Residential price indices using different sources of information

Paulo Picchetti (FGV)

May 2019

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, among which location is of fundamental importance. In the process of estimating hedonic effects, data availability, both in terms of number of observations and characteristics of dwellings, plays a crucial role in the robustness of results. This paper addresses the issue of utilizing different sources of information, which differ on the relative strengths of the information provided, by means of a statistical spatial model.

JEL: C43, E01, E31, R31

Keywords: Housing Market, Price Index, Hedonic Models, Geospatial Data, Spatial Statistics, Spatial Misalignment.

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. 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.

For example, following Eurostat(2013), given $\hat{p}_n^0 = \hat{\beta}_0^0 + \sum_{k=1}^K \hat{\beta}_k^0 z_{nk}^0$, the hedonic double-imputation Laspeyres Price Index is

$$P_{HDIL}^{0t} = \frac{\hat{\beta}_0^t + \sum_{k=1}^K \hat{\beta}_k^t \bar{z}_k^0}{\hat{\beta}_0^0 + \sum_{k=1}^K \hat{\beta}_k^0 \bar{z}_k^0}$$

With the observed (period t) prices replaced by their model-based predictions $\hat{p}_n^t = \hat{\beta}_0^t + \sum_{k=1}^K \hat{\beta}_k^i z_{nk}^t$, the hedonic double imputation Paasche Price Index is:

$$P_{HDIP}^{0t} = \frac{\hat{\beta}_0^t + \sum_{k=1}^K \hat{\beta}_k^t \bar{z}_k^t}{\hat{\beta}_0^0 + \sum_{k=1}^K \hat{\beta}_k^0 \bar{z}_k^t}$$

And the hedonic double imputation Fisher Index is found by taking the geometric mean of the Laspeyres and Paasche Indices:

$$P_{HDIF}^{0t} = \left[P_{HDIL}^{0t} P_{HDIP}^{0t}\right]^{1/2}$$

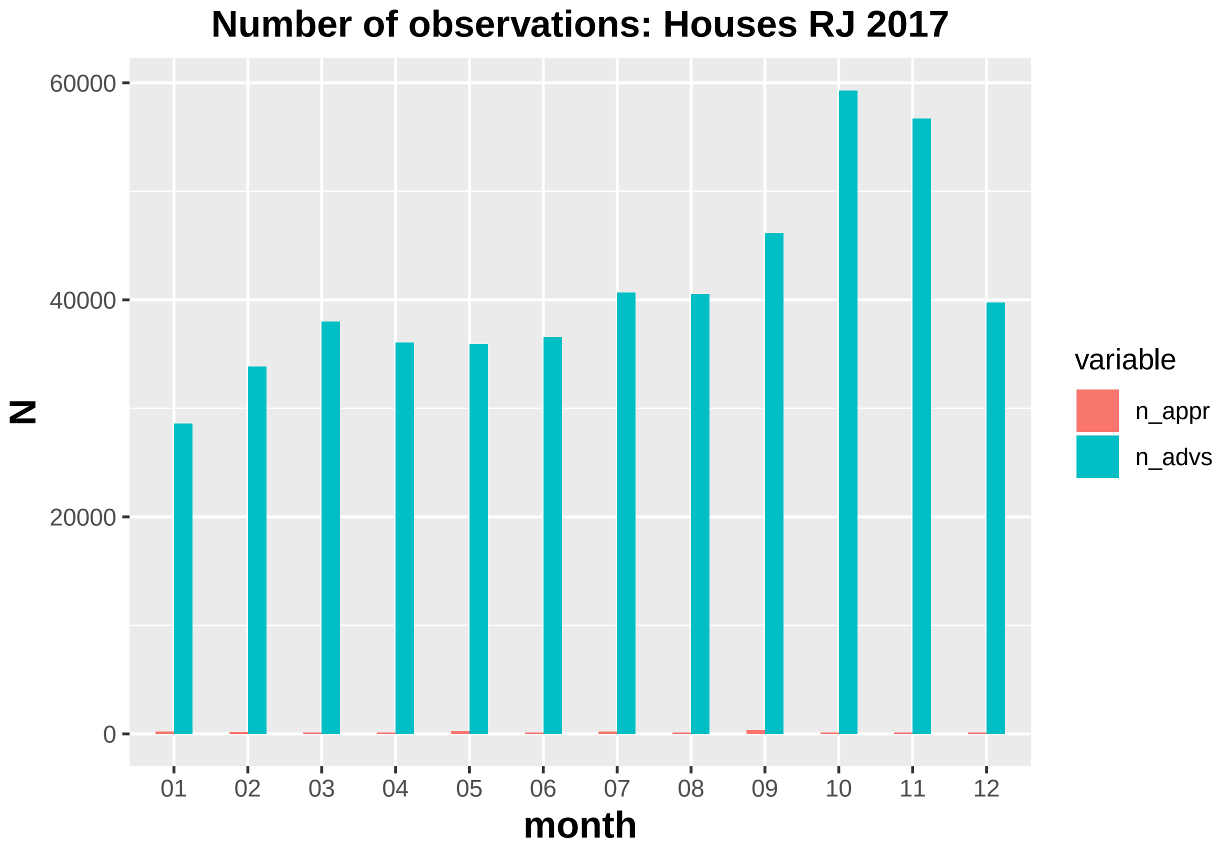
In what follows, we propose a methodology for estimating the necessary quantities for the above formulas leveraging the complementary information from different data sources.

