On the Effect of Preferential Sampling in Spatial Prediction

Alan E. Gelfand, Sujit K. Sahu, and David M. Holland *

June 6, 2012

Abstract

The choice of the sampling locations in a spatial network is often guided by practical demands. In particular, many locations are preferentially chosen to capture high values of a response, for example, air pollution levels in environmental monitoring. Then, model estimation and prediction of the exposure surface become biased due to the selective sampling. Since prediction is often the main utility of the modeling, we suggest that the effect of preferential sampling lies more importantly in the resulting predictive surface than in parameter estimation. We take demonstration of this effect as our focus.

In particular, our contribution is to offer a direct simulation-based approach to assessing the effects of preferential sampling. We compare two predictive surfaces over the study region, one originating from the notion of an 'operating' intensity driving the selection of monitoring sites, the other under complete spatial randomness. We can consider a range of response models. They may reflect the operating intensity, introduce alternative informative covariates, or just propose a flexible spatial model. Then, we can generate data under the given model. Upon fitting the model and interpolating (kriging), we will obtain two predictive surfaces to compare with the known *truth*. It is important to note that we need suitable metrics to compare the surfaces and that the predictive surfaces are random, so we need to make *expected* comparisons. We also present an examination of real data using ozone exposures. Here, what we can show is that, within a given network, there can be substantial differences in the spatial prediction using preferentially chosen locations vs. roughly randomly selected locations and that the latter provide much improved predictive validation.

Keywords: fitting model; hierarchical model; informative covariate; intensity; sampling model; spatial point pattern.

1 Introduction

The choice of the sampling locations in a spatial network is often guided by practical demands such as the need to monitor air pollution levels near their most likely sources and in areas of high population density. Air pollution surfaces constructed solely on the basis of data obtained from these networks are likely to be biased if they are not adjusted for the effects of the choice of the monitoring sites. For example, if, due to locations, monitors tend to record high levels of exposure, interpolation of levels for low population density areas or locations away from sources such as power stations may be upwardly biased. That is, if the sampling locations are

*Alan E. Gelfand is Professor, Institute of Statistics and Decisions, Duke University, Durham, NC, USA (Email: alan@stat.duke.edu). Sujit K. Sahu is senior lecturer, Mathematics Academic Unit, Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, UK. (Email: S.K.Sahu@soton.ac.uk). David M. Holland is senior statistician, U.S. Environmental Protection Agency, National Exposure Research Laboratory, Research Triangle Park, NC, USA (Email: holland.david@epa.gov).

preferentially chosen to capture high (or low) values of a response, for example, air pollution levels, then subsequent model estimation and prediction of the exposure surface can become biased due to the selective sampling. In the sequel, we use the term "bias" informally but with the intention of capturing departure from what the exposure surface would look like if we interpolated given that the locations were selected under complete spatial randomness, see e.g. Diggle (2003).

Since prediction is often the main utility of the modeling, we suggest that the effect of preferential sampling lies more importantly in the resulting predictive surface than in parameter estimation. We adopt this as our focus, taking a direct simulation approach to assess the effect. Our basic idea is to compare two predictive surfaces. One originates from the notion of an 'operating' intensity driving the selection of monitoring sites. The other considers what would have been predicted had the sampling intensity been uniform, i.e., complete spatial randomness, over the study region. Given a set of monitoring stations, we can consider a range of response models. They may reflect the operating intensity, introduce alternative informative covariates, or just propose a flexible spatial model. In particular, we use three stylized but representative versions of these scenarios. Regardless, we can generate data under the given model. Then, upon fitting the model and interpolating (kriging), we will obtain two predictive surfaces to compare. Under this simulation, we will know the "truth" and so, can compare our predictive surfaces to it. Two remarks here are: (i) we need suitable metrics to compare the surfaces and (ii) the predictive surfaces are random, so we need to make *expected* comparisons.

We also include a real data example employing ozone exposures. However, with observational data, we can not know the truth. So, what we can show is: (i) within a given network, there can be substantial differences in the spatial prediction using preferentially chosen locations vs. roughly randomly selected locations and (ii) data from a randomly selected sample of sites from the network yields much better predictive validation than preferentially sampled data. Here, we recognize that neither set of sites actually arose from specification of an intensity; the exercise is only suggestive of what can happen in practice with observational data.

As a convincing, motivating example consider a model where space, denoted by t, is one dimensional. The model is written as $Z(t) = a + b\cos(t \mod 2\pi) + \epsilon(t)$ where a and b are unknown parameters and $\epsilon(t)$ is a mean zero Gaussian process. Information contained in the data $Z(t)$ for $t = 1, \ldots, n$, regarding a and b, can vary between two very different functions of a and b depending on the set of t's where we observe the $Z(t)$ process. If we only observed the process at $t \approx 2\pi k$, we would only see observed values near $a + b$. On the other hand, if we take all the observations near $t \approx \pi(2k + 1)$, we would only see values near $a - b$.

Recent discussion on preferential sampling has been sparked by the work of Diggle *et al.* (2010) who proposed a joint hierarchical model for the response and the locations. In particular, they adopt a model for the intensity that drives the locations which is assumed to be a spatial Gaussian process realization. Then, they employ this same Gaussian process realization to explain the responses. This may not be a sensible practical specification. Pati *et al.* (2011) generalize this approach in a Bayesian hierarchical setting, introducing common covariates into both the intensity for locations and the mean of the response model with two spatial Gaussian processes, one for the intensity and one for the response. It is unclear how well the use of these *informative* covariates in the regression model corrects for the preferential sampling bias introduced by these covariates in the location model.

We note that while preferential sampling often operates in practice, it is rare that sampling sites would be drawn randomly, using an explicit intensity function. In fact, there is a substantial literature on spatial design. See, e.g., Müller *et al.* (2001) or, from a Bayesian perspective, Pilz and Spöck (2008). As a result, we doubt that complete spatial randomness ever operates in practice. Rather, geometric ideas like space filling designs (Nychka and Saltzman, 1998) or spatially-balanced designs (Theobald *et al.*, 2007) offer non-model based, non-preferential, deterministic strategies. With regard to preferential sampling, if interest is in levels at certain locations, then it would be inappropriate to discourage sampling at those locations. Furthermore, if the available data is preferentially sampled but is the only data that can be expected, then, presumably it would be analyzed. Our point is only that one might not feel comfortable

with the potential bias in predictions made from it. Again, we employ intensities to provide stochastic models for preferential and non-preferential sampling.

In this regard, Lee *et al.* (2011) apply the preferential sampling approach to build representative air quality indicators and their associated uncertainty measures from multiple air pollutants. The general area of environmental exposure modeling given the non-random monitoring sites has seen a lot of activity in the recent literature, see e.g., the very recent work of Sheppard *et al.* (2011). There is also a substantially Bayesian literature, for example, Cocchi *et al.* (2007) who developed hierarchical model for daily average PM_{10} concentration levels and Sahu *et al.* (2007) where a hierarchical auto-regressive model was developed for daily maximum 8-hour average ozone concentration levels.

With regard to studying point patterns, there is an enormous literature, summarized in the books by, e.g., Diggle (2003) and Illian *et al.* (2008). In studying species abundance over large spatial regions, Chakraborty *et al.* (2010) use point pattern analysis to address preferential sampling in the context of sampling effort. Incorporating land transformation, they distinguish an operating intensity from a potential intensity. See also earlier environmental settings employing point pattern data modeling, such as Hooten *et al.* (2003) and Latimer *et al.* (2006).

As we detail in Section 2, we conduct 12 experiments arising from two choices for the sampling model by three choices for the fitting model by two choices for the intensity under a given sample size. Again, comparison between the intensities is in the predictive space. Specifically, predictive surfaces are compared under the same response (or first stage data model) but with different intensities for the point pattern of sites (or second stage specification). We note that we are not interested in testing whether complete spatial randomness is an acceptable hypothesis for the sites. Rather, we are assuming that preferential sampling implies this is not the case and that we are trying to reveal its impact.

The format of the paper is as follows. Section 2 details the broad technical issues in modeling, distinguishing our approach from the method proposed by Diggle *et al.* (2010). The specific details of our approach are provided in Section 3. Section 4 takes up general simulation issues for us while Section 5 considers metrics for comparing surfaces. Section 6 lays out the specific simulation design that we use to compare different sampling schemes. Section 7 describes the results from the simulation study while Section 8 provides a preferential sampling assessment for ozone data. Summary remarks and directions for future work are given in Section 9.

2 Technical issues in modeling

As noted in the Introduction, we take a direct approach to investigate the effects of preferential sampling. In this regard, we treat the sampling locations as random, rather than fixed as is often done in spatial and spatio-temporal data modeling, see e.g. Banerjee *et al.* (2004) and also Cressie and Wikle (2011). Hence, we have a multilevel specification where we model intensity, then locations given intensity, then process given locations and finally, observations given process. Here, process is the spatial environmental process of interest over the study region, e.g., a climate process such as temperature or precipitation or a pollutant process such as ozone or particulate matter. We investigate such hierarchical specifications within the Bayesian framework, adding priors for the parameters introduced at each modeling stage.

