Modeling Local Property Price Response to Building Replacement

Qing Liu[1]
[1]Rhode Island School of Design

Abstract

We describe and implement a technique for modeling property price spatially according to building types in the local vicinity. This technique enables us to model what happens to local property prices when one kind of property is replaced with another, relevant to urban planning and city simulation. For a case study, we examine properties in the city of Providence, Rhode Island.

1 Introduction

In this paper, we are motivated by the question of what happens to property prices in the local vicinity when a property of a given type (such as a restaurant) is replaced with another (such as an apartment building). We are motivated to answer this question because the authors are in the process of building a city simulation game that models how various components of a city interact.

There is a rich literature on modeling property prices, but thus far, we are not aware of any papers that explicitly describe how to model these kinds of spatial building type-specific 'replacement' effects. Many previous techniques have been proposed to model individual property prices, including hedonic models that make use of individual attributes of the property (such as lot size, square footage, number of bedrooms, distance to landmarks) (Dubin, 1998; Bourassa et al., 1999); some of these models also include spatial attributes (Liu et al., 2016).

To the best of our knowledge, no previous work has explicitly described how a property price responds when another property in the local vicinity is altered in some way (in this work, we focus specifically on building type). In order to achieve, this, we propose to use the coefficients of a linear regression model augmented with building type-specific spatially lagged exogenous regressors. We show in a case study on Providence, Rhode Island property assessments that these coefficients are qualitatively sensible.

2 Spatial Regression Model

We first use a standard ordinary least squares linear regression model augmented with spatially lagged exogenous regressors to model property price:

$$P_i = \alpha + \beta X_i + \delta \sum_j w_{ij} V_i + \epsilon_i$$

.

The price of a property i is P_i, which is a linear function of a set of other attributes collectively called X_i; another set of attributes V_i that we refer to as *spatial variables*; estimable parameters α, β, and δ; and a noise term ϵ_i. The term $\delta \sum_j w_{ij} V_i$ represents the spatially lagged components of the linear model; w_{ij} is the ij-th cell of a spatial weights matrix W. Intuitively, the term $\delta \sum_j w_{ij} V_i$ captures the average value of V_i in the surroundings of location i, also known as the spatial lag of V_i. Spatially lagging the variables does not violate any assumptions on which ordinary least squares relies, so these spatial lag variables can be introduced exogenously. We credit the description of this model to (Arribas-Bel, 2016). The parameters α, β, and δ can be estimated from property data and then used to compute the effect of replacing one kind of property with another.

Specifically, in our setting, for the spatial matrix W, we implement an adjacency of matrix of the k nearest neighbors, where k is the number of neighbors and the distance is defined as Euclidean norm with respect to the longitude / latitude centroid coordinate of a given property. In this work, we select $k = 30$.

For our spatial variables V_i, we choose to use binary indicator variables that encode the building type. For a given set of building types $\mathbf{T}$ composed of types $t_1, \ldots, t_n$, the indicator variable I_k is 1 if the property is of building type k and 0 otherwise. V_i is composed of n binary building type indicator variables (so there are n building type annotations per property). Consequently, each of the components of δ (consisting of n esimated coefficients, one per building type) can be interpreted as the weight of a spatial effect on the property price with respect to each of the n building types.

So the components of $\delta = [\delta_1, \ldots, \delta_n]$ (the coefficients of the spatial terms we are interested in) present a path towards computing price changes of buildings in the vicinity defined by the k nearest neighbors. Suppose a property of type q is replaced with a property of type r. Then the price change of a building in the vicinity corresponds to $\delta'_r - \delta'_q$ (the primes denote appropriate normalization for a single property). The subtraction of δ'_q indicates that a property of type q disappeared, while the addition of δ'_r indicates that a property of type r appeared.

