A Model Based Approach to Produce Commercial and Residential Property Price indices

Paulo Fernando Mahaz Simões[a,*]

[a] *Instituto Brasileiro de Geografia e Estatística - IBGE - Av República do Chile, 500, 9º andar, centro, Rio de Janeiro, RJ, cep 20031-170*

Abstract

In this work, we produce hedonic double imputation Laspeyres property price indices based on the creation of pseudo housing units. The proposed matching of multiple units makes the arrangement of a panel data comprised by transaction prices feasible, is imperative to estimate parameters via longitudinal mixed effects models and to circumvent some issues that arise in real estate studies, such as the lack of sufficient information on transaction prices and difficulties to perform analysis based on repeat sales for the short run or for nowcasting purposes. In contrast to traditional modelling approaches, which consider determinant variables of the prices as time invariant, here we relax this hipothesis and prices are modeled under the assumption that the status of some predictors, such as the floor or the size of an apartment can now be assumed as time-varying covariates. Also, the adoption of mixed effects models accomodates a specific statistical inference issue, that is exactly the treatment of the additional intra group variance that arises with the matching of different properties when we create pseudo housing units. Finally, we use a data set of transaction prices collected from 2014 to 2016 to estimate monthly house price indices for a specific area of the city of Rio de Janeiro. The findings suggest that the implemented methodology can be very useful to estimate price variations in the real estate market.

Keywords: Laspeyres House Price Indices. Hedonic Mixed Effects Models. Pseudo Housing Units. Panel Data. Nowcasting.

1. Introduction

Despite the importance of the housing sector for economic policy and the fact that property prices in Brazil have been experimenting substantial changes since 2005, the country does not have proper official statistics to measure changes in real estate prices, neither at the national nor at city levels. More generally, some issues that are present not only in Brazil but also in many other development countries can be enumerated for the lack of

* *Corresponding author: Paulo Fernando Mahaz Simões - Email: paulo.mahaz@ibge.gov.br*

information on house price indices. Two of them are interlinked and require special attention: the first has to do with difficulties to obtain organized and reliable registers or data sets either on transaction prices or on appraisal ones; the second is related to limitations of the methods employed, which normally face adversities in treating the heterogeneity of the available data and in properly answer some research questions of scientific interest, such as summaries of trends, variance estimations and analysis of economic and social predictors that influence changes in real estate prices. These problems are especially relevant in situations where information on transaction prices are constrained and when a large number of confounding variables are present. Therefore, although considerable research has been devoted to estimate house price indices, some incongruence between employed methods and data set characteristics remains and requires special attention.

Traditionally, the study of data sets taking into account the temporal aspect of the samples in real estate analysis has been treated by methods based on repeat sales (Bailey et al., 1963; Case and Shiller, 1989), although a series of critics has emerged in literature due to both estimation and sample selection bias. Additionally, the repeat sales approach requires information on prices of houses or apartments sold at least twice. Under this assumption, the panel data available for these studies are usually very sparse over time, since repeat sales of the same properties are infrequent.

Regardless of methodological improvements and the development of more elaborated models, where the analysis of spatial and temporal effects in house price dynamics have taken important rule (Nagaraja et al., 2011; Silver and Graf, 2014), challenges still persist and one relevant gap in the construction of house price indices is concerned to solve disagreements between methods and data, taking into consideration the peculiar characteristics of the real estate items that impair a large number of matching and direct comparisons between different housing units.

An important aspect in the construction of some types of price indices is the matching of sampling units to ensure the calculus of pure price variations, which excludes variations associated with quality changes. In the consumer price indices (CPI), for example, it is easy, in general, to obtain and compare samples of similar items. In the heterogeneous real estate market, however, the matching of similar units, as preconized by the repeat sales methodology, is not a trivial task. The problem, in fact, can be essentially very complex. Even in the hypothetical case where the analyst could access transaction prices of the same property in two periods (say, t and t-k), it would not be possible to guarantee the evaluation of the pure price variation by considering the division of the nominal prices in t by its correspondent in t-k (where k is the size of an appropriate lag), since some aspects that influence house prices such as depreciation, internal improvements, repairs or the presence of a new shopping center in the neighboring could affect property values. The construction of quality adjusted price indices, such as those proposed by hedonic techniques, is important to control for confounding variables that affect prices (Diewert, 2009).

To work around, we propose the estimation of hedonic double imputation house price

indices focusing on the longitudinal aspects of the data by adopting mixed effects models. Since longitudinal data is comprised by a large number of short time series, we make the methodology feasible with the creation of a panel data of transaction prices obtained from real estate agencies. The proposed matching of multiple units is crucial to the formation of panel data of transaction prices, is imperative to estimate parameters by using longitudinal mixed effects models and circumvents some problems that arise in real estate studies, such as the lack of sufficient information on transaction prices and difficulties to performing analysis based on repeat sales for the short run or even for nowcasting pruposes. In addition, prices are modeled under the assumption that the status of some predictors, such as the floor of an apartment, that are traditionally considered as been *time-invariant*, can now be assumed as *time-varying* covariates. Also, the adoption of mixed effects models accomodate an specific statistical inference issue, that is exactly the treatment of the additional intra group variance that arises with the matching of different properties when we create pseudo housing units.

In this work, we use a sample of 306 transaction prices collected from January 2014 to December 2016 to estimate monthly house price indices for a specific area of the city of Rio de Janeiro for the year 2016. The findings suggest that the implemented methodology can be very useful to estimate price variations in the real estate market. The article is set out as follows. Section 2 is dedicated to literature review. Section 3 describes the creation of pseudo housing units. In section 4 we comment on mixed effects models methodology. In section 5 we show some results and in section 6 we present some discussion and concluding remarks.

2. Literature Review

The problem of inferring price trends is not recent and a series of works has devoted attention to analyze methods for house prices appreciations, with some of them less intensive in estimation procedures, such as the non parametric stratification, mean and median approaches. However, two another relevant technics are quite diffused in the construction of house price indices: the hedonic (Griliches, 1971) and the repeat sales methodologies (Bailey et al., 1963; Case and Shiller, 1989), each of them with important advantages and some limitations. An important objective related to the former is the control of varying characteristics of the properties. This issue has been studied by several authors since 1970s (Case and Quigley, 1991) but has suffered of difficulties in obtain sufficient data, which would consist of sale prices at different points in time and different locational and property characteristics, an essential requirement for the implementation of hedonic models (Clapp and Giaccotto, 1994). Maybe, these uncertainties and lack of information on housing data had been probably at least partly responsible for the greater diffusion of repeat sales method. Although in its genuine version the method was based exclusively on properties sold at least twice, the repeat sales implicitly involved quality control of price variations.

Concernning econometric studies more directly related to house prices, modeling mean and covariance in the real estate context experimented a great development in the 1960s, with the model proposed in the work of Bayle, Murth & Nurse (1963), hereafter called BMN (Bailey et al., 1963). Recognizing the difficulties in estimating house prices variation due to heterogeneity in house's characteristics, they introduced a version of the repeat sales method in a regression context, comparing log prices between consecutive sales of same properties in different occasions based on the hypothesis of independent errors terms with homogeneous variance. The method of ordinary least squares (OLS) was adopted to estimate the parameters of interesting. An underlying hypothesis of the work was that the error terms were normally distributed with zero mean and constant variance, that is, $\epsilon \approx N(0, \sigma^2)$.

