

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 and real data, a variety of sample sizes, location layouts, dependence structures, and response types are considered. 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 locations into sampling. We also found that model-based approaches tend

^{*}Corresponding Author: Michael Dumelle (Dumelle.Michael@epa.gov)

Preprint submitted to *Methods in Ecology and Evolution*

April 13, 2022

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 generating the sample. Each paradigm has a deep historical context (Sterba, 2009) and its own set of benefits and drawbacks (Hansen et al., 1983, p. @brus1997random). In this manuscript, we compare the design-based and model-based approaches for finite population spatial data.

Spatial data are data that have some sort of spatial index, usually via coordinates. 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)).

While comparisons between design-based and model-based approaches have been studied in spatial contexts, our contribution is comparing design-based approaches specifically built for spatial data 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 conterminous United States. An infinite population contains an infinite number of population units; an example is locations within a single lake.

82 The rest of the manuscript is organized as follows. In Section 1.1, we introduce
 83 and provide relevant background for design-based and model-based approaches
 84 to finite population spatial data. In Section 2, we describe how we intend to
 85 compare performance of the approaches using simulated and real data. In Section
 86 3, we present analysis results for the simulated and real data. And in Section 4,
 87 we end with a discussion and provide directions for future research.

88 *1.1. Background*

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

93 *1.1.1. Comparing Design-Based and Model-Based Approaches*

94 The design-based approach assumes the population is fixed. Randomness
 95 is incorporated via the selection of population units according to a sampling
 96 design. A sampling design assigns a probability of selection to each sample
 97 (subset of population units). Some examples of commonly used sampling designs
 98 include simple random sampling, stratified random sampling, and cluster sam-
 99 pling. The inclusion probability of a population unit follows by summing each
 100 sample's probability of selection over all samples that contain the population
 101 unit. Inclusion probabilities are later used to estimate population parameters.

102 When samples are chosen in a manner such that the layout of sampled units
 103 reflects the layout of the population units, we call the resulting sample spatially
 104 balanced. By “reflecting the layout of the population units”, we mean that if
 105 population units are concentrated in specific areas, the units in the sample should
 106 be concentrated in the same areas. Because spatially balanced samples reflect
 107 the layout of the population units, they are not necessarily spread out in space in

108 some equidistant manner. One approach to selecting spatially balanced samples
 109 is the generalized random tessellation stratified (GRTS) algorithm (Stevens and
 110 Olsen, 2004), which we discuss in more detail in Section 1.1.2.

111 Fundamentally, the design-based approach combines the randomness of the
 112 sampling design with the data collected via the sample to justify the estimation
 113 and uncertainty quantification of fixed, unknown parameters of a population (e.g.,
 114 a population mean). Treating the data as fixed and incorporating randomness
 115 through the sampling design yields estimators having very few other assumptions.
 116 Confidence intervals for these types of estimators are typically derived using
 117 limiting arguments that incorporate all possible samples. Sample means, for
 118 example, are asymptotically normal (Gaussian) by the Central Limit Theorem
 119 (under some assumptions). If we repeatedly select samples from the population,
 120 then 95% of all 95% confidence intervals constructed from a procedure with
 121 appropriate coverage will contain the true fixed population mean. Särndal et al.
 122 (2003) and Lohr (2009) provide thorough reviews of the design-based approach.

123 The model-based approach assumes the population is a random realization of
 124 a data-generating stochastic process (superpopulation). Randomness is formally
 125 incorporated through distributional assumptions on this process. Strictly speak-
 126 ing, randomness need not be incorporated through random sampling, though
 127 Diggle et al. (2010) warn against preferential sampling. Preferential sampling
 128 occurs when the process generating the data locations and the process being
 129 modeled are not independent of one another. To guard against preferential sam-
 130 pling, model-based approaches can implement some form of random sampling,
 131 though it is common for model-based approaches to sample non-randomly. When
 132 model-based approaches do implement random sampling, the inclusion proba-
 133 bilities are ignored when analyzing the sample (in contrast to the design-based
 134 approach, which relies on these inclusion probabilities to analyze the sample).

135 Instead of estimating fixed, unknown population parameters, as in the design-
136 based approach, often the goal of model-based inference is to predict a realized
137 variable. For example, suppose the realized mean of all population units (the
138 realized population mean) is the variable of interest. Instead of a fixed, unknown
139 mean, we are predicting the value of the mean, a random variable. Prediction
140 intervals are then derived using assumptions of the data-generating stochastic
141 process. If we repeatedly generate realizations from the same process and select
142 samples, then 95% of all 95% prediction intervals constructed from a procedure
143 with appropriate coverage will contain their respective realized means. Cressie
144 (1993) and Schabenberger and Gotway (2017) provide thorough reviews of model-
145 based approaches for spatial data. In Fig. 1, we provide a visual comparison
146 of the design-based and model-based approaches (Ver Hoef (2002) and Brus
147 (2021) provide similar figures). This figure contrasts the design-based approach
148 with a fixed population and random sampling to the model-based approach with
149 random populations and non-random sampling.

