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)

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.

The rest of the manuscript is organized as follows. In Section 1.1, we introduce and provide relevant background for design-based and model-based approaches to finite population spatial data. In Section 2, we describe how we intend to compare performance of the approaches using simulated and real data. In Section 3, we present analysis reslts for the simulated and real data. 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 (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 sample's probability of selection over all samples that contain the 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

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 a

124 data-generating stochastic process. Randomness is formally incorporated through

125 distributional assumptions on this process. Strictly speaking, randomness need

126 not be incorporated through random sampling, though Diggle et al. (2010) warn

127 against preferential sampling. Preferential sampling occurs when the process

128 generating the data locations and the process being modeled are not independent

129 of one another. To guard against preferential sampling, model-based approaches

130 can implement some form of random sampling, though it is common for model-

131 based approaches to sample non-randomly. When model-based approaches

132 do implement random sampling, the inclusion probabilities are ignored when

133 analyzing the sample (in contrast to the design-based approach, which relies on

134 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 realizations 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). This figure contrasts the design-based approach with a fixed population and random sampling to the model-based approach with random populations and non-random sampling.

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 lacking spatial balance (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 the spatial balance of a sample, Stevens and Olsen (2004) proposed loss metrics based on Voronoi polygons (i.e., Dirichlet Tessellations). After the GRTS algorithm was developed, several other spatially balanced sampling algorithms emerged, includ-
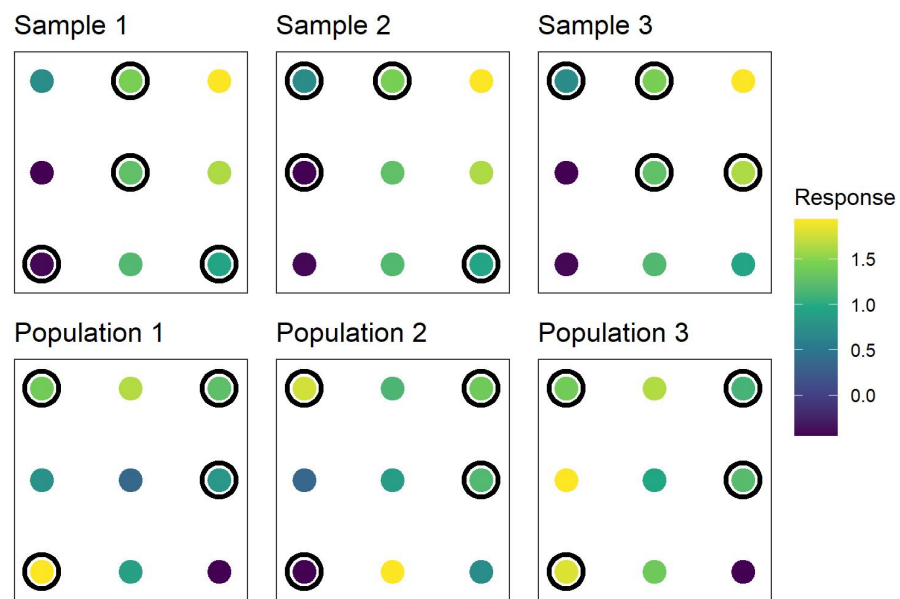
Figure 1: A visual comparison of the design-based and model-based approaches. In the top row, the design-based approach is highlighted. There is one fixed population with nine population units and three random samples of size four (points circled are those sampled). The response values at each site are 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 $\mathcal{A}_k \equiv \{a_1...a_k : a_1 = 0, 1, 2, 3; ...; a_k = 0, 1, 2, 3\}$.

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     $\{00, 01, 02, 03, 10, ..., 13, 20, ..., 23, 30, ..., 33\}$. 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 is given by

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

206   where z_i is the value of the ith population unit in the sample, π_i is the inclusion

207   probability of the ith 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.

210   It is also important to quantify the uncertainty in $\hat{\tau}_{ht}$. Horvitz and Thompson

211   (1952) and Sen (1953) provide variance estimators for $\hat{\tau}_{ht}$, but these estimators

212   have two drawbacks. First, they rely on calculating π_{ij}, the probability that

213   population unit i and population unit j are both in the sample – this quantity

214   can be challenging if not impossible to calculate analytically for GRTS samples.

