Journal Title
XX(X):1–18
©The Author(s) 2020
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Exploring a Multilevel Approach with Spatial Effects to Model Housing Price in San José, Costa Rica

Eduardo Pérez-Molina[1]

Abstract

A multilevel model of the housing market for San José Metropolitan Region (Costa Rica) was developed, including spatial effects. Hierarchical relations (lower level units nested into higher level) were modeled by specifying multilevel models with random intercepts and a conditional autoregressive (CAR) term to include spatial effects from neighboring units at the higher level (districts). The random intercepts and conditional autoregressive models presented the best fit to the data. Variation at the higher level accounted for 16% of variance in the random intercepts model and 28% in the CAR model. The sign and magnitude of regression coefficients proved remarkably stable across model specifications. Multilevel and CAR models represented an important improvement in modeling housing price, despite most of the variation still occurring at the lower level, by improving the overall model fit and expanding the interpretation of model results. However, the CAR specification only represented a limited advance over the random intercepts formulation.

Keywords

Multilevel, Conditional autoregressive model, Housing prices, San José-Costa Rica

[1] Programa de Investigación en Desarrollo Urbano Sostenible, Universidad de Costa Rica

Corresponding author:

Eduardo Pérez Molina, Universidad de Costa Rica Programa de Investigcioón en Desarrollo Urbano Sostenible, Bo Los Profesores, Calle B, No. 13, Mercedes, Montes de Oca, 115103, Costa Rica.
Email: eduardo.perezmolina@ucr.ac.cr

Introduction

Real estate property has long been conceived as a composite good (Kain and Quigley 1970; Cheshire and Sheppard 1995) for which the total price can be decomposed into hedonic prices, determined each by separate characteristics of the property (Rosen 1974). Real estate price models assume the most relevant components are structural features of the building, locational qualities of the plot, location specific environmental aspects, and social and neighborhood traits (Anselin and Lozano-Gracia 2009). These characteristics can be additionally differentiated as either compositional or contextual (Liu and Roberts 2012). Compositional elements are determined by the property itself, i.e., the structural features of the building and locational attributes of the plot (particularly accessibility). Contextual refers to the characteristics of the area immediately surrounding the location (including environmental and social neighborhood characteristics).

The main objective of this paper is to characterize the housing market of the San José Metropolitan Region in Costa Rica. Such aim involves two dimensions: on the one hand, a hedonic price model can decompose the contributions of diverse factors to the overall housing price variability. On the other, spatial dependence and spatial heterogeneity occurring simultaneously at different scales call for careful consideration when specifying quantitative models of property prices; indeed, to estimate the hedonic price model with data requires this specification to avoid bias in the results.

Two research questions have been proposed to organize the resarch reported in this paper: (1) To what extent do contextual and compositional effects of location contribute to explain variation in the housing prices? (2) How does the introduction of multilevel effects change other theoretically key covariates, in particular travel time to the central business district (CBD)?

In order to explore these issues, a set increasingly complex multilevel models of housing price has been proposed and estimated for a spatially explicit data set of 1264 real estate listings of the San José Metropolitan Region (see supplemental information for a description of the data).

The simplest model considered (*VPM*), a variance partitioning model (Chi et al. 2020), quantifies the average value of housing price in each level considered. Three levels were introduced: property listing locations nested into districts, in turn nested into municipalities. The second model was specified as multilevel with property listings nested into districts, for which a random effect (*MLRE*) in the intercept was calculated. This model included both contextual and compositional determinants of housing price. Finally, a multilevel model with covariantes, district level random intercepts, and a CAR term considering adjacency between districts (*CARM*) was calculated, following Osland et al. (2016) and Bivand et al. (2017). This model was designed to capture any effects at a larger geographical extent (i.e. neighboring districts' characteristics affecting price at all locations within any given district). A Gaussian response model (*GaLM*) without spatial effects was used as benchmark. Random slopes models, with CAR, as proposed by Dong et al. (2016), and without, following Gelfand et al. (2007), were discarded because of poor fit (the main problem being the large amount of effective parameters, relative to the available data).

While multilevel models have long been advocated as useful for hedonic price modeling of real estate (Jones 1991), they do not fully account for spatial heterogeneity problems (Chasco and Le Gallo 2012; De Aguiar et al. 2014). Specifically, interdependence has been found between the zonal random effects of multilevel models (Bivand et al. 2017), a problem of special relevance when considering continuous spatial data such as housing prices. Conditional autoregressive (CAR) and simulatenously autoregressive (SAR) processes have been proposed to account for this bias (Dong and Harris 2015; Osland et al. 2016; Bivand et al. 2017). Some earlier examples of spatial smoothing in the random effect or coefficients of multilevel models do exist (Gelfand et al. 2007; Brunauer et al. 2013); but given the usual structure of housing price data and a relatively long history of CAR applications to spatial problems (Lawson 2013), the paucity of hedonic models of real estate prices with CAR is surprising (Bivand et al. 2017).

The main difference between models was found in the fit to the data as measured by the deviance information criterion, DIC (with the *GaLM* being the weakest model). The *MLRE* and *CARM* models resulted in the best fit to the data, with hardly any difference between them in terms of deviance and effective parameters. Regression coefficients were robust to model specifications. Key determinants, in particular travel time to the CBD, showed expected signs and were significant for all models that included covariates. The main effect of introducing more complex spatial structures, through multilevel and CAR specifications, was to slightly weaken travel time to CBD, distance to nearest bus stop, and number of jobs per district as predictors of housing price –which suggests these variables were, to a certain extent, acting as proxies for urban dynamics occurring at the higher (district) level. Finally, the improvements introduced by the CAR specification, relative to the multilevel random intercepts model *MLRE*, were limited.

Background

The analysis of housing prices with hedonic models requires the use of spatially explicit data (Anselin 1998), to control for bias introduced by spatial heterogeneity and, more importantly, because the key drivers of housing price account for differences in space. The Alonso-Mills-Muth model (Glaeser 2008) describes the basic mechanism through which housing prices arise (as a result from multiple interactions between urban agents) in a city: by means of a trade-off for consumers between accessibility to the CBD and living area; properties with poor accessibility to the CBD must compensate urban agents *via* larger living areas. In addtition, environmental and social externalities, i.e. the contextual characteristics, have been found to capitalize into land values and, hence, into housing prices (Grieson and White 1989; Henneberry and Barrows 1990).

