A NON–STATIONARY NON-GAUSSIAN HEDONIC SPATIAL MODEL FOR HOUSE SELLING PRICES

Victor De Oliveira[1]

Department of Management Science and Statistics

The University of Texas at San Antonio

San Antonio, TX 78249, USA

victor.deoliveira@utsa.edu

Mark D. Ecker

Department of Mathematics

University of Northern Iowa

Cedar Falls, IA 50614, USA

ecker@math.uni.edu

September 12, 2016

Abstract

This work proposes a hedonic random field model to describe house selling prices over the period 2000–2005 in Cedar Falls, Iowa. This real estate market presents two distinctive features that are not described by commonly used stationary Gaussian random field models: (a) the city encompasses within its limits a hog lot which acts as an externality, negatively affecting the selling price of nearby houses, and (b) the distribution of house selling prices displays heavy tails. A non-stationary and non-Gaussian random field model is constructed by multiplying two independent Gaussian random fields, where the factors are tailored so the resulting model describes the aforementioned distinctive features. We also propose an empirical diagnostic to assess the fit of the proposed model to a given dataset.

Key words: Geostatistics; hedonic models; hog lot; localized externality; spatial correlation.

JEL Classifications: C16, C21

[1]This project was funded by the University of Texas at San Antonio, Office of the Vice President for Research.

1 Introduction

Models that describe house selling prices over a period of time in a certain region or market are generically called *hedonic models*. Due to the inherent heterogeneity of house characteristics and market uncertainty, this type of data is often described using statistical models, such as regression models, hierarchical models and random field models (Basu and Thibodeau, 1998; Pace, Barry, Gilley and Sirmans, 2000; Malpezzi, 2003). There are numerous factors that affect the price of a house, the most prominent ones being house–specific characteristics, such as living area, number of rooms and age of the dwelling, and neighborhood characteristics, such as quality of schools, quality of the public services and travel time to main job centers. Empirical studies have found that these house and neighborhood characteristics used in hedonic regression models explain a large fraction of the selling price variability, but often there is still substantial unexplained variability. This unexplained variability is partly attributable to the exclusion of some relevant house and/or neighborhood characteristics from the model (due to lack of such information), and partly due to the methods used to appraise houses value that use nearby 'comparables'. As a result of these, housing selling prices are also influenced by spatial effects, so two houses with the same characteristics tend to appraise more similarly when they are located near each other than when they are located far apart. Models that account for spatial effects include geostatistical random fields, which have been applied to describe spatial variation of house selling prices by Basu and Thibodeau (1998), Dubin (1998) and Ecker and De Oliveira (2008), among many others.

For traditional geostatistical models the spatial association between the prices of two houses is modeled as a function of the separation vector of the houses' locations (called stationary association), or more often as a function of the distance between their locations (called isotropic association); see Cressie (1993) and Diggle and Ribeiro (2007). A situation in which this type of stationary models is not adequate is when the region of interest contains a localized externality, such as a hog lot, nuclear power plant or a highly desirable school, which exert an impact on the value of nearby houses. As an example, people in general will be willing to pay a premium to be close to a highly desirable school, while those homes close to a nuclear power plant or a hog lot will suffer some depreciation in value. In this cases the spatial association between selling prices of two houses depends not only on the distance separating them, but also on the distances between the houses and the externality. For such cases, there is a need to construct models that incorporate the latter information as an factor to explain the

spatial variation of houses' selling prices.

Nearly all statistical models assume that these spatial factors affect only the average selling price, in the mean structure of the hedonic model. As an example, Isakson and Ecker (2008) demonstrate this hedonic modeling approach in their study of the impact of hog lots on the selling prices of nearby houses. A statistically significant and positive coefficient associated with an explanatory variable measuring proximity to the externality source is consistent with the presence of a point source, negative externality. Furthermore, one might reasonably expect to see also an increase in variability for houses closer to a hog lot. For example, buyers and sellers of homes closer to the hog lot may be more/less aware of, or bothered/not bothered by its presence, producing as a result a wider range in selling prices for homes close to the hog lot. This extra variability motivates the need for the variance and/or correlation structure of house prices, in addition to the mean structure, to depend upon the distance to the externality. Therefore, the hedonic model may be improved by modeling proximity to the externality in the disturbance term, not only through the mean structure of the model, but also through a non–stationary covariance structure that includes a point source effect.

In this work we explore a dataset of house selling prices over the period 2000–2005 in Cedar Falls, Iowa, a small mid–west city with a population of about 40,000. This dataset presents several interesting challenges for their modeling. First, the city encompasses within its limits several hog lots, two of which are nearby the the most populous part of the city, so these act as externalities negatively affecting selling prices. Second, for some years the distributions of (log) selling prices of houses display heavy tails, due to the heterogeneity of the market that includes some very expensive houses and some very inexpensive ones. Commonly used models constructed on stationarity and Gaussianity assumptions do not represent these behaviors, so there is a need for more general models tailored to describe such data.

This work proposes a non–stationary and non–Gaussian random field model for the (log) selling prices of houses that aims at representing the Cedar Falls real estate market data features described above. The model is written as a spatial trend that accounts for house characteristics and the effect of proximity to the hog lot, plus an error process that accounts for spatial effects. The error process is modeled as the product of two independent Gaussian random fields, where the two factors are tailored to produce a model with probabilistic features that mimic those found in the data. Some of the probabilistic features of such products of Gaussian variables are investigated and it is shown

that, under some conditions, the resulting variables have zero–mean symmetric distributions with tails heavier than Gaussian. By judiciously modeling the mean and covariance functions of the two Gaussian random fields we obtain a model where both the variance and correlation structures may depend on proximity to the hog lot. We describe in detail three variants of this strategy to build the error process, which result in models that are non–Gaussian and non–stationary in both mean and covariance. The proposed model is similar to those proposed in Hughes–Oliver and Gonzalez–Farias (1999) and Palacios and Steele (2006). It generalizes the former, but unlike the latter, it accounts for non–stationarity in the covariance function. Finally, we propose an empirical diagnostic, in the spirit of exploratory data analysis, to assess the adequacy of the proposed model to a given dataset.

