Hedonic residential property price estimation using geospatial data: a machine-learning approach

Paulo Picchetti

paulo.picchetti@fgv.br

Instituto Brasileiro de Economia/Escola de Economia de São Paulo
Fundação Getulio Vargas - Brazil

April 24, 2017

Abstract

The issue of heterogeneity among samples of dwellings over time for calculating residential property price index is well known. One of the main approaches to circumvent it necessarily involves estimating a hedonic price model, which seeks to explain prices by means of a set of observable covariates related to intrinsic characteristics of homes. Location is among the fundamental characteristics within this set of variables, and recent developments in computation have made it easy to associate exact geographic coordinates to each record in a dataset of homes. Consideration of this information involves flexibility in the functional form of the estimated hedonic models. This paper contributes to this context, by considering an increasingly adopted estimation algorithm originated in the fast growing machine-learning literature.

JEL: C43, E01, E31, R31 **Keywords:** Housing Market, Price Index, Hedonic Models, Geospatial Data, gradient boosting machine.

1 Introduction

Constructing a price index for residential properties involves the well-known challenges of large heterogeneity among properties, and sparse transaction data for particular properties. The hedonic approach aims to circumvent these limitations allowing the comparison between heterogeneous dwellings by attributing prices to their individual observable characteristics. In this paper, we implement a hedonic estimation for these prices using the gradient boosting algorithm, which has been receiving a lot of attention in the fast-growing statistical learning literature.

Essentially, this approach imputes prices for individual dwellings by learning the relationship between observed prices and characteristics from a representative sample of houses in a very flexible, non-parametric form. A large number of regression trees is estimated sequentially for the original data set, and the resulting residuals obtained by comparing observed and predicted values, based on a set of observable characteristics. Finally, each estimated regression tree receives a weight proportional to its predictive accuracy, and final predictions are obtained and subjected to cross-validation measures.

Imputation of these predictions for the observed characteristics of dwellings in different points in time results in the necessary information for the chosen price-index formula. Information on both prices and characteristics are provided by a large data set of appraisals, which are required by prudential regulation as part of loan agreements in Brazil. The proposed price index is calculated from more than 1.5 million appraisals on a monthly frequency, starting in January 2014. Among the various characteristics present in the appraisals, such as construction area, number of bedrooms, age, etc. there is also information on services available in the neighborhood and general common facilities for apartments.

Each address is mapped into the corresponding latitude/longitude, which serves as very precise geographical information for the learning algorithm. The flexible nature allowed by this algorithm for the relationships between characteristics and prices result in a detailed map of values controlling for every other characteristic, for nine of the major cities in Brazil. For each of these cities, the index can be computed hierarchically, first for the whole city and then for particular sub-regions of interest.

2 Gradient Tree Boosting Algorithm

Generic gradient boosting at the m-th step fits a decision tree $h_m(x)$ to pseudo-residuals. Let J_m be the number of its leaves. The tree partitions the input space into J_m disjoint regions $R_{1m}, \ldots, R_{J_m m}$ and predicts a constant value in each region. Using the indicator notation, the output of $h_m(x)$ for input x can be written as the sum: $h_m(x) = \sum_{j=1}^{J_m} b_{jm} I(x \in R_{jm})$, where b_{jm} is the value predicted in the region R_{jm}. In case of usual regression trees, the trees are fitted using least-squares loss, and so the coefficient b_{jm} for the region R_{jm} is equal to just the value of output variable, averaged over all training instances in R_{jm}.

Friedman(2002) proposes to modify this algorithm so that it chooses a separate optimal value γ_{jm} for each of the tree's regions, instead of a single γ_m for the whole tree. He calls the modified algorithm "TreeBoost". The coefficients b_{jm} from the tree-fitting procedure can be then simply discarded and the model update rule becomes:

$$F_m(x) = F_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm}),$$

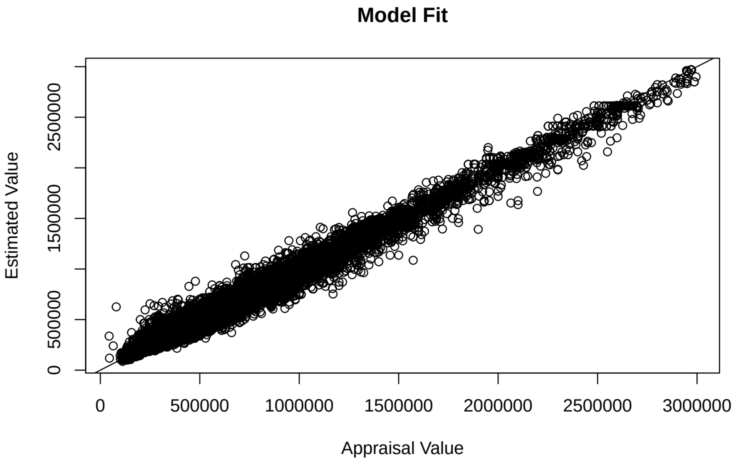
$$\gamma_{jm} = \arg\min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, F_{m-1}(x_i) + \gamma).$$

3 The Dataset

While data on actual transaction prices is the ideal information for calculating a resdidential price index, it is not realistically available on a regular frequency for Brazil. However, prudential regulation on banks establishes that loan operations should provide an independent appraisal of the price for the dwelling being financed. Since this appraisals must be based on a comprehensive set of characteristics for each home, they provide the necessary information for estimating hedonic price models, which in turn circumvent the issue of heterogeneity of houses across time. The dataset considered here contains around 5k monthly appraisals for houses and apartments in the city of São Paulo, from january 2014 through march 2017. At each month, the model is estimated using a 12-month moving window including month dummies.

