Socioeconomic-Vegetation Relationships in Urban, Residential Land: The Case of Denver, Colorado

Jeremy Mennis

Abstract

This research investigates the relationship between socioeconomic status and remotely sensed vegetation intensity in residential land in the Denver, Colorado metropolitan area. Land-cover data derived from aerial photography and normalized difference vegetation index data (NDVI) derived from Landsat ETM+ imagery were integrated with U.S. Bureau of the Census tract-level data and analyzed using choropleth mapping and multivariate statistics. Association rule mining, a data mining technique, is used to explore nonlinear relationships among variables. Results indicate that higher vegetation intensity is associated with socioeconomic advantage in both sparsely populated, large lot suburban developments, as well as in older, urban neighborhoods. This pattern likely reflects residents' ability to pay for the cost of maintaining high vegetation intensity, suburban lawn ecosystem vegetation in a semi-arid grassland environment. Additionally, residential choices may be limited by a home price structure that is closely related to the concentration of vegetation in the residential landscaping.

Introduction

Investigating the relationship between socioeconomic and ecological characteristics in urban regions is important for two reasons. First, as the world's population continues to concentrate in cities, there is an increasing recognition that understanding the interaction among socioeconomic and ecological processes in urban areas is a key to predicting the impact of urban growth on the global environment (Grimm *et al.*, 2000). Second, from a local urban planning policy perspective, understanding socioeconomic-ecological relationships can inform sustainable growth management plans. This is particularly true in cases where urban growth is straining the capacity of resources such as water and energy, as well as in areas where there is concern about the impact of urban sprawl on people's quality of life (e.g., air quality, access to parks, and open space).

Remote sensing has been widely used for monitoring urban areas, as remotely sensed imagery can provide spatially and temporally continuous data on urban land-cover (Mesev, 2003). Most urban remote sensing research has focused on capturing the extent, growth, composition, and morphological characteristics of cities (Jensen and Cowen, 1999; Karathanassi *et al.*, 2000; Civco *et al.*, 2002; Herold *et al.*, 2003; Rashed *et al.* 2003). Remote sensing has also been used in urban ecological analysis for characterizing urban vegetation (Small, 2001; Wilson *et al.*, 2003). Recently, remotely sensed imagery has been applied to a number of social science applications, from modeling population distribution (Yuan *et al.*, 1997; Sutton *et al.*, 2001) and the environmental characteristics of population concentrations (Pozzi and Small, 2002) to the spatial analysis of crime (Chen *et al.*, 2004) and demography (Weeks *et al.*, 2000).

The objective of the present research is to investigate the relationship between socioeconomic status and remotely sensed vegetation intensity in the Denver, Colorado metropolitan area (Figure 1). This area, which also includes the cities of Boulder and Longmont, is the most densely populated portion of the Front Range urban corridor that extends along the interface of the Rocky Mountains and Great Plains. This semi-arid grassland region has undergone extensive urban growth over the past twenty-five years (Riebsame *et al.*, 1997). The rapid pace of urban development has made urban sprawl a major public issue in the region, particularly regarding the formation of policies on water resource management, open space preservation, and maintaining quality-of-life.

The research methods used here include choropleth mapping and univariate and multivariate statistics, as well as the data mining technique association rule mining. Association rule mining can identify relevant relationships among variables that may not be captured by conventional analytical approaches. This technique was initially developed for business applications but is shown here to be useful for the analysis of urban ecology using socioeconomic data and remotely sensed imagery.

Linking Socioeconomic and Remotely Sensed Data

Interest in applying remotely sensed data to socioeconomic analyses has expanded as recognition of the utility of remote sensing for identifying the landscape effects of socioeconomic processes has grown (Liverman *et al.*, 1998; Fox *et al.*, 2003). Much of this research has focused on linking census- and survey-based socioeconomic data to remotely sensed land-use or land-use change data, particularly for modeling the drivers of deforestation in rural areas (Pfaff, 1999; Walsh *et al.*, 1999; Geoghegan *et al.*, 2001). Many of these studies have employed multivariate statistics to model land-cover change using household- and census unit-level data. Seto and Kaufmann (2003) extend this approach for econometric modeling of rural to urban land conversion, as indicated by Landsat TM imagery, in the area surrounding Hong Kong, China.

Department of Geography and Urban Studies, Temple University, 1115 W. Berks St., 309 Gladfelter Hall, Philadelphia, PA 19122 (jmennis@temple.edu).

Photogrammetric Engineering & Remote Sensing
Vol. 72, No. 8, August, 2006, pp. 911–921.

0099-1112/06/7208–0911/$3.00/0

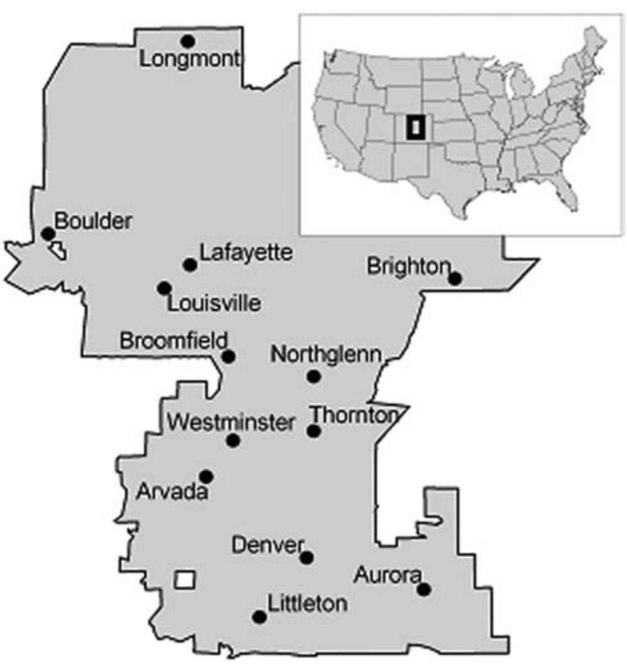

Figure 1. The Denver, Colorado area study region.

Other researchers integrating socioeconomic and remotely sensed data for urban analysis have focused on linking indicators of socioeconomic status with vegetation character as captured by normalized difference vegetation index (NDVI) data (Lo and Faber, 1997). Healthy vegetation tends to reflect strongly in the near infrared wavelengths (NIR) and relatively weakly in the visible wavelengths (VIS) (Tucker, 1979). The NDVI accounts for this by calculating a ratio of the two bands as (NIR − VIS)/(NIR + VIS). The NDVI can be considered a general measure of vegetation intensity, or *greenness*, and is a function of a host of vegetation characteristics, including density, health, and fractional cover, as well as other characteristics such as soil moisture, and has also been used to indicate net primary production and climate (Bannari *et al*., 1995; Carlson and Ripley, 1997). Vegetation indices, most prominently NDVI, have been widely used for analysis of vegetation and vegetation change from global to regional scales (Carlson *et al*., 1994; Myneni *et al*., 1998).

Of particular relevance to the present research is the work of Lo and Faber (1997), who use principle component analysis to investigate the relationship of Landsat TM-derived NDVI and surface temperature with a series of U.S. Bureau of the Census socioeconomic variables such as population density and median home value for Athens-Clarke County, Georgia. These authors found that higher NDVI values were associated with socioeconomic advantage (e.g., wealth and higher educational attainment) and low population density, and that NDVI can provide a quality of life measure incorporating both socioeconomic and biophysical characteristics. In a longitudinal study of socioeconomic and remotely sensed vegetation change in Detroit, Michigan from 1975 to 1992, Ryznar and Wagner (2001) found that vegetation intensity increased most dramatically in poor neighborhoods where population decreased, i.e., neighborhood abandonment occurred. Changes in income and racial composition, however, were not related to change in vegetation intensity in this study. Because Ryznar and Wagner (2001) do not report the association of vegetation intensity with socioeconomic character for a single moment in time, their results cannot be directly compared with those of Lo and Faber (1997) as to why the two studies suggest a somewhat opposite relationship between socioeconomic character and vegetation intensity. It should be noted, however, that the two study regions are quite different; one focuses on a major Midwest U.S. industrial city undergoing rapid depopulation in the urban core (Ryznar and Wagner, 2001) while the other focuses on a small, university-oriented town in the Southeast U.S. undergoing moderate growth (Lo and Faber, 1997).

