Sankhyā B : The Indian Journal of Statistics
https://doi.org/10.1007/s13571-020-00233-y

Bayesian Spatial Modeling for Housing Data in South Africa

Bingling Wang and Sudipto Banerjee
University of California, Los Angeles (or UCLA), Los Angeles, USA
Rangan Gupta
University of Pretoria, Pretoria, South Africa

Abstract

Spatial process models are being increasingly employed for analyzing data available at geocoded locations. In this article, we build a hierarchical framework with multivariate spatial processes, where the outcomes are "mixed" in the sense that some may be continuous, some binary and others may be counts. The underlying idea is to build a joint model by hierarchically building conditional distributions with different spatial processes embedded in each conditional distribution. The idea is simple and the resulting models can be fitted to multivariate spatial data using straightforward Bayesian computing methods such as Markov chain Monte Carlo methods. Bayesian inference is carried out for parameter estimation and spatial interpolation. The proposed models are illustrated using housing data collected in the Walmer district of Port Elizabeth, South Africa. Inferential interest resides in modeling spatial dependencies of dependent outcomes and associations accounting for independent explanatory variables. Comparisons across different models confirm that the selling price of a house in our data set is relatively better modeled by incorporating spatial processes.

AMS (2000) *subject classification.* Primary 62F15; Secondary 91B72.
Keywords and phrases. Bayesian inference; Hierarchical models; Multivariate spatial models; Point-referenced data; Spatial processes

1 Introduction

With the emergence of Geographic Information Systems (GIS) and related technologies, spatial analysts and data scientists are increasingly faced with the task of analyzing data sets with variables that are referenced with respect to the geographic coordinates where they have been observed. Two common measures of referencing are *point-referenced*, where each observation is associated with the coordinates (e.g., longitude-latitude or some

planar projection thereof) of the point where it has been recorded, and *areal* where a regional aggregate or summary of the outcome (e.g., counts, rates or proportions) is recorded. Much of the traditional spatial econometrics literature have focused upon areal models (see, e.g., LeSage and Pace 2009).

Our current article is concerned with multivariate point-referenced data sets, where each location yields measurements on multiple variables. There is a substantial literature on multivariate spatial modeling that is too vast to be reviewed here (see, e.g., Chapters 8 and 9 in Banerjee et al. (2014)). The most common setting deals with situations where all the outcomes can be treated as continuous Gaussian variables (perhaps after suitable transformations) and are modeled using multivariate Gaussian processes. Methods such as linear coregionalization model (LCM) have strong limitations as they imply symmetric cross-covariances for the variables under study. In the multivariate case, the number of parameters in the model, and hence the computational requirement, increases quickly. Moreover, ensuring identifiability of model parameters is not immediate. These and other limitations have been pointed out by Marcotte (2012), who proposed certain generalizations with potential benefits for continuous variables. The current manuscript departs from the usual settings and considers multiple "mixed" outcomes in the sense that some are continuous and some are discrete (e.g., binary, counts and so on). The inference will be focused on estimating the impact of certain predictors for each outcome, quantifying the strength of spatial association for each of these variables, and predicting each variable across the geographic region of interest.

Our application pertains to so called hedonic price models in real estate economics. We model three dependent outcomes: (i) Y_1: the log-transformed selling price of a house (continuous variable), (ii) Y_2: a variable indicating whether the property has a swimming pool or not (binary variable), and (iii) Y_3 the number of bedrooms in the house (count variable). Each of these outcomes are posited to be spatially dependent, i.e., they tend to be similar at proximate locations, and to be associated among themselves. In addition, there are other independent variables (or predictors) that impact these outcomes. Our modeling framework will seek to evaluate these relationships based upon a data set of the Walmer district of Port Elizabeth, South Africa. We will provide further details on the data in Section 5. The decision to use this data set is primarily due to the detailed nature of the data (and its availability), along with the fact that hedonic modeling of house prices is primarily restricted to the developed economies; see, Du Preez et al. (2013) & Du Preez et al. (2016) for detailed literature reviews in this regard. Hence,

our application aims to provide a different perspective to hedonic price modeling for an emerging economy based on a Bayesian spatial framework with multiple outcomes.

We pursue building a multivariate spatial process model using conditional specifications in a hierarchical framework. There is a detailed account of multivariate spatial process models in Chapter 28 of Gelfand et al. (2010). An advanced presentation of linear models for multivariate spatial or temporal data can be found in part D of the book Multivariate Geostatistics by Wackernagel (2003). Cressie and Zammit-Mangion (2016) developed a conditional approach for multivariate spatial-model construction. Genton and Kleiber (2015) reviewed the main approaches for building cross-covariance functions in multivariate models. Inference will be carried out within the Bayesian paradigm (see, e.g., Gelman et al., 2013), where the unknowns in the model (e.g., parameters and spatial processes) are assigned probability laws (prior distributions) and we learn about these parameters from their posterior distributions. Bayesian inference is appealing as they deliver exact inference with direct and easy to interpret probability statements on parameter estimates, but can be difficult to compute for spatial process models due to limited information in the data for some process parameters. To ensure easier implementation, we provide a framework that can be computed using standard Bayesian computing languages such as BUGS (http://www.openbugs.net) or JAGS (http://mcmc-jags.sourceforge.net) from within the `R` (https://www.r-project.org) statistical computing environment.

2 Multivariate Spatial Process Models

Let $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ be the set of spatial locations where the data is recorded. Let $Y_i(s)$, $i = 1, 2, 3$, denote outcome i at location s and let $x(s)$ be a $p \times 1$ vector of independent variables (or predictors) recorded at location s. We will let y_i denote the $n \times 1$ vector of outcomes for each $i = 1, 2, 3$ and X to be the $n \times p$ matrix of predictors with each row being $x^\top(s)$. A general hierarchical model will be formulated as

$$
\begin{aligned}
&IG(\tau^2 \,|\, a_\tau, b_\tau) \times \prod_{i=1}^{3} \Big\{ IG(\sigma_i^2 \,|\, a_{\sigma_i}, b_{\sigma_i}) \times N\big(w_i \,|\, 0, \sigma_i^2 R_i(\phi_i)\big) \Big\} \\
&\quad \times \prod_{j=1}^{n} \Big\{ Poi\big(Y_3(s_j) \,|\, x^\top(s_j)\theta_3 + w_3(s_j)\big) \times Ber\big(Y_2(s_j) \,|\, \Phi(\gamma Y_3(s_j) + x^\top(s_j)\theta_2 + w_2(s_j))\big) \\
&\qquad \times N\big(Y_1(s_j) \,|\, x^\top(s_j)\theta_1 + \alpha Y_2(s_j) + \beta Y_3(s_j) + w_1(s_j), \tau^2\big) \Big\}, \qquad (2.1)
\end{aligned}
$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution (yielding the probit link) and each $w_i = \{w_i(s) : s \in \mathcal{S}\}$ is

an $n \times 1$ vector with zero mean and a $n \times n$ spatial correlation matrix $R_i(\phi_i)$ whose jkth element is specified according to an exponential spatial correlation function $\rho(\phi_i; \|s_j - s_k\|)$. Each $w_i(s)$ is a spatial Gaussian process, independent across $i = 1, 2, 3$, that is specified by a variance parameter σ_i^2 and a correlation function $\rho(\phi; \|s - s'\|)$. Note that the prior distributions for the regression parameters $\{\theta_1, \theta_2, \theta_3, \alpha, \beta, \gamma\}$ in the hierarchical model are all assumed to be flat.