150 1.1.2. *Spatially Balanced Design and Analysis*

151 We previously mentioned that the design-based approach can be used to
152 select spatially balanced samples. Spatially balanced samples are useful because
153 parameter estimates from these samples tend to vary less than parameter es-
154 timates from samples lacking spatial balance (Barabesi and Franceschi, 2011;
155 Benedetti et al., 2017; Grafström and Lundström, 2013; Robertson et al., 2013;
156 Stevens and Olsen, 2004; Wang et al., 2013). The first spatially balanced sam-
157 pling algorithm to see widespread use was the generalized random tessellation
158 stratified (GRTS) algorithm (Stevens and Olsen, 2004). To quantify the spatial
159 balance of a sample, Stevens and Olsen (2004) proposed loss metrics based on
160 Voronoi polygons (i.e., Dirichlet Tessellations). After the GRTS algorithm was
161 developed, several other spatially balanced sampling algorithms emerged, includ-

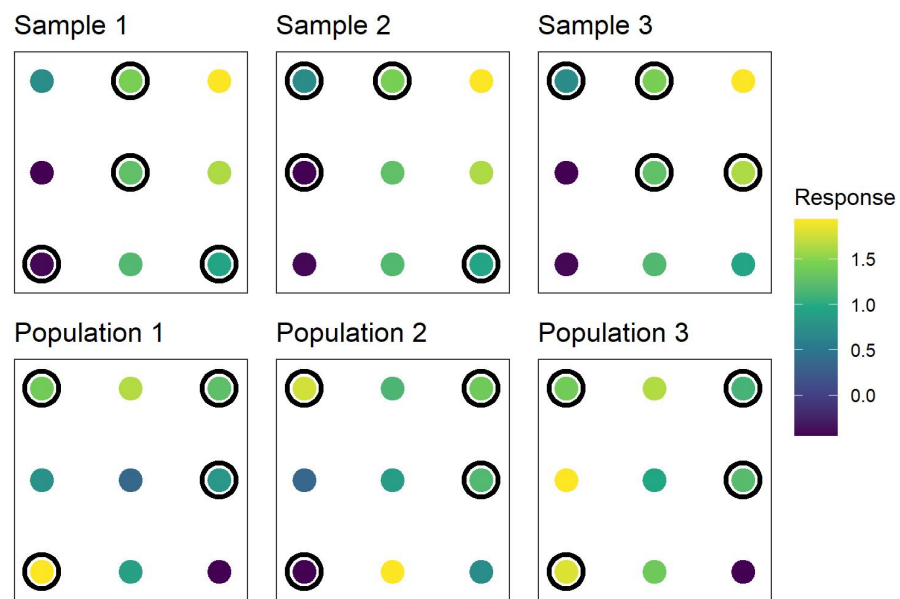


Figure 1: A visual comparison of the design-based and model-based approaches. In the top row, the design-based approach is highlighted. There is one fixed population with nine population units and three random samples of size four (points circled are those sampled). The response values at each site are fixed. 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 response values at each site are random.

ing 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 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 additional population units that can be sampled when a population unit
originally selected can no longer be sampled. A couple reasons why an originally
selected site can no longer be sampled include its location being physically
inaccessible or on private land that the researcher does not have permission to
access.

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 cell is randomly
assigned a 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

189 k steps, until all level- k cells have inclusion probability sums no larger than one.

190 Then $\mathcal{A}_k \equiv \{a_1 \dots a_k : a_1 = 0, 1, 2, 3; \dots; a_k = 0, 1, 2, 3\}$.

191 After determining \mathcal{A}_k , it is placed into hierarchical order. Hierarchical order
 192 is a numeric order that first sorts \mathcal{A}_k by the level-one addresses from smallest
 193 to largest, then sorts \mathcal{A}_k by the level-two addresses from smallest to largest, and so
 194 on. For example, \mathcal{A}_2 in hierarchical order is the set
 195 $\{00, 01, 02, 03, 10, \dots, 13, 20, \dots, 23, 30, \dots, 33\}$. Because hierarchical ordering sorts
 196 by level-one cells, then level-two cells, and so on, population units that have
 197 similar hierarchical addresses tend to be nearby one another in space. Next each
 198 population unit is mapped to a one-dimensional line in hierarchical order where
 199 each population unit's inclusion probability equals its line-length. If a level- k
 200 cell has multiple population units in it, they are randomly placed within the
 201 cell's respective line segment. A uniform random variable is then simulated in
 202 $[0, 1]$ and a systematic sample is selected on the line, yielding n sample points for
 203 a sample size n . Each of these sample points falls on some population unit's line
 204 segment, and thus that population unit is selected in the sample. For further
 205 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 τ is a population total, the Horvitz-Thompson estimator for τ , denoted by $\hat{\tau}_{ht}$, is given by

$$\hat{\tau}_{ht} = \sum_{i=1}^n z_i \pi_i^{-1}, \quad (1)$$

206 where z_i is the value of the i th population unit in the sample, π_i is the inclusion
 207 probability of the i th population unit in the sample, and n is the sample size. An
 208 estimate of the population mean is obtained by dividing $\hat{\tau}_{ht}$ by N , the number
 209 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 π_{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 π_{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 contribution from each sample unit (population unit in the sample) to the overall variance is dominated by local variation. Thus the local neighborhood variance estimator is a weighted sum of variance estimates from each sample unit’s local neighborhood. These local neighborhoods contain the sample unit itself and its three nearest neighbors among all other sample units. 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 more details, see Ver Hoef (2008). Let $\mathbf{z} \equiv \{z(s_1), z(s_2), \dots, z(s_N)\}$ be an $N \times 1$ response vector at locations s_1, s_2, \dots, 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

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

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

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

where σ_1^2 is the variance parameter that quantifies the spatially dependent

variability, σ_2^2 is the variance parameter that quantifies that spatially independent variability, ϕ 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, σ_1^2 is often called the partial sill, σ_2^2 is often called the nugget, and ϕ 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 β and δ (Patterson and Thompson, 1971). Minus twice the REML log-likelihood of the sampled sites is given by

$$\ln |\Sigma| + (z_s - X_s \tilde{\beta})^T \Sigma_{ss}^{-1} (z_s - X_s \tilde{\beta}) + \ln |X_s^T \Sigma_{ss}^{-1} X_s| + (n - p) \ln(2\pi), \quad (4)$$

where $\tilde{\beta} = (X_s^T \Sigma_{ss}^{-1} X_s)^{-1} X_s^T \Sigma_{ss}^{-1} z_s$ and Σ_{ss} is the covariance matrix of the sampled sites. Minimizing Equation 4 yields $\hat{\delta}_{reml}$, the REML estimates of δ . Then β_{reml} , the REML estimate of β , is given by $(X_s^T \hat{\Sigma}_{ss}^{-1} X_s)^{-1} X_s^T \hat{\Sigma}_{ss}^{-1} z_s$, where $\hat{\Sigma}_{ss}$ is Σ_{ss} evaluated at $\hat{\delta}_{reml}$.

