Incorporating Remote Sensing Information in Modeling House Values: A Regression Tree Approach

Danlin Yu and Changshan Wu

Abstract

This paper explores the possibility of incorporating remote sensing information in modeling house values in the City of Milwaukee, Wisconsin, U.S.A. In particular, a Landsat ETM+ image was utilized to derive environmental characteristics, including the fractions of vegetation, impervious surface, and soil, with a linear spectral mixture analysis approach. These environmental characteristics, together with house structural attributes, were integrated to house value models. Two modeling techniques, a global OLS regression and a regression tree approach, were employed to build the relationship between house values and house structural and environmental characteristics. Analysis of results indicates that environmental characteristics generated from remote sensing technologies have strong influences on house values, and the addition of them improves house value modeling performance significantly. Moreover, the regression tree model proves as a better alternative to the OLS regression models in terms of predicting accuracy. In particular, based on the testing dataset, the mean average error (MAE) and relative error (RE) dropped from 0.202 and 0.434 for the OLS model to 0.134 and 0.280 for the regression tree model, while the correlation coefficient between the predicted and observed values increased from 0.903 to 0.960. Further, as a nonparametric and local model, the regression tree method alleviates the problems with the OLS techniques and provides a means in delineating urban housing submarkets.

Introduction

Urban analysis has become an important research topic across a range of disciplines due to the continuous increase of population residing in urban environments (United Nations, 1997; Newman and Kenworthy, 1999). To satisfy the demands of urban analysis, innovative remote sensing technologies and their applications in urban environments have emerged recently (Carlson, 2003; Mesev, 2003). One major aspect of urban remote sensing lies in direct estimation of urban biophysical parameters and socio-economic characteristics. Urban biophysical parameter estimation includes extracting urban land-use and land-cover features (Treitz *et al.*, 1992; Harris and Ventura, 1995), deriving impervious surface distribution (Rashed *et al.*, 2003; Wu and Murray, 2003), evaluating urban vegetation fraction (Small, 2001; 2002), and determining urban surface temperature (energy patterns) and associated heat-island effects (Lo *et al.*, 1997; Weng *et al.*, 2004). Socio-economic characteristic estimation includes population density estimation (Lo, 1995; Sutton *et al.*, 1997; Harvey, 2002a; 2002b), housing information estimation (Forster, 1983; Weber and Hirsch, 1992), employment distribution estimation (Lo, 2004), quality of life index estimation (Lo, 1997), and urban racial segregation estimation (Yu and Wu, 2004). The other aspect of urban remote sensing aims at incorporating remote sensing information, along with other spatial data, into urban prediction models. In particular, remote sensing generated land-use information has been incorporated into cellular automata models to predict future urban development (Yeh and Li, 2003). Dobson *et al.* (2000) reported the results of the LandScan global population project, in which population distribution was modeled with remote sensing generated variables (e.g., land-cover and nighttime lights) and other spatial variables (e.g., road network, slope, and census counts). Weeks *et al.* (2004) applied remote sensing information, together with socio-economic information from census data, to predict fertility patterns in Cairo, Egypt using a spatially filtered regression model.

This study intends to achieve two major objectives. The first is to explore the possibility of incorporating remote sensing information in modeling house values. The other objective is to address technique issues of hedonic models. Traditionally, house values are modeled with house structural and/or locational attributes (such as building area and number of bathrooms) through ordinary least squares (OLS) regression analyses. Such models are also called hedonic regression models, and have been applied extensively in housing market studies (Dubin, 1998; Mulligan, 2002). These models generally focus on building the relationship between house values/prices and various house structural characteristics and/or locational attributes. For environmental attributes, the appraisal community focuses more on the influence of environmental risks/contamination on property values (Roddewig and Keiter, 2001; Jackson, 2003; 2004). Other environment related neighborhood attributes, such as exposure to water views (Benson *et al.*, 2000) and distance to

Danlin Yu is with the Department of Earth & Environmental Studies, Montclair State University, Montclair, NJ 07043 (yud@mail.montclair.edu) and formerly with the University of Wisconsin – Milwaukee, Department of Geography, Milwaukee, WI 53201

Changshan Wu is with the University of Wisconsin – Milwaukee, Department of Geography, Milwaukee, WI 53201

Photogrammetric Engineering & Remote Sensing
Vol. 72, No. 2, February 2006, pp. 129–138.

0099-1112/06/7202–0129/$3.00/0

coastal beach (Major and Lusht, 2004), are included in hedonic models. However, inclusion of environmental information that can be generated from remote sensing imagery in the hedonic models has attracted less attention. The environmental information, as pointed out by Forster (1983), might have essential influences on residents' decision making, therefore affecting house values. Hence, evaluating the influence of environmental characteristics, which can be effectively extracted from remote sensing imagery, on house values merits further investigation.

In addition, traditional hedonic models are usually constructed and calibrated through linear OLS regression techniques. Linear regression techniques, though widely adopted in the literature (Kim, 2003), with multiple house and locational attributes involved, might produce problems such as collinearity, residual heteroscedasticity, and spatial dependency that may adversely affect modeling results (Anselin, 1988; Basu and Thibodeau, 1998; Dublin, 1998; Pace *et al.*, 1998; Orford, 1999; 2000). Thus, it is imperative to investigate new modeling techniques instead of relying solely on the linear models.

In this paper, environmental characteristics, including the fractions of vegetation, impervious surface, and soil, were generated from a Landsat ETM+ image. These environmental characteristics, together with house structural attributes, were integrated to model house values in the City of Milwaukee. A global OLS regression and a regression tree approach were developed to model house values. Results were mapped to explore potential spatial patterns.

The remainder of this paper is organized as follows. The next section describes the study area and utilized data, including house values and structural information extracted from the 2003 Master Property (MPROP) dataset of the City of Milwaukee, and environmental characteristics generated from Landsat ETM+ imagery. Next, two modeling techniques, a linear OLS regression and a regression tree approach, are described followed by results obtained from these two techniques, respectively and the spatial patterns of the analytical results. Finally, conclusions and future research are given.

