# A comparison of design-based and model-based approaches for finite population spatial data.

Michael Dumelle[*,a], Matt Higham[b], Jay M. Ver Hoef[c], Anthony R. Olsen[a], Lisa Madsen[d]

[a] *United States Environmental Protection Agency, 200 SW 35th St, Corvallis, Oregon, 97333*
[b] *Saint Lawrence University Department of Mathematics, Computer Science, and Statistics, 23 Romoda Drive, Canton, New York, 13617*
[c] *Marine Mammal Laboratory, Alaska Fisheries Science Center, National Oceanic and Atmospheric Administration, Seattle, Washington, 98115*
[d] *Oregon State University Department of Statistics, 239 Weniger Hall, Corvallis, Oregon, 97331*

**Abstract**

1. The design-based and model-based approaches to frequentist statistical inference rest on fundamentally different foundations. In the design-based approach, inference relies on random sampling. In the model-based approach, inference relies on distributional assumptions. We compare the approaches for finite population spatial data.
2. We provide relevant background for the design-based and model-based approaches and then study their performance using simulated and real data. In the simulated data, a variety of sample sizes, location layouts, dependence structures, and response types are considered. In the simulated and real data, the population mean is the parameter of interest and performance is measured using statistics like bias, squared error, and interval coverage.
3. When studying the simulated and real data, we found that regardless of the strength of spatial dependence in the data, the Generalized Random Tessellation Stratified (GRTS) algorithm, which explicitly incorporates spatial locations into sampling, tends to outperform the Simple Random Sampling (SRS) algorithm, which does not explicitly incorporate spatial

[*]Corresponding Author: Michael Dumelle (Dumelle.Michael@epa.gov)

locations into sampling. We also found that model-based approaches tend to outperform design-based approaches, even for skewed data where the model-based distributional assumptions are violated. The performance gap between these approaches is small GRTS samples are used but large when SRS samples are used. This suggests that the sampling choice (whether to use GRTS or SRS) is most important when performing design-based inference.

4. There are many benefits and drawbacks to the design-based and model-based approaches for finite population spatial data that practitioners must consider when choosing between them. We provide relevant background contextualizing each approach and study their properties in a variety of scenarios, making recommendations for use based on the practitioner's goals.

**Keywords**

Design-based inference; Finite Population Block Kriging (FPBK); Generalized Random Tessellation Stratified (GRTS) algorithm; Local neighborhood variance estimator; Model-based inference; Restricted Maximum Likelihood (REML) estimation; Spatially balanced sampling; Spatial covariance

## 1. Introduction

When data cannot be collected for all units in a population (i.e., population units), data are collected on a subset of the population units – this subset is called a sample. There are two general approaches for using samples to make frequentist statistical inferences about a population: design-based and model-based. In the design-based approach, inference relies on randomly assigning some population units to be in the sample (random sampling). Alternatively, in the model-based approach, inference relies on distributional assumptions about the underlying stochastic process that generated the sample. Each paradigm has a deep historical context (Sterba, 2009) and its own set of benefits and drawbacks (Hansen et al., 1983). In this manuscript, we compare the design-based and model-based approaches for finite population spatial data.

Spatial data are data that incorporate the locations of the population units into either the sampling or estimation process. De Gruijter and Ter Braak (1990) and Brus and DeGruijter (1993) give early comparisons of design-based and model-based approaches for spatial data, quashing the belief that design-based approaches could not be used for spatially correlated data. Since then, there have been several general comparisons between design-based and model-based approaches for spatial data (Brus and De Gruijter, 1997; Brus, 2021; Ver Hoef, 2002, 2008). Cooper (2006) reviews the two approaches in an ecological context before introducing a "model-assisted" variance estimator that combines aspects from each approach. In addition to Cooper (2006), there has been substantial research and development into estimators that use both design-based and model-based principles (see e.g., Sterba (2009) and Cicchitelli and Montanari (2012), and for Bayesian approaches, see Chan-Golston et al. (2020) and Hofman and Brus (2021)).

Certainly comparisons between design-based and model-based approaches have been studied in spatial contexts. Our contribution is comparing design-based approaches that incorporate spatial locations into sampling and analysis to model-based approaches. Though the broad comparisons we draw between design-based and model-based approaches generalize to finite and infinite populations, we focus on finite populations. A finite population contains a finite number of population units (we assume the finite number is known); an example is lakes (treated as a whole with the lake centroid representing location) in the contiguous

United States. An infinite population contains an infinite number of population units; an example is locations within a single lake.

The rest of the manuscript is organized as follows. In Section 1.1, we introduce and provide relevant background for the design-based and model-based approaches to finite population spatial data. In Section 2, we describe how we compare performance of the approaches with a simulation study and an analysis of real data that contains mercury concentration in lakes located in the contiguous United States. In Section 3, we present results from the simulation study and the mercury concentration analysis. And in Section 4, we end with a discussion and provide directions for future research.

### *1.1. Background*

The design-based and model-based approaches incorporate randomness in fundamentally different ways. In this section, we describe the role of randomness for each approach and the subsequent effects on statistical inferences for spatial data.

#### *1.1.1. Comparing Design-Based and Model-Based Approaches*

The design-based approach assumes the population is fixed. Randomness is incorporated via the selection of population units according to a sampling design. A sampling design assigns a probability of selection to each sample (a subset of population units). Some examples of commonly used sampling designs include simple random sampling, stratified random sampling, and cluster sampling. The inclusion probability of a population unit follows by summing each population unit's probability of selection in each sample containing that population unit. Inclusion probabilities are later used to estimate population parameters.

When samples are chosen in a manner such that the layout of sampled units reflects the layout of the population units, we call the resulting sample "spatially balanced." By "reflecting the layout of the population units", we mean that if population units are concentrated in specific areas, the units in the sample should be concentrated in the same areas. Because spatially balanced samples reflect the layout of the population units, they are not necessarily "spread out" in space in some equidistant manner. One approach to selecting spatially balanced samples is the Generalized Random Tessellation Stratified (GRTS) algorithm (Stevens and Olsen, 2004), which we discuss in more detail in Section 1.1.2.

Fundamentally, the design-based approach combines the randomness of the sampling design with the data collected via the sample to justify the estimation and uncertainty quantification of fixed, unknown parameters of a population (e.g., a population mean). Treating the data as fixed and incorporating randomness through the sampling design yields estimators having very few other assumptions. Confidence intervals for these types of estimators are typically derived using limiting arguments that incorporate all possible samples. Sample means, for example, are asymptotically normal (Gaussian) by the Central Limit Theorem (under some assumptions). If we repeatedly select samples from the population, then 95% of all 95% confidence intervals constructed from a procedure with appropriate coverage will contain the true fixed population mean. Särndal et al. (2003) and Lohr (2009) provide thorough reviews of the design-based approach.

The model-based approach assumes the population is a random realization of a data-generating stochastic process called a superpopulation. Randomness is formally incorporated through distributional assumptions on the superpopulation. Strictly speaking, randomness need not be incorporated through random sampling, though Diggle et al. (2010) warn against preferential sampling. Preferential sampling occurs when the process generating the data locations and the process being modeled are not independent of one another. To guard against preferential sampling, model-based approaches can implement some form of random sampling, though it is common for model-based approaches to not implement random sampling. When model-based approaches do implement random sampling, the inclusion probabilities are ignored when analyzing the sample (in contrast to the design-based approach, which relies on these inclusion probabilities to analyze the sample).