2 Literature Review

Models for the analysis of geostatistical data are described in Cressie (1993) and Diggle and Ribeiro (2007), where most models treated in these and other textbooks are stationary. Reviews on strategies to construct non–stationary random processes appear in Treviño (1992), Sampson (2010) and Fouedjio (2016).

Recent spatial models that describe non–stationary random field models aimed at describing processes driven by a shock point source include Hughes–Oliver, Gonzalez–Farias, Lu and Chen (1998), Hughes–Oliver and Gonzalez–Farias (1999), Martin, Di Battista, Ippoliti and Nissi (2006), Ecker and De Oliveira (2008) and Ecker, De Oliveira and Isakson (2013). The common thread among most of these works is the modeling of spatial effects by the combination of two independent Gaussian processes. One of them a stationary process representing the spatial effect on selling prices in an ideally 'externality–free' scenario, and the the other a non–stationary process representing the 'shock' on selling prices due to the presence of the externality. Ecker and De Oliveira (2008) and Ecker, De Oliveira and Isakson (2013) combined the processes additively, which results in a Gaussian random field, while Hughes–Oliver and Gonzalez–Farias (1999) combined the processes multiplicatively, which results in a non–Gaussian random field.

Most works in the literature on this area assume that the (natural log scaled) house selling prices are normally distributed. In particular, Ecker and De Oliveira (2008) develop a model where the covariance function is a mixture of an isotropic correlation function, the distance between house sales and the point source externality, and the discrepancy between two sites' distances to the externality. Their resulting

covariance structure has three parameters more than a standard Geostatistical model (see Ecker, 2003), but assumes the response variable, the log–scaled selling price, to be normally distributed. Lack of normality of the data, or the residuals, from a statistical model, is often handled by taking a non–linear transformation of the response variable, the house selling price. Standard regression analyses assume, then, that the natural log scaled house selling price is normally distributed. However, normality is often not a reasonable assumption, even for the transformed data. In particular, the dataset used in both Ecker and De Oliveira (2008) and Ecker, De Oliveira and Isakson (2013), which consists of house sales in Cedar Falls, Iowa in the early 2000s, appears to have distributions with tails heavier than those of normal distributions, even after taking a natural log transformation. Furthermore, the more expensive homes and the more inexpensive homes in the Cedar Falls dataset have selling prices that are not well explained by their location specific variables and warrant the need for a heavier tailed distribution than normal distribution to more accurately model these data. Indeed, preliminary exploratory analysis of the regression residuals (see Section 3) indicates that t-like–distributions might be anticipated to improve the explanation and prediction of these house prices.

The multiplicative model proposed by Hughes–Oliver and Gonzalez–Farias (1999) is a non–Gaussian random field model with no closed–form expression for its likelihood. Because of this, they fitted the model using a surrogate Gaussian likelihood determined by the true mean and covariance functions derived from the model. This is inefficient at best and inappropriate at worst, since such fitting does not take into account the heavy tail nature of the distribution of the data. This is likely to result in a fitted model that does not predict well extreme selling prices, namely, very expensive or very inexpensive houses, since these are viewed as outliers under a Gaussian model. Palacios and Steel (2006) proposed a model for the error process formed by a scale mixture of Gaussian random fields aimed at representing data with outliers and non–Gaussian tail behaviors. In this model both the Gaussian variables and the mixing variables, assumed to be log–Gaussian, are spatially dependent and share the same stationary correlation structure. It was shown that the resulting Gaussian–log–Gaussian error process is stationary and has symmetric distributions with tails that resemble those of t distributions with degrees of freedom as low as about four. The model can also be interpreted as the product of two independent stationary random fields, one Gaussian and the other log–Gaussian.

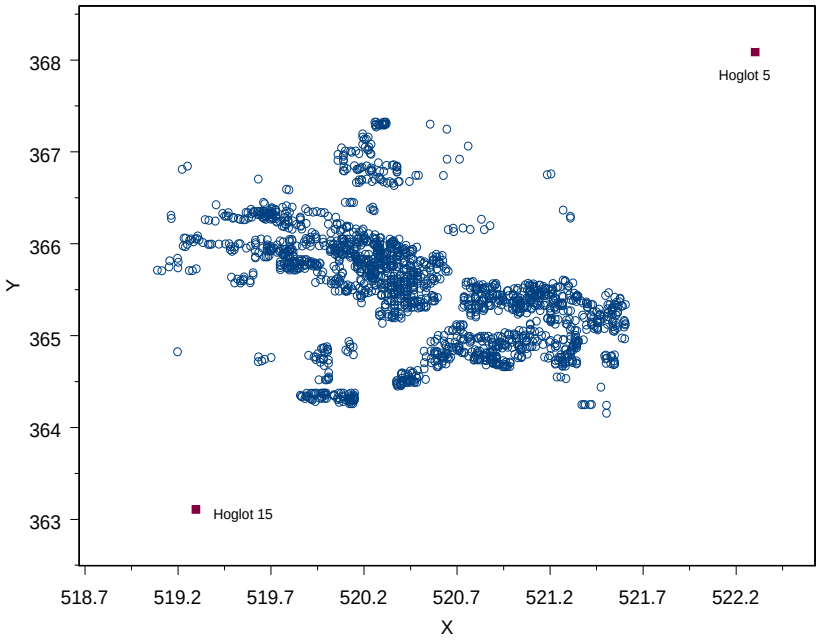

Figure 1: Locations of houses sold in Cedar Falls (○) during the period 2000-2005, and the locations of the two closest hog lots (□).

3 The Data

The dataset used in this analysis consists of 2,297 arms–length single family house sales during the period 2000-2005 in the small mid–western city of Cedar Falls located in Black Hawk County, Iowa. The locations of these houses are displayed in Figure 1. The Black Hawk County contains within its boundaries 19 hog lots, two of which are located near the most populous part if the city; the two hog lots closest to the city are displayed in Figure 1. The dataset considered here, obtained from the Black Hawk County Board of Supervisors, consists of the house selling price plus several variables collected from each house sold: 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 10,000 feet. In addition, we also computed the distance from each house sold to the two closest hog lots. Summary statistics are given in Table 3.

