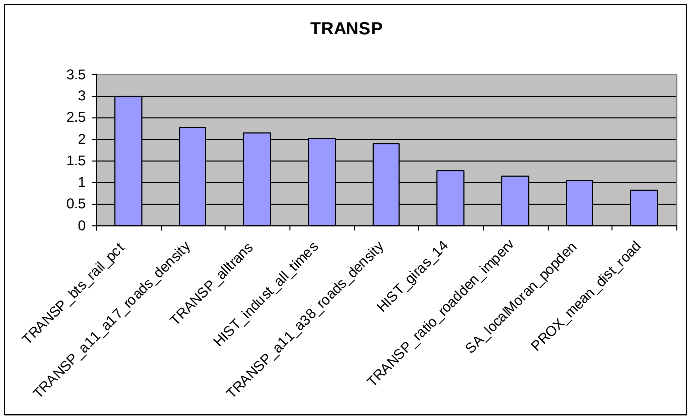


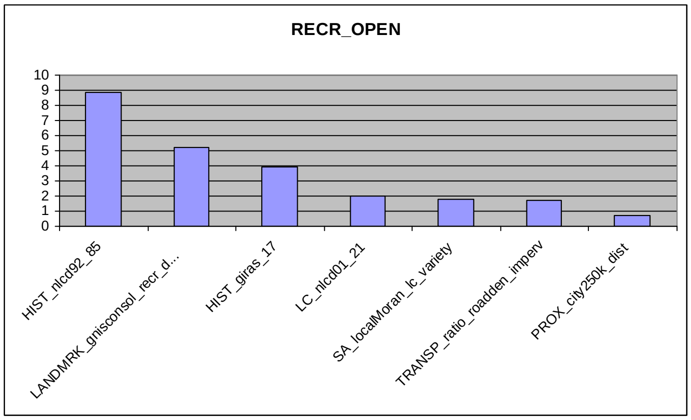

MFRES
8
7
6
5
4
3
2
1
0
SA_localMoran_popden
PROX_city250k_dist
PROX_major_road_dist
CENS_hu_median_numb_rooms
SA_localMoran_allnatveg
HIST_nlcd92_22
PROX_city100k_dist

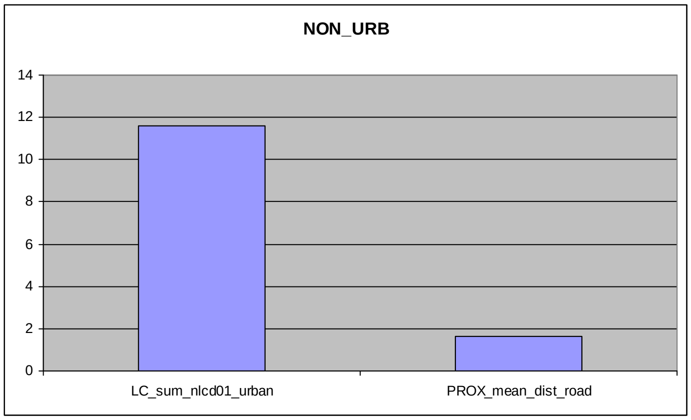

COMMERC
12
10
8
6
4
2
0
LC_nlcd01_24
HIST _giras_12
PROX_city250k_dist
TRANSP _bts_faf2_pct
LC_ratio_popden_nlcd2324
PROX_cost_10k_city
PROX_city100k_dist


INDUST
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
TRANSP _ratio_roadden_imperv
PROX_city250k_dist
HIST_nlcd92_23
PROX_city100k_dist
TRANSP _bts_rail_pct
HIST _giras_13
LC_ratio_popden_nlcd2324


INSTIT


TRANSP


RECR_OPEN


NON_URB
14
12
10
8
6
4
2
0
LC_sum_nlcd01_urban
PROX_mean_dist_road

Appendix F

Importance rankings of 188 predictor variables: lower rankings = higher importance. The rank is the result of executing a Random Forests prediction, then averaging the ranks of the %IncMSE and IncNodePurity importance measures (so ties are possible). Rankings >= 50 are not shown, as those variables had little or no useful effect on model performance. Variables are given in alphabetical order, which organizes them by category (same list and same order as in Appendix C).

Variable	SFRES_L	SFRES_M	SFRES_S	MFRES	COMMERC	INDUST	INSTIT	TRANSP	RECR_OPEN	NON_URB
CENS_hu_median_numb_rooms	8	-	18	18	10	-	-	-	-	-
CENS_hu_median_year_struct_built	-	-	-	46	-	-	-	-	-	48
CENS_hu_pct_bottledgas	-	-	-	-	-	-	-	-	-	-
CENS_hu_pct_lacking_complete_plumbing	-	-	-	-	-	-	-	-	-	-
CENS_huden	-	23	6	2	49	-	12	-	30	29
CENS_median_hh_income	22	-	43	47	-	-	19	42	31	19
CENS_pct_5_or_more_units_in_structure	-	-	13	20	7	-	30	27	-	-
CENS_pct_foreignborn	-	29	-	-	42	41	27	-	-	-
CENS_pct_households_with_ss_income	-	-	-	-	-	-	49	-	-	-
CENS_pct_hu_5ormore_person_household	28	-	-	-	-	-	-	-	-	-
CENS_pct_hu_occupied	-	-	-	-	-	-	-	-	-	-
CENS_pct_hu_one_person_household	-	-	-	39	18	-	41	-	-	-
CENS_pct_hu_owneroccupied	-	21	12	17	26	-	9	44	45	-
CENS_pct_nonwhite	-	-	46	28	14	-	25	40	-	-
CENS_pct_pop_below_poverty_lev	-	-	-	47	-	-	40	33	-	-
CENS_pct_publictransport_to_work	-	30	49	43	-	-	49	11	33	39
CENS_pct_walkbike_to_work	-	-	31	-	-	-	3	-	-	-
CENS_pden_change90_00	25	-	-	42	44	29	22	-	40	-
CENS_popden	-	17	10	3	-	-	-	-	-	-

Variable	SFRES_L	SFRES_M	SFRES_S	MFRES	COMMERC	INDUST	INSTIT	TRANSP	RECR_OPEN	NON_URB
HIST_commerc_all_times	11	-	-	-	1	6	-	-	-	-
HIST_delta_natveg_1970_2001	-	-	-	-	-	-	-	-	-	21
HIST_giras_11	18	15	40	-	35	-	-	-	-	17
HIST_giras_12	-	-	-	-	10	9	36	-	33	-
HIST_giras_13	-	-	-	-	-	5	-	25	-	-
HIST_giras_14	-	-	-	-	-	-	-	3	-	-
HIST_giras_17	-	-	-	-	-	-	-	-	4	-
HIST_highresid_92_and_01	-	41	20	22	-	-	-	-	-	-
HIST_highresid_all_times	33	17	2	11	30	-	-	-	-	27
HIST_indust_all_times	5	-	-	-	6	15	-	9	-	-
HIST_nlcd92_11	-	-	-	-	-	-	-	-	-	-
HIST_nlcd92_21	47	2	8	45	-	-	-	35	-	15
HIST_nlcd92_22	-	-	42	15	-	-	-	-	48	-
HIST_nlcd92_23	-	-	41	-	7	2	-	5	-	-
HIST_nlcd92_31	-	-	-	-	-	-	-	-	-	-
HIST_nlcd92_32	-	-	-	-	-	-	-	-	-	-
HIST_nlcd92_33	-	-	-	-	-	-	-	-	-	-
HIST_nlcd92_85	-	-	-	-	-	-	-	-	1	34
HIST_nlcd92_91	-	-	-	-	-	13	-	-	-	24
HIST_nlcd92_92	-	-	-	-	-	28	-	-	-	42
HIST_recr_all_times	-	-	-	-	-	-	-	-	2	46
HIST_sum_giras_ag	-	-	-	-	-	-	-	-	-	-
HIST_sum_giras_allveg	-	-	-	-	-	-	-	-	-	10
HIST_sum_giras_comm_ind	45	-	-	-	18	19	-	-	43	-
HIST_sum_giras_urban	46	-	-	-	-	-	-	-	-	3
HIST_sum_nlcd92_ag	-	-	-	-	-	-	-	-	-	-
HIST_sum_nlcd92_allveg	-	-	-	-	-	-	-	44	32	11

Variable	SFRES_L	SFRES_M	SFRES_S	MFRES	COMMERC	INDUST	INSTIT	TRANSP	RECR_OPEN	NON_URB
HIST_veg1970_urban01	-	-	-	-	-	-	-	-	-	13
LANDMRK_gnis_comind_density	-	-	-	-	-	-	-	-	-	-
LANDMRK_gnis_indust_density	-	-	-	-	-	12	-	-	-	-
LANDMRK_gnis_inst_grid	-	-	-	-	-	-	-	-	7	-
LANDMRK_gnis_recr_grid	-	-	-	-	-	-	2	-	-	-
LANDMRK_gnis_shopping_density	-	-	-	-	-	-	-	-	-	-
LANDMRK_gnisconsol_instit_density	-	-	-	-	-	-	1	-	-	-
LANDMRK_gnisconsol_recr_density	-	-	-	-	48	-	-	-	3	-
LC_nlcd01_11	-	-	-	-	-	-	-	-	-	22
LC_nlcd01_21	49	12	-	-	-	-	-	-	-	-
LC_nlcd01_22	-	9	34	-	-	-	-	-	20	31
LC_nlcd01_23	19	42	4	-	-	-	-	-	-	-
LC_nlcd01_24	-	-	-	44	1	36	-	-	-	-
LC_nlcd01_82	-	-	-	-	-	-	-	-	-	-
LC_nlcd01_90	-	-	-	-	-	-	-	-	-	-
LC_nlcd01_95	-	-	-	-	-	-	-	-	-	-
LC_nlcd01_imperv_mean	9	-	-	40	21	48	-	-	-	13
LC_nlcd01_imperv_range	-	-	-	-	-	-	-	-	-	-
LC_nlcd01_imperv_stdev	-	37	16	47	-	45	-	20	13	5
LC_ratio_huden_imperv	-	-	34	4	46	11	8	19	-	-
LC_ratio_popden_nlcd2324	-	-	49	24	34	2	31	14	23	-
LC_sum_nlcd01_2122	16	1	-	-	-	-	-	-	-	12
LC_sum_nlcd01_ag	-	-	-	-	-	-	-	-	-	-
LC_sum_nlcd01_allnatveg	-	-	-	-	-	-	-	-	-	7
LC_sum_nlcd01_urban	-	34	-	-	-	-	-	-	-	1
MISC_area_km2	-	39	23	13	-	45	-	-	29	-
MISC_maritime	-	-	-	-	-	-	-	-	-	-