Instead of estimating fixed, unknown population parameters, as in the design-based approach, often the goal of model-based inference is to predict a realized variable. For example, suppose the realized mean of all population units (the realized population mean) is the variable of interest. Instead of a fixed, unknown mean, we are predicting the value of the mean, a random variable. Prediction intervals are then derived using assumptions of the data-generating stochastic process. If we repeatedly generate response values from the same process and select samples, then 95% of all 95% prediction intervals constructed from a procedure with appropriate coverage will contain their respective realized means. Cressie (1993) and Schabenberger and Gotway (2017) provide thorough reviews of model-based approaches for spatial data. In Fig. 1, we provide a visual comparison of the design-based and model-based approaches (Ver Hoef (2002) and Brus (2021) provide similar figures).

#### *1.1.2. Spatially Balanced Design and Analysis*

We previously mentioned that the design-based approach can be used to select spatially balanced samples. Spatially balanced samples are useful because parameter estimates from these samples tend to vary less than parameter estimates from samples that are not spatially balanced (Barabesi and Franceschi, 2011; Benedetti et al., 2017; Grafström and Lundström, 2013; Robertson et al., 2013; Stevens and Olsen, 2004; Wang et al., 2013). The first spatially balanced sampling algorithm to see widespread use was the Generalized Random Tessellation Stratified (GRTS) algorithm (Stevens and Olsen, 2004). To quantify

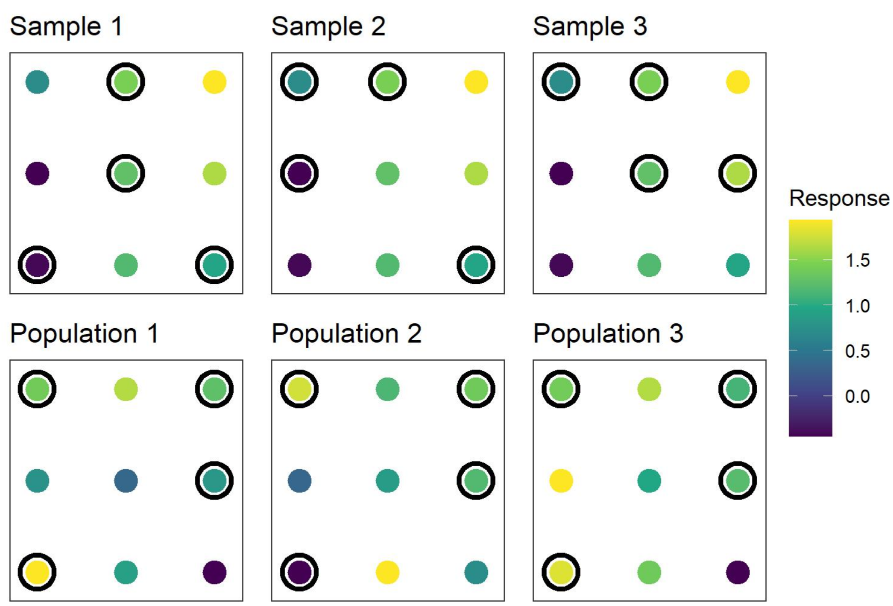



Figure 1: A visual comparison of the design-based and model-based approaches. In the top row, the design-based approach is highlighted. There is one fixed population with nine population units and three random samples of size four (points circled are those sampled). The response values at each site are fixed, but we obtain different estimates for the mean response in each random sample. In the bottom row, the model-based approach is highlighted. There are three realizations of the same data-generating stochastic process that are all sampled at the same four locations. The data-generating stochastic process has a single mean, but the mean of the nine population units is different in each of the three realizations.

the spatial balance of a sample, Stevens and Olsen (2004) proposed loss metrics based on Voronoi polygons (Dirichlet Tessellations). After the GRTS algorithm was developed, several other spatially balanced sampling algorithms emerged, including stratified sampling with compact geographical strata Walvoort et al. (2010), the Local Pivotal Method (Grafström et al., 2012; Grafström and Matei, 2018), Spatially Correlated Poisson Sampling (Grafström, 2012), Balanced Acceptance Sampling (Robertson et al., 2013), Within-Sample-Distance Sampling (Benedetti and Piersimoni, 2017), and Halton Iterative Partitioning Sampling (Robertson et al., 2018). In this manuscript, we select spatially balanced samples using the Generalized Random Tessellation Stratified (GRTS) algorithm because it is readily available in the `spsurvey` **R** package (Dumelle et al., 2022) and naturally accommodates finite and infinite sampling frames, unequal inclusion probabilities, and replacement units (replacement units are population units that can be sampled when a population unit originally selected can no longer be sampled).

The GRTS algorithm selects samples by utilizing a particular mapping between two-dimensional and one-dimensional space that preserves proximity relationships. First the bounding box of the domain is split up into four distinct, equally sized squares called level-one cells. Each level-one is randomly assigned an level-one address of 0, 1, 2, or 3. The set of level-one cells is denoted by $\mathcal{A}_1$ and defined as $\mathcal{A}_1 \equiv \{a_1 : a_1 = 0, 1, 2, 3\}$. Within each level-one cell, the inclusion probability for each population unit is summed, and if any of these sums exceed one, a second level of cells is added. Then each level-one cell is split into four distinct, equally sized squares called level-two cells. Each level-two cell is randomly assigned a level-two address of 0, 1, 2, or 3. The set of level-two cells is denoted by $\mathcal{A}_2$ and defined as $\mathcal{A}_2 \equiv \{a_1 a_2 : a_1 = 0, 1, 2, 3; a_2 = 0, 1, 2, 3\}$. The inclusion probabilities within each level-two cell are summed, and if any of these sums exceed one, a third level of cells is added. This process continues for $k$ steps, until all level-$k$ cells have inclusion probability sums no larger than one. Then $\mathcal{A}_k \equiv \{a_1...a_k : a_1 = 0, 1, 2, 3; ...; a_k = 0, 1, 2, 3\}$.

After determining $\mathcal{A}_k$, it is placed into hierarchical order. Hierarchical order is a numeric order that first sorts $\mathcal{A}_k$ by the level-one addresses from smallest to largest, then sorts $\mathcal{A}_k$ by the level-two addresses from smallest to largest, and so on. For example, $\mathcal{A}_2$ in hierarchical order is the set $\{00, 01, 02, 03, 10, ..., 13, 20, ..., 23, 30, ..., 33\}$. Because hierarchical ordering sorts by level-one cells, then level-two cells, and so on, population units that have similar hierarchical addresses tend to be nearby one another in space.Next each population unit is mapped to a one-dimensional line in hierarchical order where each population unit's inclusion probability equals its line-length. If a level-$k$ cell has multiple population units in it, they are randomly placed within the cell's respective line segment. A uniform random variable is then simulated in $[0, 1]$ and a systematic sample is selected on the line, yielding $n$ sample points for a sample size $n$. Each element in this systematic sample falls on some population unit's line segment, and thus that population unit is selected in the sample. For further details regarding the GRTS algorithm, see Stevens and Olsen (2004).

