Spatial hedonic modeling adjusted for preferential sampling

Lucia Paci

Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Milan, Italy.

E-mail: lucia.paci@unicatt.it

Alan E. Gelfand

Department of Statistical Science, Duke University, NC, USA.

María Asunción Beamonte

Facultad de Economía y Empresa, Universidad de Zaragoza, Spain.

Pilar Gargallo

Facultad de Economía y Empresa, Universidad de Zaragoza, Spain.

Manuel Salvador

Facultad de Economía y Empresa, Universidad de Zaragoza, Spain

Abstract.
Hedonic models are widely used to predict selling prices of properties. Originally, they were proposed as simple spatial regressions, i.e., a spatially referenced response regressed on spatially referenced predictors. Subsequently, spatial random effects were introduced to serve as surrogates for unmeasured or unobservable predictors and were shown to provide better out-of-sample prediction. However, what has been ignored in the literature is the fact that the locations (and times) of the sales are random and, in fact, are an observation of a random point pattern. Here, we first consider whether there is stochastic dependence between the point pattern of locations and the set of responses? If so, a second question is whether incorporating a log intensity for the point pattern of locations into the hedonic modeling enables improvement in the prediction of selling price? We connect this problem to what is referred to as preferential sampling. Through model comparison we illuminate the role of the point pattern data in the prediction of selling price. Using two different years of property sales from Zaragoza, Spain, we employ both the full database as well as an intentionally biased subset to elaborate this story.

Keywords: Bayesian inference; log-Gaussian Cox process; Markov chain Monte Carlo; nearest neighbor Gaussian process; real estate transactions; shared process models

1. Introduction

Hedonic modeling refers to the well-established objective of specifying regression models to explain selling prices of properties. There is a rich literature on such modeling which we review below. With residential properties, these models typically employ regressors capturing features of the property as well as features of the neighborhood where the property is located. Furthermore, recognizing the old maxim regarding property sales - “location, location, location” - spatial random effects have been introduced. Typically, such modeling is dynamic, with property sales aggregated to say quarters or years in order to see the temporal evolution of the sales market. These random effects are customarily modeled using Gaussian processes and can be viewed as surrogates for unmeasured or unobservable predictors carrying spatial and/or temporal information. Regardless, the random effects provide local adjustment to the model and the flexibility of the Gaussian processes results in improved prediction of selling prices.

A key feature ignored in such hedonic modeling is the fact that the locations and times of dwelling transactions are random. That is, selling price is modeled conditionally on location and time; randomness is introduced only in the regression given predictors at the location and time. Additionally, when random effects are introduced, they are also associated with locations and times that are viewed as fixed. In this regard, Paci et al. (2017) recognized that, for a given market and a given period, the collection of property sales comprises a realization of a random point pattern over space and time. That is, the number of transactions is random and, given this number, the locations and times of the transactions are random.

Investigation of selling price of properties in the context of a point pattern of property sales takes us to what has been referred to as the preferential sampling setting. Preferential sampling is a relatively recently recognized issue (Diggle et al., 2010) which we review below. The basic idea is as follows. In attempting to learn about a spatial surface over a region using data obtained from a finite set of locations, if the sampling locations arise in a "biased" fashion, such bias can impair our ability to accurately interpolate the surface. A standard setting is geostatistical modeling (see e.g. Cressie and Wikle 2011; Banerjee et al. 2014). Illustratively, consider the objective of inferring about a surface of environmental exposures. If environmental monitoring networks introduce design bias by tending to place monitors in locations where environmental levels are expected to be high, then interpolation based upon observations from these stations will necessarily produce high predictions.

For environmental exposure, we may have a designed set of sampling locations which, in turn, may introduce preferential sampling. The obvious remedy lies in spatial design of the locations, e.g., a random or space-filling design (Nychka and Saltzman, 1998; Pronzato and Müller, 2012) for locations over the region of interest will preclude such bias. However, in many settings, design is not possible. Consider our context; sites where responses (selling prices) arise occur at random, both in number and in location, i.e., again, as a realization of a spatial (or space-time) point pattern. To clarify, in environmental exposure applications with fixed monitoring networks, sampling locations are really not random but are viewed as a point pattern in order to reveal their bias through an associated intensity. Conversely, we have a truly random point pattern of property transactions which may include bias that was not designed. Then, the question becomes a stochastic one: is the realization of the locations independent of the realization of the selling prices? If not, then we have what is called preferential sampling.

With hedonic modeling, fixing a time window, our data consists of a spatial point pattern of sales of properties and, associated with each transaction, we have a selling price. We also have an available collection of predictors which can explain both sales activity as well as selling prices. Then, adjusted for covariate information, the question is whether variation in sales activity over a region, captured through an intensity, is related to variation in selling price over the region, captured through a hedonic model?

This need not be the case, say if the mix of higher and lower priced property transactions is roughly the same in subregions of high activity as in subregions of low activity. However, it can be the case that subregions of high activity are associated primarily with higher selling prices or primarily with lower selling prices. Analogous to the environmental example above, this corresponds to over-sampling of subregions where selling prices tended to be high or over-sampling of subregions where prices tended to be low, respectively. Our dataset provides a full census of property transactions. Now, if preferential sampling arises (in fact, it does as we show in Section 4), then it will be attributable to economic forces which provide bias rather than the foregoing design bias. These economic forces may be observable as, for instance, after the deep financial crisis that began in 2008. Banks had to sell off many properties (so-called toxic assets) because of mortgage defaults, at very low prices. However, these forces, e.g., interest rate, unemployment rate and consumer price indexes do not vary over the city so they cannot be used for studying the spatial variation of selling price. *Regional* economic forces are typically unobserved or unmeasured but can reveal a preferential sampling effect, as in our ensuing analysis. It is also the case that other local covariate information could mitigate possible preferential sampling issues. Examples include information, e.g., about amenities available in neighborhoods, about the so-called “built” environment associated with neighborhoods, about access to transportation or to health care in neighborhoods. Such covariates are difficult to obtain and beyond the scope of our study.

Continuing our example, suppose that for regions in space where there is lower selling price, we tend to see higher sales activity. Two potential scenarios of this sort are: (i) the case of a high rate of foreclosures or mortgage defaults in portions of the city which would lead to consequential under-pricing in property sales or (ii) a large governmental development of subsidized housing in portions of the city, again leading to under-pricing. If there is over-sampling of such regions, then we may tend to see underpricing in prediction of selling price. This would result in the presence of a preferential sampling effect. Furthermore, if we view the log intensity for the point pattern as a regressor, a negative coefficient in a hedonic model can help to push up prediction where selling prices tend be lower but also pull down prediction where they tend to be higher. Similarly, if there is over-sampling of higher priced properties, now a positive coefficient attached to the intensity can help to explain selling price.

So, two general questions arise here. Is there bias in the set of property transaction locations which affects inference with regard to selling prices of properties? If there is such bias, can we revise our hedonic model, using the spatial behavior of transaction activity, to improve the prediction of selling price? Our contribution is to illuminate the stochastic connection between the point pattern of sales locations and the associated selling prices†. We can illustrate this issue with both the full database and an intentionally biased subset of a collection of transactions discussed in Section 2. There is evident value to buyers, seller, realtors, banks, etc. in achieving better understanding of the nature of this relationship, and how it affects inference regarding selling price. However, to date, such connection has not been studied in the literature. Our objective is *not* to remove preferential sampling bias. Rather, we seek to identify whether it is present and attempt to provide a better spatial prediction model if it is.

We elaborate the above through three stories. First, working with the full database, under a simple hedonic model, that is, a spatial regression specification with no random effects, inclusion of the log intensity of the point pattern of sales locations as a regressor yields a significant regression coefficient. There is a preferential sampling effect which has a clear interpretation and which, again, we attribute to unmeasured economic forces. Second, with this database,

†Stochastic dependence is different from functional dependence. The latter merely reflects the fact that a selling price is attached to a location.

if we adopt a hedonic model with spatial random effects (a so-called geostatistical model) and then introduce the log intensity of the point pattern of sales locations as a regressor, we find no preferential sampling effect, no improvement in prediction of selling price. This may not be surprising due to the flexibility provided by the spatial random effects. Moreover, when we have the complete inventory of real estate transactions, there is no *design*‡ to bias the sampling. So, in the absence of a very strong economic effect, there is no reason to expect a preferential sampling effect under this model. As a result, our third story finds us introducing consequential sampling bias into our sample of transaction locations. Then, we are able to demonstrate a preferential sampling effect in the hedonic model with spatial random effects and to improve prediction taking this effect into account.

Each story emerges from a particular specification of the joint model for selling prices and locations of sales, applied to a set (or subset) of residential property transactions in Zaragoza, Spain. The hedonic models also allow for the inclusion of Gaussian processes (GPs) to capture local adjustment. Similarly, the spatial point pattern modeling of sales activity includes a Gaussian process in the intensity, yielding a log Gaussian Cox process (Møller et al., 1998), capturing more detail in the point pattern intensity. Through the model specifications we can learn about the presence of preferential sampling; through out-of-sample model comparison, we can evaluate whether a model incorporating preferential sampling better predicts selling price.

In brief, we can employ a nonhomogeneous Poisson process (NHPP) model or a log Gaussian Cox process (LGCP) model for the intensity surface of the point pattern of sales locations. In fact, Paci et al. (2017) demonstrated that the LGCP is preferred to the NHPP so, in the sequel, we only use the former. For the hedonic model, we consider four choices: (i) a spatial regression, (ii) a spatial regression incorporating preferential sampling, (iii) a geostatistical hedonic model, and (iv) a geostatistical model incorporating preferential sampling. In terms of the *stories* above, comparison of (i) vs. (ii) reveals a preferential sampling effect for the full dataset. Comparison of (ii) vs. (iii) reveals that the geostatistical model slightly outperforms the preferential sampling model for the full dataset. Comparison of (iii) with (iv) reveals that adding preferential sampling to the geostatistical model does not improve its performance with the full set of transactions. Then, introducing intentional bias into the set of transactions in the form of over-sampling of expensive properties, we can make the same comparisons; now comparison of (iii) with (iv) reveals improved predictive performance with a biased subset of transactions.