The fundamental approach is to use simulation to reveal what can happen under preferential sampling and what can happen if informative covariates are introduced to attempt to remedy bias. The examples we present in Section 4 are simplified, not of necessity but rather to facilitate illumination of the effects. Within the context of simulation and hierarchical modeling, we need to specify both the model for the intensity and the model for the process. We will assume that the data given process are conditionally independent, i.e., that we have a nugget. For the intensity, we consider two choices. One is a preferential sampling form motivated by "sampling where people are", i.e., by a population density surface. The other is complete spatial randomness.

As noted in the Introduction, in practice, we do not fit an intensity model; we assume the

locations are fixed and we focus on describing the process that is used to explain the response. This leads to envisioning a *true* response surface (which, again, is assumed to be observed up to white noise) along with a *fitting* surface which, in practice, is never true but supplies a process model. We consider two scenarios for the true surface. One is that it arises from an externality, i.e., there is a pollution source at a given location and exposure decays in distance from the source. The second is a process model that uses an informative covariate; in our case, we take it to be population density. The notion here is that, if sampling locations are drawn as a reflection of population density, then we should use population density in the response model to "correct" the preferential sampling bias. We consider three scenarios for the fitting surface, the two foregoing choices for the sampling surface along with, arguably, the most common choice, introduction of spatial random effects through a Gaussian process (GP). So, altogether, we have 2 `point pattern models` $\times 2$ `sampling models` $\times 3$ `fitting models` $= 12$ simulation cases. Each one will be replicated 100 times to enable suitable averaging to learn about expected performance and expected differences between each of these surfaces under preferential sampling vs. complete spatial randomness. Thus, the proposed method is quite different from the work of Diggle *et al.* (2010) and Pati *et al.* (2011) and is very computationally demanding as is elaborated in Section 3.

We note that the goal of working with these various simulation cases is to be able to distinguish between the effect of preferential sampling and the effect of using the "wrong" model. But then, this takes us to the remaining ingredient, the selection of suitable metrics to make desired comparisons; we take this up in Section 5. As noted in the Introduction, our approach seems novel in the context of preferential sampling. Moreover, it is attractive in explicitly revealing the impact of preferential sampling on prediction. Though it is offered in a fairly simple setting, it can serve as a suggestive template for further investigations. Suitably modified, it enables us to work with real data, as in Section 8. With real data, we do not have the true response surface. But, we can use the above ideas to approximate a predictive comparison between preferential and non-preferential sampling.

An obvious but worthwhile remark here is that preferential sampling affects the choice of locations but the true response surface is *not* affected by the choice of sampling locations. Typically, this is manifested in practice such that the choice of response model has nothing to do with the choice of the sampling locations. Nonetheless, what we infer about the response surface is affected by the choice of sampling locations. Our circumstance is akin to familiar regression settings where inference regarding a nonlinear (even a linear) relationship between response Y and covariate X can be affected by what levels we have drawn for X. Reiterating, we do allow the possibility of the response surface and the point process surface being similar as both can share a set of common covariates. Any such covariate, e.g., population density with an environmental monitoring network, is referred to as an 'informative covariate' for both the response and the intensity process. Again, the point is that a common covariate can influence both the response and the locations but the sampling locations cannot influence the response surface. The foregoing simulation design will enable us to assess how successful, in terms of correcting bias in the response model, the introduction of an informative covariate is.

Throughout the paper we consider the number of sampling locations, n, to be fixed. That is, to envision practical use of our approach, we would imagine having in place an existing monitoring network of a given size. With a non-homogeneous Poisson process (NHPP) model (see, e.g. Diggle, 2003) we have conditional independence of the locations given the intensity. More general models for the intensity, e.g., incorporating clustering or inhibition (e.g., Illian *et al.*, 2008), are not considered here.

3 Our approach

3.1 Point process models

Suppose that the network consists of n fixed monitoring sites $\mathbf{s}_1, \ldots, \mathbf{s}_n$ within a study domain D. In practice, these sites may have been chosen without carrying out any formal sampling

design but according to considerations such as population density and proximity to air pollution sources. Let $\mathbf{x}(\mathbf{s})$ denote the vector of levels of these covariates at location $\mathbf{s}$. Then, under a NHPP model for the sites, the underlying sampling intensity might take the form:

$$\lambda_1(\mathbf{s};\boldsymbol{\alpha}_1) = \exp\left(\mathbf{x}(\mathbf{s})'\boldsymbol{\alpha}_1\right)$$

where $\boldsymbol{\alpha}_1$ denotes the unknown parameters.

We seek to compare the implications of the operating intensity $\lambda_1(\mathbf{s};\boldsymbol{\alpha}_1)$ with any other intensity of interest $\lambda_2(\mathbf{s};\boldsymbol{\alpha}_2)$ on the predictive surfaces. To make a meaningful comparison, we assume that the two intensities are matched in scale, i.e., that we have

$$\int_D \lambda_1(\mathbf{u};\boldsymbol{\alpha}_1)d\mathbf{u} = \int_D \lambda_2(\mathbf{u};\boldsymbol{\alpha}_2)d\mathbf{u}. \tag{1}$$

The second intensity $\lambda_2(\mathbf{s};\boldsymbol{\alpha}_2)$ can be specified similarly to the first one, but possibly with different covariates arising from different considerations. The default alternative which we focus on is the homogeneous Poisson process (HPP) yielding $\lambda_2(\mathbf{s};\boldsymbol{\alpha}_2) = \lambda_2,\ \ \mathbf{s} \in D$. The constraint (1) implies that, in this case,

$$\begin{array}{rcl} \int_D \lambda_1(\mathbf{u};\boldsymbol{\alpha}_1)d\mathbf{u} & = & \int_D \lambda_2(\mathbf{u};\boldsymbol{\alpha}_2)d\mathbf{u} \\ & = & \lambda_2 \int_D d\mathbf{u} \\ & = & \lambda_2|D| \end{array}$$

where $|D| = \int_D d\mathbf{u}$ denotes the area of the region D. Thus we take $\lambda_2 = \frac{\int_D \lambda_1(\mathbf{u};\boldsymbol{\alpha}_1)d\mathbf{u}}{|D|}$. In fact, since we condition on a fixed n, in the HPP case, the resulting conditional density is $1/|D|$ over $\mathbf{s} \in D$, regardless of λ_2.

In general the network of sites $\mathbf{s}_1, \ldots, \mathbf{s}_n$ is thus assumed to be a realization of a point process having the density function

$$f(\mathbf{s}|\boldsymbol{\alpha}) = \frac{\lambda(\mathbf{s};\boldsymbol{\alpha})}{\int_D \lambda(\mathbf{u};\boldsymbol{\alpha})d\mathbf{u}}, \quad \mathbf{s} \in D$$

where the intensity function $\lambda(\mathbf{s};\boldsymbol{\alpha})$ is either of the two intensities λ_1 or λ_2. The Bayesian model for this stage of specification is completed by assuming a suitable prior $\pi(\boldsymbol{\alpha})$ for $\boldsymbol{\alpha}$. The resulting posterior distribution is $\pi(\boldsymbol{\alpha}|\mathbf{s}_1, \ldots, \mathbf{s}_n) \propto \prod_{i=1}^n f(\mathbf{s}_i|\boldsymbol{\alpha})\pi(\boldsymbol{\alpha})$ and the prior predictive distribution for n locations becomes

$$\pi(\mathbf{s}_1, \ldots, \mathbf{s}_n) = \int \prod_{i=1}^n f(\mathbf{s}_i|\boldsymbol{\alpha})\pi(\boldsymbol{\alpha})d\boldsymbol{\alpha}.$$

In the simulations below, for convenience, we assume that the λ's are known and hence no MCMC model fitting is necessary. However, in practical situations, a proposed parametric intensity surface $\lambda_1(\mathbf{s})$ would be unknown and would have to be estimated from the observed realization of the sampling locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$. Several methods exist for this task, see e.g., Diggle (2003), Illian *et al.* (2008) and, from a Bayesian perspective, Chakraborty *et al.* (2010). This estimation can be performed independently of the fitting of the response model, i.e., of the estimation for the response model parameters.

In particular, the covariate surfaces will typically be tiled to some spatial resolution so fitting and sampling requires working only with $\{\mathbf{x}_j\}$ indexing the resulting grid cells. That is, $\mathbf{x}(\mathbf{s}) = \mathbf{x}_j$ if $\mathbf{s} \in A_j$. So, once fitted, to generate a point pattern under a given $\lambda(\mathbf{s};\boldsymbol{\alpha})$, we only need to take a maximum over a finite number of cells in order to use a standard rejection/thinning algorithm, see e.g. Lewis and Shedler (1979).