After selecting a sample and collecting data, unbiased estimates of population means and totals can be obtained using the Horvitz-Thompson estimator (Horvitz and Thompson, 1952). If $\tau$ is a population total, the Horvitz-Thompson estimator for $\tau$, denoted by $\hat{\tau}_{ht}$, is is given by

$$\hat{\tau}_{ht} = \sum_{i=1}^{n} Z_i \pi_i^{-1}, \tag{1}$$

where $Z_i$ is the value of the $i$th population unit in the sample, $\pi_i$ is the inclusion probability of the $i$th population unit in the sample, and $n$ is the sample size. An estimate of the population mean is obtained by dividing $\hat{\tau}_{ht}$ by $N$, the number of population units.

It is also important to quantify the uncertainty in $\hat{\tau}_{ht}$. Horvitz and Thompson (1952) and Sen (1953) provide variance estimators for $\hat{\tau}_{ht}$, but these estimators have two drawbacks. First, they rely on calculating $\pi_{ij}$, the probability that population unit $i$ and population unit $j$ are both in the sample – this quantity can be challenging if not impossible to calculate analytically for GRTS samples. Second, these estimators tend to ignore the spatial locations of the population units. To address these two drawbacks simultaneously, Stevens and Olsen (2003) proposed the local neighborhood variance estimator. The local neighborhood variance estimator does not rely on $\pi_{ij}$ and estimates the variance of $\hat{\tau}$ conditional on the random properties of the GRTS sample – the idea being that this conditioning should yield a more precise estimate of $\hat{\tau}$. They show that the each observation's contribution to the overall variance is dominated by local variation. Thus the local neighborhood variance estimator is a weighted sum of variance estimates from each observation's local neighborhood. These local neighborhoods contain observation itself and its three nearest neighbors. For more details, see Stevens and Olsen (2003).

#### *1.1.3. Finite Population Block Kriging*

Finite Population Block Kriging (FPBK) is a model-based approach that expands the geostatistical Kriging framework to the finite population setting (Ver Hoef, 2008). Instead of developing inference based on a specific sampling design, we assume the data are generated by a spatial stochastic process. We summarize some of the basic principles of FPBK next – for technical details, see Ver Hoef (2008). Let $\mathbf{z} \equiv \{\mathrm{z}(s_1), \mathrm{z}(s_2), ..., \mathrm{z}(s_N)\}$ be an $N \times 1$ response vector at locations $s_1$, $s_2$, . . . , $s_N$ that can be measured at the $N$ population units. Suppose we want to use a sample to predict some linear function of the response variable, $f(\mathbf{z}) = \mathbf{b}'\mathbf{z}$, where $\mathbf{b}'$ is a $1 \times N$ vector of weights (e.g, the population mean is represented by a weights vector whose elements all equal $1/N$). Denoting quantities that are part of the sampled population units with a subscript $s$ and quantities that are part of the unsampled population units with a subscript $u$, let

$$\begin{pmatrix} \mathbf{z}_s \\ \mathbf{z}_u \end{pmatrix} = \begin{pmatrix} \mathbf{X}_s \\ \mathbf{X}_u \end{pmatrix} \boldsymbol{\beta} + \begin{pmatrix} \boldsymbol{\delta}_s \\ \boldsymbol{\delta}_u \end{pmatrix}, \tag{2}$$

where $\mathbf{X}_s$ and $\mathbf{X}_u$ are the design matrices for the sampled and unsampled population units, respectively, $\boldsymbol{\beta}$ is the parameter vector of fixed effects, and $\boldsymbol{\delta} \equiv [\boldsymbol{\delta}_s \ \boldsymbol{\delta}_u]'$, where $\boldsymbol{\delta}_s$ and $\boldsymbol{\delta}_u$ are random errors for the sampled and unsampled population units, respectively.

FPBK assumes $\boldsymbol{\delta}$ in Equation 2 has mean-zero and a spatial dependence structure that can be modeled using a covariance function. This covariance function is commonly assumed to be non-negative, second-order stationary (depending only on the separation vector (e.g., distance) between population units), isotropic (independent of direction), and decay with distance between population units (Cressie, 1993). Henceforth, it is implied that we have made these same assumptions regarding $\boldsymbol{\delta}$, though Chiles and Delfiner (1999), pp. 80-93 discuss covariance functions that are not second-order stationary, not isotropic, or not either. A variety of flexible covariance functions can be used to model $\boldsymbol{\delta}$ (Cressie, 1993); one example is the exponential covariance function (Cressie (1993) provides a thorough list of spatial covariance functions). The $i, j$th element of the exponential covariance matrix, $\text{cov}(\boldsymbol{\delta})$, is

$$\text{cov}(\delta_i, \delta_j) = \begin{cases} \sigma_1^2 \exp(-h_{i,j}/\phi) & h_{i,j} > 0 \\ \sigma_1^2 + \sigma_2^2 & h_{i,j} = 0 \end{cases}, \tag{3}$$

where $\sigma_1^2$ is the variance parameter that quantifies the spatially dependent variability, $\sigma_2^2$ is the variance parameter the quantifies that spatially independent variability, $\phi$ is the distance parameter that measures the distance-decay rate of the covariance, and $h_{i,j}$ is the Euclidean distance between population units $i$ and $j$. In geostatistical literature, $\sigma_1^2$ is often called the partial sill, $\sigma_2^2$ is often called the nugget, and $\phi$ is often called the range.

The parameters in Equation 2 can be estimated using a variety of techniques, but we focus on using restricted maximum likelihood (Harville, 1977; Patterson and Thompson, 1971; Wolfinger et al., 1994). REML is preferred over maximum likelihood (ML) because ML estimates can be badly biased for small sample sizes, due to the fact that ML makes no adjustment for the simultaneous estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ (Patterson and Thompson, 1971). Minus twice the REML log-likelihood of the sampled sites is given by

$$\ln|\boldsymbol{\Sigma}| + (\boldsymbol{z}_s - \boldsymbol{X}_s\tilde{\boldsymbol{\beta}})^T\boldsymbol{\Sigma}_{ss}^{-1}(\boldsymbol{z}_s - \boldsymbol{X}_s\tilde{\boldsymbol{\beta}}) + \ln|\boldsymbol{X}_s^T\boldsymbol{\Sigma}_{ss}^{-1}\boldsymbol{X}_s| + (n-p)\ln(2\pi), \quad (4)$$

where $\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}_s^T\boldsymbol{\Sigma}_{ss}^{-1}\boldsymbol{X}_s)^{-1}\boldsymbol{X}_s^T\boldsymbol{\Sigma}_{ss}^{-1}\boldsymbol{z}_s$ and $\boldsymbol{\Sigma}_{ss}$ is the covariance matrix of the sampled sites. Minimizing Equation 4 yields $\hat{\boldsymbol{\delta}}_{reml}$, the REML estimates of $\boldsymbol{\delta}$. Then $\boldsymbol{\beta}_{reml}$, the REML estimate of $\boldsymbol{\beta}$, is given by $(\boldsymbol{X}_s^T\hat{\boldsymbol{\Sigma}}_{ss}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}_s^T\hat{\boldsymbol{\Sigma}}_{ss}^{-1}\boldsymbol{z}_s$, where $\hat{\boldsymbol{\Sigma}}_{ss}$ is $\boldsymbol{\Sigma}_{ss}$ evaluated at $\hat{\boldsymbol{\delta}}_{reml}$.