We considered other models, such as those including endogenous variables (Arribas-Bel, 2016), along with other spatial regression models implemented in (Rey and Anselin, 2007), but it was not clear to us how to recover and consequently manipulate parameters specific to building type. In the future, we would like to further investigate other suitable models.

3 Case Study: Providence Properties

We scraped data from a 2019 Providence property assessment database (of Providence, a). The relevant attributes we scraped for each property included the building price (not including land price), living area, year built, and building type.

In order to extract the longitude / latitude locations of each property, we joined our scraped data with an existing data set of Providence parcel boundaries (of Providence, b) recorded in 2017. After applying some standard data cleaning techniques, including removal of outliers and adjustment of negative prices, we accumulated a data set of 36151 Providence properties.

There were approximately a hundred building types for all Providence properties appearing in the raw data; we binned these into a set of 14 distinct categories in order to simplify our model (otherwise we would require approximately a hundred indicator variables in V_i). Specifically, the set of categories was: 1. residential, 2. mixed (corresponding to mixed use residential and commercial properties), 3. retail, 4. apartments, 5. industrial, 6. office, 7. school, 8. auto shop, 9. religious, 10. food, 11. charitable organizations (including non-profits), 12. government, 13. medical (including hospitals and medical offices), and 14. gas mart.

We added two additional features to the dataset of 36151 properties: the number of trees within 500 feet of the property in 2016 and the median income according to Providence census tracts in 2017. We credit (Dillon, 2016) for the data set of trees and credit (Berke, 2019) for the census tract income processing procedure.

We now describe the specific form of the model we use in Section 2. We employ two models. Both models encode X_i as the living area (in square footage), the age of the property (normalized year), the number of trees within approximately 500 feet of the property, and the median income according to census tract. We chose these regressors for the base linear model because they represent a mix of social, environmental, and property-specific attributes that seem highly relevant to property price.

The two models differ in V_i; the first version encodes all indicator variables corresponding to each of the 14 building types except residential, whereas the second version only contains an indicator variable corresponding to residential. Empirically we found that this separation was necessary, because introducing the indicator variable corresponding to residential caused the linear model computation to be numerically unstable. This may be because the frequency of residential properties far outnumbers all other kinds of property; our compromise was to introduce a second model only containing this indicator variable. After estimation, we found that four property types had somewhat high p-values, indicating questionable statistical significance: mixed, retail, food, government, and gas mart, with p-values of 0.59, 0.07, 0.08, 0.77, and 0.66, respectively. Price changes involving these types should be interpreted with a grain of salt. All other p-values were below 0.05. The R^2 value for the residential model is 0.5211, whereas the R^2 value for the other model is 0.5478: both R^2 values are relatively high for the social sciences.

After normalizing δ (dividing each coefficient from the linear model by k), the price change according to replacement for each pair of property types was computed. These price changes are presented in Table 1 and correspond to price changes for the 'average' property in Providence. A single cell represents the price change for replacing a property type on the left hand side with a property type on the top side.

Qualitative trends can be observed from this table. Reading down the columns, replacing a building with a school, religious building, charitable building, or medical building often increases the price of the property, whereas replacing a building with an industrial warehouse, auto shop, or gas mart often results in decreasing the price of the property. The building replacements that cause negative price changes all seem sensible, whereas for the positive price changes, school in particular makes sense: property prices often increase around educational institutions. We found that price increases around charitable, religious, and medical buildings were interesting and not entirely expected; these findings indicate that our model can discover trends we previously did not know about.

4 Conclusion and Future Work

We described and implemented a technique for modeling property price spatially according to building types in the local vicinity, which ultimately enables us to model what happens to local property prices when one kind of property is replaced with another. To the best of our knowledge, we are not aware of previous work in the literature examining the effects of property replacement on local price. For a case study, we examined properties in the city of Providence, Rhode Island and described some qualitative phenomena. All of our code and data processing scripts are available on request.

