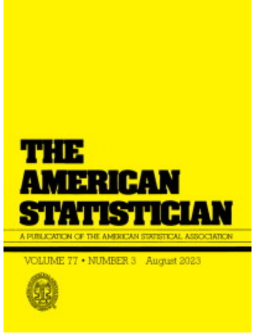

THE AMERICAN STATISTICIAN

A PUBLICATION OF THE AMERICAN STATISTICAL ASSOCIATION

VOLUME 77 • NUMBER 3 August 2023

Taylor & Francis
Taylor & Francis Group

The American Statistician

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/utas20

A Look into the Problem of Preferential Sampling through the Lens of Survey Statistics

Daniel Vedensky, Paul A. Parker & Scott H. Holan

THE AMERICAN STATISTICIAN
2023, VOL. 77, NO. 3, 313–322: Statistical Practice
https://doi.org/10.1080/00031305.2022.2143898

Taylor & Francis
Taylor & Francis Group

A Look into the Problem of Preferential Sampling through the Lens of Survey Statistics

Daniel Vedensky[a], Paul A. Parker[b], and Scott H. Holan[a,c]

[a]Department of Statistics, University of Missouri, Columbia, MO; [b]Department of Statistics, University of California Santa Cruz, Santa Cruz, CA; [c]U.S. Census Bureau, Washington, D.C.

ABSTRACT

An evolving problem in the field of spatial and ecological statistics is that of preferential sampling, where biases may be present due to a relationship between sample data locations and a response of interest. This field of research bears a striking resemblance to the longstanding problem of informative sampling within survey methodology, although with some important distinctions. With the goal of promoting collaborative effort within and between these two problem domains, we make comparisons and contrasts between the two problem statements. Specifically, we review many of the solutions available to address each of these problems, noting the important differences in modeling techniques. Additionally, we construct a series of simulation studies to examine some of the methods available for preferential sampling, as well as a comparison analyzing heavy metal biomonitoring data.

ARTICLE HISTORY
Received March 2022
Accepted October 2022

KEYWORDS
Bias correction; Informative sampling; Official statistics; Pseudo-likelihood; Spatial statistics

1. Introduction

Rooted in survey methodology, issues surrounding *informative sampling* (IS) have experienced extensive research in recent years. The problem arises when the probability of selecting a unit to the sample is correlated with the response (Pfeffermann and Sverchkov 2007). If left unaddressed when specifying a statistical model, this issue may introduce substantial biases. Although many solutions have been proposed for this problem, it remains an extremely active area of ongoing research. One reason for the continued interest is that official statistical agencies disseminate tabulations from key surveys, such as the American Community Survey, that support the distribution of billions of dollars in funds annually (*https://www.census.gov/programs-surveys/acs/about.html*). Consequently, adequate modeling of survey data, that properly accounts for the survey design, is increasingly important.

In contrast, more recently, the problem of *preferential sampling* (PS) has been studied in the context of ecological and spatial statistics. Specifically, PS arises when there is dependence between the process or response that is being modeled and the process that gives rise to the data locations (Diggle, Menezes, and Su 2010). Again, it has been observed that substantial bias may be introduced when PS is unaccounted for in the model. There are many important applications that use PS data, ranging from species distribution modeling (Pennino et al. 2019), to pollution monitoring (Zidek, Shaddick, and Taylor 2014), to sports analytics (Jiao, Hu, and Yan 2019). Diggle, Menezes, and Su (2010) introduced the foundational model for PS through a shared process for the data locations and responses. Outside of model development, Dinsdale and Salibian-Barrera (2019) propose alternative Monte Carlo estimates to those used by Diggle, Menezes, and Su (2010), whereas da Silva Ferreira and Gamerman (2015) consider the problem of optimal sampling design under effects of PS. Finally, Watson (2021) develops a test to detect PS.

These two problems have very similar definitions, yet there are some key differences. First, the sample domain is different within each problem. PS occurs in geostatistical settings, where locations are sampled from a continuous, often multi-dimensional, domain. Conversely, surveys sample individual units (e.g., people, households, establishements, etc.) from a finite population. Another important distinction is the scope of the data included in a sample. Survey data is most often accompanied by a survey weight that is inversely proportional to the unit probability of selection. These selection probabilities are known for all sampled units based on the survey design and sampling frame. However, geostatistical data are rarely accompanied by probabilities of selection. Much of the literature around PS was developed for ecological applications. However, in some cases, data for ecological applications may come from a known sample design (e.g., see, Irvine et al. 2018).

There is a vast literature surrounding various methodologies that may be used to account for IS with unit-level survey data. Parker, Janicki, and Holan (2019) give an overview of these potential approaches. One popular solution is to use an exponentially weighted pseudo-likelihood (Binder 1983; Skinner 1989). Other approaches include nonlinear regression on

CONTACT Daniel Vedensky ✉ *dvedensky@mail.missouri.edu* Department of Statistics, University of Missouri, 146 Middlebush Hall, Columbia, MO 65211-6100.

Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/TAS*.

the survey weights (Si, Pillai, and Gelman 2015) and inferring a population level model through specification of a model for the weights (Pfeffermann and Sverchkov 2007). Each of these methods relies on the reported survey weights to account for the survey design.

Despite the similarities between these two problems, the development of methodology for PS has happened mostly independently from the IS literature. This is likely due to the fact that IS solutions rely on reported survey weights, which are not typically available in PS applications. The most common approach to handling PS is to model the sampled locations as a point process and then use a shared latent process for the point process model and the response model (Diggle, Menezes, and Su 2010; Pati, Reich, and Dunson 2011). However, Zidek, Shaddick, and Taylor (2014) take an approach that is similar to the pseudo-likelihood method that is often used in the IS literature. Their solution is to exponentially weight the likelihood by the inverse of the estimated probabilities of selection. This indicates that although the comparison has not been explored explicitly, it is likely that researchers in the PS field have, to some extent, recognized the similarity to IS.

Our goal in this article is to make an explicit connection between PS and IS. We argue that despite some differences in the problem statements, these are both manifestations of a general problem for which similar methodologies may be used. It is our hope that this will help to foster future research both within and across these areas.

The remainder of this article proceeds as follows. In Section 2 we introduce notation and give formal problem statements for both IS and PS and classify solutions to these problems into three general groups. Then, in Section 3, we explore methods for both PS and IS that rely on the use of covariates to adjust for the sampling design. In Section 4 we investigate methods that use nonlinear regression on the selection probabilities, while in Section 5 we explore exponentially weighted likelihood approaches. We conduct a simulation study to compare a subset of methods for PS in Section 6. Note that we do not conduct simulations for IS, as this was previously done by Parker, Janicki, and Holan (2019). We also illustrate a subset of methods for PS in Section 7 using the heavy metal biomonitoring data discussed in Diggle, Menezes, and Su (2010). Finally, we provide concluding discussion in Section 8.

