THE UNIVERSITY OF TEXAS AT SAN ANTONIO, COLLEGE OF BUSINESS

Working Paper SERIES

December 10, 2007

0030MSS-600-2007

BAYESIAN SPATIAL MODELING OF HOUSING PRICES SUBJECT

Mark D. Ecker
Department of Mathematics
University of Northern Iowa
Cedar Falls, IA 50614, U.S.A.
ecker@math.uni.edu

Victor De Oliveira
Department of Management Science and Statistics
The University of Texas at San Antonio
San Antonio, TX 78249, U.S.A
victor.deoliveira@utsa.edu

Copyright ©2006 by the UTSA College of Business. All rights reserved. This document can be downloaded without charge for educational purposes from the UTSA College of Business Working Paper Series (business.utsa.edu/wp) without explicit permission, provided that full credit, including © notice, is given to

UTSA COLLEGE OF BUSINESS KNOWLEDGE FOR A NEW WORLD

ONE UTSA CIRCLE
SAN ANTONIO.

BAYESIAN SPATIAL MODELING OF HOUSING PRICES SUBJECT TO A LOCALIZED EXTERNALITY

Mark D. Ecker
Department of Mathematics
University of Northern Iowa
Cedar Falls, IA 50614, U.S.A.
ecker@math.uni.edu

Victor De Oliveira[1]
Department of Management Science and Statistics
The University of Texas at San Antonio
San Antonio, TX 78249, U.S.A
victor.deoliveira@utsa.edu

December 10, 2007

Abstract

This work proposes a non-stationary random field model to describe the spatial variability of housing prices that are affected by a localized externality. The model allows for the effect of the localized externality on house prices to be represented in the mean function and/or the covariance function of the random field. The correlation function of the proposed model is a mixture of an isotropic correlation function and a correlation function that depends on the distances between home sales and the localized externality. The model is fit using a Bayesian approach via a Markov chain Monte Carlo algorithm. A dataset of 437 single family home sales during 2001 in the city of Cedar Falls, Iowa, is used to illustrate the model.

Key words: Geostatistics; Hedonic regression; Monte Carlo; Random field; Real estate data.

JEL Code: C11, C16

[1] Corresponding author. Email: victor.deoliveira@utsa.edu; Phone: +1 210 4586592

1 Introduction

A variety of statistical techniques have been proposed in the literature to model housing prices in a certain region/market over a particular period of time, which are generically called *hedonic* regression models. These include linear regression models and random field models. This work proposes a model from the latter class intended for situations where there is a *localized externality* in the region that influences the selling price of a house. Examples of such externalities include proximity to a major highway, a nuclear power plant and an airport.

Numerous factors affect the selling price of a house. The most obvious are house-specific characteristics, such as living area, number of rooms, size of the parcel of land and age of the dwelling. Unless the region is urbanistically homogeneous, there will be neighborhood-specific characteristics that would also affect the selling price of the house, such as quality of schools, quality of public services and median travel time to main job centers. These house and neighborhood characteristics used in hedonic regression models explain a large fraction of the selling price variability, but many studies have found that there is still substantial unexplained variability (Pace, Barry and Sirmans, 1998; Bowen, Mikelbank and Prestegaard, 2001). This unexplained variability is mainly attributable to the exclusion of some relevant house/neighborhood characteristics from the model (due to lack of such information) and the methods currently employed to appraise houses' value that use nearby 'comparables'. As a result housing prices are also influenced by spatial effects, meaning that, *ceteris paribus*, two houses located near each other tend to appraise more similarly than two houses located far apart.

Spatial effects have been modeled using random fields, such as lattice models (Pace and Gilley, 1997; Kim, Philips and Anselin, 2003) and geostatistical models (Basu and Thibodeau, 1998; Dubin, 1998). The latter class provides some advantages over lattice models in terms of model interpretability and ability for prediction. Some comparisons between the two modeling approaches appear in Case, Clapp, Dubin and Rodriguez (2004) and Militino, Ugarte and Garcia-Reinaldos (2004). Models for spatio-temporal data have been proposed by Pace, Barry, Gilley and Sirmans (2000) and Gelfand, Ecker, Knight and Sirmans (2004).

For traditional geostatistical models, the spatial association between the selling price of two houses is modeled as a function of the distance between their locations (called isotropic association). This is a reasonable assumption in many situations, but not so for situations when there is a localized externality in the region that exerts an effect on housing prices. For the problem considered here it is expected that, *ceteris paribus*, two houses located about the same distance of the externality location tend to appraise similarly, even if they are not located near each other. Isotropic and stationary models do not represent such behavior.

We consider a cross-section of housing prices data collected during 2001 in Ceder Falls, Iowa, a region that contains a *hoglot*. It is anticipated that the distance of a house to the hoglot exerts a (negative) effect on its selling price, and as a result the mean structure of housing prices and/or the covariance structure between housing prices would depend on the distance to the hoglot. Modeling this kind of data requires the use of non-stationary processes that posses the aforementioned behavior.

A review of the spatial modeling literature reveals a scarcity of models to describe this type of data. In modeling spatial environmental data driven by a 'point source', Hughes-Oliver, Gonzalez-Farias, Lu and Chen (1998) proposed a correlation model that accounts for this kind of effect, but it lacks practical motivation and the model parameters must satisfy somewhat awkward restrictions to guarantee validity. Hughes-Oliver and Gonzalez-Farias (1999) also proposed to model the effect of an externality in the form of a point source by combining two basic independent processes. One stationary representing what the process would be in the case the point source were absent (interpreted as a base line), and the other non-stationary representing a 'shock' to the system exerted by the presence of the point source. The two process components can be combined *multiplicatively* or *additively.* Hughes-Oliver and Gonzalez-Farias (1999) proposed a particular form of non-stationary 'shock' process by combining it multiplicatively with the baseline stationary process. More recently, Martin, Di Battista, Ippoliti and Nissi (2006) proposed a model for situations where there are several point sources, focusing on modeling the mean function of the process. They considered the cases when the locations of the point sources are known or unknown, but in the latter case serious estimation and identification problems were reported.