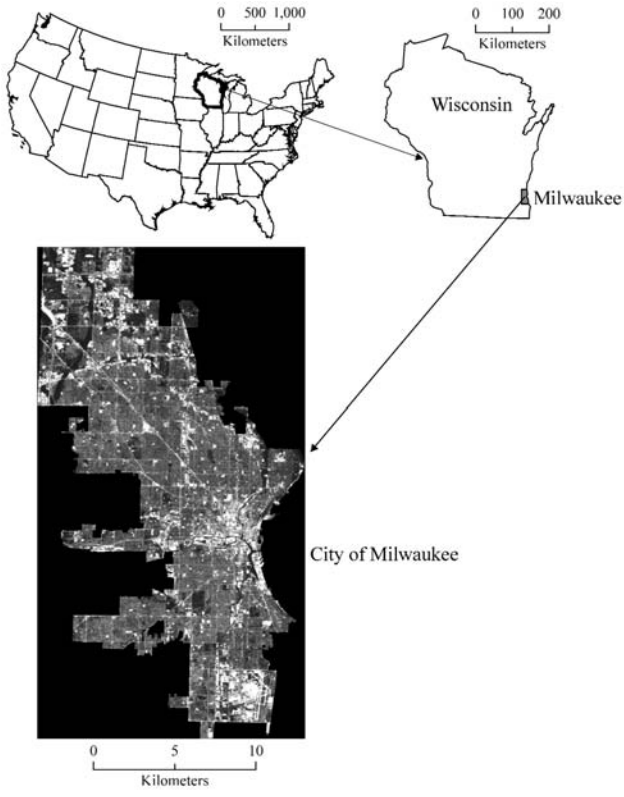

Figure 1. Location of the City of Milwaukee (Color version available at the ASPRS Web Site, URL: *www.asprs.org*).

Study Area and Data

Study Area

The City of Milwaukee was selected as our study area. Located on the western shore of Lake Michigan (Figure 1), Milwaukee has a population of about 597,000 according 2000 Census. Population grew rapidly during the early 20th Century, primarily due to immigration from Central and Eastern Europe. The City's boundary has been more or less fixed since the late 1950s, and after the completion of the current highway network in the late 1960s, Milwaukee stepped into a relatively stable period of property development.

Milwaukee is usually referred to as a "hyper-segregated" city (Massey and Denton, 1993; Yu and Wu, 2004). Spatially, African Americans are highly concentrated in the areas near the current downtown to the west (Yu and Wu, 2004), with a high residential density and poor environmental conditions. Residential segregation is still a discernible social issue, and has profound impacts on housing markets in the city.

Data

House Values and Structural Characteristics

House values and structural characteristics were extracted from the 2003 Master Property (MPROP) data file of Milwaukee. The MPROP data file has approximately 160,000 records of all real properties within the city boundary. Each record contains more than 80 various attributes including house's location, assessed value, owner information, and physical characteristics. With the emphasis of this paper focusing on owner-occupied single-family houses, 68,906 records with various house attributes were extracted and utilized for this study. From the MPROP data file, it is noticed that only assessed house values, instead of sale prices, can be obtained. Although the sales price can be calculated through the conveyance fee and the conveyance date that are recorded (Kim, 2003), the calculation might introduce new uncertainty in evaluating houses' market prices. In addition, according to Wisconsin law, the assessed value and the market value of a house cannot vary by more than 10 percent. Therefore, the assessed house values were utilized as an approximation for current house market values.

Environmental Characteristics from Landsat ETM+ Imagery

In addition to the house structural characteristics obtained from the MPROP data file, environmental characteristics were generated from a Landsat ETM+ image (see Figure 1) acquired on 09 July 2001. This image was provided by the WisconsinView project (WisconsinView, 2004), and has an average root mean square error of 0.2 pixels after georeferenced with ground reference information collected from aerial photographs. With this Landsat ETM+ image, three environmental characteristics, the fractions of vegetation, impervious surface, and soil for each pixel, were generated using the normalized spectral mixture analysis method proposed by Wu (2004). In particular, a brightness normalization method (see Equation 1) was applied to reduce or

eliminate brightness variation within a land-cover type (e.g., spectra variation within different soil types):

$$\overline{R}_b = \frac{R_b}{\mu} \times 100 \quad (1)$$

where $\overline{R}_b$ is the normalized reflectance for band b in an ETM+ pixel; R_b is the original reflectance for band b; and μ is the average reflectance for that pixel. With the normalized ETM+ image, three endmembers, vegetation, impervious surface, and soil, were selected to model urban land-cover types. These endmembers were obtained based on the feature space representation of the normalized image after principal component (PC) transformation, in company with visual interpretation of the ETM+ imagery. Detailed information about endmember calculation can be found in Wu (2004). Subsequently, a spectral mixture analysis method (Equation 2) was applied to calculate the fraction of each endmember for every ETM+ pixel:

$$\overline{R}_b = \sum_{i=1}^{N} \overline{f}_i \overline{R}_{i,b} + e_b \quad (2)$$

where $\overline{R}_b$, which was calculated using Equation 1, is the normalized reflectance for each band b in a pixel; $\overline{R}_{i,b}$, which was determined from analyzing the image spectra, is the normalized reflectance of endmember i in band b for that pixel; $\overline{f}_i$ is the fraction of endmember i, and it is constrained that the sum of endmember fractions equals one and each fraction is greater or equal to zero; and e_b is the model residual. The fractions of endmember vegetation, impervious surface, and soil in an ETM+ pixel can be obtained through a least squares method in which the model residual e_b is minimized. By applying this normalized spectral mixture analysis method in the study area, the fractions of vegetation, impervious surface, and soil for each pixel were calculated (Figure 2).

Data Aggregation

With the extracted house values, structural attributes, and environmental characteristics, it is necessary to preprocess these data such that they can share a unified spatial unit. Since it is unrealistic and impractical to obtain detailed remote sensing information for a single house, a model to incorporate remote sensing information has to be constructed on aggregated data. Ideally, the aggregated spatial units should be relatively homogeneous in describing the housing information as well as capable to handle subtle remote sensing information. Census block group was chosen to accomplish the aggregation since block groups are delineated primarily according to neighborhood homogeneity. There are in total 591 census block groups in the City of Milwaukee. However, only 571 of them contain owner-occupied single-family houses. The 20 block groups (mainly in the downtown area of the city) that do not contain the required data were hence deleted from the analysis. The aggregation of house attributes is essentially an average of individual data items within a census block group. For continuous variables, such as house values and building areas, the average represents their arithmetic mean within a census block group. The average of dummy variables, such as whether or not have air conditioners, changes to represent the percentage of houses in a census block group that have the attributes (e.g., air conditioners).