Data Sources and Preprocessing

Socioeconomic and housing characteristics data for 2000 were acquired from the U.S. Bureau of the Census at the tract level. In order to focus on socioeconomic-vegetation relationships in residential areas, the study region is restricted to tracts in urban, suburban, and exurban areas; the foothills and mountains are excluded, as are large, non-residential developed areas such as the airport and military installations. Two other tracts within the study region are also excluded, one of which has zero population within it and the other is occupied by the University of Colorado in Boulder. This latter tract was excluded because, although students do reside within its boundaries, it is an extreme outlier in its combination of socioeconomic, housing, and vegetation characteristics. There are 399 tracts in the study region.

Vegetation intensity data were derived from imagery acquired by the Enhanced Thematic Mapper Plus (ETM+) sensor carried aboard the Landsat 7 satellite. This 27 July 1999 cloud-free ETM+ image was orthorectified and processed to derive a NDVI image (Figure 2). Land-cover data were acquired from the U.S. Geological Survey (USGS) Front Range Infrastructure Resources Project (FRIRP) (Figure 2). These data encode 1996 and 1997 land-cover in vector polygon format as manually digitized from digital orthophotographic quadrangles (DOQs) and ancillary data such as wetlands inventory (Stier, 1999). Land-cover is attributed according to a five-stage hierarchical classification (Anderson *et al*., 1976). Note that although the data are referred to as land *cover* data, they are classified according to a mixture of cover types (e.g., vegetated) as well as uses (e.g., commercial). Raster elevation data were acquired from the USGS at a 30-meter horizontal resolution. Vector data on the locations of limited and unlimited access primary roads were acquired from the Environmental Systems Research Institute (ESRI) streets database.

These data were then processed within a geographic information system (GIS) to derive a set of secondary variables used in the analysis. The elevation data were used to derive a 30 m resolution raster data layer encoding percent slope. The transportation data were used to derive two 30 m resolution raster data layers encoding the distance from each grid cell to the nearest limited and unlimited access primary roads, respectively. The land-cover data layer was used to extract residential land, as distinguished from areas that are not developed (e.g., water, vegetated) or are developed but not residential (e.g., industrial), as well as land classified as commercial/light industry (one class) or mixed urban (Figure 2). The mixed urban classification includes both commercial/light industry and residential land-uses, as one might find in an urban downtown area.

A 30 m resolution raster data layer describing the density of residential land was also generated. Note that this data layer does not describe population density, but the concentration of residential land-use. The residential density data were generated by converting the residential land-cover data to raster format, creating a second *empty* raster, counting the number of residential cells within a 1 km radius

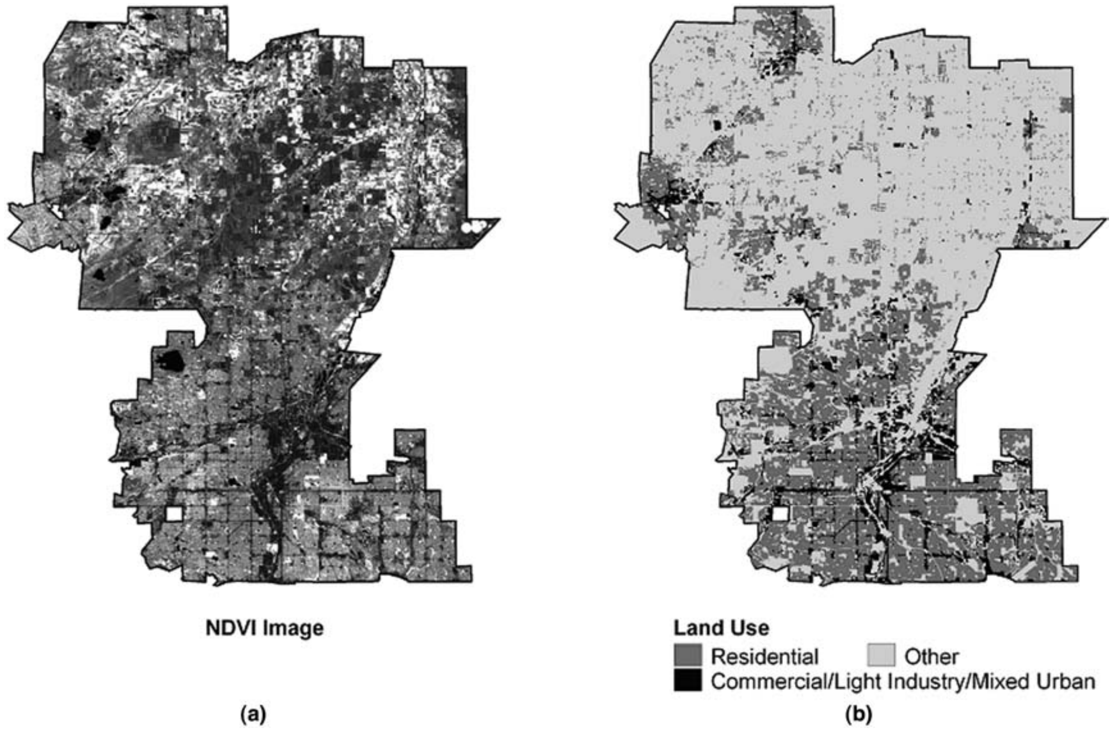

Figure 2. NDVI (a) and land-cover data (b) used in the analysis, clipped to the study region. Lighter gray in the NDVI image indicates higher vegetation intensity.

of the centroid of each grid cell in the empty raster, and assigning the sum to that empty raster grid cell. A residential density data layer grid cell with a relatively low value indicates a low degree of residential concentration while a high value indicates a high degree of residential concentration. This methodology was also used to generate a 30 m resolution raster layer of density of land classified as commercial/light industry or mixed urban.

The purpose of generating the land-cover density layers is to describe the regional land-cover character of the tract. While it is possible to simply calculate the percentage of each tract occupied by a particular land-cover, this approach would not take into account the fact that a tract may be adjacent to a large area of a particular land-cover but not itself be occupied by that land-cover. For example, in urban areas, where tracts are small, it is certainly possible that a largely residential tract may be located nearby, or even surrounded by, a heavily industrialized area. The methodology of calculating land-cover density for each tract, as described above, would capture this characteristic, while the simple calculation of the percentage of the tract occupied by a certain land-cover would not. Admittedly, the 1 km bandwidth used to generate the land-use density surface is somewhat arbitrary, but is used here as a coarse indicator of proximity.