Variable	SFRES_L	SFRES_M	SFRES_S	MFRES	COMMERC	INDUST	INSTIT	TRANSP	RECR_OPEN	NON_URB
MISC_ned30m_elev	-	47	-	-	-	-	44	-	16	33
MISC_ned30m_slope	-	-	-	-	-	-	-	-	14	-
MISC_padcat1_2	-	-	-	-	-	-	-	-	-	-
MISC_vg2000_mean	3	6	36	37	27	-	33	29	33	36
PROX_airport_crossing_dist	40	49	19	21	-	-	39	26	-	49
PROX_allcomind_gnis	-	-	-	19	24	48	29	-	-	-
PROX_allinst_gnis	-	-	-	-	-	32	3	-	-	45
PROX_allrec_gnis	-	-	-	-	-	-	-	-	11	-
PROX_city100k_dist	26	13	15	14	17	21	24	27	-	-
PROX_city20k_dist	-	-	37	7	7	-	31	-	41	-
PROX_city250k_dist	19	2	9	5	20	18	9	24	19	34
PROX_cost_10k_city	-	38	17	5	14	-	37	-	-	-
PROX_cost_50k_city	36	10	25	35	49	-	-	49	-	-
PROX_cost_gnis_instit	-	-	-	-	-	7	7	31	45	9
PROX_cost_gnis_recr	-	-	-	-	-	-	18	43	-	40
PROX_expand8rds_inters_2324	-	-	-	-	36	22	-	-	-	-
PROX_interstate_road_dist	12	39	25	22	-	-	20	6	-	-
PROX_major_road_dist	-	-	21	15	40	26	-	47	-	-
PROX_mean_dist_road	11	19	4	36	25	14	13	22	18	3
PROX_patch_2ha	-	-	-	-	13	-	-	-	-	-
PROX_prim_road_dist	-	-	33	-	41	-	-	37	-	-
SA_acf1_400_1600m_bufs	-	-	-	-	-	-	27	-	-	-
SA_acf2_400_1600m_bufs	-	-	-	-	-	-	-	-	-	-
SA_diff_urbanbuf800m	21	-	-	-	-	-	-	39	-	-
SA_localMoran_allnatveg	-	16	6	12	23	34	11	-	-	31
SA_localMoran_alltransp	32	-	-	8	-	-	26	10	-	-
SA_localMoran_dist_road	-	8	29	24	31	31	14	17	-	41

Variable	SFRES_L	SFRES_M	SFRES_S	MFRES	COMMERC	INDUST	INSTIT	TRANSP	RECR_OPEN	NON_URB
SA_localMoran_gnis_inst	-	-	23	27	28	42	6	-	-	-
SA_localMoran_gnis_recr	26	-	44	26	47	-	-	-	15	47
SA_localMoran_imperv	-	44	14	9	39	-	-	16	28	-
SA_localMoran_lc_entropy	-	26	25	32	-	-	-	-	5	-
SA_localMoran_medianrooms	13	-	3	29	43	40	16	34	-	23
SA_localMoran_medianyear	-	-	45	30	31	38	-	-	-	-
SA_localMoran_nlcd2122	-	14	38	40	-	-	-	-	11	-
SA_localMoran_nlcd2324	10	46	32	10	-	-	-	20	44	-
SA_localMoran_popden	-	6	1	1	-	47	34	15	38	-
SPCAT_area_mn_22_24	-	-	-	-	-	37	43	41	-	-
SPCAT_area_sd_21	35	24	-	-	-	-	-	-	-	43
SPCAT_circle_mn_21	-	31	-	-	-	-	-	-	-	-
SPCAT_circle_mn_22_24	-	-	-	-	-	-	-	-	-	-
SPCAT_circle_sd_21	-	47	-	-	-	-	-	-	-	-
SPCAT_ed_22_24	-	28	-	-	-	35	46	31	-	18
SPCAT_expan_zone_diff_21	13	-	-	-	-	-	-	-	-	-
SPCAT_expan_zone_diff_22	-	20	-	-	-	-	-	-	-	-
SPCAT_expan_zone_diff_23	-	31	11	-	44	-	-	-	-	-
SPCAT_expan_zone_diff_24	-	-	-	-	31	-	-	-	-	-
SPCAT_flattening_22_24	-	-	-	-	-	-	-	-	-	-
SPCAT_foc_annulus	1	-	-	-	-	-	-	-	-	-
SPCAT_frac_mn_22_24	44	-	-	-	-	-	-	-	-	-
SPCAT_frac_sd_21	-	31	-	-	-	-	-	-	-	-
SPCAT_lc_entropy	-	25	22	-	-	-	-	-	42	5
SPCAT_lpi_21	-	-	-	-	-	-	-	-	27	-
SPCAT_np_21	13	4	-	-	-	-	-	-	-	-
SPCAT_np_22_24	5	-	-	-	-	-	-	-	-	-

Variable	SFRES_L	SFRES_M	SFRES_S	MFRES	COMMERC	INDUST	INSTIT	TRANSP	RECR_OPEN	NON_URB
SPCAT_para_mn_22_24	-	-	-	-	-	20	45	30	-	-
SPCAT_pd_21	-	5	-	-	-	-	-	-	-	-
SPCAT_pd_22_24	5	-	-	31	-	44	-	46	49	-
SPCAT_riit21_interior	-	-	-	-	-	-	-	-	36	-
SPCAT_riit22_interior	-	-	-	-	-	-	-	-	9	-
SPCAT_riit23_edge	-	-	-	-	-	-	-	-	-	-
SPCAT_riit23_interior	-	-	29	-	-	-	-	-	-	-
SPCAT_riit23_transitional	-	-	-	-	-	-	-	-	-	-
SPCAT_riit2324_edge	-	36	-	-	-	-	-	-	-	-
SPCAT_riit2324_transitional	-	21	-	-	-	17	-	-	-	-
SPCAT_riit24_interior	-	-	-	34	3	10	42	38	-	-
SPCAT_shape_index_21	30	-	-	-	-	-	-	-	22	-
SPCAT_shape_index_22_24	1	26	38	-	-	-	-	-	-	20
SPCON_area_cv_slopeclass1	-	-	-	-	-	-	-	-	-	-
SPCON_area_cv_slopeclass2	-	-	-	-	-	-	-	48	37	16
SPCON_area_cv_slopeclass3	-	-	-	-	-	-	23	-	-	-
SPCON_area_sd_slopeclass2	-	-	-	-	-	25	-	-	-	-
SPCON_circle_cv_slopeclass3	-	-	-	-	-	-	-	-	-	-
SPCON_circle_mn_slopeclass2	-	-	-	-	-	-	-	-	-	-
SPCON_circle_sd_slopeclass3	-	-	-	-	-	-	-	-	-	-
SPCON_cohesion_slopeclass2	41	-	-	-	-	-	17	-	-	-
SPCON_cohesion_slopeclass3	-	-	-	-	-	-	-	-	24	-
SPCON_ed_slopeclass1	3	-	-	-	-	-	34	-	-	30
SPCON_ed_slopeclass2	-	-	-	-	-	-	-	35	-	2
SPCON_ed_slopeclass3	-	11	-	-	-	-	-	-	-	26
SPCON_focal_diff_33_77	-	-	-	-	22	-	-	-	-	-
SPCON_focal33_gt50_is_std	22	-	-	-	-	32	15	-	-	-

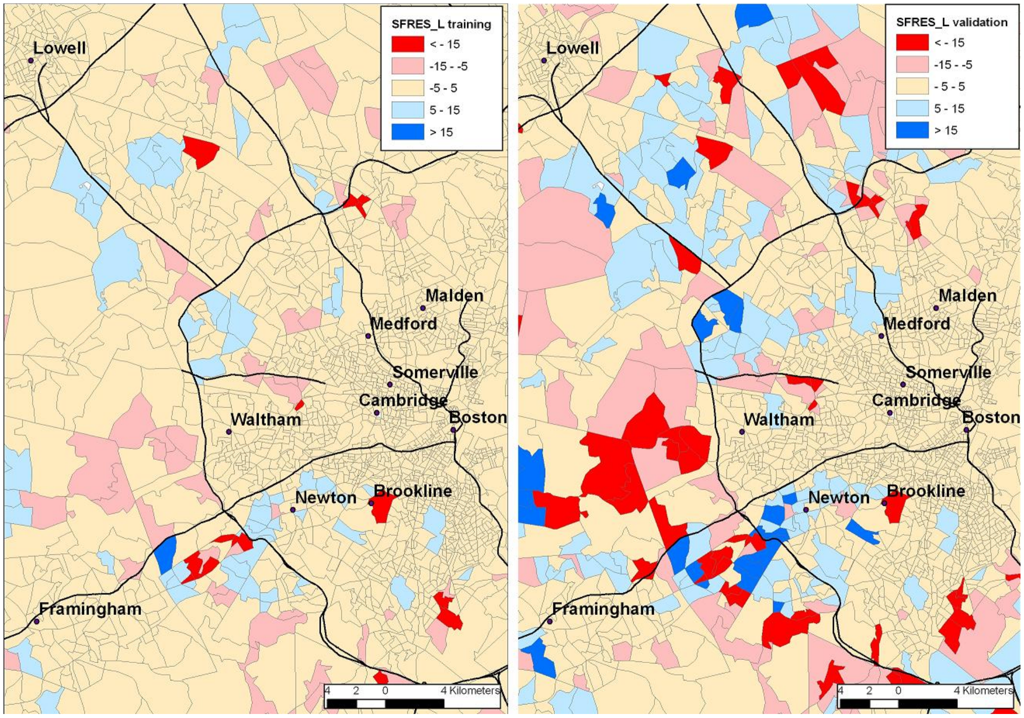
Variable	SFRES_L	SFRES_M	SFRES_S	MFRES	COMMERC	INDUST	INSTIT	TRANSP	RECR_OPEN	NON_URB
SPCON_focal77_gt50_is_cv	37	-	-	-	38	38	21	-	-	-
SPCON_frac_mn_slopeclass2	-	-	-	-	-	-	-	-	-	-
SPCON_frac_mn_slopeclass3	-	-	-	-	-	-	-	-	38	-
SPCON_frac_sd_slopeclass2	30	-	-	-	-	-	-	-	26	-
SPCON_frac_sd_slopeclass3	-	-	-	-	-	-	-	-	45	-
SPCON_is_slope_max	16	-	-	-	-	48	-	17	17	44
SPCON_is_slope_mean	-	-	-	-	-	-	-	-	-	8
SPCON_is_slope_std	-	45	-	-	-	-	-	22	6	38
SPCON_is_variety	-	42	47	32	-	23	-	-	-	-
SPCON_lpi_slopeclass1	-	-	-	-	-	-	-	-	25	-
SPCON_lpi_slopeclass2	-	-	48	-	-	-	-	-	-	37
SPCON_lpi_slopeclass3	-	35	-	-	-	-	-	-	-	-
SPCON_lpi_slopeclass4	42	-	-	-	-	-	-	-	8	-
SPCON_moran_gt50_is_30m	-	-	-	-	14	42	-	-	-	-
SPCON_moran_gt50_is_60m	-	-	-	-	-	-	-	-	-	-
SPCON_moran_gt50_is_adjusted	39	-	-	-	4	7	38	-	-	-
SPCON_para_mn_slopeclass2	-	-	-	-	-	-	-	-	-	-
SPCON_para_mn_slopeclass3	-	-	-	-	-	-	-	-	-	-
SPCON_para_mn_slopeclass4	34	-	-	-	-	-	-	-	-	-
SPCON_para_sd_slopeclass2	-	-	-	-	-	-	-	-	-	-
SPCON_para_sd_slopeclass3	-	-	-	38	-	-	-	-	-	-
SPCON_pd_slopeclass1	48	-	-	-	-	-	47	-	-	24
SPCON_pd_slopeclass2	29	-	-	-	-	30	47	-	-	-
SPCON_pd_slopeclass3	37	-	-	-	-	24	-	-	-	27
SPCON_pd_slopeclass4	-	-	-	-	-	-	-	-	21	-
TRANSP_a11_a17_roads_density	-	-	-	-	-	-	-	1	-	-
TRANSP_a11_a38_roads_density	-	-	-	-	29	26	-	8	-	-