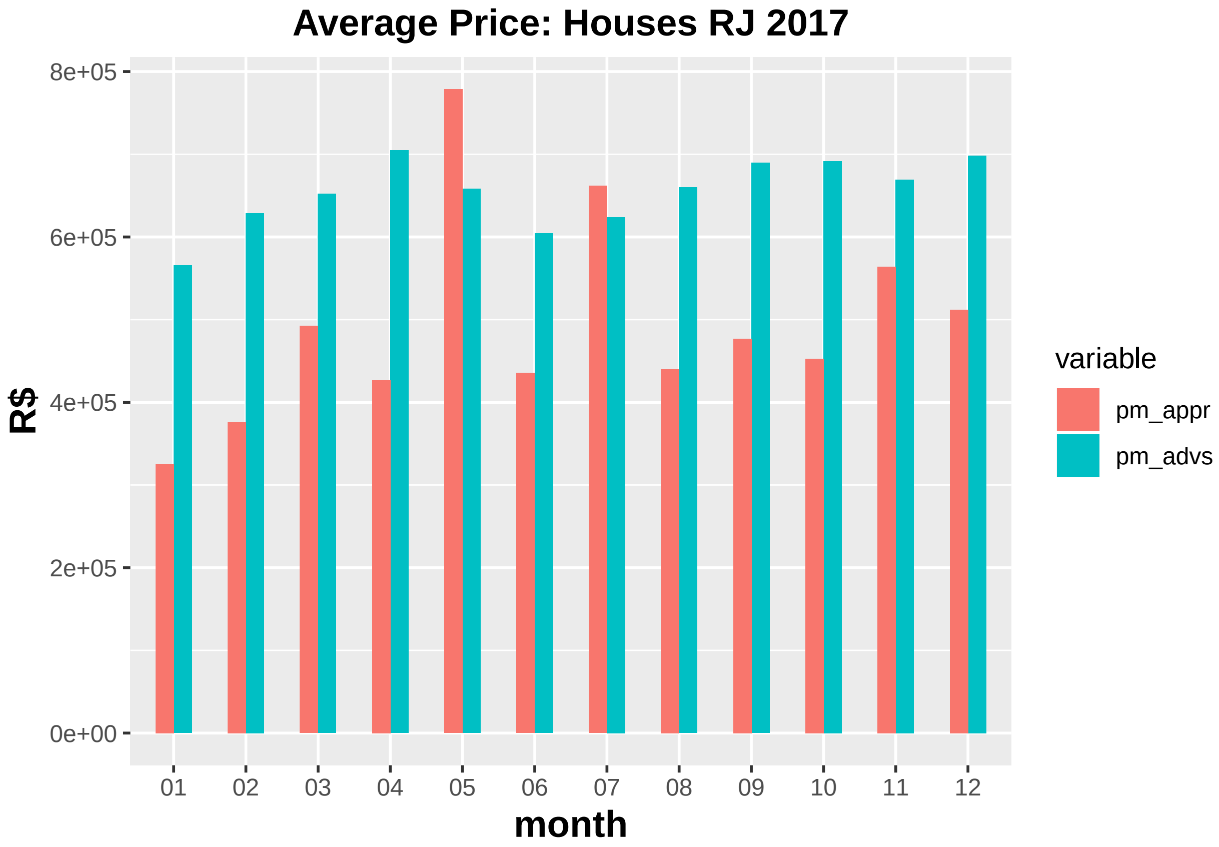
Data

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. This information set is very rich on the characteristics of each dwelling, but is limited to loan operations in each month, meaning that the number of data points as well their geographical coverage are somewhat limted.

On the other hand, advertisments provide a much larger dataset consistently throughout time. However, the information on the characteristics of the advertised units is considerably smaller than the one contained in teh appraisals backing the financing operations, while asking prices do not approximate actual transaction prices as well.