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

In this section we describe how we used simulated and real data to investigate performance between simple random sampling without replacement (SRS) and GRTS sampling as well as performance between design-based (DB) and model-based (MB) inference. In SRS and GRTS sampling, all population units had equal inclusion probabilities. The important distinction between SRS and GRTS is that SRS ignores spatial locations while sampling but GRTS explicitly incorporates them. Together, the two sampling plans (SRS and GRTS) combined with the two inference approaches (DB and MB) yielded four sampling-inference combinations: SRS-DB, SRS-MB, GRTS-DB, and GRTS-MB. For SRS-DB, the Horvitz-Thompson estimator (1) was used to estimate means and the commonly-used SRS variance formula (Lohr, 2009; Särndal et al., 2003) was used to estimate the variance. This variance formula is given by

$$\frac{f[\sum_{i=1}^n (z_i - \bar{z})^2]}{n(n-1)}, \quad (5)$$

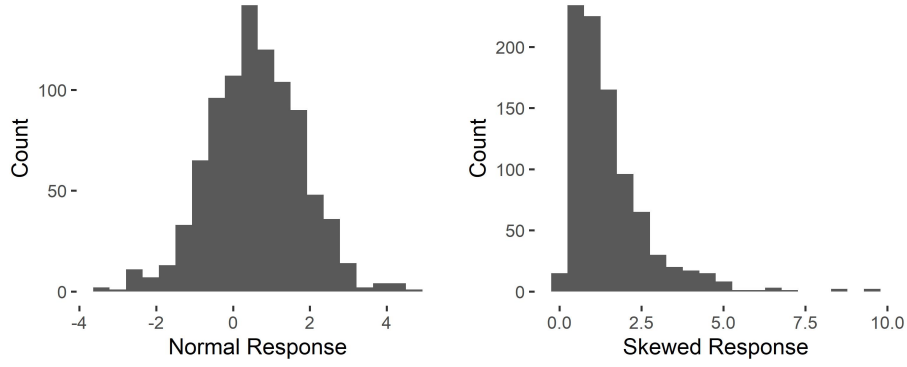
where z_i is the i th response value, \bar{z} is the mean of all z_i , n is the sample size, N is the population size, and $f = (1 - n/N)$ (f is often called the finite population correction factor). For GRTS-DB, the Horvitz-Thompson estimator was used to estimate means and the local neighborhood variance was used to estimate

variances. For SRS-MB and GRTS-MB, FPBK was used to estimate means and variances and parameters were estimated using restricted maximum likelihood.

We used simulated data to compare the sampling-inference combinations across many realized populations from the same data-generating stochastic process. With the simulated data, we were in control of the data-generating stochastic process and the random sampling process. We used real data from the 2012 National Lakes Assessment (USEPA, 2012) to compare the sampling-inference combinations within a single realized population (which is typically the case in reality). With the real data, we were in control of only the random sampling process.

2.1. Simulated Data

We evaluated performance of the four sampling-inference combinations 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 skewed. For the normal response type, the response was simulated using mean-zero random errors with the exponential covariance (Equation 3) for three proportions of dependent random error (DRE): 0% DRE, 50% DRE, and 90% DRE. Recall the proportion of DRE is represented by $\sigma_1^2/(\sigma_1^2 + \sigma_2^2)$, where σ_1^2 and σ_2^2 are the DRE variance and independent random error (IRE) variance from Equation 3, respectively. The total variance, $\sigma_1^2 + \sigma_2^2$, was always 2. The distance parameter was always $\sqrt{2}/3$, chosen so that the correlation in the DRE decayed



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

Figure 2: Histograms of realized populations simulated for the normal and skewed responses using the random layout and 50% DRE.

to nearly zero at $\sqrt{2}$, the largest possible distance between two population units in the domain. For the skewed 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 skewed random variable whose total variance was 2. The skewed responses were used to evaluate performance of the sampling-inference approaches for data that were not normal but were still estimated using REML, which relies on a normal log-likelihood. Figure 2 shows an example of a realized population for the normal and skewed responses using the random layout and 50% DRE.

In each of the 36 simulation scenarios, there were 2000 independent simulation trials. Within each simulation scenario and 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. With the model-based analyses, covariance parameters were estimated (using REML) separately for each trial. After all 2000 trials, we summarized the long-run performance of the sampling-inference combination in each scenario by calculating mean bias,

320 root-mean-squared error, and interval coverage. Mean bias is taken as the average
 321 deviation between each trial's estimated (or predicted) mean and its realized
 322 mean: $\frac{1}{n} \sum_{i=1}^{2000} (\hat{\mu}_i - \mu_i)$, where i indexes simulation trials. Root-mean-squared
 323 error is taken as the square root of the average squared deviation between each
 324 trial's estimated (or predicted) mean and its realized mean: $\sqrt{\frac{1}{n} \sum_{i=1}^{2000} (\hat{\mu}_i - \mu_i)^2}$.
 325 Interval coverage is taken as the proportion of simulation trials where the
 326 realized mean is contained in its 95% confidence (or prediction) interval. These
 327 intervals are constructed using the normal distribution – justification comes
 328 from the asymptotic normality of means via the central limit theorem (under
 329 some assumptions). Quantifying these metrics is important because together,
 330 they give us an idea of the accuracy (mean bias), spread (RMSE), and validity
 331 (interval coverage) of the sampling-inference combinations.

