Computers, Environment and Urban Systems 39 (2013) 81–92

Contents lists available at SciVerse ScienceDirect

Computers, Environment and Urban Systems

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

Boosting the predictive accuracy of urban hedonic house price models through airborne laser scanning

Marco Helbich [a,*], Andreas Jochem [a], Werner Mücke [b], Bernhard Höfle [a]

[a] Institute of Geography, University of Heidelberg, Germany
[b] Institute of Photogrammetry and Remote Sensing, Vienna University of Technology, Austria

ARTICLE INFO

Article history:
Received 10 September 2012
Received in revised form 4 January 2013
Accepted 5 January 2013
Available online 28 March 2013

Keywords:
Real estate
Airborne laser scanning
LiDAR
GIS
Solar radiation
Hedonic regression
Generalized additive model
Vienna (Austria)

ABSTRACT

This paper introduces an integrative approach to hedonic house price modeling which utilizes high density 3D airborne laser scanning (ALS) data. In general, it is shown that extracting exploratory variables using 3D analysis – thus explicitly considering high-rise buildings, shadowing effects, etc. – is crucial in complex urban environments and is limited in well-established raster-based modeling. This is fundamental in large-scale urban analyses where essential determinants influencing real estate prices are constantly missing and are not accessible in official and mass appraiser databases. More specifically, the advantages of this methodology are demonstrated by means of a novel and economically important externality, namely incoming solar radiation, derived separately for each flat. Findings from an empirical case study in Vienna, Austria, applying a non-linear generalized additive hedonic model, suggest that solar radiation is significantly capitalized in flat prices. A model comparison clearly proves that the hedonic model accounting for ALS-based solar radiation performs significantly superior. Compared to a model without this externality, it increases the model's explanatory power by approximately 13% and additionally reduces the prediction error by around 15%. The results provide strong evidence that explanatory variables originating from ALS, explicitly regarding the immediate 3D surroundings, enhance traditional hedonic models in urban environments.

1. Introduction

Real estate markets are constantly in motion, thus leading to an increased risk awareness by investors, mortgage lenders, etc. Accordingly, the predictive accuracy of economic models has gained much attention and has stimulated research (e.g. Basu & Thibodeau, 1998; Bateman, Jones, Lovett, Lake, & Day, 2002; Bourassa, Cantoni, & Hoesli, 2010; Brunauer, Lang, Wechselberger, & Bienert, 2010; Case, Clapp, Dubin, & Rodriguez, 2004; Dubin, Pace, & Thibodeau, 1999; Goodman & Thibodeau, 2003; Helbich, Brunauer, Hagenauer, & Leitner, 2013; Pace, 1998; Páez, Fei, & Farber, 2008). Hedonic price modeling (Rosen, 1974) is an extensively applied framework for mass appraisal and price index construction. These models can be improved in two ways: (a) Through novel estimation techniques (e.g. Brunauer et al., 2010; Koschinsky, Lozano-Gracia, & Piras, 2011) and (b) by ancillary structural, locational, and neighborhood variables on the basis of Geographic Information System (GIS) algorithms (e.g. Hamilton & Morgan, 2010), which have the potential to mitigate violations of model assumptions and advance model reliability. However, recent studies are limited in that they use the raster and 2D vector data model when computing GIS-based variables (e.g. Bin, Crawford, Kruse, & Landry, 2008; Bourassa, Hoesli, & Sun, 2004; Hamilton & Morgan, 2010; Kong, Yin, & Nakagoshi, 2007; Lake, Lovett, Bateman, & Day, 2000; Orford, 2010; Paterson & Boyle, 2002).

Nowadays, ALS – also referred to as airborne LiDAR – data are increasingly available because of steadily declining costs, particularly in urban environments. Since the proliferation and substantial advances in ALS as state-of-the-art technology for 3D topographic data acquisition (Vosselman & Maas, 2010), it appears that GIS-models to derive explanatory variables based on the raster or 2D vector data model have serious weaknesses. Instead of utilizing the full richness and high resolution of ALS technology, the data are aggregated to representations using single valued (elevation) functions such as digital elevation models (DEMs) resulting in a loss of information (Vosselman & Maas, 2010). In general, raster-DEMs are further differentiated into digital terrain models (DTMs) of the bare Earth without objects (e.g. vegetation and buildings) and digital surface models (DSMs), which include objects above the ground (Höfle & Rutzinger, 2011). The consideration of the upper hull (topography) of the surface is particularly appropriate if the location of interest is located in highly complex urban environments, characterized by sudden variations in heights, shadowing effects, eaves, building shapes, etc. (Hachema, Athienitis, &

* Corresponding author. Address: Institute of Geography, University of Heidelberg, Berliner Straße 48, 69120 Heidelberg, Germany.
E-mail address: helbich@uni-heidelberg.de (M. Helbich).

0198-9715/$ - see front matter
http://dx.doi.org/10.1016/j.compenvurbsys.2013.01.001

Fazioc, 2012; Jochem, Höfle, & Rutzinger, 2011; Jochem, Höfle, Rutzinger, & Pfeifer, 2009; Lukac, Zlaus, Seme, Zalik, & Stumberger, 2012). The modeled determinants depend on the scale of aggregation and are merely a rough approximation of reality, potentially resulting in counterintuitive signs of the estimated regression parameters of the hedonic model and hence leading to erroneous conclusions (e.g. Lake et al., 2000).