In this work we propose a model similar to the one studied by Hughes-Oliver and Gonzalez-Farias (1999), but combine the two process components additively rather multiplicatively. The effect of the localized externality on housing prices is modeled through the mean and covariance functions of the process. For the resulting process, the correlation between the prices of two houses depends on both the distance between the houses and the difference between the distances of the houses to the localized externality. A Bayesian approach is used to make

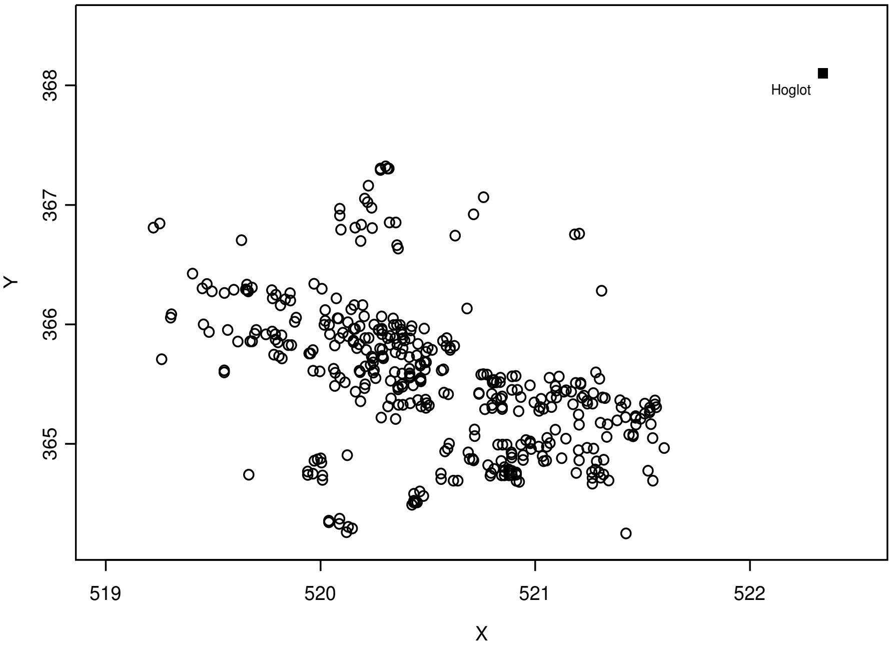

Figure 1: Locations of houses sold in Cedar Falls in 2001 (○) and the hoglot.

inference about the model parameters using a Markov chain Monte Carlo algorithm.

The organization of the paper is as follows. Section 2 describes the housing prices dataset in Cedar Falls, a region with a hoglot, and carries an exploratory data analysis that motives the model construction. Section 3 describes the proposed non-stationary geostatistical model that represents the spatial variation suggested by the exploratory data analysis. Section 4 describes the model fitting and presents the results of applying the model to the Cedar Falls dataset. Conclusions are given in Section 5 and an Appendix provides some technical details.

2 Data and Exploratory Analysis

The dataset used in this analysis consists of 437 arms-length single family home sales during 2001 in the small mid-western city of Cedar Falls, Iowa. A schematic description of the region and the location of the sold houses are displayed in Figure 1. The region contains a hoglot located in the northeast, with all parcels located within 8.4 miles of it, and an average distance to the hoglot of about 6 miles. There are 18 other hoglots in the Black Hawk County (to

Table 1: Summary statistics

Variable	Mean	Standard Deviation
Living Area (square feet)	1320.2	486.3
Number of Rooms	5.75	1.55
Parcel Size (acres)	0.24	0.61
Year Built	1958	27.85
Distance to Hoglot (miles)	6.05	0.80
Selling Price (U.S. dollars)	135548	70727

which Cedar Falls belongs), but the one in Figure 1 is the closest to the city. The dataset was obtained from the Black Hawk County Board of Supervisors. Besides the selling price, the variables collected with each sale were: living area, number of rooms, size of the parcel of land on which the house is built, year the dwelling was built and the location of the house (with spatial coordinates in units of 10000 feet).

The original sales were parced by selecting homes with a selling price between \$32000 and \$400000, with at least 3 rooms and no more than 12 rooms, with at least 500 square feet of living area and lot sizes of at least 3000 square feet. The typical house was built in 1958, with a mean price of about \$136000 and a median price of about \$109000. Most homes were sold in the downtown area, where the mean parcel size and living area were only about a quarter of an acre and 1320 square feet, respectively. Newer and much larger houses have been built in the northwest and southern portions of the city. Summary statistics are given in Table 1.

An ordinary least squares regression was run using log selling price as the response variable and log living area, number of rooms, log parcel size, year built and the distance to the hoglot as explanatory variables; Table 2 summarizes the regression results. All explanatory variables are strongly significant (at the 0.001 level) and the signs of their corresponding regression coefficients are all positive (as expected), so larger and newer houses tend to sell for more. Also, the estimated coefficient corresponding to distance to the hoglot is positive, so houses closer to the hoglot tend to sell for less. An R^2 statistic of 0.71 indicates that, overall, these explanatory variables provide a reasonable explanation for home log selling prices.