In some cases, to achieve more flexibility, we introduce a Gaussian process (GP) into the model for the intensity surface, yielding a so-called Cox process. (See, Illian *et al.*, 2008 or Diggle *et al.*, 2010, in this regard.) We might even introduce heterogeneity in the uncertainty associated with the surface, uncertainty which depends upon covariate levels. In light of the above, we

sample point patterns from an intensity given the tiling associated with the covariates $\mathbf{x}(\mathbf{s})$'s. Hence, we can use the same discretization to accommodate the GP, resulting in a finite set of spatial random effects with a joint multivariate normal distribution, following, e.g., Banerjee *et al.* (2004). However, with many grid cells, we obtain a high dimensional distribution. Dimension reduction, perhaps employing predictive processes, see e.g. Banerjee *et al.* (2008) may be used to address the computational burden. In the sequel, under our simulation-based perspective, we omit the GP component in the interest of simplicity as well as reducing noise that may obscure differences we seek to reveal.

3.2 Response Models

For observed data $Z(\mathbf{s})$ at a location $\mathbf{s}$, we assume the customary hierarchical model:

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + \epsilon(\mathbf{s}), \tag{2}$$

where $\epsilon(\mathbf{s})$ is assumed to follow the Gaussian error distribution with mean zero and variance σ^2_ϵ, independent across the locations. The spatial process model for the mean, given a set of location specific covariates $\mathbf{x}(\mathbf{s})$, is given by:

$$\mu(\mathbf{s}) = \mathbf{x}(\mathbf{s})'\boldsymbol{\beta} + w(\mathbf{s}) \tag{3}$$

where we assume a Gaussian process prior for $w(\mathbf{s})$ with zero mean and, for convenience, an isotropic covariance function $\sigma^2_w \rho(||\mathbf{s} - \mathbf{s}'||; \phi)$ independently of $\epsilon(\mathbf{s})$, see e.g. Banerjee *et al.* (2004). Here, we use $\mathbf{x}(\mathbf{s})$ generically, as in Section 3.1, recognizing that different components of $\mathbf{x}(\mathbf{s})$ may be used in the intensity model vs. in the response model. Our simulation adopts, for illustrative purposes, the exponential correlation function $\rho(d; \phi) = \exp(-\phi\, d)$ (but any choice can be considered). The specification is completed by assuming a prior distribution, $\pi(\boldsymbol{\theta})$ for the unknown parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \phi, \sigma^2_w, \sigma^2_\epsilon)$.

Let $\mathbf{z} = (z(\mathbf{s}_1), \ldots, z(\mathbf{s}_n))$ and $\boldsymbol{\mu} = (\mu(\mathbf{s}_1), \ldots, \mu(\mathbf{s}_n))$ denote the data and the mean response vector at n locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$. The joint posterior distribution of $\boldsymbol{\theta}$ and $\mathbf{w}$, where $\mathbf{w}$ is the vector of random effects, is given by $\pi(\boldsymbol{\theta}, \mathbf{w}|\mathbf{z}) \propto f(\mathbf{z}|\mathbf{w}, \boldsymbol{\theta})\pi(\mathbf{w}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$, and the posterior predictive distribution of $Z(\mathbf{s}_0)$ is given by

$$f(z(\mathbf{s}_0)|\mathbf{z}) = \int f(z(\mathbf{s}_0)|w(\mathbf{s}_0), \boldsymbol{\theta})\pi(w(\mathbf{s}_0)|\mathbf{w}, \boldsymbol{\theta}, \mathbf{z})\pi(\boldsymbol{\theta}, \mathbf{w}|\mathbf{z}) d\mathbf{w} d\boldsymbol{\theta}. \tag{4}$$

3.3 Joint modeling

We now consider both the locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$ and the data $z(\mathbf{s}_1), \ldots, z(\mathbf{s}_n)$ to be random. Evidently, we have to specify this model through a distribution for locations and then a conditional distribution for the data given the locations. Below, we have two choices for λ, i.e., $\lambda_k, k = 1, 2$ and three choices for $\boldsymbol{\theta}$, i.e., $\boldsymbol{\theta}_r, r = 1, 2, 3$. Suppressing r and k for the moment, we have the following Bayesian hierarchical model:

$$f(\mathbf{z}|\mathbf{w}, \boldsymbol{\theta}, \mathbf{s}_1, \ldots, \mathbf{s}_n)\, \pi(\mathbf{w}|\boldsymbol{\theta}, \mathbf{s}_1, \ldots, \mathbf{s}_n)\, f(\mathbf{s}_1, \ldots, \mathbf{s}_n|\boldsymbol{\alpha})$$

yielding the joint posterior distribution of $\mathbf{w}$, $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ given by

$$\pi(\mathbf{w}, \boldsymbol{\theta}, \boldsymbol{\alpha}|\mathbf{s}_1, \ldots, \mathbf{s}_n, \mathbf{z}) \propto \prod_{i=1}^{n} f(z(\mathbf{s}_i)|w(\mathbf{s}_i), \boldsymbol{\theta}, \mathbf{s}_i)\pi(\mathbf{w}|\boldsymbol{\theta}, \mathbf{s}_1, \ldots, \mathbf{s}_n)$$
$$\cdot\, \pi(\boldsymbol{\theta}) \prod_{i=1}^{n} f(\mathbf{s}_i|\boldsymbol{\alpha})\pi(\boldsymbol{\alpha}).$$

The factorization shows that $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ can be estimated separately, given the sampling locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$. This is apart from whether covariate x appears in the distribution for $z(\mathbf{s}_i$, $\mathbf{s}_i$, or both.

4 Simulation issues

In implementing a simulation study such as we have proposed, we can envision two settings: (i) the no data case and (ii) the data case. We describe both cases in detail here but, in the sequel, confine ourselves to the no data case.

For us, the *no data* setting means no observed $\mathbf{s}$'s and hence, no observed z's. We undertake what might be referred to as a *preposterior* analysis, see e.g., Dawid (1984). That is, for the purposes of our simulation study, we do not need to condition on a fixed dataset; we are only interested in comparing the effects of two different intensities. However, in Section 8, we work with real data and there we do have observed $\mathbf{s}$'s and z's. As noted above, we consider two sampling models, each with no spatial random effects. Thus, given $\boldsymbol{\theta}_r, r = 1, 2$, there is a true mean surface $\mu_r(\mathbf{s}) = \mathbf{x}_r(\mathbf{s})^T\boldsymbol{\beta}, \mathbf{s} \in D$. To study the effect of preferential sampling, we will compare $\mu_r(\mathbf{s})$ with a collection of simulated posterior predictive surfaces.

In particular, this collection will be created for each of three fitting models ($\boldsymbol{\theta}_r$) with each of two intensity models ($\boldsymbol{\alpha}_k$) for a fixed sample size n. In our experiments we have used $n = 100$, although we have investigated with $n = 50$ and $n = 300$ where the results did not change substantially at all. That is, we will create $b = 1, \ldots, B$ predictive surfaces for $3 \times 2 = 6$ fitting scenarios. These will be compared with the truth under both $r = 1$ and $r = 2$, using the metrics presented in Section 5 below. The simulation will require two loops, an outer loop over $b = 1, \ldots, B$ to generate the replicates and an inner loop over $l = 1, \ldots, L$, generating posterior samples associated with that replicate, in order to obtain the posterior mean estimated surface for that replicate.

Hence, under sampling model given by $\boldsymbol{\theta}_r$ and given by intensity $\boldsymbol{\alpha}_k$, we would draw $\boldsymbol{\alpha}_k^*$ from $\pi_k(\boldsymbol{\alpha})$. Alternatively, we could fix $\boldsymbol{\alpha}_k$. Regardless, we draw $\mathbf{s}_1^*, \mathbf{s}_2^*, \ldots, \mathbf{s}_n^*$ from $f(\mathbf{s}|\boldsymbol{\alpha}_k)$. Under response model r, given $\boldsymbol{\theta}_r$, draw $\mathbf{w}^* = (w(\mathbf{s}_1^*), w(\mathbf{s}_2^*), \ldots, w(\mathbf{s}_n^*))$, if needed, and, finally, $\mathbf{z}^* = (z(\mathbf{s}_1^*), z(\mathbf{s}_2^*), \ldots, z(\mathbf{s}_n^*))$ becomes our data from the sampling model.

Now, given a fitting model, for a grid of $\tilde{\mathbf{s}}_j$ over D, obtain $E(Z(\tilde{\mathbf{s}}_j)|\mathbf{z})$. This requires sampling from $f(z(\tilde{\mathbf{s}}_j)|\mathbf{z})$. This is standard MCMC based model fitting, see e.g. Banerjee *et al.* (2004).

We repeat the simulation B times. Hence, up to discretization, these become the set of B mean surfaces to compare with the $\mu_r(\mathbf{s})$ surface. Note that we could introduce further randomness by randomly drawing $\boldsymbol{\theta}_r$. In this case, each simulated posterior mean surface replicate has its own "true" mean surface, $\mu_{r,b}(\mathbf{s}) = \mathbf{x}_r(\mathbf{s})^T\boldsymbol{\beta}_{r,b}$.