With the model formulation in Equation 2, the Best Linear Unbiased Predictor (BLUP) for $f(\mathbf{b}'\mathbf{z})$ and its prediction variance can be computed. While details of the derivation are in Ver Hoef (2008), we note here that the predictor and its variance are both moment-based, meaning that they do not rely on any distributional assumptions. Distributional assumptions are used, however, when constructing prediction intervals.

Other approaches, such as k-nearest-neighbors (Fix and Hodges, 1989; Ver Hoef and Temesgen, 2013) and random forest (Breiman, 2001), among others, could also be used to obtain predictions for a mean or total from finite population spatial data. Compared to the k-nearest-neighbors and random forest approach, we prefer FPBK because it is model-based and relies on theoretically-based variance estimators leveraging the model's spatial covariance structure, whereas k-nearest-neighbors and random forests use ad-hoc variance estimators (Ver Hoef and Temesgen, 2013). Additionally, Ver Hoef and Temesgen (2013) compared FPBK, k-nearest-neighbors, and random forest in a variety of spatial data contexts, and FPBK tended to perform best.

## 2. Materials and Methods

### *2.1. Simulated Data*

REWRITE AS SIMPLE RANDOM SAMPLING AND WITHOUT REPLACEMENT

USE SRS / GRTS - DB / MB

ADD LOHR REFERENCE AND FORM FOR SRS VARIANCE WITHOUT REPLACEMENT WITH FPC

We used a simulation study to investigate performance of four sampling-analysis combinations. The first sampling-analysis combination was IRS-Design. In IRS-Design, samples were selected with the Independent Random Sampling (IRS) algorithm. The IRS algorithm ignores the spatial locations of the population units, thus the IRS samples were not spatially balanced. In IRS-Design, samples were analyzed using the design-based approach via the Horvitz-Thompson mean estimator and an IRS variance estimator that ignored the spatial locations of the units in the sample. The second sampling-analysis combination was IRS-Model, where samples were selected with the IRS algorithm and analyzed using the model-based approach while estimating the covariance parameters ($\boldsymbol{\delta}$) and fixed effects ($\boldsymbol{\beta}$ using restricted maximum likelihood (REML). The third sampling-analysis combination was GRTS-Design, where samples were selected with the GRTS algorithm and analyzed using the design-based approach via the Horvitz-Thompson mean estimator and the local neighborhood variance estimator (which does incorporate the spatial locations of the units in the sample). The fourth and final sampling-analysis combination was GRTS-Model, where samples were selected with the GRTS algorithm and analyzed using the model-based approach while estimating the covariance parameters ($\boldsymbol{\delta}$) and fixed effects ($\boldsymbol{\beta}$ using restricted maximum likelihood (REML). These sampling-analysis combinations are also provided in Table 1. Lastly we note that for both the IRS and GRTS samples, equal inclusion probabilities were assumed for all population units. When IRS assumes equal inclusion probabilities for all population units, the algorithm is equivalent to simple random sampling (SRS).

| | Design | Model |
|---|---|---|
| IRS | IRS-Design | IRS-Model |
| GRTS | GRTS-Design | GRTS-Model |

Table 1: Sampling-analysis combinations in the simulation study. The rows give the two types of sampling designs and the columns give the two types of analyses.

CHANGE LOGNORMAL VERBAGE TO SKEWED – look for DRE acronym

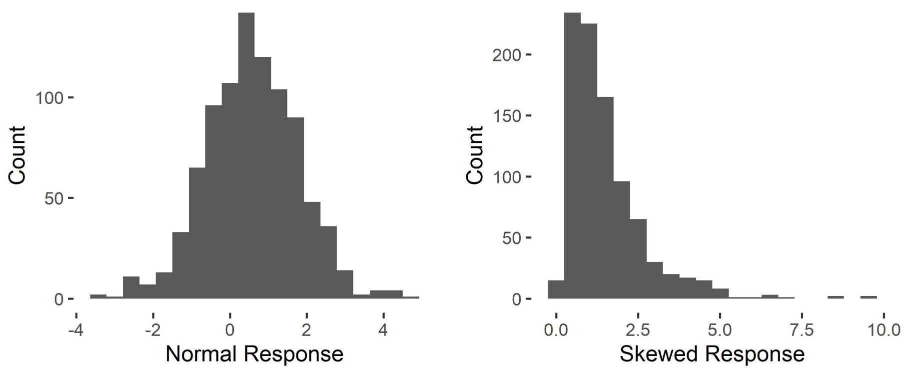
Performance for the four sampling-analysis combinations was evaluated in 36 different simulation scenarios. The 36 scenarios resulted from the crossing of three sample sizes, two location layouts (of the population units), two response types, and three proportions of dependent random error (DRE). The three sample sizes ($n$) were $n = 50, n = 100$, and $n = 200$. Samples were always selected from a population size ($N$) of $N = 900$. The two location layouts were random and gridded. Locations in the random layout were randomly generated inside the unit square ($[0, 1] \times [0, 1]$). Locations in the gridded layout were placed on a fixed, equally spaced grid inside the unit square. The two response types were normal and lognormal. For the normal response type, the response was simulated using mean-zero random errors with the exponential covariance

(Equation 3) for varying proportions of dependent random error. The proportion of dependent random error is represented by $\sigma_1^2/(\sigma_1^2 + \sigma_2^2)$, where $\sigma_1^2$ and $\sigma_2^2$ are the dependent random error variance (partial sill) and independent random error variance (nugget) from Equation 3, respectively. The total variance, $\sigma_1^2 + \sigma_2^2$, was always 2. The range was always $\sqrt{2}/3$, chosen so that the correlation in the dependent random error decayed to nearly zero at $\sqrt{2}$, the largest possible distance between two population units in the domain. For the lognormal response type, the response was first simulated using the same approach as for the normal response type, except that the total variance was 0.6931 instead of 2. The response was then exponentiated, yielding a lognormal random variable whose total variance was 2. The lognormal responses were used to evaluate performance of the sampling-analysis approaches for data that were skewed (i.e., not normal).

| | | | |
|---|---|---|---|
| Sample Size (n) | 50 | 100 | 200 |
| Location Layout | Random | Gridded | - |
| Proportion of Dependent Error | 0 | 0.5 | 0.9 |
| Response Type | Normal | Lognormal | - |

Table 2: Simulation scenario options. All combinations of sample size, location layout, response type, and proportion of dependent random error composed the 36 simulation scenarios. In each simualtion scenario, the total variance was 2.

In each of the 36 simulation scenarios, there were 2000 independent simulation trials. In each trial, IRS and GRTS samples were selected and then design-based and model-based analyses were used to estimate (design-based) or predict (model-based) the mean and construct 95% confidence (design-based) or 95% prediction (model-based) intervals. Then we recorded the bias, squared error, standard error, and interval coverage for all sampling-analysis combinations. After all 2000 trials, we summarized the long-run performance of the combinations by calculating mean bias, rMS(P)E (root-mean-squared error for the design-based approaches and root-mean-squared-prediction error for the model-based approaches), MStdE (mean standard error), and the proportion of times the true mean is contained




(a) Histogram of a single realized population for the normal response.

(b) Histogram of a single realized population for the skewed response.