Furthermore, real estate market data is often structured in a spatial hierarchy (Jones 1991; Dong et al. 2015): locations of housing units are grouped into larger spatial units such as neighborhoods, in turn sorted into municipalities, regions, and countries. This nested hierarchical grouping introduces levels into the relationships exhibited between variables of the data and, since the groups are spatial, the multilevel effects can be spatially interpreted.

The information contributed by the hierarchical structure can be introduced into hedonic price models by way of multilevel specifications (Jones 1991; Gelfand et al. 2007; Bivand et al. 2017). They can be used to include regression intercepts or slopes that vary for each of the different higher order spatial units (Jones 1991). Examples of previous applications of multilevel modeling to real estate property include random intercept models by Brunauer et al. (2013); Osland et al. (2016); Bivand et al. (2017); Nordvik et al. (2019) and the random slope models of Jones (1991); Gelfand et al. (2007); Chasco and Le Gallo (2012); De Aguiar et al. (2014).

The sources of spatial variation, as described previously, include both contextual and compositional effects (Liu and Roberts 2012). From an econometric perspective, this poses the need to account for spatial dependence and spatial heterogeneity in the data generating process (Anselin and Lozano-Gracia 2009) at different scales. While it has been argued that multilevel specifications could be sufficient to control for these effects, Chasco and Le Gallo (2012) tested a three level model with spatially lagged determinants (of a cross section of housing prices in Madrid, Spain) and found that despite a better performance in explaining variability, spatial dependence persisted in the model. De Aguiar et al. (2014), working with six years of property transactions in Belo Horizonte, Brazil, reached a similar conclusion after fitting a two level random slopes model that includes the effects of a quality of life index for the city.

Another approach to spatial dependence and spatial heterogeneity is to include various spatial autocorrelation structures in the multilevel random effect: through multilevel CAR models (Osland et al. 2016; Bivand et al. 2017; Nordvik et al. 2019), multilevel SAR models (Dong et al. 2015; Dong and Harris 2015; Cellmer et al. 2019), multivariate kernel convolution (Gelfand et al. 2007), or P-splines (Brunauer et al. 2013). These spatial effects have generally been specified for the higher level of the model (autocorrelation among neighborhoods) and as additive components of the random effect, save for Gelfand et al. (2007) who constructed a random slopes model of Singapore's apartments market.

In the context of these developments, this paper contributes to the literature by applying the multilevel models previously developed by Osland et al. (2016) for Stavanger in Norway, and extended by Bivand et al. (2017) and Nordvik et al. (2019), to a more complex urban system, the San José Metropolitan Region of Costa Rica. This polycentric and larger metropolitan area, in a middle income Latin American country (and consequently with larger social polarization, relative to Norway), can be used to consider the suitability of the model for a different spatial structure in the housing market (and also for substantive analysis of San José's housing market, which has not been systematically tackled to date).

Methodology and Data

Geographical Setting

The San José Metropolitan Region comprises four cities (San José, Alajuela, Cartago, and Heredia) with diverse functional and physical links between them (Pujol-Mesalles

2005). It is located at coordinates $9°56'N$ and $84°5'W$, in a tectonic and volcanic depression with important variation in elevation, climate, and other physical variables (Bergoeing 2017). The canyon of the Virilla river, the Carpintera volcanic-sedimentary mountain range, and more generally rivers and irregular relief, act as physical barriers to connectivity between the different human settlements, which contributes to explain its polycentric and dispersed human activity patterns (Pujol-Mesalles 2005). Urban growth in the region has been traditionally dispersed and has shown some signs of consolidation over the last 15 years (Pujol-Mesalles and Pérez-Molina 2012), although in part because of a very sparse baseline.

Administratively, the boundaries of the San José Metropolitan Region include an area of $1779km^2$, of which 24.8% are within its urban growth boundary (Consejo Nacional de Planificación Urbana 2013) and substantial exurban development is permitted by land use regulations beyond this limit. The region is divided into 31 municipalities, in turn subdivided into 166 districts (in 2017). There is no regional level governance structure in San José. Land use planning takes place at the municipal level with little coordination between different jurisdictions, although a regional plan does exist defined by the national government (Madrigal and Pérez 2012).

According to the latest census (2011), the region had a population of 2.2 million people, approximately half in the metropolitan area of San José and the other half equally distributed among Alajuela, Cartago, and Heredia. Population growth has been declining from a yearly rate of 3.0% for 1984-2000 to 0.85% for 2000-2011 (Consejo Nacional de Planificación Urbana 2013). While no direct income measures exist for the metropolitan area, yearly average household income for the Central Region (82% of its population lived, in 2011, within the San José Metropolitan Region) had been estimated at $US\$24915$ per household with a yearly growth rate of 2.95% for nominal income, 2010-2019 (Centro Centroamericano de Pobación 2012; Instituto Nacional de Estadística y Censos 2020).

In synthesis, while population growth likely declined over the last decade, income increased substantially in the region. As a consequence, one should have expected from the Alonso-Mills-Muth model an expansion of the region's urban footprint and a generalized increase of land rent –and therefore housing price– for San José; see the static comparative analysis of Brueckner (1987). From a spatial perspective, one may hypothesize income and other variables strongly related to it (like transportation) to cause differences in housing prices.

Data for Hedonic Price Modeling

Data on housing prices was compiled from an aggregator website of real estate listings. The full description of the data base compilation process can be found in the annexed supplemental material. A total of 1264 real estate listings of detached houses were available from the data. All records have data on house price, living area, and plot area. The categorization was volunteered by the users of the site. It is important to note, as Chasco and Le Gallo (2012) also do, that real estate listings represent an approximation to actual transaction values; they can be thought of as both the expectation of the seller

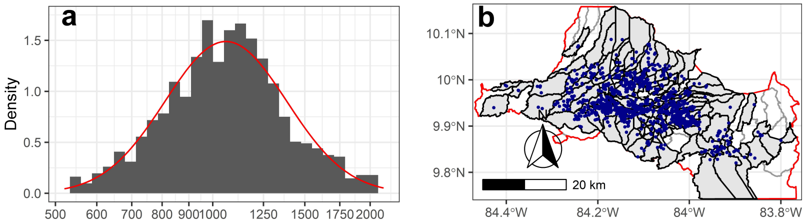

Figure 1. Housing real estate listings in the San José Metropolitan Region: location and histogram of price per m^2 of living area.

and as an opening bid for sale price negotiations. However, the spatial structure of this expectation should correspond to some extent with final sale price.