As mentioned above, a hedonic pricing model is customarily employed to measure the contribution of individual house characteristics to the overall composite value of the housing asset. Location along with other house characteristics are typically introduced into the mean structure in the hedonic model resulting in a spatial regression (Ridker and Henning, 1967; Li and Brown, 1980; Dubin and Sung, 1990; Anselin and Lozano-Gracia, 2009; Ahlfeldt and Kavetsos, 2014). However, sometimes important neighborhood characteristics are unavailable; in these cases, spatial effects may be introduced into the error structure, the so-called geostatistical approach (Dubin, 1988; Basu and Thibodeau, 1998; Gelfand et al., 2003), or as random effects using autoregressive specifications (Gelfand et al., 1998; Beamonte et al., 2010).

Preferential sampling was introduced by Diggle et al. (2010) to account for stochastic dependence between the point pattern of locations and sample of response data observed at these locations through the use of a *shared* process specification. The authors proposed a class of

‡Here, design refers to intentionally over-sampling higher priced properties or to over-sampling lower priced properties.

stochastic models for preferentially sampled geostatistical data with likelihood-based inference. A richer formulation in a Bayesian framework was proposed by Pati et al. (2011). Gelfand et al. (2012) discussed the differences in spatial prediction using randomly selected locations versus preferentially chosen locations and showed how the latter provide biases in prediction. Ferreira and Gamerman (2015) analyzed the effects of preferential sampling in the process of obtaining an optimal design in geostatistical models. Lee et al. (2015) studied the impact of preferential sampling on the inference of health effects in the air pollution epidemiology. As mentioned, many studies of preferential sampling in the literature are in the context of air pollution exposure assessment (Pati et al., 2011; Lee et al., 2011; Shaddick and Zidek, 2014). Other examples arise in veterinary parasitology, where the location of the sampled farms is not independent of the spatial distribution of parasites (Cecconi et al., 2016) and in studies of population distributions (Conn et al., 2017). However, to our knowledge, we are the first to investigate preferential sampling in the context of real estate markets.

Our motivating dataset consists of a complete census of residential property sales in Zaragoza, Spain during 2006-2014. We work with sales of apartments in a region around the city center. For each sale we have the geo-coded location and associated selling price. In addition, we have characteristics of the individual apartments and have constructed a file of neighborhood characteristics to assign to each apartment. We work with two years, one pre- and one post- the deep financial crisis that started in 2008.

The remainder of the paper is organized as follows. In Section 2 we describe the dataset. In Section 3 we first discuss some basic ideas behind the preferential sampling. Then, specific modeling choices are detailed as well as the use of the nearest neighbor Gaussian process to facilitate model fitting. In Section 4 we analyze the selling prices of the full set of transactions in Zaragoza. Section 5 illustrates the benefit of adjusting for preferential sampling when we have a biased subset of the residential sales data. Finally Section 6 provides discussion and indications for future work.

2. The residential sales data

We consider residential sales in Zaragoza, Spain, during the 9 year period $2006-2014$. Zaragoza is a city situated in the northeast part of Spain with a population of $664,953$ inhabitants as of 2014. The dataset has been provided by the *Colegio Nacional de Registradores de España* and contains, for every sale, information on location (UTM coordinates), selling price, living area, year of building construction and the exact day of the transaction.

The area of the municipality is large and for urban designs/plans the town hall divides the city into 89 urban polygons. Some of them, considered rural neighborhoods/areas, are distant from the city center and, in some cases, are disconnected from the rest of the city. For this study, we have considered 52 of those polygons, all of them around the city center, following the dense distribution of population in old cities in Spain. We have discarded properties other than apartments, the dominant type of dwelling everywhere in the city. Family houses, townhouses, etc. are located only in a few polygons. Large areas with no housing within the polygons have been removed (e.g. parks, the University of Zaragoza campus, hospital areas, etc.). What is retained is, essentially, a full census of apartment sale transactions in and around the city center.

Figure 1 shows the average price per year and the number of transactions per year. For the period analyzed in this work, it is important to keep in mind the deep economic financial

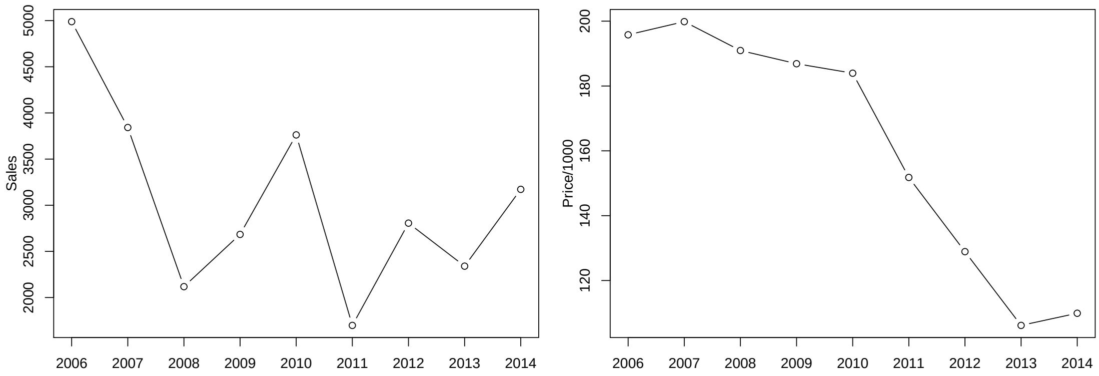

Figure 1. Number of transactions (left panel) and average price (right panel) per year.

crisis that, starting in 2008, sank the economy in Spain and many other countries, leading to devastating consequences especially in the real estate market. A clear drop in the prices is shown starting from 2010. Paci et al. (2017) analyzed the spatio-temporal point pattern of sales over the 9 year period. In this work, we bring in selling price with interest, as described above, in assessing preferential sampling with regard to selling price. Therefore, we focus on relating the spatial point pattern of sales to the spatial surface of selling price over the city. So, we abandon temporal modeling, instead analyzing selling prices for two different years, 2007 and 2012, pre- and post- crisis, respectively. Figure 2 shows the number of sales and the average price by urban polygons in 2007 (top panels) and 2012 (bottom panels); we have $2,969$ sales in 2007 and $2,186$ in 2012§.

Features of the sold properties are available to explain selling price, including the size of the property, its age and a binary regressor which records the presence of *accessory elements* (AE). Accessory elements are considered as complements to the apartment such as a garage space or a storage area. Specifically, $AE = 1$ if *no* accessories, $AE = 0$ if there are. Paci et al. (2017) partitioned all of the transactions into four size classes. Here, we consider all the sales as our point pattern and employ log size as a covariate in the hedonic model. Figure 3 displays the scatterplot of log size and log price for 2007 (left panel) and 2012 (right panel), showing the expected strong relationship. We also consider the age of the property defined as the difference between the year of transaction and the year of building construction. Then, a logarithmic transformation is applied and the scaled variable is retained. The weak relationship between log price and log age is displayed in Figure 4. Usually, economic variable are also employed in hedonic modeling, such as interest rates, unemployment rates and consumer price indexes. However, such variables do not vary over the city and so they cannot be used for studying the spatial variation of selling price.

Finally, we introduce demographic covariates at spatial level (polygon), i.e., total population, percentage of youth population (≤ 14 years old), elderly population (≥ 65 years old) and the percentage of foreign population. These proposed covariates are discussed in Gonzalez and Ortega (2013) and Eichholtz and Lindenthal (2014). They were also employed in the purely

§While we see spatial pattern in the selling prices, it is important to note that they have not yet been adjusted to reflect differences in the regressors associated with each sale.