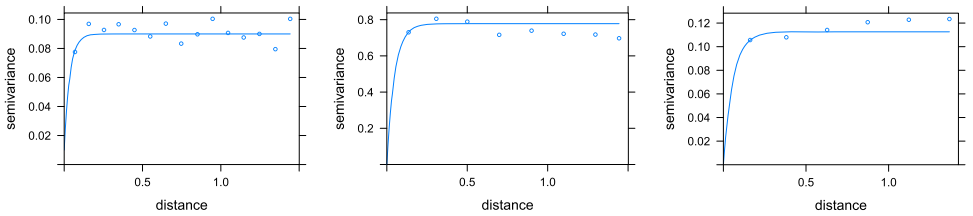
Some special cases are worth noting. For example, a model representing spatial dependence only in $Y_1(s)$ and $Y_2(s)$ but not in Y_3, will set $w_3(s) \equiv 0$ in (2.1). Similarly, for spatial dependence in $Y_1(s)$ and $Y_3(s)$, but not in Y_2, we will set $w_2(s) \equiv 0$. For spatial dependence only in $Y_1(s)$ we will set both $w_2(s)$ and $w_3(s)$ to zero, and for a complete non-spatial model we will assume $w_i(s) \equiv 0$ for each $i = 1, 2, 3$.

The customary strategy for implementing (2.1) is by Markov chain Monte Carlo (see, e.g., Robert and Casella 2004; Brooks et al. 2011). However, in full generality where all process parameters are assumed unknown, such algorithms encounter problems with convergence due to the presence of highly auto-correlated chains. This is especially true when dealing with non-Gaussian components in the data likelihood, as is the case for y_2 and y_3 in (2.1). It is, therefore, practical to assume some of the process parameters to be fixed. In particular, we assume that the spatial range (correlation decay) parameters, i.e., the ϕ_i's, are fixed using some preliminary exploratory analysis using variograms.

The *variogram* is usually defined for continuous processes $Y(s)$ as

$$\mathrm{E}[Y(s+h) - Y(s)]^2 = \mathrm{Var}[Y(s+h) - Y(s)] = 2\gamma(h) \ ,$$

where it is assumed that $E[Y(s+h) - Y(s)] = 0$ (intrinsically stationary). The function $\gamma(h)$ is called the *semi-variogram* and is assumed to be a function only of the separation between the locations. In practice, often the variogram is assumed to depend only upon the distance between locations so that $\gamma(h) = \gamma(\|h\|)$. The underlying intuition is that we expect small differences (more similarity) between measurements of the process at short distances and larger differences (less similarity) as $\|h\|$ grows larger. Empirical variograms are computed as $\hat{\gamma}(d) = \frac{1}{2|N(d)|} \sum_{(s_i, s_j) \in N(d)} [Y(s_i) - Y(s_j)]^2$, where $N(d)$ is the set pairs of points such that $\|s_i - s_j\| = d$ and $|N(d)|$ is the number of location pairs in the set. The function $\hat{\gamma}(d)$ is plotted against distance using raw data binned over neighborhoods of points separated by fixed distances (see, e.g., Cressie 1993; Banerjee et al. 2014) and are automated in a number of freely available `R` packages such as `geoR` and `gstat`.



(a) $resid(lm(Y_1 \sim Y2 + Y3 + X_{1\cdots p}))$ (b) $resid(probit(Y_2 \sim Y3 + X_{1\cdots p}))$ (c) $resid(poisson(Y_3 \sim X_{1\cdots p}))$

Figure 1: Fitted empirical variograms

To obtain estimates of the process parameters, a parametric form of the variogram is posited and its parameters are estimated using weighted least squares to minimize the distance between the parametric and the empirical variogram. In our context, we first obtain the residuals of the three outcomes by performing corresponding regression models, then the residual empirical variograms are approximated with exponential variogram model. The empirical variogram of least squares residuals for $Y_1(s)$ (log transformed selling price) is presented in Fig. 1a. Note that $Y_2(s)$ and $Y_3(s)$ are discrete variables. Y_2 is a binary variable indicating presence of a swimming pool and Y_3 is a count variable giving the number of bedrooms on each property. To obtain the semivariograms for the residuals of these two variables, we performed a probit regression and a Poisson regression respectively with the rest of the variables as covariates. The empirical variograms for the residuals corresponding to Y_2 and Y_3 are presented in Fig. 1b and c. Table 1 shows the prior estimates of the process parameters and fixed ϕ values.

3 Bayesian Inference

Statistical inference with full uncertainty quantification is obtained by drawing samples from the joint posterior distribution proportional to (2.1). We use Markov chain Monte Carlo (MCMC) algorithms constructed using the JAGS modeling language and implemented within the `R` statistical computing environment (https://cran.r-project.org/web/packages/rjags). More

Table 1: Estimates of the process parameters

Prior estimates	τ^2	σ_1^2	σ_2^2	σ_2^2	Fixed parameters	ϕ_1	ϕ_2	ϕ_3
	0.023	0.073	0.778	0.12		75.1	62.3	36.3

specifically, we fit the following five models, each emerging from (2.1), associated with different hypotheses:

- Model 1: Residual spatial dependencies do not exist for all three variables; in (2.1) we have $w_1, w_2, w_3 \equiv 0$.
- Model 2: Residual spatial dependence exists for Y_1 only; in (2.1) we have $w_2, w_3 \equiv 0$.
- Model 3: Residual spatial dependence exists for Y_1 and Y_2, but not Y_3; in (2.1) we have $w_3 \equiv 0$.
- Model 4: Residual spatial dependence exists for Y_1 and Y_3, but not Y_2; in (2.1) we have $w_2 \equiv 0$.
- Model 5: Residual spatial dependencies exist for all three variables.