Housing price divided by living area was chosen as the dependent variable, to be explained by accessibility to the CBD and other covariates. Both total price, controlled for living area, and price per square meter of living area have been used as dependent variables in hedonic price modeling. The choice of variable can influence the results, as discussed by Chi et al. (2020) –they argue for normalized values as more accurate. Housing price normalized per living area was selected becuase of theoretical considerations, as this is the form of price discussed in the Alonso-Mills-Muth theory of urban location because the size of the dwelling is a decision of urban agents (Glaeser 2008).

Figure 1.a shows the log-transformed histogram of standardized housing price, which resembles a lognormal distribution with a mean of 1098.3 and standar deviation of 300.2 (see table 1). When interpreted in the context of the region's household income, these values (in combination with the housing unit's size, measured by the living area) point to a pervasive affordability problem in the regional housing market (Agüero-Valverde et al. 2020). Given a recent context of rising income for the higher income households and increasing marginal cost of transport, it is perhaps unsurprising housing values are relatively high. Real estate listings are spatially concentrated along the East-West axis that forms the central corridor along which San José has developed and which through is crossed by the main roads of the San José Metropolitan Region (see figure 1.b).

The data representing the selected determinants is summarized in table 1. These can be grouped into three categories. First, variables of accessibility (in particular, accessibility to the CBD), which form the core of the theoretical argument built around the Alonso-Mills-Muth model. Second, one variable (house plot area) controlling for a characteristic prized by consumers. Plot area has been conceived as an endogenous characteristic within the Alonso-Mills-Muth model (Brueckner 1983); however, in practice and because the spatial equilibrium is imperfectly materialized, this trait must be controlled for. Third, characteristics of the neighborhoods –as represented by the district data and which are capitalized into the housing price: crime rate, number of jobs in the district, and a proxy

Table 1. Descriptive statistics of housing property prices and selected determinants.

Variable	Mean	Std. Dev.	Min.	Max.
Housing price per living area ($US\$/m^2$)	1098.3	300.2	520.3	2122.6
House plot area	581.9	1186.0	46.8	20000.0
Estimated travel time to CBD (min)	15.9	8.6	0.7	56.0
Euclidean distance to nearest major road (m)	1331.3	1402.6	0.0	9427.9
Euclidean distance to nearest bus route (m)	238.6	371.1	0.0	3354.1
Crime rate (homicides per 100 thousand residents in 2018)	10.4	12.6	0.0	80.2
Number of jobs in district (for 2018)	9192	10536.9	3	32887
Number of households with unsatisfied housing needs (Census 2011)	1243	1787.1	33	10120

variable for poverty. The poverty proxy variable is the number of unsatisfied needs for housing from the 2011 census (Méndez and Bravo 2014). Further details on variable sources can be found in the supplemental material.

Model Specification and Estimation

The data that describes the housing price and its determinants has been structured into the linear model (*GaLM*) as follows, for records at locations i (lower level units) and districts j (higher level units):

$$\begin{aligned} y_i &\sim \mathcal{N}\left(\mu_i,\, \sigma_e^2\right) \\ \mu_i &= \Sigma_k \beta_k X_{ik} + \Sigma_p \gamma_p W_{i*p} \end{aligned} \tag{1}$$

with y_i the natural logarithm of housing price per living area, in $US\$/m^2$, and where x_{i1} is the natural logarithm of the house plot area, x_{i2} the natural logarithm of the estimated travel time to the CBD, x_{i3} the square of x_{i2}, x_{i4} the natural logarithm of Euclidean distance to main roads, x_{i5} the natural logarithm of Euclidean distance to bus routes; w_{j0} is the intercept, w_{j1} the crime rate (homicides per 100 thousand residents for each district), w_{j2} the number of jobs in the district, and w_{j3} a poverty proxy variable, as described previously (w_{j1}, w_{j2}, w_{j3} were all transformed into their natural logarithm; since w_{j1} contains 0.00 for many districts, 0.001 was added to all values). β_k and γ_p are, respectively, the coefficients for variables with variation at the lower level (β_k) and higher level only (γ_p). The subscript $i*$ denotes the district j in which the record i is located. The errors are assumed to be independent and identically distributed (i.i.d.) and to follow a normal distribution with mean 0.00 and variance σ_e^2.

Two important points should be raised with regard to the *GaLM*. First, the effect of accessibility to the CBD in San José has been specified as non-linear (it has a quadratic form). This is thought to be a characteristic of the urban system that is being described. Second, three variables (w_{j1}, w_{j2}, w_{j3}) only have variation between districts: records within the same spatial unit all have the same level of these variables. This introduces a potential bias into the *GaLM*, depending on how large the within-district association is (this is the bias that should be controlled for with the multilevel specification).

The *GaLM* was extended to include correlation among records belonging to the same district, the multilevel effect that encodes the spatial variation. The spatial multilevel effect *MLRE* can be incorporated through the definition of an additive term η: a random effects model with random intercepts was specified by defining η as following a systematic component, $\Sigma_p \gamma_p W_{jp}$ and a normally distributed term u_j with mean 0.00 and variance σ_u^2:

$$\begin{aligned} y_i &\sim \mathcal{N}\left(\mu_i, \sigma_e^2\right) \\ \mu_i &= \Sigma_k \beta_k X_{ik} + \Delta \eta_j \\ \eta_j &= \Sigma_p \gamma_p W_{jp} + u_j \\ u_j &\sim \mathcal{N}\left(0, \sigma_u^2\right) \end{aligned} \tag{2}$$

The *CARM* model was specified as an extension of model *MLRE*. The model follows the same equation 2, except u_j is defined as following an intrinsic multivariate CAR. The prior distribution for the u_j parameter was defined as in Osland et al. (2016):

$$u_j | u_{-j} = \mathcal{N}\left(\frac{\Sigma_{q\ j} weight_{jq} u_q}{weight_{j+}}, \frac{1}{weight_{j+} \tau_u}\right) \tag{3}$$

where $q \sim j$ designates the q higher level units (districts) that are neighbors of district j, $weight_{jq}$ is a value equal to 1.00 (since equal weights are assumed for all neighbors), and $weight_{j+}$ is the sum of all weights of the neighbors of j.

The *VPM* model is a simplification of equation 2: it follows the same form, except no covariates were included, neither at the lower (property) level nor at the district level. However, a third (higher) geographical unit, the municipality, was introduced:

$$\begin{aligned} y_i &\sim \mathcal{N}\left(\mu_i, \sigma_e^2\right) \\ \mu_i &= \beta_0 + u_j + v_l \\ u_j &\sim \mathcal{N}\left(0, \sigma_u^2\right) \\ v_l &\sim \mathcal{N}\left(0, \sigma_s^2\right) \end{aligned} \tag{4}$$

with v_l a municipal level random intercept for the 30 municipalities with data (out of 31 that form the region).