Returning to the *data* setting, with an observed set of $z(\mathbf{s}_i)$ at locations $\mathbf{s}_i, i = 1, \ldots, n$, we would avoid specifying $\boldsymbol{\alpha}$'s and $\boldsymbol{\theta}$'s, rather obtaining posterior distributions for them. So, given k, fit $f(\mathbf{s}_1, \ldots, \mathbf{s}_n|\boldsymbol{\alpha}_k)\pi(\boldsymbol{\alpha}_k)$ to obtain $\pi(\boldsymbol{\alpha}_k|\mathbf{s}_1, \ldots, \mathbf{s}_n)$. Given r, fit $f(\mathbf{z}|\mathbf{w}_r, \boldsymbol{\theta}_r)\pi(\mathbf{w}_r|\boldsymbol{\theta}_r)\pi(\boldsymbol{\theta}_r)$ to obtain $\pi(\boldsymbol{\theta}_r|\mathbf{z})$. Then, under λ_k, draw $\boldsymbol{\alpha}_k^*$ from $\pi(\boldsymbol{\alpha}_k|\{\mathbf{s}_i, i = 1, \ldots, n\})$, then draw $\{\mathbf{s}_i^*, i = 1, \ldots, n\}$ given $\boldsymbol{\alpha}_k^*$, then draw $\boldsymbol{\theta}_r^*$ from $\pi(\boldsymbol{\theta}_r|\mathbf{z})$. If needed, draw $\{w(\mathbf{s}_i^*)\}$ given $\boldsymbol{\theta}_r^*$ and finally, $\mathbf{z}^*$. Again, $\mathbf{z}^*$ is viewed as data from the sampling model. As in the no data case, given $\mathbf{z}^*$ and a fitting model, we can now obtain a set of posterior mean surfaces. Since now there is no overall truth, we would define the "true" sampling model as the mean surface associated with $\boldsymbol{\theta}_r^*$. That is, because $\boldsymbol{\theta}_r$ is random, as above, each posterior mean surface replicate has its own true mean surface.

It may be useful to present the above scheme in algorithmic format. It takes the following form:

No Data Case

1. Specify a fixed sample size n, a sampling model $\boldsymbol{\theta}_r$ and an intensity sampling model $\boldsymbol{\alpha}_k$. Assume that either $\boldsymbol{\theta}_r$ and $\boldsymbol{\alpha}_k$ are fixed and known or, if not, draw these from their respective prior distributions.
2. Obtain the true surface $\mu_r(\tilde{\mathbf{s}}_j) = \mathbf{x}_r(\tilde{\mathbf{s}}_j)^T\boldsymbol{\beta}$ for a grid of $\tilde{\mathbf{s}}_j$ over D.
3. Start simulation replicate b. (This is the outer loop in the above discussion.)
4. Draw a set of n locations $\mathbf{s}_1^*, \mathbf{s}_2^*, \ldots, \mathbf{s}_n^*$ from $f(\mathbf{s}|\boldsymbol{\alpha}_k)$.
5. If the sampling model, $\boldsymbol{\theta}_r$ is spatial, draw $\mathbf{w}^* = (w(\mathbf{s}_1^*), w(\mathbf{s}_2^*), \ldots, w(\mathbf{s}_n^*))$, from the Gaussian process model, see immediately below (3).

6. Draw $\mathbf{z}^* = (z(\mathbf{s}_1^*), z(\mathbf{s}_2^*), \ldots, z(\mathbf{s}_n^*))$ from the response model (2).
7. Choose a fitting model, say $\boldsymbol{\theta}_j$, and fit it to the data $\mathbf{z}^*$ using standard MCMC methods.
8. From the MCMC model fitting, for a grid of $\tilde{\mathbf{s}}_j$ over D, obtain $\hat{\mu}^{(b)}(\tilde{\mathbf{s}}_j) = E(Z(\tilde{\mathbf{s}}_j)|\mathbf{z})$ where the expectation is over the posterior predictive distribution $f(z(\tilde{\mathbf{s}}_j)|\mathbf{z})$, see (4). This is achieved by L posterior predictive draws $z^{(l)}(\tilde{\mathbf{s}}_j)$, $l = 1, \ldots, L$, which is the inner loop in the above discussion.
9. End simulation replicate b.
10. At the end of the simulation replicate compare the true surface obtained in step 2 with the fitted surface in step 8.

The above algorithm is run for two point pattern models, $\alpha_k, k = 1, 2$, and two sampling models $\boldsymbol{\theta}_r, r = 1, 2$ in step 1 with 3 fitting models $\boldsymbol{\theta}_j, j = 1, 2, 3$, see details in Section 6, giving us the 12 possibilities mentioned in Section 2.

In the data setting only the steps 1 and 2 of the algorithm are changed as follows. The sample size of the data is set as n. A point process model is fit independently to the observed locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$ to obtain the posterior distribution $p(\boldsymbol{\alpha}|\mathbf{s}_1, \ldots, \mathbf{s}_n)$. Now a value $\boldsymbol{\alpha}$ is simulated from this posterior distribution instead of the prior distribution. Similarly, a response model $\boldsymbol{\theta}$ is independently fitted to the data and subsequently $\boldsymbol{\theta}$ is drawn from this posterior distribution instead of the prior distribution. In step 2, the true response surface is obtained as the posterior predictive surface from the fitted response model.

5 Metrics for comparison of predicted surfaces

Following the previous section, we seek pairwise comparison between a posterior mean surface and a true surface. Denote the former by $\hat{\mu}(\mathbf{s}), \mathbf{s} \in D$, the latter by $\mu(\mathbf{s}), \mathbf{s} \in D$. (In fact, each will have a subscript indicating the fitting model for the former, the sampling model for the latter.) Comparison can be made globally or locally. A local metric at $\mathbf{s}_0$ makes comparison between $\hat{\mu}(\mathbf{s}_0)$ and $\mu(\mathbf{s}_0)$. A global metric will provide an integrated comparison over the $\mathbf{s}_0 \in D$.

There are many possible candidate metrics that can facilitate comparison. With interest in bias, we first define the probability of over prediction at $\mathbf{s}_0$,

$$POP(\mathbf{s}_0) = P\left(\hat{\mu}(\mathbf{s}_0) > \mu(\mathbf{s}_0)\right).$$

Evidently, $POP(\mathbf{s}_0) = .51$ is better than $POP(\mathbf{s}_0) = .52$ but also $POP(\mathbf{s}_0) = .49$ is better than $POP(\mathbf{s}_0) = .48$. In other words, $POP(\mathbf{s}_0) + PUP(\mathbf{s}_0) = 1$, where $PUP(\mathbf{s}_0)$ is the probability of under prediction at $\mathbf{s}_0$, so we need to employ $|POP(s_0) - .5|$. In fact, let's call $LPB(\mathbf{s}_0) = |P(\hat{\mu}(\mathbf{s}_0) > \mu(\mathbf{s}_0)) - .5| = |POP(\mathbf{s}_0) - .5|$ the "local probability bias" at s_0 and $\frac{1}{|D|}\int_D LPB(\mathbf{s})d\mathbf{s}$ the GPB, the "global probability bias". LPB and GPB are employed below.

Next, since our simple examples work with Gaussian models, let us use squared error for prediction error, and define the local prediction error at $\mathbf{s}_0$,

$$LPE(\mathbf{s}_0) = E(\hat{\mu}(\mathbf{s}_0) - \mu(\mathbf{s}_0))^2$$

and, thus, the global prediction error, $GPE = \frac{1}{|D|}\int_D LPE(\mathbf{s})d\mathbf{s}$. In fact, we can partition the local and global prediction errors into $\texttt{Bias}^2$ plus variance to clarify better where differences are.

The expectation and the probability calculations are performed with respect to the distribution of $\hat{\mu}(\mathbf{s}_0)$. From Section 4, under a fixed $\boldsymbol{\theta}_r$, we can use the replicates over B to obtain Monte Carlo approximations. That is,

$$\hat{POP}(\mathbf{s}_0) = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}\left(\hat{\mu}^{(b)}(\mathbf{s}_0) - \mu(\mathbf{s}_0) > 0\right)$$

where $\mathbf{1}(A) = 1$ if A is true and zero otherwise. Thus, we obtain an estimator, $\hat{LPB}(\mathbf{s}_0)$ of $LPB(\mathbf{s}_0)$. Similarly, we can use a Monte Carlo approximation to obtain $\hat{LPE}(\mathbf{s}_0)$, an estimator of $LPE(\mathbf{s}_0)$; we omit details.

We can also develop local "two-sample t-tests" at each interpolated location $\mathbf{s}_0$, using the $\hat{\mu}^{(b)}(\mathbf{s}_0)$ as the random samples. We can also record the number of significant positive differences, negative differences, and non-significances over the set of interpolated locations in the study region D. We do not pursue this path further here.