215   Second, these estimators tend to ignore the spatial locations of the population

216   units. To address these two drawbacks simultaneously, Stevens and Olsen (2003)

217   proposed the local neighborhood variance estimator. The local neighborhood

218   variance estimator does not rely on π_{ij} and estimates the variance of $\hat{\tau}$ conditional

219   on the random properties of the GRTS sample – the idea being that this

220   conditioning should yield a more precise estimate of $\hat{\tau}$. They show that the

221   contribution from each sample unit (population unit in the sample) to the overall

222   variance is dominated by local variation. Thus the local neighborhood variance

223   estimator is a weighted sum of variance estimates from each sample unit's local

224   neighborhood. These local neighborhoods contain the sample unit itself and

225   its three nearest neighbors among all other sample units. For more details, see

226   Stevens and Olsen (2003).

227   *1.1.3. Finite Population Block Kriging*

228   Finite population block kriging (FPBK) is a model-based approach that

229   expands the geostatistical Kriging framework to the finite population setting

230   (Ver Hoef, 2008). Instead of developing inference based on a specific sampling

231   design, we assume the data are generated by a spatial stochastic process. We

232   summarize some of the basic principles of FPBK next – for more details, see

233   Ver Hoef (2008). Let $\mathbf{z} \equiv \{z(s_1), z(s_2), ..., z(s_N)\}$ be an $N \times 1$ response vector

234   at locations s_1, s_2, . . . , s_N that can be measured at the N population units.

235   Suppose we want to use a sample to predict some linear function of the response

236   variable, $f(\mathbf{z}) = \mathbf{b}'\mathbf{z}$, where $\mathbf{b}'$ is a $1 \times N$ vector of weights (e.g, the population

237 mean is represented by a weights vector whose elements all equal $1/N$). Denoting

238 quantities that are part of the sampled population units with a subscript s and

239 quantities that are part of the unsampled population units with a subscript u,

240 let

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

241 where $\mathbf{X}_s$ and $\mathbf{X}_u$ are the design matrices for the sampled and unsampled

242 population units, respectively, $\boldsymbol{\beta}$ is the parameter vector of fixed effects, and

243 $\boldsymbol{\delta} \equiv [\boldsymbol{\delta}_s \ \boldsymbol{\delta}_u]'$, where $\boldsymbol{\delta}_s$ and $\boldsymbol{\delta}_u$ are random errors for the sampled and unsampled

244 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, jth element of the exponential covariance matrix, $\mathrm{cov}(\boldsymbol{\delta})$, is

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

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

variability, σ_2^2 is the variance parameter the 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 $\boldsymbol{\beta}$ 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

264 spatial data. Compared to the k-nearest-neighbors and random forest approach,

265 we prefer FPBK because it is model-based and relies on theoretically-based

266 variance estimators leveraging the model's spatial covariance structure, whereas

267 k-nearest-neighbors and random forests use ad-hoc variance estimators (Ver Hoef

268 and Temesgen, 2013). Additionally, Ver Hoef and Temesgen (2013) compared

269 FPBK, k-nearest-neighbors, and random forest in a variety of spatial data

270 contexts, and FPBK tended to perform best.

271 **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

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

272 where z_i is the ith response value, $\bar{z}$ is the mean of all z_i, n is the sample size, N

273 is the population size, and $f = (1 - n/N)$ (f is often called the finite population

274 correction factor). For GRTS-DB, the Horvitz-Thompson esetimator was used

275 to estimate means and the local neighborhood variance was used to estimate

276 variances. For SRS-MB and GRTS-MB, FPBK was used to estimate means and

277 variances and parameters were estimated using restricted maximum likelihood.

278 We used simulated data to compare the sampling-inference combinations

279 across many realized populations from the same data-generating stochastic

280 process. With the simulated data, we were in control of the data-generating

281 stochastic process and the random sampling process. We used real data from

282 the 2012 National Lakes Assessment (USEPA, 2012) to compare the sampling-

283 inference combinations within a single realized population (which is typically

284 the case in reality). With the real data, we were in control of only the random

285 sampling process.