Variable	SFRES_L	SFRES_M	SFRES_S	MFRES	COMMERC	INDUST	INSTIT	TRANSP	RECR_OPEN	NON_URB
TRANSP_a21_a28_roads_density	-	-	-	-	-	-	-	12	-	-
TRANSP_alltrans	42	-	-	-	-	-	-	1	-	-
TRANSP_bts_faf2_pct	-	-	-	-	5	15	-	13	-	-
TRANSP_bts_portfac_pct	-	-	-	-	-	-	-	-	-	-
TRANSP_bts_rail_pct	-	-	-	-	12	4	-	4	-	-
TRANSP_culdesac_density	-	-	-	-	-	-	-	-	-	-
TRANSP_ratio_roadden_imperv	22	-	28	-	37	1	3	6	10	-

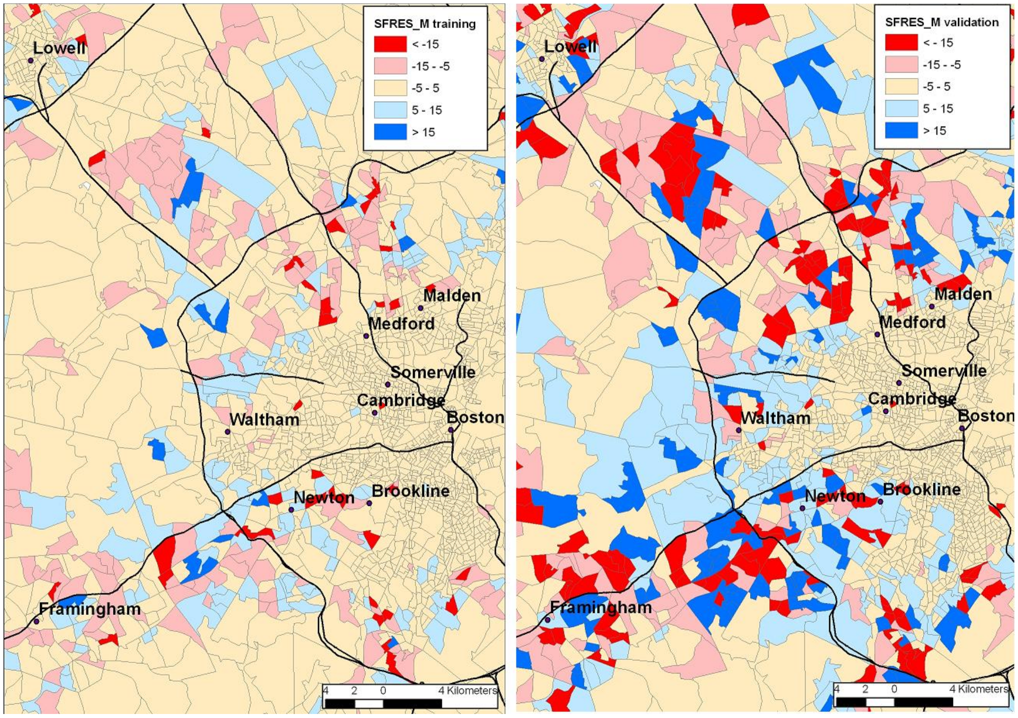
Appendix G

Mapped residuals (percent) of stand-alone 10-class models for training (left) and validation (right), for each class.

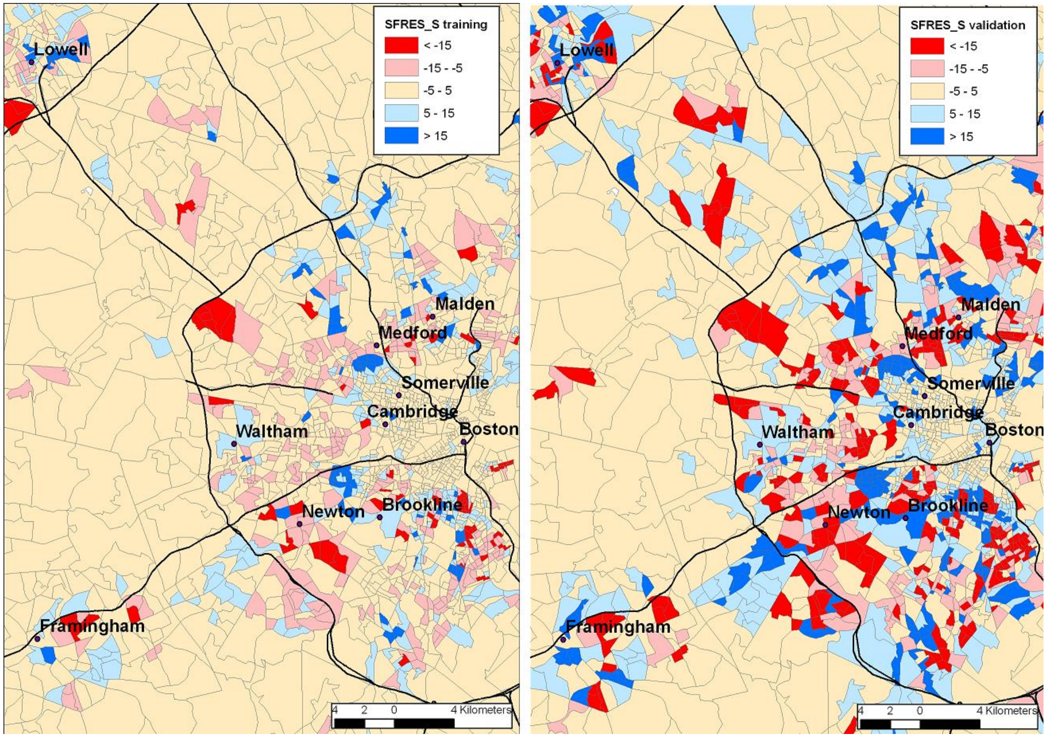
SFRES_L:



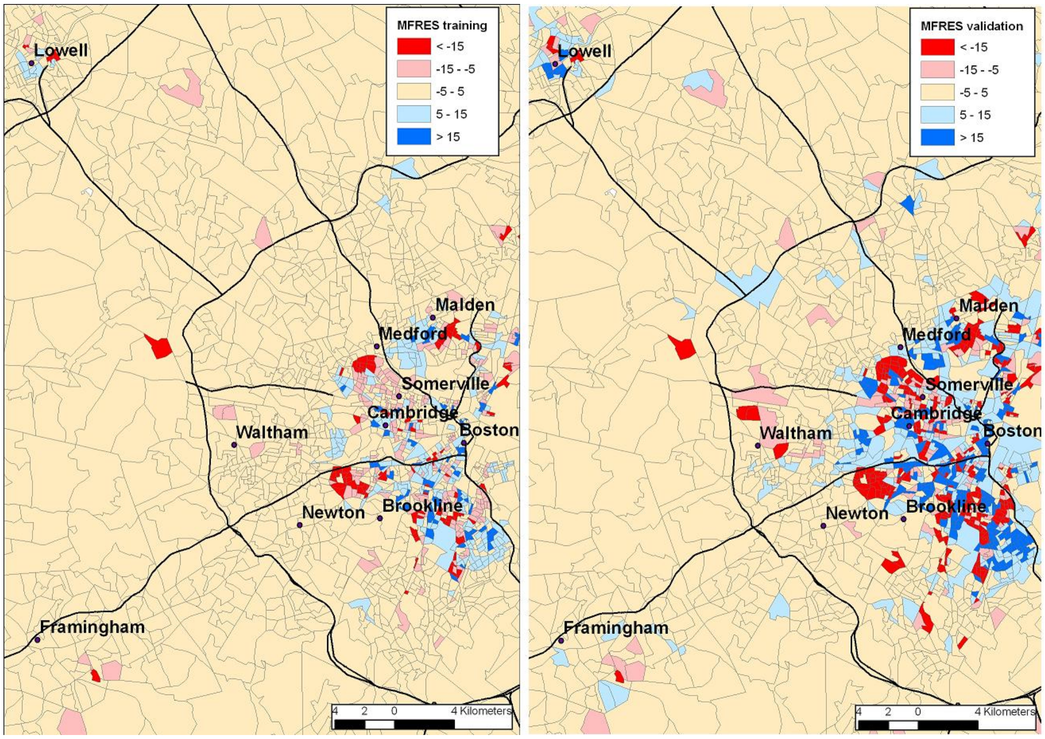
SFRES_M:



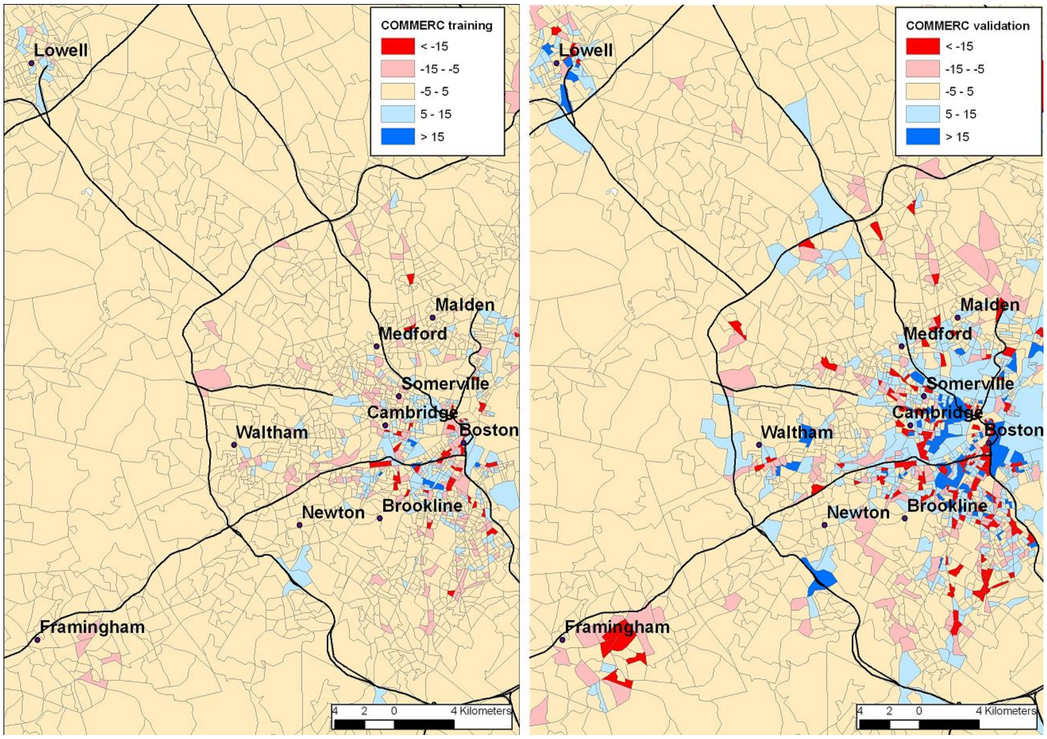
SFRES_S:



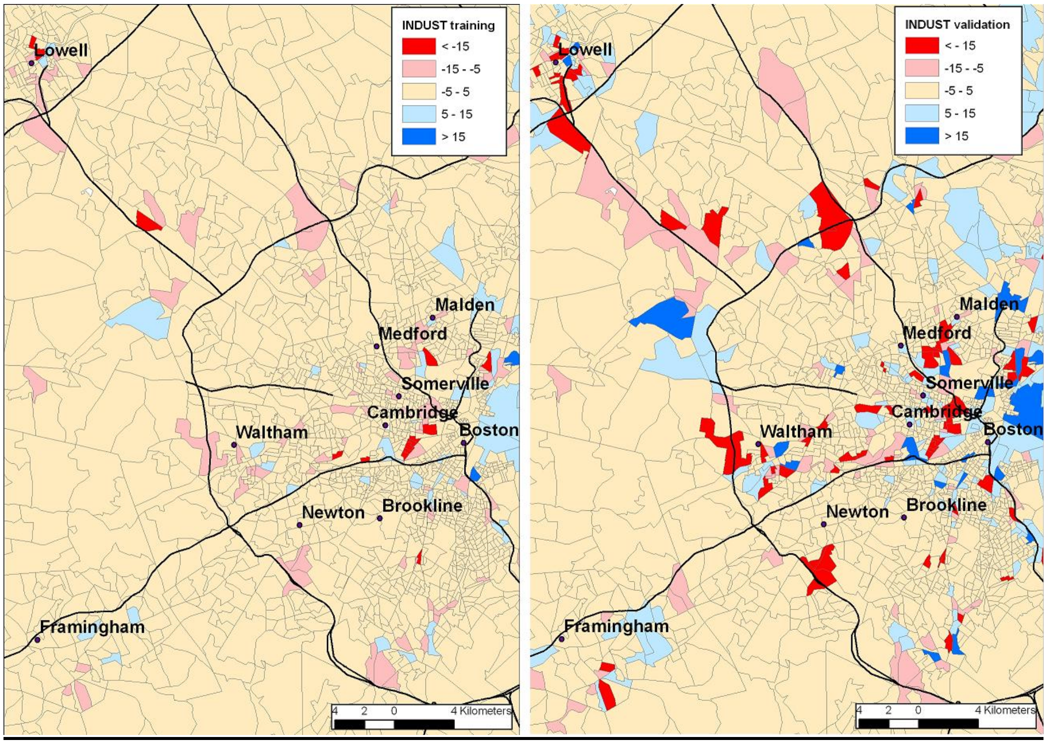
MFRES:


MFRES training
< -15
-15 - -5
-5 - 5
5 - 15
> 15
Lowell
Malden
Medford
Somerville
Cambridge
Boston
Waltham
Newton
Brookline
Framingham
4 2 0 4 Kilometers
MFRES validation
< -15
-15 - -5
-5 - 5
5 - 15
> 15
Lowell
Malden
Medford
Somerville
Cambridge
Boston
Waltham
Newton
Brookline
Framingham
4 2 0 4 Kilometers

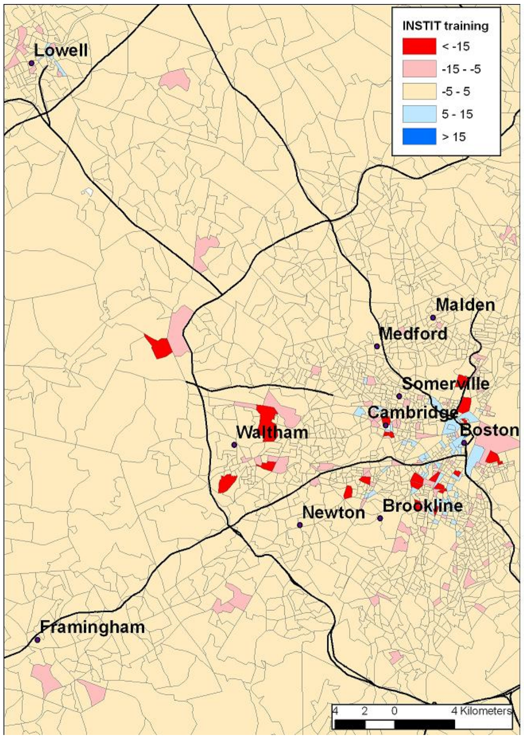
COMMERC:

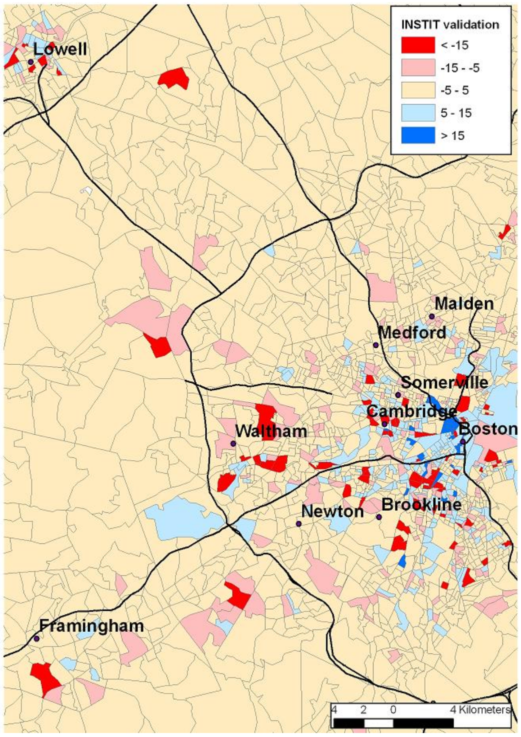


INDUST:

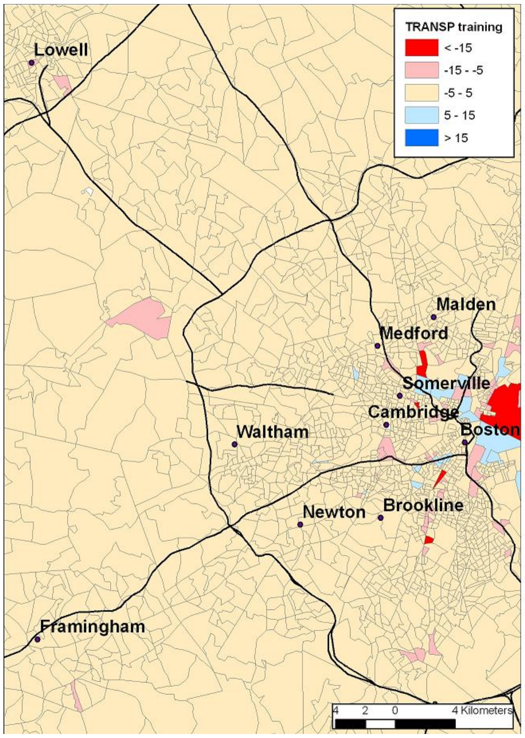


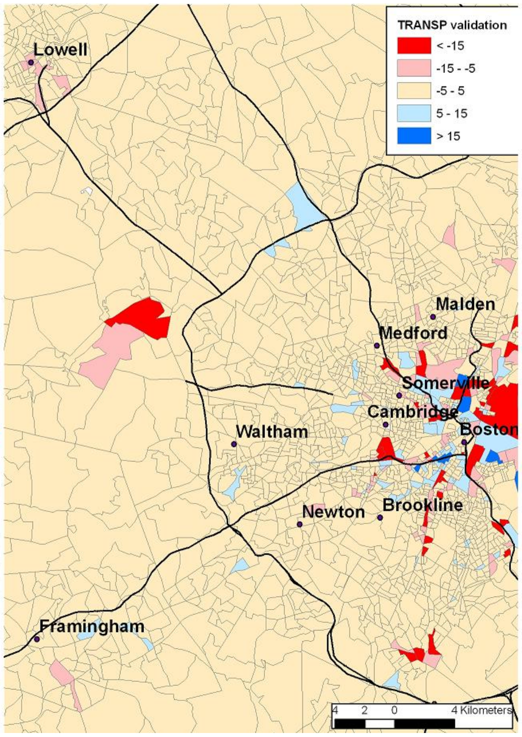
INSTIT:



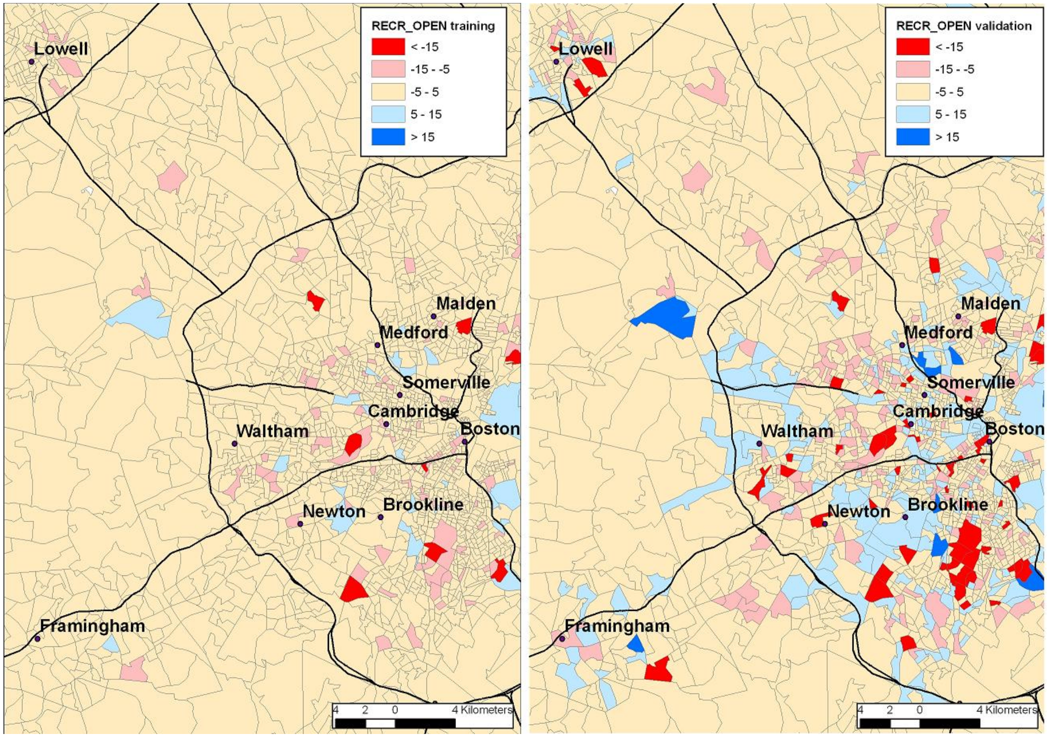


TRANSP:

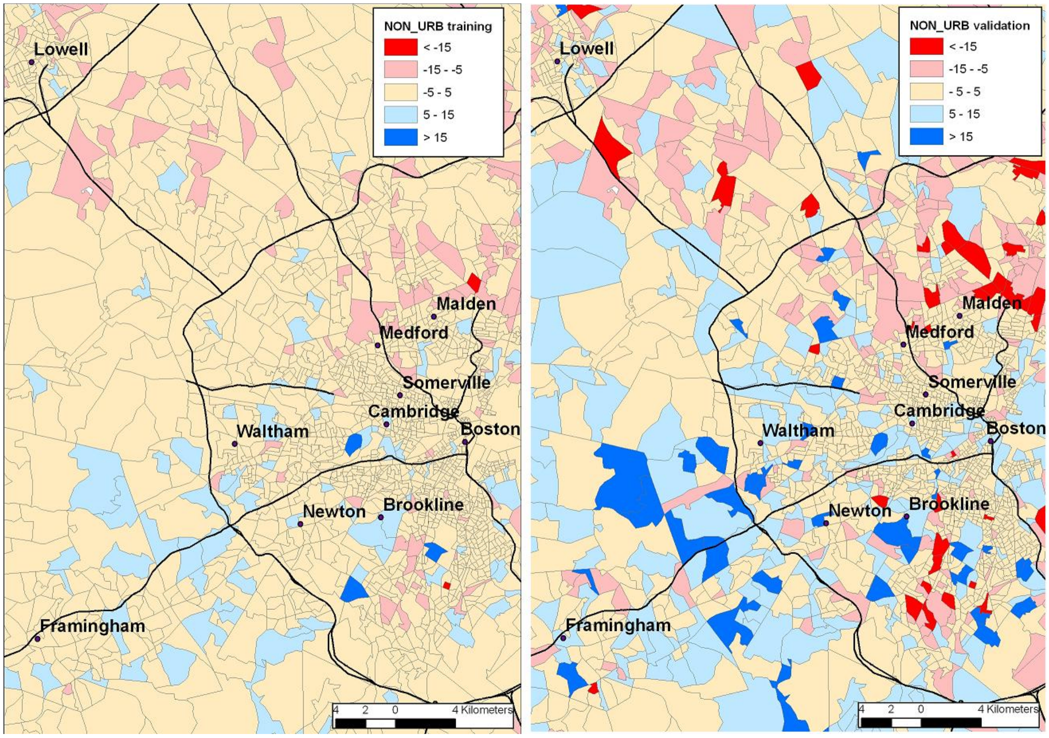




RECR_OPEN:



NON_URB:



References

References

Alonso, W., 1964. Location and land use; toward a general theory of land rent. Harvard University Press, Cambridge, MA, 204 pages.

Anderson, J.R., Hardy, E.E., Roach, J.T., and Witmer, R.E., 1976. A land use and land cover classification system for use with remote sensor data: U.S. Geological Survey Professional Paper 964, 41 p.

Aplin, P., 2003. Comparison of simulated IKONOS and SPOT HV imagery for classifying urban areas. Remotely Sensed Cities (Victor Mesev, editor), Taylor and Francis, London, UK, pp 24-45.

Atlanta Regional Commission, 2008. Atlanta Regional Commission home page; accessed in July 2008 at www.atlantaregional.com.

Aubrecht, C., Steinnocher, K., Hollaus, M., and Wagner, W., 2009. Integrating earth observation and GIScience for high resolution spatial and functional modeling of urban land use, Computers, Environment and Urban Systems, 33(1), pp 15-25.

Barnsley, M.J., and Barr, S.L., 1996. Inferring urban land use from satellite sensor images using kernel-based spatial reclassification, Photogrammetric Engineering and Remote Sensing, 62(8), pp 949-958.

Barnsley, M.J., and Barr, S.L., 1997. Distinguishing urban land-use categories in fine spatial resolution land-cover data using graph-based, structural pattern recognition systems, Computers, Environment and Urban Systems, 21(3), pp 209-225.

Barnsley, M.J., Steel, A., and Barr, S.L., 2003. Determining urban land use through an analysis of the spatial composition of buildings identified in LIDAR and multispectral image data. Remotely Sensed Cities (Victor Mesev, editor), Taylor and Francis, London, UK, pp 83-108.

Barr, S.L., Barnsley, M.J., and Steel, A., 2004. On the separability of urban land-use categories in fine spatial scale land-cover data using structural pattern recognition, Environment and Planning B: Planning and Design, 31, pp 397-418.

Bauer, T. and Steinnocher, K., 2001. Per-parcel land use classification in urban areas applying a rule-based technique, Geo-Information Systems, 14(6), pp 24-27.

Bell, L.M., Byrne, S., Thompson, A., Ratnam, N., Blair, E., Bulsara, M., Jones, T., and Davis, E., 2007. Increasing body mass index z-score is continuously associated with complications of overweight in children, even in the healthy weight range, Journal of Clinical Endocrinology & Metabolism, 92(2), pp 517-522.

Benson, B.J. and MacKenzie, M.D., 1995. Effects of spatial resolution on landscape structure parameters, Landscape Ecology, 10(2), pp 113-120.

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

Bowersox, M.A. and Brown, D.G., 2001. Measuring the abruptness of patchy ecotones: a simulation-based comparison of landscape pattern statistics, Plant Ecology, 156(1), pp 89-103.

Breiman, L., 1996. Bagging Predictors, Machine Learning, 24(2), pp 123-140.

Breiman, L., 2001. Random Forests, Machine Learning, 45(1), pp 5-32.

Breiman L. and Cutler, A., 2009. Random Forests, accessed in December 2009, at http://www.stat.berkeley.edu/~breiman/RandomForests/.

Brivio, P.A. and Zilioli, E., 2001. Urban pattern characterization through geostatistical analysis of satellite images. Remote Sensing and Urban Analysis (Donnay, Barnsley, Longley, editors), Taylor and Francis, London, UK, pp 39-53.

Brown, D.G., Addink, E.A., Duh, J.D., and Bowersox, M.A., 2004. Assessing uncertainty in spatial landscape metrics derived from remote sensing data. Remote Sensing and GIS Accuracy Assessment (Lunetta, R., Lyon, J.G., Eds), Boca Raton, FL, CRC Press, pp 221-232.

BTS, 2009. Bureau of Transportation Statistics National Transportation Atlas Database; accessed in June 2009 at http://www.bts.gov/publications/national_transportation_atlas_database/2008/.

Burgess, E.W., 1925 (reprinted 2008). Urban Ecology. Book chapter: The growth of a city: an introduction to a research project (excerpt from The City, 1925). Springer US, New York, pp 71-78.

Burt, J.E. and Barber, G.M., 1996. Elementary Statistics for Geographers. The Guilford Press, New York, pp 106-111.

Cadenasso, M.L., Pickett, S.T.A, and Schwarz, K., 2007. Spatial heterogeneity in urban ecosystems: reconceptualizing land cover and a framework for classification, Frontiers in Ecology and the Environment, 5(2), pp 80-88.

Campbell, J., 1996. Introduction to Remote Sensing. The Guilford Press, New York, pp 550-583.

Carleer, A.P., and Wolff, E., 2006. Urban land cover multi-level region-based classification of VHR data by selecting relevant features, International Journal of Remote Sensing, 27(6), pp 1035-1051.

Carlisle, D.M., Falcone, J., Wolock, D.M., Meador, M.R., and Norris, R.H., 2009. Predicting the natural flow regime: models for assessing hydrological alteration in streams, River Research and Applications (2009, in press).

