Remote Sensing-based House Value Estimation Using an Optimized Regional Regression Model

Zhenyu Lu, Jungho Im, Lindi J. Quackenbush, and Sanglim Yoo

Abstract

This study proposed a new method to predict residential property value using remote sensing data as a major data source substitute to traditional inputs in house price estimation models. An optimized regional regression (ORR) approach was proposed in this study. This approach integrated a differential evolution optimization algorithm along with the ordinary least square regression to improve house value prediction accuracy. In addition to ORR, four other regression methods, random forest, Cubist regression trees, geographically weighted regression, and global ordinary least square, were also employed to provide a comparison. Results showed that models using remote sensing data are capable of acquiring accurate house price information. In addition, the volume of residential buildings proved to be an efficient substitute for total living area, the most important variable of the house price estimation model (i.e., a hedonic model). The ORR approach yielded the most accurate predictions followed by the geographically weighted regression. Further investigation indicated that the ORR approach has three major advantages: it is effective, stable, and the results are readily interpretable.

Introduction

With the fast increase of urban populations across the world, extracting socio-economic characteristics from timely remote sensing data is highly significant for planning and monitoring the urban environment. As an important component of the urban environment, and of private properties in particular, characteristics of residential houses have drawn wide public attention. The value of housing, as a composite and heterogeneous good (Cheshire and Sheppard, 1995), is determined by a variety of characteristics (e.g., structure, neighborhood, and environmental amenity). Detailed housing information, such as age, number of rooms, and neighborhood condition, has been widely used to assess house values. Obtaining accurate and timely housing value information is a necessary but challenging task.

Hedonic modeling has been commonly used in housing market analysis (Ismail, 2006; Yu *et al.*, 2007; Selim, 2009; Sunding and Swoboda, 2010). Such hedonic models typically predict house values using three sets of variables depicting structural, location/neighborhood, and environmental amenity characteristics of residential houses. Among the three groups of variables, structural variables, such as the type of house, total living area, lot size, number of rooms, and existence of central heating, have generally been more important than the other variables for hedonic modeling in the literature (Yoo *et al.*, 2012). Unlike the usually uniform and solid contribution of structural variables, the contribution of neighborhood and environmental variables to house values varies in different regions (Kong *et al.*, 2007; Yu and Wu, 2006).

A range of studies have reported accurate results when using hedonic methods to estimate the implicit price of residential property (Shultz and King, 2001; Dehring and Dunse, 2006; Selim, 2009). The modeling accuracy reported by these studies can be partly attributed to the availability of high-quality structural data, such as house construction and parcel data. However, these high-quality datasets are not globally accessible because of issues such as privacy and cost. Some of these high-quality data are not continuously updated, and might not be accurate enough to reflect the current status of the residential property. For example, the total living area of the property might not be updated when an expansion is done by the owner. In addition, housing properties are not generally ready for GIS analysis, which requires considerable time and effort to make them available in GIS form. Exploring timely generation of suitable substitutes for these datasets from remote sensing data is worthwhile.

Remote sensing devices are capable of capturing urban characteristics over a large geographic area in a timely manner, and have provided data that several studies have integrated with structural attributes to estimate housing values (Yu and Wu, 2006; Hamilton and Morgan, 2010). Compared to obtaining neighborhood and environmental variables, identifying structural characteristics of residential houses from remote sensing data is a much more challenging task. Some structural variables, such as number of rooms or existence of central heating, cannot be directly obtained from remote sensing data. However, remote sensing is capable of providing information, such as footprint area and volume of residential houses, which can serve as surrogates for structural variables such as total acreage of a residential property and total living area of a residential property. For example, Lu *et al.* (2011) used residential building volume and footprint area derived from lidar data to estimate population. They reported that their estimation model explained 80 percent of the residential

Zhenyu Lu is with AnchorQEA, LLC., Liverpool, NY 13088.

Jungho Im is with the School of Urban and Environmental Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan, South Korea (ersgis@unist.ac.kr).

Lindi J. Quackenbush is with the Department of Environmental Resources Engineering, State University of New York College of Environmental Science and Forestry, Syracuse, NY 13210.

Sanglim Yoo is with the Department of Environmental Science, State University of New York College of Environmental Science and Forestry, Syracuse, NY 13210.

Photogrammetric Engineering & Remote Sensing
Vol. 79, No. 9, September 2013, pp. 809–820.

0099-1112/13/7909–809/$3.00/0

population variance. Some detailed structural information has been acquired using advanced remote sensing data.

Traditional hedonic models are established using a global ordinary least square (GOLS) regression method, which applies a single model to the entire study area. The performance of the GOLS approach can be adversely impacted by multi-collinearity, spatial autocorrelation, and heteroscedasticity (Ismail, 2006; Selim, 2009; Yu and Wu, 2006). The adverse impact of spatial autocorrelation for the GOLS approach might be more significant in house price estimation than multi-collinearity or heteroscedasticity (Ismail, 2006). In fact, house prices exhibit high spatial autocorrelation for three reasons: (a) residential houses in a neighborhood often have similar structural characteristics, either because they were constructed by the same company or because residents have similar social status and income; (b) neighborhood conditions, such as safety and traffic issues are similar; and (c) houses within a neighborhood usually share the same environmental amenities, such as recreation places, parks, and water bodies.

There are two major types of approaches employed to reduce the adverse impacts of multi-collinearity, spatial autocorrelation, and heteroscedasticity and improve the performance of traditional hedonic models. The first approach is to employ machine-learning based modeling techniques, such as regression trees, random forest, and neural networks. These machine-learning approaches have proved reliable with higher predication accuracy compared to traditional hedonic modeling. Yu and Wu (2006) employed a regression tree approach in modeling house values. They found the regression tree approach improved prediction accuracy by reducing the adverse effects of collinearity among the independent variables. Selim (2009) mitigated potential non-linearity in the hedonic models by using an artificial neural network (ANN) approach. Though the machine-learning approaches are effective in improving the performance of traditional hedonic models, they might have limited ability to solve the spatial autocorrelation and heteroscedasticity issues since they do not take spatial and locational information into account.

The second type of approach is to employ a submarket (market segmentation) technique, which is especially efficient in reducing adverse impacts of spatial autocorrelation and heterogeneity (Goodman, 1998; Goodman and Thibodeau, 2003; Yu *et al*, 2007; Wu and Sharma, 2012). A housing submarket is defined as a set of dwellings that are reasonably similar with each other but are different from dwellings in other submarkets. From a methodological perspective, the submarket approach provides an expansion method that allows aspatial models to encompass spatial contexts (Casetti, 1997). There are two methods used to define submarkets: *a priori* definition and data-driven methods. *A priori* classifications are based on existing spatial divisions or predefined criteria, such as aggregated census blocks, zip codes, and local government boundaries. Data-driven methods classify submarkets through techniques such as factor analysis, partitioning algorithms, and hierarchical clustering.

Geographically weighted regression (GWR) is a popular expansion model for spatial data analysis and has been employed in hedonic applications (Wu and Sharma, 2012). Many of the applications revealed that GWR tends to perform better than traditional hedonic models based on GOLS. However, while promising results have been reported, GWR has two major limitations (Yu *et al*., 2007): (a) the interior interpolation procedure inevitably introduces new errors; and (b) overfitting may occur since GWR tunes toward the best fit of the calibration samples.

In this study, a differential evolution (DE) optimization approach was applied to divide the study area into relatively homogeneous submarkets. Theoretically similar to housing submarkets where the price per unit of housing quantity is constant (Goodman, 1998), the submarket in this study represents a geographic area in which the house transaction price is similar and can be accurately expressed through a local prediction model. DE was proposed by Storn and Price (1997) and yielded comparable performance to genetic algorithms (GA), but outperformed GA in computation time and efficiency. Though DE has been widely investigated in other fields, there have been very limited efforts applying DE in remote sensing. The objectives of this study were to: (a) employ machine-learning based optimization methods (DE and GA) to delineate submarkets with homogeneity in terms of house transaction price; (b) propose an optimized regional regression (ORR) model for house price estimation using remote sensing data as a major data source; and (c) evaluate the ORR model in comparison with GOLS, GWR, and two machine learning approaches, i.e., Cubist and random forest.

