Landscape and Urban Planning 107 (2012) 293–306

Contents lists available at SciVerse ScienceDirect

Landscape and Urban Planning

journal homepage: www.elsevier.com/locate/landurbplan

# Variable selection for hedonic model using machine learning approaches: A case study in Onondaga County, NY

Sanglim Yoo[a,*], Jungho Im.[a,b,1,2], John E. Wagner[a,3]

[a] College of Environmental Science and Forestry, State University of New York, Syracuse, NY 13210-2778, USA
[b] School of Urban and Environmental Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan 689-798, South Korea

## HIGHLIGHTS

- Application of machine learning regression methods to hedonic price function to select variables.
- Comparison of selection results of machine learning methods with traditional ordinary least squares method.
- Propose more practical approaches for the selection of important variables for hedonic price function.

## ARTICLE INFO

*Article history:*
Received 22 September 2011
Received in revised form 5 June 2012
Accepted 7 June 2012
Available online 3 July 2012

*Keywords:*
Hedonic model
Variable selection
Machine learning
Cubist
Random Forest
Environmental amenities

## ABSTRACT

Based on the theoretical foundation of hedonic methods, positive relationships between various types of environmental amenities and house sales price have been investigated. However, as hedonic theory does not provide any arguments in favor of specific sets of independent variables, this lack of theoretical support led researchers to select independent variables from empirical results and intuitive information of previous studies. In previous hedonic studies, the most widely used selection criterion was stepwise selection for multiple regression with ordinary least square (OLS) regression for model fitting. The objective of this study is to apply machine learning approaches to the hedonic variable selection and house sales price modeling. Two rule-based machine learning regression methods including Cubist and Random Forest (RF) were compared with the traditional OLS regression for hedonic modeling. Each regression method was applied to analyze 4469 house transaction data from Onondaga County, NY (USA) with two different neighborhood configurations (i.e., 100 m and 1 km radius buffers). Results showed that the RF resulted in the highest accuracy in terms of hedonic price modeling followed by Cubist and the traditional OLS method. Each regression method selected different sets of environmental variables for different neighborhood. Since the variables selected by RF method led to make an in-depth hypothesis reflecting the preferences of house buyers, RF may prove to be useful for important variable selection for the hedonic price equation as well as enhancing model performance.



## 1. Introduction

A major purpose of modern urban planning is the orderly arrangement of parts of the city, so that each part could perform its functions with minimum economic cost and conflicts. In urban area, the intense demand for the services that are provided by environmental amenities is much higher than rural or suburban areas. Therefore, the issue of measuring the demand for environmental amenities has attracted attention from policy decision makers. Specifically, in terms of open space, the questions of what kind of environmental amenities they provide, and how to measure and estimate economic values of these amenities have become a major concern.

The first question has been primarily investigated by the discipline of ecology, while the second and third questions have been discussed by the discipline of economics. Economists have applied various methodologies for estimating economic values of and measuring amenities provided by open space. One of the traditional ways to answer these questions is by looking for clues in related property values. The use of property value differentials arising from the heterogeneity around each property is called the hedonic property method. Applied to open space valuation, this method measures the increases in values of houses in the neighborhoods nearby open space parcels (Loomis, Rameker, & Seidl, 2004).

* Corresponding author. Tel.: +1 315 430 8209; fax: +1 315 470 6535.
*E-mail addresses:* sayoo@syr.edu (S. Yoo), ersgis@unist.ac.kr, imj@esf.edu (J. Im.), jewagner@esf.edu (J.E. Wagner).
[1] Tel.: +82 52 217 2824.
[2] Tel.: +1 315 470 4709.
[3] Tel.: +1 315 470 6971.

0169-2046/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.landurbplan.2012.06.009

While the links between house transaction price and open space amenities have been examined by past hedonic research, less attention has been paid to the question of what set of environmental variables affect valuing open space economically. As hedonic theory does not provide any arguments in favor of a specific set of independent environmental, site specific structural or neighborhood variables nor a specific hedonic price functional form (Anderson, 2000; Freeman, 2003), this lack of theoretical guidance makes the empirical selection of variables less straightforward.

Many hedonic studies have made independent variable selection decisions using cumulative results from prior empirical studies as well as significance level and statistical estimation methods. The most widely used variable selection criterion is stepwise selection for multiple regression. To build models with higher accuracy and to find independent variables that are highly related to the dependent variable for interpreting and estimating house prices, a more detailed and flexible variable selection criteria is needed. Recently, machine learning approaches, such as Cubist (RuleQuest Research Inc.) and Random Forest (RF), have been used in statistical data mining for prediction and regression analysis; however, to the best of our knowledge, Yu and Wu (2006) is the only hedonic study that used Cubist to model house prices and RF has never been used to model house prices. Against this backdrop, this research applied machine learning methods for hedonic variable selection. The objectives of this research were to (1) apply two rule-based machine learning approaches – Cubist and RF – as well as a classical linear ordinary least squares (OLS) regression method to select important variables for the hedonic price function, and (2) evaluate the three regression methods in terms of modeling performance.

## 2. Background

### 2.1. Theoretical framework: theory of hedonic methods

Economists' consideration of the association between residential property value and environmental amenities has a long history. Ridker (1967) and Ridker and Hennings (1967) provide the first empirical evidence that environmental disamenities such as air pollution, water pollution and noise affect residential property value in an urban area (Freeman, 2003). Rosen (1974) presented a general theoretical framework for using hedonic prices to analyze the demand and supply of attributes for different products. Since Rosen (1974), hedonic price theory has provided a coherent basis for describing the market price of a house as a function of the level of characteristics embedded in each house. It is now widely accepted that housing is a composite and heterogeneous good (Cheshire & Sheppard, 1995). It is composed not only of characteristics relating to the structure itself, such as type of house, size, number of rooms, existence of central heating (i.e., structural variables), but also characteristics determined by location, such as school district a house is located, accessibility to a certain attraction point (i.e., neighborhood variables), but also environmental variables.

According to Freeman (2003), the general hedonic price function for housing is of the form;

$$P(A) = f(S, N, E) \quad (1)$$

where $P$ is the actual property sales price; $A$ is the attributes or characteristics of the house; $S$ is a vector of structural characteristics such as square footage of a house, and number of stories; $N$ is a vector of locational and neighborhood characteristics such as population density, school quality, and distance to major road; and $E$ is a vector of environmental amenity characteristics such as distance to environmental amenities, and accessibility to nearest park. The marginal implicit price function of a characteristic can be found by differentiating Eq. (1) with respect to that characteristic and can be interpreted as the additional amount that must be paid by any individual for a higher level of that characteristic, other things being equal. Based on this simple hedonic price function, in the past four decades, there have been a large number of both theoretical and empirical studies of measuring use values of non-market amenities in monetary term relying on hedonic theory.

Using hedonic methods, positive economic relationships between house sales price and various types of open space, such as urban parks (Anderson & West, 2006; Crompton, 2001; Dehring & Dunse, 2006; Irwin, 2002; Luttik, 2000; Morancho, 2003; Smith, Poulos, & Kim, 2002; Troy & Grove, 2008), land in conservation easements (Irwin, 2002), agricultural croplands (Geoghegan, 2002; Smith et al., 2002), forests (Geoghegan, 2002; Smith et al., 2002; Tyrväinen, 1997; Tyrväinen & Miettinen, 2000), and golf courses (Shultz & King, 2001; Smith et al., 2002) have been investigated. In addition to the types of open space, recently several hedonic studies have measured amenity values of spatial configurations of open space patches using landscape indices, including patch density and patch size index (Cho, Jung, & Kim, 2008; Cho, Kim, Roberts, & Jung, 2009; Kong, Yin, & Nakagoshi, 2007), patch richness index (Kong et al., 2007), edge density index (Cho et al., 2008), fractal dimension index (DiBari, 2007; Geoghegan, Wainger, & Bockstael, 1997; Poudyal, Hodges, Tonn, & Cho, 2009), Shannon's diversity index (Acharya & Bennett, 2001; Geoghegan et al., 1997; Poudyal et al., 2009), and interspersion and juxtaposition index (DiBari, 2007). Remote sensing-derived environmental characteristics such as soil fraction and impervious surface fraction (Yu & Wu, 2006) were also evaluated through a hedonic framework. Detailed descriptive overviews of open space valuation literature are found in Fausold and Lilieholm (1999) and McConnell and Walls (2005). Brander and Koetse (2011) conducted meta-analysis of open space valuation literature. Waltert and Schläpfer (2010) systematically assessed the results of peer-reviewed literature that investigated the effects of landscape amenities.

### 2.2. Methodological Issues inHedonic Methods: Selection of Variables

The preferred source of data is systematically collected information on actual sales prices of individual houses, along with relevant characteristics of each house (Freeman, 2003). Hedonic theory does not provide any arguments in favor of a specific set of independent structural, neighborhood, or environmental variables (Anderson, 2000; Freeman, 2003). This lack of theoretical guidelines hampers the empirical testing of hypothesis.

Most hedonic studies have selected variables from the results and theoretical and intuitive information of previous empirical studies as well as classical statistical methods (Anderson, 2000). Because the objective of the hedonic analysis is to determine the effect of one attribute on property values, other things being equal, the hedonic price function should include all structural, neighborhood, or environmental characteristics that enter the utility function of a household (Freeman, 2003; Tyrväinen & Miettinen, 2000). In practice, multicollinearity among independent variables often makes this impractical (Anderson & West, 2006). The consequence of multicollinearity among independent variables is that estimated coefficients for the collinear variables are unstable and have large variances (Wu, Adams, & Plantinga, 2004). The most suggested and widely applied solutions to the problem include dropping high collinear variables from the model, obtaining more data, and formalizing relationships among regressors or parameters (Kennedy, 1998).