2.1. Simulated Data

287 CHANGE LOGNORMAL VERBAGE TO SKEWED – look for DRE acronym

288 We evaluated performance of the four sampling-inference combinations in

289 36 different simulation scenarios. The 36 scenarios resulted from the crossing of

290 three sample sizes, two location layouts (of the population units), two response

291 types, and three proportions of dependent random error (DRE). The three sample

292 sizes (n) were $n = 50, n = 100$, and $n = 200$. Samples were always selected from

293 a population size (N) of $N = 900$. The two location layouts were random and

294 gridded. Locations in the random layout were randomly generated inside the

295 unit square $([0, 1] \times [0, 1])$. Locations in the gridded layout were placed on a fixed,

296 equally spaced grid inside the unit square. The two response types were normal

297 and skewed. For the normal response type, the response was simulated using

298 mean-zero random errors with the exponential covariance (Equation 3) for three

299 proportions of dependent random error (DRE): 0% DRE, 50% DRE, and 90%

300 DRE. Recall the proportion of DRE is represented by $\sigma_1^2/(\sigma_1^2 + \sigma_2^2)$, where σ_1^2

301 and σ_2^2 are the DRE variance and independent random error (IRE) variance from

302 Equation 3, respectively. The total variance, $\sigma_1^2 + \sigma_2^2$, was always 2. The distance
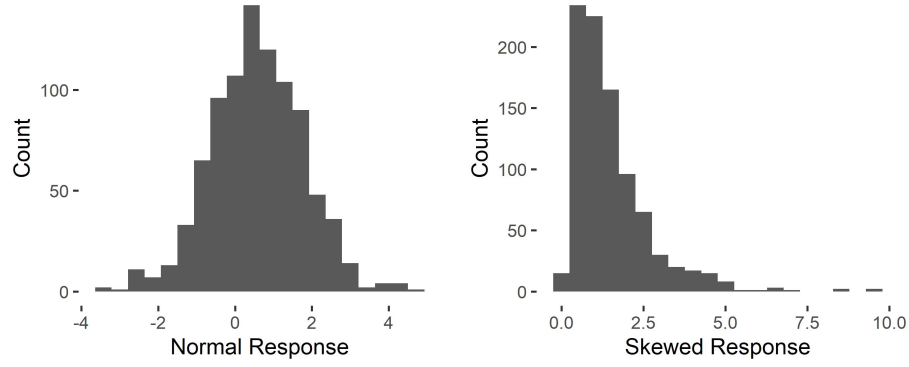
(a) Histogram of a realized population for the normal response.

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

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

parameter was always $\sqrt{2}/3$, chosen so that the correlation in the DRE decayed 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

320 the sampling-inference combination in each scenario by calculating mean bias,

321 root-mean-squared error, and interval coverage. Mean bias is taken as the average

322 deviation between each trial's estimated (or predicted) mean and its realized

323 mean: $\frac{1}{n}\sum_{i=1}^{2000}(\hat{\mu}_i - \mu_i)$, where i indexes simulation trials. Root-mean-squared

324 error is taken as the square root of the average squared deviation between each

325 trial's estimated (or predicted) mean and its realized mean: $\sqrt{\frac{1}{n}\sum_{i=1}^{2000}(\hat{\mu}_i - \mu_i)^2}$.

326 Interval coverage is taken as the proportion of simulation trials where the

327 realized mean is contained in its 95% confidence (or prediction) interval. These

328 intervals are constructed using the normal distribution – justification comes

329 from the asymptotic normality of means via the central limit theorem (under

330 some assumptions). Quantifying these metrics is important because together,

331 they give us an idea of the accuracy (mean bias), spread (RMSE), and validity

332 (interval coverage) of the sampling-inference combinations.

333 *2.2. National Lakes Assessment (Real) Data*

334   The United States Environmental Protection Agency (USEPA), states, and

335 tribes periodically conduct National Aquatic Research Surveys (NARS) to assess

336 the water quality of various bodies of water in the contiguous United States.

337 One component of NARS is the National Lakes Assessment (NLA), which

338 measures various aspects of lake health and water quality. We focus on analyzing

339 zooplankton multi-metric indices (ZMMI) and mercury concentrations in parts

340 per billion (Hg ppb) from the 2012 NLA. For ZMMI, data were collected at 1035

341 unique lakes. At less than 10% of lakes, two ZMMI replicates were collected.

342 These were averaged for the purposes of our study so that each lake had one

343 measurement for ZMMI. For Hg ppb, data were collected at 995 unique lakes

344 (and there were no replicates like for ZMMI). The ZMMI and Hg ppb data are

345 shown spatially and as a histogram in Figure 3. The ZMMI data are relatively

346 symmetric, and the Hg ppb data are skewed. Also in Figure 3 are separate spatial

347   semivariograms for ZMMI and Hg ppb. The spatial semivariogram quantifies

348   the average squared differences (semivariance) of responses whose separation

349   (distance) falls within some distance class. The spatial semivariance is closely

350   related to the spatial covariance, and spatial semivariograms are often used to

351   gauge the strength of spatial dependence in data. Both ZMMI and Hg ppb seem

352   to have moderately strong spatial dependence (Figure 3), as the semivariance

353   increases steadily with distance (meaning that observations nearby one another

354   tend to be more similar than observations far apart from one another).