Study Site and Data

Study Site

The study area is a part of the Syracuse Metropolitan Statistical Area (SMSA), which includes the City of Syracuse, the villages of East Syracuse and Solvay, and the towns of Onondaga, Dewitt, Salina, and Geddes in Onondaga County in central New York State (NYS). The SMSA is the fifth most populous region in NYS. According to the 2010 US Census (US Census Bureau, 2010), there were 742,603 residents in the SMSA, with approximately 20 percent of them in the City of Syracuse. In the past 10 years, the population in Syracuse slightly decreased (by 2,176), while Onondaga County had minor growth (by 8,690). Similar to most northeastern US cities, the growth pattern of Syracuse shows a trend of "sprawl without growth" where population density decreased with the sprawl of urban areas.

There were a total of 657 housing transactions in 2010 within the study site. The transaction price of these residential properties ranged from $10,000 to $810,000 USD with a median price of $111,500. Transaction prices of residential properties show a noticeable spatial pattern in that houses located in the southeast were sold at a relatively higher price (Plate 1).

Data

Four types of data were used in this study: residential property transaction records, airborne lidar, land-use and land-cover (LULC) raster data, and several publicly accessible GIS datasets such as census unit boundary, locations of recreation places, and highways. Residential property transactions were summarized from arms-length transaction records of Onondaga County single family residences sold in 2010. In this study, actual sales price was used to calibrate and validate hedonic models instead of assessed value because sales price better represents buyer preference (Yoo *et al*., 2012). Location of each transacted house was identified as a point using a geocoding approach in ArcGIS®.

Lidar data with a point density of 2.8 points/m^2 was collected by Kucera International, Inc. in April 2010. Two lidar-derived raster surfaces (first return and bare earth) were generated using a triangulated irregular network (TIN) approach. To be consistent with the LULC data, the lidar-derived raster surfaces were generated with a 1-foot pixel size. A first return height (FRH) surface was created using the first return surface minus the bare earth surface.

The LULC dataset with a 1-foot pixel size for the greater Syracuse area was created by the University of Vermont Spatial Analysis Laboratory using the 2010 lidar data and 2009 NYS 4-band, leaf-off orthophotography along with

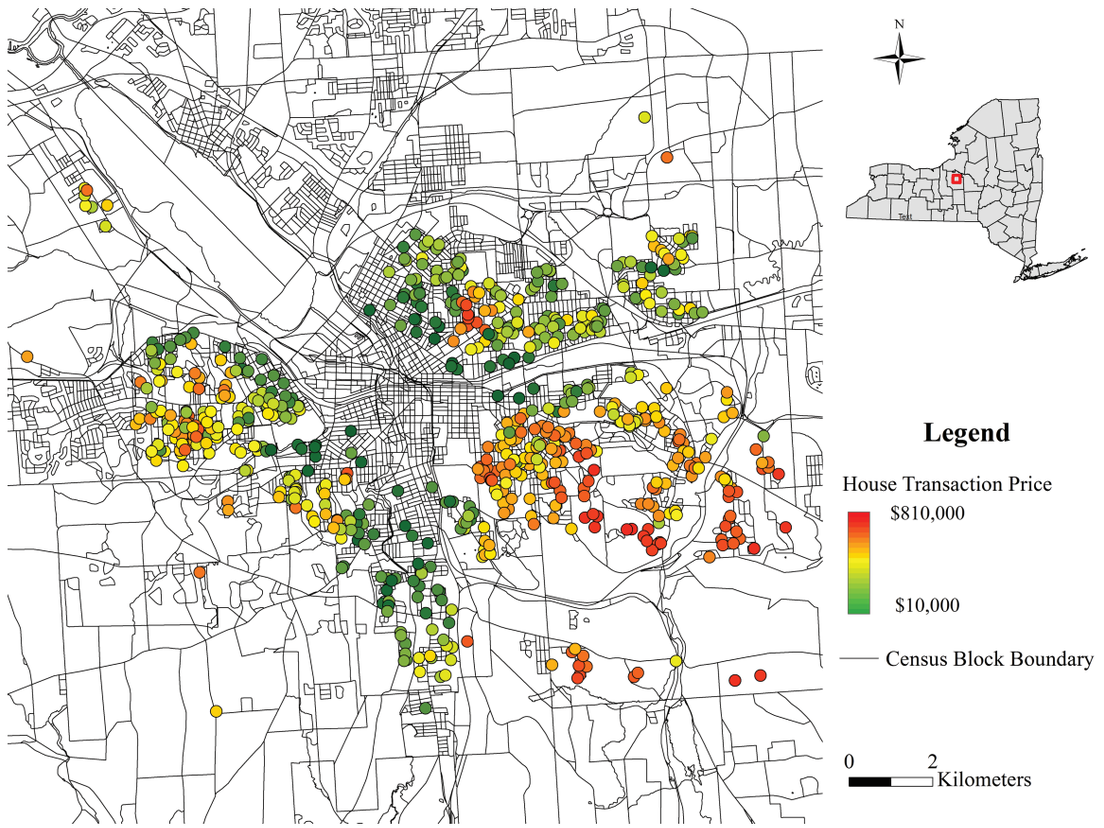


Plate 1. Residential house transaction price of the study area in 2010.

several auxiliary GIS datasets (roads, parcels, and hydrography). Seven LULC classes (tree canopy, grass/shrub, bare earth, water, buildings, roads, and other paved surfaces) were mapped using object-based image analysis. Although no formal accuracy assessment was conducted for this dataset, it underwent thorough manual quality control, with 38,855 manual corrections made to the classification.

Environmental amenities, such as location and boundaries of recreation places, are difficult to obtain from remote sensing data. To obtain environmental amenity and neighborhood information, several publically accessible GIS datasets were included in this study to supplement the remote sensing-derived information. The 2000 NYS census block polygons, and statewide highway data were downloaded from the NYS GIS Clearinghouse (*http://gis.ny.gov/gisdata/*). Datasets containing boundaries of State, County, and municipal recreation areas in NYS were also downloaded.

Table 1 describes the dependent and independent variables employed in this study. There were 16 independent variables including two structural variables, six neighborhood variables, and eight environmental variables. Compared to other studies (Yu and Wu, 2006; Selim, 2009; Yoo *et al.*, 2012), this study employed fewer structural variables, since it is not feasible to acquire information such as number of bathrooms or existence of central heating from remote sensing data. Instead, we employed surrogate information, i.e., footprint area and total volume of residential houses. The environmental variables included area ratios of four LULC classes, i.e., tree crown, grass, road, and other paved surfaces. The building class was excluded since it was not defined as an environmental variable in hedonic modeling and was already used to create structural variables. Water and bare earth were also excluded because very few neighborhoods contain these two LULC classes, which would limit benefit prediction. Previous studies have investigated using political boundaries, such as census block group or census tract (Shultz and King, 2001), local neighborhood association boundaries (Poudyal *et al.*, 2009), and fixed-radius circular buffers (Geoghegan, 2002; Yoo *et al.*, 2012) to capture environmental amenity characteristics around a residential property. However, there is no common definition of neighborhood for acquiring environmental characteristics of residential properties. This study used three neighborhood boundaries: one defined by political boundaries and two based on circular buffers. The political boundary used was the census block that contained the transacted property, and the circular buffers used radii of 100 m and 1 km around each residential property. The 100 m buffer is expected to describe the visual zone around a house while 1 km represents a typical 10 to 20 minute walking distance (Yoo *et al.*, 2012).

We calculated various environmental variables and spatial metrics based on the three kinds of neighborhood definition. The percentage of tree crown, grass/shrub, road, and other paved surface within the neighborhood was calculated. Area percentage of the LULC types in the neighborhood was used instead of area, which was widely employed in previous studies, to normalize the size of the neighborhood, enabling comparison of environmental variables generated from different sized neighborhoods.