Chen, K., 2002. An approach to linking remotely sensed data and areal census data, International Journal of Remote Sensing, 23(1), pp 37-48.

Chicago Sun Times, February 1, 2004. Author: Tom Whitmire. Thoroughly modern mountain: filmmakers snub Blue Ridge for less-developed Carpathians.

Chisholm, M.,1968. Johann Heinrich von Thünen. In Readings in economic geography: the location of economic activity (Smith, Taaffe, and King, editors), Rand McNally, Chicago, IL, pp 34-40.

Civco, D.L., Hurd, J.D., Wilson, E.H., Arnold, C.L., and Prisloe, M.P., 2002. Quantifying and describing urbanizing landscapes in the northeast United States, Photogrammetric Engineering and Remote Sensing, 68(10), pp 1083-1090.

Civco, D.L., Chabaeva, A., and Hurd, J., 2006. A comparison of approaches to impervious surface characterization, International Geoscience and Remote Sensing Symposium (IGARSS), article number 4241508, pp 1398-1402.

Comber, A.J., 2008. The separation of land cover from land use using data primitives, Journal of Land Use Science, 3(4), pp 215-229.

Conservation Biology Institute, 2009. Protected Areas Database (PAD) Version 4. Conservation Biology Institute, accessed in June 2009 at *http://www.consbio.org/*.

Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., and Lawler, J.J., 2007. Random forests for classification in ecology, Ecology, 88(11), pp 2783-2792.

Dean A.M, and Smith, G.M., 2003. An evaluation of per-parcel land cover mapping using maximum likelihood class probabilities, International Journal of Remote Sensing, 24(14), pp 2905-2920.

De'ath, G., 2007. Boosted trees for ecological modeling and prediction, Ecology, 81 (1), pp 243-251.

Debeir, O., Van den Steen, I., Latinne, P., Van Ham, P., and Wolff, E., 2002. Textural and contextual land-cover classification using single and multiple classifier systems, Photogrammetric Engineering and Remote Sensing, 68(6), pp 597-605.

De Lira, Y., Marroquin, J.L, and Hernandez, A., 2006. Monogenic bandpass filters for the segmentation of urban zones in satellite sensor imagery, International Journal of Remote Sensing, 27(10), pp 2087-2092.

Delpierre, N., Dufrene, E., Soudani, K., Ulrich, E., Cecchini, S., Boe, J., and Francois, C., 2009. Modelling interannual and spatial variability of leaf senescence for three deciduous tree species in France, Agricultural and Forest Meteorology, 149(6-7), pp 938-948.

De Wit, A.J.W., and Clevers, J.G.P.W., 2004. Efficiency and accuracy of per-field classification for operational crop mapping, International Journal of Remote Sensing, 25(20), pp 4091-4112.

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

Dong, L. and Wu, B., 2007. Extraction of residential information from high-spatial resolution image integrated with upscaling methods and object multi-features, Proceedings of SPIE – The International Society for Optical Engineering, 6786, Article number 678605.

Dong, P., 2000. Test of a new lacunarity estimation for image texture analysis, International Journal of Remote Sensing, 21(17), pp 3369-3373.

Environmental Research Systems Institute (ESRI), 2006. ArcInfo Users Manual, v 9.2.

Environmental Research Systems Institute (ESRI), 2009a. ESRI Maps and Data; accessed in March 2009 at http://www.esri.com/data/data-maps/index.html.

Environmental Research Systems Institute (ESRI), 2009b. Business Analyst product suite description; accessed in March 2009 at http://www.esri.com/getting_started/business/business_analyst.html

Estes J.E., Hajic, E.J., and Tinney, L.R.,1983. Fundamentals of image analysis: Analysis of visible and thermal infrared data. Manual of Remote Sensing, Second Edition (Robert N. Colwell, editor). ASPRS, Bethesda, MD, pp 987-1124.

European Environment Agency, 2009. CORINE land cover; accessed in March 2009 at http://www.eea.europa.eu/publications/COR0-landcover.

Falcone, J.A., and Gomez, R.B., 2005. Mapping Impervious Surface Type and Sub-Pixel Abundance Using Hyperion Hyperspectral Imagery, Geocarto International, 20 (4), pp 3-10.

Falcone, J., and Pearson, D., 2006. Land-Cover and Imperviousness Data for Regional Areas near Denver, Colorado; Dallas-Fort Worth, Texas; and Milwaukee-Green Bay, Wisconsin – 2001. U.S. Geological Survey Data Series 2006-221. Available online at: *http://pubs.usgs.gov/ds/2006/221/*.

Falcone, J.A., Stewart, J.S., Sobieszczyk, S., Dupree, J.A., McMahon, G., and Buell, G.R., 2007. A Comparison of Natural and Urban Characteristics and the Development of Urban Intensity Indices across Six Geographic Settings: U.S. Geological Survey Scientific Investigations Report 2007-5123. Available online at: *http://pubs.usgs.gov/sir/2007/5123/*.

Falcone, J.A., Carlisle, D.M., and Weber, L.C., 2009. Quantifying human disturbance in watersheds: variable selection and performance of a GIS-based disturbance index for predicting the biological condition of perennial streams. Ecological Indicators, 10(2), pp 264-273.

Fellman, J., Getis, A., and Getis, J., 1992. Human Geography. Wm. C. Brown Publishers, Dubuque, IA, 532 pages.

Forster, B.C, 1985. An examination of some problems and solutions in monitoring urban areas from satellite platforms, International Journal of Remote Sensing, 6(1), pp 139-151.

Forster, B.C, 1993. Coefficient of variation as a measure of urban spatial attributes, using SPOT HRV and Landsat TM data, International Journal of Remote Sensing, 14(12), pp 2403-2409.

Fotheringham, A.S. and Wong, D.W.S., 1991. The Modifiable Areal Unit Problem in multivariate statistical analysis, Environment and Planning A, 23, pp 1025-1044.

GeoEye Inc., 2009. GeoEye Inc. Imaging Solutions; accessed in March 2009 at www.geoeye.com.

Geographic Names Information System (GNIS), 2009. U.S. Board on Geographic Names home page; accessed in March 2009 at http://geonames.usgs.gov/index.html.

GeoLytics, 2001. Census 2000 and street 2000: East Brunswick, NJ, GeoLytics, Inc., 2 CDROMS.

Gong, P., Marceau, D.J., and Howarth, P.J., 1992. A comparison of spatial feature extraction algorithms for land-use classification with SPOT HRV data, Remote Sensing of Environment, 40, pp 137-151.

Graham, R., and Koh, A., 2002. Digital Aerial Survey: Theory and Practice. Whittles Publishing, Caithness, UK, pp 146-161.

Greenhill, D.R., Ripke, L.T., Hitchman, A.P., Jones, G.A., and Wilkinson, G.G., 2003. Characterization of suburban areas for land use planning using landscape ecological indicators derived from IKONOS-2 multispectral imagery, IEEE Transactions on Geoscience and Remote Sensing, 41(9), Part I, pp 2015-2021.

Griffith, D.A., and Wong, D.W., 2007. Modeling population density across major US cities: a polycentric spatial regression approach, Journal of Geographical Systems, 9(1), pp 53-75.

Gustafson, E.J., 1997. Quantifying landscape spatial pattern: what is the state of the art, Ecosystems, 1, pp 143-156.

Haack, B. 1987. An assessment of Landsat MSS and TM data for urban and near-urban land-cover digital classification, Remote Sensing of Environment, 21, pp 201-213.

Haack, B., Guptill, S., Holz, R., Jampoler, S., Jensen, J., and Welch, R., 1997. Urban analysis and planning. Manual of Photographic Interpretation. ASPRS, Bethesda, MD, pp 517-554.

Haralick, R.M., Shanmugam, K., and Dinstein. I., 1973. Texture features for image classification, IEEE Transactions on Systems, Man, and Cybernetics, 3, pp 610-612.

Hardin, P.J., Jackson, M.W., and Jensen, R.R., 2008. Modelling housing unit density from land cover metrics: A midwestern example, Geocarto International, 23 (5), pp 393-411.

Harris, P.M. and Ventura, S.J., 1995. The integration of geographic data with remotely sensed imagery to improve classification in an urban area, Photogrammetric Engineering and Remote Sensing, 61(8), pp 993-998.

Heinz Center, 2008. Landscape pattern indicators for the nation; accessed in March 2009 at http://www.heinzctr.org/publications/PDF/Landscape_Pattern_Indicators_12A.pdf.

Herold, M., Scepan, J., and Clarke, K.C., 2002. The use of remote sensing and landscape metrics to describe structures and changes in urban land uses, Environment and Planning A, 34, pp 1443-1458.

Herold, M., Liu, X., and Clarke, K.C., 2003. Spatial metrics and image texture for mapping urban land use, Photogrammetric Engineering and Remote Sensing, 69(9), pp 991-1001.

Homer, C., Dewitz, J., Fry, J., Coan, M., Hossain, N., Larson, C., Herold, N., McKerrow, A., VanDriel, J.N., and Wickham, J., 2007. Completion of the 2001 National Land Cover Database for the conterminous United States, Photogrammetric Engineering and Remote Sensing, 73(4), pp 337-341.

Hsu, S. 1978. Texture-tone analysis for automated land-use mapping, Photogrammetric Engineering and Remote Sensing, 44(11), pp 1393-1404.

Jacquez, G.M., Maruca, S., Fortin, M.-J., 2000. From fields to objects: A review of geographic boundary analysis, Journal of Geographic Systems, 2, pp 221-241.

Jensen, J.R., 1996. Introductory Digital Image Processing. Prentice-Hall, Upper Saddle River, NJ, 318 pages.

Jensen, J.R., and Cowen, D.C., 1999. Remote sensing of urban/suburban infrastructure and socio-economic attributes, Photogrammetric Engineering and Remote Sensing, 65(5), pp 611-622.

Jensen, J.R., Qui, F., and Patterson, K., 2001. A neural network image interpretation system to extract rural and urban land use and land cover information from remote sensor data, Geocarto International, 16(1), pp 21-30.

Johnsson, K., 2004. Segment-based land-use classification from SPOT satellite data, Photogrammetric Engineering and Remote Sensing, 60(1), pp 47-53.

Kachouie, N.N., Li, J., Alirezaie, J., 2004. A hybrid texture segmentation method for mapping urban land use, Geomatica, 58(1), pp 9-21.