For models including $w_1(s)$, i.e., Models 2-5, we collapse the parameter space by integrating out w_1 from (2.1). The resulting model is

$$\begin{aligned} &IG(\tau^2 \mid a_\tau, b_\tau) \times \prod_{i=1}^{3} IG(\sigma_i^2 \mid a_{\sigma_i}, b_{\sigma_i}) \times \prod_{i=2}^{3} N(w_i \mid 0, \sigma_i^2 R_i(\phi_i)) \\ &\quad \times \prod_{j=1}^{n} \Big\{ Poi(Y_3(s_j) \mid x^\top(s_j)\theta_3 + w_3(s_j)) \times Ber(Y_2(s_j) \mid \Phi(\gamma Y_3(s_j) + x^\top(s_j)\theta_2 + w_2(s_j))) \Big\} \\ &\quad \times N(Y_1 \mid X\theta_1 + \alpha Y_2 + \beta Y_3, \sigma_1^2 R_1(\phi_1) + \tau^2 I_n) \,, \end{aligned} \tag{3.1}$$

where Y_i denotes the $n \times 1$ vector with j-th element $Y_i(s_j)$ for $j = 1, 2, 3$, X is $n \times p$ with rows $x^\top(s_i)$ and I_n is the $n \times n$ identity matrix. The `rjags` code for the five models are supplied in the supplement Supplementary.

Once the posterior samples for $\Omega = \{\theta_1, \theta_2, \theta_3, \alpha, \beta, \gamma, \tau^2, \sigma_1^2, \sigma_2^2, \sigma_3^2\}$ are obtained, we can recover the exact posterior distribution of w_1 by sampling from

$$P(w_1|Y) = P(w_1|Y_1) \propto \int P(w_1|\Omega, Y_1) P(\Omega|Y_1) d\Omega, \tag{3.2}$$

This distribution comes out to be $N(Bb, B)$, where $B = \left(\frac{1}{\sigma_1^2} R_1^{-1}(\phi_1) + \frac{1}{\tau^2}\right)^{-1}$ and $b = \frac{1}{\tau^2}\big(Y_1 - \alpha Y_2 - \beta Y_3 - X\theta_1\big)$.

Subsequently, the posterior predictive distribution at an arbitrary location can be computed using composition sampling. Let s_0 be a new location, and w^*, Y^*, X^* be the features at the new location. We compute $P(w_i^*|Y)$ as:

$$P(w_i^*|Y) \propto \int P(w_i^*|w, \Omega, Y) P(w|\Omega, Y) P(\Omega|Y) d\Omega dw \,. \tag{3.3}$$

Composition sampling consists of three steps. First obtain posterior samples $\Omega \sim P(\Omega|Y)$, then recover w_i as described in (3.2). Finally, since we assume independent Gaussian processes for all w's, the distribution $P(w_i^*|w, \Omega, Y)$ can be simplified to $P(w_i^*|w_i, \Omega)$, and is derived from a multivariate normal distribution:

$$\begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ w_1^* \\ w_2^* \\ w_3^* \end{pmatrix} \sim MVN \left[\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \bigoplus_{i=1}^3 \sigma_i^2 R_i(\phi_i) & \bigoplus_{i=1}^3 \sigma_i^2 r_i(s_0) \\ \left[\bigoplus_{i=1}^3 \sigma_i^2 r_i(s_0)\right]^T & \bigoplus_{i=1}^3 \sigma_i^2 \end{pmatrix} \right]$$

where $\bigoplus_{i=1}^3 f_i = \begin{pmatrix} f_1 & 0 & 0 \\ 0 & f_2 & 0 \\ 0 & 0 & f_3 \end{pmatrix}$, and $r_i(s_0)$ is a vector with ith element as $\rho(\phi_i; \|s_i - s_0\|)$.

The conditional distribution has the form:

$$P(w_i^*|w, \Omega, Y) = P(w_i^*|w_i, \Omega) \sim N\left(\Sigma_{w_i, w_i^*}^T \Sigma_{w_i}^{-1} w_i, \Sigma_{w_i^*} - \Sigma_{w_i, w_i^*}^T \Sigma_w^{-1} \Sigma_{w_i^*, w_i}\right) \tag{3.4}$$

where $\Sigma_{w_i} = \sigma_i^2 R_i(\phi_i)$, $\Sigma_{w_i^*} = \sigma_i^2, \Sigma_{w_i, w_i^*} = \Sigma_{w_i^*, w_i} = \sigma_i^2 r_i(s_0)$. After obtaining w_i^*, prediction of the responses can be easily carried out by sampling the conditional expectation $E[Y^*|Y, \Omega, w^*]$.

For comparing these models, we use the Widely Applicable Information Criterion (WAIC) proposed by Watanabe (2010); also see Gelman et al. (2014). WAIC is a Bayesian criterion that closely approximates cross-validation in a computationally convenient way. It can be calculated using only training samples which is useful for model selection. WAIC is computed as $waic = -2\hat{elpd}_{waic}$ (Vehtari and Gelman, 2014), where $\hat{elpd}_{waic}$ is the estimated expected log point-wise predictive density, defined as $\hat{elpd}_{waic} = \hat{lpd} - \hat{p}_{waic}$, where $\hat{lpd}$ is the computed log pointwise predictive density, $\hat{lpd} = \sum_{i=1}^{n} \log\left(\frac{1}{S}\sum_{s=1}^{S} p(y_i|\theta^s)\right)$ and $\hat{p}_{waic}$ is the simulation-estimated effective number of parameters, computed by summing posterior variance of log predictive density over all the data points, $\hat{p}_{waic} = \sum_{i=1}^{n} V_{s=1}^{S}\big(logp(y_i|\theta^s)\big)$, where V is the sample variance $V_{s=1}^{S}(a_s) = \frac{1}{S-1}\sum_{s=1}^{S}(a_s - \bar{a})^2$.

4 Simulation

We conducted some simulation experiments to evaluate our multivariate spatial process models. We simulated several data sets each with six variables $(Y_1, Y_2, Y_3, X_1, X_2, X_3)$ under different spatial effects settings at 200 locations.