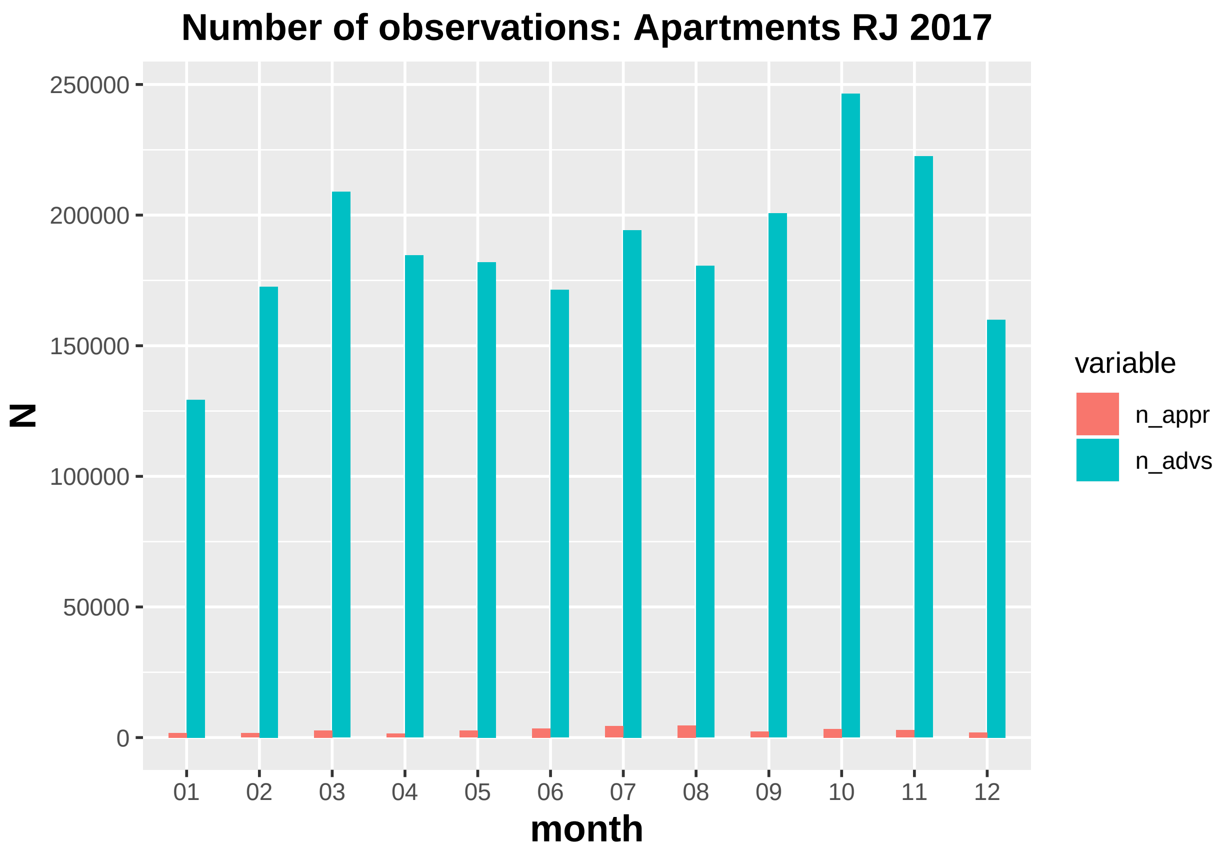
In this paper, we analyze data for the city of Rio de Janeiro. Below, we compare the absolute number of observations for houses on both datasets for each month in 2017, as well as their average prices.