The parameters to be estimated in the models defined in equations 1 to 4 are: β_k for the coefficients corresponding to variables with variation at the lower level, γ_p for the coefficients corresponding to variables with variation at the higher level, σ_e^2 the lower level i.i.d. residuals, and σ_u^2 or σ_s^2, the dispersion measures at the higher (district and municipal) level.

The β_k and γ_p coefficients were modeled using uninformative normal distributions as priors. The residuals, in turn, were modeled through a hyperprior τ_e equal to $1/\sigma_e^2$. The hyperprior follows an uninformative Gamma distribution. σ_u^2, the higher level variance which is present in the three multilevel models, and σ_s^2 (of the *VPM* model, equation

Table 2. Model comparison criteria.

	GaLM	*MLRE*	*CARM*	*VPM*
Mean Deviance	$9.363 \cdot 10^1$	$-7.692 \cdot 10^1$	$-5.542 \cdot 10^1$	9.045
Deviance std. error	$2.002 \cdot 10^{-2}$	$8.721 \cdot 10^{-2}$	$7.37 \cdot 10^{-1}$	$1.140 \cdot 10^{-1}$
pD	$1.002 \cdot 10^1$	$7.194 \cdot 10^1$	$5.683 \cdot 10^1$	$6.179 \cdot 10^1$
DIC	$1.037 \cdot 10^2$	-4.987	1.415	$7.083 \cdot 10^1$
Mean VPC	–	$1.664 \cdot 10^{-1}$	$2.838 \cdot 10^{-1}$	$9.127 \cdot 10^{-2}$
VPC std. error	–	$1.827 \cdot 10^{-4}$	$2.745 \cdot 10^{-4}$	$2.832 \cdot 10^{-4}$

The VPC estimate for the *VPM* model refers only to the district level variation (for comparison purposes); mean VPC for both levels in this model is $2.031 \cdot 10^{-1}$ with a standard error of $2.145 \cdot 10^{-4}$.

4, that formalizes the municipal level variance) were also modeled with hyperpriors τ_u equal to $1/\sigma_u^2$ and τ_s equal to $1/\sigma_s^2$ that follow uninformative Gamma distributions.

Models were estimated using WinBUGS (Spiegelhalter et al. 2003), called through the R2BUGS package (Sturtz et al. 2005). Three chains of 400000 iterations each were simulated with the first 25000 iterations discarded as burn-in. The resulting chains were thinned by selecting one out of every 25 iterations to reduce autocorrelation problems. The final posterior samples consisted of three chains of 15000 simulated instances each. Convergence was evaluated through visual inspection of the MCMC trace plots for the model parameters, autocorrelation plots, and the Gelman and Rubin diagnostic (these plots are annexed in the supplemental material; for models with multilevel effects, six randomly selected parameters are reported as illustration, although all parameters were visually inspected for convergence and can be consulted in the full model output files available online, see supplemental material).

Results

Model Estimates

The results for the estimated models are presented in tables 2 (model fit) and 3 (parameter means and 95% posterior intervals), and figure 2 (patterns for systematic variation at the higher level units).

The DIC of the four models is shown in figure 2. Two models (*MLRE* and *CARM*) have substantially lower DIC values than the other two (*GaLM* and *VPM*). The two best models do not differ markedly: in terms of fit alone, the higher order CAR structure did not improve the model fit. Additionally, the two worst models (*GaLM* and *VPM*) in terms of DIC also are broadly similar, which suggests the spatial multilevel structure may to a large extent control for the spatial effects determined by the covariates.

In table 2, the variance partitioning coefficient (VPC) is also reported. This is the fraction of variance ascribed to the higher (σ_u^2, district) or lower (σ_e^2) level units; specifically, the VPC is defined as the sum of σ_u^2 and σ_e^2 into the higher level variance, σ_u^2. As can be seen in figure 2, the *GaML* model includes by construction all of the variance in the lower level, because of the absence of a σ_u^2 term. More interestingly, the

Table 3. Regression model results for natural logarithm of housing price per m^2 of living area (mean, 2.5% and 97.5% percentiles).