We can use the estimators, $\hat{LPB}(\mathbf{s}_0)$, $\hat{LPE}(\mathbf{s}_0)$, to compare the two sampling intensities, λ_1 and λ_2. These local metrics can also be plotted over the study region to further illuminate the effect of sampling locations on the predictive surfaces.

To obtain the global measures, GPB and GPE (in terms of $\texttt{Bias}^2$ and variance), we integrate the local measures over D using discretization. These global measures are also employed in the simulation study presented in Section 6.

As a last comment here, we may also be interested in the effect of preferential sampling on the regression model. Evidently, we can compare regression coefficients with the truth, assuming that the fitting model is the same as the true model. Moreover, as long as the fitting model is the same, we can compare coefficients arising from the different intensities. With the simple simulation models in Section 6, Regression comparison will not be very interesting. However, as a general remark, suppose we consider a single covariate. Then, the effect would depend heavily on how strong the relationship is between the response and covariate. If the relationship is strong, then what we might expect is what happens in standard regression settings. We will be observing a biased set of covariate levels rather than the full support for the covariate, which can lead to bias in the resulting estimation of the relationship.

6 Specifics of the simulation design

We present a stylized illustration using three simple models to articulate differences and effects in the clearest way, attempting to avoid confounding due to sources such as identifiability, approximation, nonlinearity, heterogeneous variances, etc. We consider the study domain, D, to be the unit square where both coordinates take values in the interval $[0, 1]$. We suppose that there is a single pollution source at a point Q with coordinates $\mathbf{q}$ (Figure 1). We also suppose that there are three cities with center locations, $\mathbf{c}_1$, $\mathbf{c}_2$ and $\mathbf{c}_3$ where the $\mathbf{c}$'s are distinct and different from $\mathbf{q}$. That is, we have multiple population centers within the study region D where the population centers are not connected with the pollution source.

With this source, the exposure surface at any location $\mathbf{s}$ in D, denoted by $x(\mathbf{s})$, is given by $x(\mathbf{s}) = \exp(-\phi_q||\mathbf{s} - \mathbf{q}||)$ where $||\cdot||$ denotes the Euclidean distance and ϕ_q is assumed to be a known positive constant. That is, exposure decays inversely to distance, with ϕ_q dictating the rate of decay. This naive exposure specification can be relaxed, but, again, we are only offering an illustrative setting.

We also define what could be viewed as an informative covariate through a *population density* surface given by $p(\mathbf{s}) = \exp(-\phi_c d_{\mathbf{s}})$ where $d_{\mathbf{s}}$ is the minimum distance between $\mathbf{s}$ and the three city center locations $\mathbf{c}_1$, $\mathbf{c}_2$ and $\mathbf{c}_3$. The parameter ϕ_c, assumed to be a known positive constant, determines the population intensity. Decreasing population density away from the city center is a customary specification though, again, our isotropic choice is naive and only illustrative. We can choose the $\mathbf{c}$'s and ϕ's to obtain very different $x(\mathbf{s})$ and $p(\mathbf{s})$ surfaces. This specification will allow us to consider pollution levels for three cities having similar population density but with varying distances from the pollution source, Q. We note that, in practice, $p(\mathbf{s})$ is often obtained from census data and is available as a tiled surface at some census unit scale.

Returning to Section 3.2, with $x(\mathbf{s})$ and $p(\mathbf{s})$ as above, we work with the following three

models:

$$\text{M1: } \mu(\mathbf{s}) = \gamma_0 + \gamma_1 x(\mathbf{s}). \quad (5)$$

$$\text{M2: } \mu(\mathbf{s}) = \beta_0 + \beta_1 p(\mathbf{s}). \quad (6)$$

$$\text{M3: } \mu(\mathbf{s}) = \mu + w(\mathbf{s}). \quad (7)$$

Intentionally, models M1 and M2 do not include the spatial random effects $w(\mathbf{s})$. They play the role of true process models and, as noted above, are used as sampling as well as fitting models. Model M3 includes spatial random effects, modeled using a mean 0 Gaussian process, as is usually done in spatial settings. As seems natural, we assume $x(\mathbf{s})$ and $p(\mathbf{s})$ are positively associated with the true surface $Y(\mathbf{s})$, i.e. $\gamma_1 > 0$ and $\beta_1 > 0$ in the above models.

Model M1 asserts that there is a single pollution source, Q, providing the exposure which is away from all three cities at a various distances (see Figure 1). Model M2 describes the case that the pollution levels are only attributable to high population density areas. Again, these two models will be used to simulate data and to fit. Model M3 offers no covariates and provides a customary model fitting situation, a model which is wrong but may be useful.

Hence, we obtain six possible combinations arising from two simulation models and three fitting models. Each of these six modeling combinations are fitted under each of two intensities for the locations. The intensity for preferential sampling (PS) would be given by $\lambda_1(\mathbf{s}, \boldsymbol{\alpha}) \propto p(\mathbf{s})$, i.e. $\log(\lambda_1(\mathbf{s}, \boldsymbol{\alpha})) = \alpha_0 + \alpha_1 \log(p(\mathbf{s}))$. In fact, we discretize the study domain D using a regular rectangular grid of $100 \times 100 = 10,000$ points and, from the $p(\mathbf{s})$ surface, calculate $p(\mathbf{s})$ at each of these points. The n locations under PS are randomly drawn without replacement from these 10,000 points where the probability of selection for $\mathbf{s}$ is proportional to the population density $p(\mathbf{s})$. Under CSR, we draw a simple random sample of size n from the 10,000 grid points covering the unit square.

We now discuss the results that we can *expect* from each of the six modeling combinations. We use the label MrMk (r=1, 2, k=1, 2, 3) to denote the case when the simulation model is Mr and the fitting model is Mk.

1. M1M1. Under PS there will be very few sampling locations near the pollution source, unlike under CSR. Hence high pollution levels will not be sampled under PS. As a result, the posterior mean of γ_1 will over estimate the true simulation value to compensate for the low observed values of the regressor $x(\mathbf{s})$ at locations away from the pollution source, Q. Hence, the predictions under the PS for locations near the point Q will be higher because of the high $x(\mathbf{s})$ values together with an upwardly biased estimate of γ_1. These high predicted values near Q will lead to both LPB and LPE being high near Q. However, these over prediction will not occur under the CSR and prediction surfaces constructed under CSR will be more accurate. See Figure 2 for a practical illustration of this result.
2. M1M2. Here the incorrect model, $\beta_0 + \beta_1 p(\mathbf{s})$, is fitted to observations from the sampling model $\gamma_0 + \gamma_1 x(\mathbf{s})$. Under PS, the response at the observed sampling locations, will be essentially uncorrelated with the associated values of the fitting regressor $p(\mathbf{s})$, so β_1 will be estimated to be near zero. Again, this would appear to be the case under CSR since $p(\mathbf{s})$ does not explain the response regardless of which $\mathbf{s}$'s we draw. It seems that fitting with an informative covariate should be of no benefit in this case and there should not be much story for us here.
3. M1M3. Here the fitted spatial model will attempt to make use of the spatial random effects as a surrogate for the true $x(\mathbf{s})$. In this regard, the random effects surface, due to its flexibility, should be able to capture the true simple linear relationship. However, it will do so through the smoothness implicit in the GP. So, information from a more representative set of sampling locations under the CSR will provide a better surface to interpolate than that obtained under PS. In other words, since we usually use random effects models when kriging, we expect to do better under CSR than under PS.
4. M2M1. Under PS, as in the M1M2 case, the γ_1 will be estimated to be near zero. Again, it would appear to be the case that, under PS or under CSR, $x(\mathbf{s})$ does not explain the

response regardless of which $\mathbf{s}$'s we draw. It seems that there should not be much story for us here.

5. M2M2. Both under the PS and CSR there will be sampling locations which have both high (at the center locations) and low (locations away from the centers) $p(\mathbf{s})$ values. Hence, we expect both PS and CSR to be indistinguishable here. That is, this case is different from the M1M1 case and is not of much interest in this study. And, again, fitting with an informative covariate would not seem to offer any benefit. Again, CSR may do better than PS with regard to prediction because of the more representative set of locations, hence covariate levels.
6. M2M3. As in the M1M3 case, better spatial interpolation under CSR will compensate for the absent covariate $p(\mathbf{s})$ and as a result, we expect to predict better under CSR than under PS.

7 Findings for a specific simulation illustration

In our illustration we continue to use the hypothetical configuration of three cities and a single pollution source as discussed above, see Figure 1. Specifically, we take $\phi_c = 5$ and $\phi_q = 1.8$ so that the two cities C1 and C2 are affected similarly by the pollution levels, and about half of the area in the city C3 is affected – providing two different regions of equal area within this city.

To simulate from the models M1 and M2, in (5) and (6) above, we need to fix values of the slope and the intercept parameter as well as the error variance σ^2_ϵ. We choose the intercept to be zero and take the slope parameter equal to 2. We also take $\sigma^2_\epsilon = 1$ and consider $n = 100$ for our illustration. A brief sensitivity study is added below. In all our implementations we take $B = 100$ simulation replications in the outer loop and $L = 5000$ posterior samples in the inner loop after discarding first 1000 iterations, see Section 4.