332 *2.2. National Lakes Assessment (Real) Data*

333 The United States Environmental Protection Agency (USEPA), states, and
 334 tribes periodically conduct National Aquatic Research Surveys (NARS) to assess
 335 the water quality of various bodies of water in the contiguous United States.
 336 One component of NARS is the National Lakes Assessment (NLA), which
 337 measures various aspects of lake health and water quality. We focus on analyzing
 338 zooplankton multi-metric indices (ZMMI) and mercury concentrations in parts
 339 per billion (Hg ppb) from the 2012 NLA. For ZMMI, data were collected at 1035
 340 unique lakes. At less than 10% of lakes, two ZMMI replicates were collected.
 341 These were averaged for the purposes of our study so that each lake had one
 342 measurement for ZMMI. For Hg ppb, data were collected at 995 unique lakes
 343 (and there were no replicates like for ZMMI). The ZMMI and Hg ppb data are
 344 shown as spatial maps and as histograms in Figure 3. The ZMMI data tend
 345 to be highest near the coasts, lowest in the Central United States, higher near
 346 the coasts, are relatively symmetric, and have a mean of 55.05. The Hg ppb

data tend to be highest in the Northeastern United States, lowest elsewhere, are skewed, and have a mean of 103.16 ppb. Also in Figure 3 are separate spatial semivariograms for ZMMI and Hg ppb. The spatial semivariogram quantifies the the halved average squared differences (semivariance) of responses whose separation (distance) falls within some distance class. The spatial semivariance is closely related to the spatial covariance, and spatial semivariograms are often used to gauge the strength of spatial dependence in data. Both ZMMI and Hg ppb seem to have moderately strong spatial dependence (Figure 3), as the semivariance increases steadily with distance (meaning that observations nearby one another tend to be more similar than observations far apart from one another).

We studied performance of the four sampling-inference combinations by selecting 2000 random IRS and GRTS samples of size $n = 50$, $n = 100$, and $n = 200$ from the realized ZMMI and Hg ppb populations and then analyzing the samples using MB and DB inference. In total, there were six separate scenarios (two responses and three sample sizes). We used the same evaluation metrics as for the simulated data: mean bias, RMSE, and interval coverage. Mean bias is taken as the average deviation between each sample's estimated (or predicted) mean and the population mean (of ZMMI or Hg ppb): $\frac{1}{n} \sum_{i=1}^{2000} (\hat{\mu}_i - \mu)$, where i indexes simulation trials and μ is the population mean. Root-mean-squared error is taken as the square root of the average squared deviation between each sample's estimated (or predicted) mean and its population mean: $\sqrt{\frac{1}{n} \sum_{i=1}^{2000} (\hat{\mu}_i - \mu)^2}$. Interval coverage is taken as the proportion of simulation trials where the population mean is contained in its 95% confidence (or prediction) interval. These intervals are constructed using the normal distribution.

