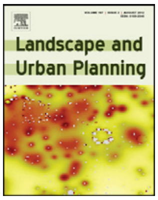
Contents lists available at SciVerse ScienceDirect

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;

$$P(A) = f(S, N, E) \tag{1}$$

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