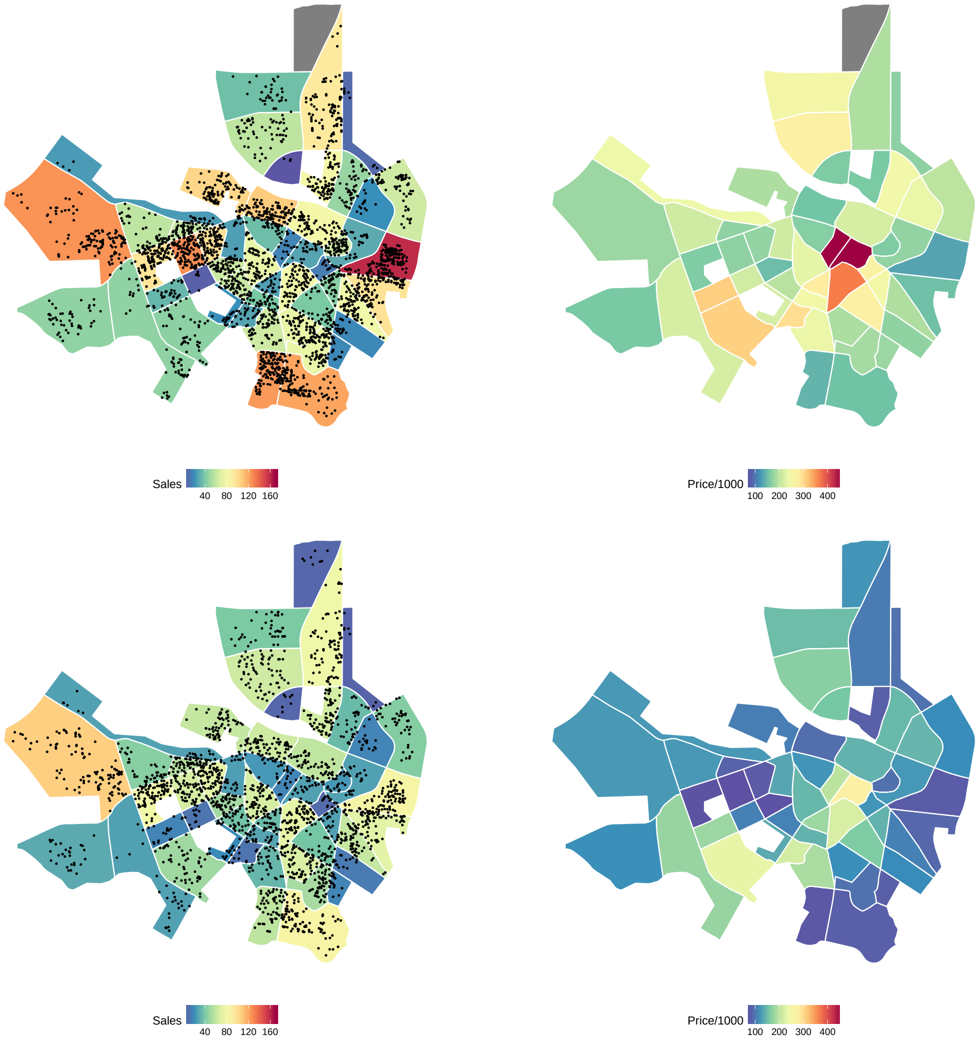

Figure 2. Number of sales (left panels) and average price (right panels) by urban polygons in 2007 (top panels) and 2012 (bottom panels). Grey color denotes no sales.

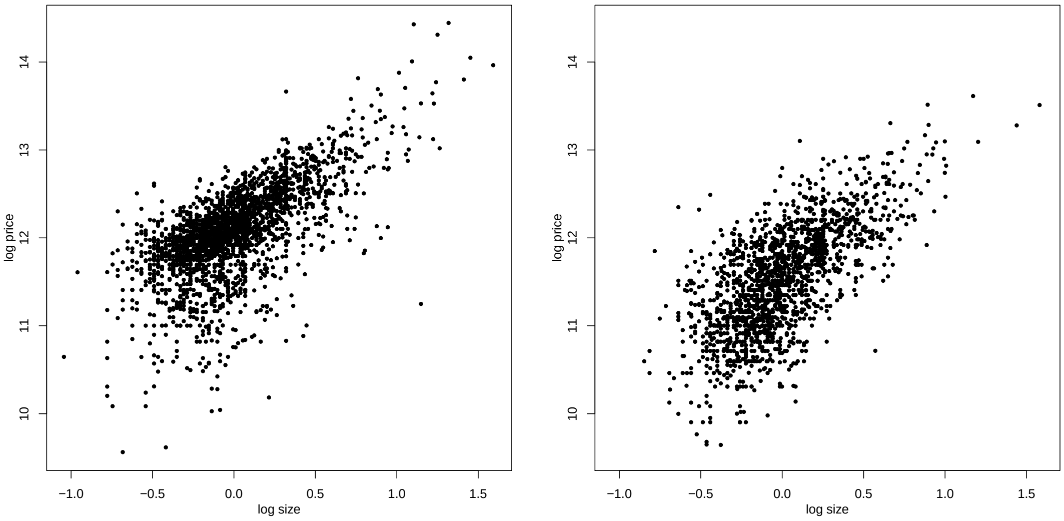

Figure 3. Scatterplot of log size and log price for 2007 (left panel) and 2012 (right panel).

point pattern analysis presented in Paci et al. (2017).

3. Modeling

3.1. Basic ideas

We begin with some basic modeling ideas in our setting. Let $\mathcal{Y}$ denote the set of selling prices and let $\mathcal{S}$ denote the spatial point pattern of sales. In Paci et al. (2017), $\mathcal{S}$ is modeled using a LGCP driven by an intensity $\lambda(\mathbf{s})$, i.e., $[\mathcal{S}|\lambda(\mathbf{s})], \mathbf{s} \in D$, where D is the spatial domain. Note that, if we simply overlay on this model a hedonic model for $\mathcal{Y}$, e.g. $[\mathcal{Y}|\mathcal{S}, \text{parameters}]$, then we can fit both models separately. $\mathcal{S}$ is viewed as fixed in the hedonic model; hence, $\mathcal{S}$ does not inform about $\mathcal{Y}$ and we do not introduce the notion of preferential sampling.

When the locations are viewed as random along with the responses at the locations, we have a marked point process. There are many ways to model marked point processes. The preferential sampling framework sets out the joint model for the locations and marks through a model for locations and then a model for marks given locations. Specifically, Diggle et al. (2010) suggest thinking in terms of three processes, i.e., the "measurement" process from which we sample $\mathcal{Y}$, the "point" process which yields $\mathcal{S}$, and the "field" process, say Ω which drives the measurement and point processes. At a high level, we can formalize the modeling through $[\mathcal{Y}|\mathcal{S}, \Omega, \text{parameters}][\mathcal{S}|\Omega, \text{parameters}][\Omega|\text{parameters}]$¶. According to Diggle et al. (2010), preferential sampling involves stochastic dependence, as opposed to functional dependence, between the point process and the measurement process. Hence, the field process adds the option of introducing a latent spatial random field driving both the measurement and the point processes, that is, a "shared" component model (Cecconi et al., 2016). In other words, we are asking whether there is a common latent spatial process driving dependence in selling price as a function of

¶Often, the set of locations may not have been developed randomly. However, in the context here, it is viewed as if it were a realization of a spatial point process.

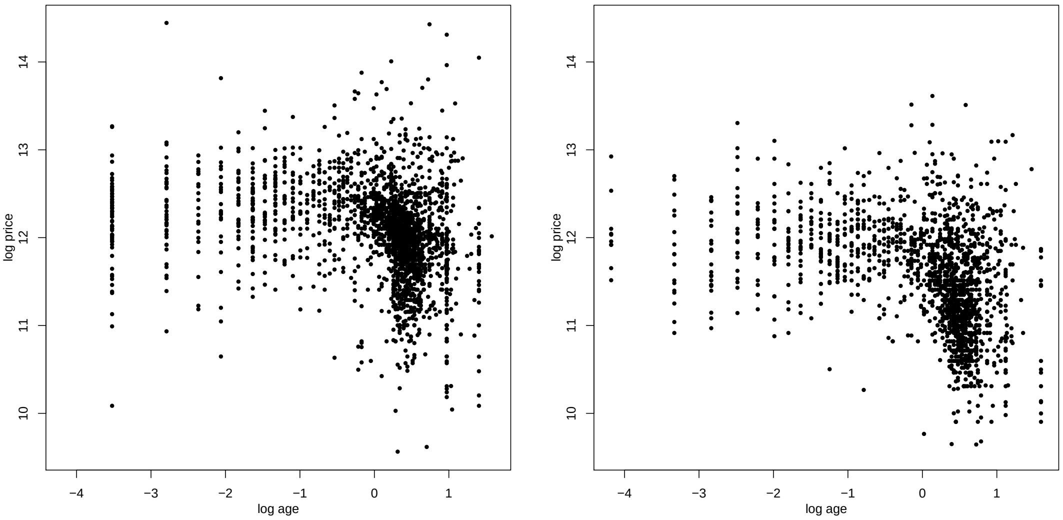

Figure 4. Scatterplot of log age and log price for 2007 (left panel) and 2012 (right panel).

proximity as well as the intensity for sales locations as a function of proximity. By analogy, we can imagine having a point pattern of tree locations along with diameter for each tree and asking if diameters for neighboring trees are strongly dependent. For trees, the answer may not be clear. For selling prices, adjusted for differences in regressors, this is the goal of our investigation.

Additionally, we can have covariates that are relevant in explaining both the selling price and the sales activity. Being more explicit, Paci et al. (2017) developed a LGCP model with intensity driven by demographic covariates as above (as well as economic covariates and a space-time GP). As noted in Section 2, such covariates arose by a partitioning of Zaragoza into a collection of areal units (urban polygons), assigning covariate levels to each unit over windows of time and then, if say $\mathbf{s} \in A_j$, setting $\mathbf{X}(\mathbf{s}) = \mathbf{X}(A_j)$, where $j = 1, \ldots, 52$. In this regard, none of the covariates used in modeling $\mathcal{S}$ are associated with individual properties. Rather, they are characteristics of the areal units in given time windows. Such covariates can also be employed in the hedonic modeling, viewed as “neighborhood” characteristics expected to influence selling price.

Evidently, we can have additional covariates that are only in the selling price model. More precisely, in constructing a hedonic model, we work at the level of the individual sale and so we customarily introduce features associated with the individual property sold. As noted in Section 2, the property specific information includes size, age, and presence of accessory elements.