	Mean	2.5% perc.	97.5% perc.	Mean	2.5% perc.	97.5% perc.
		GaML			*MLRE*	
β_1	$7.85 \cdot 10^{-2}$	$6.33 \cdot 10^{-2}$	$9.36 \cdot 10^{-2}$	$7.20 \cdot 10^{-2}$	$5.59 \cdot 10^{-2}$	$8.80 \cdot 10^{-2}$
β_2	$6.13 \cdot 10^{-1}$	$4.26 \cdot 10^{-1}$	$8.01 \cdot 10^{-1}$	$5.06 \cdot 10^{-1}$	$2.64 \cdot 10^{-1}$	$7.56 \cdot 10^{-1}$
β_3	$-1.13 \cdot 10^{-1}$	$-1.48 \cdot 10^{-1}$	$-7.79 \cdot 10^{-2}$	$-9.45 \cdot 10^{-2}$	$-1.42 \cdot 10^{-1}$	$-4.91 \cdot 10^{-2}$
β_4	$-1.11 \cdot 10^{-2}$	$-2.07 \cdot 10^{-2}$	$-1.47 \cdot 10^{-3}$	$-1.20 \cdot 10^{-2}$	$-2.24 \cdot 10^{-2}$	$-1.65 \cdot 10^{-3}$
β_5	$1.12 \cdot 10^{-2}$	$4.63 \cdot 10^{-3}$	$1.78 \cdot 10^{-2}$	$8.40 \cdot 10^{-3}$	$1.83 \cdot 10^{-3}$	$1.50 \cdot 10^{-2}$
γ_1	5.71	5.43	6.00	5.93	5.53	6.34
γ_2	$4.14 \cdot 10^{-4}$	$-3.36 \cdot 10^{-3}$	$4.19 \cdot 10^{-3}$	$-6.36 \cdot 10^{-5}$	$-6.42 \cdot 10^{-3}$	$6.41 \cdot 10^{-3}$
γ_3	$1.88 \cdot 10^{-2}$	$9.34 \cdot 10^{-3}$	$2.83 \cdot 10^{-2}$	$2.03 \cdot 10^{-2}$	$4.06 \cdot 10^{-3}$	$3.67 \cdot 10^{-2}$
γ_4	$-1.95 \cdot 10^{-2}$	$-3.46 \cdot 10^{-2}$	$-4.27 \cdot 10^{-3}$	$-2.59 \cdot 10^{-2}$	$-5.42 \cdot 10^{-2}$	$2.14 \cdot 10^{-3}$
σ_u^2	–	–	–	$1.11 \cdot 10^{-2}$	$6.25 \cdot 10^{-3}$	$1.76 \cdot 10^{-2}$
σ_e^2	$6.31 \cdot 10^{-2}$	$5.83 \cdot 10^{-2}$	$6.81 \cdot 10^{-2}$	$5.51 \cdot 10^{-2}$	$5.07 \cdot 10^{-2}$	$5.98 \cdot 10^{-2}$
		CARM			*VPM*	
β_0	–	–	–	6.92	6.88	6.96
β_1	$6.96 \cdot 10^{-2}$	$5.38 \cdot 10^{-2}$	$8.54 \cdot 10^{-2}$	–	–	–
β_2	$3.83 \cdot 10^{-1}$	$1.18 \cdot 10^{-1}$	$6.47 \cdot 10^{-1}$	–	–	–
β_3	$-7.23 \cdot 10^{-2}$	$-1.25 \cdot 10^{-1}$	$-1.98 \cdot 10^{-2}$	–	–	–
β_4	$-1.13 \cdot 10^{-2}$	$-2.18 \cdot 10^{-2}$	$-7.23 \cdot 10^{-4}$	–	–	–
β_5	$8.06 \cdot 10^{-3}$	$1.60 \cdot 10^{-3}$	$1.45 \cdot 10^{-2}$	–	–	–
γ_1	6.07	5.67	6.48	–	–	–
γ_2	$8.59 \cdot 10^{-4}$	$-4.55 \cdot 10^{-3}$	$6.22 \cdot 10^{-3}$	–	–	–
γ_3	$1.08 \cdot 10^{-2}$	$-4.37 \cdot 10^{-3}$	$2.61 \cdot 10^{-2}$	–	–	–
γ_4	$-1.17 \cdot 10^{-2}$	$-3.49 \cdot 10^{-2}$	$1.14 \cdot 10^{-2}$	–	–	–
σ_s^2	–	–	–	$8.438 \cdot 10^{-3}$	$2.163 \cdot 10^{-3}$	$1.837 \cdot 10^{-2}$
σ_u^2	$2.28 \cdot 10^{-2}$	$1.11 \cdot 10^{-2}$	$4.03 \cdot 10^{-2}$	$6.798 \cdot 10^{-3}$	$1.863 \cdot 10^{-3}$	$1.430 \cdot 10^{-2}$
σ_e^2	$5.60 \cdot 10^{-2}$	$5.16 \cdot 10^{-2}$	$6.08 \cdot 10^{-2}$	$5.899 \cdot 10^{-2}$	$5.427 \cdot 10^{-2}$	$6.410 \cdot 10^{-2}$

variance of the random intercept models (with and without CAR) is relatively small but not unimportant, with a VPC for *MLRE* of 0.16 and a larger value of 0.28 for *CARM* (almost twice as much); if only the district levels are included, as in the *VPM* model, 0.09 of the total variance can be explained –a figure that doubles when considering also a second, higher level of municipalities.

The VPC estimates are in general lower than the equivalents estimated by Bivand et al. (2017) when replicating models by Osland et al. (2016) (a VPC in the order of 0.58 for Stavanger, Norway, although with accessibility to the CBD being a higher level covariate) and Dong et al. (2015) (a VPC of 0.38 for Beijing), but the difference is reasonable and should be ascribed to the complexity of a polycentric region and to a very local effect of spatial dependence, likely present in San José and not incorporated into the models.

Estimated coefficients for the four proposed models are shown in table 3. Overall, travel time to San José is the most important determinant (about an order of magnitude greater than other coefficients); the 95% posterior intervals of both β_2 and β_3 exclude 0.00 for all models and are consistent in their signs (positive for β_2 and negative for β_3, see table 3). Taken jointly, these two coefficients suggest the relation between housing price and accessibility to the CBD follows a concave curve: it rises near the CBD, where the urban environment is physically deteriorated –particularly in the municipality of San José proper– and at some point, the slope changes downward, reflecting the Alonso-Mills-Muth model's trade-off between accessibility and land rent.

In general, determinants with variation at the lower level (β_k) are more significant and larger in magnitude than variables with variation at the higher level (γ_p), as can be seen in table 3. For models *GaLM*, *MLRE*, and *CARM*, no β_k coefficient has 0.00 within its 95% posterior interval whereas half the γ_p coefficients do –and the γ_k coefficients for the *MLRE* and *CARM* models, which are a better fit for the data, are as a rule not significant. However, it is important to note that accessibility, the most important set of determinants in the estimated models, presents variation at the lower (individual property) level. Thus, this result coincides with previous findings by Osland et al. (2016), who found socioeconomic variables to be relatively weak but accessibility to become stronger explanations of housing price variance when accounting for spatial dependency (albeit their accessibility measure was at the zonal, not individual level). A proper accounting of hierarchical structure (and spatial dependence) clarifies the role of theoretically key covariates, regardless of their level in the hierarchy.

When comparing coefficients across model specifications, magnitudes remain remarkably stable with three important exceptions: travel time to CBD (β_2 and β_3), distance to nearest bus stop (β_5), and number of jobs in the district (γ_3). The first two determinants decrease in magnitude as more spatial effects are considered: they are, in absolute terms, largest for the *GaLM* model and smallest for the *CARM* model, with the coefficients of the *MLRE* model being intermediate. Number of jobs has larger coefficients for the *GaLM* and *MLRE* models, decreasing for the *CARM* model. This characteristic suggests accessibility was serving to control for certain urban patterns, particularly for socioeconomic self-sorting: it is very likely the best districts (higher level units) in terms of accessibility attract both firms (and jobs) and higher income households, which in turn would result in a more attractive spatial context.

Finally, with regard to the *VPM* model, one should note σ_s^2 (municipal level variance) and σ_u^2 (district level variance) present similar values. The multilevel variance in the higher levels should be, therefore, apportioned between these two levels. An evident extension of the *MLRE* and *CARM* models will be to include municipal level random effects and covariates, although the main contribution of these is likely to explanation of the housing price rather than prediction, since most of this variations is already controlled for in the district level random effects (crucially, land use regulation is a municipal level covariate not currently included in the models).