In order to facilitate statistical analysis, the land-cover, environmental, and NDVI data were used to develop a set of variables aggregated to a single spatial unit, the census tract. This approach of aggregating remotely sensed to the spatial units at which aggregated socioeconomic data are available mirrors that of a number of other studies (Lo and Faber, 1997; Seto and Kaufmann, 2003). One of the tract-level variables captures vegetation intensity as indicated by NDVI values. The other variables, referred to as explanatory variables, indicate socioeconomic and environmental characteristics that may be related to vegetation intensity. These variables are calculated by summarizing each variable for the residential and mixed urban land within each tract. For example, tract-level mean elevation was calculated as the mean of all the elevation data grid cells falling within the residential and mixed urban land of each tract. Aggregating data only for the residential and mixed urban land within each tract facilitates finding socioeconomic-vegetation intensity relationships, which may be obscured if NDVI values from commercial or other land where people do not reside are included. This preprocessing resulted in the following variables:

• Vegetation Intensity	Mean NDVI
• Mean Elevation	Mean elevation in meters
• Mean Slope	Mean slope in percent
• Distance to Limited	Mean distance in meters to the nearest limited access highway
• Distance to Unlimited	Mean distance in meters to the nearest unlimited access highway
• Residential Density	Mean density of residential land in number of grid cells
• Commercial Density	Mean density of commercial/light industry/mixed urban land in number of grid cells
• Population Density	Persons/km^2 (total population/area of residential land)
• Median Income	Median yearly household income in U.S. dollars
• Percent Minority	Percent of population who do not self-identify as white, non-Hispanic
• Educational Attainment	Percent of population over the age of 25 with a high school diploma or equivalency
• Number of Rooms	Median number of rooms per housing unit
• Home Year	Median year housing unit was built
• Home Value:	Median value of owner-occupied housing unit (USD)

Table 1. Descriptive Statistics for Tract-level Variables and Correlation with Vegetation Intensity

Variable	Minimum	Maximum	Mean	St. Dev.	Corr.
Vegetation Intensity	−0.12	0.36	0.19	0.06	
Mean Elevation	1,506	1,817	1,636	50	0.26***
Mean Slope	0.2	6.2	1.8	0.96	0.06
Distance to Limited	270	26,237	4,947	5,566	0.14***
Distance to Unlimited	192	5,921	1,185	1,075	−0.02
Residential Density	202	2,995	1,915	566	0.28***
Commercial Density	10	2,191	404	359	−0.21***
Population Density	322	19,810	3,861	2,134	−0.40***
Median Income	7,411	112,596	49,126	17,540	0.27***
Percent Minority	3	95	33	23	−0.32***
Educational Attainment	37	99	83	14	0.30***
Number of Rooms	2.0	9.0	5.5	1.4	0.28***
Home Year	1939	1997	1968	15	−0.14***
Home Value	9,999	495,600	174,563	70,116	0.32***

*** = significance < 0.005.

Table 1 presents descriptive statistics for these variables. Figure 3 presents choropleth maps of a select set of these variables. Note that even though the maps assign a grayscale value to entire tracts for display purposes, the data are in fact calculated only for the residential and mixed urban area of each tract.

Analytical Methods

As a first step, NDVI is summarized by land-cover to investigate whether there are indeed differences in vegetation intensity among different land-covers. The variables are then entered into a series of analyses to explore the relationship between vegetation intensity and each of the explanatory variables. First, the choropleth maps of the variables presented in Figure 3 are visually examined. Second, correlation is used to indicate the strength of the relationship between each of the explanatory variables with vegetation intensity. Because a number of the explanatory variables are not normally distributed, Kendall's tau-b correlation was employed. Multivariate regression is then used to test the explanatory power and interaction among the explanatory variables with regards to vegetation intensity. Note that two pairs of the explanatory variables are highly correlated (Pearson $r > 0.80$, significance < 0.0005), educational attainment-percent minority and number of rooms-median income. Collinearity diagnostics, including the Variance Inflation Factor (VIF), were used to ensure that multi-collinearity was not problematic in any of the models.

Association rule mining is then used to explore nonlinear relationships among the explanatory variables in predicting vegetation intensity. Association rule mining is a data mining technique that seeks to identify rules in a transactional database (Agrawal *et al.*, 1993). An association rule takes the form $A \rightarrow B$ where A (the antecedent) and B (the consequent) are sets of predicates. Association rule mining has conventionally been applied to business data, such as supermarket transactions where rules regarding purchasing behavior are mined. An illustrative example of a rule using supermarket data is "if bagels are purchased then cream cheese is also purchased," where the purchase of a bagel is the antecedent and the purchase of cream cheese is the consequent. Association rule mining has also been adapted for spatial data where a spatial relationship is encoded in the antecedent or consequent (Koperski and Han, 1995; Mennis and Liu, 2005). GIS is used in the present research to integrate a variety of data sources (e.g., census and imagery) to mine the spatial coincidence of explanatory and vegetation intensity variable values.

Association rule mining can generate an enormous number of rules from even moderately sized data sets. There has thus been considerable research on extracting *interesting* rules from rule-sets generated from association rule mining. The term *interesting* does not have a formal definition, but is often used in the association rule mining literature to connote those rules that are of interest to an analyst, as opposed to those rules that would be considered trivial or obvious (Tan *et al.*, 2002). Researchers have developed a number of approaches to extracting interesting rules, including rule *templates* and *metrics*. The rule template approach allows the analyst to specify which variables or cases can appear in the antecedent and/or consequent of a rule (Fu and Han, 1995). Rule metrics, of which many have been developed, are measurements of various aspects of rule quality (Tan *et al.*, 2002).

The rule metrics used in the present research include *support, confidence*, and *lift*. Support is the percentage of all transactions in which the antecedent occurs. Confidence is the percentage of antecedent transactions in which the consequent occurs (i.e., its meaning is not equivalent to that of statistical confidence). Lift indicates how much more often than expected the consequent occurs when paired with the antecedent than one would expect if the antecedent and consequent were not related. As an example, consider the bagel and cream cheese rule above. Say there are 100 total transactions, with 10 transactions containing the purchase of a bagel and 20 transactions containing the purchase of cream cheese. Of those 20 cream cheese purchases, eight were accompanied by a bagel purchase. The support is .10 (10 bagel purchases/100 transactions). The confidence is .80 (8 cream cheese purchases/10 bagel purchases). The lift is 4.0 (.80 confidence/(20 cream cheese purchases/100 transactions)).

For the purpose of association rule mining in the present research, each tract is considered a transaction in the database. The antecedent is a set of predicates composed of tract attributes, such as population density or distance to limited. Because association rule mining works with categorical, not numeric data, each of the tract-level variables was transformed into ordinal data using a five-class quantile classification prior to the rule mining. In order to extract interesting rules, a template is used in which generated rules are restricted to those with vegetation intensity values ranking in the highest and lowest twentieth percentile in the consequent. Likewise, rule antecedents are restricted to combinations of the explanatory variables with values ranking in the highest and lowest twentieth percentile. Note that this template generates rules which identifies associations which concern the extreme values of the variables, thus potentially capturing nonlinear relationships among socioeconomic character and vegetation intensity, or relationships that occur only in a subset of the data in tracts with the highest or lowest variable values. Such relationships would not be captured by the multivariate regression because of the assumption of linearity in the relationships.

Rules are also restricted to those with a lift greater than 2.0 and a support greater than 2.0 percent. While this support value may seem low, note that the probability of two randomly distributed, quantile-classified, ordinal variables having values that co-occur in a given tract is 4 percent (i.e., 20% * 20%). Redundant rules with identical consequents are filtered out by identifying rules that do not differ significantly ($p < 0.05$) in the confidence of the rule given additional predicates in the antecedent (Huang and Webb, 2004). This rule mining approach facilitates the identification of interesting and non-redundant rules that express which of high and low values of the explanatory variables are associated with particularly high or low vegetation intensity. The association rule mining software Magnum Opus (Webb, 1995; Webb and Associates Pty. Ltd., 2001) was used for this analysis.