Several years later, Case & Shiller (Case and Shiller, 1989) proposed a modification in the BMN regression method, arguing that the variance of the error terms would not be constant across houses but increased with the interval between sales. They presented the construction of quarterly indexes of existing home prices between 1970 and 1986 using what they called weighted repeat sales method (WRS). A three step regression procedure was used to estimate the indexes. In the first stage, the log price of the second sale minus the log price of the first sale was regressed on a set of dummy variables, one for each time period in the sample except the first. From the first stage, a vector of residuals was calculated. In the second stage, a weighted regression of the squared residuals from the first stage was run on a constant term and the time between sales. In the third stage they employed generalized least squares regression (a weighted version), repeating the stage one after dividing each observation by the square root of the fitted value in the second stage. Therefore, the authors treated heteroscedastic problems in the prices.

Case & Quigley (Case and Quigley, 1991) combined in a single joint estimation procedure information on repeat sales of unchanged properties, on repeate sales of improved properties, and on single sales. For the treatment of heteroscedastic problems and to obtain more efficient estimates, two-stage generalized least squares estimation were used. The authors dedicated special attention to evaluate the accuracy of the price indexes derived from the models. This accuracy depended on the variance-covariance matrices of the estimated parameters.

Shiller (Shiller, 1991) also criticized geometric repeat sales (GRS) index obtained from the BMN model. Extending previous work (Case and Shiller, 1989), he observed that the variance of the error term would depend on the interval between sales, reinforcing that a more efficient estimator would be a weighted regression and suggested the use of generalized least squares to treat heteroscedasticity . It was proposed then an arithmetic repeat sales estimators with several variants: a value-weighted arithmetic repeat sales estimator (VW-ARS), the equally weighted arithmetic repeat sale (EW-ARS), and the interval-weighted and the hedonic-variable-augmented variations on these.

Kain & Quigley (Kain and Quigley, 1970) estimated the market value, or the implicit prices, of specific aspects of the bundles of residential services consumed by urban house-

holds. They obtained quantitative estimates by regressing market price of owner and renter-occupied dwelling units on measures of the qualitative and quantitative dimensions of the housing bundle and verified the influence of neighborhood schools on the value of residential properties; Griliches (Griliches, 1971) studied quality change problems in the house prices scope and used hedonic techniques in the adjustment of price variations.

Clapp & Giaccotto (Clapp and Giaccotto, 1994), studied the relationship between assessed values (AV) and repeat sales methods and economic determinants of house prices at local level. The methods were compared by modeling true price appreciation using data from Hartford, Connecticut. They found a high correlation (0.8) between the two approaches in the annual rate of change but important differences in price indices derived from them over periods of 2 to 10 quarters. They suggested that results could be improved by combining AV and repeat sales.

Wang & Zorn (Wang and Zorn, 1997) studied the aims and the targets of a series of methods related to house price indices and presented an idealized population architecture to represent a sampling process from the housing stock. Additionally, they recognized that growth rates would differ across properties and that, for each time t, there would exist a distribution of growth rates in the population. Moreover, they commented on a series of fundamental concepts behind house price indices.

Dorsey et al. (Dorsey et al., 2010) constructed hedonic house prices considering ZIP code information for the metropolitan regions of Los Angeles and San Francisco in the US. So as to circumvent selection bias problems in traditional surveys that have used repeat sales approaches, they got data on 1.1 million assessed values and explanatory variables from a mortgage company.

Goetzmann & Peng (Goetzmann and Peng, 2002) analyzed the implications of cross-sectional heteroscedasticity in the repeat sales regression (RSR) that was essentially geometric averages of individual asset returns because of the logarithmic transformation of price relatives and showed that the cross-sectional variance of asset returns affected the magnitude of the bias in the average return estimate for each period, while reducing the bias for the surrounding periods. They suggested an unbiased maximum likelihood alternative to the repeat sales regression that directly estimated index returns, which they called MLRSR. The unbiased MLRSR estimators were analogous to the RSR estimators but were arithmetic averages of individual asset returns. They showed that MLRSR could be more accurate than RSR.

Nagaraja et al. (Nagaraja et al., 2011), in the sphere of repeat sales regression, developed a statistical model for predicting individual house prices utilizing information regarding sale price, time of sale and zip code. They used data for single-family home sales for twenty U.S. metropolitan areas from July 1985 through September 2004. The model combined a fixed effect for time, that could be converted into a house price index, and a random effect for the ZIP code variable with an autoregressive component.

Silver & Graf (Silver and Graf, 2014) estimated commercial property price indexes paying attention to some key points of interest such as problems of sparse data, spatial

spillovers, and weighting.

Deng et al. (Deng et al., 2014) proposed a matching of similar housing units to obtain commercial property price indices based on quantile regressions. Authors used a dataset containing information on prices from 1995 to 2010.

Guo et al. (Guo et al., 2014) also developed a strategy to match two very similar new sales within a defined matching space. They argued that the methodology would be very useful to be applied in densely populated cities with large housing complexes. They applied the method in a study for the Chendu City, in China and used data from 2006 to 2011.

Sedgley et al. (Sedgley et al., 2008) incorporated spatial dependency in a hedonic model based on housing, neighborhood, demographic, and school quality attributes readily available on the Internet for home sales in Howard County, Maryland. The study provided guidance on how to test for spatial heterogeneity and non-normality of error terms before proceeding with hedonic analysis.

Glennon et al. (Glennon et al., 2018) highlighted the importance of improving the accuracy of property-level valuations. In a study using data from Florida, they demonstrated that forecast combination methods reduce the estimated bias and found that even the simplest forecast combination methodology, such as a simple average, has the potential to significantly improve value estimates.

3. Pseudo Housing Units

In this work, we focus on the longitudinal aspects of house prices. But taking into account that one distinctive feature of longitudinal data is that they are comprised by a large number of short time series and also considering that we do not have a natural database with such characteristics, we suggest here a construction of a panel dataset containing transaction prices based on pseudo housing units. The proposed procedure links different residential properties and is crucial to the implementation of mixed effects models as described in section 4.

To illustrate the architecture of a panel data based on pseudo housing units, we start with the ideas presented in Wang and Zorn (Wang and Zorn, 1997) about the sampling process of housing units in a idealized population and extend their matrices representation. Let Y_{ij} be the price of property i $(i = 1, ..., N)$ on occasion j $(j = 1, ..., J)$. In a idealized population with N properties and considering J reference periods, we would represent the housing stock prices, in a initial moment of the analysis, as showed in Table 1, where rows represent the unknown prices of same units along the J periods of interesting.

Table 1: Idealized Housing Stock

0	1	2	3	4	...	J
Y_{10}	Y_{11}	Y_{12}	Y_{13}	Y_{14}	...	Y_{1J}
Y_{20}	Y_{21}	Y_{22}	Y_{23}	Y_{24}	...	Y_{2J}
Y_{30}	Y_{31}	Y_{32}	Y_{33}	Y_{34}	...	Y_{3J}
Y_{40}	Y_{41}	Y_{42}	Y_{43}	Y_{44}	...	Y_{4J}
Y_{50}	Y_{51}	Y_{52}	Y_{53}	Y_{54}	...	Y_{5J}
Y_{60}	Y_{61}	Y_{62}	Y_{63}	Y_{64}	...	Y_{6J}
...	...	...	...	...	...	...
Y_{N0}	Y_{N1}	Y_{N2}	Y_{N3}	Y_{N4}	...	Y_{NJ}

Fonte: Author

Taking into account that new properties are constructed and old properties are demolished, some units are added and others are dropped in a second stage idealized representation. Table 2 illustrates a new characteristic of the housing stock. It is interesting to note that the third real estate unit was demolished in $j = 2$ and then removed from the data base; Also, the fiveth unit was constructed and included in the housing stock at the same time period.