2. Problem Statements

In order to describe the similarity between IS and PS, while still recognizing the key differences, we use a similar although slightly different notation for each problem. In general, we represent the response data, or observations, with Z. Under IS problems, these data are sampled from a finite population, $\mathcal{U} = 1, \ldots, N$. The sample, of size n, is denoted $\mathcal{S} \subset \mathcal{U}$. Typically, survey units are observed in one of $d = 1, \ldots, D$ geographic areas or domains (e.g., census tract, county, etc.). Thus, we index the data for a given observational unit in domain d as Z_{jd}, where $j \in \mathcal{S}$. In contrast to this, preferentially sampled data are typically sampled from a continuous domain, $\mathcal{D}$ (often $\mathcal{D} \subset \mathcal{R}^2$). We denote these spatially referenced data as $Z(s)$, where $s \in \mathcal{D}$.

We also define a latent process as either Y_{jd} in the context of survey data or $Y(s)$ in the context of data sampled from a continuous spatial domain. Defining a latent process model has become fairly standard in the context of hierarchical modeling (Cressie and Wikle 2011; Banerjee, Carlin, and Gelfand 2015) and facilitates conditional model specification. In many cases this aids in model development and estimation. The latent process may be a function of a vector of covariates ($\boldsymbol{x}_{jd}$ or $\boldsymbol{x}(s)$) as well as spatially correlated random effects ($\boldsymbol{\eta}_d$ or $\boldsymbol{\eta}(s)$). Note that in the context of survey data, units are nested within geographic areas and, thus, can be mapped to their corresponding area-level random effects through the use of incidence vectors.

One unique aspect of survey datasets is that they are often accompanied by unit survey weights, w_{jd}. These weights are typically assumed to be the inverse probabilities of selection, $w_{jd} = 1/p_{jd}$; however, they may include adjustments for various reasons such as nonresponse. The selection probabilities, and thus the weights, are usually known from the survey design. When data are sampled from a continuous domain, as is the case with most PS problems, there is no probability of selection. However, for a Poisson point process, the intensity is defined as the limit of the probability of observing a point in a decreasing area, and thus serves as a natural analogy. In some cases, we will construct weights for these continuously sampled data, $w(s) = 1/p(s)$, where $p(s)$ is the underlying intensity function evaluated at $s \in \mathcal{D}$. We note that in this case the intensity must usually be estimated, as it is not defined by a known sampling design. In some cases, it may be desirable to consider a discretized space (e.g., areal or lattice spatial domains). In this case, the population may be viewed as a finite collection of areas, for which a true probability of selection may be known, yielding a setup that is strikingly similar to those considered in the case of survey data. However, for the most part, our discussion here will be concerned with the case of data sampled from a continuous spatial domain.

For geostatistical data, we work with the likelihood

$$f(\boldsymbol{Z}, s) = \int f(\boldsymbol{Z}, s, \boldsymbol{Y}) d\boldsymbol{Y}.$$

In most cases, we assume that the distribution of the locations is independent of the process and responses. This is the case under uniform random sampling, and results in the likelihood

$$\int f(\boldsymbol{Z}, s, \boldsymbol{Y}) d\boldsymbol{Y} = \int f(\boldsymbol{Z}|\boldsymbol{Y}, s) f(s) f(\boldsymbol{Y}) d\boldsymbol{Y}$$
$$\propto \int f(\boldsymbol{Z}|\boldsymbol{Y}, s) f(\boldsymbol{Y}) d\boldsymbol{Y}.$$

Thus, when the data locations are independent from the process and response values, the location model can be ignored. However, when independence is not met (i.e., in the case of PS) either the factorization $\int f(\boldsymbol{Z}|s, \boldsymbol{Y}) f(s|\boldsymbol{Y}) f(\boldsymbol{Y}) d\boldsymbol{Y}$ or $\int f(\boldsymbol{Z}|s, \boldsymbol{Y}) f(\boldsymbol{Y}|s) f(s) d\boldsymbol{Y}$ must be used.

Similarly, for survey samples, when the distribution of sample inclusion indicators is independent of the response values and any latent process, the survey design may be ignored. In the case of surveys that exhibit IS, the survey design must be considered in the model.

3. Use of Informative Covariates in the Model

One of the most basic ways to correct for an informative design in the survey setting is to include all of the design variables in the model, typically as covariates. In this case, the response, conditioned on the covariates, is independent of the selection probabilities and inference can be based on this conditional distribution. For instance, including a fixed effect for strata could correct for a stratified design. Little (2012) outlines such an approach in a Bayesian setting, in which case it is possible to calculate a posterior predictive distribution for the unsampled population. This predictive distribution can then be used for inference on population quantities of interest.

Another corrective measure from the survey literature that relies on the use of design variables is post-stratification (Little 1993). This approach assumes the population contains C categories, or poststratification cells, each with a known population size N_c, $c = 1, \dots, C$. Observations within each cell are assumed to be independent and identically distributed. These categories are generally determined by cross-classifications of levels of categorical covariates or of continuous covariates that have been discretized. For example, a post-stratified estimator for the population mean, $\bar{z}_p$, could be calculated as

$$\bar{z}_P = \sum_{c=1}^{C} \frac{N_c}{N} \bar{z}_{cS},$$

where $\bar{z}_{cS}$ denotes the observed mean response for sampled units in cell c.

Gelman and Little (1997) and Park, Gelman, and Bafumi (2006) combine poststratification with a Bayesian multi-level model, which allows for parameter estimates of cells with no sampled units. For example, with binary data a model of the following form could be used

$$\begin{aligned} z_{jd}|p_{jd} &\sim \text{Bernoulli}(p_{jd}) \\ \text{logit}(p_{jd}) &= \boldsymbol{x}'_{jd}\boldsymbol{\beta} \\ \boldsymbol{\beta} &= (\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_\ell) \\ \boldsymbol{\gamma}_\ell &\overset{\text{ind}}{\sim} N_{m_\ell}(0, \sigma^2_\ell \boldsymbol{I}_{m_\ell}), \ell = 1, \dots, L, \end{aligned}$$

where $\boldsymbol{x}_{jd}$ is a vector of dummy variables for L categorical predictor variables with m_ℓ classes in variable ℓ. Bayesian inference is performed on this model to get a probability p_c for each cell, $c = 1, \dots, C$. The number of positive responses within cell c is estimated as $N_c p_c$, and any higher level aggregate estimates can be made by aggregating the corresponding cells. Importantly, only first order covariate effects are modeled, without considering interactions. For example, if race and age category were two available covariates, some combinations of age and race may have no data points. In this case, the model may still be estimable as long as there are some observations within each race category as well as some observations within each age category.

A major impediment to using these design-variable based approaches is the fact that all design variables are rarely known. For example, for data users outside of a statistical agency, full design knowledge may not be attainable, and even within a statistical agency, one may not have access to all design variables. In some cases, where all design variables are available, their inclusion may complicate the likelihood, or even make it intractable.

In the PS case, there may not be a formal sampling design and therefore no design variables in the IS sense. However, Gelfand, Sahu, and Holland (2012) introduce the analogous notion of an "informative covariate"— a covariate that is correlated with both the response values and the choice of sampling locations. They take the example of including population density as a covariate in a model that predicts pollution levels. They carry out a simulation study that, in part, assesses how effectively the use of such an informative covariate may reduce bias due to PS. In fact, they find it does little to remedy the bias, let alone make up for the difference between a completely randomized sampling scheme and a PS scheme.