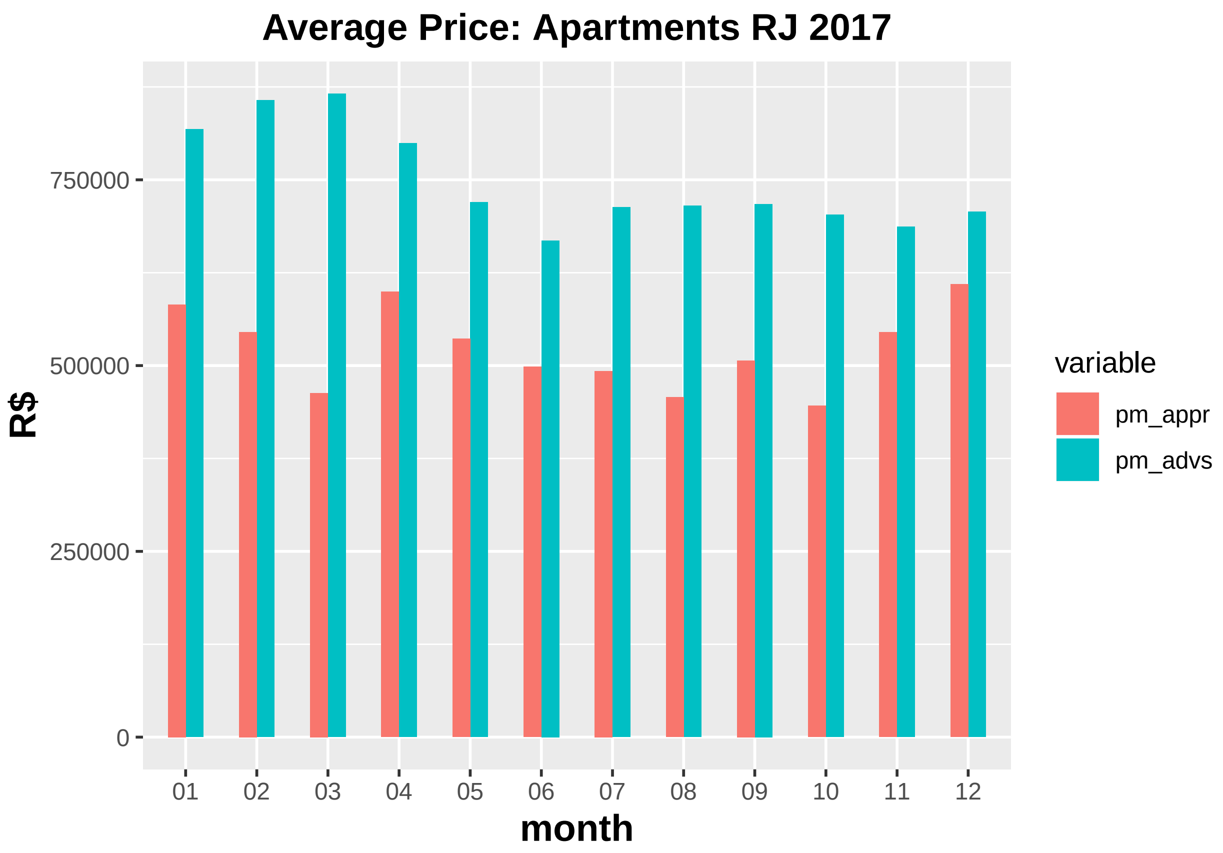




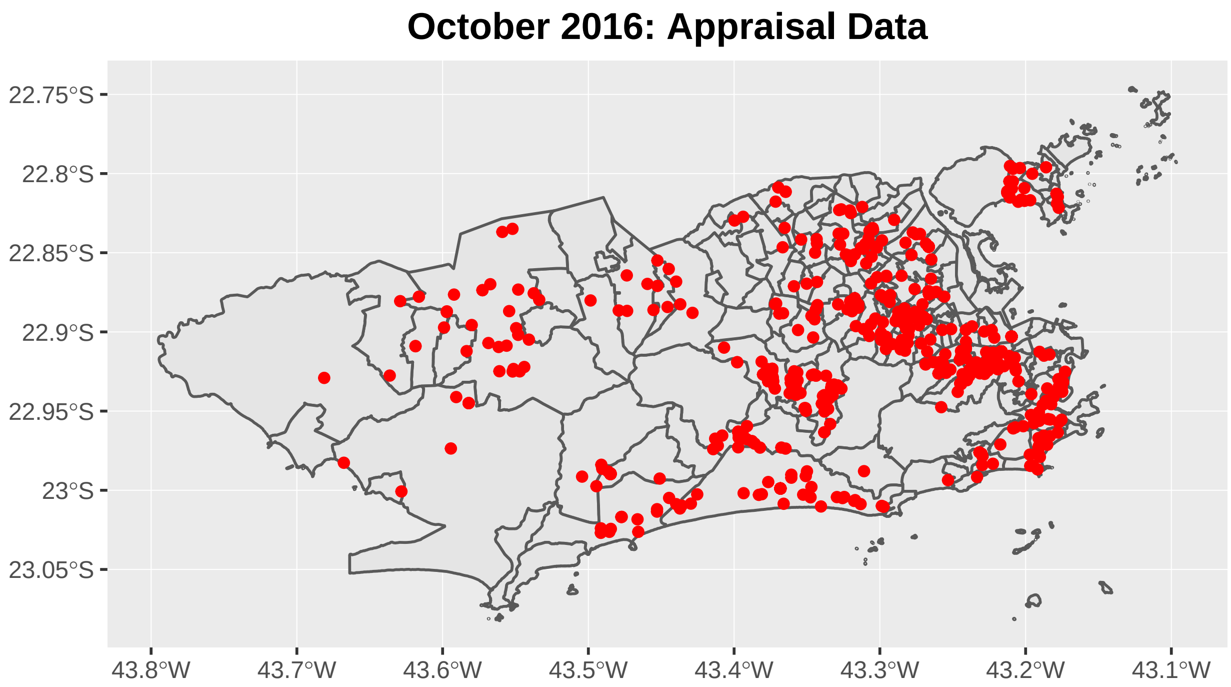
The number of observations of the ads dataset dwarfs the one from the appraisals dataset. It is interesting to observe that mean advertised prices are, as expected, above appraisal prices for almost every period, but also that the difference among them is not constant across these periods, which indicates considerable heterogeneity between the datasets, both within and between months.

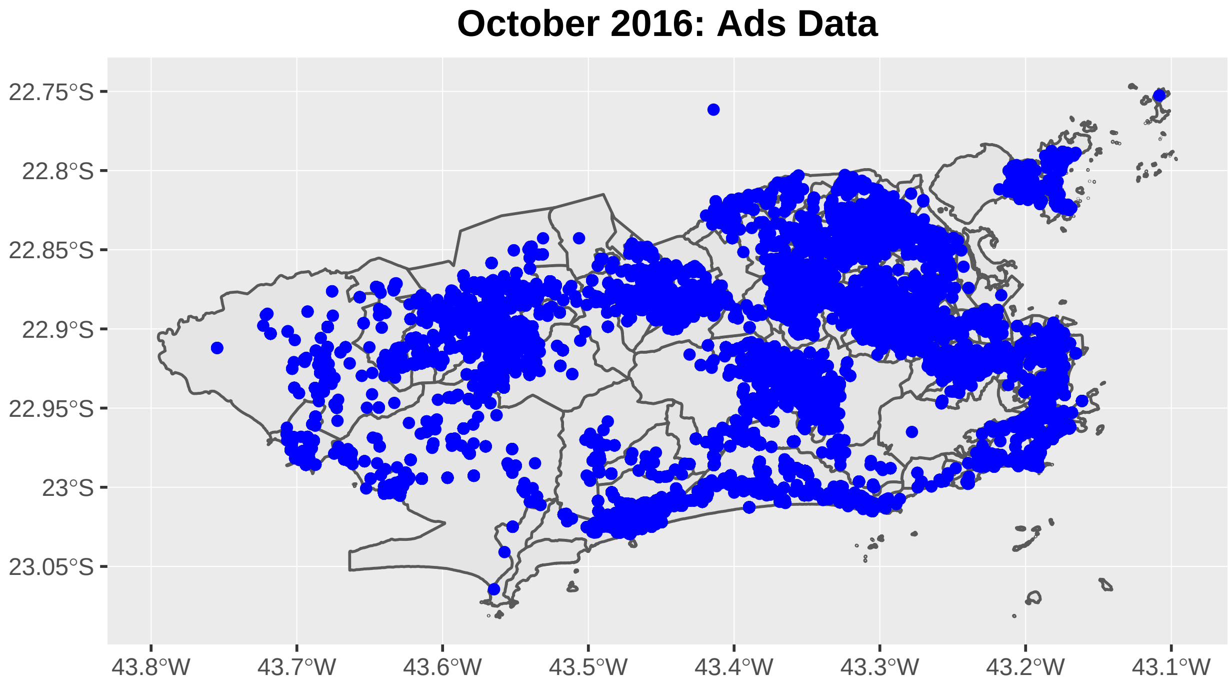
The same patterns appear when we look at the graphs for apartments only.