Setting 1 (spatial effects: w_1, w_2, w_3):

$$Y_1 \sim N(\mu_1, 0.5), \mu_1 = \alpha Y_2 + \beta Y_3 + b_{10} + b_{11}X_1 + b_{12}X_2 + b_{13}X_3 + w_1$$

$$Y_2 \sim Ber(\Phi(\gamma Y_3 + b_{21}X_1 + b_{22}X_2 + b_{23}X_3 + w_2))$$

$$Y_3 : Pois(b_{31}X_1 + b_{32}X_2 + b_{33}X_3 + w_3)$$

Setting 2 (spatial effects: w_1, w_3):

$$Y_1 \sim N(\mu_1, 0.5), \mu_1 = \alpha Y_2 + \beta Y_3 + b_{10} + b_{11}X_1 + b_{12}X_2 + b_{13}X_3 + w_1$$

$$Y_2 \sim Ber(\Phi(\gamma Y_3 + b_{21}X_1 + b_{22}X_2 + b_{23}X_3))$$

$$Y_3 : Pois(b_{31}X_1 + b_{32}X_2 + b_{33}X_3 + w_3)$$

Setting 3 (spatial effects: w_1, w_2):

$$Y_1 \sim N(\mu_1, 0.5), \mu_1 = \alpha Y_2 + \beta Y_3 + b_{10} + b_{11}X_1 + b_{12}X_2 + b_{13}X_3 + w_1$$

$$Y_2 \sim Ber(\Phi(\gamma Y_3 + b_{21}X_1 + b_{22}X_2 + b_{23}X_3 + w_2))$$

$$Y_3 : Pois(b_{31}X_1 + b_{32}X_2 + b_{33}X_3)$$

We compare the five models in each of the above settings. Prior estimates are obtained from empirical variograms. Table 2 shows the comparisons in

Table 2: WAIC and computation time comparisons

Model	M1(no w)	M2(w1)	M3(w1w2)	M4(w1w3)	M5(w1w2w3)
Setting 1	840.0	479.8	480.9	471.0	464.2
Setting 2	836.3	474.9	475.2	454.6	463.8
Setting 3	673.5	314.2	308.9	312.2	315.8
Machine specs	MacBook Pro, 2 GHz Dual-Core Intel Core i5, 8 GB 1867 MHz LPDDR3				
Computation time (100 iterations)	2.179 s	7.993 s	12.597 s	12.272 s	16.499 s

the simulation study. We see substantial improvement by incorporating spatial effects in the model. In the first simulation setting, the best model is M5 with all three spatial effects included. And in the second setting where the true spatial effects are w_1 and w_2, the best model is M3 which only includes two spatial effects w_1, w_2. Similarly, in the third setting where the true spatial effects are w_1 and w_3, the best model is M4 which includes the spatial effects w_1, w_3. Computation times presented in Table 2 are based upon running 100 iterations for each model. Estimating the additional spatial effects will increase the computational complexity and is more time consuming. Hence, we exclude them.

Figure 2: Mapping of houses in Walmer district

Table 3: Description of variables in Housing data

Variables	Type	Description
Price	Continuous	Recorded sales price of the house in Rand
Year.of.Sale	Discrete	The year that the house was sold
Size.Erf	Continuous	The size of the erf (land registered in a deeds registry) in square meters
Stories	Count	Number of stories in the house
Bath	Count	Number of bathrooms (including partial bathrooms) in the house
Bed	Count	Number of bedrooms in the house
Swim	Binary	The presence of a swimming pool
Staff	Binary	The presence of staff quarters
Bachelor	Binary	The presence of a bachelor's flat/apartment
Aircon	Binary	The presence of air conditioning,
Garage	Binary	The presence of a garage
Irrigation	Binary	The presence of a irrigation system
Dining	Binary	The presence of a separate dining room
Living.Rooms	Count	Number of living rooms
Tennis	Binary	The presence of tennis court
Wall	Binary	The presence of a boundary wall enclosing the property
Elec.Gate	Binary	The presence of an electric gate for gaining access to the property
Security	Binary	The presence of a security system
ElecFence	Binary	The presence of an electrified barrier installed on top of the boundary wall
Dist.Social.Housing	Continuous	Distance to a social housing development (Walmer Township) in metres

Table 4: SLR of lasso selected variables

Coefficients	Estimate	Std. Error	t value	$Pr(>\|t\|)$	Significance
`Size.Erf`	0.026	0.005	5.651	7.1e-08	***
`Bath`	0.045	0.028	1.606	0.110	
`Bed`	0.058	0.035	1.66	0.099	.
`Swim`	0.308	0.067	4.581	9.3e-06	***
`Staff`	0.104	0.054	1.942	0.054	.
`Irrigation`	0.198	0.049	4.047	8e-05	***
`Dist.Social.Housing`	0.011	0.004	2.520	0.013	*

5 Data Analysis

Our housing market data comes form the district of Walmer, Port Elizabeth, South Africa. Previous studies using this data set include (Du Preez et al., 2013), and (Du Preez and Sale, 2015), where (linear and nonparametric) hedonic price models were applied to the district of Walmer.[1] This district has 2625 properties in total. A sample of 168 houses that have been traded at least once during 1995 and 2009 is used here for the analysis. Our primary interest is the recorded sales prices of the houses. The ABSA house price index was used to inflate the sales price of each individual house to constant 2009 rands. Eighteen other structural and neighborhood characteristics were collected on each individual house: `Year.of.Sale, Size.Erf, Stories, Bath, Bed, Swim, Staff, Bachelor, Aircon, Garage, Irrigation, Dining, Living.Rooms, Tennis, Wall, Elec.Gate, Security, ElecFence, Dist.Social.Housing`. See more details of the variables in Table 3. The average sales price between Year 1995 and Year 2009 is R 1,618,497, range from R 193,600 to R 4,926,800. The house in the dataset has an average of 1773 square metres in size, 3.7 bedrooms, 2.7 bathrooms, 1.9 living rooms, 1.2 stories, and is located 1797 metres from the Walmer Township. The majority of houses have a garage, a swimming pool, a separate dining room, and are built with boundary wall, electric gate, and security system , although less than half of the houses have air-conditioning, tennis court, or electric fencing. Provided the physical addresses, we obtain the coordinates of each individual house from the `GoogleMaps` package in `R` and plot the map of houses in Fig. 2.

Our primary interest is the selling price so the main outcome Y_1 is chosen to be the log-transformed 2009 home prices. A log-transformation is applied as the original data indicated right skewness. We choose the other two

[1]We would like to thank Dr. Mari Du Preez and Dr. Michael Sale for kindly sharing the data set with us.

outcomes Y_2 and Y_3 using some preliminary analysis. First, we perform a lasso regression of Y_1 on the remaining variables, then least squares estimates are obtained for the subset of variables from lasso; see Table 4. We selected two types of variables that are more significant in their own category: a binary variable `Swim` as Y_2 and a count variable `Bed` as Y_3.