To investigate how the mean and variance of log selling price of a house might vary with distance to the hoglot, we performed a similar regression analysis as before, but now omitting

Table 2: OLS regression results

Variable	Parameter Estimate	P-value
Log Living Area	0.4159	< 0.0001
Number of Rooms	0.0857	< 0.0001
Log Parcel Size	0.0896	< 0.0001
Year Built	0.0052	< 0.0001
Distance to Hoglot	0.1189	< 0.0001

the distance to the hoglot as an explanatory variable. All the other explanatory variables remained strongly significant. The plot of residuals versus distance to the hoglot, given in Figure 2a, indicates that almost all houses within about five miles of the hoglot have negative residuals, so these houses tends to sell for less after adjusting for house-specific characteristics. For houses located more than about five miles from the hoglot the distribution of the residuals

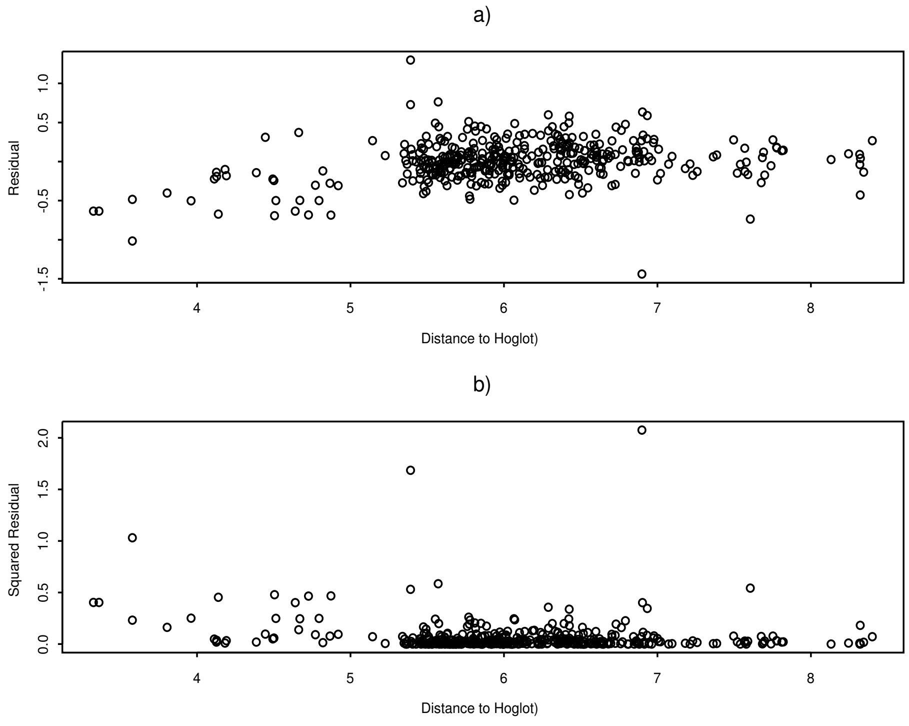

Figure 2: (a) Residuals versus distance to the hoglot, and (b) Squared residuals versus distance to the hoglot.

Table 3: Concentric circle effects of distance

Distance to Hoglot	Number of Sales	γ_2 Estimate
8.5	437	0.7644
8	429	0.7643
7.5	412	0.7844
7	400	0.7784
6.5	331	0.8231
6	216	0.6025
5.5	74	0.1326
5	39	0.4945

is close to being symmetric around zero. Hence, a distance of about five miles to the hoglot seems to act as a 'change point' for the mean log selling price of a house.

The plot of squared residuals versus distance to the hoglot, given in Figure 2b, suggests a moderate increase in variability as houses get closer to the hoglot. A possible explanation for this variability increase is that when sellers or buyers are aware of presence of the hoglot, the house might sell at a significant discount (compared to properties whose seller and buyer are unaware of the presence of the hoglot). To investigate from another point of view this apparent increase in variability with proximity to the hoglot, we considered concentric circles of increasing radii centered at the hoglot. For each circle and with all houses located within it, we linearly regressed log selling price on all explanatory variables, except distance to the hoglot. Then we used the squared residuals from this analysis, e^2, to fit the model $e^2 = \gamma_1 + \exp(-\gamma_2 * \texttt{distance to hoglot}) + \epsilon$; the results are displayed in Table 3. All the estimates of γ_2 are significant at the 0.01 level, and the estimates of γ_2 from the data in the circles of radii 6 or larger are close in magnitude. This provides additional support to the findings in Figure 2b, suggesting that house log selling price variability increases with proximity to the hoglot.

We now examine the data to explore for evidence and nature of spatial correlation. For that we use the Empirical Semivariogram Contour (ESC) plot (Ecker and Gelfand, 1999) based on residuals from the OLS regression that omits the hoglot variable. This plot provides information on how the correlation structure changes with direction. Let $\mathbf{h} = \mathbf{s} - \mathbf{u}$ be the separation