There are two objectives for variable selection. The first is to identify all the important variables, even with some redundancy, highly related to the dependent variable for explanatory and interpretation purpose, and the second is to find a

**Table 1**
Summary of variables used in this study.

| Variable | Definition | Source | Unit |
|---|---|---|---|
| Sales price of residential property in USD | Residential property sales price in $ | Post-Standard, 2000 | $ |
| *Independent variables* | | | |
| Structural variables | | | |
| ACRE | Total acreage of a residential property, in acre | Onondaga County | Acre |
| BED | Number of bedrooms in a residential property | Onondaga County | # |
| BATH | Number of full bathrooms in a residential property | Onondaga County | # |
| AGE | Age of a residential property, up to the year transacted (2000) | Onondaga County | Year |
| SQFT | Total living area of a residential property, in square foot | Onondaga County | $ft^2$ |
| Neighborhood variables | | | |
| SCHOOL | Sum of middle level (Grade 8) English language arts and Mathematics scores of all students in each school district based on New York State School Report Card, 2000–2001 | New York State, Dept. of Education | |
| DIST_ROAD | Euclidian distance to the major road | | ft |
| DIST_SYR | Driving distance to the County seat | | ft |
| DIST_MUNI | Euclidean distance to the nearest city and municipal park | | m |
| DIST_COUNT | Euclidean distance to the nearest County park | | m |
| DIST_STATE | Euclidean distance to the nearest state park | | m |
| POPDEN | Population density in a Census block | US Census Bureau | $\#/ft^2$ |
| PER_WHITE | Total number of white population in a Census Block, in percent | US Census Bureau | % |
| MEDHHINC | Median household income in 1999 | US Census Bureau | $ |
| TBACDEGR | Total number of population have bachelor's degree among population 25 years and over in a Census Block | US Census Bureau | # |
| Environmental amenity variables | | | |
| PARK | 0 when no park observed within a buffer, 1 when the nearest park is the city park, 2 when municipal park, 3 when County park, 4 when state park | Onondaga County | |
| DIST_OS | Euclidian distance to the nearest open space | | m |
| LU11 | Total area of open water laud use category in 100 m and 1 km radius buffer | NLCD 2001 | $m^2$ |
| LU20 | Total area of developed land use category in 100 m and 1 km radius buffer | NLCD 2001 | $m^2$ |
| LU31 | Total area of barren land use category in 100 m and 1 km radius buffer | NLCD 2001 | $m^2$ |
| LU40 | Total area of forest (deciduous, evergreen and mixed) land use category in 100 m and 1 km radius buffer | NLCD 2001 | $m^2$ |
| LU52 | Total area of shrub land use category in 100 m and 1 km radius buffer | NLCD 2001 | $m^2$ |
| LU71 | Total area of grass/herbaceous land use category in 100 m and 1 km radius buffer | NLCD 2001 | $m^2$ |
| LU80 | Total area of agricultural land use category in 100 m and 1 km radius buffer | NLCD 2001 | $m^2$ |
| LU90 | Total area of wetland land use category in 100 m and 1 km radius buffer | NLCD 2001 | $m^2$ |
| NP | Number of patches index | | # |
| SHAPE_MN | Mean shape index | | None |
| FRAC_MN | Mean fractal dimension index | | None |
| CONTAG | Contagion index | | % |
| SHDI | Shannon's diversity index | | # |
| PR | Patch richness index | | # |

sufficient parsimonious set of important variables for good prediction of the dependent variable (Genuer, Poggi, & Tuleau-Malot, 2010; Grömping, 2009). In prediction or variable selection with the second objective, the key is to avoid redundancy and obtain a parsimonious prediction model (Grömping, 2009). That means it is not so important that the model contains all relevant variables, as long as prediction works well.

To achieve these two objectives, selection of important variables as well as validation of a model with these variables is essential. Previously conducted hedonic studies have commonly adopted a statistical selection criteria based on significance level such as stepwise selection method (e.g., Conway & Lathrop, 2005; Dunse & Jones, 1998; Garrod & Willis, 1994; Kong et al., 2007). Even though stepwise selection is the most widely used statistical selection method for linear OLS regression, there are some cases in which stepwise regression does not provide the best possible selection of variables. As this method adds and drops variables one at a time, it is possible to miss the optimal model (Tufféry, 2011). Considering the fact that selection of important variables is the key to build a successful hedonic model, it is imperative to compare the result with that of other selection methods. As little research has applied machine learning approaches to build a hedonic model as well as to select independent variables that are highly related to the dependent variable for hedonic price model (e.g. Yu & Wu, 2006), this study is unique and innovative in that it examined the applicability of rule-based machine learning approaches to the hedonic framework.

## 3. Materials and Methods

### 3.1. Study Area

This study was conducted in the Onondaga County, New York, USA. The location of study area is illustrated in Fig. 1. Onondaga County is a part of the Syracuse Metropolitan Statistical Area which includes Onondaga, Oswego and Madison County and anchored

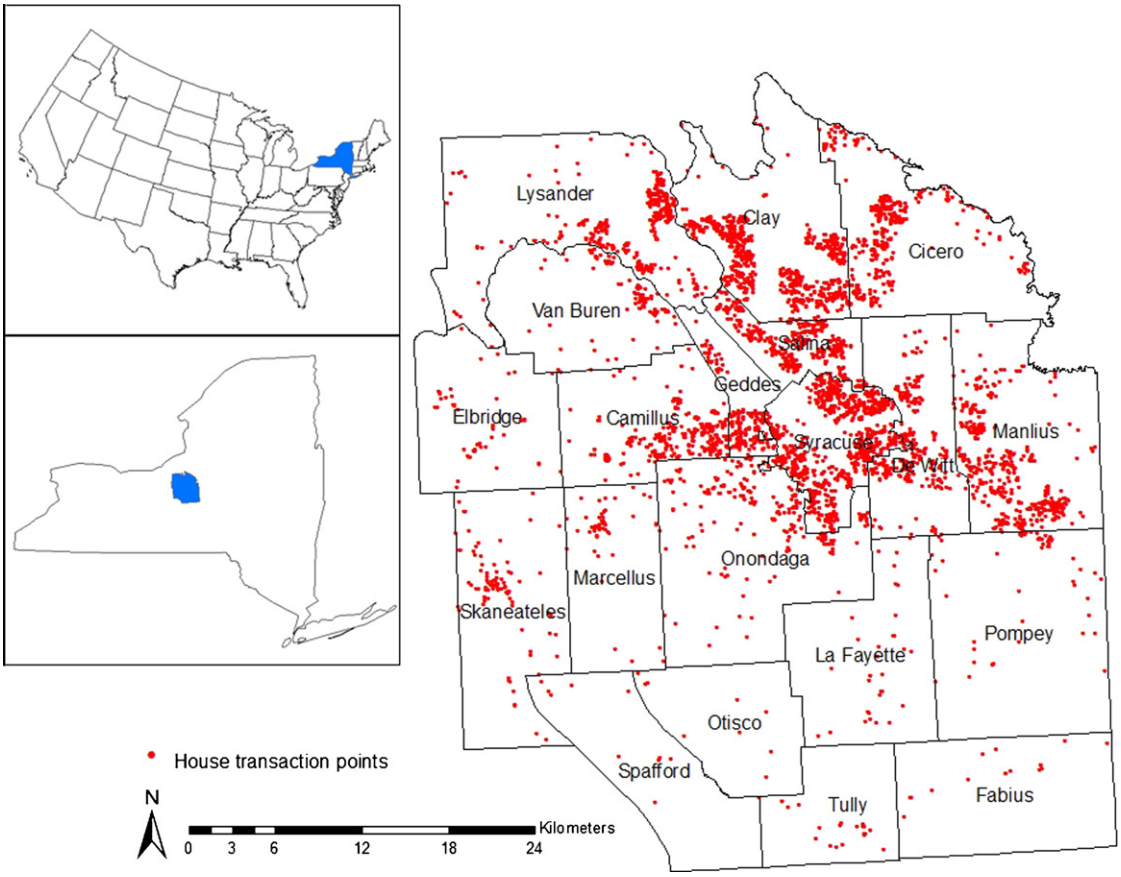


**Fig. 1.** Overview of study area and house transaction points.

by City of Syracuse, county seat of Onondaga County. According to the Census 2000 data, the Onondaga County has a total area of 2100 km$^2$ (806 square miles), of which 2000 km$^2$ (780 square miles) is land and 65 km$^2$ (25 square miles) is water. Population was 458,336 and the population density was 218.26 people/km$^2$ (Census 2000). According to analysis by the Brookings Institution, the City of Syracuse showed strong retention of its population similar to most Northeastern cities. They also described the growth pattern of this region as "sprawl without growth" pointing out the trend that Syracuse's population has declined while that of some suburban towns have grown (Pendall, 2003).

### 3.2. Data

This study was conducted using arms-length transaction records of single family residences in the Onondaga County, New York, USA from January 2000 to December 2000. Hedonic studies require large data sets, which are laborious to collect (Tyrväinen & Miettinen, 2000). There were 4469 single-family houses sold in Onondaga County between January 2000 and December 2000. House transaction points are illustrated in Fig. 1. This research employed six primary data sources including: residential property transaction data provided by the local newspaper the Post-Standard, real estate property information and park and recreation areas information from Syracuse – Onondaga County Planning Agency, Census 2000, US Census 2000 TIGER/line, 2001 National Land Cover Data set (NLCD 2001), and New York State school report cards for the 2000–2001 school years. Residential property transactions listed in the Post-Standard, which covers Syracuse metropolitan area including Onondaga County, were used as the starting point and additional information on the structure of each house were collected through Syracuse – Onondaga GIS on the Web.

**Table 2**
NLCD 2001 land use classification and land use category used in this study.

| NLCD 2001 class code | NLCD 2001 description | Land use code | Description |
|---|---|---|---|
| 11 | Open Water | 11 | Open water |
| 21 | Developed, Open space | 20 | Developed land |
| 22 | Developed, Low intensity | | |
| 23 | Developed, Medium intensity | | |
| 24 | Developed, High intensity | | |
| 31 | Barren land | 31 | Barren land |
| 41 | Deciduous forest | 40 | Forest |
| 42 | Evergreen forest | | |
| 43 | Mixed forest | | |
| 52 | Shrub/scrub | 52 | Shrub land |
| 71 | Grassland/herbaceous | 71 | Herbaceous land |
| 81 | Pasture hay | 80 | Agricultural land |
| 82 | Cultivated crops | | |
| 90 | Woody wetlands | 90 | Wetlands |
| 95 | Emergent herbaceous wetlands | | |

This data set included location of a residential property, property type, sales price, sales date, structural variables (e.g. number of stories, number of bathrooms), and neighborhood characteristics (e.g., school district). Geographical Information System (GIS) was used to integrate real property information with Census 2000, NLCD 2001, 2000 TIGER/line data and other reference information. The data set included 1 dependent variable; i.e., house transaction price, and 31 independent variables. All variables and their definition are given in Table 1. Land use codes in the legend followed that of NLCD 2001 (Table 2).

### 3.3. Defining neighborhood

This study uses two different buffer sizes around each residential property to capture the spatial pattern effects of neighborhood

within a hedonic model: the first neighborhood is given by a 100 m radius buffer and the second is 1 km radius buffer. The first buffer describes the visual zone around a house and the latter is more representative of a typical walking distance.

There is no universal standard for defining neighborhoods for this type of analysis. Previous studies relied on political boundaries, such as census block group, census tract (e.g. Shultz & King, 2001), boundaries defined by the local neighborhood associations (e.g. Poudyal et al., 2009) or a circular area with a fixed radius around the house (e.g. Acharya & Bennett, 2001; Geoghegan et al., 1997; Geoghegan, Lynch, & Bucholtz, 2003; Hite, Jauregui, Sohngen, & Traxler, 2006; Kong et al., 2007; Sander, Polasky, & Haight, 2010). Using political boundaries is problematic when houses close to each other, e.g. across the street, could belong to a different neighborhood just because a street lies between them. Moreover, neighborhood association defined boundaries were not available for this study area. The buffer method with a fixed radius around the house is also often criticized because it considers everything within a certain distance the same even though several different neighborhoods with different level of amenities, social ties and other characteristics could exist within that circular area. However, the buffer method is a better representation of each house buyer's actual preference for surrounding environment than using political boundaries. In particular, when multiple radii buffer are used, the buffer method can show changing preferences of house buyers over space.