Figure 2: Histograms realized populations for the simulated data.

in its 95% confidence (design-based) or 95% prediction (model-based) interval. The 95% intervals were constructed using the normal distribution. Justification for this comes from the asymptotic normality of means via the Central Limit Theorem (under some assumptions). Quantifying mean bias and rMS(P)E is important because they help us understand how far (under different loss metrics) the estimates (design-based) or predictions (model-based) tend to be from the true mean. Quantifying MStdE is important because it helps us understand how precise intervals tend to be. Quantifying interval coverage is important because it helps us understand how often our 95% intervals actually contain the true mean.

The IRS algorithm, IRS variance estimator, GRTS algorithm, and local neighborhood variance estimator are available in the `spsurvey` **R** package (Dumelle et al., 2022). FPBK is available in the `sptotal` **R** package (Higham et al., 2021).

### *2.2. National Lakes Assessment Data*

The United States Environmental Protection Agency (USEPA), states, and tribes periodically conduct National Aquatic Research Surveys (NARS) to assess the water quality of various bodies of water in the contiguous United States.

One component of NARS is the National Lakes Assessment (NLA), which measures various aspects of lake health and water quality (USEPA, 2012). We will analyze mercury concentration data collected at 986 lakes from the 2012 NLA. Although we can calculate the true mean mercury concentration values for these 986 lakes, here we will explore whether or not we can obtain an adequately precise estimate (design-based) or prediction (model-based) for the realized mean mercury concentration if we sample only 100 of the 986 lakes. For each of the four familiar sampling-analysis combinations (IRS-Design, IRS-Model, GRTS-Design, and GRTS-Model), we estimate (design-based) or predict (model-based) the mean mercury concentration and construct 95% intervals from this sample of 100 lakes and compare to the true mean mercury concentration from all 986 lakes.

## 3. Results

### *3.1. Simulated Data*

The mean bias was nearly zero for all four sampling-analysis combinations in all 36 scenarios, so we omit a more detailed summary of those results here. Tables for mean bias in all 36 simulation scenarios are provided in the supporting information.

Fig. 3 shows the relative rMS(P)E of the four sampling analysis combinations using the random location layout with "IRS-Design" as the baseline. The relative rMS(P)E is defined as

$$\frac{\text{rMS(P)E of sampling-analysis combination}}{\text{rMS(P)E of IRS-Design}},$$

When there is no spatial covariance (Fig. 3, "Prop DE: 0" row), the four sampling-analysis combinations have approximately equal rMS(P)E and using the GRTS algorithm or a model-based analysis does not result in much, if any, loss in efficiency compared to IRS-Design. When there is spatial covariance

(Fig. 3, "Prop DE: 0.5" and "Prop DE: 0.9" rows), GRTS-Model tends to have the lowest rMS(P)E, followed by GRTS-Design, IRS-Model, and finally IRS-Design, though the difference in relative rMS(P)E among GRTS-Model, GRTS-Design, and IRS-Model is relatively small. As the strength of spatial covariance increases, the gap in rMS(P)E between IRS-Design and the other sampling-analysis combinations widens. Finally we note that when there is spatial covariance, IRS-Model has a much lower rMS(P)E than IRS-Design, suggesting that the poor design properties of IRS are largely mitigated by the model-based analysis. These rMS(P)E conclusions are similar to those observed in the grid location layout, so we omit a grid location layout figure here. Tables for rMS(P)E in all 36 simulation scenarios are provided in the supporting information.

Fig. **??** shows the relative MStdE of the four sampling-analysis combinations using the random location layout with "IRS-Design" as the baseline. The relative MStdE is defined as

$$\frac{\text{MStdE of sampling-analysis combination}}{\text{MStdE of IRS-Design}},$$

Many general takeaways regarding MStdE are similar to general takeaways regarding rMS(P)E: there seems to be no benefit to using IRS, even when there is no spatial covariance; as the strength of spatial covariance increases, the gap in MStdE between IRS-Design and the other sampling-analysis combinations widens; and IRS-Model outperforms IRS-Design by a noticeable margin. These fact that the rMS(P)E and MStdE findings are similar is not particularly surprising because the mean bias for all sampling-analysis combinations was nearly zero, thus rMS(P)E is driven by the standard error of the estimators (design-based) or predictors (model-based). We do note that between GRTS-Design and GRTS-Model, GRTS-Design had lower MStdE when there was no spatial covariance or a medium amount of spatial covariance (Fig. **??**, "Prop DE: 0" and "Prop DE:

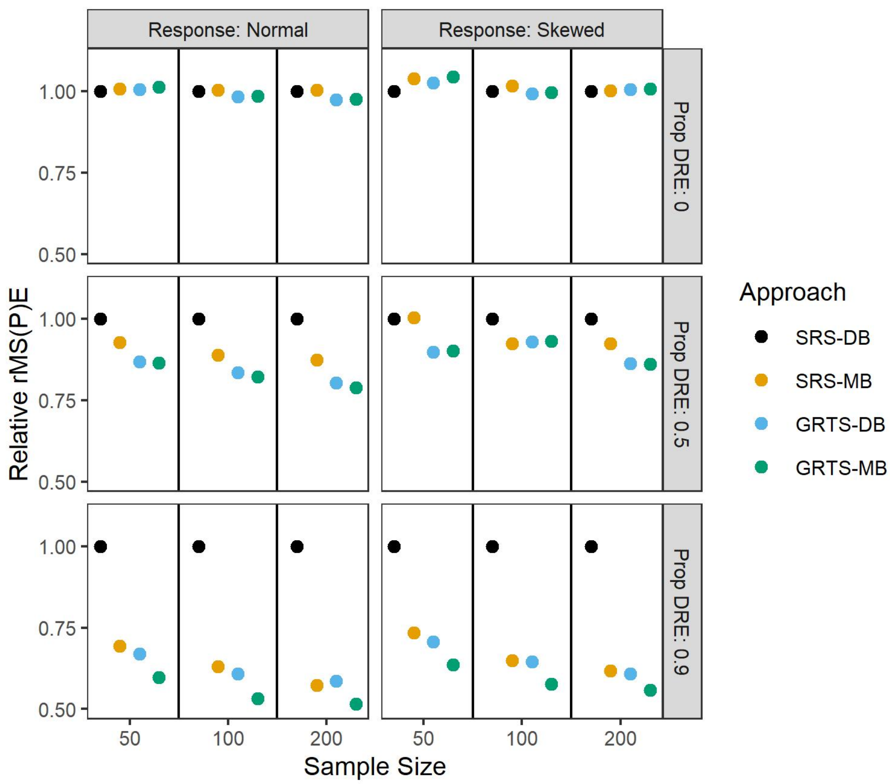



Figure 3: Relative rMS(P)E in the simulation study for the four sampling-analysis combinations. The rows indicate the proportion of dependent error and the columns indicate the response type.

0.5” rows), and GRTS-Model had lower MStdE when there was a high amount of spatial covariance (Fig. **??**, “Prop DE: 0.9” row). These MStdE conclusions are similar to those observed in the grid location layout, so we omit a grid location layout figure here. Tables for MStdE in all 36 simulation scenarios are provided in the supporting information.