Number of observations: Apartments RJ 2017
N
250000
200000
150000
100000
50000
0
01
02
03
04
05
06
07
08
09
10
11
12
month
variable
n_appr
n_advs

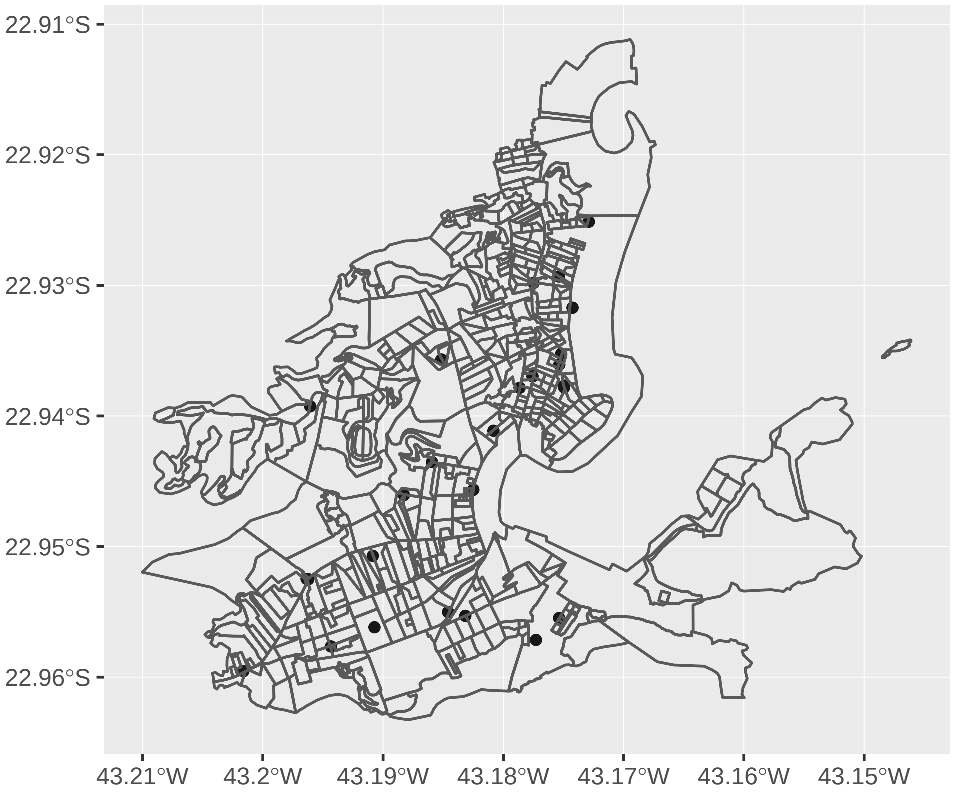


Looking at just the data from October 2016, we see that the differences in number of observations from both datasets also translates into very distinct regional coverage.

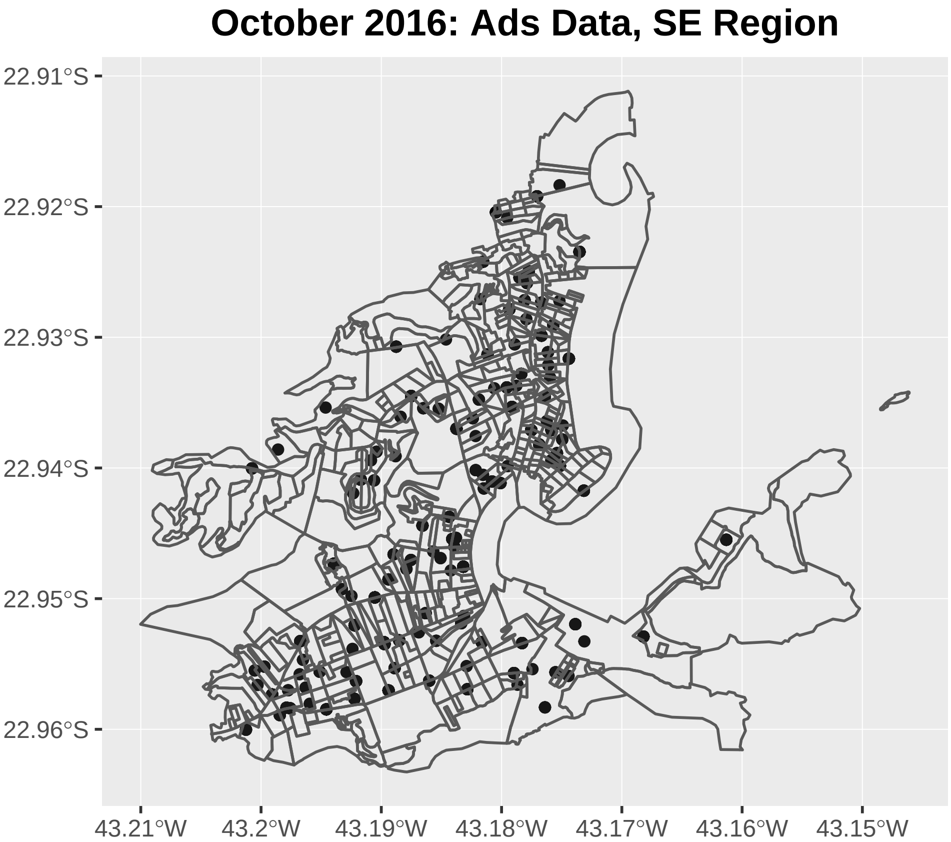

October 2016: Appraisal Data
22.75°S
22.8°S
22.85°S
22.9°S
22.95°S
23°S
23.05°S
43.8°W
43.7°W
43.6°W
43.5°W
43.4°W
43.3°W
43.2°W
43.1°W