For the selection of buffer size, we paid close attention to the buffer sizes used in previous hedonic studies and the average parcel size in Onondaga County, NY. In previous hedonic studies various buffer sizes have been used; for example, Irwin and Bockstael (2001) and Irwin (2002) used a 400 m radius buffer, Acharya and Bennett (2001) used 1mile (approximately 1600 m) and 0.25 mile (approximately 400 m) radius buffers, Geoghegan et al. (1997) used 100 m and 1000 m radius buffers, Geoghegan (2002) chose a 1600 m radius buffer, Geoghegan et al. (2003) used 100 m and 1600 m radius buffers, Hite et al. (2006) used 0.25 mile (approximately 400 m) and 0.5 mile (approximately 800 m) radius buffer, and Sander et al. (2010) used 100 m, 250 m, 500 m and 700 m radius buffers. None of them supported the use of a specific distance, thus this study considered the average parcel size in Onondaga County. The average parcel size was calculated as 11171.18 $m^2$, thus we selected 100 m and 1 km radius buffers because both buffers cover several parcels respectively. 100 m and 1 km radius buffers were used by Geoghegan et al. (1997, 2003) and Sander et al. (2010) found they effectively capture the effects of spatial pattern around each residential property. We believe buffers are an appropriate method because they reflect the perception of the surrounding landscape for actual residents which can be incorporated into a hedonic framework. Defining neighborhoods by selecting 100 m radius buffer for visible distance and 1 km radius buffer for walking distance, we expect different sets of variables to be selected reflecting individual house buyers' preferences within each buffer.

Examples of the 100 m and 1 km radius buffers and the land use information within the buffers are illustrated in Fig. 2. Each circular area shows the spatial configuration of a buffer. Descriptive statistics of variables in 100 m and 1 km buffers are summarized in Tables 3 and 4 respectively.

For these two different buffers, various environmental amenity variables such as area of each land use type, spatial metrics including number of patches (NP), patch richness index (PR), shannon's diversity index (SHDI), shape index (SHAPE_MN), fractal dimension index (FRAC_MN), and contagion index (CONTAG) were calculated. FRAGSTATS, public domain software developed by McGarigal and Marks (1995) was used to calculate spatial metrics for various landscapes. ArcGIS 9.3.1and SAS 9.2 were used to calculate and

**Table 3**
Descriptive statistics for the variables for the 100 m radius buffer ($N = 4469$).

| Variable | Mean | Std. Dev. | Minimum | Maximum |
|---|---|---|---|---|
| PRICE | 104738.36 | 66558.27 | 2500 | 630,000 |
| SCHOOL | 1413.73 | 30.95432 | 1364 | 1471 |
| ACRE | 0.454417 | 1.373426 | 0.04 | 71.87 |
| BATH | 1.480868 | 0.619298 | 1 | 5 |
| AGE | 40.1629 | 32.57673 | 1 | 201 |
| SQFT | 1710.63 | 656.2797 | 516 | 7176 |
| BED | 3.1729693 | 0.7422073 | 1 | 9 |
| DIST_SYR | 15632.53 | 9885.69 | 7920 | 66,528 |
| DIST_ROAD | 127.9524 | 130.1069 | 0.061221 | 1481.59 |
| DIST_MUNI | 1060.36 | 1185.57 | 7.9549194 | 10114.22 |
| DIST_COUNT | 4667.86 | 2561.63 | 18.6600753 | 15871.45 |
| DIST_STATE | 3723.28 | 2461.58 | 13.2016187 | 16759.25 |
| POPDEN | 1185.43 | 1156.53 | 9.1301743 | 7014.92 |
| PER_WHITE | 91.4936644 | 15.9162771 | 6.2397373 | 484.0000000 |
| MEDHHINC | 53995.08 | 18828.13 | 6875 | 110,266 |
| TBACDEGR | 175.5863 | 114.0613 | 0 | 574 |
| DIST_OS | 191.1533 | 252.4415 | 0 | 1681.46 |
| PARK | 0.0358022 | 0.2450365 | 0 | 4 |
| NP | 3.8925934 | 1.7544192 | 1 | 11 |
| SHAPE_MN | 1.2485948 | 0.1504752 | 1 | 2.2929000 |
| FRAC_MN | 1.0546449 | 0.0244788 | 1.0068 | 1.1781 |
| CONTAG | 37.0699623 | 24.3918515 | 0.768 | 100 |
| SHDI | 0.73411 | 0.3099647 | 0 | 1.8988 |
| PR | 2.8675319 | 1.1821163 | 1 | 8 |
| LU11 | 90.10434 | 892.2864 | 0 | 17544.19 |
| LU20 | 24826.65 | 9803.14 | 0 | 31374.99 |
| LU31 | 111.0002 | 1527.24 | 0 | 31374.95 |
| LU40 | 1822.32 | 4152.3 | 0 | 31055.17 |
| LU52 | 879.9569 | 2752.97 | 0 | 28716.92 |
| LU71 | 172.6039 | 1066.17 | 0 | 19544.49 |
| LU80 | 2913.93 | 6826.06 | 0 | 31374.96 |
| LU90 | 534.6365 | 2229.68 | 0 | 30257.68 |

**Table 4**
Descriptive statistics for the variables for the 1 km radius buffer ($N = 4469$).

| Variable | Mean | Std. Dev. | Minimum | Maximum |
|---|---|---|---|---|
| PRICE | 104828.24 | 66558.27 | 2500 | 630,000 |
| SCHOOL | 1413.73 | 30.9543218 | 1364 | 1471 |
| ACRE | 0.4544174 | 1.3734261 | 0.04 | 71.87 |
| BATH | 1.4808682 | 0.6192984 | 1 | 5 |
| AGE | 40.1626762 | 32.5769962 | 1 | 201 |
| SQFT | 1710.65 | 656.2965866 | 516 | 7176 |
| BED | 3.1729693 | 0.7422073 | 1 | 9 |
| DIST_SYR | 15632.53 | 9885.69 | 7920 | 66528 |
| DIST_ROAD | 127.9524342 | 130.1068951 | 0.061221 | 1481.59 |
| DIST_MUNI | 1060.36 | 1185.57 | 7.9549194 | 10114.22 |
| DIST_COUNT | 4667.86 | 2561.63 | 18.6600753 | 15871.45 |
| DIST_STATE | 3723.28 | 2461.58 | 13.2016187 | 16759.25 |
| POPDEN | 1185.43 | 1156.53 | 9.1301743 | 7014.92 |
| PER_WHITE | 91.4936644 | 15.9162771 | 6.2397373 | 484. |
| MEDHHINC | 53995.08 | 18828.13 | 6875 | 110,266 |
| TBACDEGR | 175.5862609 | 114.0612928 | 0 | 574 |
| DIST_OS | 191.1532501 | 252.4414744 | 0 | 1681.46 |
| PARK | 1.3508615 | 1.1141084 | 4 | 4 |
| NP | 128.2081002 | 29.1258209 | 27 | 230 |
| SHAPE_MN | 1.4606322 | 0.0514388 | 1.28 | 1.64 |
| FRAC_MN | 1.0751077 | 0.0059852 | 1.0517 | 1.0949 |
| CONTAG | 45.7467162 | 6.5036289 | 28.4856 | 75.3884 |
| SHDI | 1.6710593 | 0.3099647 | 0.8167 | 2.3616 |
| PR | 10.9474155 | 2.3252074 | 4 | 15 |
| LU11 | 41225.59 | 140485.11 | 0 | 1519289.53 |
| LU20 | 1928240.03 | 918045.64 | 35798.5 | 6187780.3 |
| LU31 | 9307.52 | 58092.34 | 0 | 1096465.79 |
| LU40 | 415286.04 | 346850.34 | 0 | 2156304.83 |
| LU52 | 120724.74 | 124720.15 | 0 | 940864.08 |
| LU71 | 21448.02 | 40451.61 | 0 | 683500.06 |
| LU80 | 374406.02 | 460858.59 | 0 | 2538549.96 |
| LU90 | 184452.49 | 224435.23 | 0 | 2419746.26 |

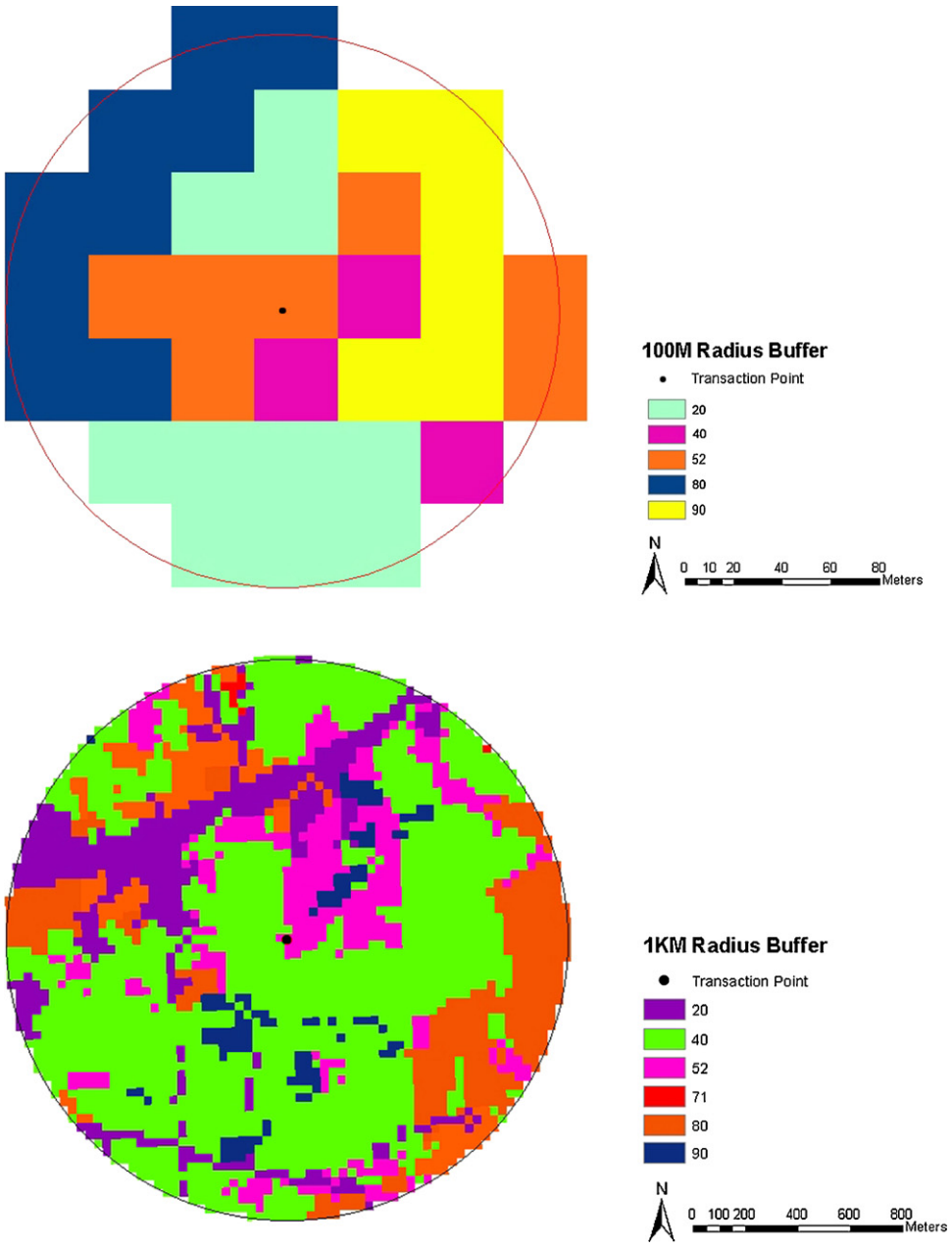


**Fig. 2.** Examples of a 100 m buffer and a 1 km radius buffers with the associated land use information.

integrate environmental amenity variables. Structural and neighborhood variables remain the same within these two buffers.