Table 2: Changes in Housing Stock

0	1	2	3	4	...	J
Y_{10}	Y_{11}	Y_{12}	Y_{13}	Y_{14}	...	Y_{1J}
Y_{20}	Y_{21}	Y_{22}	Y_{23}	Y_{14}	...	Y_{2J}
Y_{30}	Y_{31}	-	-	-	...	-
Y_{40}	Y_{41}	Y_{42}	Y_{43}	Y_{44}	...	Y_{4J}
		Y_{52}	Y_{53}	Y_{54}	...	Y_{5J}
Y_{60}	Y_{61}	Y_{62}	Y_{63}	Y_{64}	...	Y_{6J}
...	...	...	...	...	...	...
Y_{N0}	Y_{N1}	Y_{N2}	Y_{N3}	Y_{N4}	...	Y_{NJ}

Fonte: Author

Turning to some aspects of a real situation, a data base on transaction prices would be also be affected by both the a lack of information on some units. Since real estate units are infrequently sold, we could suppose in a third stage, a reduction in the available information, which would be, though, comprised by information only on sold properties (Table 3).

Table 3: Prices of Sold Properties

0	1	2	3	4	...	J
-	Y_{11}	-	-	Y_{14}	...	Y_{1J}
Y_{20}	-	-	Y_{23}	-	...	-
-	-	-	-	-	...	-
		-	Y_{53}	-	...	Y_{5J}
-	Y_{61}	-	Y_{63}	-	...	Y_{6J}
...	...	...	...	...	...	...
Y_{N0}	-	-	-	Y_{N4}	...	-

Fonte: Author

Finally, some factors, such as the absence of properties not included in the sample and the lack of information for sold ones would act to remove a large number of data from our initially idealized population, leading to an additional reduction in the available data base of transaction prices (Table 4).

Table 4: Prices of Sold Properties

0	1	2	3	4	...	J
-	-	-	-	Y_{14}	...	-
Y_{20}	-	-	-	-	...	-
-	-	-	-	-	...	-
		-	Y_{53}	-	...	Y_{5J}
-	Y_{61}	-	-	-	...	-
...	...	...	...	...	...	...
-	-	-	-	Y_{N4}	...	-

Fonte: Author

However, sales of housing units is a infrequent event and a panel data based on transaction prices and characteristics of sold properties for a convenient sample size of real estate units, comprised by repeated measures of exactly same units, is not feasible, unless one admits to work with large lags between sales, as is the case of repeat sales methodology. To workaround and so as to apply longitudinal mixed effects models formulation, taking advantage of its important characteristics, such as the treatment of the data covariance structure and the separation of cross-sectional and longitudinal effects, we suggest the chaining of similar real estate units sold in a specific area during a determined time period to develop a residential property price index. The objective is the construction of a panel data based on short time series of transaction prices from similar units. By similar units we mean housing units located at the same condos or in a 100 meters radio distance

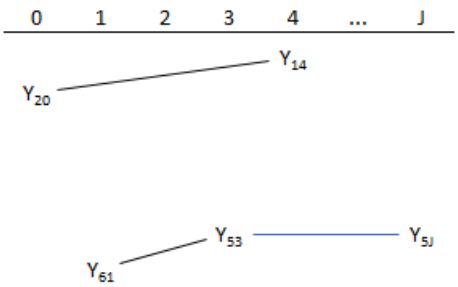

Figure 1: Pseudo Housing Units

Table 5: Chainning Prices of Similar Units

0	1	2	3	4	...	J
				Y_{14}		
Y_{20}						
			Y_{53}			Y_{5J}
	Y_{61}					
				Y_{N4}		

Fonte: Author

4. Methodology

The matching of different dwellings proposed in section 3 allows us to identify the constructed series as representing the tendency of the transaction prices of *pseudo housing units*, which should be viewed as an hypothetical property (i) that could have their physical characteristics or the status of their attributes modified at any time during the survey. Under this assumption, the response vector y_i for the pseudo housing unity or hypothetical property i in a model with fixed and random effects can be written as:

$$y_i = X_i\beta + Z_i b_i + \epsilon_i \tag{1}$$

where $i = 1, ..., N$ represents a housing unit with $j = 1, ..., n_i$ possible different numbers of repeat sales, y_i is the n_i x 1 dependent variable vector for unit i; β is the p x 1 vector of fixed regression parameters; X_i is a n_i x p covariate matrix for property i; Z_i represents the n_i x r design matrix for the random effects, b_i is the r x 1 vector of random effects and ε_i is the n_i x 1 vector of disturbances.

Commonly, we adopt $b_i \sim N(0, \Sigma_b)$ and estimate coefficients under the distributional assumption of conditionally independent errors (Hedeker and Gibbons, 2006), that is: ($\varepsilon_i \sim$

$N(0, \sigma^2_{\varepsilon} I_{n_i})$), since the inclusion of random effects, in many situations, are sufficient to account for the covariance of the data. In this case, we have:

$$\begin{pmatrix} \mathbf{y_i} \\ \mathbf{b_i} \end{pmatrix} \sim N \left(\begin{bmatrix} X_i \\ 0 \end{bmatrix}, \begin{bmatrix} Z_i \Sigma Z_i^{'} + \sigma^2_{\varsigma} I_n & Z_i \Sigma_b \\ \Sigma_b Z_i^{'} & \Sigma_b \end{bmatrix} \right) \tag{2}$$

and then:

$$V(y_i) = Z_i \Sigma_i Z_i^{'} + \sigma^2 I_i \tag{3}$$

where I_i is the identity matrix. The above mixed effects models formulation offer a series of advantages. Two of them relies on the possibility of treating the complex covariance of the data in a more parsimonious way, by adopting a reduced number of parameters. Another convenience is that we can analyze the price trajectories of specific pseudo housing units over time. In this case, however, the requirement is the estimation of individual effects b_i. Fitzmaurice Fitzmaurice et al. (2011) shows that if we admit multivariate normal distribution, it is possible to predict the conditional mean of b_i given Y_i if $\hat{\beta}$ is knew. Under this hypothesis,

$$E(b_i|Y_i) = GZ_i^{'} \Sigma_i^{-1} (Y_i - X_i \hat{\beta}) \tag{4}$$

with $\Sigma_i = Cov(Y_i) = Z_i G Z_i^{'} + R_i$[1].