The original sales dataset was parced by selecting homes with a selling price between \$32,000 and \$400,000, with at least 3 rooms and no more than 12 rooms, with at least 500 square feet of living

Table 1: Summary Statistics

Variable	Mean	Standard Deviation
Living Area (square feet)	1369.3	545.1
Number of Rooms	5.96	1.64
Parcel Size (acres)	0.24	0.61
Year Built	1961	28.8
Distance to Hog Lot 5 (miles)	6.2	0.82
Selling Price (U.S. dollars)	152722	85858

area and lot sizes of at least $3,000$ square feet. The typical house was built in 1961, with a mean price of approximately $\$153,000$ and a median price of $\$125,000$. 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 $1,370$ square feet, respectively, but newer and much larger houses have been built in the northwest and southern portions of the city.

In this section we explore the possible effect of hog lot 5 (see Figure 1) on the selling prices of houses, and disregard the possible effect of the other hog lot. See Martin et al. (2006) for possible approaches to deal with multiple point sources, and the difficulties that such modeling may encounter. 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 and year built as explanatory variables; the distance to the hog lot was not included. 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

Table 2: OLS regression results

Variable	Parameter Estimate	P-value
Log Living Area	0.481	< 0.0001
Number of Rooms	0.085	< 0.0001
Log Parcel Size	0.142	< 0.0001
Year Built	0.007	< 0.0001

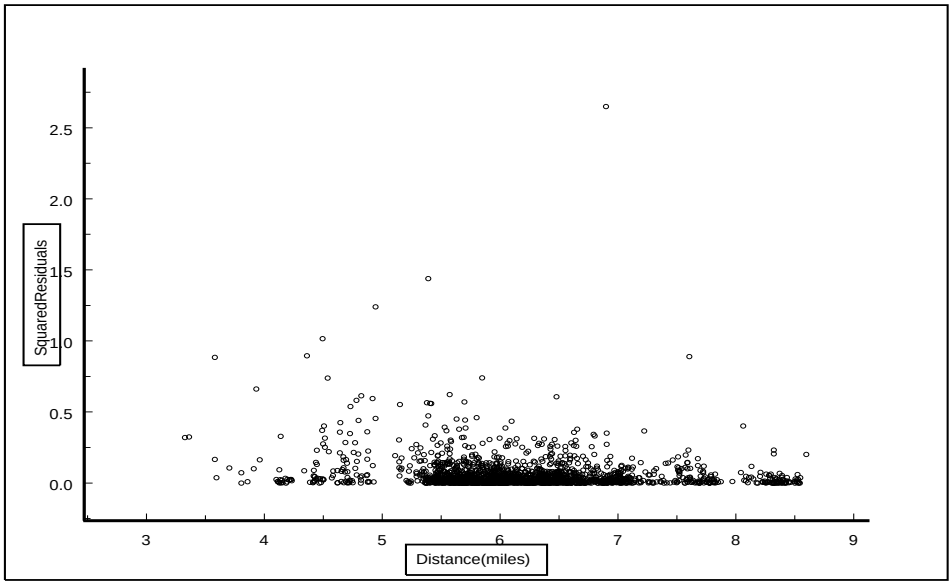

Figure 2: Plot of squared residuals versus distance to the hog lot 5.

all positive (as expected). An R^2 statistic of 0.77 indicates that, overall, these explanatory variables provide a reasonable explanation for house log selling prices.

Now Figure 2 plots the square of the residuals from the above regression versus distance to the hog lot. The plot suggests a moderate tendency of the residuals variability to increases for houses that are located closer to the hog lot. To further investigate this apparent pattern, we fitted the simple linear regression $\texttt{squaredresidual} = \beta_0 + \beta_1 \texttt{distancetohoglot} + \texttt{error}$, resulting in the estimate $\hat{\beta}_1 = -0.036$ being highly significant. All of these suggest that the variance of the (log) selling price is not constant, but rather increases for houses located closer to the hog lot. In addition, Figure 3 displays the histogram of the residuals, showing that these have distributions close to being symmetric, while Figure 4 displays the qq–plot of these residuals. The type of deviation from the straight line at the extreme quantiles indicate that these residuals have distributions with tails heavier than Gaussian, which suggest that the same holds for the distributions of the (log) selling prices. In the next section we describe a possible strategy to construct a model that represent the aforementioned data features.

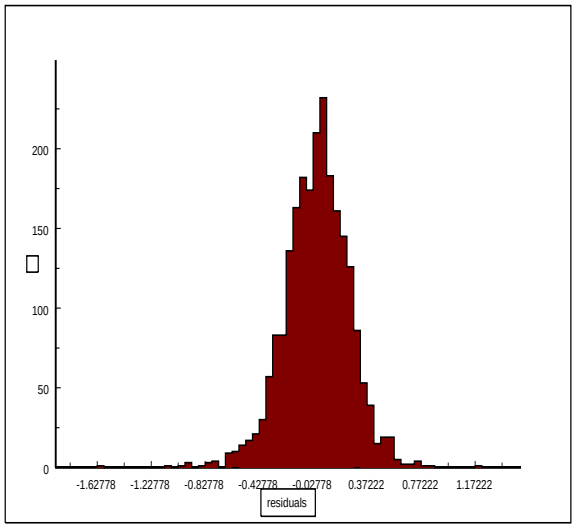

Figure 3: Histogram of the residuals from the regression in Table 2.