The estimated mean posterior of η for the *MLRE* and *CARM* models are shown in figure 2. In addition, the random effect deviations from the overall mean, at municipal and district level, of the *VPM* model are also included. The larger values are shown in dark red and the lower values, in dark blue, with intermediate values in white. Non-significant parameters (with 0.00 in the 95% posterior interval) are shown in yellow. Only districts and municipalities with data are represented in figure 2.

The η represent the total systematic variation, caused by the higher level covariates and the random effect. Additionally, they represent the variation unexplained by the lower level determinants (which is to say, accessibility and house size). They are broadly

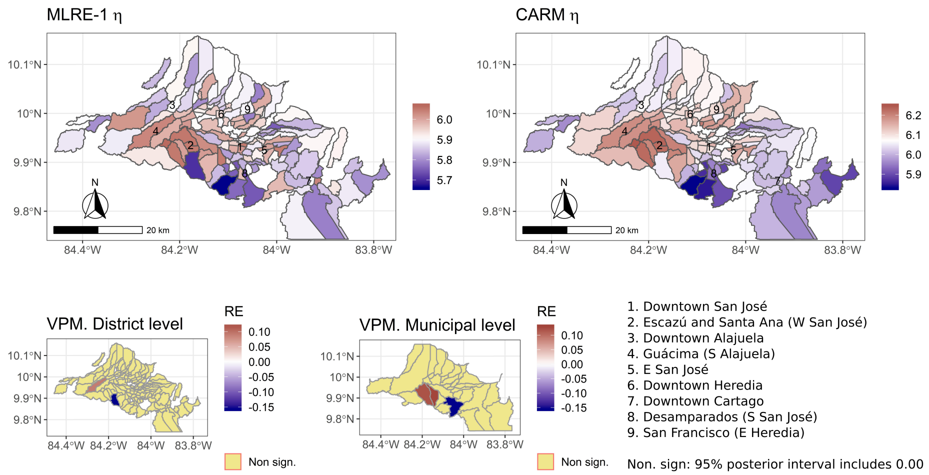

Figure 2. η for *MLRE* and *CARM* models and random effects for *VPM* model (mean values).

consistent in identifying housing market hotspots (larger values of η values for *MLRE* and *CARM*, in dark red in figure 2) and areas of lower urban activity and where housing price is less which, consequently, show up as dark blue in the η maps of figure 2.

Housing market hotspots in the San José Metropolitan Region include western San José (Escazú and Santa Ana, point 2 in figure 2), Guácima and San Rafael in southern Alajuela (point 3 in figure 2), San Pedro, Curridabat, and Tres Ríos in eastern San José (point 5 in figure 2), and to a lesser extent downtown San José (the traditional CBD, point 1 in figure 2) and eastern Heredia (the traditional CBD, point 9 in figure 2). All of these coincide with districts that, already a decade ago, concentrated residential building activity and were agglomerations of higher income households (Pujol-Mesalles et al. 2012).

Areas of lower housing prices, in turn, are located in southern San José (Desamparados, Alajuelita, and Aserrí, point 8 in figure 2) and the city of Cartago (downtown Cartago being identified by point 5 in figure 2). The former are agglomerations of low income households formed during the 1980s crisis, originally informal and largely formalized through social housing programs (Pujol-Mesalles et al. 2014). The latter is likely explained by its poor "macro" accessibility: Cartago, of the four cities that comprise the region, is the only one located in a basin that drains towards the Caribbean Sea and is separated from the rest of the region by the Carpintera mountains (Pujol-Mesalles 2005; Bergoeing 2017).

The district and municipal random effects of the *VPM* are mostly not significant (0.00 is part of the 95% posterior interval for most spatial units). The exceptions are two districts (Salitral and Guácima) and three municipalities, of which two –Santa Ana and

Escazú, in red in the *VPM* municipal level random effects map– are real estate market hotspots and the third, Desamparados, is an area of lower prices. Of the districts, Salitral in Santa Ana is the only area of low housing prices within that municipality (in part because of natural hazard risks) and Guácima in Alajuela (in red among the *VPM* district level random effects) is a concentration of relatively high income households. In this regard, it is perhaps notable that other such districts and municipalities, especially in eastern and central San José, do not show statistically significant deviations from the overall regional housing price mean.

Model Evaluation

The estimated models have allowed for a description of the role of accessibility in forming housing prices in the San José Metropolitan Region. The analysis of deviance and DIC, additionally, led to the conclusion that the *MLRE* and *CARM* models provide the best fit to the data. The main remaining question is, what might be the advantage in opting for the CAR formulation when modeling housing prices in the San José Metropolitan Region? This issue is directly linked to the principal motivation of this paper, to disentangle the effects of the spatial character and hierarchy of the data on housing prices.

An exercise in model checking, following Gelman et al. (2014), has been undertaken to compare the *MLRE* and *CARM* models with the data of the dependent variable. Adopting the expected (posterior mean and variance for normal distributions, posterior shape and rate for gamma distributions) values of the parameters for these models, 1500 replications of the housing prices (y_i^{rep}) were performed. Simulations were carried out using the R package *NIMBLE* (De Valpine et al. 2017) to draw the parameters.

For each set of replications, the minimum, 5%, 25%, 50%, 75%, and 95% percentiles were estimated and compared to the corresponding percentiles of housing price data (maximum values are also included as reference). Figure 3 shows the histogram of each simulated percentile as well as the percentiles of the housing price data itself. The choice of these percentiles as test statistics $T(y)$ follows from the models' expected capabilities. The distribution of housing price data is skewed towards the lower values; the long tail of this distribution is of no interest, as it affects a small number of housing units and hardly, if in any way, the general market equilibrium. Therefore, the models should be able to reproduce most but not necessarily all of the data, and in particular there is no need for a realistic replications of the extreme values.

From figure 3 follow three conclusions about the model: first, as expected, the median values are very well represented by the posterior simulations (generated by both models) and the extreme values, much less well represented. The probability of the simulated minimum value being less than the data minimum is actually 1.00; for the maximum, the corresponding probabilities (of the simulated instances being larger than the data) are better, between 0.85 and 0.90, largely because of the greater dispersion of the simulations in the long tail of their distributions.