3.2. Elaborating the modeling

Let $\mathbf{s}$ denote a location observed over a bounded spatial domain $D \subseteq R^2$. The collection of points that makes up the spatial point pattern is denoted by $\mathcal{S} = \{\mathbf{s}_i;\, i = 1, ..., n\}$, where $n \equiv N(D)$ is the number of points observed in D. In particular, as above, we model $\mathcal{S}$ using a LGCP driven by an intensity $\lambda(\mathbf{s})$ specified as

$$\log \lambda(\mathbf{s}) = \mathbf{X}(\mathbf{s})^T \boldsymbol{\beta} + \eta(\mathbf{s}), \tag{1}$$

where $\mathbf{X}(\mathbf{s})$ is a p-dimensional vector containing the demographic covariates discussed in the previous section at location **s** and the $\eta(\mathbf{s})$ are spatial random effects. These effects are assumed to come from a zero mean GP with covariance function of the form say $C(\mathbf{s},\mathbf{s}') = \sigma_\eta^2 \rho_\eta(\mathbf{s}-\mathbf{s}';\theta_\eta)$, where $\rho_\eta(\cdot;\theta_\eta)$ is a correlation function depending on θ_η. Paci et al. (2017) show that such modeling provides an effective LGCP model for explaining the point pattern of sales activity in the dataset described in Section 2. The LGCP likelihood for a realization $\mathcal{S}$ is given by

$$\mathcal{L}_1\left(\boldsymbol{\beta}, \boldsymbol{\eta}_D, \sigma_\eta^2; \mathcal{S}\right) = \exp\left\{-\int_D \lambda(\mathbf{s})d\mathbf{s}\right\} \prod_{i=1}^{n} \lambda(\mathbf{s}_i), \tag{2}$$

where $\lambda(\mathbf{s})$ is defined according to (1) and $\boldsymbol{\eta}_D = \{\eta(\mathbf{s}) : \mathbf{s} \in D\}$. The integral in (2) is a stochastic integral, i.e., an integral over a random realization of a stochastic process. It can never be evaluated explicitly and, in practice, it is approximated numerically using a grid of *representative points* (Banerjee et al., 2014) within D.

Turning to selling prices, let $Y(\mathbf{s})$ be the (log)selling price of a property at location $\mathbf{s}$[||]. The collection of selling prices is denoted by $\mathcal{Y}$. Let $\mathbf{W}(\mathbf{s}) = (\mathbf{V}(\mathbf{s}), \mathbf{X}(\mathbf{s}))$ where $\mathbf{V}(\mathbf{s})$ is the q-dimensional vector collecting individual property level covariates at the given location **s** introduced in Section 2. We consider four regression models for prediction of $Y(\mathbf{s})$ as follows.

(i) A simple spatial regression model

$$y(\mathbf{s}) = \mathbf{W}(\mathbf{s})^T\boldsymbol{\alpha} + \varepsilon(\mathbf{s}), \tag{3}$$

where we partition $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_V, \boldsymbol{\alpha}_X)$ and the $\varepsilon(\mathbf{s})$ are white noise errors, normally distributed with zero mean and variance τ^2. The hedonic model is fitted apart from the point pattern model, excluding the opportunity to learn about $\mathcal{Y}$ from $\mathcal{S}$ through the intensity.

(ii) A shared component model

$$y(\mathbf{s}) = \mathbf{W}(\mathbf{s})^T\boldsymbol{\alpha} + \delta\eta(\mathbf{s}) + \varepsilon(\mathbf{s}), \tag{4}$$

where $\eta(\mathbf{s})$ is as in (1). Again, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_V, \boldsymbol{\alpha}_X)$ and the $\varepsilon(\mathbf{s})$ are white noise errors, normally distributed with zero mean and variance τ^2. Here, $\eta(\mathbf{s})$ plays the role of a regressor and δ becomes the coefficient for preferential sampling effect. In the words of Diggle et al. (2010), $\delta = 0$ is non-preferential sampling while $\delta \neq 0$ indicates preferential sampling with the sign of δ indicating the direction of preferential adjustment. If $\delta > 0$ we find that, with increased expected sales activity, we predict increased selling price, vice versa for $\delta < 0$. So, fitting (4) and (1) jointly enables assessment of the *presence* of preferential sampling.

(iii) A geostatistical model

$$y(\mathbf{s}) = \mathbf{W}(\mathbf{s})^T\boldsymbol{\alpha} + \phi(\mathbf{s}) + \varepsilon(\mathbf{s}), \tag{5}$$

where again, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_V, \boldsymbol{\alpha}_X)$, the $\varepsilon(\mathbf{s})$ are white noise errors, normally distributed with zero mean and variance τ^2. Now, the $\phi(\mathbf{s})$ are spatial random effects specified by a zero mean GP with covariance function of form $C(\mathbf{s}-\mathbf{s}';\theta_\phi) = \sigma_\phi^2\rho_\phi(\mathbf{s}-\mathbf{s}';\theta_\phi)$, where $\rho_\phi(\cdot;\theta_\phi)$ is a correlation function depending on θ_ϕ. They provide local adjustment to the model in (i) and, as is customary, they are interpreted as capturing *omitted* covariates with spatial information. This

[||] We have carried out all of the ensuing analyses replacing selling price with price per square meter. Similar stories emerge so we chose to present the results only for selling price.

hedonic model is fitted apart from the point pattern model, excluding the opportunity to learn about $\mathcal{Y}$ from $\mathcal{S}$ through the intensity.

(iv) A geostatistical model with preferential sampling

$$y(\mathbf{s}) = \mathbf{W}(\mathbf{s})^T\boldsymbol{\alpha} + \phi(\mathbf{s}) + \delta\eta(\mathbf{s}) + \varepsilon(\mathbf{s}), \tag{6}$$

where $\eta(\mathbf{s})$ is as in (1) and, again, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_V, \boldsymbol{\alpha}_X)$, $\phi(\mathbf{s})$ are spatial random effects as in (5), and the $\varepsilon(\mathbf{s})$ are white noise errors, normally distributed with zero mean and variance τ^2. Again, we have the shared process idea and the coefficient δ carries the interpretation of a preferential sampling effect. The η process and the ϕ process are modeled a priori as independent. In conjunction with (1), this is the model offered in Pati et al. (2011). We see the introduction of $\eta(\mathbf{s})$ as a regressor in the geostatistical hedonic model. Because of the flexibility of $\phi(\mathbf{s})$, model (iv) need not perform better than model (iii). So, joint fitting of (6) and (1) assesses whether δ remains significant, whether we can *improve* on prediction of selling price in the presence of preferential sampling.

Under (6), the likelihood for a realization of selling prices $\mathcal{Y}$ is

$$\mathcal{L}_2\left(\boldsymbol{\alpha}, \delta, \boldsymbol{\phi}, \boldsymbol{\eta}, \sigma_\phi^2, \sigma_\eta^2, \tau^2; \mathcal{Y}\right) = \prod_{i=1}^{n} N\left(y(\mathbf{s}_i) \mid \mathbf{W}(\mathbf{s}_i)^T\boldsymbol{\alpha} + \phi(\mathbf{s}_i) + \delta\eta(\mathbf{s}_i)\,, \tau^2\right). \tag{7}$$

We offer four remarks:

Remark 1: Combining (1) with (3) or (5) implies independence of the $\mathcal{Y}$ and $\mathcal{S}$ processes. However, along with a suitable set of regressors, the flexibility of (5) may enable it to predict well ignoring the information in $\mathcal{S}$ if $\mathcal{S}$ does not show strong bias with regard to locations of sales over D. In other words, a shared process model for selling price need not be better than a geostatistical model for predicting selling price. Indeed, the geostatistical model allows selling price to have its own random effects process (GP) while the shared process model forces $\mathcal{Y}$ to inherit the same GP as for $\mathcal{S}$. As a result, a comparison of (1) and (4) (essentially just using (4) for prediction) with (1) and (5) may favor the latter. However, with bias in the locations of sales over D, comparison of (1) and (5) with (1) and (6) can now favor the latter. But, unless the preferential sampling effect is very strong, the flexibility in local adjustment available from the $\phi(\mathbf{s})$ will annihilate the local linear effect of $\eta(\mathbf{s})$ inherited from (1).

Remark 2: Combining (1) with (4) or (6) creates dependence between $\mathcal{Y}$ and $\mathcal{S}$ in the sense that $\eta(\mathbf{s})$ is shared in the model for each. When there is stochastic dependence between $\mathcal{Y}$ and $\mathcal{S}$ then not only does $\mathcal{S}$ inform about the mean surface for $y(\mathbf{s})$ but also $\mathcal{Y}$ informs about the intensity, $\lambda(\mathbf{s})$. It is natural to ask whether the data can identify both GPs in (6)? One way to clarify the identifiability issue is to note that the model in (1) informs about $\eta(\mathrm{s})$ and σ_η^2. But then, for the hedonic model in (6), we can learn about δ, the regression coefficient for $\eta(\mathrm{s})$ and therefore the spatial component of the residual, $\phi(\mathrm{s})$, as with usual geostatistical models. Nonetheless, since $\eta(\mathrm{s})$ is unobserved, unlike $\mathbf{W}(\mathrm{s})$, there could be some tradeoff between the $\delta\eta(\mathrm{s})$ term and the $\phi(\mathrm{s})$ term. This is perhaps why in the original Diggle et al. (2010) paper no $\phi(\mathrm{s})$ term was introduced, like our model (ii). Adding the $\phi(\mathrm{s})$ term to obtain our model (iv) was suggested in Pati et al. (2011) who were fortunate to have an ozone dataset which was sufficiently biased to extract a significant δ.

Remark 3: Though we assume that the $\eta(\mathbf{s})$ and $\phi(\mathbf{s})$ processes are independent, the joint model (1) and (6) specifies a joint model for sales activity and selling price that provides spatial dependence through a linear model of coregionalization (Banerjee et al., 2014). That is, if we write the shared component in the log intensity for the sales locations as $\eta(\mathbf{s}) = \sigma_\eta U_1(\mathbf{s})$ and we write the hedonic model in the form $y(\mathbf{s}) = \mathbf{W}(\mathbf{s})^T \boldsymbol{\alpha} + \delta\sigma_\eta U_1(\mathbf{s}) + \sigma_\phi U_2(\mathbf{s}) + \epsilon(\mathbf{s})$, where $\phi(\mathbf{s}) = \sigma_\phi U_2(\mathbf{s})$, then we have a coregionalization model with coregionalization matrix $A = \begin{pmatrix} \sigma_\eta & 0 \\ \delta\sigma_\eta & \sigma_\phi \end{pmatrix}$. We again see that $\delta = 0$ endows the locations model and the selling price model with spatial random effects but that the models are independent.