Karunasinghe, D.S.K. and Liong, S.Y., 2006. Chaotic time series prediction with a global model: artificial neural network, Journal of Hydrology, 323(1-4), pp 92-105.

Kivell, P., 1993. Land and the city: patterns and processes of urban change, Routledge Press, London, 223 pages.

Kuhn, S., Egert, B., Neumann, S., and Steinbeck, C., 2008. Building blocks for automated elucidation of metabolites: machine learning methods for NMR prediction, BMC Bioinfomatics, 9(400). DOI: 10.1186/1471-2105-9-400.

Lackner, M. and Conway, T.M., 2008. Determining land-use information from land cover through an object-oriented classification of IKONOS imagery, Canadian Journal of Remote Sensing, 34(1-2), pp 77-92.

Lemonsu, A., Leroux, A, Belair, S., and Mailhot, J., 2008. A general methodology of urban cover classification for atmospheric modelling, Journal of Applied Meteorology and Climatology (submitted).

Leroux, A., Gauthier, J.P., Lemonsu, A., and Mailhot, J., 2009. Automated urban land use and land cover classification for mesoscale atmospheric modeling over Canadian cities, Geomatica, 63 (1), pp 393-406.

Li, G. and Weng, Q, 2007. Measuring the quality of life in city of Indianapolis by integration of remote sensing and census data, International Journal of Remote Sensing, 28(2), pp 249-267.

Liaw, A. and Wiener. M., 2002. Classification and regression by randomForest, R News, 60(1), 2/3 pp 18-22.

Liu, X., Clarke, K, and Herold, M., 2006. Population density and image texture: a comparison study, Photogrammetric Engineering and Remote Sensing, 72(2), pp 187-196.

Lo, C. P., 2003. Zone-based estimation of population and housing units from satellite-generated land use/land cover maps. Remotely Sensed Cities (Victor Mesev, editor), Taylor and Francis, London, UK, pp 157-180.

Lu, D. and Weng, Q, 2007. A survey of image classification methods and techniques for improving classification performance, International Journal of Remote Sensing, 28(5), pp 823-870.

McGarigal, Kevin, and Marks, B.J., 1995. FRAGSTATS: Spatial pattern analysis program for quantifying landscape structure: Portland, OR, U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station, General Technical Report PNW–GTR–351, 122 p.

McGarigal, K., Cushman, S.A., Neel, M.C., and Ene, E., 2002. FRAGSTATS: Spatial Pattern Analysis Program for Categorical Maps. Computer software program produced by the authors at the University of Massachusetts, Amherst; accessed in February 2009 at *www.umass.edu/landeco/research/fragstats/fragstats.html*.

McKnight, T., 2001. Regional Geography of the United States and Canada. Prentice-Hall, Upper Saddle River, NJ, 523 pages.

Maindonald, J. and Braun, J., 2007. Data Analysis and Graphics Using R, 2nd Edition, Cambridge University Press, New York, 502 pages.

MassGIS, 2008. Download Free Data: Land Use: Massachusetts State Office of Geographic and Environmental Information; accessed in July 2008 at http://www.mass.gov/mgis/ftplus.htm.

Mather, A.S., 1986. Land Use. Longman Group, UK, Harlow, Essex, 286 pages.

McMillen, D.P., 2004. Employment densities, spatial autocorrelation, and subcenters in large metropolitan areas, Journal of Regional Science, 44(2), pp 225-243.

Mesev, V., 1998. The use of census data in urban image classification, Photogrammetric Engineering and Remote Sensing, 64(5), pp 431-438.

Mesev, V., 2003. Remotely sensed cities: an introduction. Remotely Sensed Cities (Victor Mesev, editor), Taylor and Francis, London, UK, pp 1-19.

Mesev, V., 2005. Identification and characterization of urban building patterns using IKONOS imagery and point-based postal data, Computers, Environment and Urban Systems, 29(5), pp 541-557.

Meyer, W.B., and Turner, B.L.II (eds), 1994. Changes in land use and land cover, University Press, Cambridge, 537 pages.

Mills, E.S., 1980. Urban economics, Scott Foresman, Glenview, IL, 241 pages.

Mitchell, W.W., Guptill, S.C., Anderson, K.E., Fegeas, R.G., and Hallam, C.A., 1977. GIRAS—A geographic information analysis system for handling land use and land cover data: U.S. Geological Survey Professional Paper 1059, 16 p., available online at: *http://pubs.er.usgs.gov/pubs/pp/pp1059*.

Mladenoff, D.J, White, M.A., Pastor, J., and Crow, T.R., 1993. Comparing spatial pattern in unaltered old-growth and disturbed forest landscapes, Ecological Applications, 3(2), pp 294-306.

Moeller-Jensen, L., 1990. Knowledge-based classification of an urban area using texture and context information in Landsat-TM imagery, Photogrammetric Engineering and Remote Sensing, 56(6), pp 899-904.

Moeller-Jensen, L., Kofie, R.Y., Yankson, P.W.K., 2005. Large-area urban growth observations – a hierarchical kernel based approach based on image texture, Geografisk Tidsskrift, 105(2), pp 39-47.

Muth, R.F., 1969. Cities and housing; the spatial pattern of urban residential land use. University of Chicago Press, Chicago, 355 pages.

NAICS Association, 2009. North American Industrial Classification System Association home page; accessed in March 2009 at http://www.naics.com/

National Research Council, 2007. National Land Parcel Data: A Vision for the Future. National Academies Press, Washington, DC, 172 pages.

Natural Resources Canada, 2009. Natural Resources Canada home page; accessed in March 2009 at http://www.nrcan-rncan.gc.ca/com/index-eng.php.

Nichol J., King, B., Quattrochi, D., Dowman, I., Ehlers, M., and Ding, X., 2007. Earth observation for urban planning and management, Photogrammetric Engineering and Remote Sensing, 73(9), pp 973-977.

Odum. E.P., 1971. Fundamentals of Ecology. W.B. Saunders, Philadelphia, p 144.

Openshaw, S. and Rao, L., 1995. Algorithms for reengineering 1991 census geography, Environment and Planning A, 27(3), pp 425-446.

Pal, M., 2005. Random forest classifier for remote sensing classification, International Journal of Remote Sensing, 26(1), pp 217-222.

Perepechko, A.S., Graybill, J.K., ZumBrunnen, C., and Sharkov, D., 2005. Spatial database development for Russian urban areas: a new conceptual framework, GIScience and Remote Sensing, 42(2), pp 144-170.

Pontius, R.G., Jr., Huffaker, D., and Denman, K., 2004. Useful techniques of validation for spatially explicit land-change models. Ecological Modeling, 179(4), pp 445-461.

Poulsen, M., 2002. Development of output geographies for comparative and temporal census research, Geography and Research Forum, 22, pp 61-81.

Price, C.V., Nakagaki, N., Hitt, K.J., and Clawges, R.M., 2006. Enhanced historical land-use and land-cover data sets of the U.S. Geological Survey. U.S. Geological Survey Data Series 2006-240. Available online at: http://pubs.usgs.gov/ds/2006/240/.

Qian, B. and Rasheed, K., 2004. Hurst exponent and financial market predictability, Proceedings of the 2nd IASTED International Conference on Financial Engineering and Applications. Article number 437-043, pages 203-209.

Qian, J., Zhou, Q., and Hou, Q., 2007. Comparison of pixel-based and object-oriented classification methods for extracting built-up areas in aridzone. ISPRS Workshop on Updating Geo-spatial Databases with Imagery & the 5th ISPRS Workshop on DMGISs. Available online at: *http://www.commission4.isprs.org/urumchi/papers/163-171%20Jing%20Qian.pdf*.

Quinlan, J.R., 1993. C4.5: Programs for machine learning. Morgan Kaufmann Publishers, San Mateo, CA, 302 pages.

R project, 2009. The R project for statistical computing; accessed in March 2009 at *http://www.r-project.org/*.

Ridd, M.K., 1995. Exploring a V-I-S (vegetation-impervious surface-soil) model for urban ecosystem analysis through remote sensing: comparative anatomy for cities, International Journal of Remote Sensing, 16(12), pp 2165-2185.

Riitters, K.H., J.D. Wickham, R. O'Neill, B. Jones, and E.Smith. 2000. Global-scale patterns of forest fragmentation. Conservation Ecology 4:3. [online] *http://www.consecol.org/vol4/iss2/art3*.

Rhode Island Geographic Information System (RIGIS), 2008. University of Rhode Island Department of Natural Resources Science, Environmental Data Center, accessed in July 2008 at http://www.edc.uri.edu/.

Rocha, J., Tenedorio, J.A., Encarnacao, S., and Morgado, P., 2006. Land use/cover classification through multiresolution segmentation and object oriented neural networks classification, Proceedings of SPIE – The International Society for Optical Engineering, 6366, Article number 63660A.

Rulequest, 2008. Rulequest Research Data Mining Tools; accessed in May 2008 at http://www.rulequest.com.

Salas, W.A., Boles, S.H., Frolking, S., Xiao, X., and Li, C., 2003. The perimeter/area ratio as an index of misregistration bias in land cover change estimates, International Journal of Remote Sensing, 24(5), pp 1165-1170.

Saura, S. and Martinez-Millan, J., 2001. Sensitivity of landscape pattern metrics to map spatial extent, Photogrammetric Engineering & Remote Sensing, 67(9), pp 1027-1036.

Saura, S., 2002. Effects of minimum mapping unit on land cover data spatial configuration and composition, International Journal of Remote Sensing, 23(22), pp 4853-4880.

Saura, S., 2004. Effects of remote sensor spatial resolution and data aggregation on selected fragmentation indices, Landscape Ecology, 19, pp 197-209.

Schumacher, J.V., Redmond, R.L., Hart, M.M., and Jensen, M.E., 2000. Mapping patterns of human use and potential resource conflicts on public lands, Environmental Monitoring and Assessment, 64, pp 127-137.

Segal, M.R., 2004. Machine Learning Benchmarks and Random Forest Regression., Division of Biostatistics, University of California, San Francisco, available online at: http://escholarship.org/uc/item/35x3v9t4.

Segl, K., Roessner, S., Heiden, U., and Kaufmann, H., 2003. Fusion of spectral and shape features for identification of urban surface cover types using reflective and thermal hyperspectral data, ISPRS Journal of Photogrammetry & Remote Sensing, 58, pp 99-112.