The values of the two measures, GPB and GPE with $\texttt{Bias}^2$ and variance, are provided in Table 1. First consider the M1M1 case. Figure 2 shows that CSR is especially better near the pollution source, as expected. Overall, the CSR improves on GPB. Also,the improvement in GPE is primarily reflected in reduced variance. Turning to the M1M2 case, we find a similar story for GPB with a smaller gain for GPE but, perhaps more than anticipated. In any event, the informative covariate has not remedied the bias introduced by PS. For the M1M3 case, we see little difference in GPB with gain in GPE attributable to the better spatial coverage resulting from CSR, hence better kriging, as suggested above. Analogues of Figure 2 are available for the second and third cases but are omitted in the interest of space.

Turning to the M2M1 case, the illustrative Figure 3 summarizes the local behavior. In particular, we see the reduction in $LPB(\mathbf{s})$ and variance at $\mathbf{s}$ around the pollution source. Returning to Table 1, the benefit for GPB and GPE is better than anticipated. For the M2M2 case, we see the ineffectiveness of using the informative covariate. It does not seem to remedy the GPB and, though it appears to mitigate GPE by helping with $\texttt{Bias}^2$, it still suffers increased variance relative to CSR. For the M2M3 case, again, we see little difference in GPB with benefit to GPE. Again, we suppress analogues of Figure 3 for these last two cases.

In summary, we conclude that the better spatial coverage associated with CSR compared with PS benefits the performance of the former in all of the cases. Particularly, for prediction, we see advantage to CSR, again, not surprising since the former provides *better* location of the sites leading to better kriging, averaged over the region.

7.1 Sensitivity Study

We offer a brief sensitivity study, summarized in Tables 2 and 3. Again, we take the intercept parameter to be equal to zero and run the experiment for two values of the slope parameter, $3, 6$, and two values of σ^2_ϵ, 0.1 and 1. Also, though we have experimented with three different values of the sample size n: 50, 100 and 300, the results are quite similar and so, we report the findings only for $n = 100$. For each of these 4 combinations of the slope, error variance we have

calculated the overall measures GPB and GPE, with $\texttt{Bias}^2$ and variance, for each of six MiMj, i=1, 2, j=1, 2, 3 modeling combinations, each under two different sampling situations, PS and CSR.

The sensitivity analysis reveals that our findings above are not sensitive to the parameter choices we adopted. In Table 2(a), when M1 is true, we see the benefit of CSR over PS with regard to GPB except in the M1M3 comparison where they are, essentially, indistinguishable. Table 2(b) shows that the consequential benefit of CSR over PS again lies in improved variance, leading to improved GPE. Reversing the roles, in Table 3(a), when M2 is true, with regard to GPB, we find gains for CSR relative to PS in all cases. Finally, Table 3(b) provides the same conclusion as Table 2(b); the improvement in CSR over PS emerges primarily through reduction in variance.

8 Illustration with a real data example

Given a real dataset, it is not possible to investigate differences between two different intensities, each providing a set of sites; we can only work with the dataset we have. Instead, we attempt to choose two subsets of monitoring locations from the real dataset such that they are of equal size but one tends to sample high values while the other tends to sample low values. In this way, we can demonstrate the magnitude of difference in the model based predictive maps when different sets of non-random sampling sites are used. In addition, we also consider the effect on prediction by reserving data from a portion of the sites for validation and comparing predictive mean square error performance for the preferential datasets compared with a dataset from roughly randomly sampled sites.

We consider a network of 175 ozone monitoring sites in California where we have observed the annual 4th highest daily maximum 8-hour average ozone concentration levels in the year 2008. The annual 4th highest daily maximum is employed since the federal standard for ozone levels is specified in terms of these values.
(http://www.epa.gov/ttn/naaqs/standards/ozone/s_o3_index.html) We reserve data from 51 randomly chosen sites for validation and consider the data from the remaining 124 sites for modeling. These 124 sites are partitioned into two data sets: one containing the data for the sites for which the observed ozone level was greater than the overall median ozone level from all the 175 sites and the other containing the remaining data. There were 56 sites in the data set with higher ozone levels and the other data set contained data from the remaining 68 sites. To have equal sample sizes this last dataset was randomly thinned to have 56 sites. A third data set was formed by randomly sampling data from 56 sites from the 124 modeling sites. Finally, as a fourth data set we consider the entire collection of 124 sites. Thus, the first data set is preferentially chosen to have sites with high ozone levels, the second data set to contain only sites with low values. The third data set has the same sample size as the first two but attempts to approximate spatial randomness.

For all four data sets we consider the simple GP model with a constant mean surface, i.e. the model (3) with $x(\mathbf{s}) = 1$ for all $\mathbf{s}$. After fitting all four datsets, we validate each of them with the set-aside data set from the 51 randomly chosen sites. The root mean square validation prediction errors for the four modeling scenarios are 22.7, 23.9, 18.0 and 18.0. respectively. This shows, as expected, that the first two preferentially chosen datasets exhibit much worse out-of-sample prediction and that there is not much difference in the predictive performances between the last two data sets. Figure 4 provides the four corresponding predictive surfaces and elaborates that the predictive surfaces arising from preferential sampling are quite different from those arising under approximate CSR sampling. This illustrates our main contention – that preferential sampling can have dramatic effect on predictions.

9 Summary and future work

This paper has developed a novel general way of comparing the effects of different intensity surfaces for sampling locations with regard to explaining response in a spatial data modeling problem. In particular, working with preferential sampling and complete spatial randomness, we have shown that predictions under the former can be much worse (particularly with regard to predictive variance) than the latter. We accomplished this by providing a simulation study that illuminates the situations in which this is expected to occur.

The proposed method can accommodate intensity comparisons under both the "with data" and "no data" cases as discussed in Section 3. The method thus can compare the predictive surfaces based on real data from a current network against data that would be obtained from a network of randomly sampled locations.

Extension to spatial sampling in a dynamic environment may be of interest. Future research can try to estimate the effect of preferential sampling under various common spatio-temporal models, see e.g. Sahu *et al.* (2010). The pertinent research question in this context is the effect of temporally varying networks on the dynamic predictive surfaces as well as their aggregates. A different extension would add a data assimilation wrinkle. Suppose we have a computer model providing exposures on the scale of grid cells. If we develop a fusion model for this output along with the monitoring station data, how will that affect the bias in prediction under preferential sampling?

A by-product of the proposed methods is the possible development of a Bayesian spatial network design selection criterion based on a bias reduction objective function together with an associated computational method for constructing the optimal design. This provides a novel method for spatial network design rather than the customary conditional variances and entropy approaches. See, e.g., Xia *et al.* (2006) and references therein. Discretization of the region D is necessary to implement any design strategy over a continuous domain. Then, the value of the design criterion can then be computed for each point in order to determine the point that gives, say the smallest criterion value, i.e., the optimal one to add. This clarifies the consequential computational burden to implement our simulation-based approach: a full simulation is needed for each candidate point.

REFERENCES

Banerjee, S., Carlin, B.P. and Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC, Boca Raton.

Banerjee, S., Gelfand, A. E., Finley, A. O. and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society*, Series B, **70**, 825-848.

Chakraborty, A., Gelfand, A. E., Wilson, A. M., Latimer, A. M., Silander, J. A. (2010). Modeling large scale species abundance with latent spatial processes. *Annals of Applied Statistics*, **4**, 1403-1429.

Cocchi, D., Greco, F. and Trivisano, C. (2007). Hierarchical space-time modelling of PM10 pollution. *Atmospheric Environment*, **41**, 532-542.

Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data.* John Wiley & Sons, Hoboken.

Dawid, P. (1984). Statistical theory: The prequential approach (with discussion). *J. Roy. Statist. Soc.* A, **147**, 278-292.

Diggle, P. J. (2003). *Statistical Analysis of Spatial Point Patterns.* Arnold Publishers, London.

Diggle, P.J., Menezes, R. and Su, T.L. (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society*, Series C, **59**, 191-232.

Gelfand, A. E., Schmidt, A. M., Wu, S., Silander, J. A., Latimer, A. M. and Rebelo, A. G. (2005). Modeling species diversity through species level hierarchical modeling. *Journal of the Royal Statistical Society*, Series C, **54**, 1-20.

Hooten, M.B., Larsen, D.R. and Wikle, C.K. (2003). Predicting the spatial distribution of ground flora on large domains using a hierarchical Bayesian model. *Landscape Ecology*, **18**, 487-502.

Illian, J., Antti, P., Stoyan, H., Stoyan, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*, Statistics in Practice. Wiley: Chichester.

Latimer, A. M., Wu, S., Gelfand, A.E. and Silander, J.A. (2006). Building statistical models to analyze species distributions. *Ecological Applications*, **16**, 33-50.

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, **174**, 109-126.