Results

Table 2 reports the mean and standard deviation of vegetation intensity, as well as the area as a percentage of the total study region, for different land-cover types, including the four level 1 land-covers, and selected level 2 and level 3 land-covers relevant to the research. Residential land has a higher mean vegetation intensity, and lower standard deviation, than non-residential developed land-cover,

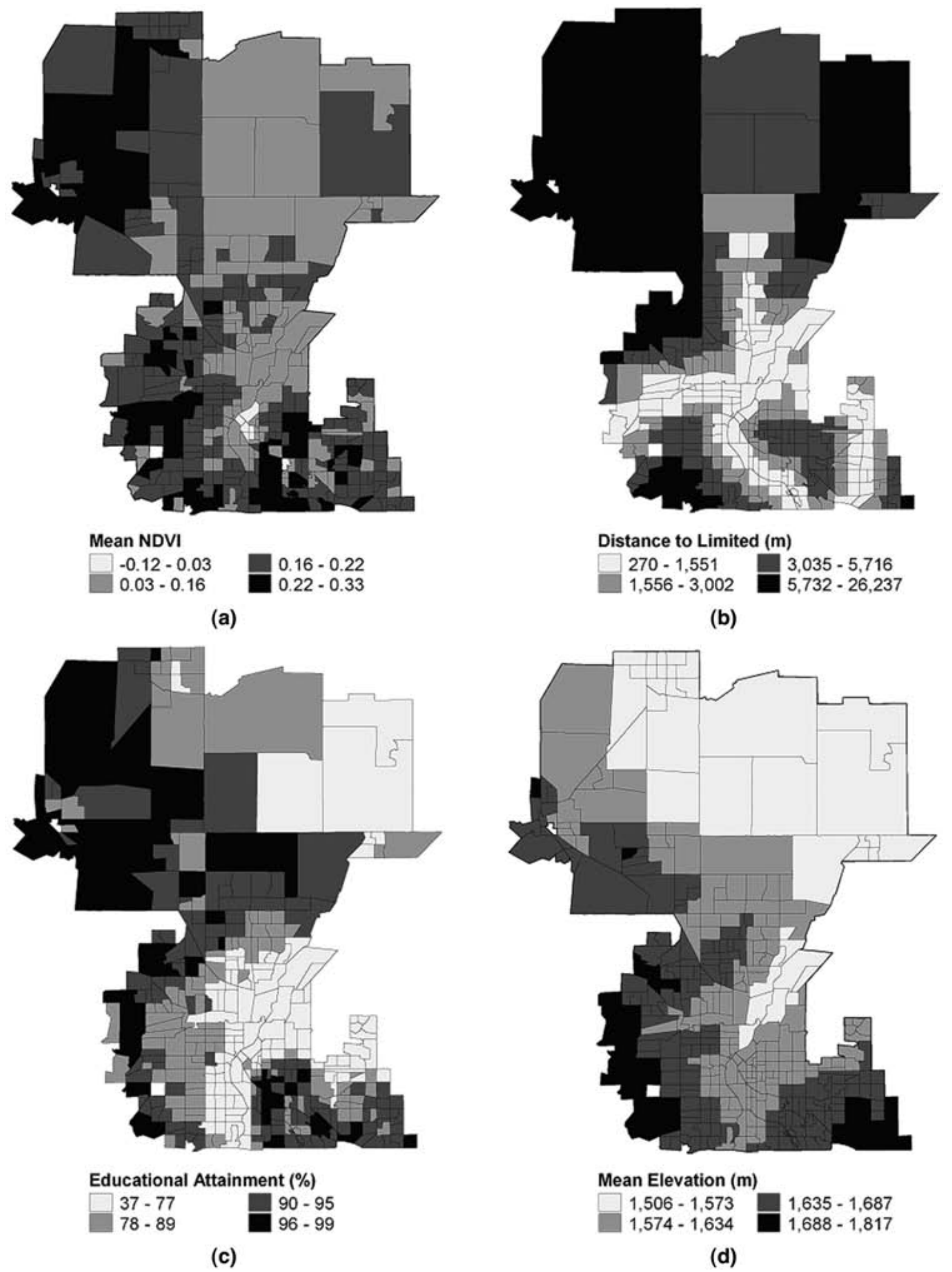

Figure 3. Selected tract-level variables used in the analysis. See text for an explanation of the derivation of each variable: (a) Mean NDVI, (b) Distance to Limited in meters, (c) Educational Attainment by percentage, and (d) Mean Elevation in meters.

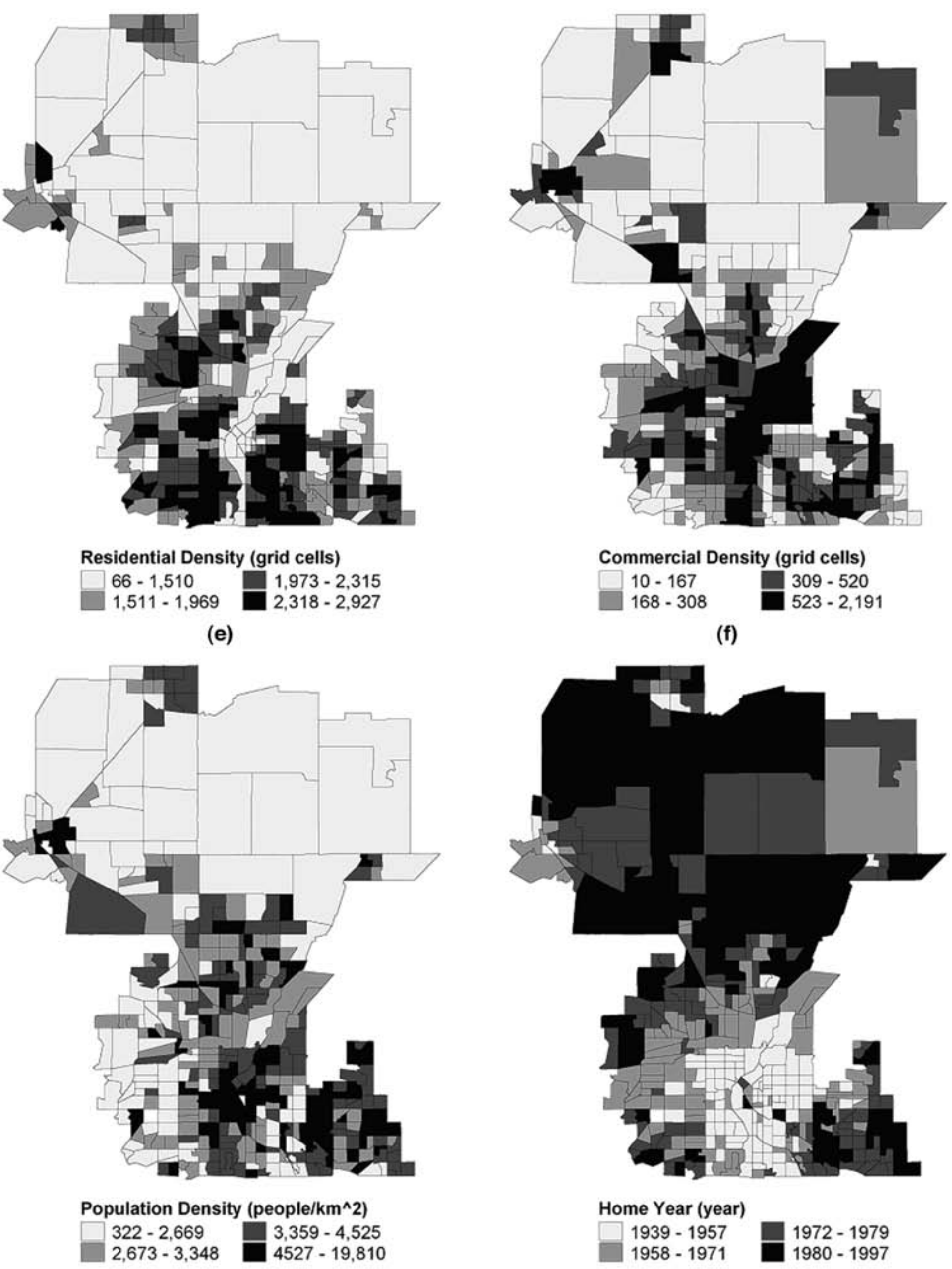

Figure 3. (Continued) Selected tract-level variables used in the analysis. See text for an explanation of the derivation of each variable: (e) Residential Density from grid cells, (f) Commercial Density from grid cells, (g) Population Density as people per square kilometer, and (h) Home Year by date built.

particularly as compared to land classified as commercial/light industrial. Notably, vegetated land-cover, which is almost exclusively herbaceous, also has a lower mean vegetation intensity than residential land. This is true for both native grassland and agricultural herbaceous vegetation cover.