Hedonic theory assumes that all essential characteristics are considered in the hedonic equation, which is seldom fulfilled due to limited data availability (Hulton, 2003; McMillen, 2010), thus resulting in model misspecification (Can, 1992). The lack of data is particularly crucial in large-scale analysis. Indeed, when it comes to this analysis, essential determinants influencing house prices are not available in traditional databases actuated by traditional federal statistical offices and rating agencies. It may also be the case that they cannot be modeled by traditional GIS algorithms (Orford, 2010). Thus, one is faced with the omitted variable bias, which states that relevant variables are missing in the model, although they influence the price significantly. Such misspecification results in, for example, ordinary least square (OLS) estimates being biased and inconsistent (see Wooldridge, 2008). A possible solution is a physical inspection of each flat[1] by an appraiser, which is only viable for a small number of objects and is strongly limited by temporal and monetary constraints. Regardless of these limitations, the subjectivity and fuzziness of such appraisals remain a problem and result in an insufficient data quality. Besides, these indices based on *in situ* data acquisition are mostly of a nominal or an ordinal nature.

Both weaknesses relate to a lack of theoretical and empirical work. Therefore, the overall objective of this research is to explore the potential of ALS for housing studies and to enhance the current methodology of hedonic house price models by utilizing ALS data in a reliable and integrative way. Moreover, this paper demonstrates, through the modeling incoming solar radiation, that 3D ALS data provide precise and objective numeric indices which can be computed in a consistent, standardized, and transferable manner, thus enhancing the predictive power of hedonic models and simultaneously mitigating model misspecifications. The case study addresses the housing market segment of owner-occupied flats in the third district of Vienna, Austria, and tests the effect of ALS-based solar radiation on flat prices in a non-linear hedonic pricing model. The main hypothesis is that accounting for the complexity of urban areas in terms of incoming solar radiation for individual flats, results in more accurate price predictions. In detail, the research at hand addresses the following main research questions:

- Is ALS capable of improving the predictive accuracy of large-scale hedonic price models?
- Does solar radiation have significant explanatory power, and does it account for a higher explained deviance, as well as for most of the reduction of the unexplained variance, respectively, in comparison to a model without this externality? If this is the case, is this covariate linearly or non-linearly related to the transaction prices of flats in Vienna?

The remainder of this paper is organized as follows: Section 2 gives an overview of hedonic modeling and the first attempts made to utilize ALS data. Following this, Section 3 introduces a 3D GIS algorithm to derive the incoming solar radiation of individual flats and a non-linear hedonic model. The potential of this method is explored by means of owner-occupied flats in Vienna (Section 4). Empirical results are discussed in Section 5, before Section 6 summarizes the implications and suggests future research avenues.

[1] A flat represents a residential apartment in a multi-level housing structure.

2. Related work

2.1. Hedonic pricing theory

Real estate is usually treated as a composite commodity traded in bundles, and valued for its utility-bearing characteristics (Rosen, 1974). Hence, households value the characteristics of a good rather than the good itself. Because property is fixed in space, a household implicitly chooses a bundle of different goods and services by selecting a specific object (Malpezzi, 2003; Sheppard, 1997). Methodologically, this is represented by the hedonic price function, which emerges from the competitive bidding of buyers (Bin, Poulter, Dumas, & Whitehead, 2011). The equilibrium between supply and demand persists when households maximize their utility, limited by their social and economic constraints (Quigley, 1985). Thus, the hedonic equation determines the functional relationship between the real estate price and its characteristics in a particular market, typically estimated by a regression equation (Sheppard, 1997). Such a model regresses the value of the property on non-traded structural and neighborhood characteristics. Assuming *ceteris paribus* conditions, the estimated coefficients mimic the implicit prices of certain characteristics and report how the price changes when one of these characteristics changes (Wooldridge, 2008).

Two challenges arise during the empirical application of hedonic regressions (McMillen, 2010): (1) The specification of the functional form, and (2) the modeling of spatial effects. Firstly, the valid model specification is not guided by economic theory, permitting either a linear or a non-linear relationship (Rosen, 1974). Pace (1998, p. 77) establishes that an incorrectly chosen functional form results in "disastrous consequences for traditional estimators" and may itself cause a spatially correlated error term (McMillen, 2010). To deal with emerging non-linearities, it is common in practice to use higher order polynomials and a log–log or semi-log model specification, further mitigating difficulties with heteroscedasticity and outliers (Malpezzi, 2003). Augmenting predictors with polynomials as parametric components, as suggested by Stevenson (2004), results in multicollinearity problems. In addition, it also distorts the fit through unevenly distributed data, and is only suitable to model the global nature of the data (Dubin, 1998; Pace, 1998). Despite the above mentioned constraints, these parametric approaches are frequently applied (e.g. Goodman & Thibodeau, 2003), although their estimation is tedious as they necessitate knowledge about the "true" functional form in advance (Brunauer et al., 2010). Pioneering attempts by Halvorsen and Pollakowski (1979) promote the more flexible Box–Cox transformation, while Cassel and Mendelsohn (1985) find contradictory evidence that this does not necessarily result in more accurate estimations. Thus, a less restrictive and more rational approach for overcoming functional specification problems is to apply non-parametric or semi-parametric models where non-linearities might be expected, as advocated by Anglin and Gençay (1996), Mason and Quigley (1996), Pace (1998), Thorsnes and McMillen (1998), and Brunauer et al. (2010). This is even more appropriate when the effect of a certain covariate is entirely unclear (e.g. solar radiation). Thorsnes and McMillen (1998) argue that fully non-parametric approaches result in imprecise estimates, and thus advise semi-parametric models which offer functional flexibility where needed, while imposing linear restrictions where appropriate. Concerning prediction capabilities, Anglin and Gençay (1996), as well as Pace (1998), achieve higher accuracies using semi-parametric models, compared to their parametric counterparts. In this context, generalized additive models (Wood, 2000, 2006) are growing in popularity. These comprise a flexible model family and result in valid and – compared to e.g. non-parametric neural networks (Do & Grudnitski, 1992) – highly interpretable models when economic processes are exceedingly complex, possibly non-linear, *a priori*