In this situation, the predictor $\hat{b_i}$ for b_i is:

$$\hat{b_i} = \hat{G} Z_i^{'} \Sigma_i^{-1} (Y_i - X_i \hat{\beta}) \tag{5}$$

and the response vector $\hat{Y_i}$ is described as:

$$\hat{Y_i} = X_i \hat{\beta} + Z_i \hat{b_i} \tag{6}$$

[1]This is the *Best Linear Unbiased Predictor*, (*BLUP*). By adopting a restrict maximum likelihood estimation for Σ_i^{-1} and substituting it in 4 we obtain the *Empirical Bayes Estimator* for the b_i. See, for example, Gelman & Hill (Gelman and Hill, 2007, p.346) e Gelman et al. (Gelman et al., 2013, p.104).

4.1. Preliminary Notes on Hedonic Double Imputation Laspeyres House Price Indices Calculation

In the hedonic double imputation Laspeyres house price indices (OECD et al., 2013) in which the objective is the calculus of house price variations between the moments t and 0, I^{0t}_{HDIL}, the observed prices of the property i in the sample S are replaced by its predicted values for both t and 0 periods. These predicted values are obtained from appropriated models and included in the index formula 7:

$$I^{0,t}_{HDIL} = \frac{\sum 1\hat{p}^t_i(S)}{\sum 1\hat{p}^0_i(S)} = \frac{\sum [\hat{\beta}^t_0 + \sum \hat{\beta}^t_k z^0_{ik}]}{\sum [\hat{\beta}^0_0 + \sum \hat{\beta}^0_k z^0_{ik}]} \tag{7}$$

where $\hat{p}^t_i(S)$ and $\hat{p}^0_i(S)$ are imputed prices for property i in the periods t and 0, respectively. Therefore, to calculate the variation for January 2016 $\hat{I}^{Jan16}$, it is necessary to estimate two models: one to access predicted prices for January 2016 and another one to obtain values for the base period, December 2015. Hence,

$$\hat{I}^{Jan16} = \frac{\sum \hat{I}^{Jan16}_i}{\sum \hat{I}^{Dez15}_i} \tag{8}$$

In this sense, the proposed framework requires, for each specif period of interest, the estimation of two regressions, both of them circumscribing part of the entire database. In this work, mixed effects models using properties data of 24 months (the current and 23 previous months) were used to estimate prices for each month. To obtain regression coefficients for January 2016, a panel data encompassing transaction prices and related regressors from February 2014 to January 2016 was employed.

4.1.1. Data and Samples

The original data set available for the study was obtained from real estate brokers and agencies. It contains variables on transaction prices and characteristics such size in square meters ($AREA$), number of rooms(R), number of bathrooms (B), year of the construction (I), internal home characteristics (PI) and facilities of the condominium (PC) for 306 apartments located in seven specific areas of the city of Rio de Janeiro, which were sold between January 2014 and December 2016. Additionally, we have adopted the variable $TIME$ to represent the month in which the property was sold and some other variables to capture geographic and social specific aspects of the city. One of them, labeled as $DIST$, refers to the distance to the beach. It is worth noting that the sample contained households located in streets where the prices were influenced by the proximity to dangerous or violent areas and the variable PDM was created to treat this issue.

By adopting the matching methodology described in section 3, a unbalanced panel data with 61 pseudo housing units was formed. We applied mixed effects models to estimate parameters coefficients and to predict prices that were used in the calculus of the monthly indicators in accordance with the double imputation hedonic Laspeyres formula.

So as to compare results, we have adopted six different samples of properties, labeled as $S1$, $S2$, $S3$, $S4$, $S5$ and $S6$. The $S1$ sample was formed by sold properties that have entered into database in December 2015. $S2$ sample was comprised by sold properties that have entered into database from November to December 2015 and so on. The $S6$ sample, for example, was comprised by housing units that were sold in the period from July to December 2015. Each of these samples received predicted values based on mixed effects models in accordance with formula 7. In summary, for a fixed sample, thirteen regressions were estimated, with each of them using a specif part of the panel.

Previous analysis not reported here have showed that some variables were not significant, to explain changes in property prices trajectories over time and then preliminary models results including these variables are omitted. In this paper, therefore, mixed effects models with conditionally independent errors are estimated using the variables *TIME*, *AREA*, *PADCOND* and *DIST*. The library *nlme* (Pinheiro et al., 2017) of the *R* software (R Core Team, 2016) were employed to estimate the models by adopting the restricted maximum likelihood method (REML) (Patterson and Thompson, 1971).

5. Specific Models and Results

Based on prior analysis of the data base and considering the importance of some predictors for the housing sector, the specification of the proposed model has the *price by square meter* as response vector and involves four independent variables, "TIME" (T), "Region" (B), "Condominium characteristics" (PC) and "Distance to sea shore" (D). The latter is important because it accounts partially for the spatial correlation of the data. We adopted the same specification for the thirteen models used to estimate indices for each month of the year 2016. Therefore, the models are described as:

$$\begin{aligned} Y_{ij} = \beta_0 + b_{0i} + \beta_1 T_{ij} + b_{1i} T_{ij} + \beta_2 B_i + \beta_3 PC_{ij} + \beta_4 D_{ij} + \epsilon_{ij} \\ b_{0i} \sim N(0, \sigma_0^2) \\ b_{1i} \sim N(0, \sigma_1^2) \\ \epsilon_{ij} \sim N(0, \sigma_\epsilon^2) \end{aligned} \tag{9}$$

where b_{0i} and b_{1i} are, respectively, random intercepts and slopes related to each pesudo housing unit (i), ϵ_{ij} are conditionally independent errors terms with hypothesized normal distribution with zero mean and varaince σ_ϵ^2. Tables 6 and 7 summarize the results for *December 2015* and *January 2016*.

Table 6: Results for December 2015

Variável	Coefficient	Estiation	Standard Deviation	p-value
Intercept	β_0	6378.54	685.37	0.000
Month	β_1	-40.89	6.68	0.000
Barra area	β_2	–	–	–
Bonsuc area	β_2	-2818.75	790.40	0.000
Freguesia area	β_2	-1459.82	988.11	0.145
Olaria area	β_2	-2945.64	1098.70	0.009
Penha area	β_2	-2788.51	903.31	0.003
Ramos area	β_2	-3579.34	840.71	0.000
V. Pen area	β_2	-2650.61	902.64	0.005
PC 1	β_3	–	–	–
PC 2	β_3	461.29	251.66	0.000
PC 3	β_3	1160.72	270.71	0.000
PC 4	β_3	1917.37	375.06	0.000
PC 5	β_3	4394.90	667.27	0.000
D (far)	β_4	–	–	–
D (near)	β_4	3160.05	635.35	0.000
D (in sea shore)	β_4	5141.09	755.09	0.000
Quality of Fit	AIC = 3458.81	BIC = 3518.97	LL = -1711.40	

Table 7: Results for January 2016

Variável	Coeficiente	Estimativa	Desvio Padrão	p-valor
Intercept	β_0	6426.32	687.30	0.000
Mês	β_1	-43.71	6.69	0.000
Barra area	β_2	–	–	–
Bonsuc area	β_2	-2720.20	811.17	0.001
Freguesia area	β_2	-1483.21	989.551	0.140
Olaria area	β_2	-2981.72	1102.08	0.009
Penha area	β_2	-2787.640	906.68	0.003
Ramos area	β_2	-3606.73	843.61	0.000
V Pen area	β_2	-2653.52	906.01	0.005
Pad Cond 1	β_3	–	–	–
Pad Cond 2	β_3	473.74	252.56	0.062
Pad Cond 3	β_3	1157.21	271.26	0.000
Pad Cond 4	β_3	1925.83	376.32	0.000
Pad Cond 5	β_3	4394.67	669.56	0.000
D far	β_4	–	–	–
D near	β_4	3146.76	637.79	0.000
D in sea shore	β_4	5140.17	757.85	0.000
Quality of fit	AIC = 3458.26	BIC = 3518.42	LL = -1711.13	

Looking over the results in Tables 6 and 7, one observes that all of the coefficients were significants in both regressions, except for the factor "Freguesia area" (p-*value* $\cong 0.145$). However, since estimated values for this level is in accordance with theoretical expectations related to differences in prices between areas, we chose to maintain it in the models.