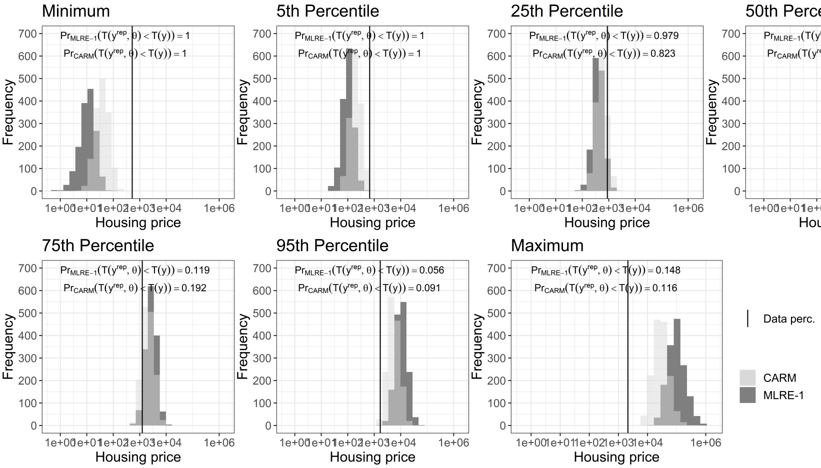

Figure 3. Model check simulation results: histograms of percentiles for y^{rep} and estimated probability of $T(y)$.

Second, the posterior distributions of $T(y^{rep})$ for the *CARM* model have peaks slightly closer to the $T(y)$ observed percentile than the $T(y^{rep})$ of the *MLRE* model –meaning the *CARM* model is somewhat better than the *MLRE* model.

Third, the simulations of both models lead to a more dispersed distribution than the data. This is evident in a very good correspondence between the median posterior simulations and the data, that increasingly deteriorates as the percentiles move away from the median towards the tail. Thus, the probability of the 25% percentile of the data being greater than the simulation and the 75% percentile of the data being less than the simulation are both between 0.10 and 0.20. The probability of the 5% percentile of the data exceeding the simulations of both models is 1.00; for the 95% percentile, the simulations are less than the data with a probability between 0.05 and 0.09. Additionally, the models seem to predict (slightly) better the housing prices larger than the median.

Discussion and Synthesis

The application of a multilevel approach to the hedonic price model of San José verified the importance of accessibility variables, in particular travel time to the CBD, in determining systematic variations of housing price. Additionally, the spatial relations structured as multiple geographic levels in the data contributed to improve the hedonic price model. When considering multilevel structure, these models outperformed the linear alternative in explaining the variation in the data. However, two important limitations have been identified: (1) the spatial autocorrelation at the higher (district) level only contributed in a very limited way, (2) the method may not have captured the part of the spatial heterogeneity that occurs at a very local scale.

The relatively simple choice of multilevel structure (two levels, the highest being districts) may have limited the models' fit to the data (indeed, the *VPM* model that included three levels showed how roughly half of the higher level variance should have been attributed to municipalities and half to districts). The introduction of a higher order municipal level is a promising future extension, especially to consider public policy (in part designed for and implemented by municipal entities). However, in terms of goodness-of-fit, neighborhood-level units to consider local externalities (a source of spatial heterogeneity not incorporated into the models reported in this paper) are likely the most important element. In the absence of official neighborhood maps, homogeneous zones that group physically similar neighborhoods for taxation purposes (defined by the Finance Ministry) may be an option. Alternatively, housing price listing locations may be thought of as a point process (Paci et al. 2017) immersed in the multilevel structure.

In sum, and returning to the research questions, most of the variation in the housing price data seems to have been explained by compositional effects of the specific location of each listing. The results suggest spatial heterogeneity operating at a more detailed scale may improve the model further, although the results themselves do verify key theoretical insights and the model structure may be profitably used for further exploration, especially of policy effects, at the regional level.

References

Agüero-Valverde J, Pujol-Mesalles R, Pérez-Molina E and Zumbado-Morales F (2020) Actualización del Plan Regulador del cantón Goicoechea. Etapa diagnóstico. Eje económico. Technical report, ProDUS-UCR, San José, Costa Rica.

Anselin L (1998) GIS research infrastructure for spatial analysis of real estate markets. *Journal of Housing Research* 9(1): 113–133. DOI:10.5555/jhor.9.1.e523670p713076p1.

Anselin L and Lozano-Gracia N (2009) Spatial Hedonic Models. In: Mills TC and Patterson K (eds.) *Palgrave Handbook of Econometrics*, volume 2. London: Palgrave Macmillan, pp. 1213–1250. DOI:10.1057/9780230244405_26.

Bergoeing JP (2017) *Geomorphology and Volcanology of Costa Rica*. Amsterdam, The Netherlands: Elsevier.

Bivand R, Sha Z, Osland L and Thorsen IS (2017) A comparison of estimation methods for multilevel models of spatially structured data. *Spat. Stat.* 21: 440–459. DOI:10.1016/j.spasta.2017.01.002.

Brueckner JK (1983) The economics of urban yard space: An "implicit-market" model for housing attributes. *J Urban Econ.* 13(2): 216–234. DOI:10.1016/0094-1190(83)90007-4.

Brueckner JK (1987) The Structure of Urban Equilibria: A Unified Treatment of the Muth-Mills Model. In: Mills ES (ed.) *Handbook of Regional and Urban Economics*, volume 2, chapter 20. Amsterdam: North Holland, pp. 821–845.

Brunauer W, Lang S and Umlauf N (2013) Modelling house prices using multilevel structured additive regression. *Stat. Model.* 13(2): 95–123. DOI:10.1177/1471082X13475385.

Cellmer R, Kobylińska K and Bełej M (2019) Application of Hierarchical Spatial Autoregressive Models to Develop Land Value Maps in Urbanized Areas. *ISPRS Int. J. Geoinf.* 8(4): 195.

DOI:10.3390/ijgi8040195.

Centro Centroamericano de Pobación (2012) Censos Nacionales de Población y Vivienda de Costa Rica 2011. ⟨https://censos.ccp.ucr.ac.cr/index.php/censos_c?censo=cr2011 ⟩. Viewed Aug. 28, 2020.

Chasco C and Le Gallo J (2012) Hierarchy and spatial autocorrelation effects in hedonic models. *Economics Bulletin* 32(1): 1474–1480.

Cheshire P and Sheppard S (1995) On the Price of Land and the Value of Amenities. *Econ.* 62(246): 247–267. DOI:10.2307/2554906.

Chi B, Dennett A, Oléron-Evans T and Morphet R (2020) Shedding new light on residential property price variation in England: A multi-scale exploration. *Env. & Plan. B* DOI: 10.1177/2399808320951212. Advance online publication.

Consejo Nacional de Planificación Urbana (2013) PlanGAM 2013. Technical report, MIVAH, San Jose, Costa Rica. https://www.mivah.go.cr/Biblioteca_PlanGAM.shtml.