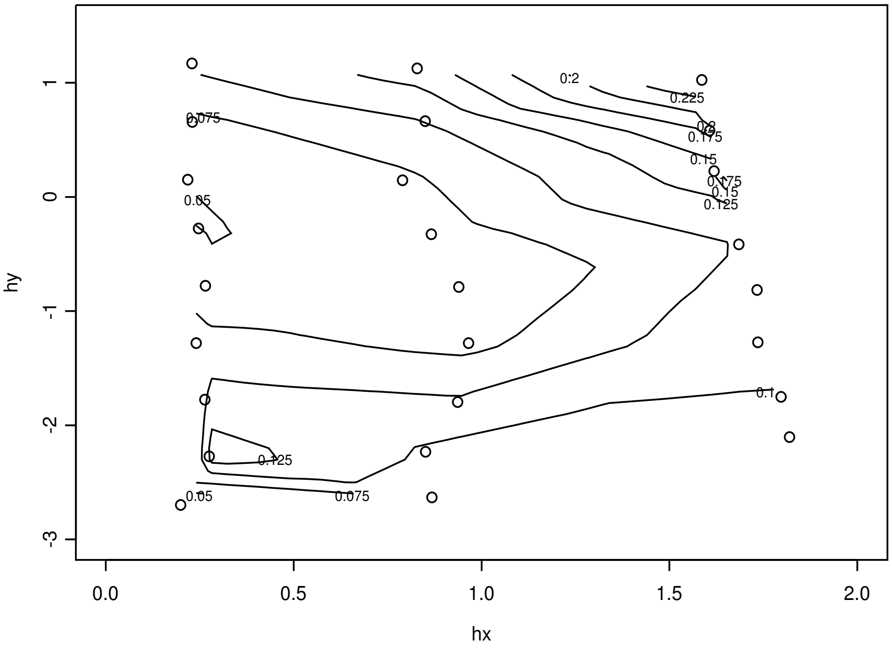

Figure 3: Empirical semivariogram contour plot

vector for the pair of locations $\mathbf{s}, \mathbf{u}$. To construct the ESC plot the plane of separation vectors $\mathbf{h} = (h_x, h_y)$ is divided into cells, and for each cell squared differences are averaged for all pairs of residuals corresponding to locations whose separation vector falls within the cell; we require $h_x \geq 0$ to remove dependence on the arbitrary ordering of two sites.

The ESC plot in Figure 3 indicates reasonably strong spatial correlation, with the highest variability in the north east corner of Figure 3. Since the hoglot is in this direction for all sites in the region, Figure 3 graphically supports a model having the covariance structure for home selling prices being a function of the hoglot distance.

We now investigate how the correlation between log selling prices of two houses may depend on their distance to the hoglot. For that we computed an empirical semivariogram plot based on residuals from the regression on all explanatory variables (including hoglot distance) that is tailored to detect the kind of spatial correlation described above. The distance range $(0, 3)$ was divided into bins, and for each bin squared differences are averaged for all pairs of residuals corresponding to locations where the absolute value of the difference between the distances of the houses to the hoglot falls within the bin. This empirical semivariogram plot, given in

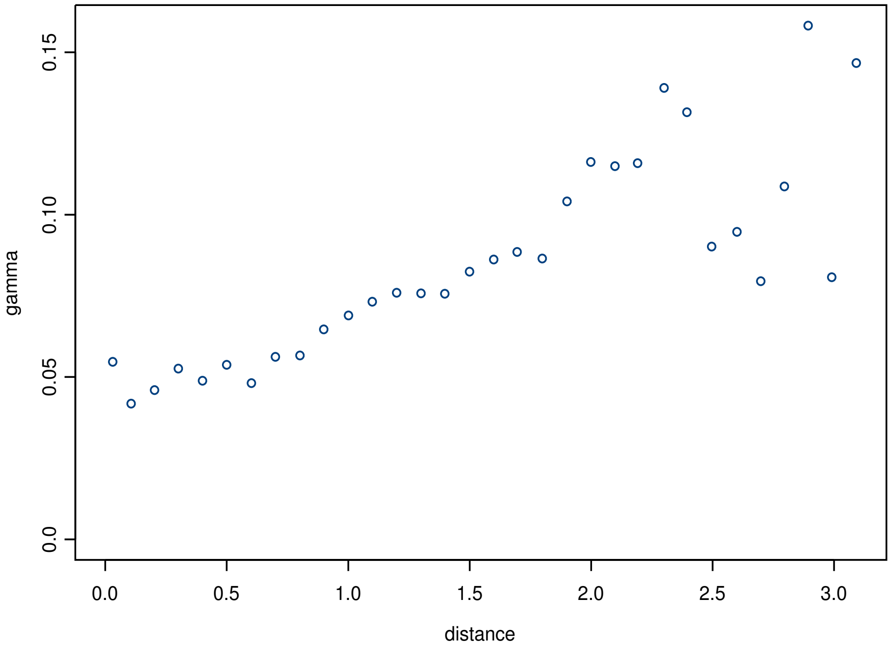

Figure 4: Hoglot distance empirical semivariogram

Figure 4, displays the classical rising up and leveling off that is typically modeled with a sill, range and nugget in the geostatistical literature (see Cressie, 1993). Figure 4 indicates that (log) selling price of houses with similar distances to the hoglot tend to be more correlated than (log) selling price of houses with not so similar distances to the hoglot. Together, Figures 2, 3 and 4 indicate that the distance to the hoglot affects not only the mean structure of (log) selling prices of houses, but also their covariance structure.

3 Model Specification

We propose a model for the log selling price of every actual or potential arms-length single family house sale over the region of interest during the year 2001, that mimics the features of the data revealed in the previous exploratory analysis. We assume houses are located at 'single points' in the region, a mathematical simplification that would have no real bearing on the conclusions since the region is predominately a residential neighborhood with lot and house sizes negligible compared to the size of the region.

We consider a model constructed under the premise that the main effects on housing prices described in the Introduction are additive. Several of these additive components explain house-specific characteristics effects and purely spatial effects on log selling prices under the hypothetical scenario that, were no hoglot were in the neighborhood, a baseline process would result. In addition, a model component explains the effect of the hoglot on log housing prices as a 'shock' process. We adopt the following notation: $D \subset R^2$ represents the region of interest, $\mathbf{s} \in D$ denotes a generic location with coordinates x and y say, and $\mathbf{s}^*$ denotes the location of the hoglot. Also, $||\mathbf{s}||$ denotes the (Euclidean) norm of $\mathbf{s}$ and $|x|$ the absolute value of $x \in R$.