Methodology

A Linear Model Specification

Following the literature of constructing a house hedonic model (Berry and Bednarz, 1975; Goodman, 1978; Thibodeau, 1989; Cheshire and Sheppard, 1995; Kim, 2003), initially six independent variables describing various house attributes were identified and retrieved from the MPROP data file. In particular, two dummy variables, AIRCD and FIREPLC, indicating whether central air-conditioners and fireplaces are present or not, and four continuous variables including floor size (FLSIZE), number of bathrooms (NOFBATH), number of stories (NOFST), and house age (HSAGE), were chosen to model house values. Intuitively, AIRCD, FIREPLC, FLSIZE, NOFBATH, and NOFST

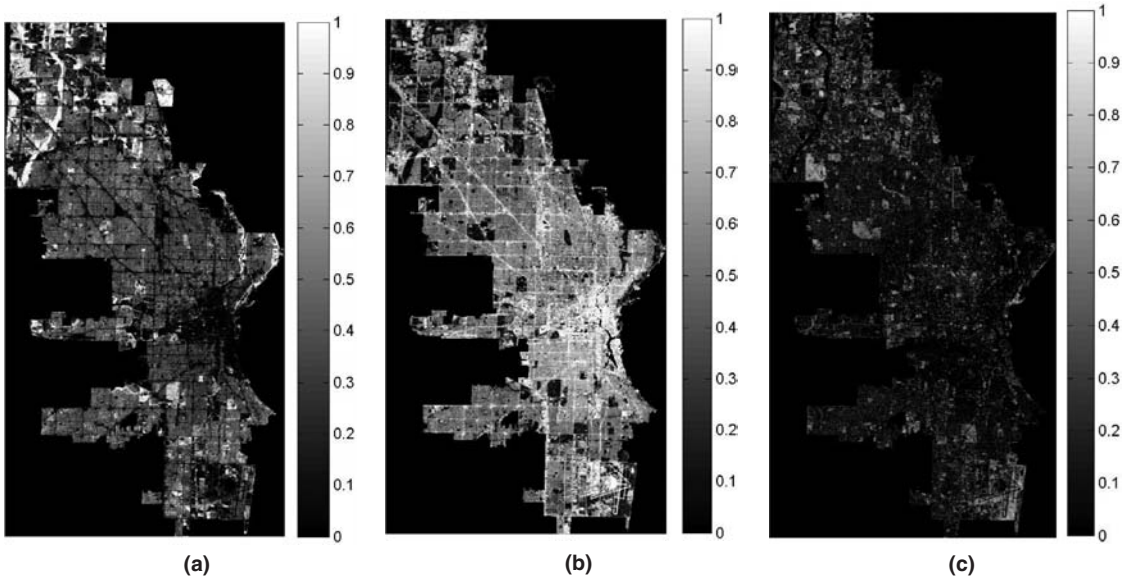

Figure 2. Fraction images of vegetation, impervious surface, and soil generated from the Landsat ETM+ image: (a) Vegetation, (b) Impervious surface, and (c) soil using the normalized spectral mixture analysis method proposed by Wu (2004).

are hypothesized to be positively related with house values; while HSAGE is hypothesized to be negatively related with house values. According to each house attribute's distribution characteristics, a classical semi-log hedonic price model is constructed below:

$$\log P_i = \beta_0 + \beta_1 \log X_{1i} + \beta_2 \log X_{2i} + \cdots + \alpha_1 D_{1i} + \alpha_2 D_{2i} + \cdots + \varepsilon_i \quad (3)$$

where the X_s are continuous variables including FLSIZE, HSAGE, NOFBATH, and NOFST; β_s is their coefficient; D_s are the aggregated dummy variables that include AIRCD and FIREPLC, and α_s is their coefficient.

To examine whether remote sensing generated environmental characteristics can improve house value modeling results, a remote sensing hedonic model is constructed by adding relevant environmental factors to the above model:

$$\log P_i = \beta_0 + \beta_1 \log X_{1i} + \cdots + \alpha_1 D_{1i} + \alpha_2 D_{2i} + \cdots + \gamma_1 R_{1i} + \gamma_2 R_{2i} + \cdots + \varepsilon_i \quad (4)$$

where R_s are environmental factors generated using remote sensing technologies, and γ_s are their coefficients. During our preliminary data analysis, we found that all of the three remote sensing derived environmental factors (fractions of vegetation, impervious surface, and soil in a census block group) project significant influences on house values. The most important factor, however, is the product of the soil fraction and impervious surface fraction (SOILIMP). This factor, representing the combined effects of soil and impervious surfaces and usually deemed to be an indicator of deteriorated environmental conditions, is hypothesized to have a negative relationship with house values.

Regression Tree Specification

In calibrating hedonic models, an ordinary least square (OLS) algorithm is often employed. However, it is noticed that the OLS calibration might possess many problems, such as collinearities and residual heteroskedesticity, that may adversely affect the model's accuracy. More importantly, as the OLS calibration is essentially a global treatment of the relationship, it assumes such a relationship will hold for any houses. In practice, this speculation is doubtful. For instance, it might be reasonable to think that certain house attributes, such as the number of bathrooms, have much more of an influence on house values for expensive houses than for inexpensive ones. Actually, many have argued that the impacts of house characteristics on house values may vary considerably within a metropolitan housing market (for example, see Straszheim, 1975; Maclennan, 1982; Rothenberg *et al.*, 1991; Orford, 2000). Hence, a functional metropolitan housing market operates as a series of interlinked housing submarkets (Maclennan, 1982). The stratification methods of such submarkets are still under debate (Adair *et al.*, 1996), which focus on whether such submarkets should be stratified by house structural characteristics or geographical areas. However, as argued by Orford (2000), such dichotomy might be problematic as locational and house structural attributes are very likely interconnected instead of separated.

Following the spirit of these arguments, a classification and regression tree (CART, Breiman *et al.*, 1984; Ripely, 1996) algorithm was employed in this study. As a non-parametric algorithm, the CART may be a better alternative to model house values without the problems that make the OLS technique untenable. Moreover, the CART is essentially a local algorithm, which may be appropriate to investigate housing submarkets and how such submarkets being divided in the city.