Remark 4: The need to approximate (2) with representative points along with the fairly large number of properties in the product term in (2) results in a high-dimensional multivariate normal vector of random effects as part of the model fitting. The additional GP for the selling prices included in (5) and (6) brings an additional high dimensional multivariate normal vector of random effects. To expedite computations we employ a nearest neighbor Gaussian process (NNGP; Datta et al. 2016) which we briefly review in the next subsection.

3.3. Replacing the GP with a NNGP

The NNGP (Datta et al., 2016) is constructed as follows. Let $\mathcal{R} = \{\mathbf{s}_1, \ldots, \mathbf{s}_r\}$ be a fixed set of r points in D, which is called 'reference set'. Let the vector $\boldsymbol{\zeta}_R = (\zeta(\mathbf{s}_1), \ldots, \zeta(\mathbf{s}_r))$ have a multivariate normal distribution with mean zero and spatial covariance matrix C_R arising from a GP over $D \subseteq R^2$. The NNGP proceeds from the factorization of the joint density of $\boldsymbol{\zeta}_R$ as the product of conditional densities, i.e.

$$p(\boldsymbol{\zeta}_R) = p\left(\zeta(\mathbf{s}_1)\right) p\left(\zeta(\mathbf{s}_2) \mid \zeta(\mathbf{s}_1)\right) \ldots p\left(\zeta(\mathbf{s}_r) \mid \zeta(\mathbf{s}_{r-1}), \ldots, \zeta(\mathbf{s}_1)\right). \tag{8}$$

Then, the conditioning sets on the right hand of (8) are replaced with carefully chosen conditioning subsets, $N(\mathbf{s}_i) \subset \mathcal{R} \setminus \{\mathbf{s}_i\}$, called 'neighbor sets', of size at most m, where $m << r$ (see e.g. Vecchia 1988; Stein et al. 2004; Datta et al. 2016).

The NNGP replaces the distribution for $\boldsymbol{\zeta}_R$ in (8) with

$$\begin{aligned} \widetilde{p}(\boldsymbol{\zeta}_R) &= \prod_{i=1}^{r} p\left(\zeta(\mathbf{s}_i) \mid \boldsymbol{\zeta}_{N(\mathbf{s}_i)}\right) \\ &= \prod_{i=1}^{r} N\left(\zeta(\mathbf{s}_i) \mid \mathbf{a}_i' \, \boldsymbol{\zeta}_{N(\mathbf{s}_i)}, v_i\right) = N\left(\boldsymbol{\zeta}_R \mid \mathbf{0}, \mathbf{K}\right) \end{aligned} \tag{9}$$

where $\mathbf{a}_i$ is the $m \times 1$ vector defined as $\mathbf{a}_i = C^{-1}_{N(\mathbf{s}_i),N(\mathbf{s}_i)} C_{N(\mathbf{s}_i),\mathbf{s}_i}$ and the variance is given by $v_i = C_{\mathbf{s}_i} - C_{\mathbf{s}_i,N(\mathbf{s}_i)} C^{-1}_{N(\mathbf{s}_i),N(\mathbf{s}_i)} C_{N(\mathbf{s}_i),\mathbf{s}_i}$. The resulting precision matrix $\mathbf{K}^{-1}$ is sparse with at most $O(rm^2)$ non-zero elements (Datta et al., 2016). Hence, the approximation in (9) produces a sparsity-inducing prior distribution for the spatial random effects over $\mathcal{R}$. Here, we choose the same reference set in the NNGP specification of $\phi(\mathbf{s})$ and $\eta(\mathbf{s})$. In particular, the reference set $\mathcal{R}$ consists of the set of the observed locations; this choice has been shown to yield good approximations to the parent GP (Datta et al., 2016).

With regard to the neighbor set $N(\mathbf{s}_i)$, any subset of $\{\mathbf{s}_1, \ldots, \mathbf{s}_{i-1}\}$ ensures a valid probability density in (9). However, the choice of the m nearest neighbors depends upon the order of the collection of the spatial points. Datta et al. (2016) showed that the inference is robust to the ordering of the locations. Since spatial locations are not ordered naturally, here we impose order by one of the coordinates (longitude). Then, we chose the neighbor set as the set of $m = 10$ nearest neighbors with respect to a Euclidean distance.

Next, let $\mathcal{U} = \{\mathbf{s}_1^*, \ldots, \mathbf{s}_q^*\}$ be any finite set in D outside $\mathcal{R}$, i.e., $\mathcal{R} \cap \mathcal{U}$ is empty. The nearest neighbor density of $\boldsymbol{\zeta}_U = \big(\zeta(\mathbf{s}_1^*), \ldots, \zeta(\mathbf{s}_q^*)\big)$ is defined conditional on $\boldsymbol{\zeta}_R$ as

$$\begin{aligned} \widetilde{p}(\boldsymbol{\zeta}_U \mid \boldsymbol{\zeta}_R) &= \prod_{j=1}^{q} p\left(\zeta(\mathbf{s}_j^*) \mid \boldsymbol{\zeta}_{N(\mathbf{s}_j^*)}\right) \\ &= \prod_{j=1}^{q} N\left(\zeta(\mathbf{s}_j^*) \mid \mathbf{a}_j'\, \boldsymbol{\zeta}_{N(\mathbf{s}_j^*)}\,, v_j\right) \end{aligned} \tag{10}$$

where $N(\mathbf{s}_j^*)$ is the set of m-neighbors of $\mathbf{s}_j^*$ which must be in the reference set $\mathcal{R}$, while $\mathbf{a}_j$ and v_j are defined analogous to (9). Datta et al. (2016) show that the finite dimensional distributions supplied through (9) and (10) determine a valid GP which is defined to be an NNGP. Here, the set $\mathcal{U}$ contains all the representative points (roughly 4500 points) used to evaluate the integral in the LGCP likelihood (2).

3.4. Fitting and prediction details

To complete the Bayesian specification we supply a prior distribution for all of the model parameters. Independent vague normal distributions are employed for all of the coefficients, the $\boldsymbol{\alpha}$'s and $\boldsymbol{\beta}$'s. Also, a vague normal prior distribution is adopted for δ, the preferential sampling coefficient.

We assume that the spatial random effects $\eta(\mathbf{s})$ come from an NNGP, arising from a parent GP with zero mean and an isotropic covariance function, with variance σ_η^2 and exponential correlation function, i.e., $C(\mathbf{s}, \mathbf{s} + h; \theta_\eta) = \sigma_\eta^2 \exp(-\theta_\eta \|h\|)$, depending upon a spatial decay parameter θ_η. We replace $\eta(\mathbf{s})$ in (1) with $\tilde{\eta}(\mathbf{s}) = \eta(\mathbf{s}) - \sigma_\eta^2/2$ so that, a priori, $E\left[\exp\{\tilde{\eta}(\mathbf{s})\}\right] = 1$ in order to *center* the random effects with regard to $\lambda(\mathbf{s})$.

Model fitting is more challenging with the additional spatial GP $\phi(\mathbf{s})$ but, for the NNGP, we only need the $\mathbf{s}_i$ associated with the sales. Hence, we assume that the spatial random effects $\phi(\mathbf{s})$ come from an NNGP, arising from a parent GP with zero mean and an exponential covariance function, depending upon a variance σ_ϕ^2 and a decay parameter θ_ϕ. To accommodate identifiability issues with the hyperparameters (Banerjee et al., 2014), we adopt an empirical Bayes approach for the decay parameters, i.e., we fix θ_η and θ_ϕ at values corresponding to spatial ranges of roughly 3.5 km and 1.5 km, respectively. We select a shorter range for selling prices than for selling intensity since we envision that price is more "local" than selling activity. Different ranges also allow us to better identify the two GPs of model in (6). We experimented with other values of the decay parameters, learning that predictive inference has little sensitivity to choices close to these values. Finally, we place vague inverse Gamma priors on the variances σ_η^2, σ_ϕ^2, and τ^2.

The joint posterior distribution of model (iv) in equation (6) is given by

$$
\begin{aligned}
p\left(\boldsymbol{\alpha},\boldsymbol{\beta},\delta,\boldsymbol{\eta}_D,\boldsymbol{\phi},\sigma_\eta^2,\sigma_\phi^2,\tau^2 \mid \mathcal{Y},\mathcal{S}\right) \;\propto\;& \mathcal{L}_1\left(\boldsymbol{\beta},\boldsymbol{\eta}_D,\sigma_\eta^2;\,\mathcal{S}\right)\times\mathcal{L}_2\left(\boldsymbol{\alpha},\delta,\boldsymbol{\phi},\boldsymbol{\eta},\sigma_\phi^2,\sigma_\eta^2,\tau^2;\,\mathcal{Y}\right)\\
\times\;& \prod_{i=1}^{n} N\left(\eta(\mathbf{s}_i)\mid \mathbf{a}_i'\,\boldsymbol{\eta}_{N(\mathbf{s}_i)}\,,v_i\right)\\
\times\;& \prod_{j=n+1}^{n+L} N\left(\eta(\mathbf{s}_j^*)\mid \mathbf{a}_j'\,\boldsymbol{\eta}_{N(\mathbf{s}_j^*)}\,,v_j\right)\\
\times\;& \prod_{i=1}^{n} N\left(\phi(\mathbf{s}_i)\mid \mathbf{b}_i'\,\boldsymbol{\phi}_{N(\mathbf{s}_i)}\,,w_i\right)\\
\times\;& \pi(\boldsymbol{\alpha})\times\pi(\boldsymbol{\beta})\times\pi(\delta)\times\pi(\sigma_\eta^2)\times\pi(\sigma_\phi^2)\times\pi(\tau^2)
\end{aligned}
\tag{11}
$$

where $\mathcal{L}_1\left(\boldsymbol{\beta},\boldsymbol{\eta}_D,\sigma_\eta^2;\,\mathcal{S}\right)$ and $\mathcal{L}_2\left(\boldsymbol{\alpha},\delta,\boldsymbol{\phi},\boldsymbol{\eta},\sigma_\phi^2,\sigma_\eta^2,\tau^2;\,\mathcal{Y}\right)$ are defined in (2) and (7), respectively, and $\pi(\cdot)$ denotes the prior distributions.