Setiawan, H., Mathieu, R., and Thompson-Fawcett, M., 2006. Assessing the applicability of the V-I-S model to map urban land use in the developing world: case study of Yogyakarta, Indonesia, Computers, Environment and Urban Systems, 30(4), pp 503-522.

Shapire, R., Freund, Y, Bartlett, P, and Lee, W., 1998. Boosting the margin: the explanation for the effectiveness of voting methods, Annals of Statistics, 26(5), pp 1651-1686.

Shmueli, G. and Koppius, O., 2007. Predictive vs. explanatory modeling in IS research, Conference on Information Systems & Technology, Seattle, WA. Online at http://www.smith.umd.edu/faculty/gshmueli/web/html/109.html.

SILVIS Lab, 2009. WUI Maps, Statistics, and GIS data library: Univ. of Wisconsin Forest and Wildlife Ecology Wildland-Urban Interface project; accessed in January 2009 at http://silvis.forest.wisc.edu/Library/WUILibrary.asp.

Siroky, D.S., 2009. Navigating Random Forests and related advances in algorithmic modeling, Statistics Survey, 3, pp 147-163.

Smith, G.M. and Fuller, R.M., 2001. An integrated approach to land cover classification: an example in the Island of Jersey, International Journal of Remote Sensing, 22(16), pp 3123-3142.

Southern California Association of Governments, 2008. Southern California Association of Governments home page; accessed in July 2008 at http://www.scag.ca.gov/.

Stehman, S.V., Wickham, J.D., Smith, J.H., and Yang, L., 2003. Thematic accuracy of the 1992 National Land-Cover Data for the eastern United States: statistical methodology and regional results, Remote Sensing of Environment, 86(4), pp 500-516.

Stow, D., Lopez, A., Lippitt, C., Hinton, S., and Weeks, J., 2007. Object-based classification of residential land use within Accra, Ghana based on QuickBird satellite data, International Journal of Remote Sensing, 28(22), pp 5167-5173.

Strobl, C., Boulesteix, A., Kneib, T., Augustin, T., and Zeileis, A., 2008. Conditional variable importance for random forests, BMC Bioinfomatics, 9:307. Available online at: *http://www.biomedcentral.com/1471-2105/9/307*.

Su, W., Li, J., Chen, Y., Liu, Z., Zhang, J., Low, T., Suppiah, I., Hashim, S., 2008. Textural and local spatial statistics for the object-oriented classification of urban areas using high resolution imagery, International Journal of Remote Sensing, 29(11), pp 3105-3117.

Sun, H., Forsythe, W., and Waters, N., 2007. Modeling urban land use change and urban sprawl: Calgary, Alberta, Canada, Networks and Spatial Economics, 7(4), pp 353-376.

Taubenboeck, H., Esch, T., and Roth, A., 2006. An urban classification approach based on object-oriented analysis of high resolution satellite imagery for a spatial structuring within urban areas, 1st EARSeL Workshop of the SIG Urban Remote Sensing, Humboldt-Universitaet zu Berlin, 2-3 March, 2006.

Theobald, D.M., 2005. Landscape patterns of exurban growth in the USA from 1980 to 2020, Ecology and Society, 10(1), 34 p.

Thrall, G.I., 1980. The consumption theory of land rent, Urban Geography, 1(4), pp 350-370.

Tiede, D., Lang, S., Albrecht, F., and Hoelbling, D., 2010. Object-based class modeling for cadastre-constrained delineation of geo-objects. Photogrammetric Engineering and Remote Sensing, 76(2), pp 193-202.

Tofallis, C., 1999. Model building with multiple dependent variables and constraints, The Statistician, 48(3), pp 371-378.

Tso, B., Olsen, R.C., 2004. Scene classification using combined spectral, textural and contextual information, Proceedings of SPIE – The International Society for Optical Engineering, 5425, pp 135-146.

U.S. Census Bureau, 2009a. Cartographic Boundary Files: U.S. Census Bureau; accessed in September 2008 at http://www.census.gov/geo/www/cob/bdy_files.html.

U.S. Census Bureau, 2009b. Summary File 3: U.S. Census Bureau; accessed in March 2009 at http://factfinder.census.gov/jsp/saff/SAFFInfo.jsp?_content=sp4_decennial_sf3.html&_title=Summary+File+3+(SF+3)&_lang=en&_sse=on.

U.S. Census Bureau, 2009c. Download Center – American Fact Finder; accessed in March 2009 at http://factfinder.census.gov/servlet/DCGeoSelectServlet?ds_name=DEC_2000_SF3_U.

U.S. Census Bureau, 2009d. 2008 TIGER/Line shapefiles; accessed in June 2009 at http://www.census.gov/geo/www/tiger/tgrshp2008/tgrshp2008.html.

U.S. Census Bureau, 2009e. Census 2000 Urban and Rural Classification; accessed in December 2009 at http://www.census.gov/geo/www/ua/ua_2k.html.

U.S. Environmental Protection Agency, 2009, EPA National Stream Report: Wadeable Streams Assessment, accessed in March 2009, at *http://www.epa.gov/owow/streamsurvey/*

U.S. Environmental Protection Agency, 2010, Urbanization and Population Change, accessed in January 2010, at http://cfpub.epa.gov/eroe/index.cfm?fuseaction=detail.viewInd&lv=list.listByAlpha&r=209832&subtop=225.

U.S. Geological Survey, 2008. Raw Data Download: U.S. Geological Survey National Atlas; accessed in September 2008 at http://www.nationalatlas.gov/atlasftp.html.

U.S. Geological Survey, 2009a, USGS Geographic Data Download, accessed in March 2009, at *http://edc2.usgs.gov/geodata/index.php*.

U.S. Geological Survey, 2009b, NAWQA Resources: Refining 1970s land-use data with population data, accessed in March 2009, at *http://water.usgs.gov/nawqa-only/bglu/bglu.html*.

U.S. Geological Survey, 2009c National Water-Quality Assessment (NAWQA) Program: U.S. Geological Survey, accessed in March 2009, at *http://water.usgs.gov/nawqa/*.

U.S. Geological Survey, 2009d, Multi-Resolution Land Characteristics Consortium, accessed in March 2009, at *www.mrlc.gov*.

Venables, W.N. and Ripley, B.D., 1999. Modern Applied Statistics with S-Plus, Springer-Verlag, New York, 501 pages.

Vogelmann, J.E., Sohl, T., and Howard, S.M., 1998. Regional characterization of land cover using multiple sources of data. Photogrammetric Engineering and Remote Sensing, 64(1), pp 45-57.

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

Walsh, S.E., Soranno, P.A., and Rutledge, D.T., 2003. Lakes, wetlands, and streams as predictors of land use/cover distribution, Environmental Management, 31(2), pp 198-214.

Walton, J.T., 2008. Subpixel urban land cover estimation: comparing Cubist, Random Forests, and Support Vector regression, Photogrammetric Engineering and Remote Sensing, 74(10), pp 1213-1222.

Wang, W., Yang, J., and Lin, Y., 2008. Open-source versus proprietary GIS on landscape metrics calculation: a case study, Proceedings of the academic track of the 2008 Free and Open Source Software for Geospatial (FOSS4G) Conference, available online:
http://www.foss4g.org/index.php/foss4g/2008/paper/view/124/47.

Watts, R.D., Compton, R.W., McCammon, J.H., Rich, C.L., Wright, S.M., Owens, T., and Ouren, D.S., 2007. Roadless space of the conterminous United States, Science, 316(5825), pp 736-738.

Weber, C., 2003. Urban agglomeration delimitation using remote sensing data. Remotely Sensed Cities (Victor Mesev, editor), Taylor and Francis, London, UK, pp 146-159.

Welch, R., 1982. Spatial resolution requirements for urban studies, International Journal of Remote Sensing, 3(2), pp 139-146.

Wu, B., Huang, B., and Fung, T., 2009a. Projection of land use change patterns using kernel logistic regression, Photogrammetric Engineering and Remote Sensing, 75(8), pp 971-979.

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

Wu, C., and Murray, A.T., 2007. Population estimation using landsat enhanced thematic mapper imagery, Geographical Analysis, 39(1), pp 26-43.

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

Wu, J., Xu, J., and Yue, W., 2006. V-I-S model for cities that are experiencing rapid urbanization and development, International Geoscience and Remote Sensing Symposium(IGARSS), 3 (1526276), pp 1503-1506.

Wu, S., Silvan-Cardenas, J, and Wang, L., 2007. Per-field urban land use classification based on tax parcel boundaries, International Journal of Remote Sensing, 28(12), pp 2777-2800.

Wu, S., Qiu, X., Usery, E.L., and Wang, L., 2009b. Using geometrical, textural, and contextual information of land parcels for classification of detailed urban land us, Annals of the Association of American Geographers, 99(1), pp 76-98.

Wyatt, P., 2004. Constructing a land use data set from public domain information, Planning Practice and Research, 19(2), pp 147-171.

Xian, G., Zhu, Z., Hoppus, M., Fleming, M., 2002. Application of decision-tree techniques to forest group and basal area mapping using satellite imagery and forest inventory data, Pecora 15/Land Satellite Information IV/ISPRS Commission I/FIEOS 2002 Conference Proceedings. Available online: http://www.isprs.org/commission1/proceedings02/paper/00005.pdf.

Yohannes, Y. and Hoddinott, J., 2006, Classification and Regression Trees: An Introduction, Technical Guide No. 3 (Washington, DC: International Flood Policy Research Institute).

Yu, D. and Wu., C., 2006. Incorporating remote sensing information in modeling house values: a regression tree approach, Photogrammetric Engineering and Remote Sensing, 72(2), pp 129-138.

Zhang, Q., and Wang, J., 2003. A rule-based urban land use inferring method for fine-resolution multispectral imagery, Canadian Journal of Remote Sensing, 29(1), pp 1-13.

Zhang, Q., Wang, J., Gong, P., and Shi, P., 2003. Study of urban spatial pattern from SPOT panchromatic imagery using textural analysis, International Journal of Remote Sensing, 24(21), pp 4137-4160.

Curriculum Vitae

James A. Falcone received his Bachelor of Arts in Environmental Science from the University of Virginia in 1979. He received a Master of Science in Geographic and Cartographic Sciences from George Mason University in 1999, and a Master of Science and Technology in Remote Sensing from the University of New South Wales (Sydney, Australia) in 2002. He is currently employed at the U.S. Geological Survey in Reston, Virginia, and works with the National Water-Quality Assessment Program.