In general, the CART algorithm is composed of two parts, i.e., the classification tree and regression tree algorithms. The former is usually used to deal with categorical dependent variables; while the latter is concerned with continuous dependent variables, which is of the interest in this study. In brevity, the regression tree algorithm conducts a recursive binary partition on the data. Depending on how the dependent variable and the independent variables interact with each other, it grows a (inverted) categorical tree by repeatedly splitting the data according to specific rules, which define the conditions for data splitting. The goal of the algorithm is to categorize the data into more homogeneous groups by uncovering the predictive structure of the problem under consideration (Breiman *et al.*, 1984). In a highly condensed form, the regression tree algorithm is implemented in two steps. First, it grows a tree by splitting the predictors (independent variables) using various splitting rules (Breiman *et al.*, 1984) to achieve the best predicting accuracy. Obviously, this step will yield a very complex tree that has many different classes with only a few cases in each (Breiman *et al.*, 1984; Steinberg and Colla 1995). Second, the large tree will be "pruned" by minimizing the cost-complexity measurement described by Brieman *et al.* (1984). After the initial tree is pruned, the new tree will assign all cases to rule-defined groups. For each group, a multivariate regression model can be established for exploring the relationship between independent and dependent variables within that group. Therefore, the regression tree algorithm is considered as a local model, in which different relationships hold for different groups. The performance of the regression tree model was reported to be more accurate than ordinary regression models (Huang and Townshend, 2003; Yang *et al.*, 2003). Furthermore, Quinlan (1993) argued that a combination of rule-based and instance-based algorithms might provide an even better model performance than a pure rule-based regression tree model. For a better prediction of a target case, the combined algorithm initially finds the most similar cases (neighboring cases) to the target case within the training dataset. Then, a rule-based model is applied to predict the target case and these neighboring cases. Finally, the difference between the predicted value and the true value of each neighboring case is evaluated and taken into account in predicting the value of the target case.

Accuracy Assessment

As a nonparametric method, the regression tree algorithm does not assume any *a priori* distribution of the variables in the model. This relaxation of variable distribution assumptions enables the algorithm to be quite robust in dealing with outliers, collinearities among independent variables, heteroskadesticity, and/or distributional error structures that might cause problems in parametric analyses (Breiman *et al.*, 1984). However, this also makes it impossible to carry out hypothesis testing as conducted in parametric models. In practice, three statistics are employed to assess the model's performance. They are the mean average error (MAE), the relative error (RE), and the product-moment correlation coefficient (R) between the predicted values and the actual values of the model's dependent variable.

The MAE measures the absolute prediction error of the model. It is obtained through the formula:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (5)$$

where n is the number of observations, y_i is the dependent variable at observation i, and $\hat{y}_i$ is the estimated value of y_i at observation i.

The relative error indicates a relative improvement of the model on the global mean, and is taken the form:

$$RE = \frac{\sum_{i=1}^{n} \left| y_i - \hat{y}_i \right|}{\sum_{i=1}^{n} \left| y_i - \bar{y} \right|} \tag{6}$$

where all the labels are as defined above with $\bar{y}$ representing the global mean of the dependent variable.

The product-moment correlation coefficient R between the actual and predicted values measures the quality of *least square fitting* for the predicted values to the actual ones. It follows:

$$R = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2 \sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2}}. \tag{7}$$

In the software package used in this study, Cubist (see **http://www.rulequest.com/cubist-info.htm** for detailed information), these three statistics are calculated automatically when a model is built. They are hence used to measure the quality of the regression tree model and compare the performance of the regression tree model against the OLS model.

Furthermore, since the judgment of the model performance is more informative using different dataset than the one that constructs it, a regression tree model is usually constructed and tested using disjointed sets of data (the training and testing datasets). In this study, the training and testing datasets are generated randomly following a 90 percent-training to 10 percent-testing division. The choice of 90 to 10 reflects our intention to use as many cases as possible to ensure better model performance for the OLS algorithm. For comparison purposes, the exact same training and testing datasets were used in the OLS and regression tree models.

Results

Linear Regression Model

Two linear regression analyses were conducted to determine whether remote sensing generated environmental characteristics would contribute in explaining the variation of house values in the City of Milwaukee. Models were constructed based on the 90 percent training dataset and validated using the 10 percent testing dataset. The first model regressed the house values on the six house attributes, and the second one added remote sensing generated environmental characteristics. Results are reported in Table 1 and Table 2. In addition to regular statistics such as *t*-values of the coefficients, *F* statistics of the models, and adjusted R-squares, the three accuracy statistics (MAE, RE, and R) were calculated for further comparisons with the regression tree model.

From the two tables, three inferences can be made:

1. The *F*-statistics of the two models indicate that the relationship between house values and the six elected house attributes and remote sensing generated environmental characteristics is significant. The adjusted R-squares in both models indicate that the constructed models can explain around 70 percent of the variation of the house values in the City of Milwaukee.
2. When comparing the adjusted R-squares of these two regression models, we found that the inclusion of remote sensing generated environmental characteristics in the second model increases the explaining power of the model by around 4 percent. Moreover, the hypothesis of negative relationship between the composite factor of soil and impervious surfaces and house values is supported by the analysis as well. This means that house values are generally low in regions where impervious surfaces and soil are abundant. As the abundance of soil and impervious surfaces generally indicates a relatively deteriorated neighborhood environmental condition, this relationship is quite intuitive. This result indicates that remote sensing generated information can be a valuable addition to the traditional house value models.
3. With a close inspection on the coefficient and p-value for each independent variable, most of the hypotheses seem to be supported by the data except for house age (HSAGE) and floor size (FISIZE). In the model without the environmental variable, house age projects highly significant positive influences on house values (Table 1), although it turns to be insignificant when the environmental variable is added (Table 2). Furthermore, the floor size (FISIZE) does not have a significant influence on house values (see Tables 1 and 2). These results seem to be quite counter-intuitive. However, correlation

TABLE 1. ORDINARY LINEAR REGRESSION WITHOUT REMOTE SENSING INFORMATION

Dependent Variable: House Price, Log Transformed.
Residual Standard Error: 0.306 (training data), 0.274 (testing data)

Coefficients:	Estimate	Std. Error	t-Value	Pr(>|t|)
Intercept	8.345	1.184	7.048	0.000
AIRCD	1.623	0.101	16.147	0.000
FLSIZE	0.056	0.181	0.310	0.756
FIREPLC	1.088	0.127	8.567	0.000
HSAGE	0.368	0.073	5.014	0.000
NOFBATH	0.692	0.179	3.871	0.000
NOFST	0.119	0.146	0.816	0.415

Adjusted R-square: 0.689
F-statistic: 190.8 on 6 and 507 DF, p-value: <2.2e-16
Accuracy statistics:

Training data	
MAE	0.224
RE	0.517
Correlation Coefficient *R*	0.832
Testing data	
MAE	0.212
RE	0.457
Correlation Coefficient *R*	0.892

TABLE 2. ORDINARY LINEAR REGRESSION WITH REMOTE SENSING INFORMATION

Dependent Variable: House Price, Log Transformed
Residual Standard Error: 0.285 (training data), 0.263 (testing data)