Fig. 4 shows the 95% interval coverage for each of the four sampling-analysis combinations in the random location layout. Within each scenario, the sampling-analysis combinations tend to have fairly similar interval coverage, though when $n = 50$ or $n = 100$, GRTS-Design coverage is usually a few percentage points lower than the other combinations. Coverage in the normal response scenarios was usually near 95%, while coverage in the lognormal response scenarios usually varied from 90% to 95% but increased with the sample size. At a sample size of 200, all four sampling-analysis combinations had approximately 95% interval coverage in both response scenarios for all dependent error proportions. These interval coverage conclusions are similar to those observed in the grid location layout, so we omit a grid location layout figure here. Tables for interval coverage in all 36 simulation scenarios are provided in the supporting information.

### *3.2. National Lakes Assessment DAta*

USE MERCURY UNITS

Fig. **??** shows a map and histogram of mercury concentration in all 986 NLA lakes. The map shows mercury concentration exhibits some spatial patterning, with high mercury concentrations in the northeast and north central United States. The histogram shows that mercury concentration is right-skewed, with most lakes having a low value of mercury concentration but a few having a much higher concentration. Fig. **??** also shows mercury concentration’s empirical semivariogram. The empirical semivariogram can be used as a tool to visualize spatial dependence. It quantifies the mean of the halved squared differences

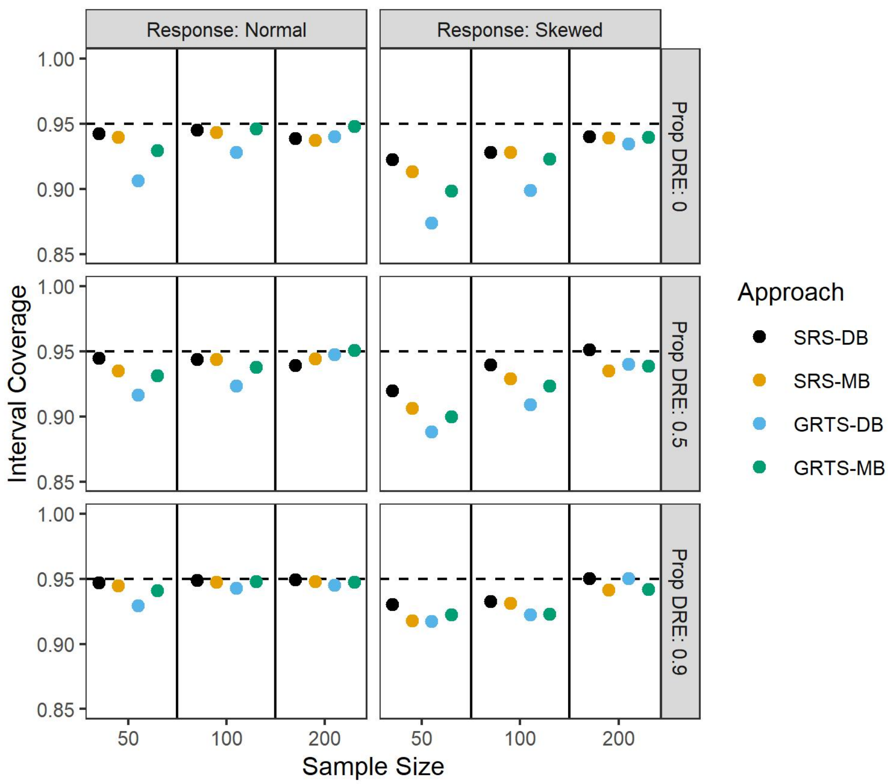



Figure 4: Interval coverage in the simulation study for the four sampling-analysis combinations. The rows indicate the proportion of dependent error and the columns indicate the response type. The solid, black line represents 95% coverage.

(semivariance) among all pairs of mercury concentrations at different distances apart. When a process has spatial covariance (exhibits spatial dependence), the mean semivariance tends to be smaller at small distances and larger at large distances. The empirical semivariogram in Fig. **??** suggests that mercury concentration exhibits spatial dependence. Lastly we note that the true mean mercury concentration in the 986 NLA lakes is 103.2 ng / g.

We selected a single IRS sample and a single GRTS sample and estimated (design-based) or predicted (model-based) the mean mercury concentration and constructed 95% confidence (design-based) and 95% (model-based) prediction intervals. For the model-based analyses, the exponential covariance was used. Table 3 shows the results from these analyses. Though we should not generalize these results to other samples from this population, we do mention a few findings. First, IRS-Design has the largest standard error. Second, compared to IRS-Design and IRS-Model, GRTS-Design and GRTS-Model are much closer to the true mean mercury concentration (have bias closer to zero) and have much lower standard errors (more precise intervals). Third, GRTS-Model has the least amount of bias and the lowest standard error (most precise interval). Finally, we note that for all sampling-analysis combinations, the true mean mercury concentration (103.2 ng / g) is within the bounds of the combination's 95% interval.

| Approach | True Mean | Est/Pred | SE | 95% LB | 95% UB |
|---|---|---|---|---|---|
| IRS-Design | 103.2 | 112.7 | 8.8 | 95.4 | 129.9 |
| IRS-Model | 103.2 | 110.5 | 7.9 | 95.0 | 125.9 |
| GRTS-Design | 103.2 | 101.8 | 6.1 | 89.8 | 113.7 |
| GRTS-Model | 103.2 | 102.3 | 5.9 | 90.8 | 113.9 |

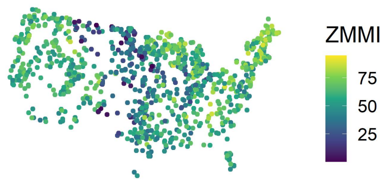
Table 3: For each sampling-analysis combination (Approach), the true mean mercury concentration (True Mean), estimates/predictions (Est/Pred), standard errors (SE), lower 95% interval bounds (95% LB), and upper 95% interval bounds (95% UB) for mean mercury concentration computed using a sample of 100 lakes in the NLA data.

### *3.3. New Application*

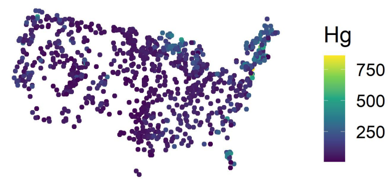
## 4. Discussion

ADD EXTRAS LIKE ANISOTROPY AND UNEQUAL INCLUSION PROBABILITIES

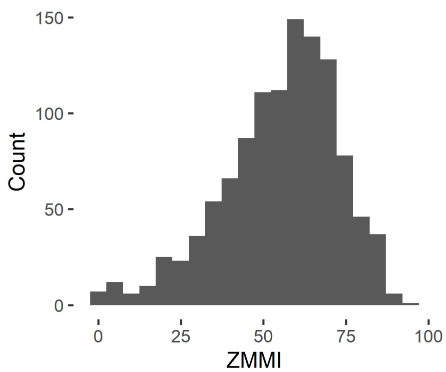
The design-based and model-based approaches to statistical inference are fundamentally different paradigms. The design-based approach relies on random sampling to estimate population parameters. The model-based approach relies on distributional assumptions to predict realized values of a stochastic process. Though the model-based approach does not rely on random sampling, it can still be beneficial as a way to guard against preferential sampling. While the design-based and model-based approaches have often been compared in the literature from theoretical and analytical perspectives, our contribution lies in studying them in a spatial context while implementing spatially balanced sampling and the design-based, local neighborhood variance estimator. Aside from the theoretical differences described, a few analytical findings from the simulation study are particularly notable. First, independent of the analysis approach, we found no reason to prefer IRS over GRTS when sampling spatial data – GRTS-Design and GRTS-Model generally had similar rMS(P)E as their IRS counterparts when there was no spatial covariance and lower rMS(P)E than their IRS counterparts when there was spatial covariance. Second, the sampling decision (IRS vs GRTS) is most important when using a design-based analysis. Though GRTS-Model still had lower rMS(P)E than IRS-Model, the model-based analysis mitigated most of the rMS(P)E inefficiencies that result from the IRS samples lacking spatial balance. Third, as the strength of spatial covariance increases, the gap in rMS(P)E and MStdE between IRS-Design and the other sampling-analysis combinations also increases, likely because IRS-Design is the only combination that ignores spatial locations in sampling and analysis. Fourth and finally, when