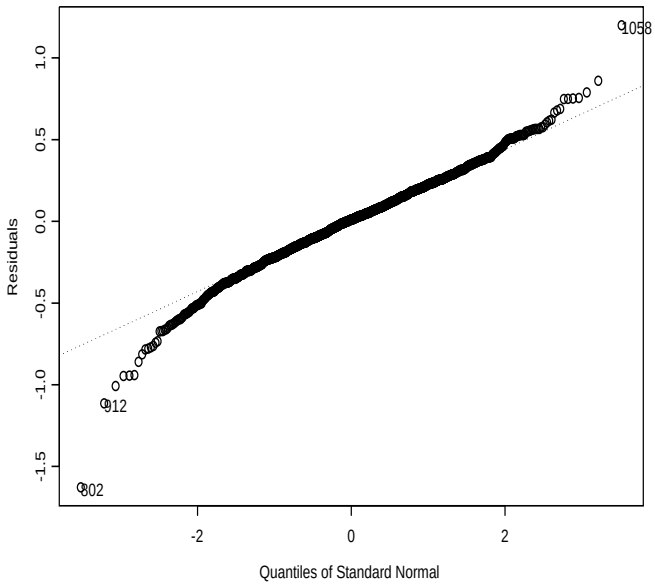

Figure 4: Quantile–quantile plot of the residuals from the regression in Table 2.

4 Modeling Strategy

Let $D \subset \mathbb{R}^2$ represent the region under study, and for a particular (fixed) time period let $Y(\mathbf{s})$ be the log of the (actual/potential) selling price of a house located at $\mathbf{s} \in D$ during this period. In addition, it is assumed that the region D contains an 'entity' located at $\mathbf{s}^*$ that exerts an effect over $Y(\mathbf{s})$ for all $\mathbf{s} \in D$; this entity is often called a localized externality or point source. Also, let $||\mathbf{s}||$ denote the Euclidean norm of $\mathbf{s}$ and $d_{\mathbf{s}} = ||\mathbf{s} - \mathbf{s}^*||$ the distance between a house and the point source.

The random field $Y(\cdot)$ will be modeled as

$$Y(\mathbf{s}) = \mu(\mathbf{s}) + W(\mathbf{s}), \qquad \mathbf{s} \in D,$$

where $\mu(\mathbf{s}) := E(Y(\mathbf{s}))$ is a deterministic spatial trend, and $W(\mathbf{s})$ is a zero–mean random field that will be called the 'error process'. The deterministic trend includes the known house characteristics that influence $Y(\mathbf{s})$ and a possible effect from the localized externality, assumed to have the traditional linear regression form

$$\mu(\mathbf{s}) = \sum_{j=0}^{p} \beta_j f_j(\mathbf{s}) + h(d_{\mathbf{s}}; \alpha),$$

where $f_0(\mathbf{s}) \equiv 1$, $f_1(\mathbf{s}), \ldots, f_p(\mathbf{s})$ are house characteristics and $h(d_{\mathbf{s}}; \alpha)$ is the contribution to the spatial trend due to the point source, where the latter is assumed to depend only on $d_{\mathbf{s}}$. Finally, $\beta_0, \beta_1, \ldots, \beta_p$ and α are unknown regression parameters.

The focus in this work is on the modeling of the error process $W(\cdot)$ in order to mimic the data features revealed in the exploratory data analysis in Section 3. Hughes–Oliver and Gonzalez–Farias (1999) proposed modeling $W(\cdot)$ by combining two basic independent processes. One of them stationary, representing what the process would had been if 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. They proposed to combine the two processes either *multiplicatively* or *additively*. Hughes–Oliver and Gonzalez–Farias (1999) proposed a specific error model constructed multiplicatively, while Ecker and De Oliveira (2008) proposed an alternative error model constructed additively. Both constructions used Gaussian random fields for the model components and the aforementioned articles showed ways in which the resulting $W(\cdot)$ mimics several of the data features described in Section 3. But the additive proposal results in a Gaussian random field, so this error process has Gaussian tails. On the other hand, as we will show below, the multiplicative proposal results in a

non–Gaussian random field with tails heavier than Gaussian, so this is the model we generalize and investigate further.

We now describe some of the probabilistic characteristics of the product of two independent random fields, and how these depend on the characteristics of the factors.

Result 1. Let $W_1(\cdot)$ and $W_2(\cdot)$ be independent random fields with mean functions $\mu_1(\mathbf{s})$ and $\mu_2(\mathbf{s})$, and covariance functions $C_1(\mathbf{s},\mathbf{u})$ and $C_2(\mathbf{s},\mathbf{u})$, respectively. Then, the mean and covariance functions of $W(\mathbf{s}) = W_1(\mathbf{s})W_2(\mathbf{s})$ are given by

$$\begin{aligned} E(W(\mathbf{s})) &= \mu_1(\mathbf{s})\mu_2(\mathbf{s}) \\ \operatorname{cov}(W(\mathbf{s}), W(\mathbf{u})) &= C_1(\mathbf{s},\mathbf{u})C_2(\mathbf{s},\mathbf{u}) + \mu_1(\mathbf{s})\mu_1(\mathbf{u})C_2(\mathbf{s},\mathbf{u}) + \mu_2(\mathbf{s})\mu_2(\mathbf{u})C_1(\mathbf{s},\mathbf{u}). \end{aligned}$$

The proof involves direct calculations. Note in particular that if both random fields $W_1(\cdot)$ and $W_2(\cdot)$ have mean 0, then $E(W(\mathbf{s})) = 0$ and $\operatorname{cov}(W(\mathbf{s}), W(\mathbf{u})) = C_1(\mathbf{s},\mathbf{u})C_2(\mathbf{s},\mathbf{u})$.

Result 2. Consider again the notation and conditions of Result 1, and assume $C_1(\mathbf{s},\mathbf{u})$ and $C_2(\mathbf{s},\mathbf{u})$ are both continuous along the 'diagonal' $\mathbf{s} = \mathbf{u}$. Then the random field $W(\mathbf{s}) = W_1(\mathbf{s})W_2(\mathbf{s})$ is mean square continuous.

The proof follows from a result in Palacios and Steel (2006).

Result 3. Consider again the notation and conditions of Result 1, and assume that $W_1(\mathbf{s})$ and $W_2(\mathbf{s})$ have symmetric distributions about $\mu_1(\mathbf{s})$ and $\mu_2(\mathbf{s})$. Then, the distribution of $W(\mathbf{s}) = W_1(\mathbf{s})W_2(\mathbf{s})$ is symmetric about $\mu_1(\mathbf{s})\mu_2(\mathbf{s})$ if and only if $\mu_1(\mathbf{s})\mu_2(\mathbf{s}) = 0$.