Conn, Thorson, and Johnson (2017) also examine the effect of including informative covariates in models for preferentially sampled data, specifically in the context of animal population models. They, too, conclude that this approach is often inadequate since predictive covariates explain only a small portion of variation present in the data in many contexts. Often, the factor driving location selection may not even be known. However, they do point out that for some sampling designs, there is theoretical justification for collecting more samples in areas where the response is expected to be higher. In this case, sampling according to a covariate and then including it in the model will lead to an ignorable, nonpreferential design. They mention post-stratification as a potential way to correct for bias, as well, with the caveat that it may not be clear how to do so when effort is allocated in a subjective manner. Unlike the IS setting, it is also unrealistic to expect knowledge of the population poststratification cell sizes, N_c, in the PS case.

4. Regressing on the Selection Process

Another common and conceptually straightforward approach to alleviating bias in IS or PS situations is to directly specify the dependence relationship between the response and sample selection process in the likelihood.

In the field of survey methodology, for example, there is a long history of use of the data model

$$Z_{jd} = \boldsymbol{x}'_{jd}\boldsymbol{\beta} + g(w_{jd}) + \epsilon_{jd}, \tag{1}$$

which attempts to directly adjust for IS through estimation of the function $g(\cdot)$. In the most simple case, Firth and Bennett (1998) consider the linear function $g(w_{jd}) = a \cdot w_{jd}$. In practice, the assumption of linearity can be quite restricting, and thus, Zheng and Little (2003) explore the use of nonlinear models for $g(\cdot)$ via penalized splines. We note that survey data is frequently used for the purpose of small area estimation, in which the area population totals $\sum_{j \in d} g(w_{jd})$ are required to construct estimates. In the case of linear $g(\cdot)$, this is not necessarily restrictive (e.g., the weights should sum to the population count, which is typically known); however, nonlinear specifications of $g(\cdot)$ should be carefully considered.

Si, Pillai, and Gelman (2015) use a flexible Gaussian process prior for $g(\cdot)$. They take a unique approach to estimation of population totals by defining poststratification cells according to the unique survey weight values. Through a multinomial data model for the observed cell sizes, they are able to generate predictions of the survey weights for all individuals outside of the sample. Vandendijck et al. (2016) extend this approach to the small area estimation setting by defining poststratification cells as unique combinations of weight and geographic area.

Unlike survey data settings, geostatistical data is not usually accompanied by known sample weights or probabilities. This can introduce further challenges when attempting to model the sampling dependence directly.

Diggle, Menezes, and Su (2010) introduce a foundational approach to handling preferentially sampled data. They assume that the data locations follow a log-Gaussian Cox process (Møller, Syversveen, and Waagepetersen 1998). That is, conditional on an underlying, unobserved, Gaussian process, $Y(s)$, the data locations follow a Poisson point process with intensity

$$p(s) = \exp\{\alpha + Y(s)\beta\}.$$

The response values are simultaneously modeled as

$$Z(s) = \mu + Y(s) + \epsilon(s),$$

where $\epsilon(s)$ is independent and identically distributed. Importantly, the latent process $Y(s)$ is shared between the location model and the response model to account for preferentially sampled data. Through simulation and analysis of lead pollution data in Galicia, Spain, they show that failure to account for PS leads to substantial prediction bias and underestimated standard errors. Additionally, they show that their approach is able to reduce this bias.

Pati, Reich, and Dunson (2011) take a very similar approach within a Bayesian framework. Again, they model the data locations as a log-Gaussian Cox process with intensity

$$p(s) = \exp\{Y(s)\},$$

conditional on the latent Gaussian process, $Y(s)$. In doing so, the response model is

$$Z(s) = \eta(s) + Y(s)\beta + \epsilon(s), \tag{2}$$

where $\eta(s)$ is an additional spatially correlated Gaussian process that adds flexibility, and again, $\epsilon(s)$ is independent and identically distributed noise. Note that $\eta(s)$ can be defined to include covariate information in the mean structure. In this framework, β controls the level of PS in the data, where $\beta = 0$ indicates no presence of PS. After conditioning on $Y(s)$ and noting that the intensity, $p(s)$ is analogous to the sample selection probability in the discrete case, (2) becomes reminiscent of the IS Model (1), with $g(w(s)) = \log\left(\frac{1}{w(s)}\right)\beta$. Both models take the approach of regressing on a function of the selection process in order to account for selection bias. Grantham et al. (2018) embed this approach into a deeper hierarchical model in order to account for informative missingness of geostatistical data.

Both Diggle, Menezes, and Su (2010) and Pati, Reich, and Dunson (2011) only consider the case of a continuous (Gaussian) response variable over a continuous spatial domain. In an effort to expand the applicability of these methods, Conn, Thorson, and Johnson (2017) extend to the case of count data on a discrete (areal) domain, as well as the case where only some parts of the domain are sampled preferentially. For the discrete domain, data locations are observed over a finite grid space, eliminating the need for the point process model. This results in a true probability of selection, although one that is still typically unobserved in practice. Similarly, Gelfand and Shirota (2019) consider the case of presence/absence as well as presence only data through a shared process method and Pennino et al. (2019) consider abundance data under PS schemes.

The general approach of regressing on the selection process through a shared latent process has dominated much of the literature in PS. However, this general approach makes up comparatively less of the literature in the IS world. This may be in part due to the fact that survey data under informative sample designs are typically accompanied by a survey weight. Considering the weights as fixed and known could allow for a broader class of methods than the case where weights or selection probabilities must be modeled, as is the case with the shared process models discussed herein.

One limitation of (2) is the assumption of linearity between $Z(s)$ and $Y(s)$. In scenarios where data is frequently observed in locations with both high and low expected response values, this linear assumption is flawed. Yet, estimation of a nonlinear function $g(Y(s))$ can be challenging when $Y(s)$ is a stochastic process itself. In contrast, the survey literature frequently considers nonlinear $g(w_{jd})$, allowable in part due to the assumption of fixed and known survey weights.

5. Weighted Likelihood Adjustments

Another common approach to dealing with informatively sampled data in the survey realm is to incorporate the survey weights into the likelihood to get a "pseudo-likelihood" of the form

$$\prod_{j\in\mathcal{S},d\in\mathcal{D}} p(z_{jd}|\boldsymbol{\theta})^{w_{jd}},$$

where, as before, the weights w_{jd} are inversely proportional to the probability of selection. Inference is performed by solving the corresponding estimating equations

$$\sum_{j\in\mathcal{S},d\in\mathcal{D}} w_{jd}\frac{\partial}{\partial\boldsymbol{\theta}}\log p(z_{jd}|\boldsymbol{\theta}) = 0 \tag{3}$$

and leads to design-consistent estimation of $\boldsymbol{\theta}$.

This strategy was introduced by Binder (1983) and Skinner (1989). As Parker, Janicki, and Holan (2019) detail, much recent work has been done in the survey literature to extend the pseudo-likelihood approach. These developments allow for the addition of random effects, the use of hierarchical models, and the use of Bayesian inference.