Let $Y(\mathbf{s})$ represent the log selling price of a house located at $\mathbf{s}$. Then

$$Y(\mathbf{s}) = \sum_{j=0}^{p} \beta_j f_j(\mathbf{s}) + g_1(d_{\mathbf{s}}; \boldsymbol{\alpha}) + W_1(\mathbf{s}) + W_2(\mathbf{s}) + \epsilon(\mathbf{s}), \quad \mathbf{s} \in D, \tag{1}$$

where $f_0(\mathbf{s}) = 1$, $f_1(\mathbf{s}), \ldots, f_p(\mathbf{s})$ are known house-specific characteristics, $(\beta_0, \ldots, \beta_p) \in R^{p+1}$ and $\boldsymbol{\alpha} \in R^q$ are unknown regression parameters, $g_1(\cdot; \boldsymbol{\alpha})$ is function of $d_{\mathbf{s}} = ||\mathbf{s} - \mathbf{s}^*||$ and $\{\epsilon(\mathbf{s}) : \mathbf{s} \in D\}$ is white noise with variance $\tau^2 \geq 0$ representing unstructured (non-spatial) variation. The random fields $\{W_1(\mathbf{s}) : \mathbf{s} \in D\}$ and $\{W_2(\mathbf{s}) : \mathbf{s} \in D\}$ are zero-mean Gaussian and independent, as well as independent of $\epsilon(\cdot)$, with respective covariance functions given by

$$\text{cov}\{W_1(\mathbf{s}), W_1(\mathbf{u})\} = (1 - \lambda)(g_2(d_{\mathbf{s}}; \boldsymbol{\gamma}) g_2(d_{\mathbf{u}}; \boldsymbol{\gamma}))^{\frac{1}{2}} K_1(||\mathbf{s} - \mathbf{u}||; \theta_1),$$

$$\text{cov}\{W_2(\mathbf{s}), W_2(\mathbf{u})\} = \lambda(g_2(d_{\mathbf{s}}; \boldsymbol{\gamma}) g_2(d_{\mathbf{u}}; \boldsymbol{\gamma}))^{\frac{1}{2}} K_2(|d_{\mathbf{s}} - d_{\mathbf{u}}|; \theta_2),$$

where $g_2(\cdot; \boldsymbol{\gamma})$ is a positive function of $d_{\mathbf{s}}$, with $\boldsymbol{\gamma}$ and $\lambda \in [0, 1]$ unknown parameters. $K_1(\cdot;\ \theta_1)$ is an isotropic correlation function in R^2 and $K_2(\cdot;\ \theta_2)$ is a stationary correlation function in R^1, with θ_1, θ_2 unknown correlation parameters. The motivation for the latter correlation function and its validity as a correlation function in R^2 is given in the Appendix. The process $W_1(\cdot)$ can be loosely interpreted as the baseline spatial random effect while the process $W_2(\cdot)$ can be loosely interpreted as the hoglot random effect. The random field $Y(\cdot)$ is Gaussian and non-stationary with

$$\text{E}\{Y(\mathbf{s})\} = \sum_{j=0}^{p} \beta_j f_j(\mathbf{s}) + g_1(d_{\mathbf{s}}; \boldsymbol{\alpha}) \quad \text{and} \quad \text{var}\{Y(\mathbf{s})\} = g_2(d_{\mathbf{s}}; \boldsymbol{\gamma}) + \tau^2.$$

Hence $g_2(\cdot; \boldsymbol{\gamma})$ models how the variance of log selling prices varies with location. We may have $g_2(d_{\mathbf{s}}; \boldsymbol{\gamma}) = \gamma_1$, in the case of homogeneous variability, or $g_2(d_{\mathbf{s}}; \boldsymbol{\gamma})$ varying with $d_{\mathbf{s}}$, as suggested by the data analysis in Section 2. Also, note that when $\tau^2 = 0$, the correlation structure of $Y(\cdot)$ is given by

$$\mathrm{corr}\{Y(\mathbf{s}), Y(\mathbf{u})\} = (1-\lambda)K_1(||\mathbf{s}-\mathbf{u}||;\theta_1) + \lambda K_2(|d_{\mathbf{s}} - d_{\mathbf{u}}|;\theta_2),$$

which is a mixture between $K_1(\cdot;\theta_1)$ that depends on the distance between the houses, and $K_2(\cdot;\theta_2)$ that depends on the absolute value of the differences between distances of houses and the hoglot; the parameter λ quantifies the importance of $K_2(\cdot;\theta_2)$ in the above mixture relative to $K_1(\cdot;\theta_2)$. A similar correlation structure occurs when $\tau^2 > 0$, but the above interpretation would not be so straightforward.

From the results of the exploratory analysis in Section 2 we use the model with $p = 4$ where $\{f_j(\mathbf{s})\}_{j=1}^4$ are the four house-specific characteristics, and

$$g_1(d_{\mathbf{s}};\alpha) = \alpha \times \min(d_{\mathbf{s}}, 5.25) \quad , \quad g_2(d_{\mathbf{s}};\boldsymbol{\gamma}) = \gamma_1 + \exp(-\gamma_2 d_{\mathbf{s}}),$$

and the correlation functions of the $W_1(\cdot)$ and $W_2(\cdot)$ processes are assumed to be exponential

$$K_1(h;\theta_1) = \exp(-\theta_1 h) \quad , \quad K_2(h;\theta_2) = \exp(-\theta_2 h); \qquad \theta_1, \theta_2 > 0.$$