Positive economic relationships were recently detected between house sale price and various types of open space,

Table 1. Summary of Variables Used in this Study

Variable		Definition	Source	Unit
Dependent Variable	Sales Price	Residential property sales price in $	Post Standard, 2010	$
Structural Variables	AREA	Total footprint area of residential house	LULC	ft^2
	Volume	Total volume of residential house	LULC & LiDAR	ft^3
Neighborhood Variables	DIST_HW	Euclidean distance to nearest interstate highway	Interstate Highway Polyline	km
	DIST_RD	Euclidean distance to nearest road	LULC	m
	DIST_WATER	Euclidean distance to nearest water body	LULC	km
	DIST_RC1	Euclidean distance to nearest state-level recreation place	Recreation polygon	km
	DIST_RC2	Euclidean distance to nearest county-level recreation place	Recreation polygon	km
	DIST_RC3	Euclidean distance to nearest municipal-level recreation place	Recreation polygon	km
Environmental Variables	LC_TR	Ratio of tree crown LULC category in neighborhood (i.e., census block, 100m or 1km radius buffer)	LULC & Census block boundary	%
	LC_GS	Ratio of grass/shrub LULC category in neighborhood (i.e., census block, 100m or 1 km radius buffer)	LULC & Census block boundary	%
	LC_IMP	Ratio of other paved surface LULC category in neighborhood (i.e., census block, 100m or 1km radius buffer)	LULC & Census block boundary	%
	LC_RD	Ratio of road LULC category in neighborhood (i.e., census block, 100m or 1km radius buffer)	LULC & Census block boundary	%
	NP	Number of patches index in neighborhood (i.e., census block, 100m or 1km buffer)	LULC & Census block boundary	#
	FRAC_MN	Mean fractal dimension index in neighborhood (i.e., census block, 100m or 1km buffer)	LULC & Census block boundary	None
	CONTAG	Contagion index in neighborhood (i.e., census block, 100m or 1km buffer)	LULC & Census block boundary	%
	SHDI	Shannon's diversity index in neighborhood (i.e., census block, 100m or 1km buffer)	LULC & Census block boundary	#

such as urban parks (Anderson and West, 2006), land in conservation easement (Irwin, 2002), agricultural cropland (Geoghegan, 2002), forests (Smith *et al.*, 2002), and golf courses (Smith *et al.*, 2002). Four spatial metrics commonly adopted in the literature (DiBari, 2007; Yoo *et al.*, 2012): number of patches (NP), Shannon's Diversity Index (SHDI), Fractal Dimension Index (FRAC_MN), and Contagion Index (CONTAG) were employed to measure the amenity value of open space patches. The four metrics were calculated using FRAGSTATS 3.3. To avoid duplicate information and collinearity of independent variables, model calibration did not include environmental amenity variables calculated using all three neighborhood definitions. Only the environmental variables calculated based on the neighborhood boundary that contributed most to the house transaction price were included in subsequent analysis.

Methodology

Figure 1 illustrates the workflow of this study. Six LULC classes (buildings, grasses, trees, water, roads, and other paved surfaces) were extracted from the LULC dataset. Structural characteristics of residential houses, i.e., footprint area and total volume, were subsequently calculated using the lidar-derived height surface and the building footprints delineated in the LULC dataset. Neighborhood and environmental variables were calculated using the location of the houses identified in the residential house transaction records, the LULC raster, and the auxiliary GIS data. Once all variables were calculated, the transaction records were randomly divided into two equally sized groups for model calibration and validation. The proposed ORR approach and four other regression-based approaches (GOLS, GWR, Cubist regression trees (RuleQuest Research, Inc.), and random forest (RF))were trained using the calibration data, and model performance was evaluated using the validation dataset.

Submarket Optimization

The DE and GA approaches were compared in terms of optimizing the division of the study area into homogenous submarkets. The DE optimization algorithm is a descendent of GA, which is an efficient and widely-used optimization method in remote sensing (Gleason and Im, 2012a; Im *et al.*, 2011; Tseng *et al.*, 2008). A major advantage of GA over other optimization methods, such as hill climbing and simulated annealing, is GA processes a population of solutions in a parallel manner (Tseng *et al.*, 2008), which results in higher efficiency in finding optimum solutions. In addition, GA employs several evolution operators (selection, crossover, and mutation) to help a population of solutions converge to a global optimum. DE is similar to GA in terms of initializing and processing a population of solutions in a parallel manner and employing several evolution operations. However, DE is simpler in

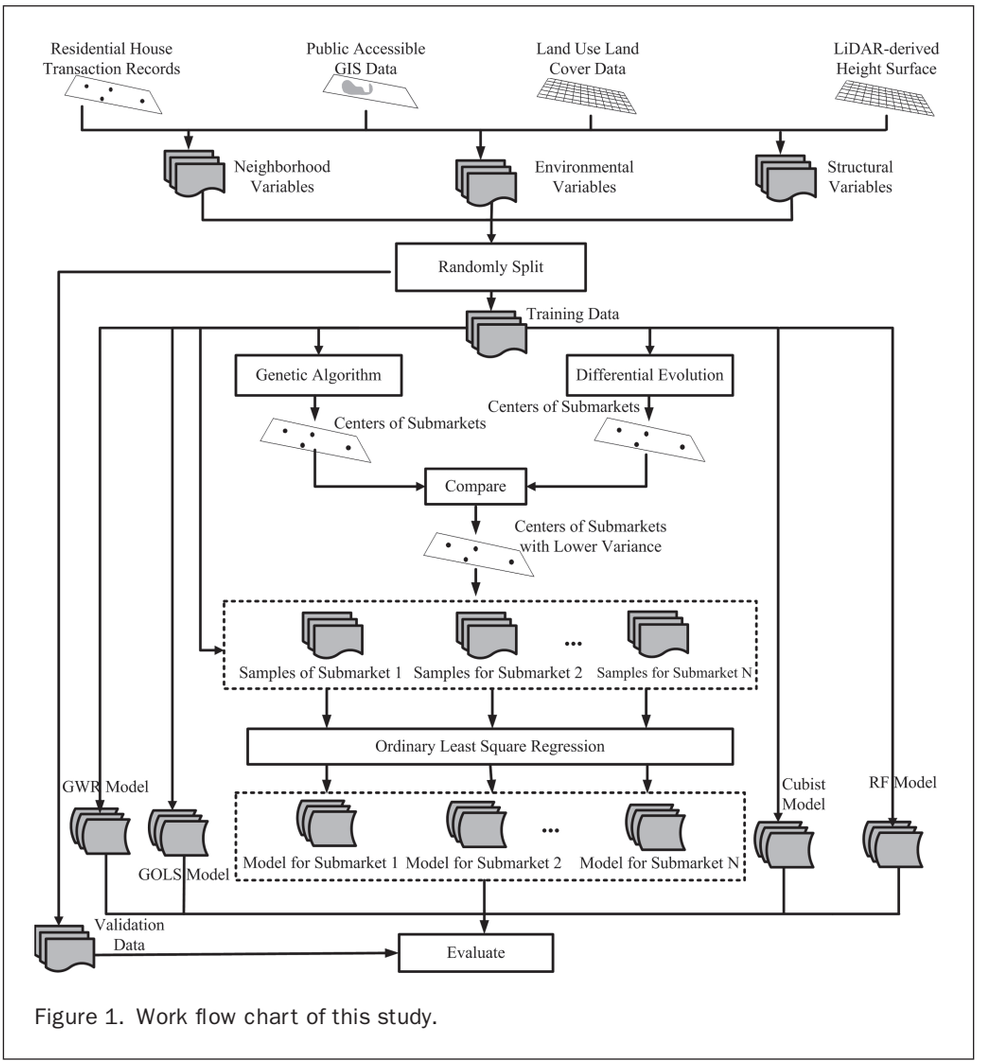


Figure 1. Work flow chart of this study.

implementation and has fewer parameters to define. Despite its simple form, DE has proved to yield comparable accuracy as GA.