In particular, Savitsky and Toth (2016) show that, pseudo-likelihoods can reasonably be used in a Bayesian context. Given a prior distribution, $\pi(\boldsymbol{\theta})$, over $\boldsymbol{\theta}$, they prove that L_1 consistency is guaranteed for a pseudo-posterior of the form

$$\widehat{\pi}(\boldsymbol{\theta}|z_{jd}, w) \propto \left[\prod_{j\in\mathcal{S},d\in\mathcal{D}} p(z_{jd}|\boldsymbol{\theta})^{\widetilde{w}_{jd}}\right]\pi(\boldsymbol{\theta})$$

for certain survey designs. In this case, the survey weights are normalized to sum to the sample size $\widetilde{w}_{jd} = \frac{w_{jd}}{\sum w_{jd}/n}$, so that the influence of each weight is on the order of the information in the sample. They note that other formulations are possible, such as including a prior for the weights, or modeling them jointly with the response, but the "plug-in" approach is the simplest and performs quite well. More recently, Williams and Savitsky (2020) have extended this method to an even wider class of sampling designs.

Pseudo-likelihoods have not been as widely adopted in the PS literature as the latent process approach of Diggle, Menezes, and Su (2010), described in Section 4. However, Zidek, Shaddick, and Taylor (2014) implement a frequentist version of the pseudo-likelihood approach in the context of air quality monitoring. They analyze time series data for black smoke pollution in the United Kingdom, where the choice of sampling sites changed preferentially over several decades. Their approach to accounting for PS draws on the "response-biased sampling" literature, such as Scott and Wild (2011), as well as design-based survey inference.

Given this official statistics perspective, they place greater emphasis on producing unbiased estimates compared to the geostatistical techniques outlined above, which focus more on prediction. They further diverge from the geostatistical setup by assuming a finite population of possible sampling locations rather than a continuous domain. This formulation is more in line with the typical survey setting, which assumes a finite population. They also modify the estimating equations in (3) to allow for covariates.

As before, an obvious hurdle to translating any IS model to the PS setting is a lack of known weights in the latter case. To handle this, Zidek, Shaddick, and Taylor (2014) use logistic regression to estimate the probability of selection for each site in the domain at each time point, t based on the sample at time $t-1$, with some differences depending on whether the set of monitoring locations increases or decreases over time.

While the use of weighted likelihoods remains far less prevalent in the PS literature, the application in Zidek, Shaddick, and Taylor (2014) suggests that many of the relevant developments in the survey literature could well be carried over to the PS problem. In fact, more recently, Schliep, Wikle, and Daw (2021) have studied the use of weighted composite likelihoods in the context of spatial kriging under biased sampling schemes.

6. Simulations

To assess some of the approaches for handling PS data outlined so far, we carry out a set of simulation studies. We consider two ways of simulating PS data and compare the performance of each method in accounting for the preferential sample.

6.1. Scenario 1: Spatially Implicit Scenario

In the first scenario, we begin by generating 1000 candidate points, $\mathbf{s}_i = (s_{1i}, s_{2i})'$, for $i = 1, \ldots, 1000$, uniformly over the unit square $\mathcal{D} = [0,1] \times [0,1]$. To take a preferential sample, we thin these candidate points with probability proportional to a function of the response at each point. We keep each point with probability

$$p(\mathbf{s}_i) = (1 - (s_{1i} - 0.5)^2 - (s_{2i} - 0.5)^2)^8. \qquad (4)$$

This selects points closer to the center of the domain with higher probability than those near the edges of the domain. We then generate values of a response, z_i, from the model

$$z(\mathbf{s}_i) = (5,2)'\mathbf{s}_i + 2\widetilde{p}(\mathbf{s}_i) + \varepsilon_i,$$

where $\varepsilon_i \sim N(0, .5)$ and $\widetilde{p}(\mathbf{s}_i)$ is probability of selection as defined in (4), now centered and scaled. Including the probability of selection in the response introduces a dependency between the sampling scheme and the response.

1. As a baseline, we fit a Bayesian linear regression model that does not attempt to correct for the PS scheme

$$\begin{aligned} \mathbf{Z}|\boldsymbol{\beta}, \sigma_z &\propto \prod_{i \in \mathcal{S}} N(z_i | \boldsymbol{s}_i'\boldsymbol{\beta}, \sigma_z) \\ \boldsymbol{\beta} &\sim N_2(0, \sigma_\beta^2 \mathbf{I}_2) \\ \sigma_z &\sim \text{Cauchy}^+(0, 10), \end{aligned}$$

   where $\mathbf{s}_i$ is the two-dimensional coordinate vector for the ith sample point and $\boldsymbol{\beta}$ is a vector of their associated regression coefficients. For our simulations, we set $\sigma_\beta^2 = 10$. This can be seen as a pseudo-likelihood model with unit weights, hence, we refer to it as "unweighted" (UW).
2. We then fit a set of weighted pseudo-likelihood models similar to those outlined in Section 5. These are specified identically to the UW model, but with the addition of scaled weights, $\widetilde{w}_i = \widetilde{w}(\mathbf{s}_i)$, that correct for the underlying selection probability, so that

$$\begin{aligned} \mathbf{Z}|\boldsymbol{\beta}, \sigma_z &\propto \prod_{i \in \mathcal{S}} [N(z_i | \boldsymbol{s}_i'\boldsymbol{\beta}, \sigma_z)]^{\widetilde{w}_i} \\ \boldsymbol{\beta} &\sim N_2(0, \sigma_\beta^2 \mathbf{I}_2) \\ \sigma_z &\sim \text{Cauchy}^+(0, 10). \end{aligned}$$

   As before, we set $\sigma_\beta^2 = 10$ in our simulations.

   We compare the performance of two different schemes for defining the weights. First, as a baseline for this scenario, we take the true, known probability of selection, $p(\mathbf{s}_i)$ at each sample point $\mathbf{s}_i$ and set $w_i = w(\mathbf{s}_i) = 1/p(\mathbf{s}_i)$, then rescale so that these weights sum to the sample size. We refer to this model as the pseudo-likelihood known weights (PKW) model.

   In practice, the true weights are not typically known, so for a comparison that might actually be used in practice, we obtain a kernel density estimate of the probability of selection using a Gaussian kernel via the `MASS` package with default bandwidth (Venables and Ripley 2002). That is, using the observed locations, $\boldsymbol{s}_i$, we construct a kernel density estimate of the sampling locations over the spatial domain. Then, for any given location, $\boldsymbol{s}_i$, we take the reciprocal of the kernel density estimate at the location to be the weight (i.e., $w(\boldsymbol{s}_i) = 1/\widehat{p}(\boldsymbol{s}_i)$, where $\widehat{p}(\boldsymbol{s}_i)$ is the kernel density estimate at location $\boldsymbol{s}_i$). As before, we rescale these weights so that they sum to the sample size. We refer to this model as the pseudo-likelihood, estimated weights (PEW) model.
3. Lastly, we fit a hierarchical model as defined in Pati, Reich, and Dunson (2011) and described in Section 4, where a latent GP is shared between the response and the point process. We refer to this model as the PRD model. Fitting this model

Table 1. Average of parameter estimates, as well as 90% CI coverage probabilities and average CI widths, for each of the four models in the spatially implicit scenario simulation.

	$\widehat{\beta_1}$			$\widehat{\beta_2}$		
Model	Mean	CI coverage	Mean CI width	Mean	CI coverage	Mean CI width
UW	6.408	2%	1.311	3.520	0%	1.314
PEW	5.278	73%	1.032	2.258	83%	1.028
PKW	5.101	61%	0.983	1.989	86%	0.976
PRD	4.604	89%	2.174	1.676	98%	2.167

NOTE: The true parameter values are $\beta_1 = 5$ and $\beta_2 = 2$.