De Aguiar MM, Simões R and Golgher AB (2014) Housing market analysis using a hierarchical–spatial approach: the case of Belo Horizonte, Minas Gerais, Brazil. *Regional Studies, Regional Science* 1(1): 116–137. DOI:1080/21681376.2014.934391.

De Valpine P, Turek D, Paciorek C, Anderson-Bergman C, Temple Lang D and Bodik R (2017) Programming with models: writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics* 26: 403–413. DOI:10.1080/10618600.2016.1172487.

Dong G and Harris R (2015) Spatial Autoregressive Models for Geographically Hierarchical Data Structures. *Geogr. Anal.* 47(2): 173–191. DOI:10.1111/gean.12049.

Dong G, Harris R, Jones K and Yu J (2015) Multilevel Modelling with Spatial Interaction Effects with Application to an Emerging Land Market in Beijing, China. *PLOS ONE* 10(6): e0130761. DOI:10.1371/journal.pone.0229751.

Dong G, Ma J, Harris R and Pryce G (2016) Spatial Random Slope Multilevel Modeling Using Multivariate Conditional Autoregressive Models: A Case Study of Subjective Travel Satisfaction in Beijing. *Ann. Am. Assoc. Geograph.* 106(1): 19–35. DOI:10.1080/00045608.2015.1094388.

Gelfand AE, Banerjee S, Sirmans C, Tu Y and Ong SE (2007) Multilevel modeling using spatial processes: Application to the Singapore housing market. *Comput. Stat. & Data Anal.* 51(7): 3567–3579. DOI:10.1016/j.csda.2006.11.019.

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A and Rubin DB (2014) *Bayesian Data Analysis*. 3rd edition. Boca Raton, FL: CRC Press.

Glaeser EL (2008) *Cities, Agglomeration, and Spatial Equilibrium*. Oxford, UK: Oxford University Press.

Grieson RE and White JR (1989) The existence and capitalization of neighborhood externalities: A reassessment. *J. Urban Econ.* 25(1): 68–76. DOI:10.1016/0094-1190(89)90044-2.

Henneberry DM and Barrows RL (1990) Capitalization of Exclusive Agricultural Zoning into Farmland Prices. *Land Econ.* 66(3): 249–258. DOI:10.2307/3146727.

Instituto Nacional de Estadística y Censos (2020) Encuesta Nacional de Hogares. ⟨https://www.inec.cr/encuestas/encuesta-nacional-de-hogares ⟩. Viewed Aug. 28, 2020.

Jones K (1991) Specifying and Estimating Multi-Level Models for Geographical Research. *Trans. Inst. Br. Geogr.* 16(2): 148–159. DOI:10.2307/622610.

Kain JF and Quigley JM (1970) Measuring the Value of Housing Quality. *J. Am. Stat. Assoc.* 65(330): 532–548. DOI:10.1080/01621459.1970.10481102.

Lawson AB (2013) *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology*. 2nd edition. Boca Raton, FL: CRC Press.

Liu N and Roberts D (2012) Do Incomers Pay More for Rural Housing? *Env. & Plan. A* 44(8): 1986–2005. DOI:10.1068/a44495.

Madrigal P and Pérez M (2012) Sobre la ineficacia y la ineficiencia de la legislación: el caso de la Gran Área Metropolitana. In: Pérez M (ed.) *Avatares del ordenamiento territorial en Costa Rica*, chapter I. San José, Costa Rica: FLACSO-Costa Rica, pp. 9–21.

Méndez F and Bravo O (2014) Costa Rica, Mapas de Pobreza 2011. In: Robles A, Chavarría R, González D, Solano E, Gutiérrez M, Morales N and Vargas J (eds.) *Costa Rica a la Luz del Censo 2011*. San José, Costa Rica: INEC, pp. 9–39.

Nordvik V, Osland L, Thorsen I and Thorsen IS (2019) Capitalization of neighbourhood diversity and segregation. *Env. & Plan. A* 51(8): 1775–1799.

Osland L, Thorsen IS and Thorsen I (2016) Accounting for Local Spatial Heterogeneities in Housing Market Studies. *J. Reg. Sci.* 56(5): 895–920. DOI:10.1111/jors.12281.

Paci L, Beamonte MA, Gelfand AE, Gargallo P and Salvador M (2017) Analysis of residential property sales using space–time point patterns. *Spatial Statistics* 21: 149–165. DOI: 10.1016/j.spasta.2017.06.007.

Pujol-Mesalles R (2005) Sistema de transporte en la región metropolitana de San José. In: Bussière Y (ed.) *Transporte urbano en Latinoamérica y el Caribe: Estudio de casos (San José, Puebla, Puerto España, Puerto Príncipe)*, chapter III. San José, Costa Rica: FLACSO, pp. 71–86.

Pujol-Mesalles R and Pérez-Molina E (2012) Crecimiento urbano en la región metropolitana de San José, Costa Rica. Una exploración espacial y temporal de los determinantes del cambio de uso del suelo, 1986–2010. Technical Report WP13RP1SP, Lincoln Insitute of Land Policy, Cambridge, MA.

Pujol-Mesalles R, Pérez-Molina E and Sánchez-Hernández L (2014) Informalidad en la vivienda de la Gran Área Metropolitana: El impacto de los proyectos de vivienda social, 2000-2011. In: Robles A, Chavarría R, González D, Solano E, Gutiérrez M, Morales N and Vargas J (eds.) *Costa Rica a la Luz del Censo 2011*. San José, Costa Rica: INEC, pp. 410–429.

Pujol-Mesalles R, Sánchez-Hernández L and Pérez-Molina E (2012) Patrones de crecimiento y concentración de actividades urbanas en la Gran Área Metropolitana de Costa Rica, 1993-2010. *Rev. Reflex.* : 191–209DOI:10.15517/RR.V0I0.1533. Special issue *Jornadas de Investigación Interdisciplinaria en Ciencias Sociales*.

Rosen S (1974) Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *J. Political Econ.* 82(1): 34–55. DOI:10.1086/260169.

Spiegelhalter D, Thomas A, Best N and Lunn D (2003) *WinBUGS user manual*. Medical Research Council Biostatistics Unit, University of Cambridge, Cambridge, UK. Version 1.4.

Sturtz S, Ligges U and Gelman AE (2005) R2WinBUGS: A Package for Running WinBUGS from R. *J. Stat. Softw.* 12(3). DOI:10.18637/jss.v012.i03.

Supplemental material

Supplemental material for this article is available online. Data and models have been made available online at the corresponding author's ResearchGate page.