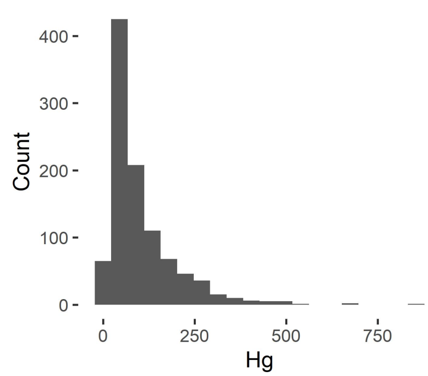

(a) Spatial map of the ZMMI population.



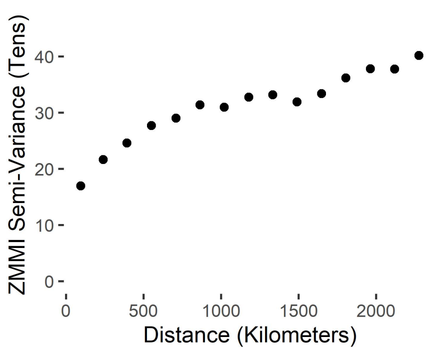

(b) Spatial map of mercury (Hg) population.



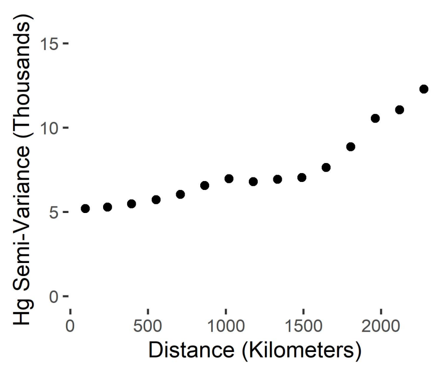

(c) Histogram of the ZMMI population.




(d) Histogram of the mercury (Hg) population.




(e) Semivariogram of the ZMMI population.




(f) Semivariogram of the mercury (Hg) population.

Figure 5: Exploratory graphics of the ZMMI and mercury (Hg) populations in the National Lakes Assessment (NLA) 2012 data.

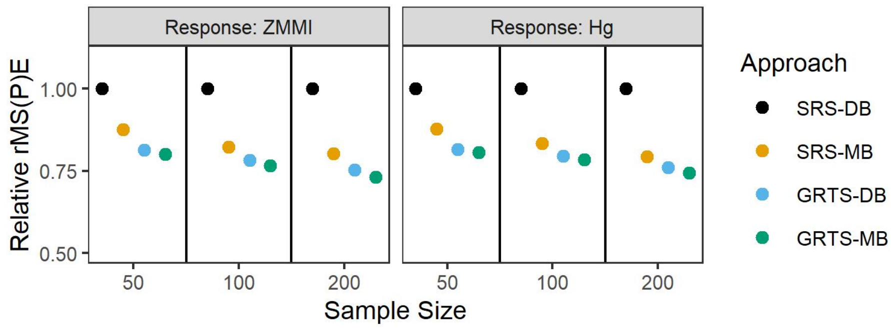



Figure 6: Relative rMS(P)E in the data study for the four sampling-analysis combinations. The rows indicate the proportion of dependent error and the columns indicate the response type.

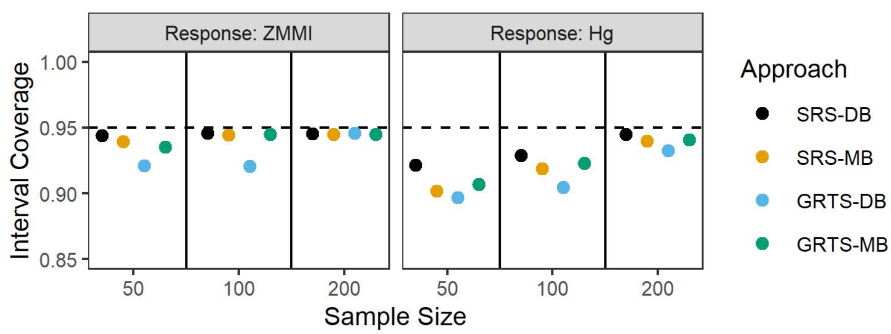



Figure 7: Interval coverage in the data study for the four sampling-analysis combinations. The rows indicate the proportion of dependent error and the columns indicate the response type. The solid, black line represents 95% coverage.

the response was normal, interval coverage for all sampling-analysis combinations was usually close to 95% for all sample sizes; when the response was lognormal, interval coverage for all sampling-analysis combinations was usually between 90% and 95% and closest to 95% when $n = 200$.

There are several benefits and drawbacks of the design-based and model-based approaches for finite population spatial data. Some we have discussed, but others we have not, and they are worthy of consideration in future research. Design-based approaches are often computationally efficient, while model-based approaches can be computationally burdensome, especially for likelihood-based estimation methods like REML that rely on inverting a covariance matrix. The design-based approach easily handles binary data through a straightforward application of the Horvitz-Thompson estimator. In contrast, analyzing binary data using a model-based approach generally requires a logistic mixed regression model, which can be challenging to estimate and interpret (Bolker et al., 2009). The design-based approach yields valid results because the sampling plan and inclusion probabilities are specified directly by the researcher, while the model-based approach may not yield valid results if the assumptions made do not not accurately capture reailty. The model-based approach, however, can more naturally quantify the relationship between covariates (predictor variables) and the response variable. The model-based approach also yields estimated spatial covariance parameters, which help better understand the dependence structure in the process in study. Model selection is also possible using model-based approaches and criteria such as cross validation, likelihood ratio tests, or AIC (Akaike, 1974). Model-based approaches are capable of more efficient small-area estimation than design-based approaches by leveraging distributional assumptions in areas with few observed units. Model-based approaches can also compute unit-by-unit predictions at unobserved locations and use them to construct informative visualizations like smoothed maps. Brus and De Gruijter (1997) provide a more thorough discussion regarding the benefits and drawbacks of the two approaches. In short, when deciding whether the design-based or model-based approach is more appropriate to implement, the benefits and drawbacks of each approach should be considered alongside the particular goals of the study.

## Acknowledgments

We would like to thank the editors and anonymous reviewers for their thoughtful comments which greatly improved the manuscript.