Table 2. MSE and mean absolute bias for the posterior mean predictive surface produced by each of the four models in the spatially implicit scenario simulation.

Model	MSE	Mean Abs. Bias
UW	2.551	1.464
PEW	0.121	0.268
PKW	0.081	0.046
PRD	0.380	0.360

requires discretizing the domain, and we follow the authors' recommendations for an equally spaced grid of 225 knots on $[-0.2, 1.2]^2$ over a square grid of 41×41 points, which ensures the grid spacings are chosen to be no larger than the standard deviation of the kernel in the convolution representation.

The UW, PKW, and PEW models are fit using Hamiltonian Monte Carlo via Stan. Each of these models is run for 5500 iterations with the first 1000 iterations discarded for burn-in. The PRD model is estimated using 60,000 iterations with 10,000 iterations discarded for burn-in. We repeat each of the simulations 100 times with an initial candidate set size of 1000 locations each time. Visual inspection of the trace plots as well as effective sample size of the sample chains indicated no lack of convergence for any of the models.

The mean estimate, credible interval (CI) coverage rate, and mean CI width for each parameter are reported in Table 1. Overall MSE and bias for the predicted surfaces are reported in Table 2 and posterior mean surface plots are provided in Figure 1. In terms of parameter estimates, all three PS models are able to greatly reduce the bias compared to the UW model, with the PL approaches performing best in this case. The unweighted model in particular overestimates both regression coefficients, resulting in a predicted surface that differs significantly from the truth. The other models are able to reduce this bias to various degrees, resulting in plots that align more with the true surface. In terms of accuracy of uncertainty estimates, as assessed by the CI coverage rate, again the PS models are able to greatly improve upon the UW model. However, for β_1 specifically, the PRD model has coverage much closer to the nominal level than the PL models. In particular, for β_1, we see that for PKW the bias is the smallest among the methods considered, with slight underestimation of uncertainty. This issue may be similar to what has been observed in the context of IS when comparing results arising from plugging in weights versus estimated weights (e.g., see León-Novelo and Savitsky 2019; Williams and Savitsky 2021). Looking at predictive ability (Table 2), the unweighted model grossly underperforms each of the corrective models and shows substantial bias. The corrective models all show greatly reduced

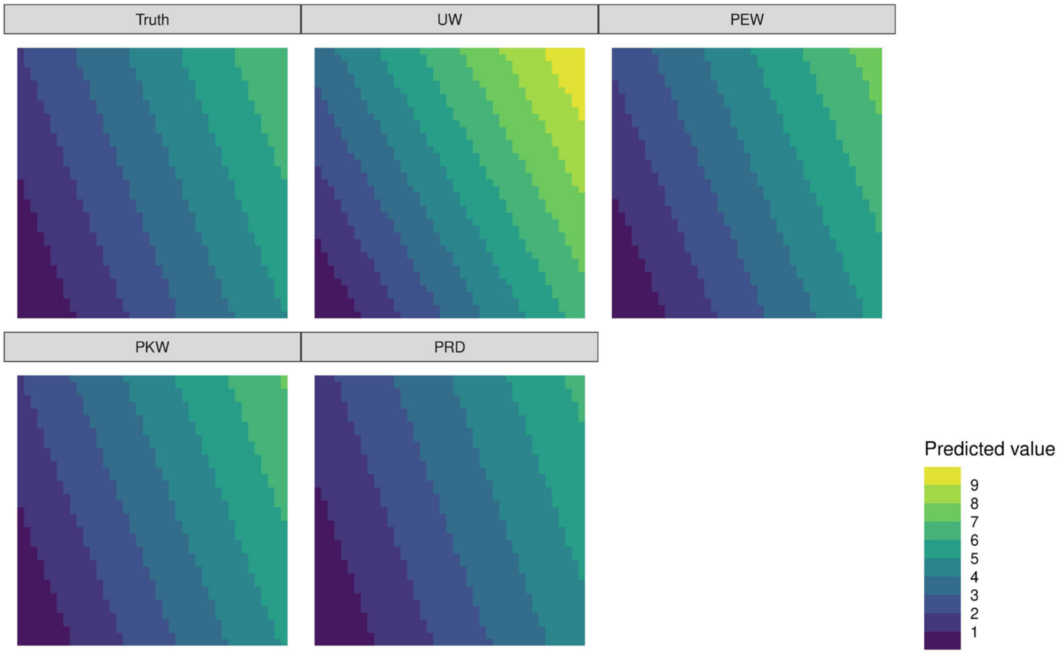

Figure 1. Plots of the predicted surfaces for each of the four models in the spatially implicit simulation scenario. The true surface is included for comparison. Each of the predicted surfaces is averaged over 100 simulations.

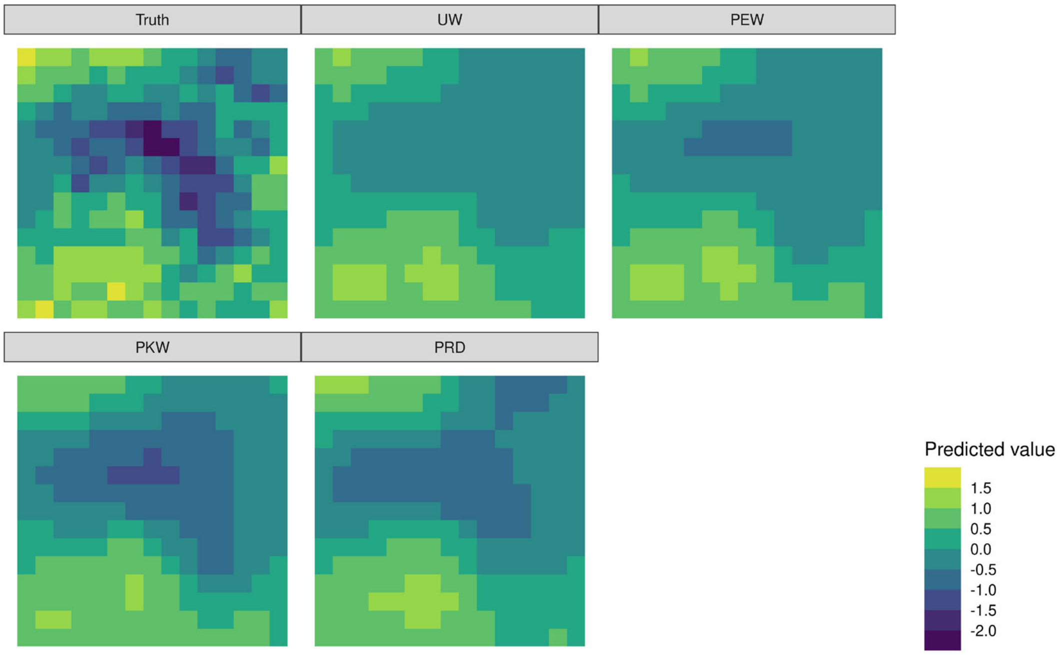

Figure 2. Plots of the posterior mean predicted surfaces for each of the four models in the spatially explicit simulation scenario. The true surface is included for comparison. Each of the predicted surfaces is averaged over 100 simulations.

bias, where the best results in terms of MSE are achieved with the PKW model. However, since full knowledge of the underlying sample location intensity surface is impractical in reality, the PEW model would be the most reasonable in practice.