Coefficients:	Estimate	Std. Error	t-value	Pr(>|t|)
Intercept	10.476	1.131	9.262	0.000
AIRCD	1.498	0.095	15.792	0.000
FLSIZE	−0.031	0.170	−0.184	0.854
FIREPLC	0.839	0.122	6.877	0.000
HSAGE	0.090	0.076	1.189	0.235
NOFBATH	0.844	0.168	5.032	0.000
NOFST	0.289	0.138	2.099	0.036
SOILIMP	−5.626	0.643	−8.747	0.000

Adjusted R-square: 0.730
F-statistic: 198.8 on 7 and 506 DF, p-value: <2.2e-16
Accuracy statistics:

Training data	
MAE	0.213
RE	0.493
Correlation Coefficient *R*	0.856
Testing data	
MAE	0.202
RE	0.434
Correlation Coefficient *R*	0.903

analyses on the six house attributes and the remote sensing generated environmental variable (SOILIMP) revealed that there exist strong collinearities among these predictors. In particular, Table 3 indicates that FISIZE is highly positively correlated with FIREPLC, NOFBATH, and NOFST, while HSAGE is highly negatively correlated with AIRCD, and is negatively correlated with SOILIMP. The existence of such collinearities among the predictors might be the reason for the results in Table 1 and 2. Therefore, there is a need for better modeling technologies to address these problems.

TABLE 3. COLLINEARITIES AMONG THE FIVE HOUSE ATTRIBUTES AND THE REMOTE SENSING INFORMATION

	AIRCD	FLSIZE	FIREPLC	HSAGE	NOFBATH	NOFST	SOILIMP
AIRCD	1.000	−0.236	0.207	−0.767*	0.325	−0.182	0.418*
FLSIZE		1.000	0.754*	0.305	0.724*	0.806*	−0.182
FIREPLC			1.000	−0.180	0.825*	0.512*	0.120
HSAGE				1.000	−0.303	0.340*	−.637*
NOFBATH					1.000	0.514*	0.180
NOFST						1.000	−0.204
SOILIMP							1.000

*: indicates significant correlation at 0.01 level between the two crossed variables.

Regression Tree Analysis

The exact same sets of data were processed using the regression tree model previously outlined. The combined rule-based and instance-based algorithm was utilized with nine nearest-neighbors for constructing the model. The data were split into eight groups according to the rules defined by Cubist, and were illustrated in Table 4 in an ascending order according to the mean of house values. Moreover, Table 4 shows rule definitions (conditions), the coefficients of independent variables, and the relative importance of each independent variable in each group. In addition, the three accuracy statistics (MAE, RE, and correlation coefficient R) are reported as well.

Analysis of these results indicates that the regression tree model has very promising performance compared to the

TABLE 4. CART REGRESSION TREE MODEL

Dependent variable: House Price, log transformed
Use instances and rules, 9 nearest neighbors are used, no extrapolation is allowed
Residual standard error: 0.218 (training data), 0.178 (testing data)

	Rule I	Rule II	Rule III	Rule IV
Conditions	AIRCD <= 0.375 FLSIZE > 7.242 HSAGE > 4.584 NOFST <= 0.363	FIREPLC <= 0.280 HSAGE > 4.584 NOFST > 0.363 SOILIMP > 0.041	AIRCD <= 0.375 FLSIZE <= 7.242 FIREPLC <= 0.280 NOFST <= 0.363 SOILIMP > 0.041	AIRCD <= 0.375 FIREPLC <= 0.280 HSAGE <= 4.584 SOILIMP > 0.041
Mean	10.341	10.718	10.780	10.909
Coefficients:				
Intercept	10.045	10.180	9.866	10.036
AIRCD	1.580 (1)*	1.260 (1)	2.370 (1)	1.840 (1)
FLSIZE	—	0.180 (5)	0.180 (5)	0.140 (5)
FIREPLC	1.140 (2)	0.090 (6)	0.540 (2)	0.090 (6)
HSAGE	—	−0.170 (3)	−0.170 (4)	−0.130 (4)
NOFBATH	—	0.090 (7)	0.090 (7)	0.850 (2)
NOFST	0.220 (4)	0.270 (4)	0.170 (6)	—
SOILIMP	−1.900 (3)	−4.400 (2)	−3.500 (3)	−1.900 (3)
	Rule V	**Rule VI**	**Rule VII**	**Rule VIII**
Conditions	AIRCD <= 0.375 NOFST <= 0.353 SOILIMP <= 0.041	AIRCD > 0.375	FIREPLC <= 0.280 NOFST > 0.353 SOILIMP <= 0.041	AIRCD <= 0.375 FIREPLC > 0.280
Mean	11.056	11.580	11.674	12.063
Coefficients:				
Intercept	9.711	6.115	7.282	11.318
AIRCD	1.430 (1)	1.290 (1)	1.330 (1)	1.710 (1)
FLSIZE	0.190 (3)	0.520 (2)	0.600 (2)	—
FIREPLC	0.090 (6)	0.400 (4)	0.090 (5)	0.300 (4)
HSAGE	−0.090 (5)	0.250 (3)	−0.090 (4)	−0.100 (5)
NOFBATH	0.730 (2)	—	0.090 (6)	1.580 (2)
NOFST	—	—	—	0.100 (6)
SOILIMP	−1.600 (4)	−1.100 (5)	−1.600 (3)	−8.700 (3)
Accuracy statistics:				
Training data	*MAE*		0.160	
	RE		0.370	
	Correlation Coefficient *R*		0.920	
Testing data	*MAE*		0.134	
	RE		0.280	
	Correlation Coefficient *R*		0.960	

*Numbers in paranthesis indicate the rank of importance of the factors under the specific rules; and "—" indicates the factor is not important enough to be included in the model under that rule.

linear regression model. The residual standard errors of the regression tree model are smaller than those from the OLS regression models. All three accuracy statistics (*MAE*, *RE* and *R*) of the regression tree model on both training and testing data have salient improvement compared to the OLS counterpart (Table 4 and Table 2). These results reveal that the regression tree model is a better alternative in terms of predicting house values with house structural attributes and environmental characteristics.

Moreover, the regression tree model seems to reduce many of the adverse effects of collinearities among the predictors. The two predictors that were found holding opposite signs as hypothesized in the linear regression models, i.e., FISIZE and HSAGE (Table 2), seem to be holding the anticipated signs in the regression tree generated model. Specifically, except for groups I and VIII, FISIZE appears in all other six groups and projects positive influences on house values. For HSAGE, except for group VI, the hypothesized negative relationship was observed in all other rules that it appears. However, within group VI, the result implies that with more than one third of the houses within that census block group having air conditioners installed, house age might project positive influences on its average house values. A possible speculation on this relationship might be that older houses with air conditioner usually indicate that they are located in a relatively well-to-do neighborhood. Houses within such a neighborhood might have historical values that can add to house values. Such subtle categorical differentiation among relationships between the dependent variable and the predictors, is masked however in the OLS regression model.