This was shown by Chen and Slud (1984). From this results follows that when $W_1(\cdot)$ and $W_1(\cdot)$ are Gaussian random fields, $W(\mathbf{s})$ has a symmetric distribution about 0, provided that either $\mu_1(\mathbf{s})$ or $\mu_2(\mathbf{s})$ is 0.

Result 4. Consider again the notation and conditions of Result 1, where in addition the random fields $W_1(\cdot)$ and $W_2(\cdot)$ are assumed Gaussian with mean 0 and variance 1. Then, $W(\mathbf{s}) = W_1(\mathbf{s})W_2(\mathbf{s})$ is a non–Gaussian random field whose distributions have tails heavier than those of normal distributions.

Proof. Let W_1, W_2 be independent random variables with standard normal distribution, and $W = W_1W_2$. It follows from Result 1 that $E(W) = E(W_1) = 0$ and $\operatorname{var}(W) = \operatorname{var}(W_1) = 1$. Also, from a result in Rohatgi (1976, p. 141) we have that the pdf of W is given by

$$f_W(w) = \int_{-\infty}^{\infty} \phi(x)\, \frac{1}{|x|}\phi\Big(\frac{w}{|x|}\Big)dx,$$

where $\phi(\cdot)$ is the pdf of the standard normal distribution, so the distribution of W is clearly non–Gaussian. Additionally, for any $t > 0$ we have

$$\begin{aligned} P(W > t) &= \int_t^\infty f_W(w)dw \\ &= \int_{-\infty}^\infty \left(\int_t^\infty \frac{1}{\sqrt{2\pi|x|^2}} e^{-\frac{w^2}{2|x|^2}} dw \right) \phi(x)dx \\ &= \int_{-\infty}^\infty \Big(1 - \Phi\Big(\frac{t}{|x|}\Big)\Big)\phi(x)dx \\ &= 1 - 2\int_0^\infty \Phi\Big(\frac{t}{x}\Big)\phi(x)dx, \end{aligned}$$

where $\Phi(\cdot)$ is the cdf of the standard normal distribution. The above expression lacks a closed–form, so we investigate its behaviour numerically. Figure 5 displays plots of $P(W > t)$ and $P(W_1 > t) = 1-\Phi(t)$ for $t > 0$ (solid and broken lines, respectively[2]). It holds that $P(W > t) > P(W_1 > t)$ for $t > 1.72$, so large positive values are more likely for W than for W_1. A similar result holds for large negative values because W and W_1 are symmetric about 0. To complement the above, Figure 6 displays histograms from $10,000$ realizations of W_1 (left) and W (right), where it is also clear that the tails of W are heavier than those of W_1.

5 A General Model

In this section we describe three variants of the approach described in the previous section to construct a spatial processes driven by a localized shock point source using two independent Gaussian random fields. The first generalizes the model proposed by Hughes–Oliver and Gonzalez–Farias (1999), while the second adapts to the multiplicative error process a technique used in Ecker and De Oliveira (2008) for the case of an additive error process. The third variant is new and results in a model with some features that differ from those of the first two variants.

Variant 1. Let $X_1(\mathbf{s})$ be a zero–mean Gaussian random field in $\mathbb{R}^2$ with covariance function

$$\mathrm{cov}(X_1(\mathbf{s}), X_1(\mathbf{u})) = \min\{s_1, u_1\} \cdot \min\{s_2, u_2\},$$

where $\mathbf{s} = (s_1, s_2)$, $\mathbf{u} = (u_1, u_2)$; this is an example of a Wiener process in $\mathbb{R}^2$. For a given strictly monotone function $g : \mathbb{R}^+ \to \mathbb{R}^+$, define the transformation $T_1 : \mathbb{R}^2 \to \mathbb{R}^2$ as

$$T_1(\mathbf{s}) = (g(d_\mathbf{s}), 1),$$

[2]The former computed by numerical quadrature.

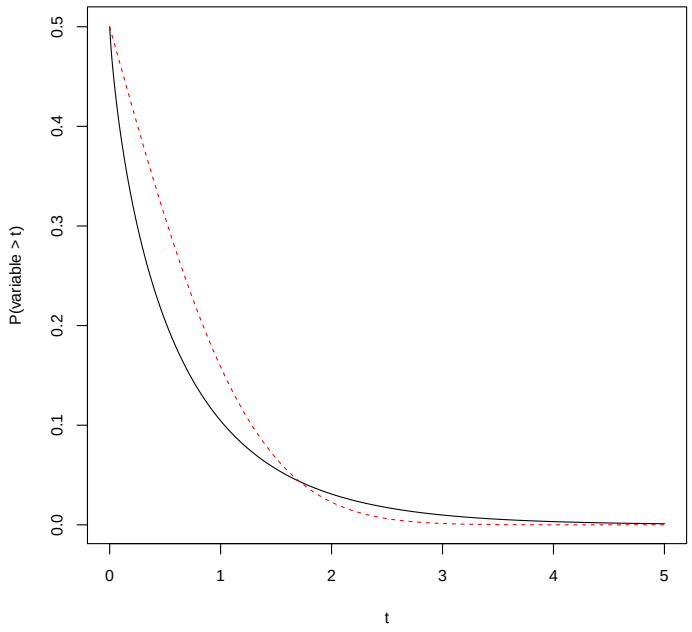

Figure 5: Plots of $P(W > t)$ (black solid line) and $P(W_1 > t)$ (red broken line).