The runtime of unweighted and pseudo-likelihood models are quite similar. Taking the unweighted as a baseline, the relative runtime for PEW to fit is only 1.019 times longer, while for PKW, it is 0.957 longer. The relative runtime for PRD is several orders of magnitude longer, but a more optimized implementation would likely narrow this discrepancy.

6.2. Simulation 2: Spatially Explicit Scenario

In the second scenario, we begin by simulating an intensity surface, $p(s)$, from a Gaussian process with a squared exponential kernel using the covariance function

$$c(s_i, s_j) = \exp\left(-2\sum_{d=1}^{D}(s_{di} - s_{dj})^2\right).$$

We then implement a PS scheme by selecting sample locations from an inhomogenous Poisson point process using the generated intensity function. Finally, observed values, $Z(s)$, are generated by adding iid Gaussian noise to the intensity evaluated at the observed locations. In other words, $Z(s) = p(s) + \epsilon(s)$. Thus, this approach is in essence generating data from the model specified by Diggle, Menezes, and Su (2010). The resulting true surface is reproduced in the first panel of Figure 2.

We fit the same three models as in Simulation 1, but now with a spatial process for the mean rather than a linear combination of spatial covariates. We choose to use a basis expansion representation of the spatial process. Thus, for this scenario, the UW model is given as

$$\begin{aligned}
Z(s)|\boldsymbol{\eta}, \sigma_z^2 &\sim \mathrm{N}(p(s), \sigma_z^2)\\
p(s) &= \sum_{k=1}^{K}\phi_k(s)\eta_k\\
\eta_k &\sim \mathrm{N}(0, \lambda_k\tau),\ k = 1, \dots, K\\
\lambda_k &\sim \mathrm{Cauchy}^+(0,1),\ k = 1, \dots, K\\
\tau &\sim \mathrm{Cauchy}^+(0,1),
\end{aligned}$$

where $\phi_k(s)$ is the value of the kth basis function evaluated at location s. An initially large number of basis functions is selected automatically at two resolutions using the FRK package (Zammit-Mangion and Cressie 2021). A horseshoe prior (Carvalho, Polson, and Scott 2010) is then placed on η_k in order to provide shrinkage for the initially large number of basis functions.

Similar to the first simulation scenario, we also compare to weighted pseudo-likelihood versions of this model. As before, we fit a pseudo-likelihood model with known weights (PKW) and a pseudo-likelihood model with weights constructed using a kernel density estimator (PEW) with the same settings as in the first simulation scenario. In addition, we compare to the method used by Pati, Reich, and Dunson (2011) (PRD).

The results of this simulation are summarized in Table 3. In this scenario, the PRD model outperforms the pseudo-likelihood models both in terms of MSE and mean absolute bias. This is to be expected here, as the data generating model is a case of the PRD model. However, each of the corrective models is able to reduce the MSE and bias relative to the UW model. Finally, the difference in MSE and mean absolute bias between the PEW and PKW models is less pronounced than in Simulation 1.

Figure 2 shows the predicted surface for each model, averaged over 100 simulations, alongside the true surface. Note the

Table 3. MSE and mean absolute bias for the posterior mean predictive surface produced by each of the four models in the spatially explicit scenario simulation.

Model	MSE	Mean Abs. Bias
UW	0.382	0.418
PEW	0.332	0.376
PKW	0.304	0.341
PRD	0.245	0.312

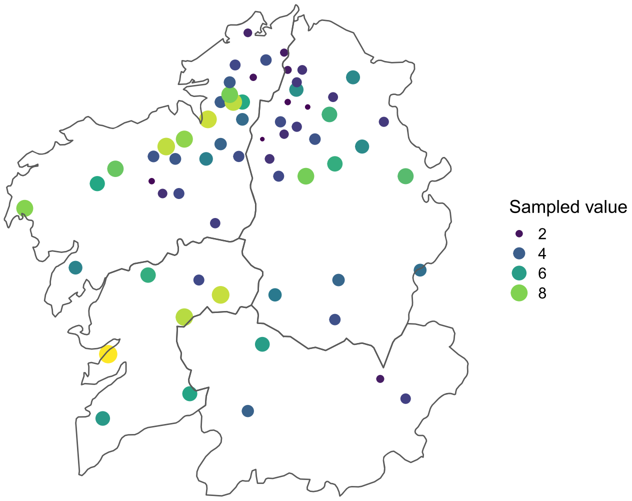

Figure 3. Plot of the sampling locations for the heavy metal bio-monitoring data in Galicia under a preferential scheme. Points are scaled proportional to the sampled value at a given location.

lower values at the center of the true surface. The unweighted model fails to capture these, instead predicting values closer to zero. In contrast, the weighted models are able to capture this behavior to varying degrees.

Taking UW as a baseline, the relative runtime for PEW is 0.913. For PKW it is 0.901, while for PRD it is 49.340.

7. Application to Heavy Metal Biomonitoring Data

To illustrate the different approaches in practice, we fit three of the models from the second simulation scenario to the heavy metal biomonitoring data from Galicia, Spain analyzed in Diggle, Menezes, and Su (2010). The response is lead concentrations in micorgrams per gram dry weight of moss. The data come from two surveys taken only a few years apart, first in 1997 and then in 2000. The 1997 data (63 sample points) uses a PS scheme, where sites with "large gradients" were more likely to be sampled. Figure 3 shows the sampled locations, with a clear concentration in the northern part of the region. The 2000 data (132 sample points) uses a regular lattice sample (i.e., not preferential) over the same domain (Figure 4). One question of interest could be whether a difference in mean response across the two samples is attributable to a true change over time, or only to the difference in sampling scheme.

With real data, we have no way of knowing the true surface, and this rules out fitting the PKW model of Section 6. However, by fitting models that do and do not assume the presence of PS to both datasets, we can evaluate the degree to which the model estimates differ relative to each other. The fit for each of the models to the nonpreferential data is shown in Figure 5, while the fit for the preferential data is shown in Figure 6. Note the difference in scale between the two figures. As expected, all three models yield similar estimates when the data is sampled in a nonpreferential manner. For the preferentially sampled data, there is much more variation in the estimates across models. The PEW model seems to give larger predictions than the UW model along the western edge, where predictions are generally higher than average. Interestingly, the PRD model results in high predicted heavy metals along the southeast edge, contradictory to both other models, as well as the nonpreferential sample. In this case, there is no true baseline to compare to, however, we might expect the true surface to be somewhat similar in the nonpreferential and the preferential sample, since they both sample the same geographic domain, although at different times.