Fitting the model using Markov chain Monte Carlo (MCMC) requires Gibbs sampling steps for updating the coefficients of price modeling and Metropolis steps for updating the coefficients of sales modeling. In general, model fitting for the LGCP is computationally demanding because of the large number of Gaussian random variables which arise in the intensity. However, we benefit from the NNGP specification for the random effects which enables easily calculated univariate normal draws for these random effects. Thus, sequential updating of each of the η's and ϕ's conditionally on their neighbors in the reference set is performed through a random walk Metropolis step and a Gibbs step, respectively.

Besides the explanation provided by the regressors, the hedonic model is used for prediction. So, we need the predictive distribution $[Y(\mathbf{s}_0)|\text{data}]$ for a new location $\mathbf{s}_0$. This will require the samples from predictive distributions for $\eta(\mathbf{s}_0)$ and $\phi(\mathbf{s}_0)$ along with usual composition sampling (Banerjee et al., 2014). We perform model comparison out-of sample. The LGCP enables random p-thinning of the point pattern (Illian et al., 2008; Leininger and Gelfand, 2017) to partition the locations roughly into a proportion p for fitting and a proportion $1-p$ for validation. This thinning induces a fitting set and a validation set of selling prices. As a last remark, with the hedonic modeling here, prediction only considers interpolation within the given spatial study area.

4. Analysis of the full set of transactions in Zaragoza

In this Section we present the results obtained from fitting models discussed in Section 3.2 for the full set of residential sales data that consists of 2356 and 1726 transactions in 2007 and 2012, respectively. First, we show posterior summaries for the hedonic models and then present a comparison of such models in terms of predictive performance. Presumably, with the full census of sales transactions, design bias should not be an issue but a preferential sampling effect may still arise, as we discussed in the Introduction.

4.1. Posterior summaries

Posterior summaries for the models (i), (ii), (iii) and (iv), each fitted with (1), for each of the two years, are shown in Tables 1-4, respectively. Results of simple hedonic model in Table 1 show that the price is higher in less populated neighborhoods, with a higher percentage of elderly people and with a lower proportion of foreigners. However, the coefficients associated with the demographic covariates become not significant when spatial random effects are included in the hedonic model, with the only exception of the percent of foreigners.

Recall that the size of the apartment and the age of the property are on the log scale; in this way the corresponding coefficients estimate the elasticity of the price with respect to these covariates. The coefficient of the size of the property is positive, indicating that the bigger the property, the higher the price. Conversely, the negative coefficient of the age of the property reflects an age penalization on the price of the dwelling. In particular, since the size coefficient is roughly 1, an increase of 1% in (log) size results in a rise of roughly 1% in the price. Analogously, since the age coefficient is roughly -0.1 in 2007 and -0.25 in 2012, an increment of 1% in the age of the property is reflected in a decrease of around 0.1% in the price in 2007 and roughly 0.25% in 2012. Finally, the AE variable presents a negative coefficient in 2012, revealing, as anticipated, that presenting accessories yields an increase in expected price.

With regard to the preferential sampling coefficient δ under model (ii), we note that it is significantly negative in 2007 and positive in 2012. There is significant evidence of a preferential sampling effect, following the interpretation of Diggle et al. (2010). There is evidence of a latent economic effect that is introducing bias into the full set of transaction locations. These effects have to be interpreted with care. In 2007, before the crisis, there was high demand with relatively higher sales activity away from the center where properties were developed (some with subsidies) with relatively lower prices. So, $\delta < 0$ for this year reflects the effect of the log intensity to pull down prices where there was high activity. By contrast, in 2012, just after the crisis, overall demand decreased but there was relatively higher sales activity in the center, where prices were relatively higher. So, $\delta > 0$ for this year reflects the effect of the log intensity to push up prices where there was high intensity.

Under model (iv), the coefficient δ is not significant for both 2007 and 2012, revealing that the sales activity does not effect the selling prices in the presence of flexible GP random effects in the hedonic model. Again, when analyzing the full dataset, there is no expectation of strong sampling bias, hence there is no reason to expect a significant preferential sampling effect in model (iv). However, this is not the case when we employ a strongly biased subsample of transaction locations, as shown in Section 5 below.

4.2. Model comparison

Model comparison focuses only on prediction of selling prices since the motivation for this work is to assess the effect of preferential sampling on the prediction of such prices. That is, we hold out a proportion of the selling prices and obtain their associated posterior predictive distributions. Predictive mean square error (PMSE) and continuous ranked probability score (CRPS; Gneiting and Raftery 2007) are used to compare performance across the four choices of hedonic models. Here, a nominal 20% of data are held out for validation using a p-thinning approach (Illian et al., 2008; Leininger and Gelfand, 2017), resulting in 613 sales held out in 2007 and 460 sales held out in 2012.

Table 1. Posterior summaries of parameters of regression model (i).

	2007			2012		
	2.5%	50%	97.5%	2.5%	50%	97.5%
α_0	12.0753	12.1203	12.1665	11.5809	11.6266	11.6729
α_{pop}	-0.0656	-0.0380	-0.0105	-0.0912	-0.0607	-0.0299
$\alpha_{\leq 14}$	-2.0475	-0.7132	0.5694	-3.2256	-2.0220	-0.8768
$\alpha_{\geq 65}$	0.2026	0.9110	1.6436	0.6864	1.4316	2.2062
α_{foreign}	-1.8420	-1.4436	-1.0531	-2.3405	-2.0259	-1.7200
α_{size}	0.9130	0.9607	1.0119	1.0863	1.1415	1.1995
α_{age}	-0.1323	-0.1161	-0.1003	-0.2706	-0.2524	-0.2345
α_{AE}	-0.0802	-0.0327	0.0165	-0.1186	-0.0703	-0.0205
τ^2	0.1327	0.1405	0.1491	0.1162	0.1241	0.1331

Table 2. Posterior summaries of parameters of shared component model (ii).

	2007			2012		
	2.5%	50%	97.5%	2.5%	50%	97.5%
α_0	12.5545	12.6743	12.8218	11.0866	11.2173	11.3229
α_{pop}	-0.0821	-0.0356	0.0163	-0.0447	0.0070	0.0563
$\alpha_{\leq 14}$	-4.2382	-2.1571	-0.1696	-4.4355	-2.6334	-0.7507
$\alpha_{\geq 65}$	-1.3251	-0.0683	1.0735	-1.2399	-0.0415	1.0574
α_{foreign}	-1.9629	-1.2785	-0.2785	-1.7051	-1.0917	-0.4734
α_{size}	0.8143	0.8653	0.9150	1.0264	1.0853	1.1428
α_{age}	-0.1215	-0.1052	-0.0882	-0.2676	-0.2492	-0.2300
α_{AE}	-0.0817	-0.0341	0.0155	-0.1253	-0.0751	-0.0263
δ	-0.8431	-0.6867	-0.5580	0.4344	0.5484	0.6838
σ^2_η	0.1260	0.1615	0.2120	0.1080	0.1525	0.2189
τ^2	0.1083	0.1158	0.1242	0.0890	0.0972	0.1061

Table 3. Posterior summaries of parameters of geostatistical model (iii).

	2007			2012		
	2.5%	50%	97.5%	2.5%	50%	97.5%
α_0	12.0021	12.0908	12.1833	11.4903	11.6032	11.7074
α_{pop}	-0.0485	0.0034	0.0654	-0.0809	-0.0229	0.0368
$\alpha_{\leq 14}$	-3.6748	-1.2990	1.0384	-3.8308	-1.4872	0.8206
$\alpha_{\geq 65}$	-1.1069	0.1606	1.4812	-1.0711	0.4124	1.9133
α_{foreign}	-2.1212	-1.2862	-0.4358	-1.3901	-0.7524	-0.1100
α_{size}	0.8297	0.8798	0.9306	0.9697	1.0265	1.0865
α_{age}	-0.1318	-0.1147	-0.0985	-0.2540	-0.2356	-0.2171
α_{AE}	-0.0925	-0.0449	0.0023	-0.1159	-0.0652	-0.0166
σ^2_ϕ	0.0432	0.0567	0.0752	0.0545	0.0724	0.0937
τ^2	0.1016	0.1088	0.1163	0.0800	0.0872	0.0950

Table 4. Posterior summaries of parameters of geostatistical model with preferential sampling (iv).