The choropleth maps shown in Figure 3 suggest that the highest residential vegetation intensity values are found in certain parts of Denver as well as in the western third of the study region in general, particularly in areas such as Arvada and Boulder. Socioeconomic advantage exhibits a similar, though certainly not identical, pattern, as evidenced by the distribution of educational attainment. Interstate highways run east-west and north-south, respectively, throughout the study region and cross in north Denver, so that the highest distance to limited values are found primarily at the western edge of the study region. Residential density, commercial density, and population density all share broadly similar patterns, being higher in Denver and its immediate vicinity, as well as in Boulder and Longmont. Although, central Denver is notable for having high population density and commercial density, yet low residential density. Clearly, most people in that area are living in land classified as mixed urban rather than purely residential. Tracts with the oldest homes are found in Denver, as well as in the *old town* sections of Boulder and Longmont. Tracts with newer

Table 2. Summary of NDVI by Selected Land-cover Classes

Level 1	Level 2	Level 3	% Area	Mean	St. Dev.
Developed			43	0.16	0.16
	Residential		27	0.20	0.11
	Non-Res. Dev.		16	0.09	0.21
		Com./Lt. Ind.	7	−0.02	0.12
	Mixed Urban		0.3	0.04	0.14
Vegetated			52	0.17	0.19
	Woody		2	0.29	0.15
	Herbaceous		50	0.17	0.19
		Natural	16	0.14	0.14
		Planted	34	0.18	0.21
Bare			2	0.06	0.15
Water			3	−0.09	0.21

neighborhoods dominate the remainder of the northern half of the study region.

The correlation results are presented in Table 1 beside the descriptive statistics. All of the explanatory variables are significantly correlated to mean NDVI, with the exception of mean slope and distance to unlimited. As vegetation intensity increases, elevation, distance to limited, residential density, median income, educational attainment, number of rooms, and home value all increase. Commercial density, population density, and home year all decrease with increasing vegetation intensity. Vegetation intensity has the strongest relationships with population density, percent minority, and home value.

Results from the multivariate regression are presented in Table 3. Because they were not significantly correlated to vegetation intensity, mean slope and distance to unlimited were not included in the multivariate regression. Model 1 in Table 3 shows the results for the regression of vegetation intensity using the census-derived socioeconomic and housing variables as independent variables (median income was excluded in the model presented here because of multicollinearity). Population density, educational attainment, and home year are highly significant, and explain 43 percent of the variation in vegetation intensity. Home value is not significant, likely because its relationship with vegetation intensity is mediated by educational attainment, with which it has a Pearson $r > 0.5$ (significance < 0.0005). Model 2 focuses on factors of development and transportation. Commercial density, residential density, and distance to limited are all highly significant, with residential density explaining the greatest amount of the variation in vegetation intensity.

Table 3. Standardized Coefficients of Multivariate Regressions of Vegetation Intensity

Ind. Variable	Model 1	Model 2	Model 3	Model 4	Model 5
Mean Elevation			0.12***	0.15***	0.12***
Distance to Limited		0.27***	0.23***	0.25***	0.23***
Residential Density		0.43***	0.23***	0.25***	0.23***
Commercial Density		0.33***	−0.27***	−0.29***	−0.25***
Population Density	−0.46***		−0.29***	−0.28***	−0.27***
Percent Minority				−0.13***	
Educational Attainment	0.42***		0.21***		0.20***
Number of Rooms					0.06
Home Year	−0.35***		−0.33***	−0.30***	−0.33***
Home Value	−0.01				
Adjusted R^2	0.43	0.45	0.61	0.60	0.61

*** = significance < 0.005.

When the census, development and transportation, and elevation variables are combined in one equation, as in Model 3 in Table 3, they account for 61 percent of the variation in vegetation intensity. All the variables that were significant in Models 1 and 2 remain significant, with home year contributing the greatest amount to the slope of the regression line and the influence of residential density and commercial density slightly reduced. Mean elevation is highly significant in Model 3, but its contribution to the slope of the regression line is by far the least among all variables. The replacement of educational attainment with percent minority, and median income with number of rooms, as presented in Models 4 and 5, respectively, does not significantly change the nature of relationships among explanatory variables presented in Model 3.

One interesting result of the multivariate regression that is not reported in Table 3 is the zero-order and partial correlation values for each of the variables. Of note is that for home year in Model 3, the zero-order correlation is −0.13 and the partial correlation is −0.35. This indicates that the influence of home year in the model actually increases after the effects of the other explanatory variables are accounted for. For all the other variables in all the models, influence is reduced by the effects of the other explanatory variables. This unusual aspect of home year is explored further in the association rule mining results.

The association rule mining resulted in over 7,900 rules when all variable value combinations are allowed. When the restrictions to identify interesting rules are applied, 48 rules are identified, 14 with high vegetation intensity in the consequent and 34 with low vegetation intensity in the consequent. Table 4 shows ten representative rules, and their lift values, with high vegetation intensity in the consequent. The first rule may be read as "if residential density is high and percent minority is low then vegetation intensity is high," and the lift for this rule is 4.7. These rules generally echo the results of the multivariate regression in finding that high vegetation intensity is associated with high residential density and socioeconomic disadvantage. One result of interest is a comparison of rules 8 and 10. Rule 8 shows that neighborhoods near highways that are primarily white tend to have high vegetation intensity.

Table 4. Results of Association Rule Mining Predicting High Vegetation Intensity

#	Slp	Lim	Res	Pdn	Inc	Min	Edu	Yr	Val	Lift
1			high			low				4.7
2			high	low						4.4
3			high				high			4.4
4						low		low		4.4
5			high		high					4.3
6			high						high	4.0
7							high	low		4.0
8		low				low				3.7
9			high					low		2.5
10	high	high								2.1

Note: Slp = Mean Slope, Lim = Distance to Limited, Res = Residential Density, Pdn = Population Density, Inc = Median Income, Min = Percent Minority, Edu = Educational Attainment, Yr = Home Year, Val = Home Value.

Table 5. Results of Association Rule Mining Predicting Low Vegetation Intensity

#	Elv	Slp	Lim	Res	Com	Inc	Min	Edu	Yr	Val	Lft
11	low					low					4.5
12				low					low		4.5
13				low				low			4.5
14	low		low								4.3
15				low			high				4.0
16		low			high		high				3.9
17		low			high				low		3.8
18									high	low	3.7
19					high	low					3.6
20		low							low		2.4

Note: Elv = Mean Elevation, Slp = Mean Slope, Lim = Distance to Limited, Res = Residential Density, Com = Commercial Density, Inc = Median Income, Min = Percent Minority, Edu = Educational Attainment, Yr = Home Year, Val = Home Value.

Interestingly, rule 10 shows that neighborhoods far from highways that have steep slopes also have high vegetation intensity. Thus, both high and low distance to limited value tracts are associated with high vegetation intensity, but the nature of this relationship is dependent on the racial and environmental setting.