In addition, being consistent with the OLS regression model, the regression tree model indicates that remote sensing-derived environmental characteristics play a significant role in modeling house values in the City of Milwaukee. In five of the eight rules generated by the regression tree algorithm, the remote sensing generated variable, SOILIMP, acts as a criterion to define the rules. In particular, the value 0.041 for SOILIMP serves as a standard for data splitting, with high-value houses grouped with low SOILIMP value and low-value houses grouped with high SOILIMP value. This result is quite intuitive because the lower fractions of soil and impervious surface indicate better environmental conditions, therefore promoting house values. Indeed, the hypothesized negative relationship between the SOILIMP factor and house values is supported under all rules. Moreover, among the seven predictors, SOILIMP ranked high (2 to 5) according to the importance in the regression tree model.

Spatial Patterns of the Analytical Results

Model Residuals in Space

In addition to the above statistical assessment of the models' performance, modeling accuracy on spatial data could be further assessed through GIS's visualization capability. From Tables 1, 2, and 4, it is noticed that the residual standard errors of the three models, i.e., the classical hedonic model, the remote sensing hedonic model and the regression tree model, keep decreasing for both training and testing data. While this fact provides solid evidence that the models are improving, it reveals little about where and how they improve. This concern is addressed through mapping the residuals of the three models. The residual maps of these three models are presented in Figure 3. The spatial distribution of each model's residuals reflects the model's improvement and where such improvement occurs as well.

Through a close inspection of Figure 3, two results stand out. First, although there are variations between the two linear models, a discernable spatial pattern of the their residuals emerges. That is, a distinct under- and over-prediction divide presents in the middle of the city (see Figure 3). The house values at the lakeside are highly under-predicted, while at its immediate western neighboring regions house values are over-predicted. Though spatially connected, these two regions are separated by the Milwaukee River with African-American population concentrated in the western regions (Yu and Wu, 2004). Such spatial patterns imply that housing markets in the City of Milwaukee can

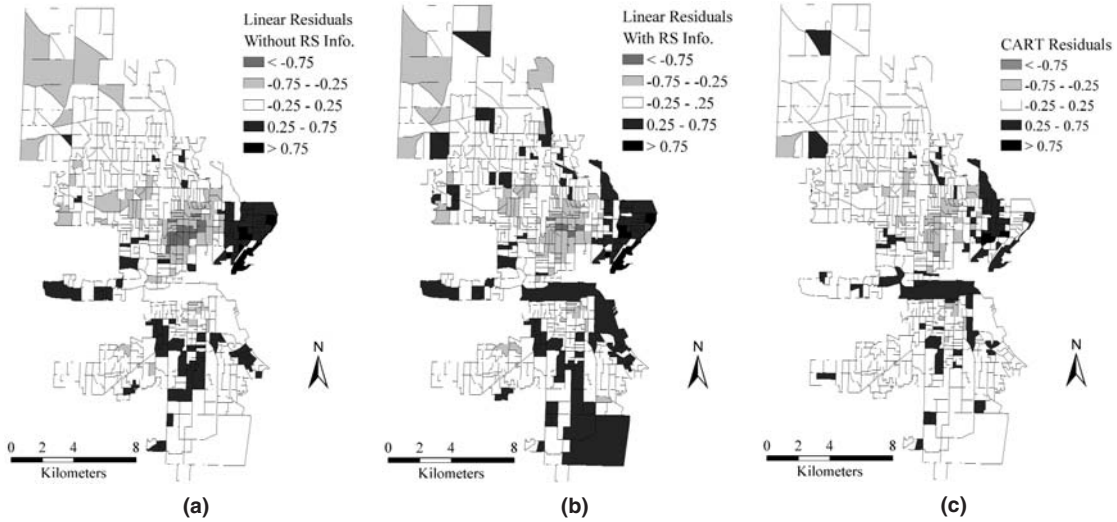

Figure 3. Spatial distribution of the model residuals: (a) Residuals from the classical hedonic model; (b) Residuals from the remote sensing information supported hedonic model; and (c) Residuals from the regression tree algorithm generated model (Color version available at the ASPRS Web Site, URL: *www.asprs.org*).

be stratified into submarkets instead of a uniformed one as suggested by the OLS models.

Second, the regression tree model provides a much better model residual surface. Comparing Figure 3a and 3b with 3c, it is quite salient that the variation of the residuals is smoothed. It is found that for about 80 percent of the census block groups (460 out of 571) the modeling residuals are within the range of −0.25 to 0.25 (representing a relatively accurate prediction), while in the linear models, only 67 percent of census block groups have model residuals within this range. Sparse over- and under-predictions using the regression tree algorithm still occur in the central part (over-prediction) and near the Lake Michigan (under-prediction), but in a much smoother way compared to its linear counterparts.

Spatial Structure of House Groups Split by Regression Tree Rules

To explore the spatial patterns of house groups that are split by the eight rules of the regression tree model, these house groups were mapped using ARCGIS®. Under the theory of the regression tree algorithm, the houses within a group are relatively homogeneous and can be modeled with a single rule. Figure 4 shows the spatial locations of house groups under regression tree rules. During the process, we find that group III and group IV have many overlapping cases. In this research, these two groups are merged to generate one single spatial group for better representation.

The spatial distribution of the rule-defined groups manifests interesting spatial patterns. There exists strong coincidence of data homogeneity and spatial homogeneity. Since the regression tree algorithm's categorization is entirely based on the house structural and environmental attributes, the coincidence of data homogeneity and spatial homogeneity provides solid evidence that in identifying housing submarkets in the City of Milwaukee, locational and structural attributes are indeed inter-connected instead of separated. That is, the implicit influences of house attributes and related environmental characteristics on house values vary along both structural and geographical lines. Moreover, the relative spatial consistence of the regression tree generated rules might present a means of delineating the boundaries of these submarkets in the City of Milwaukee.