5.1. Residual Diagnostics

To complete the analysis we show residual diagnostics for both estimated regressions based on the assumed hypotheses. The vector of residuals in the longitudinal context suggested here for each pseudo housing unit i is

$$r_i = Y_i - X_i\hat{\beta} - \hat{b_i} \tag{10}$$

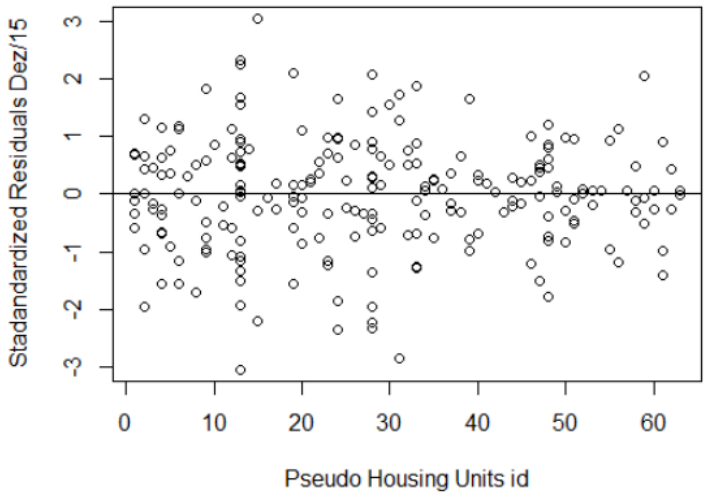

Figure 2: Standardized residual distribution - December 2015.

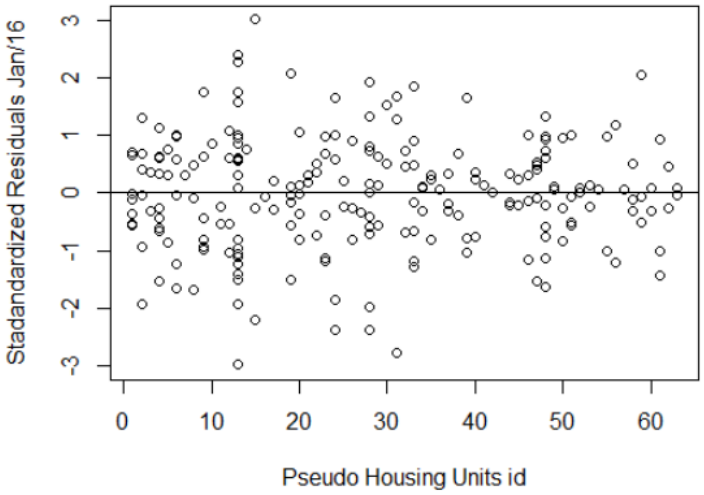

Figure 3: Standardized residual distribution - January 2016.

Figures 2 and 3 take into account disturbances for the 306 original houses. For each of the 61 pseudo housing units, vertical points refer to deviations across time related to its expected prices. We have admitted that the values are randomly distributed around zero. Furthermore, the distribution of the random intercepts and slopes for *December 2015* and *January 2016* showed in Figures 4, 5, 6 and 7 are also consistent with normality assumptions and evidence no apparently systematic pattern.

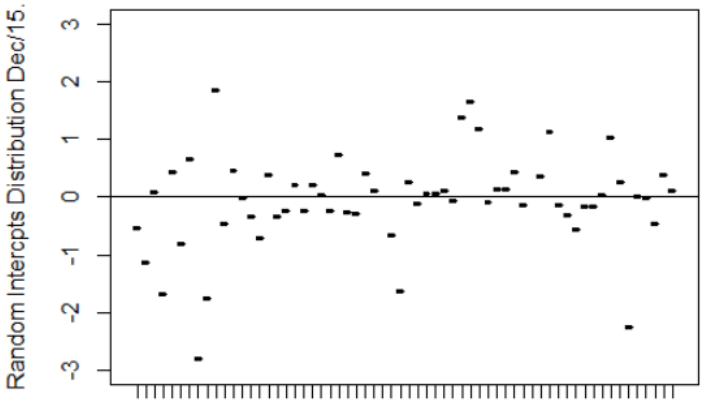

Figure 4: Random intercepts distribution - December 2015.

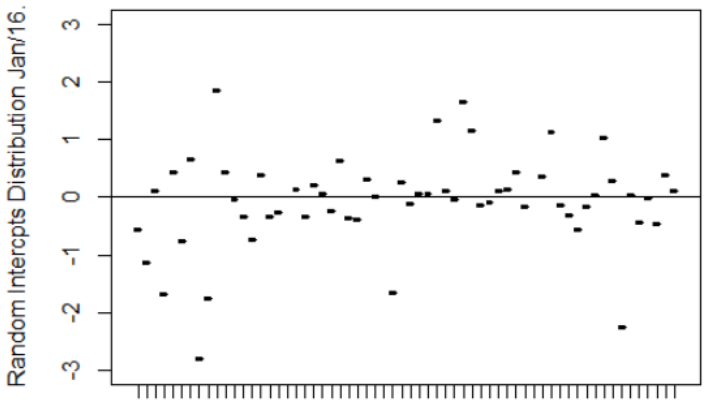

Figure 5: Random intercepts distribution - January 2016.

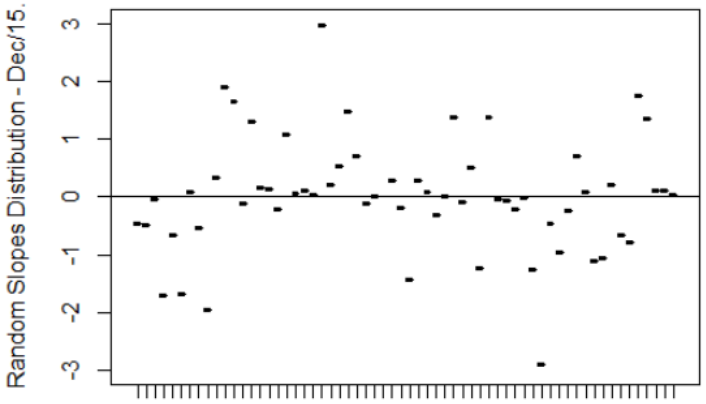

Figure 6: Random slopes distribution - December 2015.