Posterior inference for regression coefficients in all five models discussed in Section 2 are presented in Tables 5–7. The results of the five models agree closely. The variables that have positive effects on home prices are `Swim`,

Table 5: Coefficients for modeling log(price) (Y_1)

Posterior Mean (95% CI)	M1(no w)	M2(w1)	M3(w1w2)	M4(w1w3)	M5(w1w2w3)
Swim(Y_2)	0.29 (0.15, 0.42)	0.29 (0.15, 0.42)	0.28 (0.14, 0.41)	0.28 (0.15, 0.42)	0.28 (0.14, 0.42)
Bed (Y_3)	0.03 (−0.03, 0.10)	0.04 (−0.04, 0.1)	0.04 (−0.03, 0.11)	0.04 (−0.03, 0.14)	0.03 (−0.04, 0.12)
Year.of.Sale	0.7 (0.65, 0.75)	0.67 (0.54, 0.79)	0.62 (0.55, 0.71)	0.67 (0.59, 0.73)	0.49 (0.36, 0.63)
Size.Erf	0.03 (0.02, 0.04)	0.03 (0.02, 0.03)	0.03 (0.02, 0.04)	0.03 (0.02, 0.04)	0.03 (0.02, 0.03)
Stories	0.11 (−0.04, 0.26)	0.11 (−0.03, 0.26)	0.11 (−0.03, 0.25)	0.11 (−0.02, 0.25)	0.10 (−0.03, 0.23)
Bath	0.05 (−0.01, 0.10)	0.05 (−0.01, 0.10)	0.05 (−0.01, 0.10)	0.05 (−0.02, 0.11)	0.05 (−0.01, 0.11)
Staff	0.09 (−0.03, 0.20)	0.09 (−0.03, 0.20)	0.09 (−0.03, 0.20)	0.08 (−0.03, 0.19)	0.09 (−0.02, 0.20)
Bachelor	−0.11 (−0.21, 0)	−0.11 (−0.21, 0)	−0.11 (−0.21, 0)	−0.11 (−0.21, 0)	−0.11 (−0.21, 0)
Aircon	0.03 (−0.08, 0.15)	0.03 (−0.09, 0.15)	0.03 (−0.09, 0.15)	0.03 (−0.09, 0.15)	0.03 (−0.09, 0.15)
Garage	−0.05 (−0.19, 0.09)	−0.04 (−0.18, 0.09)	−0.04 (−0.18, 0.10)	−0.05 (−0.18, 0.10)	−0.04 (−0.18, 0.10)
Irrigation	0.19 (0.08, 0.29)	0.19 (0.09, 0.30)	0.19 (0.09, 0.30)	0.19 (0.09, 0.30)	0.19 (0.08, 0.30)
Dining	−0.06 (−0.18, 0.05)	−0.06 (−0.17, 0.05)	−0.06 (−0.18, 0.05)	−0.06 (−0.17, 0.05)	−0.06 (−0.17, 0.05)
Living.Rooms	−0.03 (−0.11, 0.05)	−0.03 (−0.11, 0.05)	−0.03 (−0.12, 0.04)	−0.03 (−0.10, 0.05)	−0.03 (−0.11, 0.05)
Tennis	0.04 (−0.13, 0.21)	0.03 (−0.13, 0.19)	0.03 (−0.14, 0.19)	0.03 (−0.14, 0.19)	0.03 (−0.14, 0.20)
Wall	0.12 (−0.25, 0.48)	0.09 (−0.31, 0.47)	0.10 (−0.22, 0.43)	0.10 (−0.29, 0.48)	0.10 (−0.25, 0.41)
Elec.Gate	0.13 (−0.07, 0.31)	0.14 (−0.05, 0.32)	0.14 (−0.05, 0.34)	0.15 (−0.05, 0.35)	0.14 (−0.07, 0.34)
Security	0.28 (−0.11, 0.65)	0.32 (−0.01, 0.62)	0.32 (−0.03, 0.61)	0.32 (0.02, 0.68)	0.35 (0.01, 0.68)
Elec.Fence	0.10 (−0.01, 0.22)	0.10 (−0.02, 0.22)	0.10 (−0.02, 0.22)	0.10 (−0.03, 0.22)	0.10 (−0.02, 0.22)
Dist.Social.Housing	0.01 (0, 0.02)	0.01 (0, 0.02)	0.01 (0, 0.02)	0.01 (0, 0.02)	0.01 (0, 0.02)

Table 6: Coefficients summary for modeling Swim Y_2

Posterior Estimates Mean (95% CI)	M1(no w)	M2(w1)	M3(w1w2)	M4(w1w3)	M5(w1w2w3)
Bed (Y_3)	0.07 (-0.39, 0.50)	0.03 (−0.38, 0.47)	0.07 (−0.34, 0.46)	0.03 (−0.36, 0.43)	0.02 (−0.43, 0.4)
Year.of.sale	0.41 (0.07, 0.76)	0.25 (−0.06, 0.62)	−0.22 (−0.44, −0.03)	−0.01 (−0.33, 0.23)	−0.5 (−0.77, −0.18)
Size.Erf	0.03 (−0.03, 0.08)	0.03 (−0.02, 0.08)	0.03 (−0.03, 0.08)	0.03 (−0.02, 0.08)	0.03 (−0.02, 0.08)
Stories	0.01 (−0.81, 0.92)	−0.02 (−0.82, 0.86)	0.09 (−0.76, 0.97)	0.01 (−0.83, 0.88)	0.02 (−0.78, 0.84)
Bath	0.1 (−0.24, 0.47)	0.12 (−0.24, 0.49)	0.10 (−0.24, 0.42)	0.12 (−0.2, 0.44)	0.12 (−0.19, 0.47)
Staff	0.46 (−0.12, 1.08)	0.46 (−0.13, 1.07)	0.48 (−0.12, 1.08)	0.44 (−0.13, 1.02)	0.46 (−0.14, 1.08)
Bachelor	−0.05 (−0.58, 0.46)	−0.03 (−0.57, 0.48)	−0.04 (−0.59, 0.52)	−0.03 (−0.53, 0.47)	−0.03 (−0.57, 0.52)
Aircon	0.00 (−0.63, 0.66)	−0.01 (−0.65, 0.64)	0.01 (−0.64, 0.69)	0.02 (−0.57, 0.62)	0.06 (−0.57, 0.72)
Garage	0.67 (0.07, 1.28)	0.67 (0.07, 1.23)	0.73 (0.08, 1.39)	0.65 (0.05, 1.2)	0.68 (0.07, 1.3)
Irrigation	0.11 (−0.44, 0.68)	0.08 (−0.47, 0.65)	0.11 (−0.47, 0.68)	0.11 (−0.41, 0.63)	0.11 (−0.48, 0.66)
Dining	0.34 (−0.20, 0.86)	0.35 (−0.19, 0.91)	0.37 (−0.18, 0.93)	0.34 (−0.18, 0.86)	0.31 (−0.26, 0.87)
Living.Rooms	−0.01 (−0.41, 0.39)	0.00 (−0.36, 0.38)	−0.01 (−0.4, 0.39)	0.00 (−0.38, 0.39)	0.00 (−0.4, 0.39)
Tennis	0.60 (−0.37, 1.71)	0.60 (−0.42, 1.75)	0.62 (−0.42, 1.83)	0.52 (−0.48, 1.61)	0.58 (−0.44, 1.76)
Wall	0.29 (−1.46, 1.85)	0.24 (−1.39, 1.66)	0.4 (−1.2, 2.06)	0.02 (−2.11, 1.67)	0.16 (−1.93, 1.74)
Elec.Gate	0.19 (−0.67, 0.98)	0.14 (−0.83, 1)	0.15 (−0.8, 1.05)	0.11 (−0.75, 0.94)	0.11 (−0.91, 1.06)
Security	1.48 (0.22, 2.89)	1.62 (−0.33, 3.77)	1.66 (0.22, 3.74)	1.21 (−0.18, 2.45)	1.41 (−0.01, 3)
Elec.Fence	0.96 (0.28, 1.74)	0.99 (0.28, 1.79)	0.98 (0.23, 1.79)	0.96 (0.25, 1.75)	0.99 (0.26, 1.87)
Dist.Social.Housing	−0.02 (−0.07, 0.02)	−0.02 (−0.07, 0.01)	−0.03 (−0.07, 0.02)	−0.02 (−0.07, 0.02)	−0.03 (−0.07, 0.02)