	2007			2012		
	2.5%	50%	97.5%	2.5%	50%	97.5%
α_0	11.9308	12.1489	12.315	11.4932	11.6187	11.7493
α_{pop}	-0.0471	0.0062	0.0649	-0.0912	-0.0349	0.0193
$\alpha_{\leq 14}$	-3.6739	-1.3985	0.8132	-3.9975	-1.9712	0.0715
$\alpha_{\geq 65}$	-1.0635	0.2137	1.4425	-0.6425	0.7015	2.0568
α_{foreign}	-2.2668	-1.2984	-0.3535	-1.8518	-1.2604	-0.6474
α_{size}	0.8290	0.8783	0.9292	0.9673	1.0257	1.0855
α_{age}	-0.1314	-0.1139	-0.0979	-0.2519	-0.2333	-0.2140
α_{AE}	-0.0903	-0.0438	0.0043	-0.1188	-0.0686	-0.0198
δ	-0.1383	-0.0435	0.0682	-0.0980	0.0055	0.1057
σ^2_{ϕ}	0.0445	0.0571	0.0762	0.0458	0.0570	0.0708
σ^2_{η}	0.2964	0.3623	0.4363	0.2257	0.2803	0.3616
τ^2	0.1018	0.1086	0.1159	0.0692	0.0771	0.0849

Figure 5 displays the posterior predictive distribution for the selling price of four illustrative transactions in 2007 (first row) and in 2012 (second row) from each of the four models, overlaid with a vertical line showing the actual selling price. Properties located in the city center and away from the center are shown in the left and right panels, respectively. The posterior predictive distributions obtained under models (ii), (iii) and (iv) are well centered on the actual prices, showing the clear benefit of local spatial adjustments introduced by random effects relative to the simple spatial regression model.

Results of the model comparison are presented in Table 5. Comparing the simple spatial regression model with the shared component model we can appreciate the benefit of the correction introduced by the preferential sampling in the hedonic model. The comparison of model (ii) with model (iii) tells us that the geostatistical model slightly outperforms the preferential sampling model. This is likely attributable to model (iii) not having to share $\eta(\mathbf{s})$ with (1). However, the closeness in performance between these two models suggests that the effects of omitted variables are captured roughly equally as well by $\eta(\mathbf{s})$ as $\phi(\mathbf{s})$. Finally, adding a shared component to the geostatistical hedonic model, model (iv) does not provide any improvement in terms of predictive performance. Again, we have anticipated this in our discussion above.

In summary, there is a preferential sampling effect in the absence of the spatial GP in the hedonic model and there is substantial improvement in prediction when a shared process is introduced. However, with the GP included in the hedonic model, such benefit disappears. Moreover, with a full inventory of transactions, prediction of selling prices using model (iv) is not beneficial. We next demonstrate that this need not be the case when consequential bias is introduced into the locations of the sales.

5. Introducing bias in the set of transactions

In this section we show that we can benefit by adjusting for preferential sampling when we have a strongly biased subset of the data. In the canonical environmental exposure setting, we usually have a designed set of sampling locations which may introduce preferential sampling. In other words, there is bias in the selection of monitoring site locations. For instance, in Pati et al. (2011)

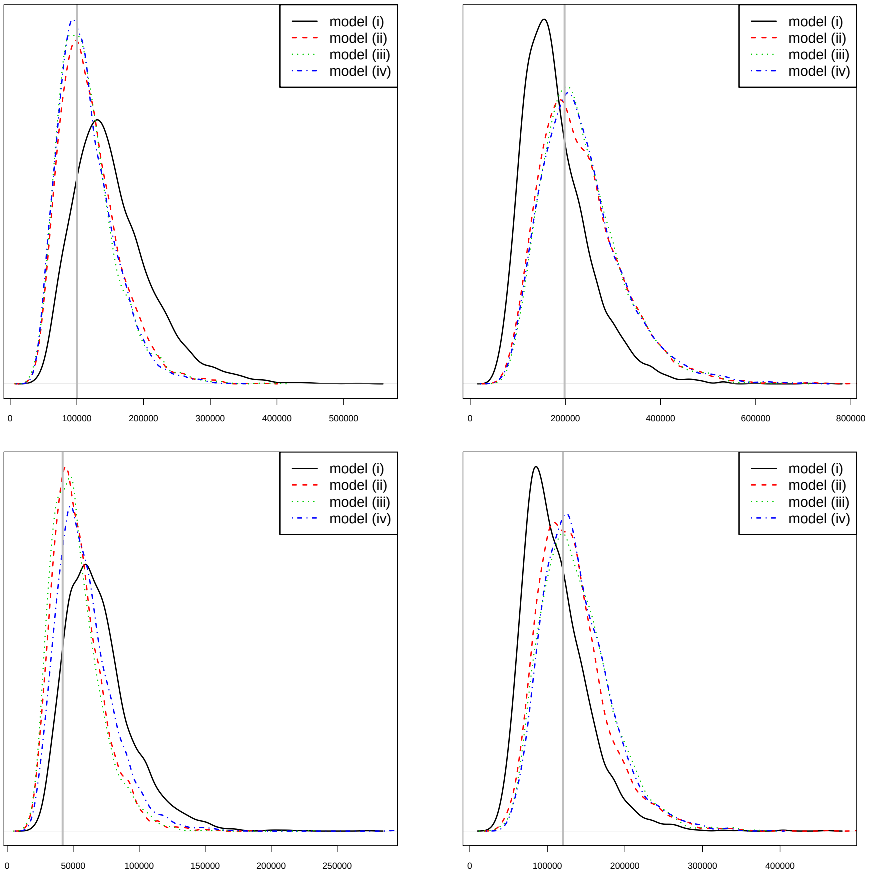

Figure 5. Posterior predictive distribution for the selling price of four illustrative transactions in 2007 (first row) and 2012 (second row) under the four models. Properties located in the city center and away from the center are shown in the left and right panels, respectively. The vertical line represents the actual selling price.

Table 5. Model comparison based on predictive mean square error (PMSE) and continuous ranked probability score (CRPS).

	2007		2012	
	PMSE	CRPS	PMSE	CRPS
model (i)	0.1487	0.2066	0.1247	0.1947
model (ii)	0.1346	0.1943	0.1121	0.1832
model (iii)	0.1328	0.1940	0.1098	0.1808
model (iv)	0.1327	0.1939	0.1124	0.1834

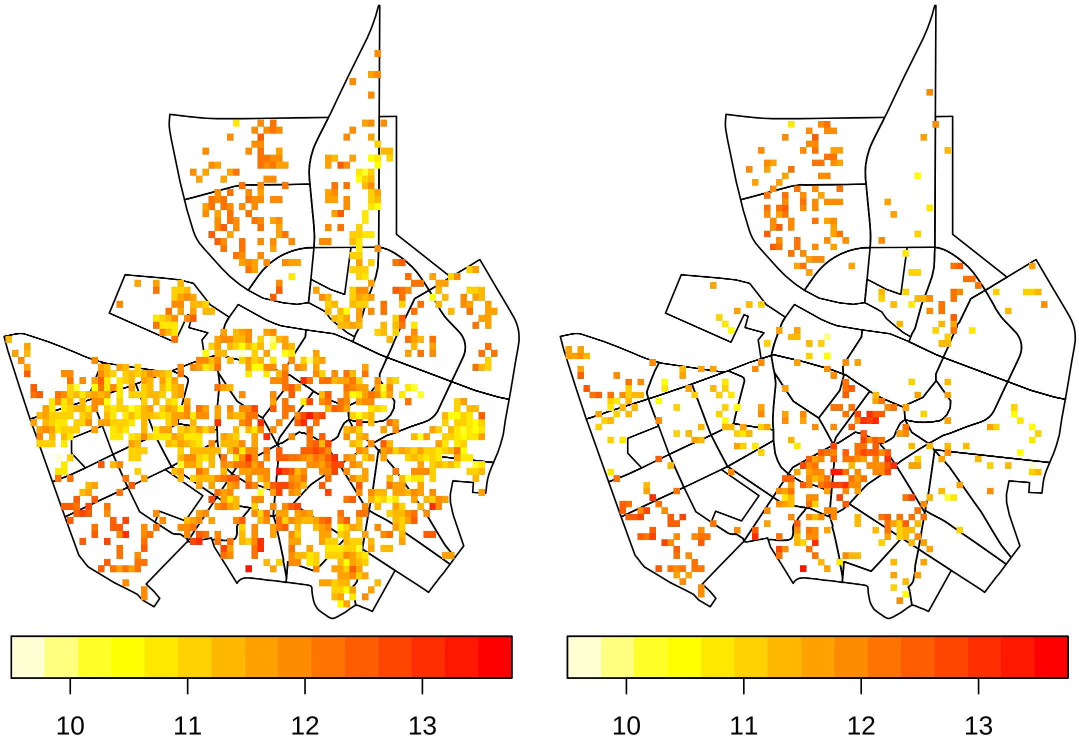

Figure 6. Log prices of all transactions (left panel) and of the biased sample of locations (right panel) over the subregion in 2012.

monitoring stations tended to over-sample locations with high exposure levels. This led to the log intensity surface associated with the designed locations becoming a significant covariate with a positive coefficient. In particular, they showed that this happened in the presence of the usual GP random effects, analogous to our geostatistical plus shared effects model (iv). Here, we develop a comparable illustration with our dataset by intentionally introducing design bias into our point pattern of transactions. We create a biased subset of transaction locations with regard to selling price, e.g., the sales locations available are more likely to be at say higher price locations.

Specifically, we draw a subsample of sales locations in 2012 in a subregion of the city. We take all the transactions from polygons which tend to have expensive apartments and we sample 1/8th of the locations from cheaper neighborhoods. Figure 6 displays the log prices of all transactions (left panel) and of the biased sample of locations (right panel) over the city center in 2012; that is, the biased sample on the right panel mimics the proposed design intention.