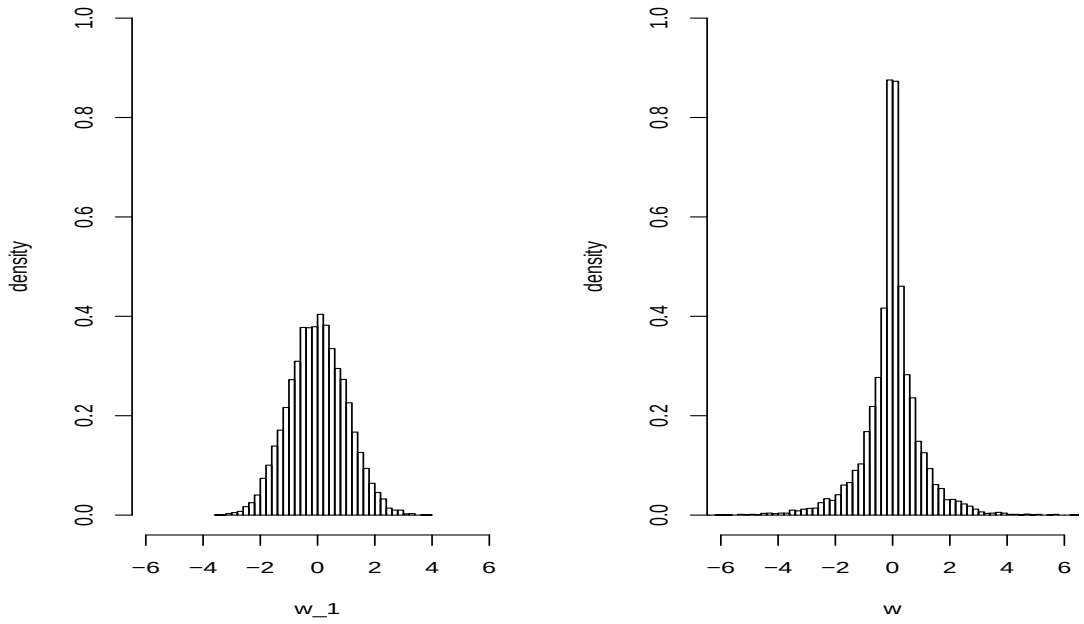

Figure 6: Histogram of $10,000$ realizations of W_1 (left) and W (right).

where $d_{\mathbf{s}} = ||\mathbf{s} - \mathbf{s}^*||$ is the Euclidean distance between $\mathbf{s}$ and the location of the point source. Now define the random field $W_2(\cdot)$ by

$$W_2(\mathbf{s}) = X_1(T_1(\mathbf{s})) = X_1((g(d_{\mathbf{s}}), 1)), \qquad \mathbf{s} \in \mathbb{R}^2. \tag{1}$$

This is a zero–mean Gaussian random field with covariance function

$$\begin{aligned} \mathrm{cov}(W_2(\mathbf{s}), W_2(\mathbf{u})) &= \min\{g(d_{\mathbf{s}}), g(d_{\mathbf{u}})\} \\ &= \begin{cases} g(\min\{g(d_{\mathbf{s}}), g(d_{\mathbf{u}})\}) & \text{if } g \text{ is increasing} \\ g(\max\{g(d_{\mathbf{s}}), g(d_{\mathbf{u}})\}) & \text{if } g \text{ is decreasing} \end{cases}, \end{aligned} \tag{2}$$

with the property that the realizations of this random field take the same value for all locations with the same distance to the point source $\mathbf{s}^*$.

We now generalize the model proposed by Hughes–Oliver and Gonzalez–Farias (1999). Let $W_1(\cdot)$ and $W_2(\cdot)$ be two independent Gaussian random fields, where $W_1(\cdot)$ is stationary with mean 0 and covariance function $\sigma_1^2 K_1\big(||\mathbf{s}-\mathbf{u}||\big)$, with $K_1(\cdot)$ a stationary and continuous correlation function in $\mathbb{R}^2$, and $W_2(\cdot)$ is the random field defined in (1). Then, the error spatial processes driven by a localized shock point source is defined as

$$W(\mathbf{s}) = W_1(\cdot)W_2(\cdot). \tag{3}$$

By Results 1 and 4, this is a *non-Gaussian* random field with mean zero and covariance function

$$\mathrm{cov}(W(\mathbf{s}), W(\mathbf{u})) = \sigma_1^2 K_1\big(||\mathbf{s} - \mathbf{u}||\big) \cdot \min\{g(d_{\mathbf{s}}), g(d_{\mathbf{u}})\}, \qquad \mathbf{s}, \mathbf{u} \in \mathbb{R}^2. \tag{4}$$

In particular, we have that $\mathrm{var}(W(\mathbf{s})) = \sigma_1^2 g(d_{\mathbf{s}})$, so the function $g(\cdot)$ controls how the variance of the process varies with distance to the point source. Possible variance function include $\exp((\gamma_1 + t)^{\gamma_2})$, $\gamma_1 + \exp(-\gamma_2 t)$ and $\gamma_1(1 + \exp(-\gamma_2 t))$, with γ_1, γ_2 unknown parameters; these were used, respectively, in the data analyses by Hughes–Oliver and Gonzalez–Farias (1999), Ecker and De Oliveira (2008) and Ecker, De Oliveira and Isakson (2013). Also, these previous works have used members of the power exponential family to model $K_1(\cdot)$. In addition,

$$\mathrm{corr}(W(\mathbf{s}), W(\mathbf{u})) = K_1\big(||\mathbf{s} - \mathbf{u}||\big) \cdot \Big(\frac{\min\{g(d_{\mathbf{s}}), g(d_{\mathbf{u}})\}}{\max\{g(d_{\mathbf{s}}), g(d_{\mathbf{u}})\}}\Big)^{1/2},$$

so the correlation function of $W(\cdot)$ is the correlation function of $W_1(\cdot)$ modulated by the ratio $\big(g(d_{\mathbf{s}})/g(d_{\mathbf{u}})\big)^{1/2}$. A possible drawback of this variant is that a single function, $g(\cdot)$ in this case,

controls both the process variance and the correlation modulation. The next approach seeks to avoid this potential limitation.