`Bed, Year.of.Sale, Size.Erf, Stories, Bath, Staff, Irrigation, Tennis, Wall, Elec.Gate, Security, Elec.Fence and Dist.Social.Housing`, and the ones that have negative effects are `Bachelor, Garage, Dining, Living.Rooms`. We see that Y_2`(Swim), Year.of.Sale, Size.Erf, Bachelor, Irrigation` and `Dist.Social.Housing` are significant variables for modeling Y_1 according to the posterior 95% confidence intervals. From the posterior estimates of the coefficients, when we have $Y_2 = 1$, Y_1 is increased by approximately 0.28 which indicates that the selling price of a house with a swimming pool is about $e^{0.28} = 1.3$ times more than a house without a swimming pool. Housing price also significantly increase with

Table 7: Coefficients summary for modeling bed Y_3

Posterior Estimates Mean (95% CI)	M1(no w)	M2(w1)	M3(w1w2)	M4(w1w3)	M5(w1w2w3)
Year.of.sale	0.01	−0.02	−0.11	0.09	−0.05
	(−0.1, 0.08)	(−0.16, 0.13)	(−0.26, −0.03)	(0.01, 0.14)	(−0.16, 0.04)
Size.Erf	0.01	0.01	0.01	0.01	0.01
	(−0.01, 0.02)	(−0.01, 0.02)	(−0.01, 0.02)	(0, 0.02)	(−0.01, 0.02)
Stories	0.05	0.04	0.05	0.05	0.07
	(−0.17, 0.25)	(−0.18, 0.25)	(−0.14, 0.26)	(−0.18, 0.26)	(−0.15, 0.29)
Bath	0.08	0.07	0.07	0.08	0.08
	(0.00, 0.15)	(0.00, 0.15)	(0.00, 0.15)	(0.01, 0.15)	(0.01, 0.16)
Staff	−0.02	−0.01	−0.01	−0.03	−0.02
	(−0.2, 0.15)	(−0.18, 0.17)	(−0.19, 0.17)	(−0.21, 0.17)	(−0.2, 0.16)
Bachelor	0.00	0.00	0.00	0.00	−0.01
	(−0.16, 0.17)	(−0.15, 0.17)	(−0.16, 0.17)	(−0.17, 0.17)	(−0.17, 0.17)
Aircon	0.06	0.06	0.06	0.06	0.05
	(−0.13, 0.24)	(−0.13, 0.24)	(−0.13, 0.23)	(−0.12, 0.24)	(−0.14, 0.24)
Garage	−0.07	−0.06	−0.07	−0.07	−0.06
	(−0.29, 0.15)	(−0.26, 0.15)	(−0.28, 0.14)	(−0.28, 0.13)	(−0.27, 0.15)
Irrigation	−0.07	−0.07	−0.07	−0.09	−0.08
	(−0.23, 0.1)	(−0.24, 0.10)	(−0.24, 0.10)	(−0.26, 0.09)	(−0.25, 0.09)
Dining	−0.02	−0.02	−0.02	−0.01	−0.03
	(−0.19, 0.16)	(−0.20, 0.16)	(−0.19, 0.17)	(−0.20, 0.18)	(−0.21, 0.16)
Living.Rooms	0.01	0.02	0.01	0.01	0.01
	(−0.12, 0.13)	(−0.11, 0.15)	(−0.11, 0.13)	(−0.12, 0.12)	(−0.11, 0.14)
Tennis	0.11	0.11	0.12	0.12	0.11
	(−0.14, 0.35)	(−0.14, 0.34)	(−0.12, 0.36)	(−0.13, 0.36)	(−0.14, 0.36)
Wall	0.25	0.19	0.24	0.2	0.24
	(−0.26, 0.90)	(−0.35, 0.86)	(−0.32, 0.77)	(−0.43, 0.72)	(−0.33, 1.07)
Elec.Gate	−0.05	−0.04	−0.04	−0.07	−0.01
	(−0.36, 0.27)	(−0.35, 0.35)	(−0.35, 0.29)	(−0.41, 0.26)	(−0.31, 0.32)
Security	0.05	0.09	0.06	0.11	0.12
	(−0.44, 0.54)	(−0.46, 0.72)	(−0.42, 0.61)	(−0.51, 0.7)	(−0.47, 0.65)
Elec.Fence	0.06	0.06	0.06	0.05	0.05
	(−0.12, 0.25)	(−0.13, 0.25)	(−0.12, 0.24)	(−0.14, 0.23)	(−0.13, 0.24)
Dist.Social.Housing	0.00	0.00	0.00	0.00	0.00
	(−0.01, 0.02)	(−0.01, 0.02)	(−0.01, 0.02)	(−0.01, 0.02)	(−0.01, 0.02)

time. Bigger sized houses tend to be more expensive. However, since most of the houses in this dataset are very close in size, the influence is relatively small. A bachelor-style flat will impact the selling price by 10%. An irrigation system seems to be a significantly important factor in evaluating the sales price of a house.

Coefficients in the Probit model with Y_2 `Swim` as response variable are shown in Table 6. The posterior estimates suggest two significant variables that are positively associated with the swimming pool. We conclude that houses with garage and electric fence are more likely to have a swimming pool. Coefficients in the Poisson model with Y_3 `Bed` as response variable are shown in Table 7. `Bath` is a significant variable, which is reasonable as

Table 8: Posterior parameters summary

Posterior Estimates Mean (sd)	M1(no w)	M2(w1)	M3(w1w2)	M4(w1w3)	M5(w1w2w3)
τ^2	0.106 (0.012)	0.012(0.014)	0.013(0.017)	0.014(0.017)	0.012(0.015)
σ_1^2	–	0.096(0.018)	0.094(0.021)	0.093(0.020)	0.095(0.019)
σ_2^2	–	–	0.089(0.038)	–	0.226(0.039)
σ_3^2	–	–	–	0.016(0.005)	0.016(0.003)

the number of bathrooms usually increases with the number of bedrooms. However, `Bed` is not strongly related to any of the remaining variables.