For future work, we plan to explore models that more faithfully represent the price landscape of properties, such as modeling of non-convex, non-continuous price regions as in (Liu et al., 2016). Another direction is to incorporate more fine-grained property-specific information for the computation of price changes: our current normalization scheme selects the 'average' property to be the median value in the data set and scales price changes against the median, whereas in reality, the situation is much more nuanced. We also plan to investigate data sets that are more in sync: we currently use three data sets spread across 2016, 2017, and 2019, but it would be helpful to use data sets all synced against the same year.

References

Dani Arribas-Bel. 2016. Geographic Data Science with PySAL and the pydata stack.

Alex Berke. 2019. Income, Race, Bikes.

S C Bourassa, F Hamelink, M Hoesli, and B D MacGregor. 1999. Defining housing submarkets. *Journal of Housing Economics*, 8:160–183.

James Dillon. 2016. Providence Tree Visualization by James Dillon PhD.

Robin A. Dubin. 1998. Predicting house prices using multiple listings data. *The Journal of Real Estate Finance and Economics*, 17(1):35–59.

Bang Liu, Borislav Mavrin, Di Niu, and Linglong Kong. 2016. House price modeling over heterogeneous regions with hierarchical spatial functional analysis. In *IEEE 16th International Conference on Data Mining, ICDM 2016, December 12-15, 2016, Barcelona, Spain*, pages 1047–1052.

City of Providence. a. City of providence online assessment database. `http://gis.vgsi.com/providenceri/`. Accessed: 2019-11-17.

City of Providence. b. Providence parcel boundaries 2017. `https://catalog.data.gov/dataset/providence-parcel-boundaries-2017`. Accessed: 2019-11-17.

Sergio J. Rey and Luc Anselin. 2007. PySAL: A Python Library of Spatial Analytical Methods. *The Review of Regional Studies*, 37(1):5–27.

	residential	mixed	retail	apartments	industrial	office	school	auto shop	religious	food	charitable	gov	medical	gas mart
residential	0	+4218	+8131	+10018	-18140	+11814	+99859	-18392	+50392	+11753	+23729	+3977	+16122	-301
mixed	-4218	0	+3913	+5799	-22358	+7596	+95640	-22610	+46174	+7534	+19511	-242	+11904	-4519
retail	-8131	-3913	0	+1886	-26272	+3683	+91727	-26523	+42261	+3621	+15597	-4155	+7990	-8432
apartments	-10018	-5799	-1886	0	-28158	+1797	+89841	-28410	+40375	+1735	+13711	-6041	+6104	-10318
industrial	+18140	+22358	+26272	+28158	0	+29954	+117999	-252	+68532	+29893	+41869	+22117	+34262	+17840
office	-11814	-7596	-3683	-1797	-29954	0	+88044	-30206	+38578	-62	+11914	-7838	+4308	-12115
school	-99859	-95640	-91727	-89841	-117999	-88044	0	-118251	-49466	-88106	-76130	-95882	-83737	-100159
auto shop	+18392	+22610	+26523	+28410	+252	+30206	+118251	0	+68784	+30145	+42121	+22369	+34514	+18091
religious	-50392	-46174	-42261	-40375	-68532	-38578	+49466	-68784	0	-38640	-26663	-46416	-34270	-50693
food	-11753	-7534	-3621	-1735	-29893	+62	+88106	-30145	+38640	0	+11976	-7776	+4369	-12053
charitable	-23729	-19511	-15597	-13711	-41869	-11914	+76130	-42121	+26663	-11976	0	-19752	-7607	-24029
gov	-3977	+242	+4155	+6041	-22117	+7838	+95882	-22369	+46416	+7776	+19752	0	+12145	-4277
medical	-16122	-11904	-7990	-6104	-34262	-4308	+83737	-34514	+34270	-4369	+7607	-12145	0	-16422
gas mart	+301	+4519	+8432	+10318	-17840	+12115	+100159	-18091	+50693	+12053	+24029	+4277	+16422	0

Table 1