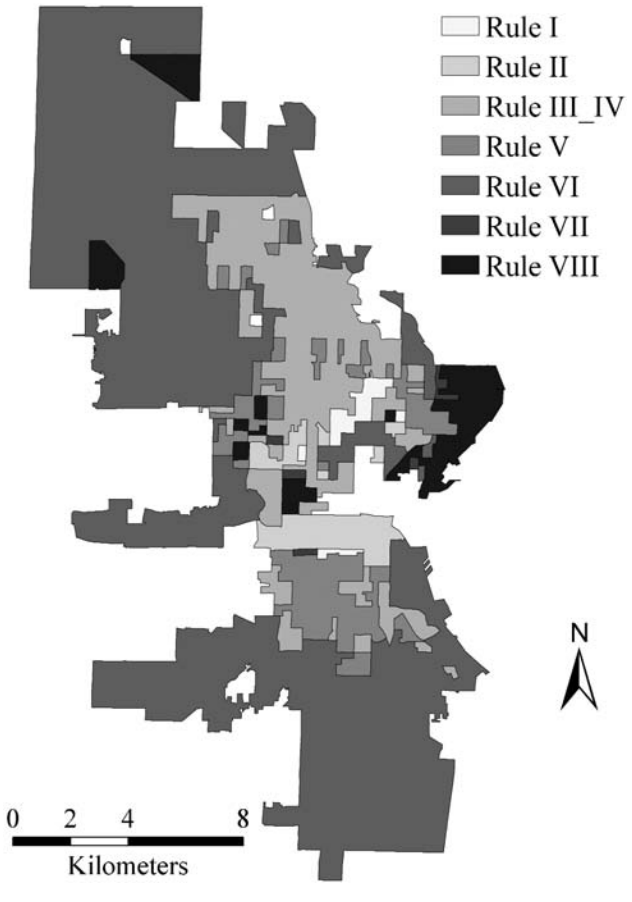

Figure 4. Spatial patterns of regression tree rules (I through VIII label the eight rules, with rule I indicating the house group with lowest values, and VIII indicating the house group with highest values) (Color version available at the ASPRS Web Site, URL: *www.asprs.org*).

Discussion and Conclusions

Aiming at incorporating remote sensing information in a house hedonic model and improving the model performance, this study constructed a remote sensing hedonic model and applied a regression tree approach utilizing data from Milwaukee. The results from the analyses were further visualized and analyzed spatially. Overall, three important conclusions can be drawn from the study.

1. At the aggregated census block group level, environmental characteristics generated from remote sensing imagery act as an important factor in modeling urban house values. In particular, the addition of the SOILIMP variable, representing the product of soil fraction and impervious surface fraction, increases the explaining power of the linear OLS model by around 4 percent. Moreover, in the regression tree model, SOILIMP also serves as an important factor in splitting house groups and modeling house values for each group. This research proves that environmental variables, which may be generated from remote sensing imagery, might serve as a valuable addition to the appraisal measurements that are regularly employed in hedonic studies. However, it is also noticed that the remote sensing imagery used in this study has a resolution of 30 meters. Although the findings are promising, the resolution might be coarse for capturing subtle urban morphological characteristics. In future studies, higher resolution imagery may provide better information in understanding urban housing markets.
2. Although the linear OLS regression technique is capable of capturing most of the essential relationships between house values and house structural attributes, it fails to capture subtle categorical differentiation among the data. Furthermore, collinearities among the predictors adversely affect the interpretation of some house value determinants, such as the floor size in this study. The regression tree approach, on the other hand, successfully recognizes the categorical differentiation among the data, and splits them accordingly. Beyond the improvement of modeling accuracy, the adverse effects of collinearities among the predictors are reduced to an acceptable level as well.
3. Although the study did not intentionally include a spatial factor in the model construction, the analyses from the regression tree approach give support to the hypothesis that there is a coincidence between data and locational homogeneity. Since the regression tree algorithm is capable of generating a series of house structural attributes-based rules, which can be deemed as attributes in defining housing submarkets, such coincidence provides some insights in delineating housing submarkets in the City of Milwaukee. More importantly, such coincidence does not seem to be present by chance. The inter-connection between house structural attributes and geographical areas in determining house values might be found in metropolitan areas other than the city of Milwaukee. More case studies on this subject need to be conducted to verify such assertion.

Acknowledgment

The authors wish to thank the anonymous reviewers and the editor, Dr. James W. Merchant, for their insightful comments and remarks.

References

Adair, A.S., J.N. Berry, and W.S. McGreal, 1996. Hedonic modeling, housing submarkets and residential valuation, *Journal of Property Research*, 13:67–83.

Anselin, L., 1988. *Spatial Econometrics: Methods and Models*, Kluwer Academic, Dordrecht, Netherlands.

Basu, A., and T.G. Thibodeau, 1998. Analysis of spatial autocorrelation in house prices, *Journal of Real Estate Finance and Economics*, 17(1):61–85.

Benson, E.D., J.L. Hansen, and A.L. Schwartz, Jr., 2000. Water views and residential property values, *The Appraisal Journal*, 68(3): 260–271.

Berry, B.J.L., and R.S. Bendnarz, 1975. A hedonic model of prices and assessments for single-family homes: Does the assessor follow the market or the market follow the assessor?, *Land Economics*, 51:21–40.

Breiman, L., J.H. Friedman, R.A. Olshen, and C.J. Stone, 1984. *Classification and Regression Trees*, Wadsworth, Inc., Monterey, California.

Carlson, T., 2003. Applications of remote sensing to urban problems, *Remote Sensing of Environment*, 86:273–274.

Cheshire P., and S. Sheppard, 1995. On the price of land and the value of amenities, *Economica*, 62:247–267.

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

Dublin, R.A., 1998. Predicting house prices using multiple listings data, *Journal of Real Estate Finance and Economics*, 17(1): 35–39.

Forster, B., 1983. Some urban measurements from Landsat data, *Photogrammetric Engineering & Remote Sensing*, 49(12): 1693–1707.

Goodman, A.C., 1978. Hedonic prices, price indices and housing markets, *Journal of Housing Research*, 3:25–42.

Harris, P.M., and S.J. Ventura, 1995. The integration of geographic data with remotely sensed imagery to improve classification in an Urban Area, *Photogrammetry Engineering & Remote Sensing*, 61(8):993–998.

Harvey, J.T., 2002a. Estimating census district populations from satellite imagery: some approaches and limitations, *International Journal of Remote Sensing*, 23(10):2071–2095.

Harvey, J.T., 2002b. Population estimation models based on individual TM pixels, *Photogrammetric Engineering & Remote Sensing*, 68(11):1181–1192.

Huang, C., and J.R.G. Townshend, 2003. A stepwise regression tree for nonlinear approximation: Applications to estimating subpixel land-cover, *International Journal of Remote Sensing*, 24(1):75–90.

Jackson, O.T., 2003. Methods and techniques for contaminated property valuation, *The Appraisal Journal*, 71(4):311–320.

Jackson, O.T., 2004. Surveys, market interviews, and environmental stigma, *The Appraisal Journal*, 72(4):300–310.