The views expressed in this manuscript are those of the authors and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency or the National Oceanic and Atmospheric Administration. Any mention of trade names, products, or services does not imply an endorsement by the U.S. government, the U.S. Environmental Protection Agency, or the National Oceanic and Atmospheric Administration. The U.S. Environmental Protection Agency and National Oceanic and Atmospheric Administration do not endorse any commercial products, services, or enterprises.

## Conflict of Interest Statement

There are no conflicts of interest for any of the authors.

## Author Contribution Statement

All authors conceived the ideas; All authors designed the methodology; MD and MH performed the simulations and analyzed the data; MD and MH led the writing of the manuscript; All authors contributed critically to the drafts and gave final approval for publication.

## Data and Code Availability

This manuscript has a supplementary **R** package that contains all of the data and code used in its creation. The supplementary **R** package is hosted on GitHub. Instructions for download at available at

https://github.com/michaeldumelle/DvMsp.

If the manuscript is accepted, this repository will be archived in Zenodo.

## Supporting Information

In the supporting information, we provide tables of summary statistics for all 36 simulation scenarios.

## References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19, 716–723.

Barabesi, L., Franceschi, S., 2011. Sampling properties of spatial total estimators under tessellation stratified designs. Environmetrics 22, 271–278.

Benedetti, R., Piersimoni, F., 2017. A spatially balanced design with probability function proportional to the within sample distance. Biometrical Journal 59, 1067–1084.

Benedetti, R., Piersimoni, F., Postiglione, P., 2017. Spatially balanced sampling: A review and a reappraisal. International Statistical Review 85, 439–454.

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H., White, J.-S.S., 2009. Generalized linear mixed models: A practical guide for ecology and evolution. Trends in ecology & evolution 24, 127–135.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

Brus, D., De Gruijter, J., 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-dased sampling strategies for soil (with discussion). Geoderma 80, 1–44.

Brus, D.J., 2021. Statistical approaches for spatial sample survey: Persistent misconceptions and new developments. European Journal of Soil Science 72, 686–703.

Brus, D.J., DeGruijter, J.J., 1993. Design-based versus model-based estimates of spatial means: Theory and application in environmental soil science. Environmetrics 4, 123–152.

Chan-Golston, A.M., Banerjee, S., Handcock, M.S., 2020. Bayesian inference for finite populations under spatial process settings. Environmetrics 31, e2606.

Chiles, J.-P., Delfiner, P., 1999. Geostatistics: Modeling Spatial Uncertainty. John Wiley & Sons, New York.

Cicchitelli, G., Montanari, G.E., 2012. Model-assisted estimation of a spatial population mean. International Statistical Review 80, 111–126.

Cooper, C., 2006. Sampling and variance estimation on continuous domains. Environmetrics 17, 539–553.

Cressie, N., 1993. Statistics for spatial data. John Wiley & Sons.

De Gruijter, J., Ter Braak, C., 1990. Model-free estimation from spatial samples: A reappraisal of classical sampling theory. Mathematical Geology 22, 407–415.

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

Dumelle, M., Kincaid, T.M., Olsen, A.R., Weber, M.H., 2022. Spsurvey: Spatial sampling design and analysis.

Fix, E., Hodges, J.L., 1989. Discriminatory analysis. Nonparametric discrimination: Consistency properties. International Statistical Review/Revue Internationale de Statistique 57, 238–247.

Grafström, A., 2012. Spatially correlated poisson sampling. Journal of Statistical Planning and Inference 142, 139–147.

Grafström, A., Lundström, N.L., 2013. Why well spread probability samples are balanced. Open Journal of Statistics 3, 36–41.

Grafström, A., Lundström, N.L., Schelin, L., 2012. Spatially balanced sampling through the pivotal method. Biometrics 68, 514–520.

Grafström, A., Matei, A., 2018. Spatially balanced sampling of continuous populations. Scandinavian Journal of Statistics 45, 792–805.

Hansen, M.H., Madow, W.G., Tepping, B.J., 1983. An evaluation of model-dependent and probability-sampling inferences in sample surveys. Journal of the American Statistical Association 78, 776–793.

Harville, D.A., 1977. Maximum likelihood approaches to variance component estimation and to related problems. Journal of the American Statistical Association 72, 320–338.

Higham, M., Ver Hoef, J., Frank, B., Dumelle, M., 2021. Sptotal: Predicting totals and weighted sums from spatial data.

Hofman, S.C., Brus, D., 2021. How many sampling points are needed to estimate the mean nitrate-n content of agricultural fields? A geostatistical simulation approach with uncertain variograms. Geoderma 385, 114816.

Horvitz, D.G., Thompson, D.J., 1952. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 47, 663–685.

Lohr, S.L., 2009. Sampling: Design and analysis. Nelson Education.

Patterson, H.D., Thompson, R., 1971. Recovery of inter-block information when block sizes are unequal. Biometrika 58, 545–554.

Robertson, B., Brown, J., McDonald, T., Jaksons, P., 2013. BAS: Balanced acceptance sampling of natural resources. Biometrics 69, 776–784.

Robertson, B., McDonald, T., Price, C., Brown, J., 2018. Halton iterative partitioning: Spatially balanced sampling via partitioning. Environmental and Ecological Statistics 25, 305–323.

Särndal, C.-E., Swensson, B., Wretman, J., 2003. Model assisted survey sampling. Springer Science & Business Media.

Schabenberger, O., Gotway, C.A., 2017. Statistical methods for spatial data analysis. CRC press.

Sen, A.R., 1953. On the estimate of the variance in sampling with varying probabilities. Journal of the Indian Society of Agricultural Statistics 5, 127.

Sterba, S.K., 2009. Alternative model-based and design-based frameworks for inference from samples to populations: From polarization to integration. Multivariate Behavioral Research 44, 711–740.

Stevens, D.L., Olsen, A.R., 2003. Variance estimation for spatially balanced samples of environmental resources. Environmetrics 14, 593–610.

Stevens, D.L., Olsen, A.R., 2004. Spatially balanced sampling of natural resources. Journal of the American Statistical Association 99, 262–278.

USEPA, 2012. National lakes assessment 2012. https://www.epa.gov/national-aquatic-resource-surveys/national-results-and-regional-highlights-national-lakes-assessment.

Ver Hoef, J., 2002. Sampling and geostatistics for spatial data. Ecoscience 9, 152–161.

Ver Hoef, J.M., 2008. Spatial methods for plot-based sampling of wildlife populations. Environmental and Ecological Statistics 15, 3–13.

Ver Hoef, J.M., Temesgen, H., 2013. A comparison of the spatial linear model to nearest neighbor (k-nn) methods for forestry applications. PlOS ONE

8, e59129.

Walvoort, D.J., Brus, D., De Gruijter, J., 2010. An r package for spatial coverage sampling and random sampling from compact geographical strata by k-means. Computers & geosciences 36, 1261–1267.

Wang, J.-F., Jiang, C.-S., Hu, M.-G., Cao, Z.-D., Guo, Y.-S., Li, L.-F., Liu, T.-J., Meng, B., 2013. Design-based spatial sampling: Theory and implementation. Environmental Modelling & Software 40, 280–288.

Wolfinger, R., Tobias, R., Sall, J., 1994. Computing gaussian likelihoods and their derivatives for general linear mixed models. SIAM Journal on Scientific Computing 15, 1294–1310.