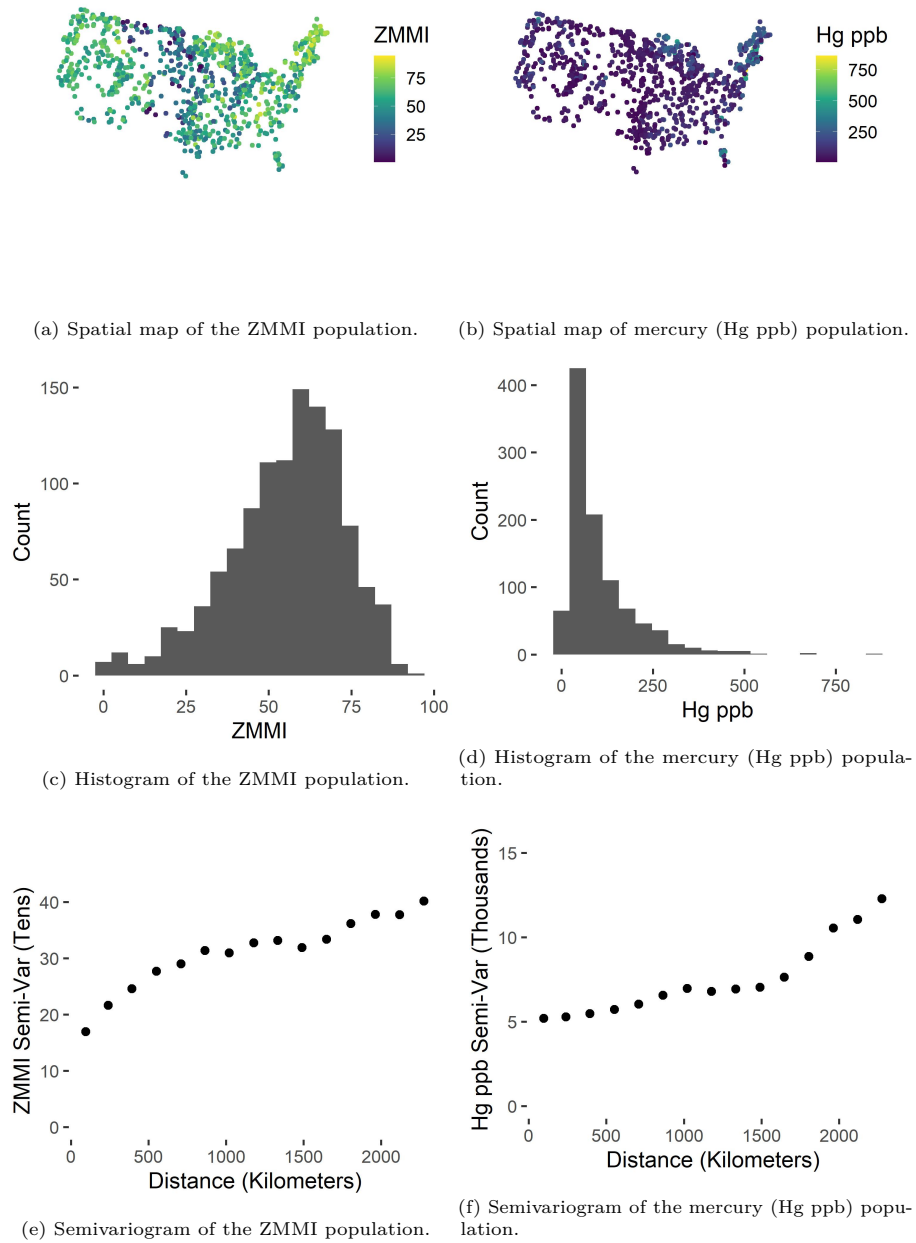


Figure 3: Exploratory graphics of the ZMMI and mercury (Hg ppb) populations in the National Lakes Assessment (NLA) 2012 data.

3. Results

3.1. Simulated Data

Mean bias was nearly zero for all four sampling-inference 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.

We define the relative RMSE as a ratio with numerator given by the RMSE for a sampling-inference combination and the denominator given by the RMSE for SRS-DB. Relative RMSEs for the random location layout are provided in Fig. 4. When there is no spatial covariance (Fig. 4, “DRE%: 0%”), the four sampling-inference combinations have approximately equal RMSE. In these scenarios, using GRTS sampling or model-based inference does not generally increase efficiency compared to SRS-DB. When there is spatial covariance (Fig. 4, “DRE%: 0%” and “DRE%: 50%”), GRTS-MB tends to have the lowest RMSE, followed by GRTS-DB, SRS-MB, and finally SRS-DB, though the difference in relative RMSE among GRTS-MB, GRTS-DB, and SRS-MB is small. As the strength of spatial covariance increases, the gap in RMSE between SRS-DB and the other sampling-inference combinations widens. Finally we note that when there is spatial covariance, SRS-MB has a much lower RMSE than SRS-DB, suggesting that the lack of efficiency from SRS is largely mitigated by model-based inference. These RMSE conclusions are similar to those observed in the grid location layout, so we omit a figure and discussion regarding the grid location layout here. Tables for RMSE in all 36 simulation scenarios are provided in the supporting information.

95% interval coverage for each of the four sampling-inference combinations in the random location layout is shown in Fig. 5. Within each simulation scenario, all sampling-inference combinations tend to have fairly similar interval

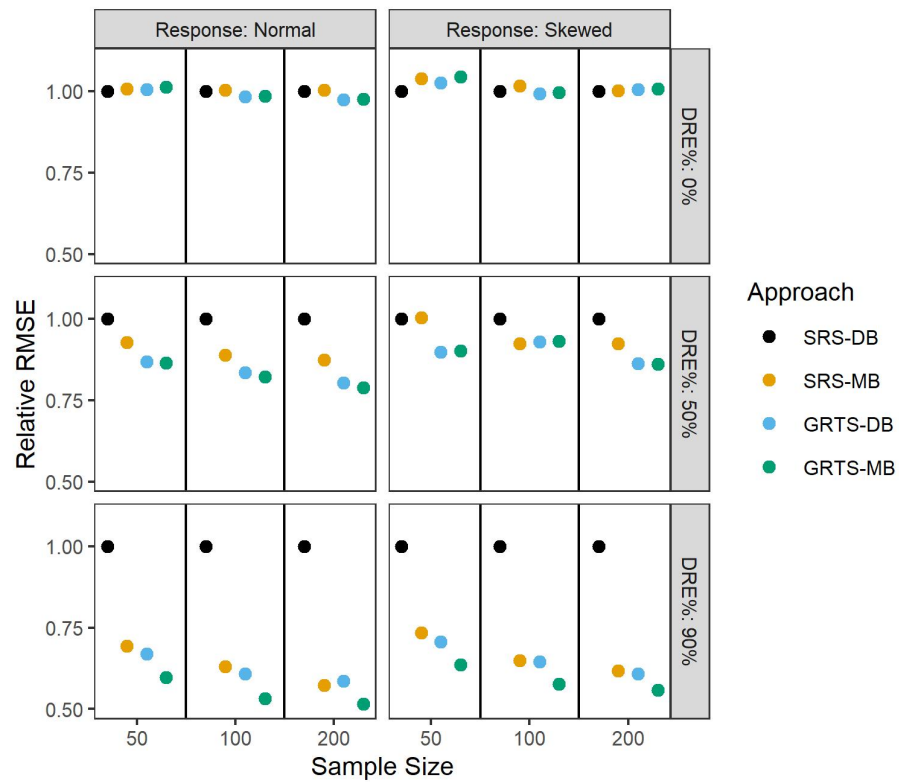


Figure 4: Relative RMSE in the simulation study for the four sampling-inference combinations and three sample sizes in the random location layout. The rows indicate the proportion of dependent error and the columns indicate the response type. The solid, black lines separate the sample sizes

coverage, though when $n = 50$ or $n = 100$, GRTS-DB coverage is usually a few percentage points lower than the other combinations, which suggests that the local neighborhood variance estimate may be slightly too small for small n . Coverage in the normal response scenarios was usually near 95%, while coverage in the skewed response scenarios usually varied from 90% to 95% but increased with the sample size. At a sample size of 200, all four sampling-inference 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 figure and discussion regarding the grid location layout here. Tables for interval coverage in all 36 simulation scenarios are provided in the supporting information.