Variant 2. Let $W_1(\cdot)$ be a Gaussian random field with mean zero and covariance function given by $(g(d_{\mathbf{s}})g(d_{\mathbf{u}}))^{1/2}K_1\big(||\mathbf{s}-\mathbf{u}||\big)$, where $g:\mathbb{R}^+\to\mathbb{R}^+$ is a strictly monotone function, and $K_1(\cdot)$ is a stationary and continuous correlation function in $\mathbb{R}^2$. Also, let now $X_2(\cdot)$ be a Gaussian process on the real line that is independent of $W_1(\cdot)$, with mean 0, variance 1 and correlation function $\text{corr}(X_2(t_1),X_2(t_2))=K_2(|t_1-t_2|)$, where $K_2(\cdot)$ is a stationary and continuous correlation function on $\mathbb{R}^1$. Similarly as in variant 1, consider the transformation $T_2:\mathbb{R}^2\to\mathbb{R}^+$ defined by $T_2(\mathbf{s})=d_{\mathbf{s}}$, and the random field in the plane defined as

$$W_2(\mathbf{s})=X_2(T_2(\mathbf{s}))=X_2(d_{\mathbf{s}}),\qquad \mathbf{s}\in\mathbb{R}^2,$$

which in this case is a zero–mean Gaussian process with covariance function $K_2(|d_{\mathbf{s}}-d_{\mathbf{u}}|)$. Possible forms for $g(\cdot)$ and $K_1(\cdot)$ are the same as those for variant 1. Then, the product error process (3) is a *non-Gaussian* random field with mean zero and covariance function

$$\text{cov}(W(\mathbf{s}),W(\mathbf{u}))=(g(d_{\mathbf{s}})g(d_{\mathbf{u}}))^{1/2}K_1(||\mathbf{s}-\mathbf{u}||)\cdot K_2(|d_{\mathbf{s}}-d_{\mathbf{u}}|).$$

Like variant 1, it holds in this case that $g(\cdot)$ controls how the variance of $W(\cdot)$ varies with distance to the point source, but unlike variant 1, the correlation modulation is now controlled by a separate term, $K_2(|d_{\mathbf{s}}-d_{\mathbf{u}}|)$ in this case.

Variant 3. Let $W_1(\cdot)$ be a Gaussian random field with mean zero and covariance function given by $\sigma^2K_1\big(||\mathbf{s}-\mathbf{u}||\big)$, with $K_1(\cdot)$ a stationary and continuous correlation function in $\mathbb{R}^2$, and $W_2(\cdot)$ a Gaussian random field independent of $W_1(\cdot)$ with mean $(g(d_{\mathbf{s}}))^{1/2}$ and covariance function $\frac{\tau^2}{\sigma^2}\mathbf{1}\{\mathbf{s}=\mathbf{u}\}$, where $g:\mathbb{R}^+\to\mathbb{R}^+$ is a strictly monotone function and $\mathbf{1}\{A\}$ is the indicator function of the event A. Possible forms for $g(\cdot)$ and $K_1(\cdot)$ are the same as those in variant 1. Again from Results 1 and 4 follow that the product error process (3) is a *non-Gaussian* random field with mean zero and covariance function given by

$$\text{cov}(W(\mathbf{s}),W(\mathbf{u}))=\sigma^2\big(g(d_{\mathbf{s}})g(d_{\mathbf{u}})\big)^{1/2}K_1\big(||\mathbf{s}-\mathbf{u}||\big)+\tau^2\mathbf{1}\{\mathbf{s}=\mathbf{u}\}.$$

Like variants 1 and 2, it holds in this case that $g(\cdot)$ controls how the variance of $W(\cdot)$ varies with distance to the point source. And like variant 1, $g(\cdot)$ affects the correlation modulation, but now the

magnitude of this modulation depends also on τ^2/σ^2. In addition, note that unlike variants 1 and 2, it holds that the above covariance function is discontinuous along the 'diagonal' $\mathbf{s} = \mathbf{u}$, so it displays the so–called nugget effect.

6 Exploratory Diagnostic

The usual diagnostics for stationary/isotropic covariance functions are not applicable to assess adequacy of the zero–mean error processes described in the previous section, given their non–stationary nature. In this section we adapt one of the traditional diagnostics to make it applicable to identify the very particular form of non–stationarity in the error process (3) described in variant 1. Slight modifications of the proposed diagnostic would make them applicable to assess the adequacy of the error processes constructed using variants 2 and 3.

The variogram of the error process (3) is given by

$$\begin{aligned} \mathrm{var}(W(\mathbf{s}) - W(\mathbf{u})) &= \mathrm{var}(W(\mathbf{s})) + \mathrm{var}(W(\mathbf{u})) - 2\,\mathrm{cov}(W(\mathbf{s}), W(\mathbf{u})) \\ &= \sigma_1^2\big(g(d_{\mathbf{s}}) + g(d_{\mathbf{u}}) - 2K_1\big(||\mathbf{s}-\mathbf{u}||\big) \cdot \min\{g(d_{\mathbf{s}}), g(d_{\mathbf{u}})\}\big), \end{aligned}$$

so for any two locations with the same distance to the point source (so $d_{\mathbf{s}} = d_{\mathbf{u}}$) it holds that

$$\mathrm{var}(W(\mathbf{s}) - W(\mathbf{u})) = 2\sigma_1^2 g(d_{\mathbf{s}})(1 - K_1(||\mathbf{s}-\mathbf{u}||)).$$

Now define the scaled process $\tilde{W}(\mathbf{s}) := W(\mathbf{s})/(\mathrm{var}(W(\mathbf{s})))^{1/2}$, $\mathbf{s} \in D$. Its semivariogram function is given by

$$\begin{aligned} \frac{1}{2}E\big((\tilde{W}(\mathbf{s}) - \tilde{W}(\mathbf{u}))^2\big) &= \frac{\frac{1}{2}E\big((W(\mathbf{s}) - W(\mathbf{u}))^2\big)}{\mathrm{var}(W(\mathbf{s}))} \\ &= 1 - K_1(||\mathbf{s}-\mathbf{u}||), \qquad \text{when } d_{\mathbf{s}} = d_{\mathbf{u}}, \end{aligned}$$

where the latter is an isotropic and continuous semivariogram with 'sill' equal to 1. Based on this observation, we suggest the following diagnostic for the adequacy of the a covariance function (4):