355   We studied performance of the four sampling-inference combinations by

356   selecting 2000 random IRS and GRTS samples of size $n = 50$, $n = 100$, and

357    $n = 200$ from the realized ZMMI and Hg ppb populations and then analyzing the

358   samples using MB and DB inference. We used the same evaluation metrics as for

359   the simulated data: mean bias, RMSE, and interval coverage. Mean bias is taken

360   as the average deviation between each sample's estimated (or predicted) mean

361   and the population mean (of ZMMI or Hg ppb): $\frac{1}{n} \sum_{i=1}^{2000} (\hat{\mu}_i - \mu)$, where i indexes

362   simulation trials and μ is the population mean. Root-mean-squared error is

363   taken as the square root of the average squared deviation between each sample's

364   estimated (or predicted) mean and its population mean: $\sqrt{\frac{1}{n} \sum_{i=1}^{2000} (\hat{\mu}_i - \mu)^2}$.

365   Interval coverage is taken as the proportion of simulation trials where the

366   population mean is contained in its 95% confidence (or prediction) interval.

367   These intervals are constructed using the normal distribution.

368   **3. Results**

369   *3.1. Simulated Data*

370   The mean bias was nearly zero for all four sampling-inference combinations

371   in all 36 scenarios, so we omit a more detailed summary of those results here.

372   Tables for mean bias in all 36 simulation scenarios are provided in the supporting

(a) Spatial map of the ZMMI population.

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

(c) Histogram of the ZMMI population.

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

(e) Semivariogram of the ZMMI population.
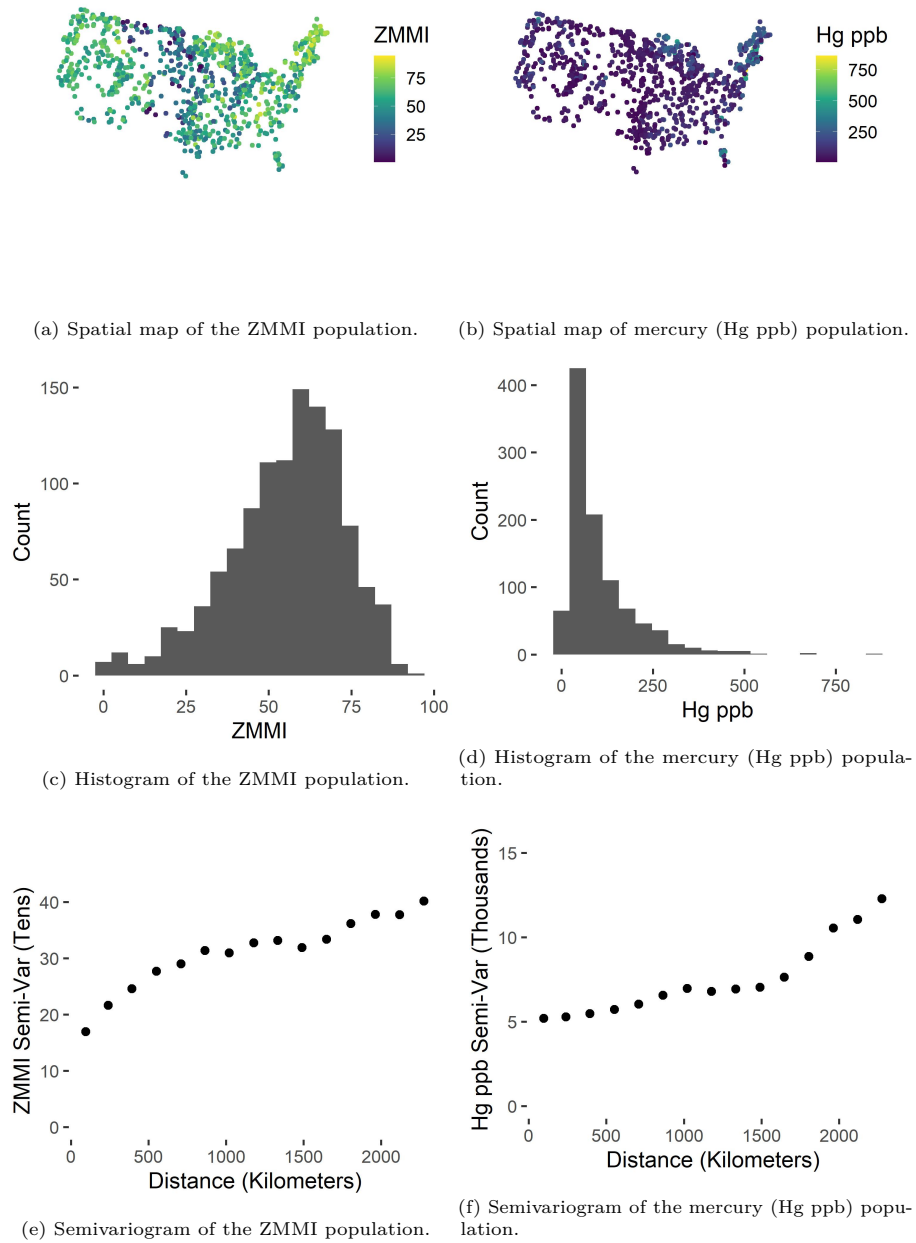
(f) Semivariogram of the mercury (Hg ppb) population.