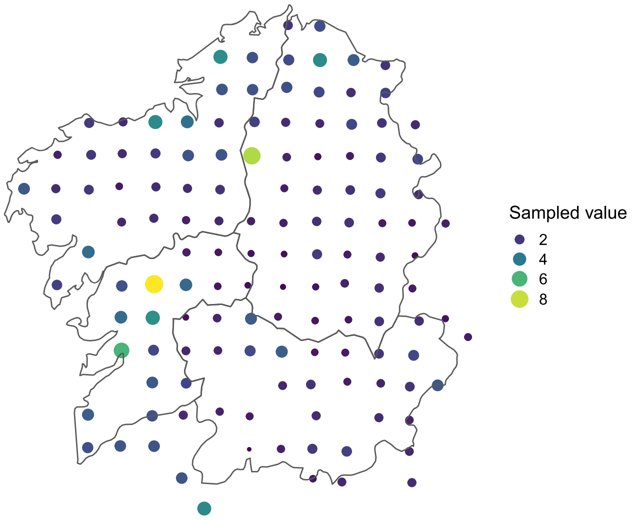

Figure 4. Plots of the sampling locations for the heavy metal bio-monitoring data in Galicia under a nonpreferential, regular lattice sample. Points are scaled proportional to the sampled value at a given location. The boundary is only approximate, so some points appear to fall outside of it, but these points do not affect the analysis.

The heavy metal data example considered here is interesting in that we have both a sample taken under preferential sampling, as well as a sample taken on a regular lattice. In this case, the samples are taken at two different time points. Thus, the true heavy metal surfaces that we predict are not necessarily the same, although we would expect similarity between the two. The important takeaway here is that the unweighted model will give similar results to the more complex models when there is no preferential sampling, however, the unweighted model will differ (i.e., have greater expected bias) when preferential sampling is present. Here, as in most real data settings, we do not know the true surface, and thus do not argue for a preference among the models that do account for the sampling strategy.

The combination of preferential and nonpreferential data is somewhat reminiscent of research around combining probability and nonprobability samples in survey statistics (e.g., see Wiśniowski et al. 2020). For example, suppose the heavy metal data were collected at the same time and thus used to predict the same surface. Then one strategy may be to fit a model for the preferentially sampled data, and then use the model fit to construct an informative prior for the non-preferentially sampled dataset. Another approach would be to model the data jointly with shared parameters.

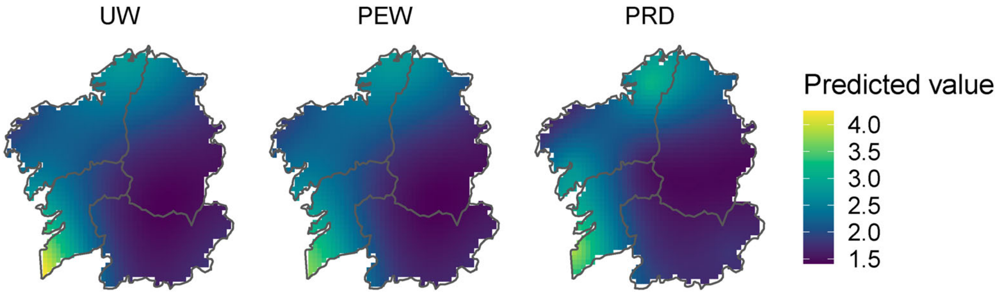

Figure 5. Predicted surfaces over Galicia for each model fit to the heavy metal bio-monitoring data under the nonpreferential, regular lattice, sampling scheme.

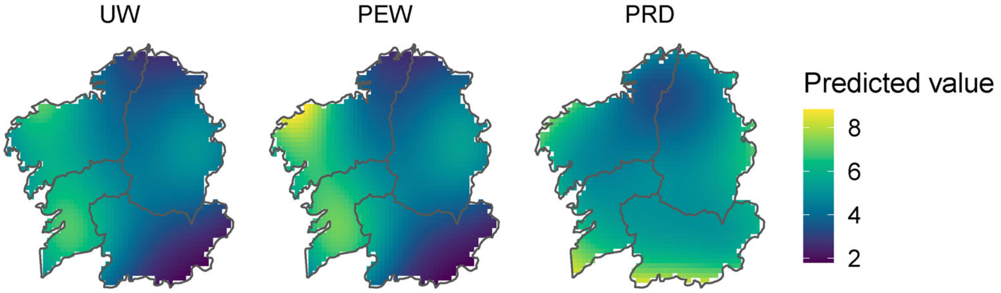

Figure 6. Predicted surfaces over Galicia for each model fit to the heavy metal bio-monitoring data under the preferential sampling scheme. Note the difference in scale from Figure 5.

8. Discussion

The problems of PS and IS have recently undergone extensive research, with many proposed methods within each domain. Although the overlap between the two problems is significant, in large part, the research has evolved down two independent tracks. In other words, little attention has been devoted to the explicit connections between these two related problems.

In this article we review many of the existing solutions to each of these problems and note the important differences in the modeling techniques. In addition, we provide a comprehensive review and comparison of the methods for addressing PS and note that a more extensive review of methods for addressing IS can be found in Parker, Janicki, and Holan (2019).

Broadly speaking, there are several approaches for handling the problems of PS and IS, including use of informative covariates in the model, regressing on the selection process, and weighted likelihood adjustments. Even though we demonstrate that there is significant overlap between the approaches used to handle each problem, one important distinction is that in the context of IS the weights are typically known from the sampling design, which is not the case for PS. In contrast, in the context of PS, the weights need to be estimated if they are to be used in a model. Nevertheless, there is literature on treating the weights as random in the context of IS, providing additional opportunities for extensions across the different domains (León-Novelo and Savitsky 2019; Williams and Savitsky 2021).

To demonstrate, compare, and contrast the different methods in the context of PS, we provide a multi-faceted simulation study. Notably, as expected, we find that the methods accounting for PS outperform the unweighted approach. This corroborates the findings in the IS context (e.g., see Parker, Janicki, and Holan 2019). Finally, through an application to heavy metal monitoring data, we illustrate the difference in analyses that account for PS relative to analyses that ignore PS.

There are several opportunities for future research. For example, methods in PS can be adapted to IS and vice versa. Specifically, as innovations in IS continue to appear in the literature, researchers working on PS problems can leverage this technology through suitable adaptations and extensions that explicitly account for the difference between the two problems noted throughout this article. In contrast, we believe that adapting methods from PS to IS will be particularly useful in contexts where the weights are not considered fixed and instead are treated as a portion of the model to be estimated. Ultimately, we envision that many of the extensions that arise will be application specific. To this end, the present work is meant to bridge the gap between the IS and PS communities and promote future research in both areas.

Supplementary Materials

The supplementary material contains R code for the application in Section 7. The file *Pseudo_LL.stan* implements the Bayesian pseudo-likelihood models specified in Section 6.2, while *MCMC.R* implements the sampler for the PRD model. *Heavy_metal_example.R* contains code for fitting these models and reproducing Figures 4–6. The data described in this section are stored in *lead97-new.txt* and *lead00-new.txt*. Shape files for plotting the boundary of Galicia are included, as well.

Acknowledgments

We thank the Associate Editor and two anonymous referees for providing comments that helped improve this article. We would also like to thank Prof. Brian J. Reich for providing the MCMC code used to fit the PRD model. This article is released to inform interested parties of ongoing research and to encourage discussion. The views expressed on statistical issues are those of the authors and not those of the NSF or U.S. Census Bureau.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

This research was partially supported by the U.S. National Science Foundation (NSF) under NSF grants SES-1853096, NCSE-2215168, and NCSE-2215169.