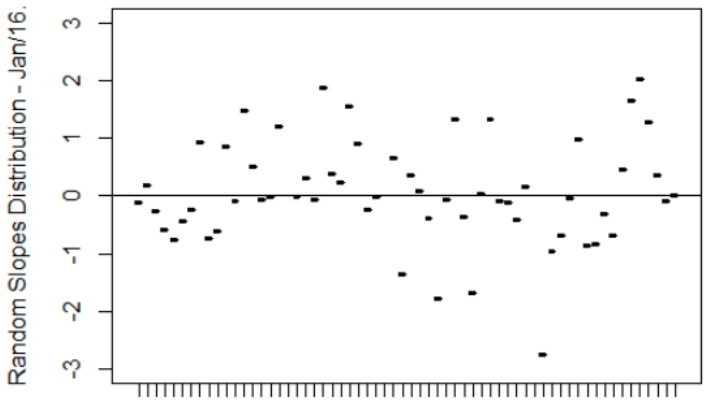

Figure 7: Random slopes distribution - January 2016.

5.1.1. One Step Forward Predictions

Although the primary interest of the suggested approach do not relies on forecasting, we have used the models to analyze short run predictions. Considering estimated regressions and real prices available, we found predicted values one step ahead for specific housing units. We have validated results by taking a look to the Mean Absolute Percentage Error (*MAPE*) and Mean Absolute Deviations (*MAD*. Table 8 resume results for the sample $S6$, that is, comprised by houses that was sold from July 2015 to December 2015[2].

Table 8: Predictions

Modelo	Reference	Predictions	MAPE (%)	MAD (R$ $/m^2$)
1	Dez/15	Jan/16	8.0229	468.17
2	Jan/16	Fev/16	7.9821	624.25
3	Fev/16	Mar/16	2.1236	182.90
4	Mar/16	Abr/a6	8.8944	583.06
5	Abr/a6	Mai/16	10.3189	416.29
6	Mai/16	Jun/16	6.0559	407.52
7	Jun/16	Jul/16	5.6451	418.31
8	Jul/16	Ago/16	-	-
9	Ago/16	Set/16	9.2607	555.9
10	Set/16	Out/16	11.4599	930.04
11	Out/16	Nov/16	5.1595	408.18
12	Nov/16	Dez/16	7.7515	383.04

[2]The MAPE statistic is calculated by $\frac{\Sigma|(y_{ij} - \hat{y}_{ij})/y_{ij}|}{n_j} * 100$, with $(y_{ij} \neq 0)$, while *MAD* is obtained by $\frac{\Sigma|y_{ij} - \bar{y}_{ij}|}{n_j}$.

Table 9: House price variations for each month of 2016 considering 6 sample types

Mês	S1	S2	S3	S4	S5	S6
Jan	-1.3616	-1.2257	-0.9809	-0.9678	-0.9665	-0.9399
Feb	-0.8593	-0.8289	-1.7305	-1.5259	-1.2960	-1.1849
Mar	-1.2923	-1.2788	-1.0337	-0.9837	-0.9106	-0.8921
Apr	-0.9461	-0.9675	-0.7374	-0.8255	-0.7968	-0.7808
May	-0.6097	-0.3694	-0.3694	-0.4981	-0.4772	-0.4240
Jun	-1.2654	-1.2507	-1.2477	-1.3173	-1.2772	-1.2045
Jul	-0.3982	0.1136	0.0307	0.0897	0.0275	-0.0115
Aug	-0.4451	-0.3675	-0.2513	-0.2249	-0.1297	-0.0572
Sep	-1.5931	-1.6205	-1.1224	-0.9200	-0.6637	-0.6117
Oct	-1.1872	-1.2907	-1.7590	-2.3445	-1.9898	-1.8832
Nov	-0.5819	-0.6782	-0.4999	-0.3636	-0.4039	-0.3805
Dec	-0.7385	-0.6300	-0.6215	-0.6777	-0.5903	-0.5849
Accumulated	-10.7215	-9.9261	-9.8645	-10.0846	-9.0894	-8.6112

5.2. Monthly Laspeyres House Price Indices Estimations

The specified models in 9 were eligible to estimate double imputation Laspeyres house price indices (*HDIL*) for each sample type ($S1$ to $S6$), as mentioned in section 4.1.1. Substituting observed prices by predicted values, the index formula becomes:

$$I_{HDIL}^{t,t-1} = \frac{\sum_{i \in S6} [\hat{b}_{0i}^{t} + \hat{\beta}_{0}^{t} + \hat{b}_{1i}^{t} T_{ij}^{S6} + \sum_{k=1}^{K} \hat{\beta}_{k}^{t} X_{ik}^{S6}]}{\sum_{i \in S6} [\hat{b}_{0i}^{t-1} + \hat{\beta}_{0}^{t-1} + \hat{b}_{1i}^{t-1} T_{ij}^{S6} + \sum_{k=1}^{K} \hat{\beta}_{k}^{t-1} X_{ik}^{S6}]} \tag{11}$$

wehere ($I_{HDIL}^{t,t-1}$) is the calculated index between periods t and $t-1$. For January 2016, the computation is:

$$I_{HDIL}^{jan16,dez15} = \frac{\sum_{i \in S6} [\hat{b}_{0i}^{jan16} + \hat{\beta}_{0}^{jan16} + \hat{b}_{1i}^{jan16} T_{ij}^{S6} + \sum_{k=1}^{K} \hat{\beta}_{k}^{jan16} X_{ik}^{S6}]}{\sum_{i \in S6} [\hat{b}_{0i}^{dez15} + \hat{\beta}_{0}^{dez15} + \hat{b}_{1i}^{dez15} T_{ij}^{S6} + \sum_{k=1}^{K} \hat{\beta}_{k}^{dez15} X_{ik}^{S6}]} \tag{12}$$

where $\hat{\beta}_0$ is the overall population intercept, $\hat{\beta}_k$ are estimated coefficients of predictors included in the models, $\hat{b}_{0i}$ and $\hat{b}_{1i}$ are, respectively, random intercepts and slopes and X_{ik} is the design matrix of fixed effects.

The application of formula 11 to the panel database comprised by pseudo housing units led to the estimates shown in Table 9.

by taking a look at Table 9, it is worth noting that house price variations obtained for each month of 2016 is similar in level and in tendency. The highest accumulated variation was observed for sample $S1$ (-10.7215%), while sample $S6$ presented the lowest value for the annual index (-8.6112)[3]. Figures 8, 9, 10, 11, 12 and 13 offer a visualization of indexes calculated using the proposed methodology for each one of the six sample sizes. Additionally, Figure 14 shows the six index tendencies together.