Table 5 reports ten representative rules with low vegetation intensity in the consequent, labeled rules 11 through 20. In agreement with the regression results, low vegetation intensity is associated with low residential density and high commercial density. Flat areas of low elevation are also associated with low vegetation intensity when paired with a high concentration of commercial land-use (rules 16, 17, and 19) or socioeconomic disadvantage (rule 11). Particularly interesting results are found for the home year variable. Low home year (i.e., an older neighborhood) is associated with low vegetation intensity when that neighborhood also has low residential density (rule 12) or high commercial density (rule 17). However, newer neighborhoods are also associated with low vegetation intensity when the home values in that neighborhood are low (rule 18). These rules may be contrasted with those in Table 4, in which older neighborhoods are associated with high vegetation intensity when paired with indicators of socioeconomic advantage (rules 4 and 7) and high residential density (rule 9).

Discussion

These results indicate that vegetation intensity in residential land is a function of a number of interrelated factors in the Front Range of Colorado. As one would expect, a high concentration of commercial land is associated with low vegetation intensity, as these areas are generally urban areas with little vegetation and large areas of impervious surfaces (roads, rooftops). This relationship may also reflect the low vegetation intensity in land classified as mixed urban, which occurs almost exclusively in downtown Denver and Boulder where it is surrounded by commercial land. A high concentration of residential land is associated with high vegetation intensity because of the presence of lawn grasses and deciduous trees associated with residential development. This *suburban lawn* ecosystem vegetation present on residential land increases NDVI values due to elevated biomass and photosynthetic activity as compared to the native shortgrass prairie vegetation or regional agricultural activity (Golubiewski and Wessman, 2006). One can speculate that extensive suburban and exurban style developments, as captured by high residential density values, are more likely to support suburban lawn ecosystem vegetation than smaller, isolated residential areas.

Lower population density is also associated with higher vegetation intensity. This is expected given the negative relationship of population density with vegetation fraction in many urban areas (Pozzi and Small, 2002). However, rule 2 indicates that the combination of sparsely populated, yet highly residential land-use, produces particularly high vegetation intensity. This suggests that large lot residential development, with managed vegetation but relatively few people, produces particularly high vegetation intensity values. This is particularly true in wealthy, white neighborhoods with high educational attainment and high home values.

Generally, older neighborhoods also tend to have higher vegetation intensity, likely because the managed vegetation is more mature. However, this is only the case when accompanied by indicators of large lot residential development and/or socioeconomic advantage. In fact, older residential neighborhoods nearby high concentrations of commercial land, or where home values are anomalously low, tend to have particularly low vegetation intensity. Figure 4 demonstrates the role of neighborhood age in determining vegetation intensity by mapping the results of rules 9 and 12. The area occupied by rule 9 (high residential density, low home year, and high vegetation intensity) is in Denver's wealthy Cherry Creek neighborhood, which contains many large deciduous trees and extensive grassy yards. Nearby is the area occupied by rule 12 (low residential density, low home year, and low vegetation intensity), which contains some of the oldest homes in the city on small lots adjacent to the downtown and other commercial centers. Contrast this pattern with the map of rules 10 and 13. The area occupied by rule 10 (high slope, high distance to limited, and high vegetation intensity) captures parts of Boulder and newer residential areas being built around the towns of Lafayette and Louisville, areas of socioeconomic advantage (i.e., see the map of educational attainment in Figure 3). Rule 13 (low residential density, low educational attainment, and low vegetation intensity), on the other hand, highlights tracts containing socioeconomically disadvantaged residents nearby areas of either commercial/light industrial land (in the southern tracts) or agricultural land (in the northern tracts).

These results suggest that, contrary to what one might expect, vegetation intensity in residential areas of the Front Range is not concentrated necessarily in older neighborhoods, nor exclusively in the suburbs, but in either older or suburban development-type neighborhoods that are also wealthy and white. It is important to note that the present study cannot reveal the process of causation by which this relationship occurs. While there is a relatively clear causal mechanism between, say, home year and vegetation intensity (intensity increases as vegetation matures and gains in biomass), it is unclear precisely how socioeconomic advantage influences vegetation intensity, or even whether the causal relationship acts in the opposite direction (i.e., vegetation intensity influences the socioeconomic status of a neighborhood).

Certainly, however, maintaining the suburban lawn ecosystem vegetation associated with higher vegetation intensity in residential areas demands significant watering and active landscape management practices (Golubiewski, 2003). Such practices are costly in the Front Range, where chronic water scarcity and urban growth have caused water restrictions and dramatic price increases in certain jurisdictions in recent years. One can speculate that, in the Front Range, residents transform their local residential landscape based on their ability to pay for the creation and maintenance

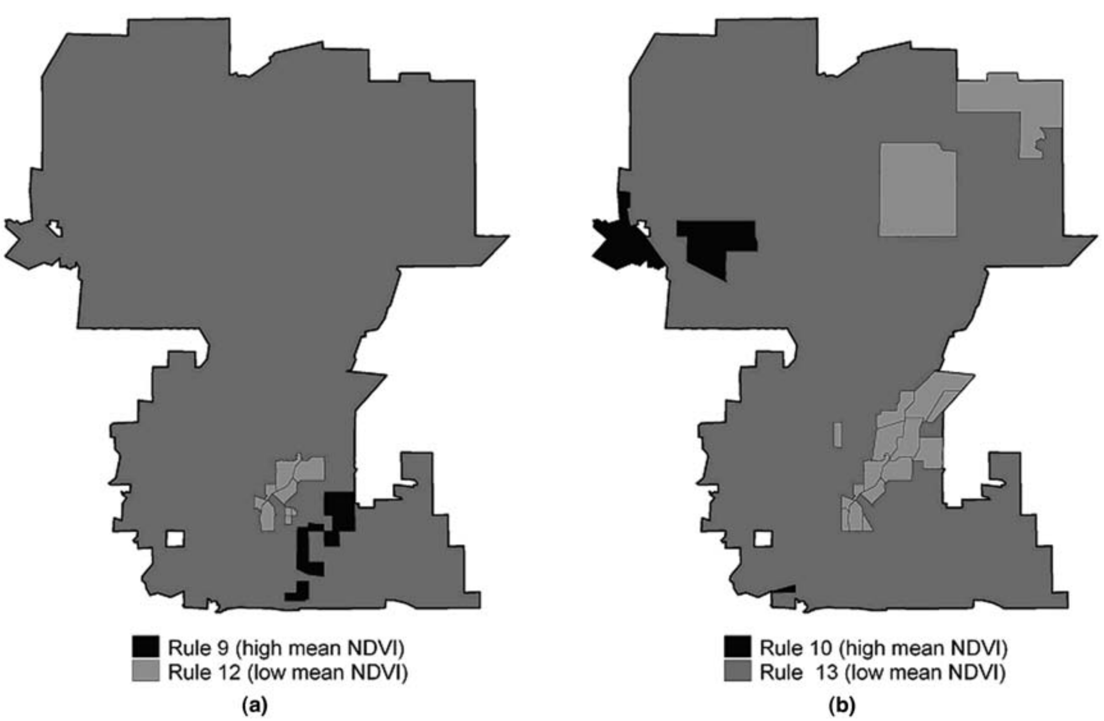

Figure 4. Maps of rules 9 and 12 (a) and rules 10 and 13 (b). The maps show the tracts where both the antecedent and consequent in each rule are *true*.

of suburban lawn ecosystem vegetation. Additionally, residential choices may be limited by a home price structure that is closely related to the concentration of vegetation in the residential landscaping.

These findings are of interest to those involved in analyzing and modeling urban growth and ecology, as they suggest that, in the Front Range, economic factors are closely related to urban vegetation intensity, and hence, ecological function. Conversely, previous research has shown that proximity to preserved open space and other environmental amenities, such as vegetation concentration, impacts property values and residential choice (Geoghegan *et al.*, 1997; Walsh, 2003). This feedback among socioeconomic status, vegetation intensity, and residential choice has ramifications on computational modeling of urban regions, as the majority of urban spatial models do not account for ecological-socioeconomic feedbacks in a sophisticated manner (Alberti and Waddell, 2002).