### 3.4. Methods

#### 3.4.1. Classical linear OLS regression with stepwise selection

Linear OLS regression is a classical parametric method which requires explicit modeling of nonlinearities and interactions. The variable selection method commonly labeled stepwise selection is the most widely used procedure and is a forward selection process because there are no variables selected at the outset. Those having the largest *F*-ratio are added one by one. If the partial sums of squares for any previously included variables do not meet a minimum criterion to stay in the model, the selection procedure changes to backward elimination and variables are dropped one at a time until all remaining variables meet the minimum criterion. Then, forward selection resumes. The variable selection process terminates when all variables in the model meet the criterion to stay and no variables outside the model meet the criterion to enter. In the SAS program, the default 'significance level to enter (SLE)', or '*F*-to-enter' value and the 'significance level to stay (SLS)', or '*F*-to-stay' are 0.15 for the stepwise regression.

The first step in choosing a functional form is to see what theory can tell us, however hedonic theory provides no or very little guidance with respect to the functional form for the price function (Freeman, 2003; Halstead, Bouvier, & Hansen, 1997; Palmquist, 2005). Thus goodness-of-fit criterion was widely applied by most of the empirical hedonic studies (Anderson & West, 2006; Cropper, Leland, & McConnell, 1988; Fotheringham, Brunsdon, & Charlton, 2002; Halstead et al., 1997; Palmquist, 2005). Even though the distribution characteristics between the dependent variable and 31 independent variables for this study clearly indicated linear–linear form of hedonic price equation, this study tested linear–linear, log–linear, linear–log and log–log functional forms of hedonic price equations for the selection of functional form based on the goodness of fit measure. Test results show the linear–linear form of

hedonic price model provides best fit to the data for both 100 m and 1 km buffers yielding the best goodness-of-fit measure. Thus, the linear–linear form of hedonic price model was used in this study. Classical linear–linear hedonic price functional form was constructed using Eq. (2):

$$P(A) = \alpha + \sum_{i=1}^{n} \beta_i \cdot S_i + \sum_{j=1}^{n} \gamma_j \cdot N_j + \sum_{k=1}^{n} \tau_k \cdot E_k + \varepsilon \quad (2)$$

where $P$ is the actual property sales price with $A$ denoting the attributes or characteristics of the house; $S$ is a vector of structural characteristics; $N$ is a vector of locational and neighborhood characteristics; $E$ is a vector of environmental amenity characteristics; $\alpha$, $\beta$, $\gamma$ and $\tau$ are the associated parameters and $\varepsilon$ is random error term; and $i$ is the number of structural variables, $j$ is the number of neighborhood variables, and $k$ is the number of environmental amenity variables. Each house buyer chose their own utility-maximizing home given this price function.

### 3.4.2. Machine learning approaches

Unlike linear OLS regression, machine learning approaches applied in this study are nonparametric and allow nonlinearities and interactions to be learned from the data without any need to model them explicitly (Grömping, 2009). Popularized by Breiman, Friedman, Olshen, and Stone (1984), regression trees have been developed and used in recent years as algorithm development has focused on enhancing predictive power. Important advancements include developing of bagging and boosting techniques as well as hybrid tree-base methods (Walker, Kellndorfer, LaPoint, Hoppus, & Westfall, 2007). Bagging or 'Bootstrap AGGregatING' (Breiman, 1996) and boosting fall into the category of ensemble learning methods where the goal is to construct a 'forest' (i.e., ensemble) of expert trees and combine them through a voting scheme (i.e. simple averaging) for the purpose of improving predictive accuracy (Dietterich, 2000; Walker et al., 2007). As the Cubist and RF have recently gained popularity for their learning ability and flexibility, this study applied two rule-based tree approaches, Cubist and RF, as a variable selection method.

*3.4.2.1. Cubist.* Cubist (RuleQuest Research Inc.) is a tool for generating rule-based regression trees developed by Quinlan (2010). The Cubist package is commercial software that has been used for data mining in a wide range of classification and regression studies. It implements a hybrid tree-based approach that combines a regression tree algorithm with local modeling using a proprietary variant of linear least squares regression. Because it is a commercial product, the exact nature of its algorithm is unknown. The analytical results of Cubist models are expressed as collections of rules, where each rule has an associated multivariate linear model. Whenever a situation matches a rule's conditions, the associated model is used to calculate the predicted value. As Cubist does not require those linear models to be mutually exclusive, output values are averaged to arrive at a final prediction. Recently, this technique has been used to estimate forest biomass (e.g. Blackard et al., 2008; Chen, Gong, Baldocchi, & Tian, 2007; Im, Jensen, Coleman, & Nelson, 2009), to predict soil properties (e.g. Minasny & McBratney, 2008), to quantify impervious surface areas (e.g. Im, Lu, Rhee, & Quackenbush, 2012), and to estimate house values with the aid of Landsat ETM + derived environmental characteristics (e.g. Yu & Wu, 2006). Cubist version 2.04a was used in this study. Cubist returns a rule-based multivariate linear model for house transaction price and the contribution of individual attributes to the model.

*3.4.2.2. Random Forest.* Random Forest (RF) is a classification and regression technique introduced by Breiman (2001) and recently has received a lot of attention in biostatistics and other fields (Grömping, 2009). A forest is an ensemble of trees, and a RF is random into two ways: (1) each tree is based on a random subset of the observations, and (2) each split within each tree is created based on a random subset of candidate variables (Grömping, 2009). RF models are based on classification and regression trees, which are series of binary rule based decisions that dictate how an input is related to its predictor variable. In addition to the bagging strategy stated above, the RF algorithm incorporates a unique approach to node splitting. While node splitting in conventional classification and regression trees is typically accomplished using all predictors, RF node splitting is achieved using a random subset of predictors chosen at each node (Breiman, 2001; Liaw & Wiener, 2002; Walker et al., 2007). When RF constructs a tree, an average of 36.8% (approximately 1/3rd of cases) of the observations are not used for any individual tree, and these hold-out cases are referred to as out-of-bag (OOB) (Breiman, 2001; Grömping, 2009; Walker et al., 2007). In an OOB method, the errors from data excluded for each regression tree generation are used to inform RF of the relative strength and correlation of that tree (Breiman, 2001). Due to the randomness of RF's clustering algorithm, RF may return slightly different result in every trial (Walton, 2008).

RF software is available from the Breiman and Cutler web site as FORTRAN source (Breiman & Cutler, 2004), as an R package (Liaw & Wiener, 2002), or as a commercial product from Salford Systems (http://www.salford-systems.com). This study used RF add-on package with R statistical software. The R software offers several options for the RF model and its outputs. Among these, the variable importance plot (VIP) and percent increase in mean squared error (MSE) provide relative importance of the independent variables in the prediction of dependent variable. To calculate percent increase in MSE and produce a VIP plot, the RF algorithm estimate the importance of each predictor by computing how much the error increases for a given tree when out-of-bag (OOB) data for each predictor are randomly permuted while all other predictors are left unchanged (Liaw & Wiener, 2002; Walker et al., 2007).Only two user-specified parameters are required to run RF: the number of trees in the forest, *ntree*, and the number of variables randomly sampled at each split, *mtry* (Genuer et al., 2010). Although Genuer et al. (2010) showed the effect of a larger value of *mtry* to the magnitude of variable importance by RF using toys data (100 observations, 100–1000 variables), Liaw and Wiener (2002)'s result based on the experiment using Boston housing data (106 observations, 13 variables) suggested that the values of *mtry* did not make dramatic differences in the results if one has a very large number of variables but expects only few to be important. Following Liaw and Wiener's (2002) suggestion which provides more similar setting to this study, we operated RF with default value of *ntree* 500 and that of *mtry* that is the square root of the number of variables for the regression.

### 3.4.3. Accuracy assessment

This study used 3587 out of 4469 transactions as training samples to build a model and used the remaining 882 transactions to validate the model. Training and validation data sets were selected randomly considering the municipality to which each residential property belongs. The training set was built by selecting 80% of transaction data from each municipality and validation was built by selecting 20% of transaction data from each municipality. The 80–20 ratio was chosen to use as many transactions possible to ensure reliable variable selection as well as better model performance. As socio-economic as well as environmental characteristics of residential properties vary depending on the location of each residential property, the proportion of samples from each municipality affects the modeling results. Thus, to reduce the possible effects caused by location specific characteristics of each residential property, the same training and validation data sets was used

**Table 5**
OLS variable selection results and parameter estimates for 100 m radius buffer.

| Step | Variable entered | Partial $R^2$ | Model $R^2$ | Parameter estimates | Pr > \|t\| | VIF |
|---|---|---|---|---|---|---|
| | (Intercept) | – | – | −1722601 | <0.0001 | 0 |
| 1 | SQFT | 0.6412 | 0.6412 | 57.74642 | <0.0001 | 2.77899 |
| 2 | AGE | 0.0555 | 0.6967 | −344.28158 | <0.0001 | 1.47741 |
| 3 | SCHOOL | 0.0322 | 0.7288 | 231.24123 | <0.0001 | 1.98059 |
| 4 | BATH | 0.0094 | 0.7383 | 14260 | <0.0001 | 2.15802 |
| 5 | MEDHHINC | 0.0075 | 0.7458 | 0.42868 | <0.0001 | 1.78028 |
| 6 | LU11 | 0.0046 | 0.7504 | 0.03485 | <0.0001 | 1.28897 |
| 7 | DIST_ROAD | 0.0037 | 0.7541 | 22.22737 | <0.0001 | 1.28788 |
| 8 | PARK | 0.0024 | 0.7565 | −2607.10056 | <0.0001 | 1.24065 |
| 9 | LU31 | 0.0016 | 0.7581 | 0.04720 | <0.0001 | 1.11559 |
| 10 | BED | 0.0015 | 0.7596 | −4310.60895 | <0.0001 | 1.71217 |
| 11 | PR | 0.0008 | 0.7604 | −2544.57596 | <0.0001 | 3.01459 |
| 12 | LU80 | 0.0010 | 0.7614 | 0.00569 | 0.0008 | 2.18258 |
| 13 | DIST_MUNI | 0.0019 | 0.7634 | −3.69314 | <0.0001 | 1.68482 |
| 14 | SHAPE_MN | 0.0009 | 0.7642 | −192646 | <0.0001 | 5.63972 |
| 15 | FRAC_MN | 0.0032 | 0.7675 | 1565988 | <0.0001 | 5.46600 |
| 16 | CONTAG | 0.0011 | 0.7685 | 305.88214 | 0.0292 | 2.95090 |
| 17 | DIST_COUNT | 0.0005 | 0.7691 | 0.84478 | 0.0004 | 1.32722 |
| 18 | LU71 | 0.0007 | 0.7698 | 0.05153 | 0.0028 | 1.26842 |
| 19 | LU52 | 0.0004 | 0.7701 | 0.02234 | 0.0007 | 2.33374 |
| 20 | LU40 | 0.0004 | 0.7705 | −0.00682 | 0.0045 | 2.50601 |
| 21 | ACRE | 0.0004 | 0.7709 | 936.89876 | 0.0228 | 1.22362 |
| 22 | DIST_OS | 0.0002 | 0.7710 | −5.97781 | 0.0716 | 2.52520 |
| 23 | NP | 0.0002 | 0.7712 | −57.35677 | 0.1136 | 3.99328 |

to build and validate models for the two buffer sizes with three regression methods.