3.2. National Lakes Assessment (Real) Data

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

The relative RMSE of both ZMMI (symmetric response) and Hg ppb (skewed response) for all four sampling-inference combinations are shown in Fig. 6. GRTS-MB has the lowest RMSE, followed by GRTS-DB, SRS-MB, and then SRS-DB. The difference in RMSE among GRTS-MB and GRTS-DB tends to be quite small. When $n = 50$, SRS-MB RMSE is approximately evenly between IRS-DB RMSE and GRTS-MB RMSE, but for the larger sample sizes ($n = 100$, $n = 200$), SRS-MB RMSE is closer to GRTS-MB RMSE. Lastly we note that GRTS-MB, GRTS-DB, and SRS-MB all have noticeably lower RMSE than SRS-DB.

95% interval coverage of both ZMMI and Hg ppb for all four sampling-inference combinations is shown in Fig. 5. When $n = 50$, interval coverage for both responses is too low, though interval coverage is higher for ZMMI

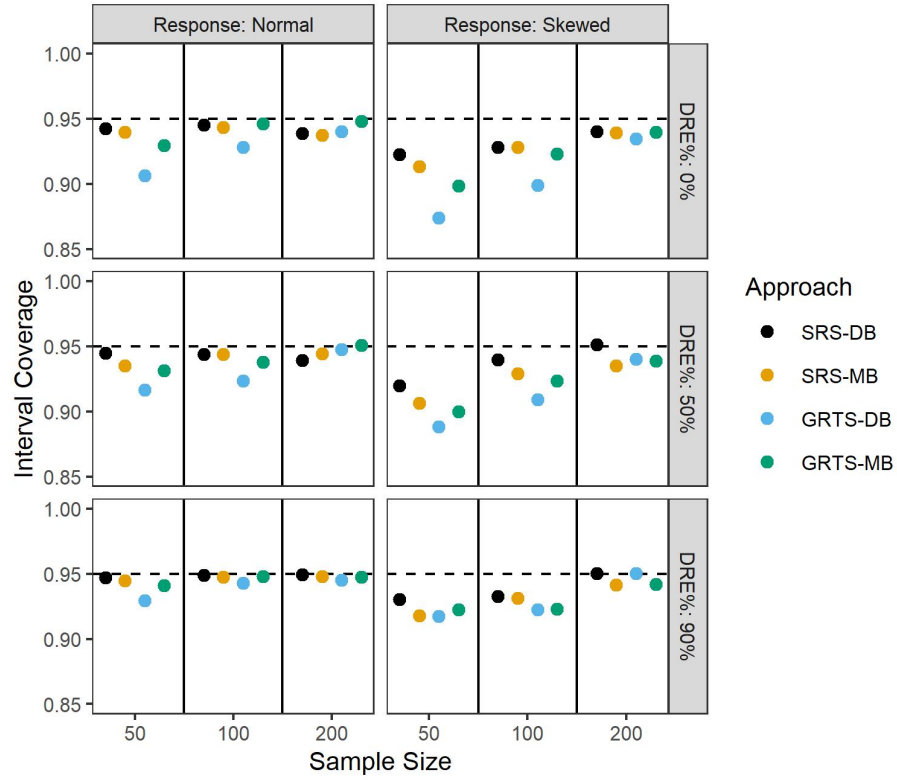


Figure 5: Interval coverage in the simulation study for the four sampling-inference combinations and three sample sizes in the random location layout. The rows indicate the proportion of dependent error and the columns indicate the response type. The solid black lines separate the sample sizes and the dashed black lines represent 95% coverage.

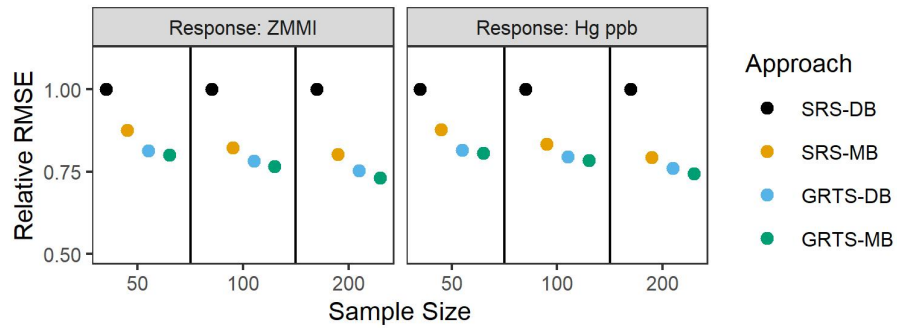


Figure 6: Relative RMSE in the data study for the four sampling-inference combinations. The rows indicate the proportion of dependent error and the columns indicate the response type.