Let $\mathbf{Y} = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))'$ denote the $n = 437$ log selling prices of houses at locations $\mathbf{s}_1, \ldots, \mathbf{s}_n \in D$. Then we have

$$\mathbf{Y} \sim \mathrm{N}_n(X_1\boldsymbol{\beta} + X_2\alpha, \Sigma(\lambda, \boldsymbol{\gamma}, \boldsymbol{\theta}) + \tau^2 I_n), \tag{2}$$

where X_1 is a known $n \times 5$ matrix defined by $(X_1)_{i1} = 1$ and $(X_1)_{ij} = f_{j-1}(\mathbf{s}_i)$, $2 \leq j \leq 5$, X_2 is a known $n \times 1$ matrix defined by $(X_2)_{i1} = \min(d_{\mathbf{s}_i}, 5.25)$, I_n is the identity matrix and $\Sigma(\lambda, \boldsymbol{\gamma}, \boldsymbol{\theta})$ is an $n \times n$ non-negative definite matrix defined by

$$\Sigma(\lambda, \boldsymbol{\gamma}, \boldsymbol{\theta})_{ij} = \Big(g_2(d_{\mathbf{s}_i};\boldsymbol{\gamma})g_2(d_{\mathbf{s}_j};\boldsymbol{\gamma})\Big)^{\frac{1}{2}}\Big((1-\lambda)K_1(||\mathbf{s}_i - \mathbf{s}_j||;\theta_1) + \lambda K_2(|d_{\mathbf{s}_i} - d_{\mathbf{s}_j}|;\theta_2)\Big).$$

Remark. For any combination of parameters, sample size and sampling locations the covariance matrix of the data is positive definite, but this is not necessarily so for $\Sigma(\lambda, \boldsymbol{\gamma}, \boldsymbol{\theta})$; for computational reasons we then require $\tau^2 > 0$. The fact that linear combinations of positive definite functions are not necessarily positive definite was pointed out by Myers and Journel (1990).

To complete the model specification, we assume the model parameters are independent *a priori* with marginal distributions given by

$$p(\boldsymbol{\beta}) \propto 1 \quad , \quad p(\alpha) \propto 1 \quad , \quad \lambda \sim \text{unif}[0,1]$$

$$\tau^2 \sim \text{IG}(2, a) \quad , \quad \gamma_i \sim \text{IG}(2, b_i) \quad , \quad \theta_i \sim \text{IG}(2, c_i), \quad i = 1, 2, \tag{3}$$

where $a, b_i, c_i > 0$ are known and $\text{IG}(p_1, p_2)$ denotes the inverse gamma distribution with mean $(p_2(p_1 - 1))^{-1}$. The priors for the model parameters are intended to be vague, where in particular the marginal prior variances of τ^2, γ_i and θ_i are infinite. Additional inverse gamma distributions could be placed on the hyper-parameters, if an extra layer of prior uncertainty were desired.

4 Model Fitting and Results

We use the Bayesian approach to fit the proposed model from Section 3, where posterior inference about the model parameters relies on a standard Monte Carlo Markov Chain (MCMC) algorithm, a Gibbs sampler with a Metropolis step. The mean parameters $\boldsymbol{\mu} = (\boldsymbol{\beta}, \alpha)$ are simulated using a Gibbs sampler step, while the covariance parameters $\Delta = (\lambda, \boldsymbol{\gamma}, \boldsymbol{\theta}, \tau^2)$ are simulated using a Metropolis sampler step. The posterior distribution of the model parameters is determined by (2) and (3) and the full conditional distribution of $\boldsymbol{\mu}$ is

$$p(\boldsymbol{\mu} \mid \Delta, \mathbf{y}) = \text{N}_{p+2}((X'(\Sigma + \tau^2 I_n)^{-1} X)^{-1}(X'(\Sigma + \tau^2 I)^{-1}\mathbf{y}), (X'(\Sigma + \tau^2 I_n)^{-1} X)^{-1}),$$

where $X = [X_1 \ X_2]$ and $\Sigma = \Sigma(\lambda, \boldsymbol{\gamma}, \boldsymbol{\theta})$ are given in (2). Likewise, the full conditional distribution of Δ is

$$\begin{aligned} p(\Delta \mid \boldsymbol{\mu}; \mathbf{y}) \quad \propto \quad & |(\Sigma + \tau^2 I_n)|^{-\frac{1}{2}} \exp\{(\mathbf{y} - X\boldsymbol{\mu})'(\Sigma + \tau^2 I_n)^{-1}(\mathbf{y} - X\boldsymbol{\mu})\} \\ & \times \, \pi(\tau^2)\pi(\gamma_1)\pi(\gamma_2)\pi(\theta_1)\pi(\theta_2), \end{aligned}$$

where $\pi(\cdot)$ represent the inverse gamma prior densities of $\tau^2, \gamma_1, \gamma_2, \theta_1$ and θ_2 respectively given in (3). This complete conditional distribution is not of standard form, so we use a Metropolis step to simulate Δ as a block. As proposal distribution for $\log(\Delta)$, we used a multivariate normal distribution with mean vector equal to previous iteration and diagonal covariance matrix with variances chosen by trial and error. Four chains were run for 25000 iterations each. Trace plots of the regression parameters (not shown) and the covariance parameters in Figure 5 suggest very reasonable mixing and convergence.