Each regression method generates its own accuracy measures, but for the consistency in comparison, root-mean-squared-error (RMSE) and Akaike Information Criterion (AIC) were calculated and compared. The RMSE unit is the same as that of the dependent variable, 2000 US dollar. Error was calculated by subtracting the actual sales price from the estimated value. RMSE indicates the absolute fit of the model to the data explaining how close the observed data points are to the model's predicted values. RMSE is a good measure of how accurately the model predicts the response and is the most important criterion for fit if the main purpose of the model is prediction. Both training RMSE and validation RMSE were calculated in this study. For a reasonable comparison between training and validation, this study also used relative RMSE (rRMSE), which represents the ratio of RMSE to the mean of actual house transaction price for the training or validation data set. In addition, AIC is the measure of the relative goodness of fit of a model indicating the information loss when a given model is used to describe reality (Burnham & Anderson, 2004). A model chosen by AIC values is the one that minimize the information loss during modeling process. Unlike RMSE and rRMSE, AIC values are determined by the number of sample size. Note that this study divided the data into a training and a validation data set, thus the only meaningful comparison among the three modeling methods is within each data set.

## 4. Results

### 4.1. Using the 100 m radius buffer

Through the linear OLS regression model based on the stepwise criteria, 23 independent variables including SQFT, AGE, ACRE, BATH, BED, SCHOOL, MEDHHINC, DIST_ROAD, DIST_MUNI, DIST_COUNT, LU11, LU31, LU40, LU52, LU71, LU80, PARK, DIST_OS, CONTAG, PR, NP, SHAPE_MN and FRAC_MN were selected. The OLS variable selection results and parameter estimates of each variable are summarized in Table 5. The training RMSE was $31607.43 with the rRMSE 30.23%. The validation RMSE was calculated as $33525.65, which was higher than the training RMSE, and the rRMSE was 31.8%. The training AIC for the OLS model was 32327.49 and the validation AIC was 8028.77 (see Table 9).

Cubist provides how individual variables contributes to the model by showing two different types of information: (1) the percentage of cases for which the variable concerned appears in a condition of an applicable rule and (2) the percentage of cases for which the variable appears in the multivariate linear model of an applicable rule. When selecting variables for the Cubist model, both of the information was used. Table 6 summarizes contribution of individual variables to build the Cubist rule and model, and as a result, a total of 28 variables including SQFT, ACRE, AGE, BATH, BED, SCHOOL, POPDEN, PER_WHITE, MEDHHINC, TBACDEGR, DIST_ROAD, DIST_SYR, DIST_MUNI, DIST_COUNT, DIST_STATE, LU11,

**Table 6**
Variables selected by the Cubist model for the 100 m radius buffer training set.

| Percent appearance in a rule conditions (%) | Percent appearance in a multivariate regression models (%) | |
|---|---|---|
| 100 | 99 | AGE |
| 59 | 73 | ACRE |
| 57 | 87 | MEDHHINC |
| 51 | 100 | SQFT |
| 41 | 97 | SCHOOL |
| 8 | 19 | LU52 |
| 6 | 1 | LU11 |
| 4 | 72 | DIST_MUNI |
| 3 | 11 | CONTAG |
| | 89 | BATH |
| | 83 | SHDI |
| | 76 | FRAC_MN |
| | 76 | SHAPE_MN |
| | 75 | DIST_COUNT |
| | 68 | LU20 |
| | 65 | TBACDEGR |
| | 63 | PER_WHITE |
| | 62 | LU40 |
| | 62 | DIST_OS |
| | 57 | DIST_SYR |
| | 56 | NP |
| | 49 | BED |
| | 19 | LU80 |
| | 15 | DIST_STATE |
| | 10 | POPDEN |
| | 10 | LU71 |
| | 10 | LU90 |
| | 8 | DIST_ROAD |

**Table 7**
OLS variable selection results and parameter estimates for 1 km radius buffer.

| Step | Variable entered | Partial $R^2$ | Model $R^2$ | Parameter estimates | Pr > \|t\| | VIF |
|---|---|---|---|---|---|---|
| | (Intercept) | – | – | −402893 | <0.0001 | 0 |
| 1 | SQFT | 0.6395 | 0.6395 | 56.39045 | <0.0001 | 2.81058 |
| 2 | AGE | 0.0569 | 0.6965 | −298.94183 | <0.0001 | 1.39081 |
| 3 | SCHOOL | 0.0320 | 0.7284 | 269.85310 | <0.0001 | 1.64673 |
| 4 | BATH | 0.0099 | 0.7383 | 15068 | <0.0001 | 2.14635 |
| 5 | MEDHHINC | 0.0078 | 0.7461 | 0.44199 | <0.0001 | 1.76792 |
| 6 | LU20 | 0.0081 | 0.7542 | −4.61198 | <0.0001 | 94.32073 |
| 7 | LU40 | 0.0029 | 0.7571 | −4.63970 | <0.0001 | 18.33277 |
| 8 | DIST_MUNI | 0.0023 | 0.7594 | −3.07403 | <0.0001 | 1.44754 |
| 9 | BED | 0.0018 | 0.7612 | −4062.97317 | <0.0001 | 1.71574 |
| 10 | DIST_ROAD | 0.0015 | 0.7627 | 14.57718 | 0.0025 | 1.38912 |
| 11 | LU90 | 0.0008 | 0.7635 | −4.45369 | <0.0001 | 6.40576 |
| 12 | CONTAG | 0.0007 | 0.7642 | 90.12062 | 0.0010 | 1.55007 |
| 13 | ACRE | 0.0005 | 0.7647 | 927.03285 | 0.0239 | 1.20259 |
| 14 | LU80 | 0.0003 | 0.7649 | −3.72503 | <0.0001 | 45.79397 |
| 15 | FRAC_MN | 0.0003 | 0.7652 | 154505 | 0.0406 | 11.90484 |
| 16 | PER_WHITE | 0.0003 | 0.7655 | 25.69858 | 0.0463 | 1.02120 |
| 17 | DIST_OS | 0.0002 | 0.7657 | 5.59514 | 0.0436 | 1.74501 |
| 18 | DIST_COUNT | 0.0002 | 0.7660 | 0.38752 | 0.0955 | 1.24048 |
| 19 | LU71 | 0.0002 | 0.7662 | −4.13761 | <0.0001 | 2.10721 |
| 20 | SHAPE_MN | 0.0002 | 0.7663 | −17922 | 0.1279 | 10.87807 |
| 21 | LU52 | 0.0002 | 0.7665 | −3.36226 | <0.0001 | 8.25342 |
| 22 | LU31 | 0.0002 | 0.7667 | −2.95841 | <0.0001 | 3.36717 |

LU20, LU40, LU52, LU71, LU80, LU90, DIST_OS, NP, SHDI, CONTAG, SHAPE_MN and FRAC_MN were selected among 31 independent variables.

The Cubist training RMSE was $23093.8 and rRMSE was 22.09%. The validation RMSE of the Cubist model was $26399.05, higher than the training RMSE, and the validation rRMSE was 25.04%. The training AIC for the Cubist model was 31359.72 and the validation AIC was 7855.68 (see Table 9).

There is no universally accepted variable selection strategy for RF. Suggested selection criteria from previous studies are limited. Díaz-Uriarte and Alvarez de Andrés (2006) suggested eliminating 20% of the variables having the smallest variable importance and building a new forest with the random variables. They stated the proportion of variables to eliminate is an arbitrary parameter of their method and does not depend on the data. Genuer et al. (2010) suggested selecting sets of variables according to the minimum prediction value given by a model and then keeping only the variables with an averaged variable importance exceeding this level. This study followed the strategy proposed by Díaz-Uriarte and Alvarez de Andrés (2006), eliminating 20% of variables having the smallest variable importance. Through above stated selection procedure, total 26 variables including SQFT, ACRE, AGE, BATH, BED, SCHOOL, POPDEN, PER_WHITE, MEDHHINC, TBACDEGR, DIST_ROAD, DIS_MUNI, DIST_COUNT, DIST_STATE, LU11, LU20, LU40, LU52, LU71, LU80, LU90, DIST_OS, NP, SHDI, CONTAG and SHAPE_MN were selected by the RF model. The RF training RMSE was calculated as $11037.21 with rRMSE = 10.56% while the validation RMSE was $13178.84 with rRMSE = 12.5%. The training AIC for the RF was 29055.47 and the validation AIC was 7319.46 (see Table 9).

### 4.2. Using the 1 km radius buffer

A total of 22 variables were selected for the linear OLS regression model including SQFT, ACRE, AGE, BATH, BED, SCHOOL, PER_WHITE, MEDHHINC, DIST_ROAD, DIST_MUNI, DIST_COUNT, LU20, LU31, LU40, LU52, LU71, LU80, LU90, DIST_OS, CONTAG, SHAPE_MN and FRAC_MN. The OLS variable selection results and parameter estimates of each variable are summarized in Table 7. The RMSE for the training set is $32539.24 and the rRMSE was 31.01%. The validation RMSE was calculated as $31265.71 with the rRMSE = 29.38%. The OLS training AIC was 35595.93 and the validation AIC was 8428.12 (see Table 9).

Table 8 presents the information that Cubist produced for the contribution of individual variable to the model, and as a result, 24 variables were selected including SQFT, ACRE, AGE, BATH, BED, SCHOOL, POPDEN, PER_WHITE, MEDHHINC, DIST_ROAD, DIST_SYR, DIST_MUNI, DIST_COUNT, DIST_STATE, LU20, LU40, LU52, LU71, LU80, LU90, DIST_OS, SHDI, SHAPE_MN and FRAC_MN were selected among total 31 independent variables. The RMSE for the training set was calculated as $24110.20 and the training rRMSE was 23.08%. The Cubist validation RMSE was $23789.96 with the rRMSE = 22.47%. The Cubist training AIC was 31485.91 and the validation AIC was 7767.96 (see Table 9).