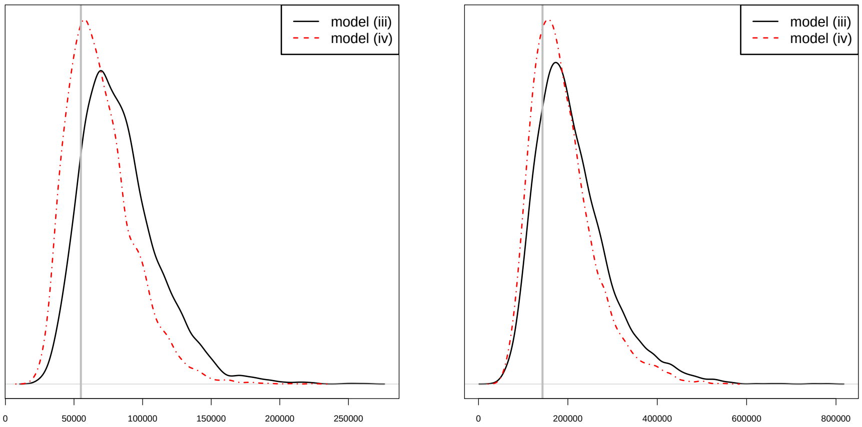

Figure 7. Posterior predictive distribution for the selling price of two illustrative transactions in 2012 under the two models. The vertical line represents the actual selling price.

Table 6. Model comparison based on PMSE and CRPS over the biased sample of transaction locations of 2012.

	PMSE	CRPS
model (iii)	0.1560	0.2213
model (iv)	0.1467	0.2140

We fit hedonic models (iii) and (iv) to the biased sample of transactions (616 transactions) and we predict selling prices for all of the transactions not sampled (1190 transactions). In this example, we show the results considering only log size and age as predictors, since the demographic covariates did not exhibit significant effects. As expected, the coefficient of the preferential sampling under model (iv) is significantly positive with 90% CI equal to [0.0655, 0.1718], raising the predicted response in areas of high activity, lowering it in areas of lower activity. So here, with a geostatistical hedonic model, we can see significant explanation associated with locations of sales activity.

Under models (iii) and (iv), each fitted to the biased sample, Figure 7 displays the posterior predictive distribution for the selling price of two illustrative held out transactions in 2012; the vertical line shows the actual selling price. In particular, the left panel is associated with a property located close to a location in the fitting set (less than 8 meters) while the right panel displays a property located far from the locations in the fitting set (more than 400 meters). In both cases, the posterior predictive distribution obtained under model (iv) is better centered and more concentrated with regard to the actual price.

Table 6 presents comparison of models (iii) and (iv) in terms of PMSE and CRPS obtained using the biased sample of transactions. The table now reveals that adding preferential sampling to the geostatistical model yields an improvement of 6% in PMSE and 3% in CRPS compared to the geostatistical hedonic model. Though the improvements are modest, the theme that emerges from both this section and the previous one is, essentially, that fitting the richer model (iv) will always be at least as good as fitting the geostatistical model (iii).

6. Summary and future work

We have considered the issue of preferential sampling in a novel context, that of extracting whether there is a preferential sampling effect on the selling price of properties, taking into consideration the fact that the locations of property sales are random. In fact, we have addressed two questions: (i) is there evidence of dependence between activity in sales locations and associated selling prices and (ii) if so, can we improve prediction of selling prices by taking advantage of this dependence. Our approach leads to a data fusion between a point pattern model for the locations of individual sales and a hedonic model for the selling prices of individual properties. We have demonstrated for both the full census of transactions and an intentionally biased sample of transactions the presence of a preferential sampling effect. We have further argued that, under the full census, a flexible geostatistical hedonic model is not outperformed by a model including a preferential sampling component. However, when there is substantial bias in the sample of locations, the latter can provide modest predictive improvement.

Future work will consider comparison with other cities (in particular, such data is available for several other cities in Spain). We can also explore retrospective investigation of the effects of economic shocks with regard to a potential preferential sampling effect. Additionally, since, for each transaction, we have both a location and a time of sale, we can attempt to extend this analysis to a continuous space-time setting, requiring jointly, a space-time hedonic model for selling price and a space-time point pattern model for property sales. Here, many modeling choices are available including dynamics in coefficients as well as separable vs. nonseparable space-time modeling for the random effects. Computation becomes more demanding due to the fact that the dimension of the latent random effects vectors will increase by an order of magnitude. This is exacerbated with a nonseparable covariance specification.

Acknowledgments

The research of Lucia Paci was partially funded by the research grant track D1 provided by the UCSC. The research of María Asunción Beamonte was funded by the Grupo consolidado DGA E-22 Modelos Estocásticos. The work of Pilar Gargallo and Manuel Salvador was supported by the Research Project ECO2016-79392-P provided by the Spanish Ministerio de Economía y Competitividad (Spain). Pilar Gargallo was also supported by grant SO2014-55780-C3-2-P provided by the Ministerio de Economía y Competitividad (Spain). We thank the Associate Editor and the anonymous reviewers for their valuable comments which have improved the paper.

References

Ahlfeldt, G. M. and Kavetsos, G. (2014) Form or function?: the effect of new sports stadia on property prices in London. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **177**, 169–190.

Anselin, L. and Lozano-Gracia, N. (2009) Spatial hedonic models. In *Palgrave Handbook of Econometrics* (eds. T. Mills and K. Patterson). London: Palgrave Macmillan.

Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2014) *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton, FL: Chapman and Hall/CRC, 2 edn.

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.

Beamonte, A., Gargallo, P. and Salvador, M. (2010) Robust bayesian inference in STAR models with neighbourhood effects. *Journal of Statistical Planning and Inference*, **140**, 3047–3057.

Cecconi, L., Grisotto, L., Catelan, D., Lagazio, C., Berrocal, V. and Biggeri, A. (2016) Preferential sampling and Bayesian geostatistics: Statistical modeling and examples. *Statistical Methods in Medical Research*, **25**, 1224–1243.

Conn, P. B., Thorson, J. T. and Johnson, D. S. (2017) Confronting preferential sampling when analysing population distributions: diagnosis and model-based triage. *Methods in Ecology and Evolution*, **8**, 1535–1546.

Cressie, N. and Wikle, C. (2011) *Statistics for Spatio-Temporal Data*. Hoboken, NJ: Wiley.

Datta, A., Banerjee, S., Finley, A. O. and Gelfand, A. E. (2016) Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, **111**, 800–812.

Diggle, P. J., Menezes, R. and Su, T.-l. (2010) Geostatistical inference under preferential sampling (with discussion). *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **59**, 191–232.

Dubin, R. A. (1988) Estimation of regression coefficients in the presence of spatially autocorrelated error terms. *The Review of Economics and Statistics*, **70**, 466–474.

Dubin, R. A. and Sung, C.-H. (1990) Specification of hedonic regressions: Non-nested tests on measures of neighborhood quality. *Journal of Urban Economics*, **27**, 97 – 110.

Eichholtz, P. and Lindenthal, T. (2014) Demographics, human capital, and the demand for housing. *Journal of Housing Economics*, **26**, 19 – 32.

Ferreira, G. d. S. and Gamerman, D. (2015) Optimal design in geostatistics under preferential sampling. *Bayesian Analysis*, **10**, 711–735.

Gelfand, A. E., Ghosh, S. K., Knight, J. R. and Sirmans, C. F. (1998) Spatio-temporal modeling of residential sales data. *Journal of Business & Economic Statistics*, **16**, 312–321.

Gelfand, A. E., Kim, H.-J., Sirmans, C. F. and Banerjee, S. (2003) Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, **98**, 387–396.

Gelfand, A. E., Sahu, S. K. and Holland, D. M. (2012) On the effect of preferential sampling in spatial prediction. *Environmetrics*, **23**, 565–578.

Gneiting, T. and Raftery, A. E. (2007) Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102**, 359–378.

Gonzalez, L. and Ortega, F. (2013) Immigration and housing booms: evidence from Spain. *Journal of Regional Science*, **53**, 37–59.

Illian, J., Penttinen, A., Stoyan, H. and Stoyan, D. (2008) *Statistical Analysis and Modelling of Spatial Point Patterns*. Wiley.

Lee, A., Szpiro, A., Kim, S. and Sheppard, L. (2015) Impact of preferential sampling on exposure prediction and health effect inference in the context of air pollution epidemiology. *Environmetrics*, **26**, 255–267.

Lee, D., Ferguson, C. and Scott, E. M. (2011) Constructing representative air quality indicators with measures of uncertainty. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **174**, 109–126.

Leininger, T. J. and Gelfand, A. E. (2017) Bayesian inference and model assessment for spatial point patterns using posterior predictive samples. *Bayesian Analysis*, **12**, 1–30.

Li, M. M. and Brown, H. J. (1980) Micro-neighborhood externalities and hedonic housing prices. *Land Economics*, **56**, 125–141.

Møller, J., Syversveen, A. R. and Waagepetersen, R. P. (1998) Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, **25**, 451–482.

Nychka, D. and Saltzman, N. (1998) Design of air-quality monitoring networks. In *Case Studies in Environmental Statistics* (eds. D. Nychka, W. W. Piegorsch and L. H. Cox), 51–76. New York, NY: Springer US.

Paci, L., Beamonte, M. A., Gelfand, A. E., Gargallo, P. and Salvador, M. (2017) Analysis of residential property sales using space-time point patterns. *Spatial Statistics*, **21**, 149 – 165.

Pati, D., Reich, B. J. and Dunson, D. B. (2011) Bayesian geostatistical modelling with informative sampling locations. *Biometrika*, **98**, 35–48.

Pronzato, L. and Müller, W. G. (2012) Design of computer experiments: space filling and beyond. *Statistics and Computing*, **22**, 681–701.

Ridker, R. G. and Henning, J. A. (1967) The determinants of residential property values with special reference to air pollution. *The Review of Economics and Statistics*, **49**, 246–257.

Shaddick, G. and Zidek, J. V. (2014) A case study in preferential sampling: Long term monitoring of air pollution in the UK. *Spatial Statistics*, **9**, 51 – 65.

Stein, M. L., Chi, Z. and Welty, L. J. (2004) Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society: Series B*, **66**, 275–296.

Vecchia, A. V. (1988) Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society: Series B*, **50**, 297–312.