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

information.

Fig. 4 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

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

When there is no spatial covariance (Fig. 4, "Prop DE: 0" row), the four sampling-inference combinations have approximately equal rMS(P)E. In these scenarios, using the GRTS algorithm or a model-based analysis does not increase efficiency compared to IRS-Design. When there is spatial covariance (Fig. 4, "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-inference 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-inference combinations using the random location layout with "IRS-Design" as the baseline. The relative MStdE is defined as

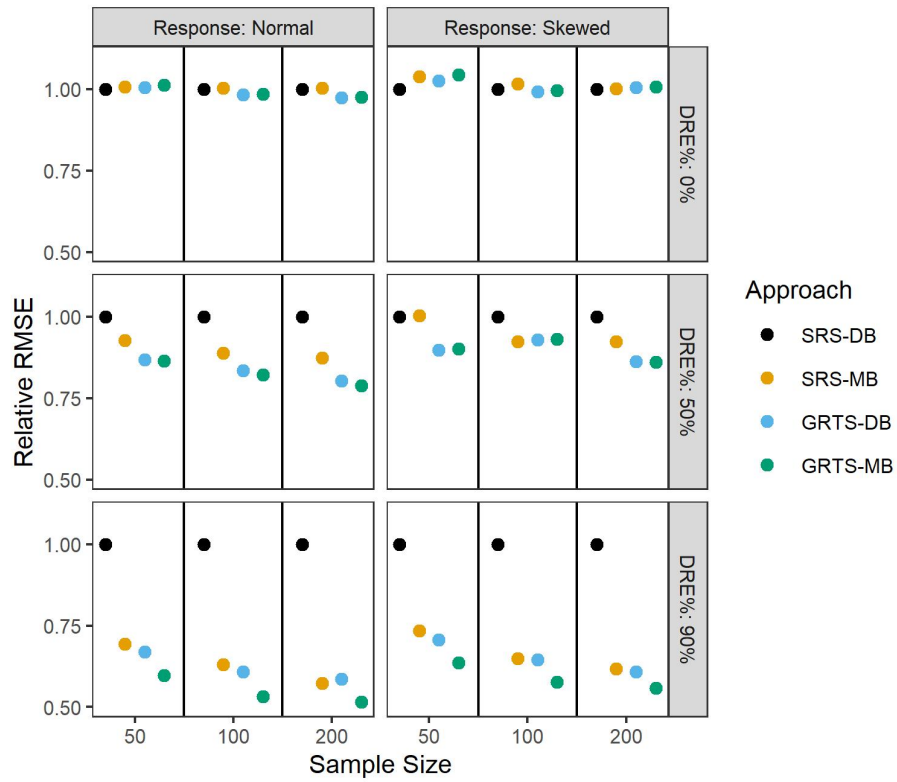$$\frac{\text{MStdE of sampling-inference combination}}{\text{MStdE of IRS-Design}},$$

Figure 4: Relative rMS(P)E in the simulation study for the four sampling-inference combinations. The rows indicate the proportion of dependent error and the columns indicate the response type.