The RF model selected 26 variables out of the 31 variables including SQFT, ACRE, AGE, BATH, BED, SCHOOL,

**Table 8**
Variables selected by the Cubist model for the 1 km radius buffer training set.

| Percent appearance in a rule conditions (%) | Percent appearance in a multivariate regression models (%) | |
|---|---|---|
| 100 | 95 | AGE |
| 69 | 96 | MEDHHINC |
| 59 | 68 | ACRE |
| 45 | 100 | SQFT |
| 32 | 93 | SCHOOL |
| 3 | 68 | LU80 |
| 3 | 70 | DIST_SYR |
| | 98 | LU20 |
| | 97 | BATH |
| | 78 | TBACDEGR |
| | 68 | SHDI |
| | 66 | DIST_COUNT |
| | 61 | PER_WHITE |
| | 57 | BED |
| | 26 | LU40 |
| | 20 | DIST_MUNI |
| | 13 | DIST_STATE |
| | 11 | POPDEN |
| | 11 | FRAC_MN |
| | 8 | LU90 |
| | 8 | DIST_ROAD |
| | 8 | LU52 |
| | 6 | DIST_OS |
| | 6 | SHAPE_MN |

**Table 9**
Summary of RMSE, rRMSE and AIC for each regression method.

| | | Linear regression | Cubist | Random Forest | |
|---|---|---|---|---|---|
| 100 m | Training | RMSE | 31607.43 | 23093.8 | 11037.21 |
| | | rRMSE (%) | 30.23 | 22.09 | 10.56 |
| | | AIC | 32327.49 | 31359.72 | 29055.47 |
| | Validation | RMSE | 33525.65 | 26399.05 | 13178.84 |
| | | rRMSE (%) | 31.80 | 25.04 | 12.5 |
| | | AIC | 8028.77 | 7855.68 | 7319.46 |
| 1 km | Training | RMSE | 32539.24 | 24110.20 | 11318.56 |
| | | rRMSE (%) | 31.01 | 23.08 | 10.84 |
| | | AIC | 35595.93 | 31485.91 | 29133.90 |
| | Validation | RMSE | 31265.71 | 23789.96 | 13289.01 |
| | | rRMSE (%) | 29.38 | 22.47 | 12.55 |
| | | AIC | 8428.12 | 7767.96 | 7324.21 |

POPDEN, PER_WHITE, MEDHHINC, TBACDEGR, DIST_ROAD, DIST_SYR, DIST_MUNI, DIST_COUNT, DIST_STATE, LU11, LU20, LU40, LU52, LU80, DIST_OS, NP, SHDI, CONTAG, SHAPE_MN and FRAC_MN. The RF training RMSE was $11318.56 and the rRMSE was 10.84%. The RF validation RMSE was $13289.01, higher than the training RMSE, and the RF validation rRMSE was 12.55%. The RF training AIC was 29133.90 and the validation AIC was 7324.21 (see Table 9).

Fig. 3 illustrates the increase in mean squared error (MSE) in percentage when each independent variable was held in the out-of-bag data, which shows relative importance of each variable in the RF model. While using the 100 m buffer, SQFT and AGE were identified as the most important variables while PARK and FRAC_MN appeared the least contributing variables. Similarly, SQFT and AGE resulted in the considerable increase in MSE when held out-of-bag using the 1 km buffer. LU71, LU90 and PARK were considered as relatively less important variables when using the 100 m buffer.

While RF contains an inherent cross-validation procedure, which can be used to identify relative importance of each independent variable, OLS and Cubist regressions do not. In order to identify the sensitivity (i.e., relative importance) of each independent variable in the OLS and Cubist models similar to that in the RF model (Fig. 3), the increase in MSE in percentage when each independent variable was excluded was calculated (Fig. 4). It should be noted that most of the variables were not considered as important in the OLS regression (i.e., <5% increase in MSE) and SQFT was dominantly the most important variable for using both 100 m and 1 km buffers (i.e., ~50% increase in MSE). In the sensitivity of the variables in the Cubist model, SQFT and AGE were identified as important variables using both 100 m and 1 km buffers.

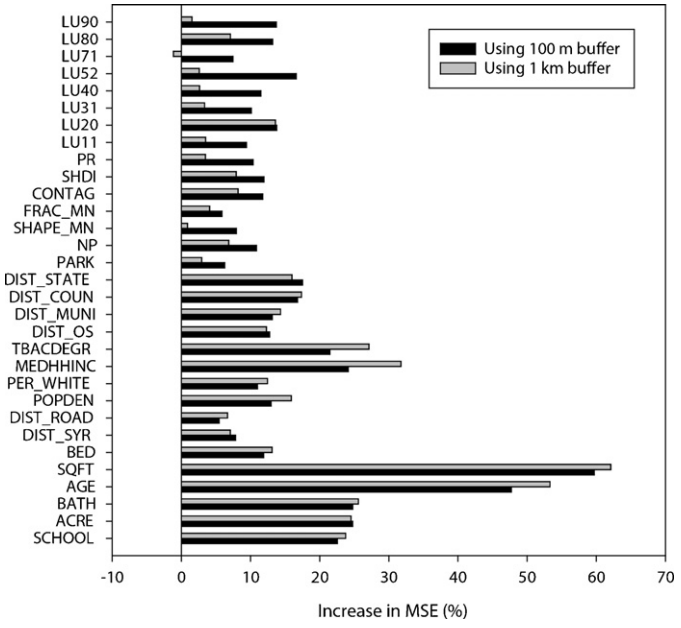


**Fig. 3.** Increase in MSE (%) when each independent variable is held in the out-of-bag data in the RF model for the 100 m and 1 km radius buffers.

Table 9 summarizes the RMSE, rRMSE and AIC values of all regression methods. Results showed that machine learning techniques, both Cubist and RF, produced lower RMSE and AIC values than the traditional OLS regression method. The validation scatter plots between the actual prices and the predicted prices by each model are illustrated in Fig. 5. These scatter plots also supported the aforementioned comparison results – better fitting of machine learning regression methods. RF provided best fit with $R^2$ of 0.97 for the 100 m buffer which indicated 97% of the house transaction price variation was explained with the model and 0.91 for the 1 km buffer which indicating 91% of the house transaction price variation was explained with the model. Cubist fit the data with $R^2$ value of 0.85 for the 100 m buffer which indicating 85% of the house transaction price variation was explained with the model and 0.84 for the 1 km buffer which indicating 84% of the house transaction price variation was explained with the model. OLS regression produced the least fit among three regression methods with $R^2$ of 0.76 for the 100 m buffer which indicating 76% of the house transaction price variation was explained with the model and 0.71 for the 1 km buffer which indicating 71% of the house transaction price variation was explained with the model. Variables selected for each model were summarized in Table 10.

Results of this study showed that the two machine learning techniques outperformed the classical OLS regression method for house price estimation. RF resulted in the higher $R^2$ values and lower RMSEs, rRMSEs and AIC values than Cubist and linear OLS regression for both sizes of buffers. In RF, there is no need for validation or a separate test set to get an unbiased estimate of the test set error, because error is estimated internally, during the forest building process (Breiman & Cutler, 2004). However to ensure the models provide accurate information about the data being modeled and to compare each regression technique, this study conducted model validation for RF and compared this with Cubist and linear OLS regression validation results.

Based on the modeling performance and selected variables, the RF method appears more suitable than Cubist for the purposes of this study. We expected different sets of variables from each category to be selected by three regression methods reflecting individual house buyers' preferences within each buffer. As summarized in Table 10, each regression technique selected different neighborhood and environmental variables for two sizes of buffers, which might reflect house buyers' preferences. For the 100 m buffer, RF selected variables more focused on land use types within a buffer. All of the variables explaining total area of open space land use categories, excluding LU31 which represent total area of barren land use category in each buffer, were selected for

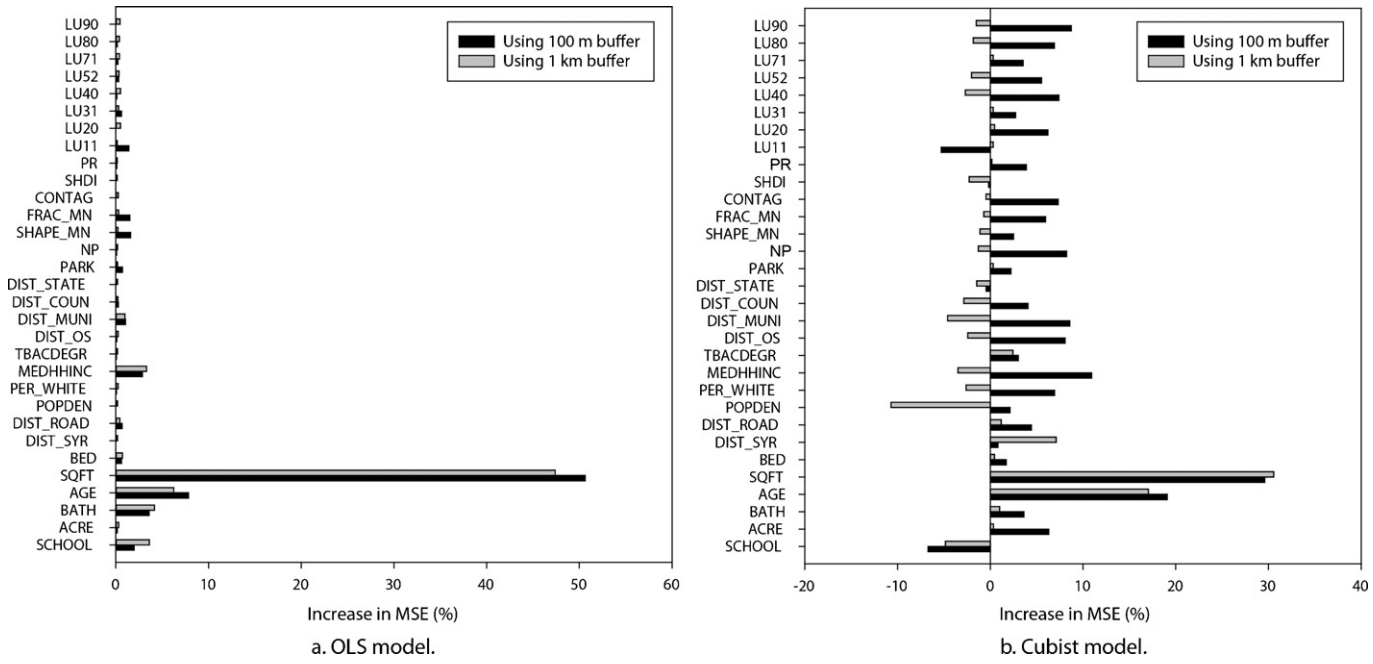


**Fig. 4.** Sensitivity analysis results for the OLS and Cubist models: increase in MSE (%) when each independent variable is excluded for the 100 m and 1 km radius buffers.

the regression model. For the 1 km buffer, RF focused more on the variables associated with land use configurations. These results led to hypothesizing that house buyers may be more concerned about land use types within a visible distance thus may prefer or dislike certain types of land use around their house within visible distance. Also, we hypothesized that house buyers may be more concerned about land use configurations around their house within walking distance. Landscape metrics NP, CONTAG, SHDI and SHAPE_MN