(symmetric response) than for Hg ppb (skewed response). When $n = 100$, ZMMI interval coverage is approximately 95% except for GRTS-DB, which has coverage around 92%, while Hg ppb interval coverage ranges from approximately 90% (GRTS-DB) to 93% (GRTS-MB). When $n = 200$, ZMMI interval coverage is approximately 95% while Hg ppb interval coverage ranges from approximately 93% (GRTS-DB) to 95% (GRTS-MB). As with the simulated data, coverages for the NLA data tended to increase with the sample sizes, coverages tended to be higher symmetric responses than skewed responses, and regularly the local neighborhood variance was slightly too small for small n , yielding slightly lower interval coverages than the other sampling-inference combinations. Recall that model-based inference defines interval coverage properties across realized populations. With the simulated data, we evaluated interval coverage across realization populations, but for the NLA data, we evaluated interval coverage within a single realization for different samples. We did find that model-based coverages were similar to the design-based coverages, however, suggesting that in many cases it is reasonable to heuristically view data from separate samples as being from approximately separate realized populations. But generally, if model-based intervals constructed from many random samples of a single realized population show improper coverage, this does not necessarily imply a deficiency in model-based inference.

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

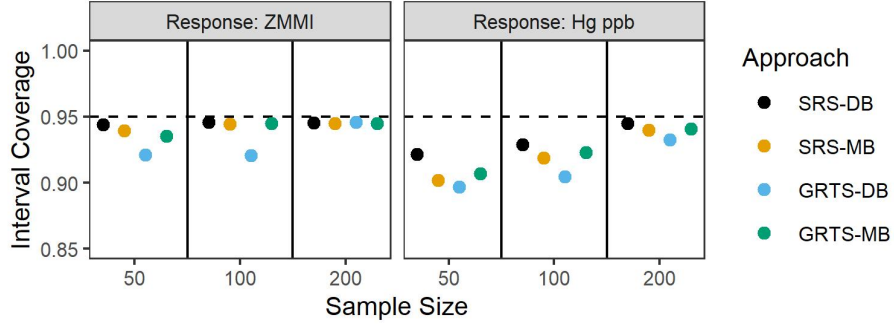


Figure 7: Interval coverage in the data study for the four sampling-inference combinations. The rows indicate the proportion of dependent error and the columns indicate the response type. The solid, black line represents 95% coverage.

452 on distributional assumptions to predict realized values of a stochastic process.
 453 Though the model-based approach does not rely on random sampling, it can still
 454 be beneficial as a way to guard against preferential sampling. While the design-
 455 based and model-based approaches have often been compared in the literature
 456 from theoretical and analytical perspectives, our contribution lies in studying
 457 them in a spatial context while implementing spatially balanced sampling and the
 458 design-based, local neighborhood variance estimator. Aside from the theoretical
 459 differences described, a few analytical findings from the simulation study are
 460 particularly notable. First, independent of the analysis approach, we found no
 461 reason to prefer IRS over GRTS when sampling spatial data – GRTS-DB and
 462 GRTS-MB generally had similar RMSE as their IRS counterparts when there
 463 was no spatial covariance and lower RMSE than their IRS counterparts when
 464 there was spatial covariance. Second, the sampling decision (IRS vs GRTS) is
 465 most important when using a design-based analysis. Though GRTS-MB still
 466 had lower RMSE than SRS-MB, the model-based analysis mitigated most of the
 467 RMSE inefficiencies that result from the IRS samples lacking spatial balance.
 468 Third, as the strength of spatial covariance increases, the gap in RMSE and
 469 MStdE between SRS-DB and the other sampling-inference combinations also

470 increases, likely because SRS-DB is the only combination that ignores spatial
 471 locations in sampling and analysis. Fourth and finally, when the response was
 472 normal, interval coverage for all sampling-inference combinations was usually
 473 close to 95% for all sample sizes; when the response was lognormal, interval
 474 coverage for all sampling-inference combinations was usually between 90% and
 475 95% and closest to 95% when $n = 200$.

476 AT LEAST HAVE DISCUSSION ABOUT MODEL BASED ASSUMPTIONS
 477 AND MOVE VALIDITY COMMENTS TO RESULTS SECTION.

478 There are several benefits and drawbacks of the design-based and model-
 479 based approaches for finite population spatial data. Some we have discussed,
 480 but others we have not, and they are worthy of consideration in future research.
 481 Design-based approaches are often computationally efficient, while model-based
 482 approaches can be computationally burdensome, especially for likelihood-based
 483 estimation methods like REML that rely on inverting a covariance matrix. The
 484 design-based approach easily handles binary data through a straightforward
 485 application of the Horvitz-Thompson estimator. In contrast, analyzing binary
 486 data using a model-based approach generally requires a logistic mixed regression
 487 model, which can be challenging to estimate and interpret (Bolker et al., 2009).
 488 The design-based approach yields valid results because the sampling plan and
 489 inclusion probabilities are specified directly by the researcher, while the model-
 490 based approach may not yield valid results if the assumptions made do not
 491 not accurately capture reality. The model-based approach, however, can more
 492 naturally quantify the relationship between covariates (predictor variables) and
 493 the response variable. The model-based approach also yields estimated spatial
 494 covariance parameters, which help better understand the dependence structure
 495 in the process in study. Model selection is also possible using model-based
 496 approaches and criteria such as cross validation, likelihood ratio tests, or AIC

497 (Akaike, 1974). Model-based approaches are capable of more efficient small-area
498 estimation than design-based approaches by leveraging distributional assumptions
499 in areas with few observed units. Model-based approaches can also compute
500 unit-by-unit predictions at unobserved locations and use them to construct
501 informative visualizations like smoothed maps. Brus and De Gruijter (1997)
502 provide a more thorough discussion regarding the benefits and drawbacks of the
503 two approaches. In short, when deciding whether the design-based or model-
504 based approach is more appropriate to implement, the benefits and drawbacks of
505 each approach should be considered alongside the particular goals of the study.