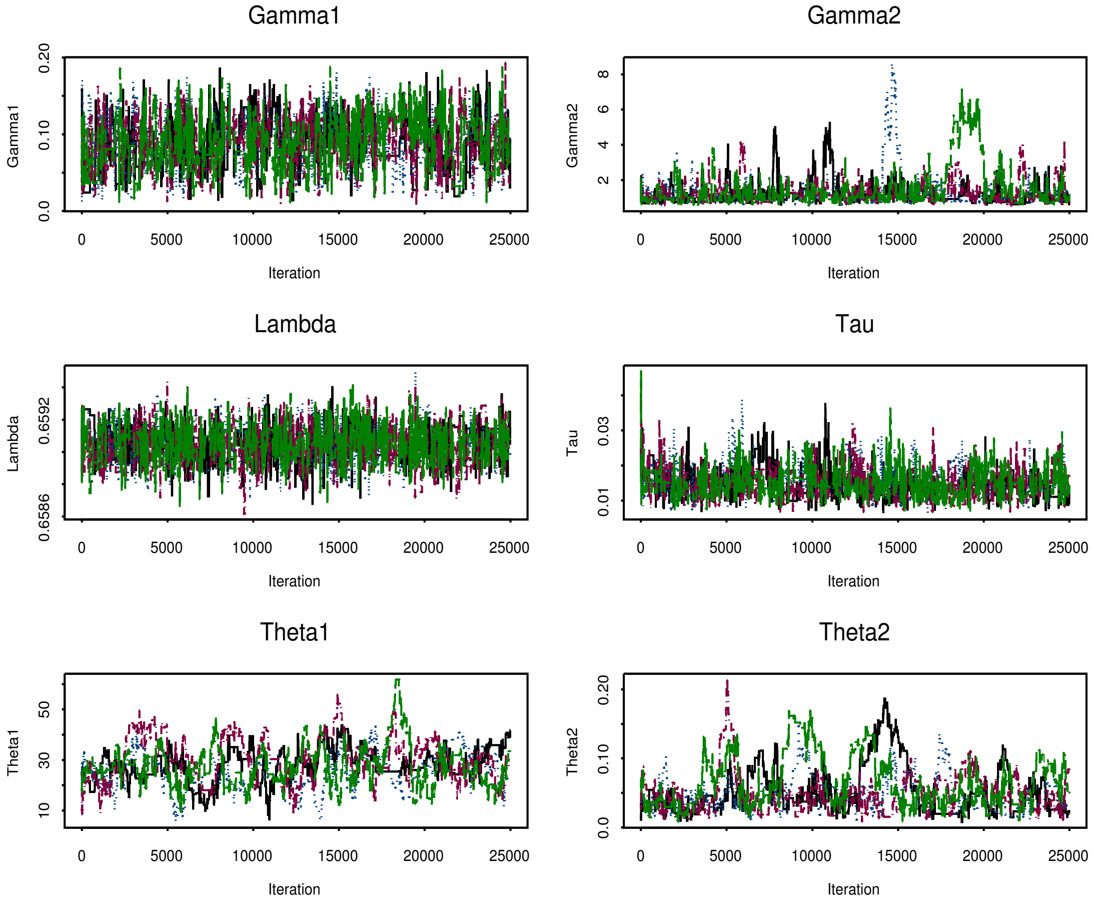

Figure 5: MCMC strings for covariance parameters

Posterior summaries for the mean parameters are given in Table 4. The 95% credible intervals of β_1 through β_4 contain only positive values, confirming the significance of all the house-specific explanatory variables. The regression parameter estimates in Table 4 are (as expected) all positive and the interpretation of the effect of the corresponding explanatory variables on log selling price is the same as that given in Section 2. In particular, Living area (through the β_1 parameter) and number of rooms (through the β_2 parameter) positively influence the selling price of a house; the bigger the house, the more value. Parcel size (β_3 parameter) is interpreted as the larger the lot size, the more valuable the home built upon it. Year built (β_4 parameter) indicates that newer homes (larger values of year built) are worth more, all other things being equal. A similar conclusion holds for the explanatory variable involving the distance to the hoglot, whereby homes in closer proximity to the hoglot are worth less.

Table 4: Posterior summaries for the model mean parameters.

	Mean	Standard Deviation	95% Interval
β_0	-4.146	1.893	(-7.460,-0.157)
β_1	0.323	0.058	(0.206,0.431)
β_2	0.204	0.034	(0.142,0.275)
β_3	0.068	0.013	(0.038,0.090)
β_4	0.005	0.001	(0.004,0.007)
α	0.457	0.120	(0.221,0.689)

Table 5: Posterior summaries for the model covariance parameters.

	Mean	Standard Deviation	95% Interval
γ_1	0.087	0.033	(0.025,0.147)
γ_2	1.338	0.904	(0.676,5.321)
θ_1	27.820	7.881	(14.358,46.530)
θ_2	0.053	0.033	(0.015,0.136)
τ^2	0.015	0.004	(0.009,0.023)
λ	0.6591	0.0001	(0.6588,0.6593)

The Bayesian estimates are close to the corresponding ordinary least squares (OLS) estimates in Table 2, except for the estimate of β_2. We note that the OLS analysis uses $d_{\mathbf{s}}$ as explanatory variable, while the Bayesian analysis uses $\min(d_{\mathbf{s}}, 5.25)$, which together with the moderate correlations between some of the explanatory variables could potential cause the discrepancy between the two estimates of β_2.

Posterior summaries for the covariance parameters are given in Table 5. Note that the estimate of θ_2 is close to zero while the estimate of θ_1 is far from zero, which means the spatial correlation between the (log) selling price of two houses depends much more on their relative distance to the hoglot, $|d_{\mathbf{s}} - d_{\mathbf{u}}|$, than on the distance between them, $||\mathbf{s} - \mathbf{u}||$. This conclusion is also supported by the estimate of λ being larger than 0.5. Therefore, the purely spatial variability of (log) selling prices (that beyond what is explained by house-specific characteristics) is mostly due to the hoglot effect.