Kennickell, A., M. Starr-McCluer, and B.J. Surette, 1999. Recent changes in U.S. family finances: results from the 1998 Survey of Consumer Finances, *Federal Reserve Bulletin*, 86:1–29.

Kim, S., 2003. Long-term appreciation of owner-occupied single-family house prices in Milwaukee neighborhoods, *Urban Geography*, 24(3):212–231.

Lo, C.P., 1995. Automated population and dwelling unit estimation from high-resolution satellite images: a GIS approach, *International Journal of Remote Sensing*, 16(1):17–34.

Lo, C.P., 1997. Application of Landsat TM data for quality of life assessment in an urban environment, *Computers, Environments, and Urban Systems*, 21(3/4):259–276.

Lo, C.P., 2004. Testing urban theories using remote sensing, *GIScience and Remote Sensing*, 41(2):95–115.

Lo, C.P., D.A. Quattrochi, and J.C. Luvall, 1997. Application of high-resolution thermal infrared remote sensing and GIS to assess the urban heat island effect, *International Journal of Remote Sensing*, 18(2):287–304.

Maclennan, D., 1982, *Housing Economics: An Applied Approach*, Longman, London.

Major, C., and K.M. Lusht, 2004. Beach proximity and the distribution of property values in shore communities, *The Appraisal Journal*, 72(4):333–338.

Massey, D.S., and Denton, N.A., 1988. The dimensions of residential segregation, *Social Forces*, 67:281–315.

Mesev, V., 2003, *Remotely Sensed Cities*, Taylor & Francis, London.

Mulligan, G.F., R. Franklin, and A.X. Esparza, 2002. Housing prices in Tucson, Arizona, *Urban Geography*, 23(5):446–470.

Newman, P., and J. Kenworthy, 1999. *Sustainability and Cities: Overcoming Automobile Dependence*, Island Press, Washington, D.C.

Orford, S., 1999. *Valuing the Built Environment: GIS and House Price Analysis*, Ashgate Publishing Ltd, Hants, United Kingdom.

Orford, S., 2000. Modelling spatial structures in local housing market dynamics: a multilevel perspective, *Urban Studies*, 37(9):1643–1671.

Pace, K.R., R. Barry, and C.F. Sirmans, 1998. Spatial statistics and real estate, *Journal of Estate Finance and Economics*, 17(1): 5–13.

Phinn, S., M. Stanford, P. Scarth, A.T. Murray, and T. Shyy, 2002. Monitoring the composition and form of urban environments based on the vegetation – impervious surface – soil (VIS) model by sub-pixel analysis techniques, *International Journal of Remote Sensing*, 23:4131–4153.

Quinlan, J.R., 1993. Combining instance-based and model-based learning. *Proceedings of the 10th International Conference of Machine Learning* (P. Utgoff, edtor), Morgan Kaufmann Publishers, Amherst, Massachusetts, pp. 236–243.

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.

Ridd, M.K., 1995. Exploring a V-I-S (Vegetation-Impervious Surface-Soil) model for urban ecosystems analysis through remote sensing: Comparative anatomy for cities, *International Journal of Remote Sensing*, 16:2165–2186.

Ripley, B.D., 1996. *Pattern Recognition and Neural Networks*, Cambridge University Press, New York.

Roddewig, R.J., and A.C. Keiter, 2001. Mortgage lenders and the institutionalization and normalization of environmental risk analysis, *The Appraisal Journal*, 69(2):119–125.

Rothenberg, J., G.C. Glaster, R.V. Butler, and J. Pitkin, 1991. *The Maze of Urban Housing Markets: Theory, Evidence and Policy*, The University of Chicago Press, Chicago.

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

Small, C., 2002. Multitemporal analysis of urban reflectance, *Remote Sensing of Environment*, 81:427–442.

Steinberg, D., and P. Colla, 1995. CART: Tree- structured Nonparametric Data Analysis, Salford Systems, San Diego, California.

Straszheim, M.R., 1975. *An Econometric Analysis of the Urban Housing Market*, National Bureau of Economic Research, New York.

Sutton, P., C. Roberts, C. Elvidge, and H. Meij, 1997. A comparison of nighttime satellite imagery and population density for the continental united states, *Photogrammetric Engineering & Remote Sensing*, 63(11):1303–1313.

Thibodeau, T.G., 1989. Housing price indices from the 1974–1983 SMSA annual housing surveys, *AREUEA Journal*, 1:100–117.

Treitz, P.M., P.J. Howarth, and P. Gong, 1992, Application of satellite and GIS technologies for land-cover and land-use mapping at rural-urban fringe: a case study, *Photogrammetric Engineering & Remote Sensing*, 58(4):439–448.

United Nations, 1997. Urban and rural areas 1996, *ST/ESA/SER.A/166*, Sales No. E.97.XIII.3.

Weber, C., and J. Hirsch, 1992. Some urban measurements from SPOT data: Urban life quality indices, *International Journal of Remote Sensing*, 13(17):3251–3261.

Weeks, J.R., A. Getis, A.G. Hill, M.S. Gadalla, and T. Rashed, 2004. The fertility transition in Egypt: Intraurban patterns in Cairo, *Annals of the Association of American Geographers*, 94:74–93.

Weng, Q., D. Lu, and J. Schubring, 2004. Estimation of land surface temperature – vegetation abundance relationship for urban heat island studies, *Remote Sensing of Environment*, 89:467–483.

WisconsinView, 2004, URL: http://www.wisconsinview.org/, Madison, Wisconsin (last date accessed: 23 October 2005).

Wu, C., and A.T. Murray, 2003. Estimating impervious surface distribution by spectral mixture analysis, *Remote Sensing of Environment*, 84:493–505.

Wu, C., 2004. Normalized spectral mixture analysis for monitoring urban composition using ETM+ image, *Remote Sensing of Environment*, 93(4):480–492.

Yang, L., C, Huang, C.G. Homer, B.K. Wylie, and M.J. Coan, 2003. An approach for mapping large-area impervious surfaces: synergistic use of Landsat-7 ETM+ and high spatial resolution imagery, *Canadian Journal of Remote Sensing*, 29(2):230–240.

Yeh A., and X. Li, 2003. Simulation of development alternatives using neural networks, cellular automata, and GIS for urban planning, *Photogrammetric Engineering & Remote Sensing*, 69(9):1043–1052.

Yu, D., and C. Wu, 2004. Understanding population segregation from Landsat ETM+ imagery: A geographically weighted regression approach, *GIScience and Remote Sensing*, 41(3):187–206.

(Received 18 August 2004; accepted 14 December 2004; revised 08 February 2005)