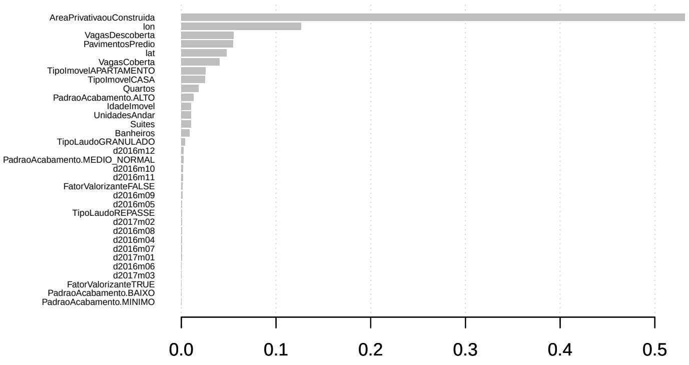
4 Results

The estimation process uses cross-validation within the sample for tuning hyper-parameters of the boosting regression tree, which allows for an ideal compromise between minimizing mean squared predictions errors and overfitting.



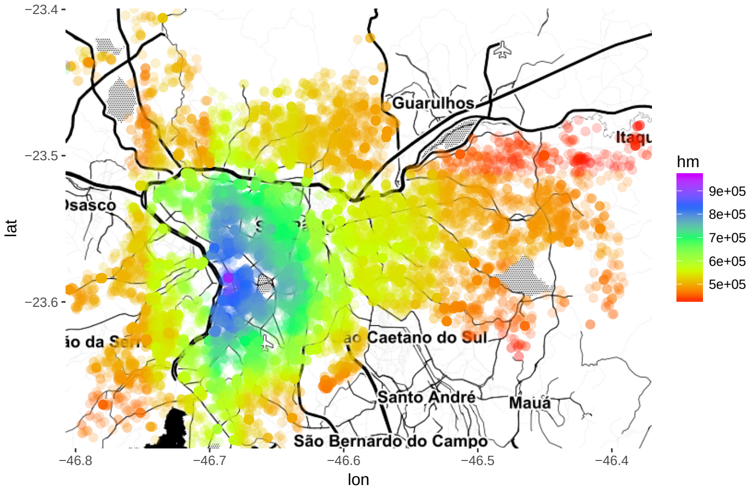
The results show a very good fit, with an out-of-sample root mean squared error within 10pct of actual (appraisal) prices. Some outliers can be seen in the residuals, which are checked for measurement and coding errors, and eventually discarded from the dataset for the calculation of the actual price index.

The relative importance of each covariate in the model for reducing total squared errors is best seen on the graph below.



Lat/lon variables are among the most relevant, and their interaction across levels of the regression trees allow a flexible estimation of a response surface in space.

The heatmap below simulates housing prices using averages for every covariate in the model, except for lat/lon variables.



The flexibility of the estimated relationship between location and prices naturally allows for incorporating the effects of spatial non-stationarity, anysotropy and discontinous jumps.

5 Discussion

Prediction accuracy for house prices using geospatial data along with their intrinsic characteristics in a flexible manner justifies considering machine-learning algorithms such as gradient tree boosting for constructing residential price indexes. Exact lat/lon information allows for a much finer resolution of spatial effects when compared to aggreation over arbitrary regions, such as the ones defined by administrative or postal code motivations. The results of the model can then be applied to any of the traditional index formulas such as time dummies or imputation. In the case of locations without observed values in particular time periods, interpolation based on exact neighborhood values should provide much better approximations. Two further extensions are worth considering. First, spatio-Temporal estimation and smoothing using machine-learning algorithms should complement exhisting approaches such as spatio-temporal kriging. Second, Bayesian hierarchical estimation [Gelman and Hill (2000)] can be applied to the hyperparameters of the gradient tree boosting algorithm considered here.

References

Edelstein, R. and Quan, D. (2006), *How Does Appraisal Smoothing Bias Real Estate Returns Measuerment ?*, The Journal of Real Estate Finance and Economics, Springer.

Fisher, J., D. Geltner, and R. Webb. (1994). *Value Indices of Commercial Real Estate: A Comparison of Index Construction Methods*, Journal of Real Estate Finance and Economics 9, 137-164.

Friedman, J. (2002): Stochastic gradient boosting, Computational Statistics and Data Analysis, Vol.38, issue 4.

Gelman, Andrew and Hill, Jennifer (2007), *Data Analysis using Regression and Multilevel/Hierarchical Models* Cambridge University Press.

Hill, R. and Scholz, M. (2013): Incorporating Geospatial Data into House Price Indexes: A Hedonic Imputation Approach with Splines, mimeo presented at the Ottawa Group Meeting.

Geltner, D., and W. Goetzmann. (2000). *Two Decades of Commercial Property Returns: A Repeated- Measure Regression-Based Version of the*

NCREIF Index, Journal of Real Estate Finance and Economics 21(1), 5-21.

IPD Australia (2009), *Methodology changes to the PCA/IPD Australia Property Index*, IPD Australia.

McAllister, P., A. Baum, N. Crosby, P. Gallimore, and A. Gray. (2003). *Appraiser Behaviour and Appraisal Smoothing: Some Qualitative and Quantitative Evidence*, Journal of Property Research 20(3), 261Y-80.