5 Conclusions

Recent advances on the modeling of real estate data has shown the importance of including spatial effects to describe housing prices variability in a certain region/market. The commonly used random fields with isotropic and stationary covariance functions do not adequately represent the housing prices variability in regions/markets that contain a localized externality affecting housing prices. For this situation this work proposes a new non-stationary random field to model the spatial variation of housing prices. The effect of the localized externality on housing prices can be modeled through the mean function, the covariance function or both. In particular, the variability of housing prices may depend on the distance to the localized externality, and the proposed correlation structure combines isotropic and non-isotropic correlations functions.

Acknowledgments

The work of the second author was partially supported by National Science Foundation Grant DMS-0719508, and a College of Business (The University of Texas at San Antonio) summer research grant.

Appendix

Motivation and Validity of $K_2(\cdot)$.

Let $X(t)$ be a zero-mean stationary process in R^1 with covariance function given by $C(|h|) = \text{cov}\{X(t+h), X(t)\}$. For an arbitrary transformation $T : R^2 \to R^1$, define $\tilde{X}(\mathbf{s}) = X(T(\mathbf{s}))$, $\mathbf{s} \in R^2$. Then $\tilde{X}(\cdot)$ is a zero-mean random field in R^2 with covariance function

$$\text{cov}\{\tilde{X}(\mathbf{s}), \tilde{X}(\mathbf{u})\} = \text{cov}\{X(T(\mathbf{s})), X(T(\mathbf{u}))\} = C(|T(\mathbf{s}) - T(\mathbf{u})|).$$

A distinctive feature of the random field so constructed is that, for any y in the range of the transformation T, $\tilde{X}(\mathbf{s})$ is constant for all $\mathbf{s}$ in $T^{-1}(\{y\})$. Then the random field $W_2(\cdot)$ defined in Section 3.1 can be obtained, using $T(\mathbf{s}) = d_{\mathbf{s}}$, as $W_2(\mathbf{s}) = (\lambda g_2(T(\mathbf{s}); \boldsymbol{\gamma}))^{\frac{1}{2}} \tilde{X}(\mathbf{s})$, and hence $K_2(|d_{\mathbf{s}} - d_{\mathbf{u}}|; \theta_2)$ is a correlation function in R^2. The random field $W_2(\cdot)$ takes the same value for all locations at the same distance from $\mathbf{s}^*$.

References

Basu, S. and Thibodeau, T.G. (1998). Analysis of Spatial Autocorrelation in House Prices. *The Journal of Real Estate Finance and Economics*, 17, 61-85.

Bowen, W.M., Mikelbank, B.A. and Prestegaard, D.M. (2001). Theoretical and Empirical Considerations Regarding Space in Hedonic Housing Price Model Applications. *Growth and Change*, 32, 466-490.

Case, B., Clapp, J., Dubin, R. and Rodriguez, M. (2004). Modeling Spatial and Temporal House Price Patterns: A Comparison of Four Models. *The Journal of Real Estate Finance and Economics*, 29, 167-191.

Cressie, N. (1993). *Statistics for Spatial Data* (rev. ed). New York: Wiley.

Dubin, R.A. (1998). Predicting House Prices Using Multiple Listings Data. *The Journal of Real Estate Finance and Economics*, 17, 35-59.

Dubin, R., Pace, R.K. and Thibodeau, T.G. (1999). Spatial Autoregression Techniques for Real Estate Data. *Journal of Real Estate Literature*, 7, 79-95.

Ecker, M. and Gelfand, A. (1999). Bayesian Modeling and Inference for Geometrically Anisotropic Spatial Data. *Mathematical Geology*, 31, 67-83.

Gelfand, A.E., Ecker, M., Knight, J. and Sirmans, C.F. (2004). The Dynamics of Location in House Price. *Journal of Real Estate Finance and Economics*, 29, 149-166.

Hughes-Oliver, J.M., Gonzalez-Farias, G., Lu, J-C., and Chen, D. (1998). Parametric Nonstationary Correlation Models. *Statistics and Probability Letters*, 40, 267-278.

Hughes-Oliver, J.M. and Gonzalez-Farias, G. (1999). Parametric Covariance Models for Shock-induced Stochastic Processes. *Journal of Statistical Planning and Inference*, 77, 51-72.

Kim, C.W., Philips, T.T and Anselin, L. (2003). Measuring the Benefits of Air Quality Improvement: A Spatial Hedonic Approach. *Journal of Environmental Economics and Management*, 45, 24-39.

Martin, R.J., Di Battista, T., Ippoliti, L. and Nissi, E. (2006). A Model for Estimating Point Sources in Spatial Data. *Statistical Methodology*, 3, 431-443.

Militino, A.F., Ugarte, M.D. and Garcia-Reinaldos, L. (2004). Alternative Models for Describing Spatial Dependence Among Dwelling Selling Prices. *The Journal of Real Estate Finance and Economics*, 29, 193-209.

Myers, D.E. and Journel, A. (1990). Variograms with Zonal Anisotropies and Noninvertible Kriging Systems. *Mathematical Geology*, 22, 779-785.

Pace, R.K. and Gilley, O.W. (1997). Using the Spatial Configuration of the Data to Improve Estimation. *The Journal of Real Estate Finance and Economics*, 17, 333-340.

Pace, R.K., Barry, R., Gilley, O.W. and Sirmans, C.F. (2000). A Method for Spatio-temporal Forecasting with an Application to Real Estates Prices. *International Journal of Forecasting*, 16, 229-246.

Pace, R.K., Barry, R. and Sirmans (1998). Spatial Statistics and Real Estate. *The Journal of Real Estate Finance and Economics*, 17, 5-13.