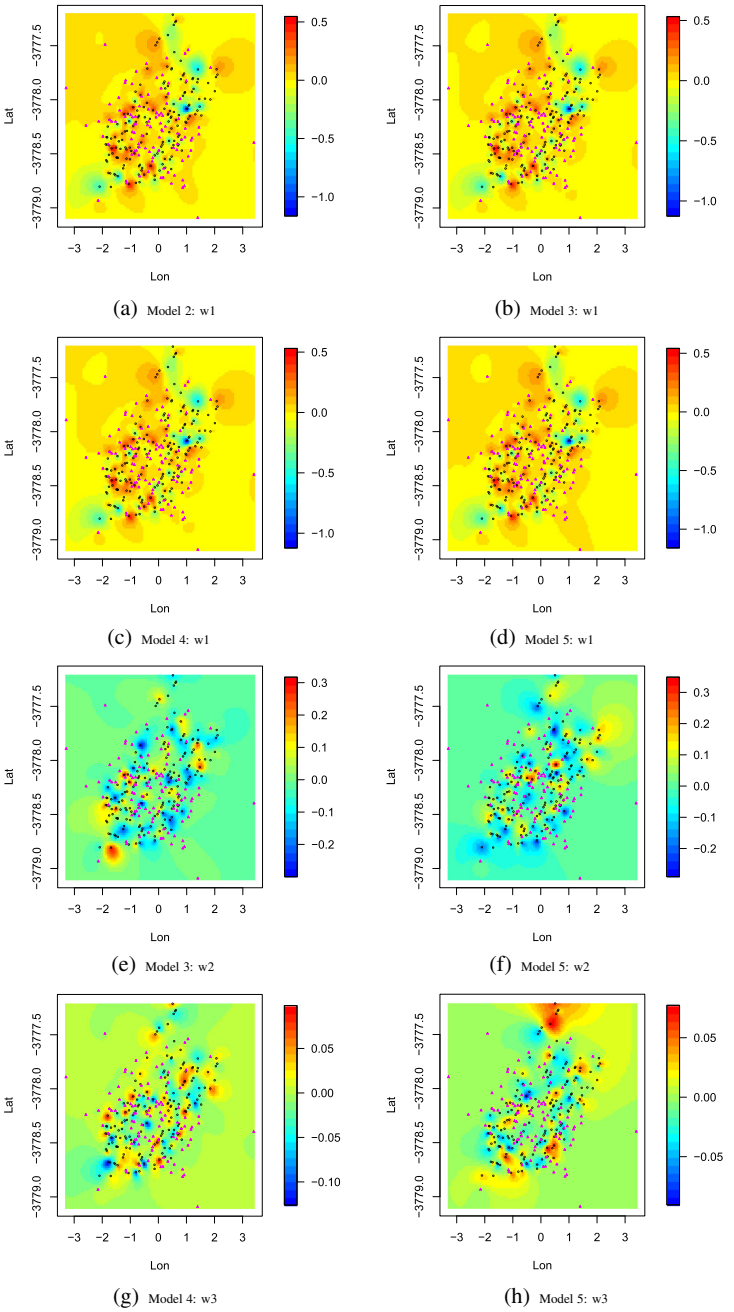
Posterior estimates of the process parameters and spatial random effects are recorded in Table 8. Table 9 shows the significance of the three posterior spatial random effects w_1, w_2, w_3 among all 168 locations.

Figure 3 depict the interpolated surfaces of spatial random effects w_1, w_2, w_3 for four spatial models (Model 2 $\sim$ Model 5). New locations are randomly simulated on the grid of the region. Predictions of spatial random effects are calculated at each new location using parameters estimated from the model. We use multilevel B-spline approximation (MBA) algorithm proposed by Lee et al. (1997) which is available in `R` package `MBA` to approximate the surface from a bivariate scatter of data points. The black dots are the original locations and the red triangles are the simulated new locations. We can see clearly the patterns of spatial dependence in Fig. 3a–h that there is obvious spatial pattern according to the image plots and the four models display similar patterns. The spatial effects of housing price are positive at most locations, however, the spatial effects of swimming pool and number of bedrooms are negative at most locations.

We run each jags model for 20000 iterations with one chain. We present the diagnostic plots of parameters α, β, γ and spatial effects at random locations in Fig. 4. The trace plots show good convergence for α, β, γ. Among the three spatial effect terms, w_1 converges well, while w_2 and w_3 converge much slower.

Table 9: Posterior spatial random effects significance summary

95% Confidence Interval	M1(no w)	M2(w1)	M3(w1w2)	M4(w1w3)	M5(w1w2w3)
w1 Proportion of not including 0	–	99%	98%	98%	96%
Significantly Positive	–	79	79	79	78
Significantly Negative	–	88	86	86	83
w2 Proportion of not including 0- -		–	96%	–	97%
Significantly Positive	–	–	74	–	85
Significantly Negative	–	–	87	–	78
w3 Proportion of not including 0	–	–	–	96%	96%
Significantly Positive	–	–	–	82	76
Significantly Negative	–	–	–	80	85



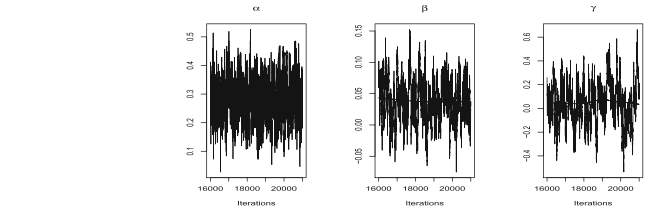
(a) Model 2: w1 (b) Model 3: w1

(c) Model 4: w1 (d) Model 5: w1

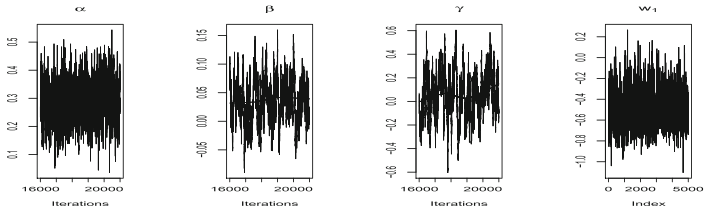
(e) Model 3: w2 (f) Model 5: w2

(g) Model 4: w3 (h) Model 5: w3

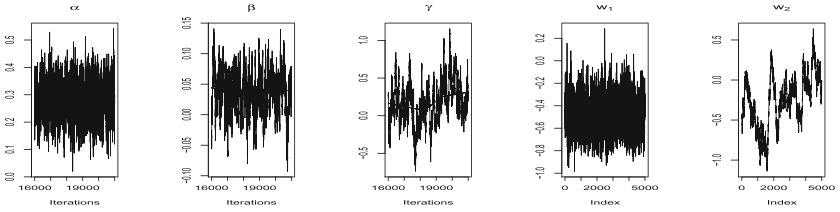
Figure 3: Interpolated surface of the mean of the spatial random effects posterior distribution and spatial effect predictions. (○: Original locations △: New locations)