The "chromosome," an abstraction of a solution to the specific problem, is the most important concept in both GA and DE. A chromosome consists of an array of genes, and each gene works as an equally important part of the solution. In this study, the chromosome is the same for both GA and DE, and represents coordinates of submarket centers. Each gene of a chromosome is a coordinate value (e.g., latitude or longitude) of a certain center. With the study area divided into N submarkets, each chromosome contained $2 \times N$ genes. The performance of a chromosome is evaluated using a fitness value: the higher the chromosome fitness, the better the solution. In this study, fitness measured the homogeneity of the submarkets and was calculated using one minus the total weighted variance of house sale prices of all submarkets divided by the variance of samples of the whole study area (see below):

$$\text{Fitness} = \frac{\mathrm{R}_n - \sum_{i=1}^{k} \frac{F_{ni}}{\mathrm{F}_n} \times \mathrm{R}_{ni}}{\mathrm{R}_n} = 1 - \frac{\sum_{i=1}^{k} \frac{F_{ni}}{\mathrm{F}_n} \times \mathrm{R}_{ni}}{\mathrm{R}_n} \qquad (1)$$

where R_n is the variance of all samples, R_{ni} is the variance of samples in the i^{th} submarket. F_n represents the total number of samples, and F_{ni} is the number of samples in the i^{th} submarket. The sample variance (R_n was calculated using $\Sigma(Y - \bar{Y})^2$ where $\bar{Y}$ is the mean of the dependent variable (sale price). The fitness value is within the range of 0 to 1; higher fitness values indicate more homogeneous submarkets in terms of sale price.

The DE method includes four key operations (initialization, crossover, mutation, and selection) to search for an optimum solution. Initialization involves generating a population of chromosomes by randomly distributing the centers of submarkets within the study area. A chromosome was added to the initialized vector of chromosomes only if the number of samples within each defined submarket exceeded a certain number. In this study, a valid submarket was tentatively defined to contain at least ten calibration samples to make sure the local model was representative. After each chromosome was evaluated using the fitness value, a mutation operation was employed to generate a mutated vector. The mutated vector is also a population of chromosomes with the same number of chromosomes as the initialized chromosome population. The gene values of each chromosome were calculated based on one of five DE mutation mechanism functions (Table 2).

TABLE 2. THE MUTATION MECHANISMS OF DE

Mutation mechanism function	Mathematical equation
Best_1	$v_{i,G} = x_{best,G} + F(x_{r1,G} - x_{r2,G})$
Rand_1	$v_{i,G} = x_{r3,G} + F(x_{r1,G} - x_{r2,G})$
Rand_to_Best_1	$v_{i,G} = x_{i,G} + F(x_{r1,G} - x_{r2,G})$
Best_2	$v_{i,G} = x_{best,G} + F(x_{r1,G} + x_{r2,G} - x_{r3,G} - x_{r4,G})$
Rand_2	$v_{i,G} = x_{r5,G} + F(x_{r1,G} + x_{r2,G} - x_{r3,G} - x_{r4,G})$

Crossover creates a trial vector by choosing some parts from a mutation vector and other parts from the initial chromosomes. A random number between 0 and 1 was generated for each gene of a chromosome. If the random number is larger than the crossover rate, the mutation vector gene is selected for the trial vector. Otherwise, the gene of the initial chromosome is selected. After mutation and crossover, the range of solutions is diversified compared to the initial chromosomes. Selection was finally carried out to identify the input chromosomes (target vector) for the next generation. The selection operation compared the fitness of individual chromosome from the vector of initialized chromosomes with the trial vector. The chromosome with higher fitness is selected as input for the next generation. When the increasing generation equals the maximum number of optimization generations, the chromosome with the highest fitness value was selected as the best solution for submarket division. Details about DE are found in Price (1999) and Storn and Price (1997).

Modeling Approaches for Residential House Value Estimation

The five regression approaches used in this study are based on the general hedonic price function for housing, which has the following form (Freeman, 2003):

$$P(A) = f(S, N, E) \tag{2}$$

where P represents residential property transaction price, A is the attributes or characteristics of the house, and S, N, and E form a vector of variables depicting the structural, neighborhood, and environmental amenity, respectively, of residential houses. The GOLS model was calibrated using a linear ordinary least square (OLS) approach. Variables with positive coefficients increase house values while variables with negative coefficients tend to decrease house values.

The GWR technique is a well-adopted spatial analysis method developed from the idea of Cassetti's (Casetti, 1997) expansion regression method. It is effective in dealing with spatial heterogeneity and autocorrelation. The GWR computations were carried out in ArcGIS® 10.0, using adaptive kernels to calibrate the GWR model. Although adaptive kernels are more computationally intensive than fixed kernels, they produce smaller local estimation variance in areas where data are sparse. The Akaike information criterion (AIC; Akaike, 1974) was employed to obtain the optimal size of nearest neighbors for the adaptive kernel. The study also included two machine learning regression methods, Cubist regression trees and random forest (RF). Compared to GOLS, these machine learning approaches have the advantage of allowing nonlinearities and interactions among independent variables. Cubist is a commercial software package that implements a hybrid tree-based approach combining a regression tree algorithm with local multivariate regression modeling. Cubist categorizes input data into relatively homogeneous groups to reduce prediction bias. Several studies have reported that Cubist outperformed OLS by significantly increasing the estimation accuracy (Im *et al.*, 2009; Lu *et al.*, 2010; Yoo *et al.*, 2012; Yu and Wu, 2006). Cubist version 2.04a was used in this study.

RF is a classification and regression technique introduced by Breiman (2001), which randomizes an ensemble of decision trees to improve prediction accuracy by combining random expert trees through voting or averaging. The RF approach has been applied in remote sensing studies for solving mapping, classification, and regression problems. Several studies have found that the RF approach yielded comparable or better results than the general regression tree approach (Gleason and Im, 2012b; Yoo *et al.*, 2012). This study used the RF add-on package in MATLAB software, which requires specification of two parameters: the number of trees in the forest (*ntree*) and the number of variables randomly sampled at each node (*mtry*) (Breiman and Cutler, 2004). We used 1,000 for *ntree* and the default *mtry* value, which is the square root of the number of independent variables.

Optimized Regional Regression

The optimized regional regression (ORR) approach proposed in this study is theoretically an expansion method. The basic assumption of the ORR model is similar to Cubist, which tries to subset input data into relatively homogenous groups to reduce prediction bias. However, the ORR model spatially divides input data into homogeneous subsets according to the delineated boundary of submarkets instead of thresholds from a subset of attributes. The ORR model might result in similar estimations if the divided submarkets overlap with the coverage of samples defined by the RT (regression tree) rulesets.

A key component of the ORR model is meaningfully dividing the whole study area into several submarkets. We compared DE and GA approaches to find submarkets centers and selected the method that generated the higher optimization accuracy. After determining the internal optimization method of the ORR model (i.e., DE or GA), the only parameter to specify is the number of submarkets, which depends on the characteristics of the study area and the number of available samples. Theoretically, an ideal division of submarkets using the ORR model implies: (a) each submarket is relatively homogeneous in terms of dependent variable values; and (b) submarkets are different from each other. A small number of submarkets might not reflect the spatial distribution of all samples, which results in underfitting; a large number of submarkets might include too many local patterns, which increases the risk of overfitting. In this study, we tested different numbers of submarkets and selected one based on the performance patterns and site characteristics.

Once each submarket was defined, OLS regression models were calibrated for each submarket using training samples within the submarket. Similarly, validation samples within a certain submarket were predicted using the local OLS model of this submarket. The ORR regression approach was implemented using the C# programming language.

Accuracy Assessment

Three performance metrics, i.e., adjusted coefficient of determination (adjusted R^2), root mean squared error (RMSE), and

relative RMSE (RRMSE) were used to evaluate the regression models. Adjusted R^2 is a modification of R^2 that considers the number of explanatory terms in a regression model. Unlike R^2 which usually increases with increasing numbers of independent variables, the adjusted R^2 increases only if a new variable improves the model more than would be expected by chance.