- Estimate the variance function $g(d_{\mathbf{s}})$ by parametric or non-parametric regression, say by $\hat{g}(d_{\mathbf{s}})$.
- Assuming $\hat{g}(d_{\mathbf{s}})$ known, compute $\hat{W}(\mathbf{s}_i) = W(\mathbf{s}_i)/(\hat{g}(d_{\mathbf{s}_i}))^{1/2}$ for each of the sampling locations $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n$, where $W(\mathbf{s}_i) = Y(\mathbf{s}_i) - \hat{\mu}(\mathbf{s}_i)$ are the residuals obtained after fitting the mean function.
- Partition the interval $(0, \max\{||s_i - s_j||\})$ into distance bins $0 < l_1 < l_2 < \ldots < l_R$, and partition the study region as $D = D_1 \cup D_2 \cup \ldots \cup D_K$, where the D_js are all annulus centered at $\mathbf{s}^*$ with increasing radii.
- For each distance l_r estimate the semivariogram of $\hat{Z}(\cdot)$ for two locations l_r units apart by

$$\hat{\gamma}(l_r) = \frac{1}{2|N_r|} \sum_{(\mathbf{s}_i, \mathbf{s}_j) \in N_r} (\hat{Z}(\mathbf{s}_i) - \hat{Z}(\mathbf{s}_j))^2,$$

  where

$$N_r = \{(\mathbf{s}_i, \mathbf{s}_j) : ||\mathbf{s}_i - \mathbf{s}_j|| \approx l_r \text{ and } \mathbf{s}_i, \mathbf{s}_j \in D_j \text{ for some } j\},$$

  and $|N_r|$ is the number of pairs in N_r.

The diagnostic consists of plotting $\hat{\gamma}(l_r)$ versus l_r. If the covariance model (4) with model components $g(\cdot)$ and $K_1(\cdot)$ is adequate for a given data set, then the semivariogram estimates $\hat{\gamma}(l_r)$ should have the typical behavior of rising with increasing distance, and then leveling off to a value close to 1.

7 Conclusions

This work proposed a strategy to construct random fields with non–constant variance functions and distributions with tails heavier than Gaussian, with the motivation of mimicking the data features of house sale prices revealed in Section 3. The strategy generalizes the multiplicative model proposed by Hughes–Oliver and Gonzalez–Farias (1999) which is a non–stationary and non-Gaussian random field. It holds that the likelihood of the parameters of such models lack a closed–form expression, so likelihood–based inference is challenging. Because of this, Hughes–Oliver and Gonzalez–Farias (1999) fitted the model using a *surrogate* Gaussian likelihood determined by the true mean and covariance functions derived from the model. This is inefficient at best and inappropriate at worst, since such fitting does not take into account the heavy tail nature of the distribution of the data. This is likely

to result in a fitted model that does not predict well extreme selling prices, namely, very expensive or very inexpensive houses, since these are viewed as 'outliers' under a Gaussian model.

In the future we plan to investigate the use of the EM algorithm and/or Bayesian data augmentation to fit the model using the true likelihood under the proposed model. More specifically, we plan to investigate a possible adaptation of the Markov chain Monte Carlo algorithm proposed by Palacios and Steel (2006) to fit the so–called Gaussian–log–Gaussian random field obtained by scale–mixing of a Gaussian random filed, where the scaling process is a log–Gaussian random field.

Finally, we also proposed a graphical diagnostic to assess the adequacy of the proposed model to a given dataset for one of the model variants that we proposed. We plan to derive similar graphical diagnostics for the other two model variants, as well as to investigate the effectiveness of these diagnostics using simulated and real data

8 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.

Chen, R.W. and Slud, E.V. (1984). On the Product of Symmetric Random Variables. *Communications in Statistics–Theory and Methods*, 13, 611-615.

Cressie, N. (1993). *Statistics for Spatial Data*. Wiley.

Diggle, P.J. and Ribeiro, P.J. (2007). *Model-based Geostatistics*. Springer.

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

Ecker, M.D. (2003). Geostatistics: Past, Present and Future. *Encyclopedia of Life Support Systems (EOLSS)*. Developed under the Auspices of the UNESCO, Eolss Publishers, Oxford, UK. (www.eolss.net)

Ecker, M.D., De Oliveira, V. and Isakson, H. (2013). A Note on Non-Stationary Point Source Spatial Model. *Environmental and Ecological Statistics*, 20, 59-67.

Ecker, M.D. and De Oliveira, V. (2008). Bayesian Spatial Modeling of Housing Prices Subject to a Localized Externality. *Communications in Statistics–Theory and Methods*, 37, 2066-2078.

Fouedjio, F. (2016). Second-Order Non-Stationary Modeling Approached for Univariate Geostatistical Data. *Stochastic Environmental Research and Risk Assessment*, to appear.

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.

Isakson, H. and Ecker, M.D. (2008). An Analysis of the Impact of Swine CAFOs on the Value of Nearby Houses. *Agricultural Economics*. 39, 1-8.

Malpezzi, S. (2003). Hedonic Pricing Models: A Selective and Applied Review. In: *Housing Economics and Public Policy*, T. O'Sullivan and K. Gibb (eds.), Wiley-Blackwell, pp 67-89.

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.

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.

Palacios, M.B. and Steel, M.J.S. (2006). Non-Gaussian Bayesian Geostatistical Modeling. *Journal of the American Statistical Association*, 101, 604-618.

Rohatgi, V.K. (1976). *An Introduction to Probability Theory and Mathematical Statistics*. Wiley.

Sampson, P.D. (2010). Constructions for Nonstationary Spatial Processes. In: *Handbook of Spatial Statistics*, A.E. Gelfand, P.J. Diggle, M. Fuentes and P. Guttorp (eds.), CRC/Press, pp 119-130.

Treviño, G. (1992). An Heuristic Overview of Non-Stationarity. In: *Non-Stationary Stochastic Processes and Their Applications*, A.G. Miamee (ed.), World Scientific Publishers, pp 48-61.