Conclusions

This research demonstrates the utility of integrating socioeconomic and remotely sensed imagery for investigating the interaction of urban ecological and social systems. Association rule mining has also been shown to be a useful tool in this investigation. While an exploratory method, association rule mining has the ability to capture relationships among variables that are nonlinear and/or occur in only a subset of the data. Such nonlinearities and data subsets may indicate thresholds in socioeconomic or ecological variables that are particularly important in the context of social-ecological system dynamics. For example, these results suggest that the influence of age of a neighborhood on vegetation intensity differs in nature depending on the co-occurrence of the socioeconomic character of that neighborhood. Recognizing such nonlinearities, and the subsets of data within which certain relationships among variables occur, can play a key role in understanding the complex interactions among drivers of urban ecosystem dynamics (Pickett *et al.*, 2001; Alberti *et al.*, 2004).

In the analysis presented here, interesting rules were found by setting criteria specified using a rule template and certain rule metrics. The motivation for selecting these criteria was to focus on identifying relationships among variables that tend to occur at the extremes of the data values, for instance to uncover what socioeconomic variables are associated with the very highest and lowest vegetation intensity values. The reasoning here is that social interactions with vegetation would likely be most easily detected in situations where the vegetation intensity was particularly enhanced or suppressed by human agency. The rule mining criteria were also used to restrict the volume of results to a manageable level as well as to simply bound the scope of analysis. However, it is certainly possible that other relevant relationships in the data exist, for instance among other subsets of the data residing in the middle three-fifths of the range for each variable. Such relationships can be investigated by relaxing the template and rule metric criteria. Investigating patterns in the data that are not expressed at the extreme values of the variables is a topic for future research.

This analysis is subject to certain limitations concerning data quality. The accuracy of the land-cover data is of particular concern, as NDVI was only calculated for land classified as residential or mixed urban. In urban and suburban areas, where large clusters of homes occur, manual vector digitizing of residential lands may be relatively straightforward. But in exurban or rural lands the delineation of residential versus naturally vegetated or agricultural land becomes more complicated, as individual homes are scattered across the landscape. A related issue is the resolution of the ETM+ imagery. At 30 m resolution, pixels likely contain a variety of land-covers, particularly in the urban areas (Jensen and Cowen, 1999). Consequently, when summarizing NDVI values within residential land polygons, pixels in the NDVI image that are in reality partially occupied by non-residential land-uses may be included. This error may be compounded by positional and attribute errors contained in the land-cover data, for which a minimum mapping unit of 2.5 acres is specified. Additionally, manual vector digitizing of land-cover necessarily involves some level of cartographic generalization, such as the smoothing of geometrically complex features and the determination of crisp boundaries between naturally *fuzzy* categories (Buttenfield and McMaster, 1991).

These data integration issues were mitigated to some extent by aggregating the residential NDVI data to the tract level, thereby reducing the effects that the error associated with any particular pixel or land-cover polygon might introduce. This presupposes that the positional error in the land-cover and other data sets is random, and not highly biased, the latter of which would be the case if the error was due primarily to problems of, perhaps, registration. While no formal testing on registration between the ETM+ imagery and land-cover data was done, a visual inspection found logically consistent spatial correspondence between the NDVI image and land-cover, tract boundary, and primary road data. Positional differences among the NDVI and land-cover data that could be observed did not appear to be biased in any particular direction.

Related to the choice of tracts as the unit of analysis is the modifiable areal unit problem (MAUP), which states that the analysis of spatially aggregated data may be impacted by the scale of data aggregation, as well as by the pattern of spatial partitioning at any one scale (Openshaw, 1983; Fotheringham and Wong, 1991). In an analysis focusing on northern Thailand, Walsh *et al*. (1999) show that the analysis of socioeconomic-remotely sensed vegetation relationships is affected by the MAUP. It is currently unknown whether the results of the present study would persist, and if so, to what degree, were the unit of analysis changed to the block group or to, say, an exhaustive tessellation of 1 km square cells. Tracts were chosen as the unit of analysis because they support the aggregation of the NDVI and land-cover data and are a common unit of analysis for socioeconomic analysis. The drawback is that tracts may be spatially heterogeneous in terms of socioeconomic status. The use of block groups instead of tracts would likely reduce this heterogeneity, though block groups in urban areas may be so small as to exacerbate the error introduced by data integration, as discussed above. The use of other spatial units, such as 1 km square cells, would necessitate the areal interpolation of the socioeconomic data, introducing another source of error.

Future research will address the issue of scale by performing the analysis at multiple scales to review the sensitivity of the results to scale variation. In addition, higher spatial and spectral resolution data may be used to identify not only general vegetation intensity but also specific types of vegetation, such as woody versus herbaceous vegetation types. This approach would improve the ability to distinguish the specific vegetation types associated with socioeconomic status and development characteristics, if any. Finally, a temporal analysis along the lines of Ryznar and Wagner's (2002) work would support an analysis of the causes of the socioeconomic-vegetation relationships found here by linking change in vegetation intensity with the transformation of land from non-residential to residential use. Understanding such causal relationships can ultimately inform research on how socioeconomic and ecological processes interact within the context of urban growth and land-cover change.

Acknowledgments

Thanks to Carol Wessman and Nancy Golubiewski for catalyzing this research, offering many insights into it, and for aid in data preprocessing.

References

Agrawal, R., T. Imielinski, and A. Swami, 1993. Mining association rules between sets of items in large databases, *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 207–216.

Alberti, M., and J.M. Marzluff, 2004. Ecological resilience in urban ecosystems: Linking urban patterns to human and ecological functions, *Urban Ecosystems*, 7:241–265.

Alberti, M., and P. Waddell, 2000. An integrated urban development and ecological simulation model, *Integrated Assessment*, 1(3):215–227.

Anderson, I.R., E.E. Hardy, J.T. Roach, and R.E. Witmer, 1976. *A Land Use and Land Cover Classification System for Use With Remote Sensor Data*, Reston, Virginia, U.S. Geological Survey Professional Paper 964.

Bannari, A., D. Morin, and F. Bonn., 1995. A review of vegetation indices, *Remote Sensing Reviews*, 13:95–120.

Buttenfield, B., and R. McMaster, 1991. *Map Generalization: Making Rules for Knowledge Representation*, London, Longman.

Carlson, T.N., R.R. Gillies, and E.M. Perry, 1994. A method to make use of thermal infrared temperature and NDVI measurements to infer surface soil water content and fractional vegetation cover, *Remote Sensing Reviews*, 9:161–173.

Carlson, T.N., and Ripley, D.A., 1997. On the relation between NDVI, fractional vegetation cover, and leaf area index, *Remote Sensing of Environment*, 62:241–252.

Chen, D., J. Weeks, and J. Kaiser. 2004. Remote sensing and spatial statistics as tools in crime analysis (F. Wang, Editor), *Geographic Information Systems and Crime Analysis*, Hershey, Pennsylvania, Idea Group Publishing.

Civco, D.L., J.D. Hurd, E.H. Wilson, C.L. Arnolld, and M.P. Prisloe, 2002. Quantifying and describing urbanizing landscapes in the Northeast United States, *Photogrammetric Engineering & Remote Sensing*, 68(10):1083–1090.

Fotheringham, A.S., and D.W.S. Wong, 1991. The modifiable areal unit problem in multivariate statistical analysis, *Environment and Planning A*, 23:1025–1044.