Lewis, P.A.W. and Shedler, G.S. (1979) Simulation of non-homogeneous Poisson processes by thinning. *Naval Research Logistics Quarterly,* **26**, 403-413.

Müller, W. G. (2001). *Collecting data: optimum design of experiments for random fields.* New York: Springer.

Nychka, D. and Saltzman, N. (1998). Design of air-quality monitoring networks. In *Case studies in Environmental Statistics,* ed. Nychka D, Cox L, Piegorsch W. Editors. Lecture Notes in Statistics, Springer Verlag: New York.

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

Pilz, J. and Spöck, G. (2008). Bayesian spatial sampling design. In *Proc. 8th International Geostatistics Congress* (J. M. Ortiz and X. Emery, Eds.), Gecamin Ltd., Santiago de Chile, 21-30.

Sahu, S. K., Gelfand, A. E. and Holland, D. M. (2007). High Resolution Space-Time Ozone Modeling for Assessing Trends. *Journal of the American Statistical Association*, **102**, 1221-1234.

Sahu, S. K., Gelfand, A. E. and Holland, D. M. (2010). Fusing point and areal space-time data with application to wet deposition. *Journal of the Royal Statistical Society*, Series C, **59**, 77-103.

Sheppard L., Burnett, R. T., Szpiro, A. A., Kim, S. Y., Jerrett, M., Pope, C. A. and Brunekreef, B. (2011). Confounding and exposure measurement error in air pollution epidemiology. *Air Quality, Atmosphere & Health*, DOI: 10.1007/s11869-011-0140-9, Springer.

Theobald, D.M., Stevens, D. L., Jr., D. White, N.S. Urquhart, A.R. Olsen, and J.B. Norman. (2007). Using GIS to generate spatially balanced random survey designs for natural resource applications. *Environmental Management*, **40**, 134–146.

Xia, G., Gelfand, A.E. and Miranda, M.L. (2006). Approximately Optimal Spatial Design Approaches for Environmental Health Data. *Environmetrics*, **17**, 363-385.

Disclaimer:

The U.S. Environmental Protection Agency's Office of research and Development partially collaborated in the research described here. Although it has been reviewed by EPA and approved for publication, it does not necessarily reflect the Agency's policies or views.

Table 1: Values of GPB, $\texttt{Bias}^2$, Variance and GPE under PS and CSR.

	Simulation Model M1							
	GPB		$\texttt{Bias}^2$		Variance		GPE	
Fitting Model	PS	CSR	PS	CSR	PS	CSR	PS	CSR
M1	0.041	0.020	0.239	0.222	0.079	0.036	0.318	0.258
M2	0.037	0.022	0.375	0.390	0.230	0.141	0.605	0.531
M3	0.095	0.089	0.238	0.206	0.185	0.140	0.423	0.346
	Simulation Model M2							
	GPB		$\texttt{Bias}^2$		Variance		GPE	
Fitting Model	PS	CSR	PS	CSR	PS	CSR	PS	CSR
M1	0.047	0.023	0.414	0.385	0.388	0.185	0.802	0.570
M2	0.035	0.020	0.244	0.236	0.065	0.038	0.309	0.274
M3	0.109	0.104	0.331	0.272	0.371	0.240	0.702	0.512

Table 2: Sensitivity analysis; results when the simulation model is M1 (see Section 7).

(a) Values of GPB under PS and CSR

	Fitting Model M1			
	$\gamma_1 = 3$		$\gamma_1 = 6$	
σ^2	PS	CSR	PS	CSR
0.1	0.005	0.002	0.005	0.002
1.0	0.042	0.020	0.041	0.020
	Fitting Model M2			
0.1	0.008	0.007	0.021	0.021
1.0	0.040	0.025	0.052	0.038
	Fitting Model M3			
0.1	0.044	0.040	0.086	0.079
1.0	0.114	0.117	0.222	0.239

(b) Values of $\texttt{Bias}^2$ (B), Variance (V) and their sum (GPE) under PS and CSR

	Fitting Model M1											
	$\gamma_1 = 3$						$\gamma_1 = 6$					
	PS			CSR			PS			CSR		
σ^2	B	V	B+V	B	V	B+V	B	V	B+V	B	V	B+V
0.1	0.225	0.009	0.234	0.209	0.004	0.213	0.255	0.010	0.264	0.239	0.004	0.244
1.0	0.259	0.085	0.344	0.251	0.040	0.291	0.258	0.078	0.337	0.245	0.039	0.284
	Fitting Model M2											
0.1	0.466	0.370	0.835	0.454	0.237	0.691	0.474	1.437	1.911	0.460	0.938	1.398
1.0	0.413	0.439	0.852	0.418	0.273	0.691	0.456	1.474	1.930	0.447	0.961	1.408
	Fitting Model M3											
0.1	0.221	0.112	0.333	0.172	0.057	0.229	0.218	0.276	0.494	0.162	0.108	0.270
1.0	0.263	0.290	0.553	0.216	0.194	0.410	0.251	0.610	0.861	0.195	0.361	0.556

Table 3: Sensitivity analysis; results when the simulation model is M2 (see Section 7).

(a) Values of GPB under PS and CSR.

	Fitting Model M1			
	$\beta_1 = 3$		$\beta_1 = 6$	
σ^2	PS	CSR	PS	CSR
0.1	0.016	0.009	0.050	0.029
1.0	0.052	0.027	0.084	0.047
	Fitting Model M2			
0.1	0.004	0.002	0.004	0.002
1.0	0.036	0.021	0.035	0.020
	Fitting Model M3			
0.1	0.161	0.107	0.432	0.298
1.0	0.163	0.168	0.878	0.620

(b) $\texttt{Bias}^2$ (B), Variance (V) and their sum (GPE) under PS and CSR.

	Fitting Model M1											
	$\beta_1 = 3$						$\beta_1 = 6$					
	PS			CSR			PS			CSR		
σ^2	B	V	B+V	B	V	B+V	B	V	B+V	B	V	B+V
0.1	0.472	0.692	1.164	0.452	0.339	0.791	0.476	2.741	3.216	0.457	1.347	1.804
1.0	0.442	0.745	1.186	0.414	0.373	0.787	0.467	2.815	3.282	0.445	1.370	1.815
	Fitting Model M2											
0.1	0.230	0.007	0.237	0.223	0.004	0.227	0.248	0.008	0.256	0.245	0.005	0.250
1.0	0.266	0.080	0.346	0.250	0.041	0.290	0.258	0.070	0.328	0.239	0.038	0.278
	Fitting Model M3											
0.1	0.232	0.341	0.573	0.203	0.187	0.39	0.220	1.016	1.237	0.182	0.508	0.690
1.0	0.351	0.685	1.036	0.277	0.413	0.69	0.254	1.860	2.113	0.212	1.056	1.268

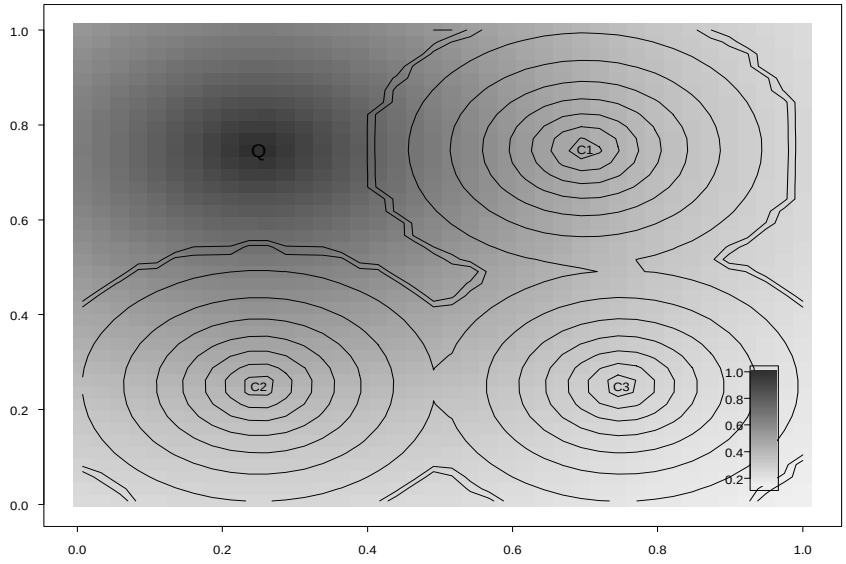

Figure 1: A simulated pollution map of the study region. It also shows the location of the point pollution source, Q, and the three city centers, C1-C3. The contours represent levels of the scaled population density.