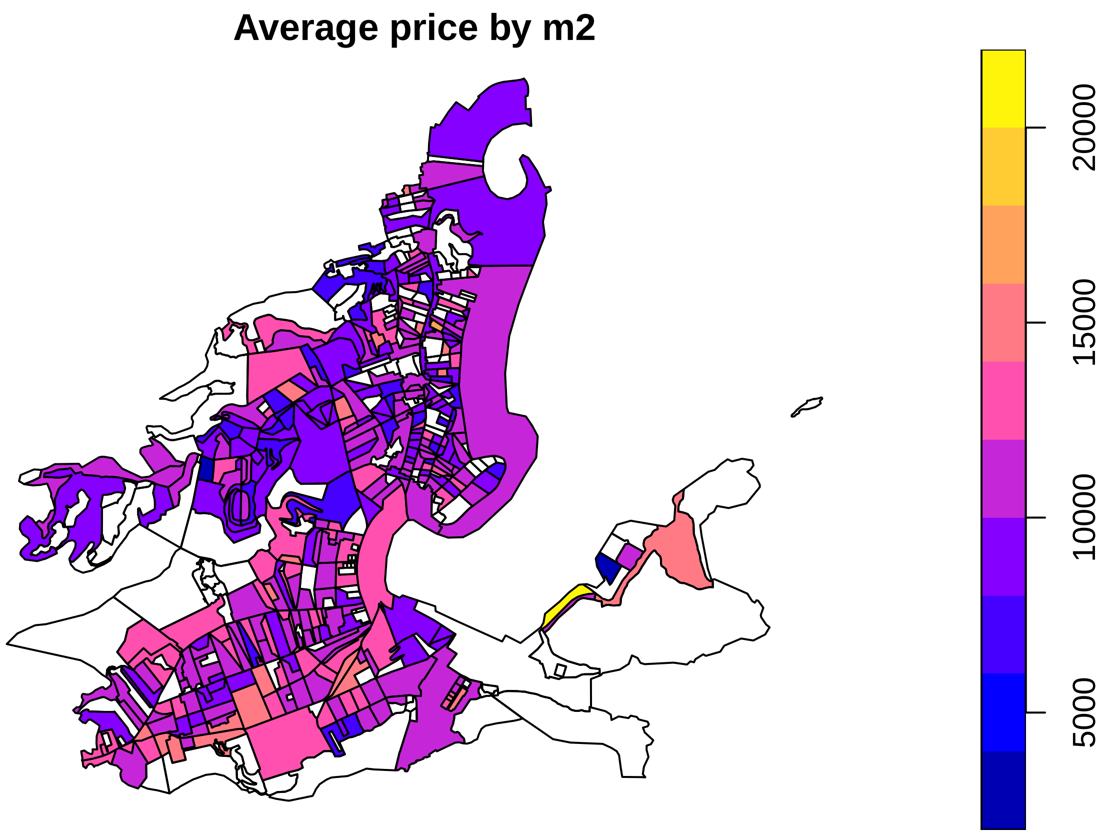
These differences become more specific when we zoom in the southeast region of the city.