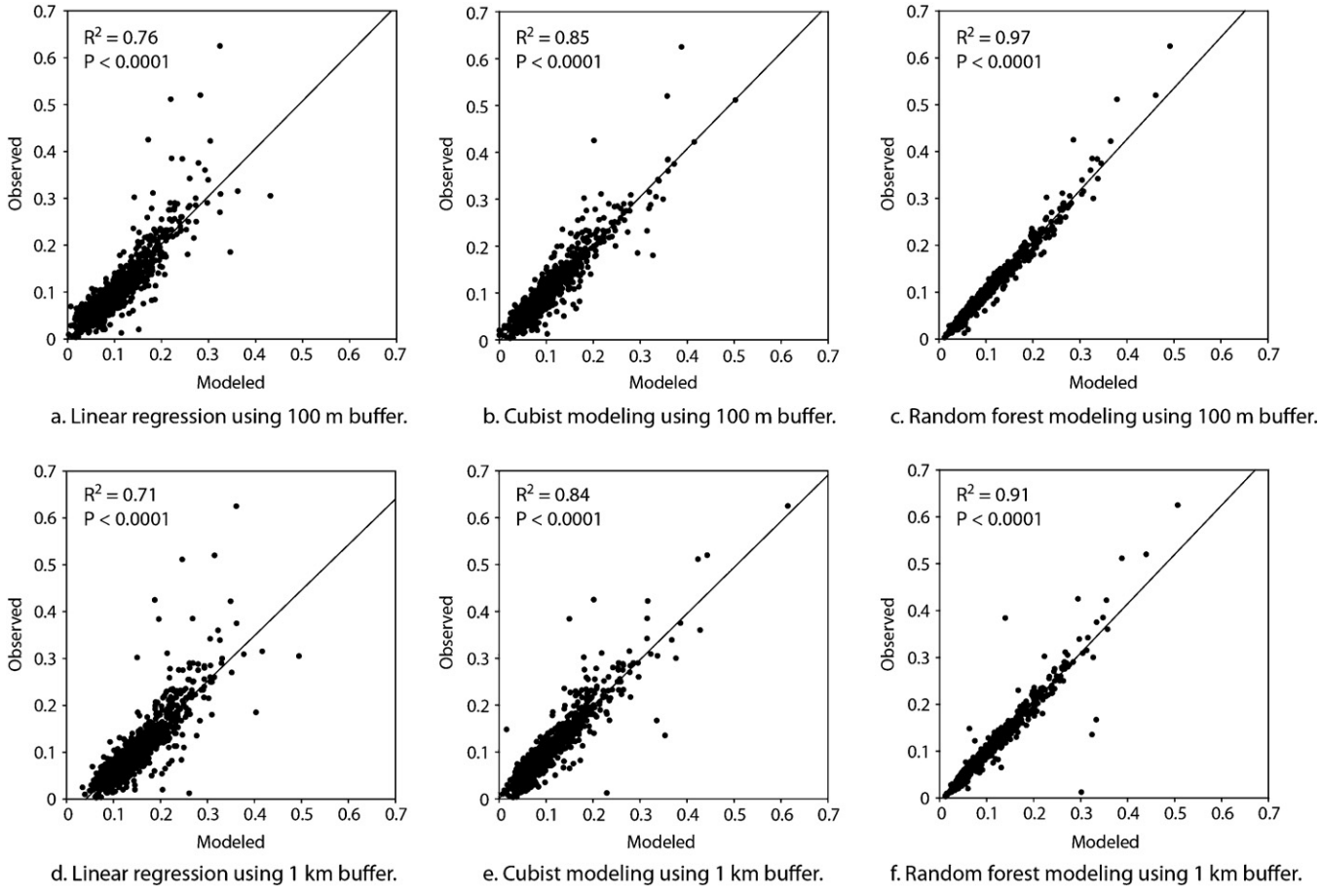


**Fig. 5.** Predicted values vs. observed values for each regression method using the validation data set (unit: million dollars).

**Table 10**
Summary of selected variables for each regression method.

| Buffer | Linear regression | Cubist | Random forest |
|---|---|---|---|
| 100 m | S: SQFT, ACRE, AGE, BATH, BED | S: SQFT, ACRE, AGE, BATH, BED | S: SQFT, ACRE, AGE, BATH, BED N: SCHOOL, POPDEN, PER_WHITE, MEDHHINC, TBACDEGR, DIST_ROAD, DIST_MUNI, DIST_COUNT, DIST_STATE A: LU11, LU20, LU40, LU52, LU71, LU80, LU90, DIST_OS, NP, SHDI, CONTAG, SHAPE_MN |
| | N: SCHOOL, MEDHHINC, DIST_ROAD, DIST_MUNI, DIST_COUNT | N: SCHOOL, POPDEN, PER_WHITE, MEDHHINC, TBACDEGR, DIST_ROAD, DIST_SYR, DIST_MUNI, DIST_COUNT, DIST_STATE | |
| | A: LU11, LU31, LU40, LU52, LU71, LU80, PARK, DIST_OS, CONTAG, PR, NP, SHAPE_MN, FRAC_MN | A: LU11, LU20, LU40, LU52, LU71, LU80, LU90, DIST_OS, NP, SHDI, CONTAG, SHAPE_MN, FRAC_MN | |
| | Total: 23 variables | Total: 28 variables | Total: 26 variables |
| 1 km | S: SQFT, ACRE, AGE, BATH, BED N: SCHOOL, PER_WHITE, MEDHHINC, DIST_ROAD, DIST_MUNI, DIST_COUNT A: LU20, LU31, LU40, LU52, LU71, LU80, LU90, DIST_OS, CONTAG, SHAPE_MN, FRAC_MN | S: SQFT, ACRE, AGE, BATH, BED N: SCHOOL, POPDEN, PER_WHITE, MEDHHINC, DIST_ROAD, DIST_SYR, DIST_MUNI, DIST_COUNT, DIST_STATE A: LU20, LU40, LU52, LU71, LU80, LU90, DIST_OS, SHDI, SHAPE_MN, FRAC_MN | S: SQFT, ACRE, AGE, BATH, BED N: SCHOOL, POPDEN, PER_WHITE, MEDHHINC, TBACDEGR, DIST_ROAD, DIST_SYR, DIST_MUNI, DIST_COUNT, DIST_STATE A: LU11, LU20, LU40, LU52, LU80, DIST_OS, NP, SHDI, CONTAG, SHAPE_MN, FRAC_MN |
| | Total: 22 variables | Total: 24 variables | Total: 26 variables |

were selected for both buffer sizes and we believe this result indicates that house buyers consider the aggregation level of various land use types around their residential property. According to the RF variable selection results, house buyers are more concerned about environmental amenities of surrounding areas in larger scale than smaller scale. In Fig. 3, the percent increases in MSE values of environmental amenity variables were relatively higher in 100 m than 1 km radius buffer. That means when an individual makes a house buying decision, that person tends to consider environmental amenities in visible distance rather than in walking distance.

Cubist selection results emphasized the importance of environmental variables more than RF, especially when using the 100 m buffer. For the 100 m buffer, five environmental variables – two land use variables and three landscape metrics – were used in rule-based models with more than 50% frequency (Table 6). For the 1 km buffer, three environmental variables including two land use variables and the land use metric SHDI were used in rule-based models with more than 50% frequency (Table 8). However, Cubist selection results for each buffer size shows difference in terms of the number of selected variables. Thus Cubist selection results did not coincide with this study's expectations.

Yu and Wu (2006) used a Cubist approach to model house prices. They compared classical OLS regression method and Cubist based on the mean average error (MAE), the relative error (RE), and the product-moment correlation coefficient ($R$) between predicted values and the actual values to evaluate the relationship between house values and environmental amenities. With the smaller MAE and RE values and larger $R$ value, they found Cubist could be a better alternative to the OLS regression model in terms of predicting house values with house structural attributes and environmental characteristics. However, their hedonic price function did not include neighborhood variables. This is problematic because hedonic theory regards a house as a bundle of attributes that cannot easily be repackaged (Anderson, 2000). Moreover, the hedonic price function should include all housing characteristics that enter the utility function of a household (Tyrväinen & Miettinen, 2000) and when important variables are not included in the final hedonic model, omitted variable bias occurs. Yu and Wu (2006) used aggregated house transaction data by averaging individual data attributes within a census block group even though their data set included two dummy variables. Their stated purpose of aggregating data into census block group was to acquire neighborhood homogeneity, but Census Block Group is a relatively large geographic area for evaluating individual house buyers' preferences. In addition, Yu and Wu (2006) used assessed value as the dependent variable and not actual sales data which is the better representative of individuals' actual preferences. While the method proposed in Yu and Wu (2006) appears to be useful to catch overall spatial patterns of house price for an area under investigation, it lacks in identifying important variables that can reflect buyers' preferences.

Together with Cubist and RF method, support vector regression (SVR) is a family of machine learning approaches (Mountrakis, Im, & Ogole, 2011). Walton (2008) applied and compared Cubist, RF and SVR and found the SVR method to be the best estimator for estimating urban land cover. Following Walton (2008), we tested SVR using Chang and Lin's (2001) LIBSVM (a library for support vector machines) and found SVR produced larger RMSE values for training and validation data sets. For the 100 m buffer, the training RMSE was $68,845 and the validation RMSE was $70,269. For the 1 km buffer, the training RMSE was $69,227 and the validation RMSE was $70,305. These RMSE values were even higher than OLS, thus we did not consider SVR in our study. The poor performance of SVR might be because parameter optimization was not included in the LIBSVM. Optimizing the parameters that are used in SVR could increase house price estimation accuracy. Even though Walton (2008) found SVR was the best estimator for urban forest canopy cover, SVR did not provide desirable results for estimating house price.

Considering that there are no theoretical arguments in favor of a specific set of independent variables for the hedonic model (Anderson, 2000), most of previous hedonic studies have selected independent variables case specifically regarding characteristics of data. We believe more attention should be paid to the rule-based machine learning approaches, because machine learning approaches are more flexible than conventional regressions, thus these are well adapted to site-specific data such as hedonic data. As the rule-based machine learning approaches are more resistant to noise in the data (Tuffféry, 2011), they enhance the model fitting to the data set than traditional regression approaches. Consistently, in this study, we found the rule-based machine learning approaches generated better fit than the classical OLS regression method and RF variable selection results were better representations of the neighborhood characteristics around a residential property. Selecting important variables for hedonic equation using RF is a practical decision because as Cubist is a commercial product, the high cost of acquiring the software may make it less applicable. In addition, its exact algorithm works as a 'black-box' and only rules and the associated regression models are reported as a result. In the light

**Table 11**
Summary of Moran's I for each regression method.

| | Linear regression | Cubist | Random Forest | |
|---|---|---|---|---|
| 100 m | Training | 0.044181 | 0.030615 | 0.025153 |
| | Validation | 0.057250 | 0.042519 | 0.038315 |
| 1 km | Training | 0.033120 | 0.044965 | 0.029366 |
| | Validation | 0.047873 | 0.038688 | 0.035369 |

of these facts, RF may prove to be useful for selecting independent variables that are highly related to dependent variable for the hedonic equation as well as enhancing model performance.