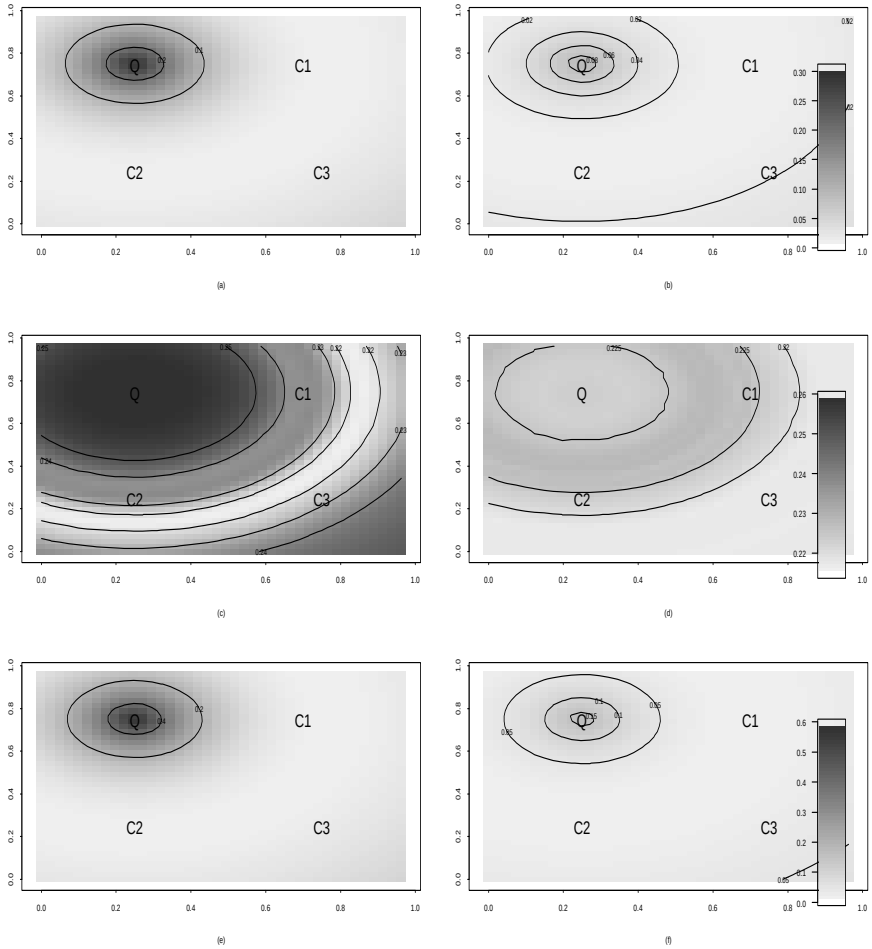

Figure 2: GPB, $\texttt{Bias}^2$ and Variance surfaces for the case M1M1. Panel (a) is the *GPB* surface under PS while Panel (b) is the same under CSR. Panel (c) is the $\texttt{Bias}^2$ surface under PS and Panel (d) is the same under the CSR. Panel (e) is the Variance surface under PS and Panel (f) is the same under the CSR.

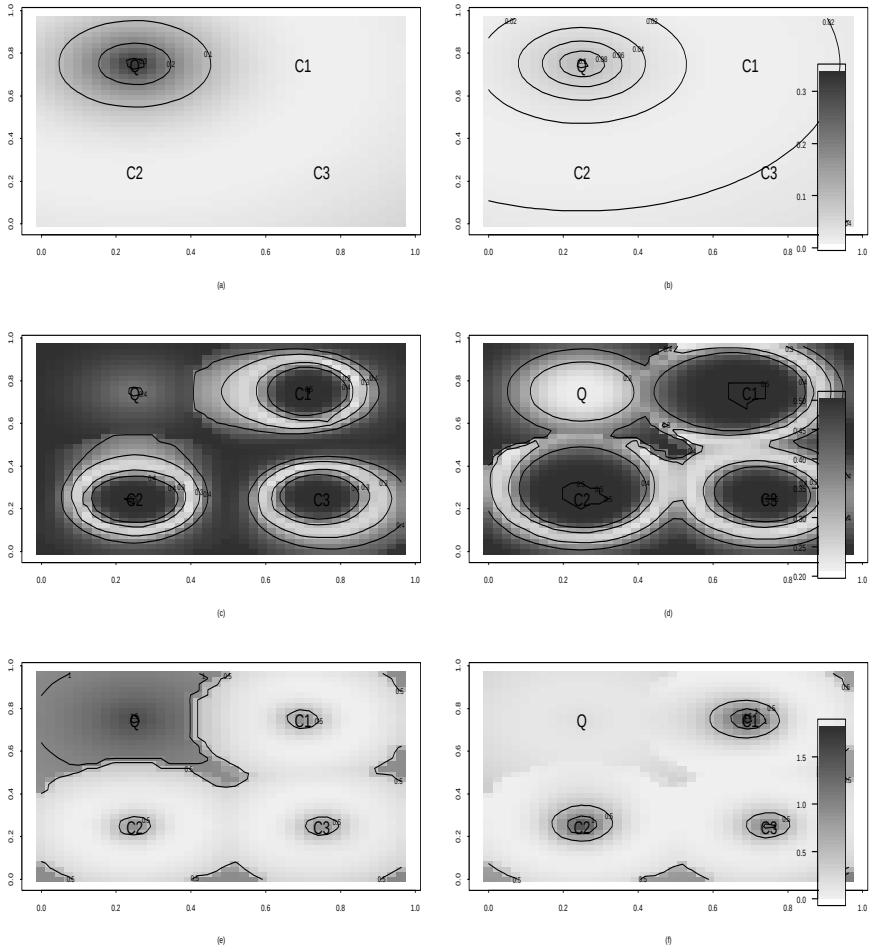

Figure 3: GPB, $\texttt{Bias}^2$ and Variance surfaces for the case M2M1. Panel (a) is the *GPB* surface under PS while Panel (b) is the same under CSR. Panel (c) is the $\texttt{Bias}^2$ surface under PS and Panel (d) is the same under the CSR. Panel (e) is the Variance surface under PS and Panel (f) is the same under the CSR.

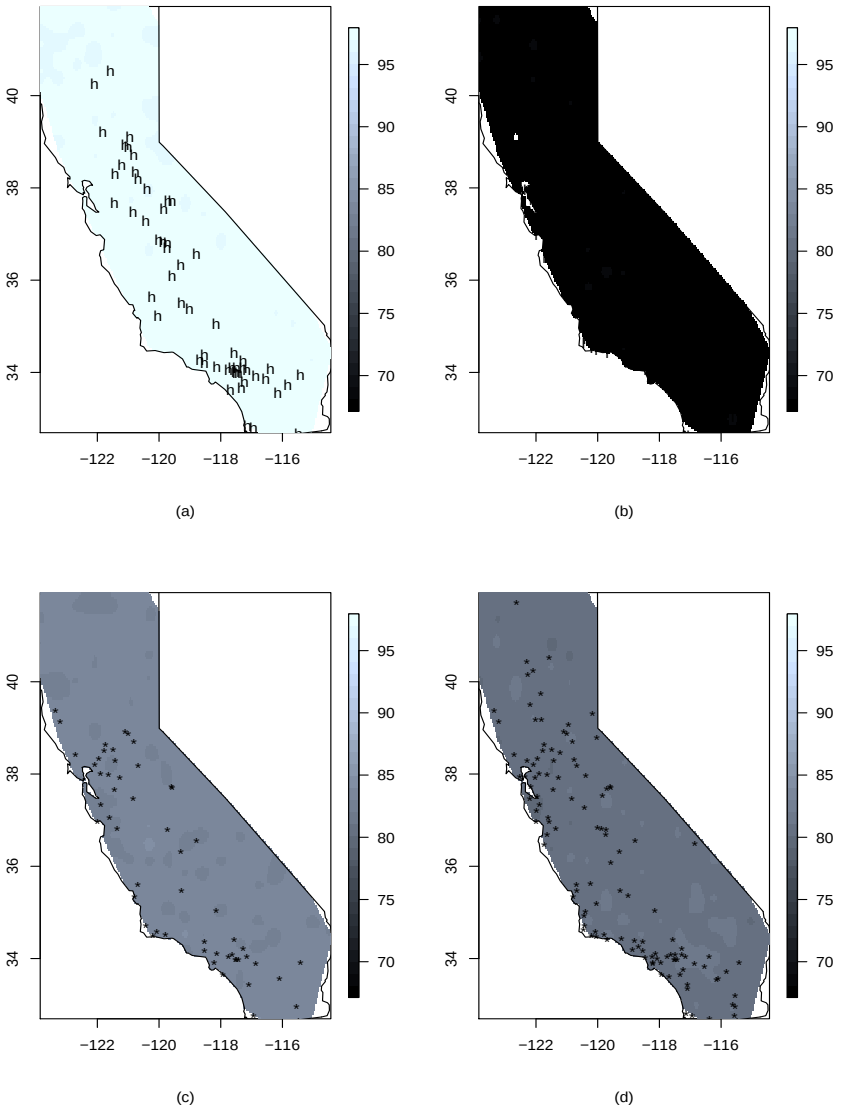

Figure 4: For the real data (Section 8), predictive maps of the annual 4th highest daily maximum 8-hour average ozone levels based on data from: (a) sites (denoted by h) with high ozone values, (b) sites with low values, (c) 56 randomly chosen sites, and (d) all the 124 modeling sites.