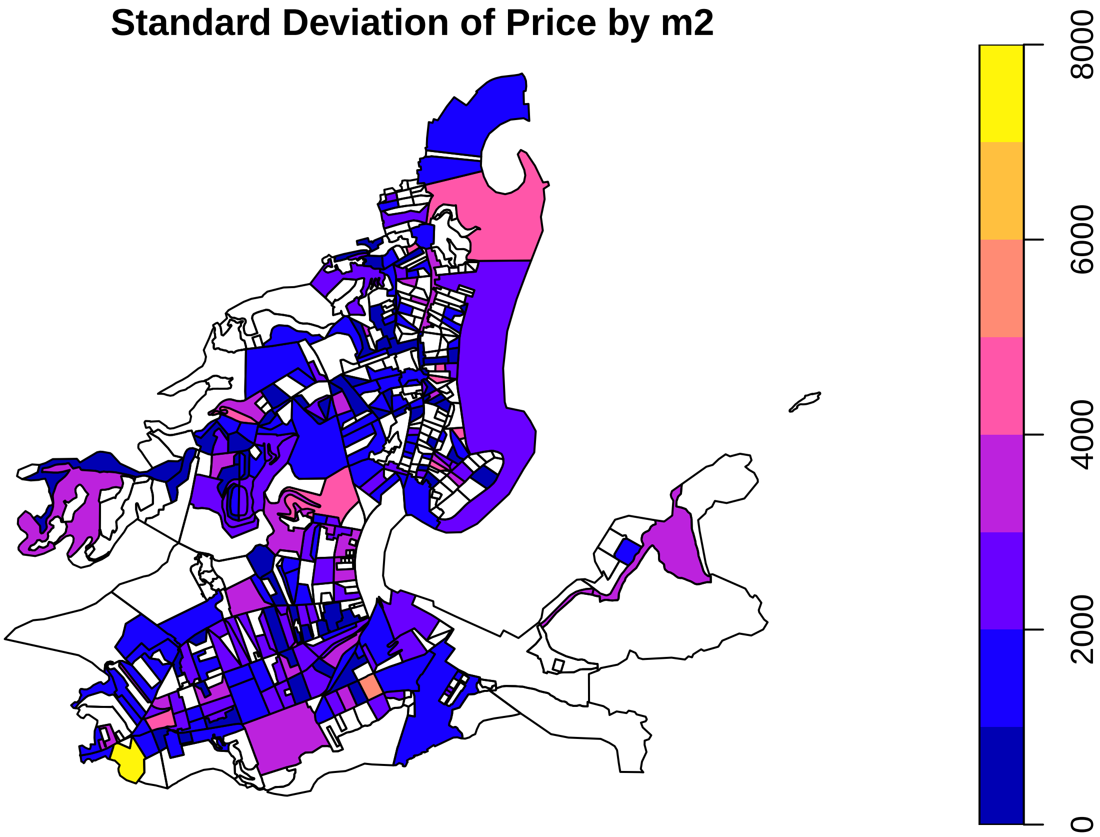


October 2016: Appraisal Data, SE Region



One could follow the strategy of aggregating individual dwelling into regions to mitigate the lack of observations and regional dispersions. Neighborhoods are too large to achieve a desired precision, but looking at census sectors we note that there is still considerable heterogeneity in prices. For instance, aggregating prices over census sectors in the southeast region of the city:


Average price by m2
20000
15000
10000
5000



Therefore, keeping the strategy of modeling individual price dwellings for estimating the hedonic price models, we propose to combine both sources of information, leveraging their relative strenghtnesses by means of a statistical spatial model. Central to the idea of this model is the strong correlation between appraisal and advertisement prices on one hand, and of prices across space on the other. The model adopted deals with the fact that prices on both datasets are correlated among themselves, but usually not measured in the same locations at each point in time.

Data Misalignment Model

The basic model (Banerjee et alli (2015)) is

$$Y(s) = X(s)\beta + \omega(s) + \epsilon(s)$$

where $Y(s) = [y_1(s), y_2(s)]$ is a vector of measurements (prices) for each dataset, in locations s. $\epsilon(s)$ represents measurement error, assumed distributed as gaussian, with a diagonal covariance matrix with parameters Ψ_1, Ψ_2. $X(s)\beta$ is the deterministic effect of each dwelling's attributes on the prices. Non-deterministic spatial effects are represented by a Gaussian Process $\omega(s)$, with spatial memory parameters ϕ_1, ϕ_2. While these parameters capture the spatial autocorrelation of each dataset, when combining these datasets we have an additional parameter measuring the correlation among the values in the datasets, ρ. The combinations among the parameters ϕ_i and ρ_i define a cross spatial covariance matrix.

The $X(s)$s represent the intrinsic characteristics of each unit, and their amount of information differs considerably among the datasets. In the appraisals dataset we have a large number of covariates, including not only type (house or apartment), area, number of bedrooms, bathrooms and garage slots, but also information on the availability of amenities such as swimming pools, playgrounds, barbecue kiosks, fireplaces, among others. In the advertisements dataset the information consistently available is much smaller, comprising type, area and number of bedrooms, and in some cases the availability of a small number of amenities.

The statistical model assumes a gaussian distributio for $Y(s)$ with mean function $\mu(\mathbf{s};\beta)$ and covariance function $C\left(\mathbf{s}-\mathbf{s}';\boldsymbol{\theta}\right)=\sigma^2\rho\left(\mathbf{s}-\mathbf{s}';\boldsymbol{\phi}\right)$ so that $\boldsymbol{\theta}=\left(\sigma^2,\boldsymbol{\phi}\right)^T$.

So we have

$$\mathbf{Y}_s\left|\boldsymbol{\beta},\boldsymbol{\theta}\sim N\left(\boldsymbol{\mu}_s(\boldsymbol{\beta}),\sigma^2 H_s(\phi)\right)\right.$$

where $\mu_s(\beta)_i=\mu\left(s_i;\beta\right)$ e $\left(H_s(\phi)\right)_{ii'}=\rho\left(\mathbf{s}_i-\mathbf{s}_{i'};\phi\right)$. Prediction for new localities in a Bayesian context result from the predictive distribution function.

$$f\left(\mathbf{Y}_{s'}|\mathbf{Y}_s\right)=\int f\left(\mathbf{Y}_{s'}|\mathbf{Y}_s,\boldsymbol{\beta},\boldsymbol{\theta}\right)f(\boldsymbol{\beta},\boldsymbol{\theta}|\mathbf{Y}_s)d\boldsymbol{\beta}d\boldsymbol{\theta}$$

From the Gaussian Process, we can write:

$$f\left(\left(\begin{array}{c}\mathbf{Y}_{s'}\\ \mathbf{Y}_{s'}\end{array}\right)|\boldsymbol{\beta},\boldsymbol{\theta}\right)=N\left(\left(\begin{array}{c}\boldsymbol{\mu}_s(\boldsymbol{\beta})\\ \boldsymbol{\mu}_{s'}(\boldsymbol{\beta})\end{array}\right),\sigma^2\left(\begin{array}{cc}H_s(\boldsymbol{\phi}) & H_{s,s'}(\phi)\\ H^T_{s,s'}(\boldsymbol{\phi}) & H_{s'}(\phi)\end{array}\right)\right)$$