506 **Acknowledgments**

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

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

517 **Conflict of Interest Statement**

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

519 **Author Contribution Statement**

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

524 **Data and Code Availability**

525 This manuscript has a supplementary **R** package that contains all of the
526 data and code used in its creation. The supplementary **R** package is hosted on
527 GitHub. Instructions for download at available at
528 <https://github.com/michaeldumelle/DvMsp>.

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

530 **Supporting Information**

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

533 **References**

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

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

538 Benedetti, R., Piersimoni, F., 2017. A spatially balanced design with proba-
539 bility function proportional to the within sample distance. Biometrical Journal
540 59, 1067–1084.

- 541 Benedetti, R., Piersimoni, F., Postiglione, P., 2017. Spatially balanced
542 sampling: A review and a reappraisal. *International Statistical Review* 85,
543 439–454.
- 544 Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R.,
545 Stevens, M.H.H., White, J.-S.S., 2009. Generalized linear mixed models: A
546 practical guide for ecology and evolution. *Trends in ecology & evolution* 24,
547 127–135.
- 548 Breiman, L., 2001. Random forests. *Machine Learning* 45, 5–32.
- 549 Brus, D., De Gruijter, J., 1997. Random sampling or geostatistical modelling?
550 Choosing between design-based and model-based sampling strategies for soil
551 (with discussion). *Geoderma* 80, 1–44.
- 552 Brus, D.J., 2021. Statistical approaches for spatial sample survey: Persistent
553 misconceptions and new developments. *European Journal of Soil Science* 72,
554 686–703.
- 555 Brus, D.J., DeGruijter, J.J., 1993. Design-based versus model-based esti-
556 mates of spatial means: Theory and application in environmental soil science.
557 *Environmetrics* 4, 123–152.
- 558 Chan-Golston, A.M., Banerjee, S., Handcock, M.S., 2020. Bayesian inference
559 for finite populations under spatial process settings. *Environmetrics* 31, e2606.
- 560 Chiles, J.-P., Delfiner, P., 1999. *Geostatistics: Modeling Spatial Uncertainty*.
561 John Wiley & Sons, New York.
- 562 Cicchitelli, G., Montanari, G.E., 2012. Model-assisted estimation of a spatial
563 population mean. *International Statistical Review* 80, 111–126.
- 564 Cooper, C., 2006. Sampling and variance estimation on continuous domains.
565 *Environmetrics* 17, 539–553.
- 566 Cressie, N., 1993. *Statistics for spatial data*. John Wiley & Sons.
- 567 De Gruijter, J., Ter Braak, C., 1990. Model-free estimation from spatial

- 568 samples: A reappraisal of classical sampling theory. *Mathematical Geology* 22,
569 407–415.
- 570 Diggle, P.J., Menezes, R., Su, T.-l., 2010. Geostatistical inference under
571 preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied*
572 *Statistics)* 59, 191–232.
- 573 Dumelle, M., Kincaid, T.M., Olsen, A.R., Weber, M.H., 2022. *Spsurvey:*
574 *Spatial sampling design and analysis.*
- 575 Fix, E., Hodges, J.L., 1989. Discriminatory analysis. Nonparametric dis-
576 crimination: Consistency properties. *International Statistical Review/Revue*
577 *Internationale de Statistique* 57, 238–247.
- 578 Grafström, A., 2012. Spatially correlated poisson sampling. *Journal of*
579 *Statistical Planning and Inference* 142, 139–147.
- 580 Grafström, A., Lundström, N.L., 2013. Why well spread probability samples
581 are balanced. *Open Journal of Statistics* 3, 36–41.
- 582 Grafström, A., Lundström, N.L., Schelin, L., 2012. Spatially balanced
583 sampling through the pivotal method. *Biometrics* 68, 514–520.
- 584 Grafström, A., Matei, A., 2018. Spatially balanced sampling of continuous
585 populations. *Scandinavian Journal of Statistics* 45, 792–805.
- 586 Hansen, M.H., Madow, W.G., Tepping, B.J., 1983. An evaluation of model-
587 dependent and probability-sampling inferences in sample surveys. *Journal of the*
588 *American Statistical Association* 78, 776–793.
- 589 Harville, D.A., 1977. Maximum likelihood approaches to variance compo-
590 nent estimation and to related problems. *Journal of the American Statistical*
591 *Association* 72, 320–338.
- 592 Hofman, S.C., Brus, D., 2021. How many sampling points are needed to
593 estimate the mean nitrate-n content of agricultural fields? A geostatistical
594 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>.

- 622 Ver Hoef, J., 2002. Sampling and geostatistics for spatial data. *Ecoscience* 9,
623 152–161.
- 624 Ver Hoef, J.M., 2008. Spatial methods for plot-based sampling of wildlife
625 populations. *Environmental and Ecological Statistics* 15, 3–13.
- 626 Ver Hoef, J.M., Temesgen, H., 2013. A comparison of the spatial linear
627 model to nearest neighbor (k-nn) methods for forestry applications. *PIOS ONE*
628 8, e59129.
- 629 Walvoort, D.J., Brus, D., De Gruijter, J., 2010. An r package for spatial
630 coverage sampling and random sampling from compact geographical strata by
631 k-means. *Computers & geosciences* 36, 1261–1267.
- 632 Wang, J.-F., Jiang, C.-S., Hu, M.-G., Cao, Z.-D., Guo, Y.-S., Li, L.-F., Liu, T.-
633 J., Meng, B., 2013. Design-based spatial sampling: Theory and implementation.
634 *Environmental Modelling & Software* 40, 280–288.
- 635 Wolfinger, R., Tobias, R., Sall, J., 1994. Computing gaussian likelihoods and
636 their derivatives for general linear mixed models. *SIAM Journal on Scientific*
637 *Computing* 15, 1294–1310.