Even though the focus of this study lies in hedonic variables selection, there are other methodological issues in hedonic methods. One of the critical issues is the existence of spatial autocorrelation in model residuals. In hedonic framework, when variables that are highly related to the dependent variable and common to all observations in a neighborhood are not included in the model, spatial autocorrelation is likely to occur (Geoghegan et al., 2003). Cubist and RF are not typically designed to incorporate spatial autocorrelation in modeling the relationship between variables, we tested the extent of spatial autocorrelation in model residuals using Moran's *I* index (see Table 11). Results show that RF reduces the training and validation spatial autocorrelation in model residuals for both buffer sizes than the OLS regression method. Cubist reduces training and validation spatial autocorrelation in model residuals for the 100 m buffer, while generated higher level of training spatial autocorrelation but lower level of validation spatial autocorrelation than OLS when using the 1 km buffer. This result confirms that RF is recommended as a variable selection method for the hedonic equation.

The importance of this study lies in the proposition of a more flexible variable selection procedure than conventional variable selection methods for hedonic modeling given individual buyers' preferences may make hedonic modeling more complicated and non-linear. This study is the first attempt to apply machine learning approaches to select important variables for hedonic modeling. We believe succeeding hedonic studies using machine learning methods will provide empirical support for the result of this study.

## References


Acharya, G., & Bennett, L. L. (2001). Valuing open space and land-use patterns in urban watersheds. *The Journal of Real Estate Finance and Economics*, *22*(2), 221–237.

Anderson, D. E. (2000). Hypothesis testing in hedonic price estimation–On the selection of independent variables. *The Annals of Regional Science*, *34*(2), 293–304.

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

Blackard, J. A., Finco, M. V., Helmer, E. H., Holden, G. R., Hoppus, M. L., Jacobs, D. M., et al. (2008). Mapping US forest biomass using nationwide forest inventory data and moderate resolution information. *Remote Sensing of Environment*, *112*(4), 1658–1677.

Brander, L. M., & Koetse, M. J. (2011). The value of urban open space: Meta-analysis of contingent valuation and hedonic pricing results. *Journal of Environmental Management*, *92*(10), 2763–2773.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Washington, DC: Chapman and Hall/CRC.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

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

Breiman, L. & Cutler, A. (2004). Random forests. Retrieved from http://www.stat.berkeley.edu/users/breiman/randomforests

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research*, *33*(2), 261–304.

Chang, C. C., & Lin, C. J. (2001). LIBSVM: A library for support vector machines. Retrieved from http://www.csie.ntu.edu.tw/~cjlin/libsvm and http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.ps.gz.

Chen, Q., Gong, P., Baldocchi, D., & Tian, Y. Q. (2007). Estimating basal area and stem volume for individual trees from lidar data. *Photogrammetric Engineering and Remote Sensing*, *73*(12), 1355–1365.

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

Cho, S. H., Jung, S., & Kim, S. G. (2008). Valuation of spatial configurations and forest type in the Southern Appalachian highlands. *Environmental Management*, *43*(4), 628–644.

Cho, S. H., Kim, S. G., Roberts, R. K., & Jung, S. (2009). Amenity values of spatial configurations of forest landscapes over space and time in the Southern Appalachian Highlands. *Ecological Economics*, *68*(10), 2646–2657.

Conway, T. M., & Lathrop, R. G. (2005). Modeling the ecological consequences of land-use policies in an urbanizing region. *Environmental Management*, *35*(3), 278–291.

Crompton, J. L. (2001). The impact of parks on property values: A review of the empirical evidence. *Journal of Leisure Research*, *33*(1), 1–31.

Cropper, M. L., Leland, B. D., & McConnell, K. E. (1988). On the choice of functional form for hedonic price functions. *The Review of Economics and Statistics*, *70*(4), 668–675.

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

Díaz-Uriarte, R., & Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, *7*(1), 3–16.

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.

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, *40*(2), 139–157.

Dunse, N., & Jones, C. (1998). A hedonic price model of office rent. *Journal of Property Valuation and Investment*, *16*(3), 297–312.

Fausold, C. J., & Lilieholm, R. J. (1999). The economic value of open space: A review and synthesis. *Environmental Management*, *23*(3), 307–320.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). *Geographically weighted regression: The analysis of spatially varying relationships*. Chichester, England: John Wiley & Sons Ltd.

Freeman, A. M. (2003). *The measurement of environmental and resource values: Theory and methods*. Washington, DC: RFF Press.

Garrod, G., & Willis, K. (1994). An economic estimation of the effect of a watershed location on property values. *Environmental and Resource Economics*, *4*(2), 209–217.

Genuer, R., Poggi, J. M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letter*, *31*(14), 2225–2236.

Geoghegan, J., Wainger, L. A., & Bockstael, N. E. (1997). Spatial landscape indices in a hedonic framework: An ecological economics analysis using GIS. *Ecological Economics*, *23*(3), 251–264.

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

Geoghegan, J., Lynch, L., & Bucholtz, S. (2003). Capitalization of open space into housing values and the residential property tax revenue impacts of agricultural easement program. *Agricultural and Resource Economics Review*, *32*(1), 33–45.

Grömping, U. (2009). Variable importance assessment in regression: Linear regression versus random forest. *American Statistical Association*, *63*(4), 308–319.

Halstead, J. M., Bouvier, R. A., & Hansen, B. E. (1997). On the issue of functional form choice in hedonic functions: Further evidence. *Environmental Management*, *21*(5), 759–765.

Hite, D., Jauregui, A., Sohngen, B. L., & Traxler, G. J. (2006). Open space at the rural-urban fringe: A joint spatial hedonic model of developed and undeveloped land values. Retrieved at SSRN from http://ssrn.com/abstract=916964

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

Im, J., Lu, Z., Rhee, J., & Quackenbush, L. J. (2012). Impervious surface quantification using a synthesis of artificial immune networks and decision/regression trees from multi-sensor data. *Remote Sensing of Environment*, *117*, 102–113.

Irwin, E. G., & Bockstael, N. E. (2001). The problem of identifying land use spillovers: Measuring the effects of open space on residential property values. *American Journal of Agricultural Economics*, *83*(3), 698–704.

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

Kennedy, P. A. (1998). *Guide to econometrics*. Cambridge, MA: MIT Press.

Kong, F., Yin, H., & Nakagoshi, N. (2007). Using GIS and landscape metrics in the 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.

Liaw, A., & Wiener, M. (2002). Classification and regression by random forest. *R News*, *2*(3), 18–22.

Loomis, J., Rameker, V., & Seidl, A. (2004). A hedonic model of public market transaction for open space protection. *Journal of Environmental Planning and Management*, *47*(1), 86–93.

Luttik, J. (2000). The value of trees, water and open space as reflected by house prices in the Netherlands. *Landscape and Urban Planning*, *48*(3–4), 161–167.

McConnell, V., & Walls, M. (2005). *The value of open space: Evidence from studies of nonmarket benefits*. Washington, DC: RFF Press.

McGarigal, K., & Marks, B.J., (1995). FRAGSTATS: spatial pattern analysis program for quantifying landscape structure. USDA Forest Service General Technical Report. PNW-GTR-351. Portland, OR.

Minasny, B., & McBratney, A. B. (2008). Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, *94*(1), 72–79.

Mountrakis, G., Im, J., & Ogole, C. (2011). Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *66*, 247–259.
Morancho, A. (2003). A hedonic valuation of urban green areas. *Landscape and Urban Planning*, *66*, 35–41.
New York State Department of Education (2002). New York State school report cards for the 2000–2001 school year [data file]. Retrieved from http://www.p12.nysed.gov/repcrd2002/home.html
Palmquist, R. (2005). Property value model. In K. G. Mäler, & J. R. Vincent (Eds.), *The handbook of environmental economics: Valuing environmental changes* (pp. 785–786). Amsterdam: North-Holland Publishers.
Pendall, R. (2003). Sprawl without growth: The upstate paradox. Retrieved from http://www.brookings.edu/reports/2003/10demographics_pendall.aspx
Poudyal, N. C., Hodges, D. G., Tonn, B., & Cho, S. H. (2009). Valuing diversity and spatial pattern of open space plots in urban neighborhoods. *Forest Policy and Economics*, *11*(3), 194–201.
Quinlan, J. R. (2010). An overview of Cubist. Retrieved from http://www.rulequest.com/cubist-win.html
Ridker, R. G. (1967). *Economic cost of air pollution, studies in measurement*. New York, NY: Frederick A. Praeger Publishers.
Ridker, R. G., & Henning, J. A. (1967). The determinants of residential property values with special reference to air pollution. *Review of Economics and Statistics*, *49*(2), 246–257.
Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *The Journal of Political Economy*, *82*, 34–55.
Sander, H., Polasky, S., & Haight, R. G. (2010). The value of urban tree cover: A hedonic property price model in Ramsey and Dakota Counties, Minnesota, USA. *Ecological Economics*, *69*(8), 1646–1656.
Shultz, S. D., & King, D. A. (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., Poulos, C., & Kim, H. (2002). Treating open space as an urban amenity. *Resource and Energy Economics*, *24*, 107–129.
Troy, A., & Grove, J. M. (2008). Property values, parks, and crime: A hedonic analysis in Baltimore, MD. *Landscape and Urban Planning*, *87*(3), 233–245.
Tufféry, S. (2011). *Data mining and statistics for decision making*. Chichester, UK: Wiley & Sons Ltd.
Tyrväinen, L. (1997). The amenity value of the urban forest: An application of the hedonic pricing method. *Landscape and Urban Planning*, *37*(3–4), 211–222.
Tyrväinen, L., & Miettinen, A. (2000). Property prices and urban forest amenities. *Journal of Environmental Economics and Management*, *39*(2), 205–223.
U.S. Census Bureau. (2000). Onondaga County, New York [State & County QuickFacts Sheet for Onondaga County, Yew York]. Retrieved from http://quickfacts.census.gov
U.S. Census Bureau. (2002). U.S. Census TIGER/Line-file [data file]. Retrieved from http://www.census.gov/geo/www/tiger/tgrcd108/tgr108cd.html
U.S. Environmental Protection Agency. (2001). 2001 National Land Cover Data [data file]. Retrieved from http://www.epa.gov/mrlc/nlcd-2001.html
Walker, W. S., Kellndorfer, J. M., LaPoint, E., Hoppus, M., & Westfall, J. (2007). An empirical InSARoptical fusion approach to mapping vegetation canopy height. *Remote Sensing of Environment*, *109*, 482–499.
Waltert, F., & Schläpfer, F. (2010). Landscape amenities and local development: A review of migration, regional economics and hedonic pricing studies. *Ecological Economics*, *70*(2), 141–152.
Walton, J. T. (2008). Subpixel urban land cover estimation: Comparing Cubist, random forests, and support vector regression. *Photogrammetric Engineering and Remote Sensing*, *74*(10), 1213–1222.
Wu, J., Adams, R. M., & Plantinga, A. J. (2004). Amenities in an urban equilibrium model: Residential development in Portland, Oregon. *Land Economics*, *80*(1), 19–32.
Yu, D., & Wu, C. (2006). Incorporating remote sensing information in modeling house values: A regression tree approach. *Photogrammetric Engineering and Remote Sensing*, *72*(2), 129–138.