Fox, J., R.R. Rindfuss, S.J. Walsh, and V. Mishra, 2003. *People and the Environment: Approaches for Linking Household and Community Surveys to Remote Sensing and GIS*, Boston, Kluewer Academic Publishers.

Fu, Y., and J. Han, 1995. Meta-rule-guided mining of association rules in relational databases, *Proceedings of the International Workshop on the Integration of Knowledge Discovery with Deductive and Object-Oriented Databases*, pp. 39–46.

Geoghegan, J., L. Wainger, and N. Bockstael, 1997. Spatial landscape indices in a hedonic framework: An ecological economics analysis using GIS, *Ecological Economics*, 23:251–264.

Geoghegan, J., S.C. Villar, P. Kelpeis, P.M. Mendoza, Y, Ogneva-Himmelberger, R.R.Chowdhury, I.B.L. Turner, and C. Vance, 2001. Modeling tropical deforestation in the southern Yucatan peninsular region: comparing survey and satellite data, *Agriculture, Ecosystems and Environment*, 85(1–3):25–46.

Golubiewski, N.E., 2003. *Carbon in Conurbations: Afforestation and Carbon Storage as Consequences of Urban Sprawl in Colorado's Front Range*, Ph.D. Dissertation, University of Colorado, Boulder, Colorado.

Golubiewski, N.E., and C.A. Wessman, 2006. Urbanization transforms prairie carbon pools: Effects of landscaping in Colorado's Front Range, *Ecological Applications*, 16(2):555–571.

Grimm, N.B., J.M. Grove, S.T.A. Pickett, and C.L. Redman, 2000. Integrated approaches to long-term studies of urban ecosystems, *BioScience*, 50(7):571–584.

Herold, M., N.C. Goldstein, and K.C. Clarke, 2003. The spatio-temporal form of urban growth: Measurement, analysis and modeling, *Remote Sensing of Environment*, 86:286–302.

Huang, S., and G.I. Webb, 2004. Efficiently identifying exploratory rules' significance, *Proceedings of the 2004 Australasian Data Mining Workshop*, pp. 169–182.

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

Karathannassi, V., C.H. Jossifidis, and D. Rokos, 2000. A texture-based classification method for classifying built areas according to their density, *International Journal of Remote Sensing*, 21(9): 1807–1823.

Koperski, K., and J. Han, 1995. Discovery of spatial association rules in geographic information databases, *Proceedings of the Fourth International Symposium on Large Spatial Databases*, pp. 47–66.

Liverman, D., E.F. Moran, R.R. Rindfuss, and P.C. Stern (editors), 1998. *People and Pixels: Linking Remote Sensing and Social Science*. Washington, D.C., National Academy Press.

Lo, C.P., and B.J. Faber, 1997. Integration of Landsat Thematic Mapper and census data for quality of life assessment, *Remote Sensing of Environment*, 62:143–157.

Mennis, J., and J.W. Liu, 2005. Mining association rules in spatio-temporal data: An analysis of urban socioeconomic and land cover change, *Transactions in GIS*, 9(1):5–17.

Mesev, V. (editor), 2003. *Remotely Sensed Cities*, London, Taylor and Francis.

Myneni, R.B., C.J. Tucker, G. Asrar, and C.D. Keeling., 1998. Inter-annual variations in satellite-sensed vegetation index data from 1981 to 1991, *Journal of Geophysical Research*, 103:6145–6160.

Openshaw, S., 1983. The Modifiable Areal Unit Problem, *Concepts and Techniques in Modern Geography*, Volume 38, Norwich, UK, Geobooks.

Pfaff, A.S.P., 1999. What drives deforestation in the Brazilian Amazon? Evidence from satellite and socioeconomic data, *Journal of Environmental Economics and Management*, 37:26–43.

Pickett, S.T.A., M.L. Cadenasso, J.M. Grove, C.H. Nilon, R.V. Pouyat, W.C. Zipperer, and R. Costanza, 2001. Urban ecological systems: Linking terrestrial ecological, physical, and socio-economic components of metropolitan areas, *Annual Review of Ecology and Systematics*, 32:127–157.

Pozzi, F., and C. Small, 2002. Vegetation and population density in urban and suburban areas in the U.S.A., *Proceedings of the Third International Symposium of Remote Sensing of Urban Areas*: pp. 489–496.

Rashed, T., J.R. Weeks, D. Roberts, J. Rogan, and R. Powell, 2003. Measuring the physical composition of urban morphology using multiple endmember spectral mixture models, *Photogrammetric Engineering & Remote Sensing*, 69(9):1011–1020.

Riebsame, W.E., H. Gosnell, and D. Theobald, 1997. *Atlas of the New West*, New York, W.W. Norton and Company.

Ryznar, R.M., and T.W. Wagner, 2001. Using remotely sensed imagery to detect urban change: Viewing Detroit from space, *Journal of the American Planning Association*, 67(3):327–336.

Seto, K.C., and R.K. Kaufmann, 2003. Modeling the drivers of urban land-use change in the Pearl River delta, China: integrating remote sensing with socioeconomic data, *Land Economics*, (79)1:106–121.

Small, C., 2001. Estimation of urban vegetation abundance by spectral mixture analysis, *International Journal of Remote Sensing*, 22(7):1305–1334.

Stier, M., 1999. *Temporal land-use and land-cover mapping*, Oral presentation at the American Society for Photogrammetry and Remote Sensing (ASPRS) Conference, Portland, Oregon, URL, *http://rockyweb.cr.usgs.gov/frontrange/land/templanduse/apsabs.htm* (last date accessed: 18 May 2006).

Sutton, P., D. Roberts, C. Elvidge, and K. Baugh. 2001. Census from heaven: An estimate of the global human population using nighttime satellite imagery, *International Journal of Remote Sensing*, 22 (16):3061–3076.

Tan, P-N., V. Kumar, and J. Srivastava, 2002. Selecting the right interestingness measure for association patterns, *Proceedings of the Eighth International Conference on Knowledge Discovery and Data Mining*, pp. 32–41.

Tucker, C.J., 1979. Red and photographic infrared linear combinations for monitoring vegetation, *Remote Sensing of Environment*, 8:127–150.

Walsh, R., 2003. Analyzing open space policies in a locational equilibrium model with endogenous landscape amenities, *Proceedings of the Association of Environmental and Resource Economists Summer Workshop: Spatial Theory Modeling and Econometrics in Environmental and Resource Economics*, URL: http://www.aere.org/meetings/0306workshop.html (last date accessed: 18 May 2006).

Walsh, S.J., T.P. Evans, W.F. Welsh, B. Entwisle, and R.R. Rindfuss, 1999. Scale-dependent relationships between population and environment in northeastern Thailand, *Photogrammetric Engineering & Remote Sensing*, 65(1):97–105.

Webb, G.I., 1995. OPUS: An efficient admissible algorithm for unordered search, *Journal of Artificial Intelligence Research*, 3:421–465.

Webb, G.I. and Associates Pty. Ltd., 2001. *Magnum Opus*, Version 1.3.

Weeks, J., M. Gadalla, T. Rashed, J. Stanforth, and A. Hill. 2000. Spatial variability in fertility in Menoufia, Egypt assessed through the application of remote-sensing and GIS technologies, *Environment and Planning A*, 32(4):695–714.

Wilson, J.S., M. Clay, E. Martin, D. Stuckey, and K. Vedder-Risch, 2003. Evaluating environmental influences of zoning in urban ecosystems with remote sensing, *Remote Sensing of Environment*, 86(3):303–321.

Yuan, Y., R.M. Smith, and W.F. Limp, 1997. Remodeling Census population with spatial information from Landsat TM imagery, *Computers, Environment and Urban Systems*, 21(3/4):245–258.

(Received 29 April 2005; accepted 15 June 2005; revised 29 June 2005)