Both RMSE and RRMSE are broadly used accuracy assessment metrics. RMSE describes the accuracy of a model prediction against observed values. RRMSE is the standardized RMSE, which in this case represents the ratio of RMSE to the mean of actual house transaction prices.

Results and Discussion

Submarket Partitioning

Figure 2 summarizes the performance of DE and GA in terms of computation time and fitness value with the increasing number of submarkets. The parameters used for the DE and GA algorithms are presented in Table 3. The comparison between the two optimization algorithms revealed that DE is more efficient than GA, yielding slightly higher accuracies with less computation time. With increasing numbers of submarkets, both the fitness value of the best solution and the corresponding computation time increased regardless of the algorithm used. The computation times for DE and GA were similar when the number of submarkets was small (≤5), but GA required much more time compared to DE when the number of submarkets was large (>5). DE slightly outperformed GA in terms of the fitness value regardless of the number of submarkets.

The number of submarkets had a more significant impact on the fitness value of both DE and GA approaches where the number of submarkets is small (≤5), with minimal improvement in fitness value after that. The total weighted variance of the four submarkets was around 10 percent of the total variance of all samples. As the number of submarkets increased from four to ten, the additional increase in fitness value was less than 0.1 (0.89 to 0.97). Figure 2 illustrates that using additional submarkets never led to lower fitness values. Accordingly, the fitness value should not be the only criteria for determining the number of submarkets, since this would always select a large number of submarkets, which may incorporate too many local patterns leading to overfitting. As mentioned in the methodology section, some *a priori* knowledge or investigation on the spatial pattern of the housing transaction

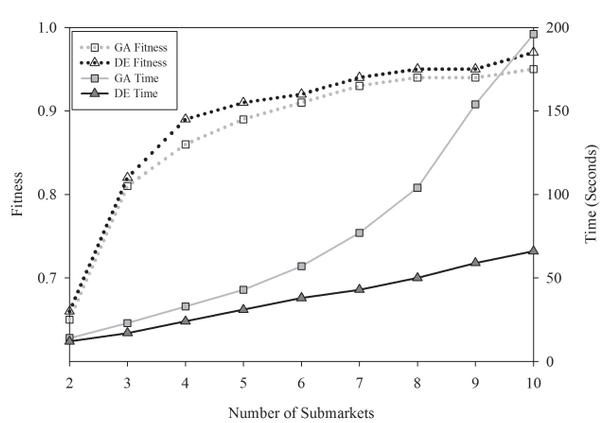


Figure 2. Comparison of accuracy and computation time of DE and GA.

Table 3. Parameter Setup of DE and GA

Parameter Name	GA	Parameter Name	DE
Generation	500	Generation	500
Number of chromosome	10	Number of chromosome	10
Selection ratio	0.3	Weight Factor	0.5
Crossover rate	0.8	Crossover rate	0.8
Mutation rate	0.2	Mutation function	Best_2

records and site characteristics are also useful in determining the number of submarkets. In this study, we used four submarkets because the increase in the number of submarkets from three to four resulted in the last jump in fitness values (from 0.82 to 0.89). Additional submarkets resulted in only a minor increase of fitness value (less than 2 percent). In addition to the fitness value trends, the transaction prices showed different patterns for the northeast, southeast, northwest, and southwest, which also suggested that four submarkets represented the patterns of this study area.

Residential House Price Modeling

Table 4 summarizes the accuracy metrics (adjusted R^2, RMSE, and RRMSE of the five regression methods, considering the three different neighborhood groups of the environmental variables. It is notable that no single group of environmental variables resulted in the highest estimation accuracy across the five regression methods. In fact, while there were small differences, the estimation results of the GOLS, GWR, RF, and ORR approaches were not significantly different when considering the three groups of environmental variables. This was because none of the environmental variables were important contributors to these three regression models. However, the Cubist method using the 100 m radius environmental variables resulted in much higher estimation accuracy than results using the other two groups of environmental variables. The reason for that is that Cubist yielded useful rulesets by splitting the number of patches in the 100 m environmental variables, which increased prediction accuracy. However, similar patterns were not found using the environmental variables determined within the census block or the 1 km neighborhood. Given that the 100 m radius buffer environmental variables resulted in the significantly better estimation for the Cubist method, and also that the other regression methods were not sensitive to the neighborhood considered, the following comparisons of the different regression methods used environmental variables calculated within the 100 m radius buffer.

Both the two machine learning algorithms (Cubist and RF) and two expansion methods (GWR and ORR) outperformed the statistical regression method (GOLS) in fitting the calibration data patterns. Of the five methods considered, RF resulted in the best fit of the calibration data (adjusted R^2 value of around 0.95) followed by ORR, GWR, and Cubist. The GOLS model performed poorly in fitting the calibration data with an adjusted R^2 value of around 0.67. However, the performance of the five regression models was quite different when they were evaluated using the validation data. The ORR model yielded the best results, explaining 77 percent of the variance of house transaction price. The GWR model yielded the second best result with an adjusted R^2 of 0.75. It is notable that methods based on expansion concepts (ORR and GWR) outperformed both the statistical and machine-learning regression algorithms in this study, which implies that including locational information is valuable in accurately estimating house values. Surprisingly,

Table 4. Comparison of the House Value Estimation Results Using the Environmental Variables Calculated Based on Three Boundary Schemes; Values in Bold Indicate that the Corresponding Model Resulted in the Highest Validation Accuracy when One of the Three Groups of Environmental Variables was Used

	Census Block			100m Radius			1000m Radius		
	Adj R^2	RMSE ($10,000)	RRMSE (%)	Adj R^2	RMSE ($10,000)	RRMSE (%)	Adj R^2	RMSE ($10,000)	RRMSE (%)
GOLS	**0.68/0.70***	**4.2/4.4**	**33.1/33.9**	0.67/0.67	4.3/4.5	33.7/34.5	0.68/0.67	4.2/4.5	32.7/34.7
Cubist	0.78/0.62	3.5/4.9	27.3/38.1	0.78/**0.71**	**3.4/4.3**	**27.1/33.0**	**0.80**/0.60	3.5/4.9	27.3/37.9
RF	0.95/0.68	1.9/4.6	15.1/35.5	0.95/0.66	1.9/4.6	15.1/35.7	0.96/0.70	1.8/4.4	13.9/34.4
GWR	0.70/0.71	4.0/4.3	31.6/32.9	**0.70**/**0.75**	**3.9**/**4.0**	**30.9**/**30.9**	0.66/0.72	4.3/4.2	34.2/32.8
ORR	0.80/0.74	3.4/4.2	26.5/32.8	0.79/**0.77**	3.4/**3.8**	26.8/**29.2**	**0.81**/0.75	**3.2**/4.0	**25.2**/30.8

* calibration/validation

the RF model made the poorest prediction, with more than a 20 percent increase in RRMSE value compared to calibration accuracy. The substantial decrease of validation accuracy indicated a significant overfitting issue in the RF model. It is interesting that the performance of the RF method in this study was contradictory to previous studies (Breiman, 2001) where RF mitigated overfitting. This inconsistent performance is probably due to the differences between datasets. If outliers comprise only a small percentage of samples, the majority of rulesets would use representative samples, thus overfitting caused by outlier samples would be negligible. However, if the samples are quite variable, the rulesets could enlarge the impact of outliers since RF requires all trees to grow to full size. In fact, the RF model has been criticized for overfitting with noisy classification and regression tasks (Segal, 2003). Segal (2003) found that maximally sized trees overfit, and suggested regulating tree size by limiting the number of splits or node size. Overfitting was less significant for the other regression models, which showed a small decrease in validation accuracy compared to calibration accuracy (around 5 percent decrease of RRMSE values). It is interesting that the validation accuracy is equal to, or even slightly higher than, the calibration accuracy for the GOLS model. The possible reason for this is that the GOLS method calibrates the regression model based on the major trend of the calibration data, which was usually consistent with the interior pattern of the validation data.