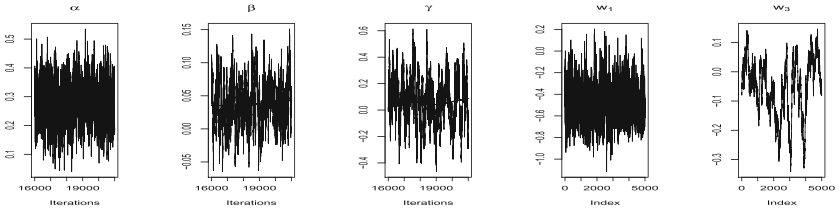
(a) Model 1



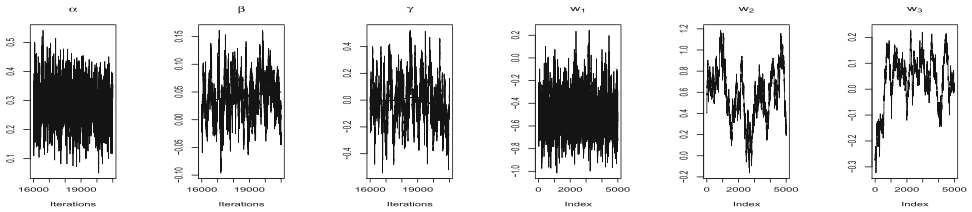
(b) Model 2



(c) Model 3



(d) Model 4



(e) Model 5

Figure 4: Trace plots of α, β, γ and spatial effects w_1, w_2, w_3 at random locations

Table 10: WAIC Comparison

Model	M1(no w)	M2(w1)	M3(w1w2)	M4(w1w3)	M5(w1w2w3)
WAIC	880.6	563.3	578.5	594.9	567.0
LPML	886.4	678.2	665.8	704.4	660.5

WAIC comparisons shown in Table 10 suggests that the model performance is greatly enhanced by incorporating random spatial effects of the selling price into the model. Adding spatial effects to `Swim` and `Bed` do not seem to produce substantial improvements. To show that the inference is consistent with other model selection criteria, we also computed PSIS-LOO CV, efficient approximate leave-one-out (LOO) cross-validation for Bayesian models using Pareto smoothed importance sampling (PSIS) proposed by Vehtari et al. (2017). WAIC and the LOO information criterion (LOOIC) give consistent comparisons among the five models as seen in the Table 10.

6 Conclusion

This study works with multivariate point-referenced data and considers multiple mixed outcomes, i.e., some are continuous and some are discrete, in a Bayesian hierarchical framework. When applied to a data set of 168 houses in the district of Walmer, Port Elizabeth, South Africa, our results revealed that among all the house characteristics, the presence of a swimming pool, the year the property was sold, the size of the property (in square meters), whether it is a bachelor's apartment, the presence of an irrigation system, and the distance to a social housing development (in meters) are significantly associated with the selling price. Furthermore, the presence of a swimming pool seems to be significantly associated with the property having a garage and an electric fence. Importantly, comparison between the different models suggest that incorporating a spatial process for selling price clearly results in better model fitting of the data, while the spatial processes for swimming pool (binary) and number of bedrooms (count) did not result in further significant improvements.

References

banerjee, s., carlin, b.p. and gelfand, a.e. (2014). Hierarchical Modeling and Analysis for Spatial Data. 2nd Edn. Chapman & Hall/CRC.

brooks, s., gelman, a., jones, g. and meng, x.-l. (2011). Handbook of Markov Chain Monte Carlo. Chapman & Hall/CRC.

cressie, n. and zammit-mangion, a. (2016). Multivariate spatial covariance models: a conditional approach. *Biometrika* **103**, 4, 915–935.

cressie, n.a.c. (1993). *Statistics for Spatial Data*. Wiley, New York.

du preez, m., balcilar, m., razzak, a., koch, s.f. and gupta, r. (2016). House values and proximity to a landfill in South Africa. *Journal of Real Estate Literature* **24**, 133–149.

du preez, m., lee, d. and sale, m. (2013). Nonparametric estimation of a hedonic price model: A South African case study. *Journal for Studies in Economics and Econometrics* **37**, 2, 41–62.

du preez, m. and sale, m. (2015). Municipal assessments versus actual sales prices in hedonic price studies. *Journal of Economic and Financial Sciences* **8**, 35–46.

gelfand, a. e., diggle, p. j., fuentes, m. and guttorp, p. (2010). Handbook of Spatial Statistics. Chapman & Hall/CRC.

gelman, a., carlin, j. b., stern, h. s., dunson, d. b., vehtari, a. and rubin, d.b. (2013). Bayesian Data Analysis. 3rd Edn. Chapman & Hall/CRC.

gelman, a., hwang, j. and vehtari, a. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing* **24**, 6, 997–1016.

genton, m.g. and kleiber, w. (2015). Cross-Covariance Functions for multivariate geostatistics. *Stat. Sci.* **30**, 2, 147–163.

lee, s., wolberg, g. and shin, s.y. (1997). Scattered data interpolation with multilevel B-Splines. *IEEE Transactions on Visualization and Computer Graphics* **3**, 3.

lesage, j. and pace, r.k. (2009). Introduction to Spatial Econometrics. Chapman & Hall/CRC.

marcotte, d. (2012). *Revisiting the Linear Model of Coregionalization*. Springer, Dordrecht, p. 67–78.

robert, c. and casella, g. (2004). Monte Carlo Statistical Methods. 2nd ed. Springer.

vehtari, a. and gelman, a. (2014). WAIC and cross-validation in Stan. Statistics and Computing.

vehtari, a., gelman, a. and gabry, j. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* **27**, 5, 1413–1432.

wackernagel, h. (2003). *Multivariate Geostatistics: An Introduction with Applications*. Springer, Berlin.

watanabe, s. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research* **11**, 3571–3594.

Publisher's Note. Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Bingling Wang
and Sudipto Banerjee
Department of Biostatistics,
University of California, Los Angeles
(or UCLA), Los Angeles, CA, USA
E-mail: binglingwang@ucla.edu
sudipto@ucla.edu

Rangan Gupta
Department of Economics, University
of Pretoria, Pretoria, South Africa
E-mail: rangan.gupta@up.ac.za

Paper received: 18 June 2018.