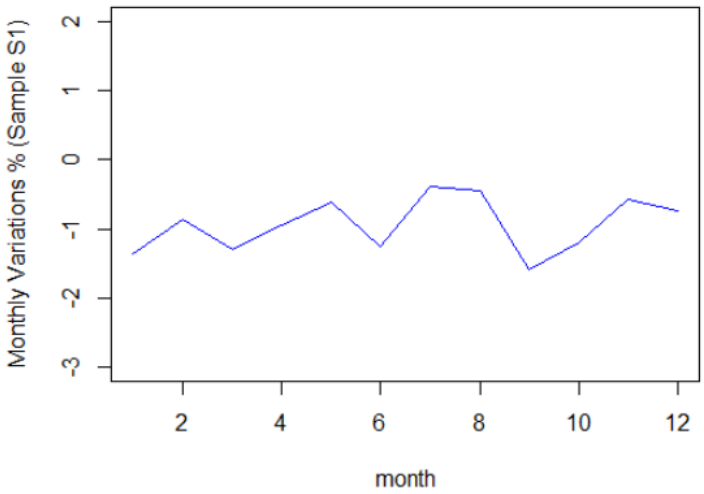

Figure 8: Monthly 2016 variations - Sample S1

[3]Nominal values

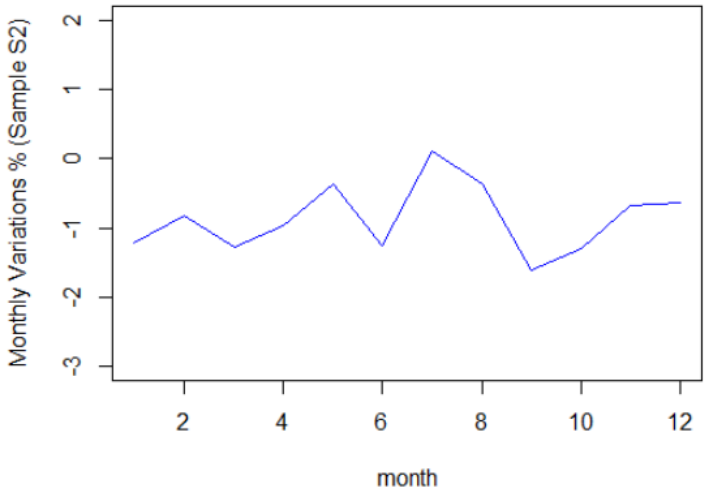

Figure 9: Monthly 2016 variations - Sample S2

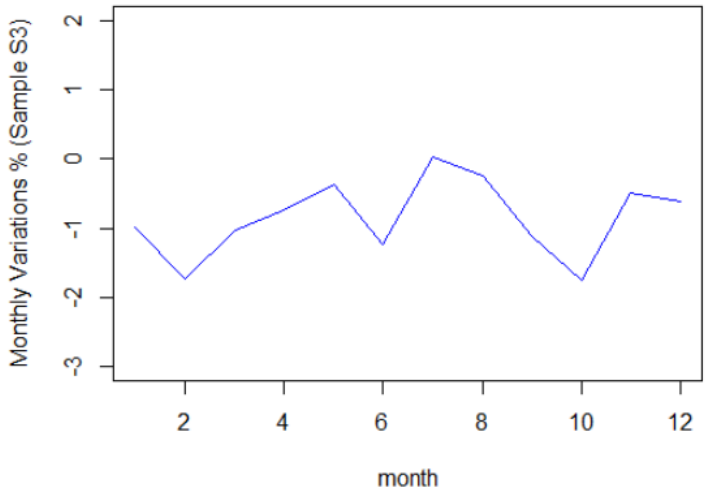

Figure 10: Monthly 2016 variations - Sample S3

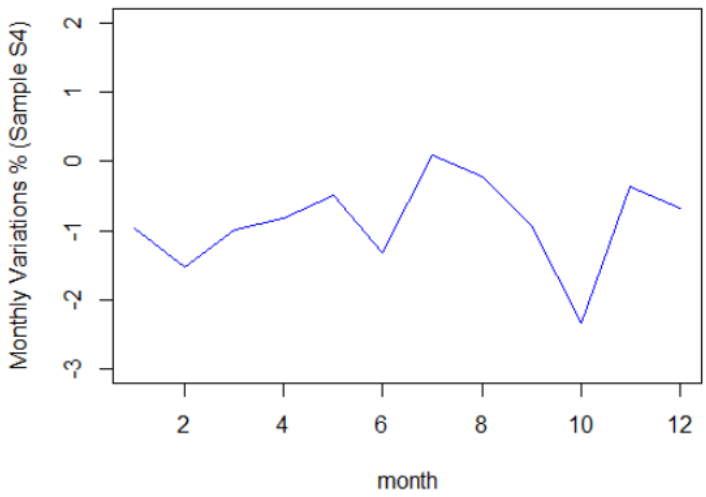

Figure 11: Monthly 2016 variations - Sample S4

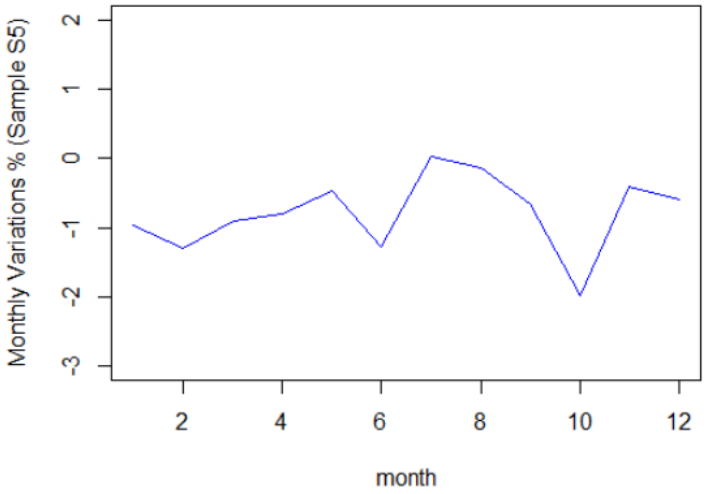

Figure 12: Monthly 2016 variations - Sample S5

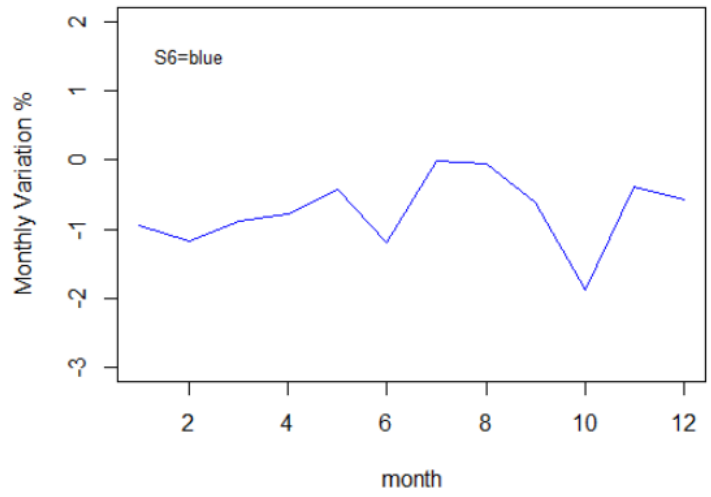

Figure 13: Monthly 2016 variations - Sample S6

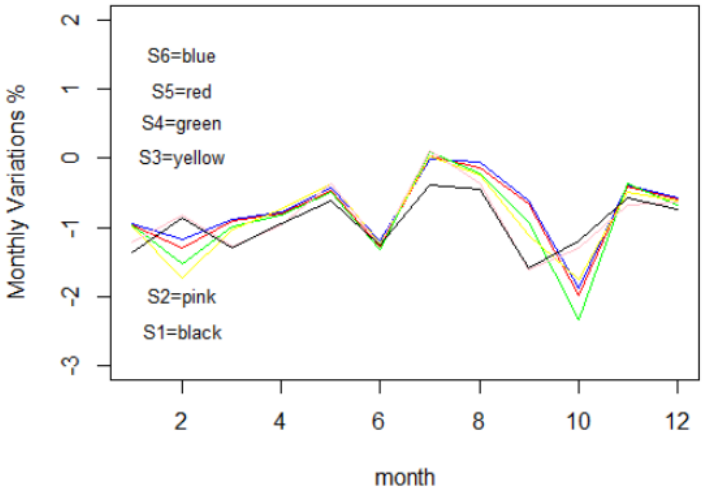

Figure 14: Comparing variations - Samples S1 to S6

5.3. Bootstrap Confidence Intervals

The accuracy of monthly hedonic double imputation Laspeyres house price indices were investigated with bootstrap confidence intervals (Efron, 1979) based on 1000 replicates for each month. In Table 10, it is possible to see upper and lower limits for the sample $S6$, which has size 60. Figure 15 displays the trajectories of the indices with 95% confidence limits.