Table 5 shows the coefficients of the explanatory variables derived from the GOLS regression methods. Variables, such as volume, DIST_HW, DIST_RD, DIST_WATER, DIST_RC2, and DIST_RC3, were significant contributors to the house transaction value estimation. It is notable that DIST_RC1 was negatively related to the house transaction value in this study. There are two possible reasons for this: (a) the contribution of state level recreation places was offset by other types of recreation places, such as county level or municipal level recreation places; and/or (b) the state level recreation places are all far away from most of the houses used, which actually cast limited impacts on the transaction price. The contribution of environmental variables is not as significant as the structural and neighborhood variables.

Figure 3 shows the validation result of the five regression models. ORR, GWR, and Cubist made relatively accurate predictions for houses with transaction price below $200,000 USD. The RF model overestimated the value of houses with low transaction price while underestimating the value of houses with high transaction price. In fact, with a limited number of high transaction price calibration samples to establish a representative pattern, all regression methods underestimated the value of houses with transaction price higher than $400,000 USD.

Table 5. Coefficients of the Explanatory Variables Derived from the GOLS Regression Method

Independent Variable	Coefficient	P Value
Areas (ft^2)	−0.002	0.017
Volume (10,000ft^3)	3.414	< 0.01
DIST_HW (km)	0.376	< 0.01
DIST_RD (m)	0.060	< 0.01
DIST_WATER (km)	0.778	< 0.01
DIST_RC1 (km)	−0.282	< 0.01
DIST_RC2 (km)	0.226	< 0.01
DIST_RC3 (km)	0.232	0.011
LC_TR (%)	−0.084	0.866
LC_GS (%)	0.057	0.908
LC_IMP (%)	−0.023	0.963
LC_RD (%)	−0.122	0.806
NP (#)	0.010	0.345
FRAC_MN	46.714	0.021
CONTAG (%)	0.251	0.410
SHDI (#)	−7.925	0.436
Constant	-54.634	0.474

Further Investigation on the ORR Model

The proposed ORR model proved to be effective in this study making accurate prediction that was comparable to the machine learning approaches and the other expansion method (GWR). However, since the ORR model was based on a stochastic mechanism, it is essential to investigate model stability through multiple-runs. A total of 30 ORR model runs were performed. The median and average adjusted R^2 of the 30 runs were 0.69 and 0.70, respectively. Of the 30 runs, more than two-thirds of the runs (22) yielded better prediction than the GOLS model (adjusted R^2 = 0.67) which confirmed that the ORR model is stable in yielding accurate results.

One of the ORR models did yield quite poor prediction with an adjusted R^2 of 0.55. This poor prediction related to poor calibration of some of the local models, which resulted in abnormal coefficients for several independent variables. These coefficients resulted in somewhat biased estimations when applied to the validation data, which decreased overall prediction accuracy. One possible solution for these biased coefficients is to include a feature selection process, which

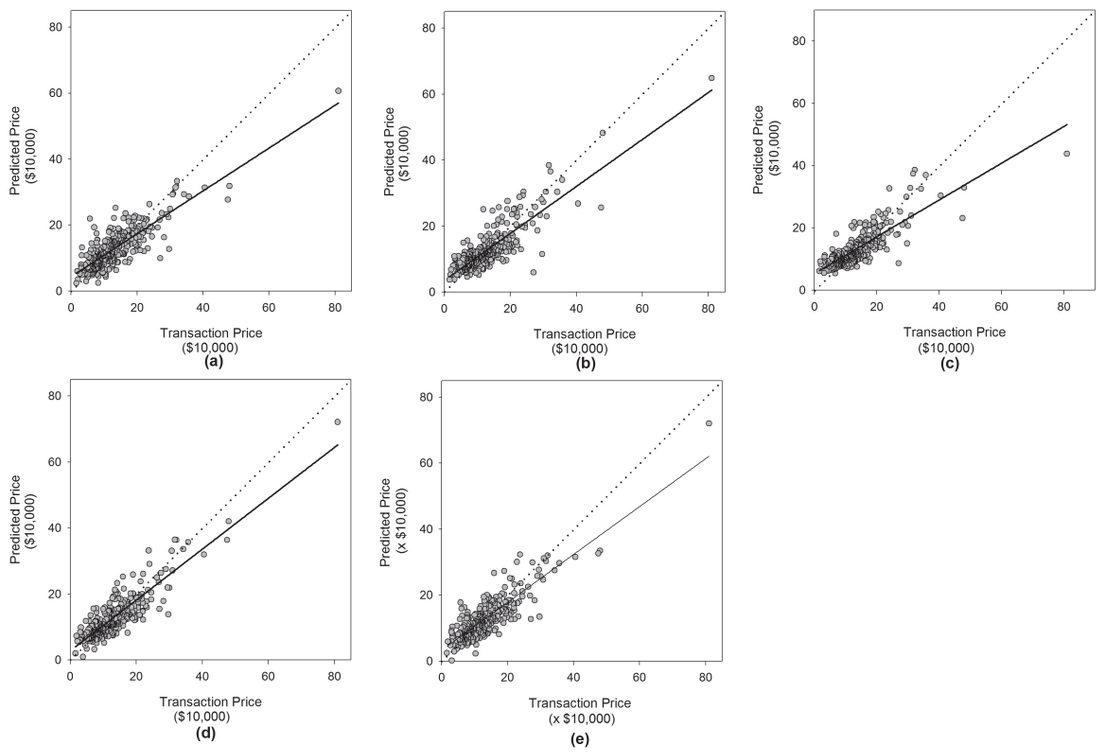


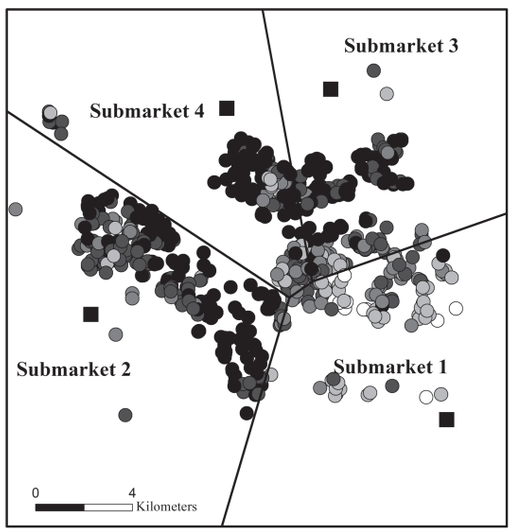
Figure 3. Validation result of the five regression models: (a) using GOLS model, (b) using Cubist model, (c) using RF model, (d) using GWR model, and (e) using ORR model.

will be a focus of future studies aimed at improving the ORR approach.

Further analysis was taken to investigate the sensitivity of the ORR model to sample size using the other four regression models to provide a comparison. With 20 percent of the total samples set aside as validation data, three tests were performed with calibration sample sizes of 459 (Test 1), 333 (Test 2), and 196 (Test 3), which used 70 percent, 50 percent, and 30 percent of the total samples, respectively. Twenty randomly sampled datasets were tested for each sample size. Among the five regression models, the GOLS model was the least sensitive to sample size. With the decreasing number of training samples, both the GOLS and GWR models yielded constant estimation results, although the accuracies were more variable when the number of calibration samples was relatively small (Test 3). Both the Cubist and ORR models were quite sensitive to sample size, with increased variability and decreased prediction accuracy as calibration sample size decreased. Im *et al.* (2011) found that the Cubist regression model is sensitive to sample size such that performance became unstable when the calibration sample size was small. In this study, several Cubist model runs resulted in quite low accuracy when the calibration sample size is small (i.e., Test 1). Compared to the Cubist results, the RF is less sensitive to sample size. In addition, the RF did not yield biased estimation as Cubist did, which implied that using an ensemble of expert trees can effectively decrease the risk of dependence on a certain split node or independent variable.