References

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2015), *Hierarchical Modeling and Analysis for Spatial Data*, London: Chapman and Hall. [314]

Binder, D. A. (1983), "On the Variances of Asymptotically Normal Estimators from Complex Surveys," *International Statistical Review*, 51, 279–292. [313,316]

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010), "The Horseshoe Estimator for Sparse Signals," *Biometrika*, 97, 465–480. [319]

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. [315,316]

Cressie, N., and Wikle, C. K. (2011), *Statistics for Spatio-Temporal Data*, Hoboken, NJ: Wiley. [314]

da Silva Ferreira, G., and Gamerman, D. (2015), "Optimal Design in Geostatistics Under Preferential Sampling," *Bayesian Analysis*, 10, 711–735. [313]

Diggle, P. J., Menezes, R., and Su, T. (2010), "Geostatistical Inference under Preferential Sampling," *Journal of the Royal Statistical Society*, Series C, 59, 191–232. [313,314,316,317,319,320]

Dinsdale, D., and Salibian-Barrera, M. (2019), "Methods for Preferential Sampling in Geostatistics," *Journal of the Royal Statistical Society*, Series C, 68, 181–198. [313]

Firth, D., and Bennett, K. (1998), "Robust Models in Probability Sampling," *Journal of the Royal Statistical Society*, Series B, 60, 3–21. [315]

Gelfand, A. E., Sahu, S. K., and Holland, D. M. (2012), "On the Effect of Preferential Sampling in Spatial Prediction," *Environmetrics*, 23, 565–578. [315]

Gelfand, A. E., and Shirota, S. (2019), "Preferential Sampling for Presence/Absence Data and for Fusion of Presence/Absence Data with Presence-Only Data," *Ecological Monographs*, 89, e01372. [316]

Gelman, A., and Little, T. C. (1997), "Poststratification into many Categories using Hierarchical Logistic Regression," *Survey Methodology*, 23, 127–135. [315]

Grantham, N. S., Reich, B. J., Liu, Y., and Chang, H. H. (2018), "Spatial Regression with an Informatively Missing Covariate: Application to Mapping Fine Particulate Matter," *Environmetrics*, 29, e2499. [316]

Irvine, K. M., Rodhouse, T. J., Wright, W. J., and Olsen, A. R. (2018), "Occupancy Modeling Species–Environment Relationships with Non-ignorable Survey Designs," *Ecological Applications*, 28, 1616–1625. [313]

Jiao, J., Hu, G., and Yan, J. (2019), "A Bayesian Joint Model for Spatial Point Processes with Application to Basketball Shot Chart," arXiv preprint arXiv:1908.05745. [313]

León-Novelo, L. G., and Savitsky, T. D. (2019), "Fully Bayesian Estimation under Informative Sampling," *Electronic Journal of Statistics*, 13, 1608–1645. [318,321]

Little, R. J. (2012), "Calibrated Bayes, an Alternative Inferential Paradigm for Official Statistics," *Journal of Official Statistics*, 28, 309–334. [315]

Little, R. J. A. (1993), "Post-Stratification: A Modeler's Perspective," *Journal of the American Statistical Association*, 88, 1001–1012. [315]

Møller, J., Syversveen, A. R., and Waagepetersen, R. P. (1998), "Log Gaussian Cox Processes," *Scandinavian Journal of Statistics*, 25, 451–482. [316]

Park, D. K., Gelman, A., and Bafumi, J. (2006), State-Level Opinions from National Surveys: Poststratification Using Multilevel Logistic Regression, in *Public Opinion in State Politics*, ed. J. E. Cohen, pp. 209–228, Redwood City, CA: Stanford University Press. [315]

Parker, P. A., Janicki, R., and Holan, S. H. (2019), "Unit Level Modeling of Survey Data for Small Area Estimation under Informative Sampling: A Comprehensive Overview with Extensions," arXiv preprint arXiv:1908.10488. [313,314,316,321]

Pati, D., Reich, B. J., and Dunson, D. B. (2011), "Bayesian Geostatistical Modelling with Informative Sampling Locations," *Biometrika*, 98, 35–48. [314,316,317,319]

Pennino, M. G., Paradinas, I., Illian, J. B., Muñoz, F., Bellido, J. M., López-Quílez, A., and Conesa, D. (2019), "Accounting for Preferential Sampling in Species Distribution Models," *Ecology and Evolution*, 9, 653–663. [313,316]

Pfeffermann, D., and Sverchkov, M. (2007), "Small-Area Estimation under Informative Probability Sampling of Areas and within the Selected Areas," *Journal of the American Statistical Association*, 102, 1427–1439. [313,314]

Savitsky, T. D., and Toth, D. (2016), "Bayesian Estimation under Informative Sampling," *Electronic Journal of Statistics*, 10, 1677–1708. [316]

Schliep, E. M., Wikle, C. K., and Daw, R. (2021), "Correcting Spatial Gaussian Process Parameter and Prediction Variance Estimation Under Informative Sampling," arXiv preprint arXiv:2108.12354. [317]

Scott, A. J., and Wild, C. J. (2011), "Fitting Regression Models with Response-Biased Samples," *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 39, 519–536. [317]

Si, Y., Pillai, N. S., and Gelman, A. (2015), "Bayesian Nonparametric Weighted Sampling Inference," *Bayesian Analysis*, 10, 605–625. [314,315]

Skinner, C. J. (1989), "Domain Means, Regression and Multivariate Analysis," in *Analysis of Complex Surveys*, eds. C. J. Skinner, D. Holt, and T. M. F. Smith, pp. 80–84, Chichester: Wiley. [313,316]

Vandendijck, Y., Faes, C., Kirby, R. S., Lawson, A., and Hens, N. (2016), "Model-based Inference for Small Area Estimation with Sampling Weights," *Spatial Statistics*, 18, 455–473. [316]

Venables, W. N., and Ripley, B. D. (2002), *Modern Applied Statistics with S* (4th ed.), New York: Springer. [317]

Watson, J. (2021), "A Perceptron for Detecting the Preferential Sampling of Locations and Times Chosen to Monitor a Spatio-Temporal Process," *Spatial Statistics*, 43, 100500. [313]

Williams, M. R., and Savitsky, T. D. (2020), "Bayesian Estimation Under Informative Sampling with Unattenuated Dependence," *Bayesian Analysis*, 15, 57–77. [317]

——— (2021), "Uncertainty Estimation for Pseudo-Bayesian Inference Under Complex Sampling," *International Statistical Review*, 89, 72–107. [318,321]

Wiśniowski, A., Sakshaug, J. W., Perez Ruiz, D. A., and Blom, A. G. (2020), "Integrating Probability and Nonprobability Samples for Survey Inference," *Journal of Survey Statistics and Methodology*, 8, 120–147. [320]

Zammit-Mangion, A., and Cressie, N. (2021), "FRK: An R Package for Spatial and Spatio-Temporal Prediction with Large Datasets," *Journal of Statistical Software*, 98, 1–48. [319]

Zheng, H., and Little, R. J. (2003), "Penalized Spline Model-based Estimation of the Finite Populations Total from Probability-Proportional-to-Size Samples," *Journal of Official Statistics*, 19, 99–117. [315]

Zidek, J. V., Shaddick, G., and Taylor, C. G. (2014), "Reducing Estimation Bias in Adaptively Changing Monitoring Networks with Preferential Site Selection," *The Annals of Applied Statistics*, 8, 1640–1670. [313,314,317]