Therefore, $\mathbf{Y}_{s'}\left|\mathbf{Y}_s,\boldsymbol{\beta},\boldsymbol{\theta}\right.$ is distributed as

$$\begin{gathered}N\left(\boldsymbol{\mu}_{s'}(\boldsymbol{\beta})+H^T_{s,s'}(\boldsymbol{\phi})H_s^{-1}(\boldsymbol{\phi})\left(\mathbf{Y}_s-\mu_s(\boldsymbol{\beta})\right)\right)\\ \sigma^2\left[H_{s'}(\boldsymbol{\phi})-H^T_{s,s'}(\boldsymbol{\phi})H_s^{-1}(\boldsymbol{\phi})H_{s,s'}(\phi)\right]\end{gathered}$$

This model is estimated using the `SpBayes` package available for the statistical software `R`. Further statistical and computational details can be found in Banerjee et alli (2015). Estimates for prices of particular houses are obtained from their posterior predictve distributions.

Results

The above methodology is applied to estimating house prices in the city of Rio de Janeiro, with data from appraisals and advertisements spanning the months between january 2014 and december 2017. The MAPE (Mean Absolute Percentage Error) evaluation metric is calculated using LOOCV (Leave one out cross validation) for both datasets. The results are compared to two other models applied individually to each dataset:

- Univariate spatial model, using the same specification and Bayesian strategy of the model presented in last section, but with latent spatial effects modeled using purely auto-regressive correlation matrices.
- Tree-based Gradient Boosting Machine model (see Picchetti (2017)), a popular machine learning algorithm, trained to predict house prices based on their intrinsic characteristics including location information.

Model	Appraisals	Advertisements
GBM	18.3	28.1
Spatial Univariate	19.2	31.6
Spatial Misaligned	16.8	25.7

MAPE for the estimates based on appraisal data are significantly lower than the ones for advertisemnt data, which is expected given the greater availability of information on covariates contained in the appraisals. The univariate spatial model performs worse then the GBM algorithm, but the multivariate misaligned data model improves the results according to the MAPE metric.

Conclusions and further research

The methodology applied above results in precision gains when estimating hedonic models of house prices, in the context of different sources of information that complement each ohter. Once the values for individual dwellings on each month have been estimated in the best possible way, one can use these estimates to calculate any of the above price formulas based on hedonic modeling.

An alternative approach (see, for example Hill et alli (2017)) is to leverage the information of correlations between house prices not only across space, but across time as well.
The spatio-temporal model is

$$\begin{aligned} y_t(\boldsymbol{s}) &= \boldsymbol{x}_t(\boldsymbol{s})^\top \boldsymbol{\beta}_t + u_t(\boldsymbol{s}) + \epsilon_t(\boldsymbol{s}), \quad \epsilon_t(\boldsymbol{s}) \overset{ind}{\sim} N\left(0, \tau_t^2\right) \\ \boldsymbol{\beta}_t &= \boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \overset{i.i.d.}{\sim} N\left(0, \boldsymbol{\Sigma}_\eta\right) \\ u_t(\boldsymbol{s}) &= u_{t-1}(\boldsymbol{s}) + w_t(\boldsymbol{s}), \quad w_t(\boldsymbol{s}) \overset{ind}{\sim} GP\left(\boldsymbol{0}, C_t\left(\cdot, \boldsymbol{\theta}_t\right)\right), \quad t = 1, 2, \ldots, N_t \end{aligned}$$

This model has some attractive features that make it worth exploring:

- The dynamic specification of the latent spatial effect $u_t(\boldsymbol{s})$ induces interactions between time and spatial effects on house prices not easily captured the direct specification of a proper parametric spatio-temporal correlation function.
- The induced smooth time trajectory of the intercept component in $\boldsymbol{x}_t(\boldsymbol{s})^\top \boldsymbol{\beta}$ provides an attractive estimate of a time-dummy price index (as in Eurostat (2013)).
- Among the coefficients estimated in $C_t\left(\cdot, \boldsymbol{\theta}_t\right)$, the covariance matrix of the gaussian process behind the latent spatio-temporal effect, is the correlation between the values in $y_t(\boldsymbol{s})$. In our example, this is the correlation between asking prices and appraisal prices. Since it is also estimated through a dynamic specification, one can gain valuable insight on the trajectory of the differences between these prices, especially during changes in the housing market cycles.

Combining this methodology with the one presented above, one can devise a two-step procedure for estimating a dynamic hedonic housing price index, first combining information from different sources for constructing a robust dataset for each period, and then applyng the spatio-temporal model on this dataset.

References

Banerjee, S., Carlin, B. and Gelfand, A. (2015): *Hierarchical Modeling and Analysis for Spatial Data*, second edition, CRC Press.

Eurostat (2013): *Handbook on Residential Property Prices Indices (RPPIs)*

Hill, R., Rambaldi, A. and Scholz,M. (2017): *Weekly Hedonic House Price Indices: An Imputation Approach from a Spatio-Temporal Model*, Fifteenth Meeting of the Ottawa Group proceedings.

Piccheti, P. (2017): *A Hedonic Imputation approach to residential property price indexes: using a boosting learning algorithm applied to appraisal data*, Fifteenth Meeting of the Ottawa Group proceedings.