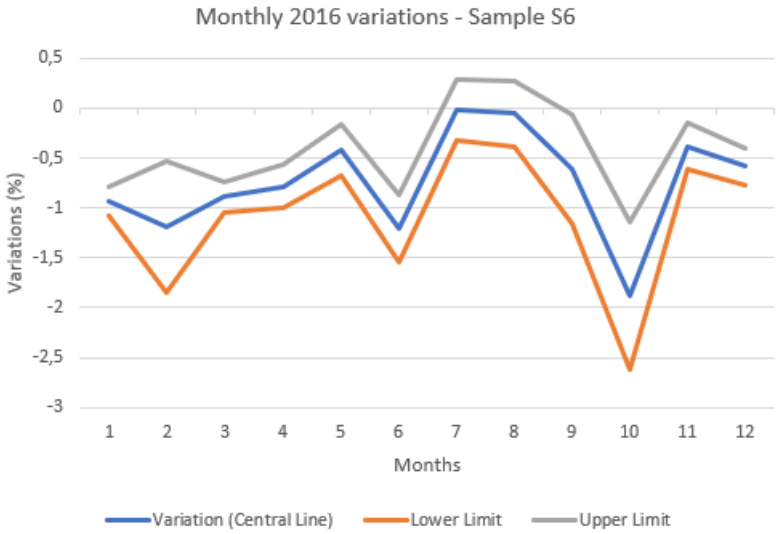

Figure 15: Monthly variations with 95% bootstrap confidence intervals - Sample S6

Table 10: 95% Confidence Limits for Monthly Indices

Mês	Variation (%)	Lower Limit	Upper Limit
Jan/16	-0.93997576	-1.0850158	-0.79493576
Feb/16	-1.18498173	-1.8455017	-0.52446173
Mar/16	-0.89217714	-1.0489771	-0.73537714
Apr/16	-0.78083960	-0.9905596	-0.57111960
May/16	-0.42403444	-0.6788344	-0.16923444
Jun/16	-1.20455500	-1.5397150	-0.86939500
Jul/16	-0.01156223	-0.3173222	0.29419777
Aug/16	-0.05720701	-0.3806070	0.26619299
Sep/16	-0.61173888	-1.1605389	-0.06293888
Oct/16	-1.88321976	-2.6201798	-1.14625976
Nov/16	-0.38051861	-0.6196386	-0.14139861
Dec/16	-0.58491403	-0.7652340	-0.40459403

6. Conclusion

The objective in this work was to present a innovative technique to analyze house price indices in a longitudinal perspective. One important hurdle faced by analysts when studying house prices is the infrequency of repeat sales of house and, consequently, time series and other statistical techniques are discarded in many studies. The proposed matching of

dwellings by adopting the pseudo housing units approach suggested here allows analyses in a longitudinal context. Furthermore, mixed effects models and the estimation procedures adopted, such as restricted maximum likelihood, complemented the framework in a formal way. Although the results were not of primary interest here, monthly double imputation Laspeyres house price indices were estimated with very satisfied outcomes.

References

Bailey, M. J., Muth, R. F., and Nourse, H. O. (1963). A regression method for real estate price index construction. *Journal of the American Statistical Association*, 58(304):933–942.

Case, B. and Quigley, J. (1991). The dynamics of real estate prices. *The Review of Economics and Statistics*, 73(1):50–58.

Case, K. E. and Shiller, R. J. (1989). The Efficiency of the Market for Single-Family Homes. *American Economic Review*, 79(1):125–137.

Clapp, J. M. and Giaccotto, C. (1994). The influence of economic variables on local house price dynamics. *Journal of Urban Economics*, 36(2):161 – 183.

Deng, Y., McMillen, D., and Sing, T. (2014). Matching indices for thinly-traded commercial real estate in singapore. *Regional Science and Urban Economics*, 47(1):86–98.

Diewert, W. E. (2009). The paris oecd-imf workshop on real estate price indexes: Conclusions and future directions. In *Price and Productivity Measurement*, volume 01, pages 87–116.

Dorsey, R. E., Hu, H., Mayer, W. J., and chen Wang, H. (2010). Hedonic versus repeat-sales housing price indexes for measuring the recent boom-bust cycle. *Journal of Housing Economics*, 19(2):75 – 93.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, 7(1):1–26.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2011). *Applied Longitudinal Analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., second edition.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis, Third Edition (Chapman & Hall/CRC Texts in Statistical Science)*. Chapman and Hall/CRC.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel / Hierarchical Models*. Cambridge University Pr.

Glennon, D., Kiefer, H., and Mayock, T. (2018). Measurement error in residential property valuation: An application of forecast combination. *Journal of Housing Economics*, 41:1–29. Exported from https://app.dimensions.ai on 2019/01/30.

Goetzmann, W. and Peng, L. (2002). The bias of the rsr estimator and the accuracy of some alternatives. *Real Estate Economics*, 30(1):13–39.

Griliches, Z. (1971). *Price Indexes and Quality Change: Studies in New Methods of Measurement*. Harvard Univ. Press.

Guo, X., Zheng, S., Geltner, D., and Liu, H. (2014). A new approach for constructing home price indices: The pseudo repeat sales model and its application in china. *Journal of Housing Economics*, 25(Supplement C):20 – 38.

Hedeker, D. and Gibbons, R. D. (2006). *Longitudinal Data Analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc.

Kain, J. F. and Quigley, J. M. (1970). Measuring the value of housing quality. *Journal of the American Statistical Association*, 65(330):532–548.

Nagaraja, C. H., Brown, L. D., and Zhao, L. H. (2011). An autoregressive approach to house price modeling. *Ann. Appl. Stat.*, 5(1):124–149.

OECD, ILO, IMF, UNECE, and the World Bank (2013). *Handbook on Residential Property Prices Indices (RPPIs)*. INTERNATIONAL MONETARY FUND.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2017). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-131.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sedgley, N., A. Williams, N., and W. Derrick, F. (2008). The effect of educational test scores on house prices in a model with spatial dependence. *Journal of Housing Economics*, 17:191–200.

Shiller, R. J. (1991). Arithmetic repeat sales price estimators. *Journal of Housing Economics*, 1(1):110 – 126.

Silver, M. and Graf, B. (2014). *Commercial Property Price Indexes: Problems of Sparse Data, Spatial Spillovers, and Weighting*. IMF Working Papers. INTERNATIONAL MONETARY FUND.

Wang, F. T. and Zorn, P. M. (1997). Estimating house price growth with repeat sales data: What's the aim of the game? *Journal of Housing Economics*, 6(2):93–118.