Although the best accuracies of the ORR model were only slightly different in Test 1 and 2, the accuracy of the ORR model in Test 3 dropped dramatically. While the ORR model significantly outperformed the GOLS model in Test 1 and Test 2, the ORR model showed similar but more variable results in Test 3. Similar patterns were detected for the GWR model. However, the accuracy of GWR model did not drop so dramatically as the ORR model in Test 3. The reason for the relatively constant performance of GWR model is that the GWR model employed the adaptive kernels which produce smaller local estimation variance in areas where data are sparse. To investigate the possible reasons for the poor performance of the ORR model with a small number of calibration samples (Test 3), an experiment was carried out to repeat Test 3 using two submarkets instead of four. Compared to the ORR model with four submarkets, the performance of the ORR model with two submarkets substantially increased with the highest adjusted R^2 of 0.79 and the average adjusted R^2 of 0.70. In addition, the ORR model with two submarkets again significantly outperformed the GOLS model. These results suggest that when the number of calibration samples is relatively small, the ORR model might not perform well with a large number of submarkets since it may not allocate enough samples to each submarket to establish a representative local model.

Figure 4 shows the four submarkets defined by the ORR model that yielded the most accurate prediction. Houses in the southeast (Region 1) submarket had higher transaction prices while houses in the west (Region 2) submarket had relatively lower price. The transaction price of houses located in the northwest (Region 4) and northeast (Region 3) submarkets were more variable, with similar levels of lower and medium transaction price. The map also lists the local models of the four submarkets. The differences in these local models suggest that it is meaningful to divide the study area into submarkets



Submarket Models:

IndependentVariable	Submarket1	Submarket2	Submarket3	Submarket4
Areas(ft2)	-0.003	0.000	0.000	-0.002
Volume(10000ft3)	**3.959***	**2.234***	**1.922***	**2.586***
DIST_HW(km)	0.353	0.109	**0.32***	0.869
DIST_RD(m)	0.095	**0.038***	0.010	**0.079***
DIST_WATER(km)	**3.344***	0.582	0.627	0.332
DIST_RC1(km)	0.125	0.076	-0.137	-0.216
DIST_RC2(km)	0.358	0.175	0.061	0.147
DIST_RC3(km)	0.263	0.048	0.369	0.038
LC_TR (%)	0.110	-2.703	-0.158	1.503
LC_GS (%)	0.412	-2.646	-0.047	1.585
LC_IMP (%)	0.391	-2.836	-0.137	1.161
LC_RD (%)	0.116	-2.855	-0.221	1.182
NP (#)	-0.074	0.000	0.009	-0.008
FRAC_MN	**156.873***	**80.333***	18.271	13.805
CONTAG (%)	0.371	-0.387	0.141	0.225
SHDI (#)	20.865	-15.159	-4.406	24.386
Constant	-258.041	220.354	-8.222	-210.810

Figure 4. Delineated submarkets of the ORR model which yielded the most accurate estimation result (coefficients in the table are those of independent variables in each submarket and * represents significant parameter where p-value <= 0.05).

to establish local models. The footprint area of houses did not contribute to the transaction price estimation in local models for Region 2 and 3, while had a minor, negative correlation to house value in the other two submarkets. All the local models included house volume, which likely provided a better approximation of total living area, since volume considers footprint area and reflects the number of stories.

Variable Importance

Variable selection is an important focus of hedonic modeling (Yoo *et al.*, 2012) and requires understanding the importance of each independent variable. In this study, importance was quantified by considering the percentage increase of RRMSE if an independent variable was excluded; the higher the RRMSE increase, the more important the variable. It is notable that GWR method is not included in this investigation since multi-collinearity should be thoroughly examined and some of the independent variables could not be used simultaneously. Figure 5 summarizes the increase of RRMSE value for each independent variable in the four regression models and clearly shows that the total volume of the residential house was the most important independent variable in all four regression methods. Yoo *et al.* (2012) concluded that the total living area of a residential property was the most important variable for house value prediction. Volume, integrating both footprint size and height, provided a valuable surrogate for living area. No other single variable was recognized as important for all four regression methods. The number of patches (NP) was important for the RF and Cubist methods, while distance to state level recreation areas (Dist_RC1), highways (Dist_HW), and roads (Dist_RD) were important for the GOLS and ORR approaches.

It is interesting that the increase of RRMSE for each independent variable is more pronounced for the ORR method than the other three methods. The reason for this is that the interior optimization function (DE) of the ORR model is based on a stochastic mechanism, yielding a slightly different result for each run. The increase of RRMSE might be caused either by excluding the variable from the model or random variance. Due to this stochastic mechanism, variable importance information given by the ORR model might be less accurate. However, the ORR model is useful in describing importance of variables in each submarket if a feature selection process is implemented. Among the five regression models, the RF method produced the most stable result when a variable was excluded from modeling (i.e., the increase of RRMSE for each independent variable). This indicates that the performance of the RF model is less dependent on any particular variable. This is also caused by RF's stochastic mechanism where a subset of samples and variables were randomly selected. Since the RF establishes a large number (1,000) of expert trees, the contribution of each independent variable might be diluted by averaging.

Conclusions

This study investigated house value estimation using remote sensing derived information as a major data source as a substitute for the extensive field-based observation typically required by current modeling techniques. Final results indicated that models based on remote sensing derived information are capable of deriving meaningful residential property values. Our estimations explained close to 80 percent of the variance of house transaction price, which proved that our proposed methodology is effective. In addition, some of the remote sensing derived variables, such as total volume, provided valuable surrogates of the well-adopted independent variables of hedonic modeling (i.e., living area). Approaches based on expansion models outperformed machine-learning and statistical methods, which confirmed that including spatial information is effective in improving house value estimation accuracy.

The ORR model proposed in this study has three major advantages compared to the other regression models considered. The first advantage is effectiveness, with the ORR approach yielding comparable prediction accuracy to the

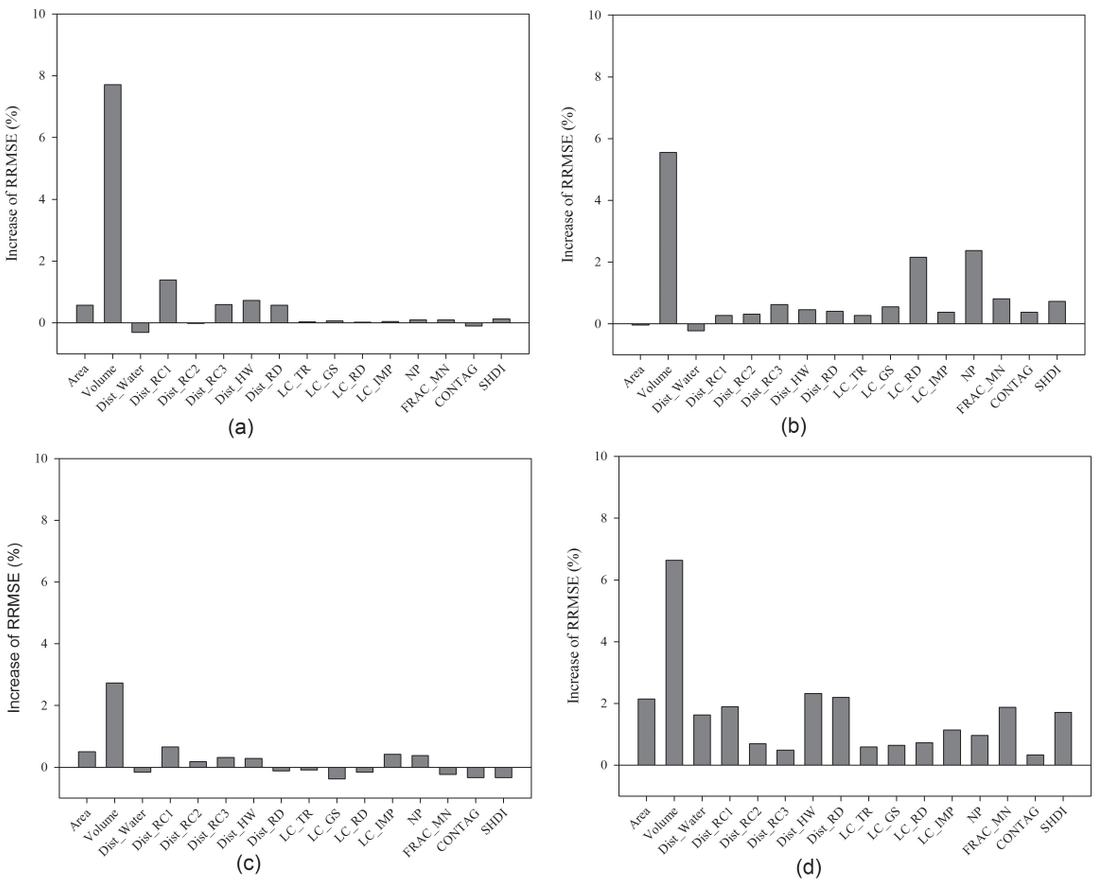


Figure 5. Increase in RRMSE (%) when each independent variable is held in the out-of-bag data in the four regression models (Except GWR).

machine learning approaches (RF and Cubist) as well as the expansion method (GWR) and outperforming them for some cases. The second advantage is robustness, with 22 out of 30 runs resulting in higher estimation accuracy than the GOLS model. The third advantage is interpretability: the information provided by both the local statistical regression model for the submarkets used by the ORR model, and the boundary of these submarkets, facilitates both spatial and statistical analysis of data patterns.

The success of the ORR model in this study suggests that machine-learning optimization techniques are helpful in delineating housing segment (e.g., submarkets) and worth further investigation in the future. The number of submarkets in the ORR model is a site-specific parameter that needs to be established according to the characteristics of study area and available samples. During the sensitivity analysis of the ORR model, we observed overfitting when the number of training samples was small and the number of submarkets was relatively large. Future work in improving the ORR model will explore how to automate the selection of the appropriate number of submarkets. In addition, to make the ORR model more interpretable and robust, we plan to include a feature selection process for each local model.

References

Anderson, S.T., and S.E. West, 2006. Open space, residential property values, and spatial context, *Regional Science and Urban Economics*, 36:773–789.

Breiman, L., 2001. Random forests, *Machine Learning*, 45(1):5–32.

Breiman, L., and A. Cutler, 2004. Notes on setting up and using random forests, URL: *http://oz.berkeley.edu/users/breiman/Using_random_forests_v4.0.pdf* (last date accessed: 07 June 2013).

Casetti, E., 1997. The expansion method, mathematical modeling, and spatial econometrics, *International Regional Science Review*, 20:9–33.

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

Dehring, C., and N. Dunse, 2006. Housing density and the effect of proximity to public open space in Aberdeen, Scotland, *Real Estate Economics*, 34(4):553–566.

DiBari, J.N., 2007. Evaluation of five landscape-level metrics for measuring the effects of urbanization on landscape structure: The case of Tucson, Arizona, USA, *Landscape and Urban Planning*, 79:308–314.

Freeman, A.M., 2003. *The Measurement of Environmental and Resource Values: Theory and Methods*. RFF Press, Washington, D.C., 491 p.

Geoghegan, J., 2002. The value of open spaces in residential land use, *Land Use Policy*, 19(1):91–98.

Gleason, C.J., and J. Im, 2012a. A fusion approach for tree crown delineation from lidar data, *Photogrammetric Engineering & Remote Sensing*, 78(7):679–692.

Gleason, C.J., and J. Im, 2012b. Forest biomass estimation from airborne LiDAR data using machine learning approaches, *Remote Sensing of Environment*, 125:80–91.

Goodman A.C., 1998. Housing market segmentation, *Journal of Housing Economics*, 7:121–143.

Goodman, A.C., and T.G. Thibodeau, 2003. Housing market segmentation and hedonic prediction accuracy, *Journal of Housing Economics*, 12:181–201.

Hamilton, S.E., and A. Morgan, 2010. Integrating lidar, GIS and hedonic price modeling to measure amenity values in urban beach residential property markets, *Computers, Environment and Urban Systems*, 34:133–141.

Im, J., J.R. Jensen, M. Coleman, and E. Nelson, 2009. Hyperspectral remote sensing analysis of short rotation woody crops grown with controlled nutrient and irrigation treatments, *Geocarto International*, 24(4):293–312.

Im, J., Z. Lu, and J.R. Jensen, 2011. A genetic algorithm approach to moving threshold optimization for binary change detection, *Photogrammetric Engineering & Remote Sensing*, 77(2):167–180.

Ismail, S., 2006. Spatial autocorrelation and real estate studies: A literature review, *Malaysian Journal of Real Estate*, 1(1):1–13.

Irwin, E.G., 2002. The effect of open space on residential property value, *Land Economics*, 78(4):465–480.

Kong, F., H. Yin, and N. Nakagoshi, 2007. Using GIS and landscape metrics in hedonic price modeling of the amenity value of urban green space: A case study in Jinan City, China, *Landscape and Urban Planning*, 79:240–252.

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

Lu, Z., J. Im, and L.J. Quackenbush, 2011. A volumetric approach to population estimation using lidar remote sensing, *Photogrammetric Engineering & Remote Sensing*, 77(11):1145–1156.

Odland, J., 1988. *Spatial Autocorrelation*, Sage Publication, Newbury Park, California.

Poudyal, N.C., D.G. Hodges, B. Tonn, and S.H. Cho, 2009. Valuing diversity and spatial pattern of open space plots in urban neighborhoods, *Forest Policy and Economics*, 11(3):194–201.

Price, K.V., 1999. An Introduction to Differential Evolution, *New Ideas in Optimization* (D. Corne, M. Dorigo, and F. Glover, editors) McGraw-Hill, London.

Segal, M.R., 2003. *Machine Learning Benchmarks and Random Forest Regression*, Center for Bioinformatics & Molecular Biostatistics, Kluwer Academic, The Netherlands.

Selim, H., 2009. Determinants of house prices in Turkey: Hedonic regression versus artificial neural network, *Expert Systems with Applications*, 36:2843–2852.

Shultz, S.D., and D.A. King, 2001. The use of census data for hedonic price estimates of open-space amenities and land use, *The Journal of Real Estate Finance and Economics*, 22(2):239–252.

Smith, V.K., C. Poulos, and H. Kim, 2002. Treating open space as an urban amenity, *Resource and Energy Economics*, 24:107–129.

Storn, R., and K. Price, 1997. Differential evolution-a simple and efficient heuristic for global optimization over continuous spaces, *Journal of Global Optimization*, 11:341–359.

Sunding, D., and A.M. Swoboda, 2010. Hedonic analysis with locally weighted regression: An application to the shadow cost of housing regulation in Southern California, *Regional Science and Urban Economics*, 40:550–573.

Tseng, M., S. Chen, G. Hwang, and M. Shen, 2008. A genetic algorithm rule-based approach for land-cover classification, *ISPRS Journal of Photogrammetry and Remote Sensing*, 63:202–212.

U.S. Census Bureau. 2010. Onondaga County, New York (State & County QuickFacts Sheet for Onondaga Count, New York), URL: *http://quickfacts.census.gov* (last date accessed: 07 June 2013).

Wu, C.S., and R. Sharma, 2012. Housing submarket classification: The role of spatial contiguity, *Applied Geography*, 32:746–756.

Yoo, S., J. Im, and J.E. Wagner, 2012. Variable selection for hedonic modeling using machine learning approaches: a case study in Onondaga County, NY, USA, *Landscape and Urban Planning*, 107:293–306.

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

Yu, D., Y.D. Wei, and C.S. Wu, 2007. Modeling spatial dimensions of housing prices in Milwaukee, WI, *Environment and Planning B: Planning and Design*, 34:1085–1102.

(Received 19 July 2012; accepted 05 April 2013; final version 16 April 2013)