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-inference 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-inference 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: 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. 5 shows the 95% interval coverage for each of the four sampling-inference combinations in the random location layout. Within each scenario, the sampling-inference 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-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 grid location layout figure here. Tables for interval coverage
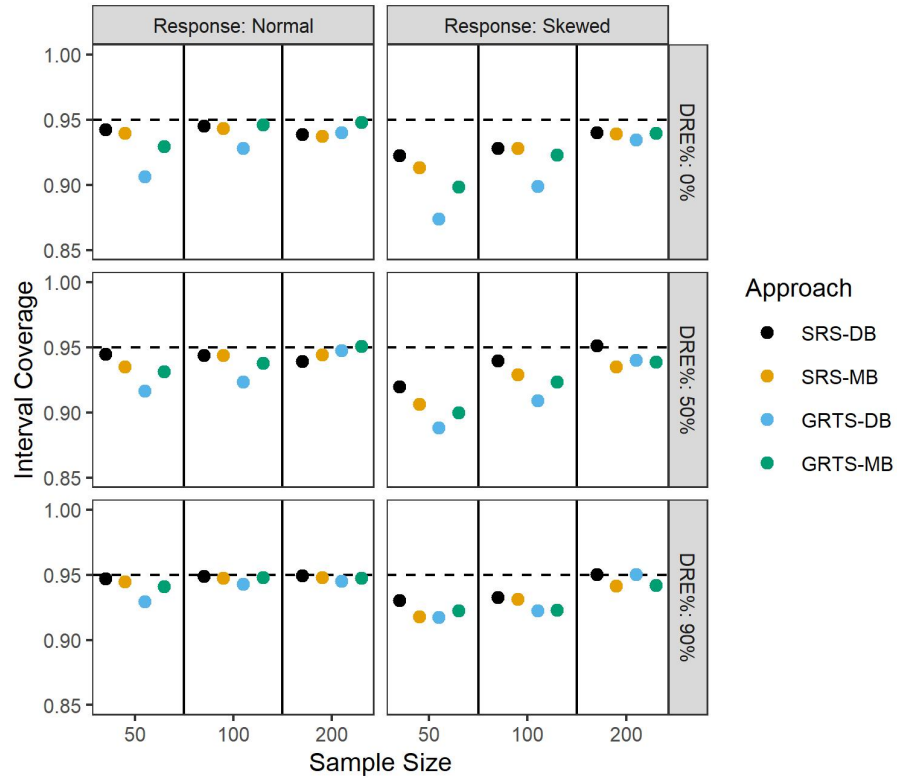
Figure 5: Interval coverage in the simulation study for the four sampling-inference combinations. The rows indicate the proportion of dependent error and the columns indicate the response type. The solid, black line represents 95% coverage.

416 in all 36 simulation scenarios are provided in the supporting information.

417 *3.2. National Lakes Assessment DAta*

418 USE MERCURY UNITS

419 Fig. **??** shows a map and histogram of mercury concentration in all 986 NLA

420 lakes. The map shows mercury concentration exhibits some spatial patterning,

421 with high mercury concentrations in the northeast and north central United

422 States. The histogram shows that mercury concentration is right-skewed, with

423 most lakes having a low value of mercury concentration but a few having a much

424 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 (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 1 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-inference combinations, the true mean mercury concentration (103.2 ng / g) is within the bounds of the combination's 95% interval.

3.3. New Application

4. Discussion

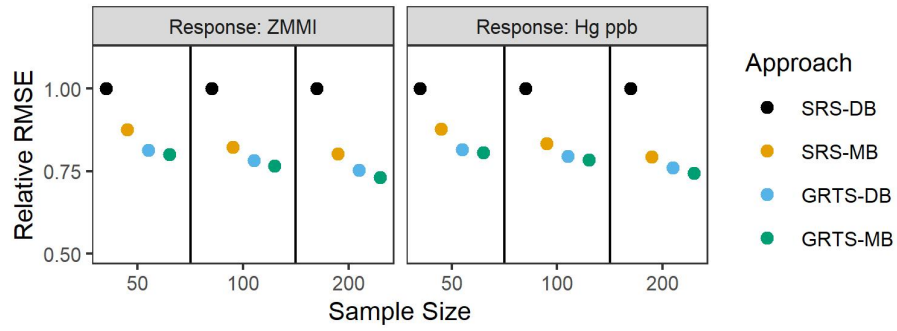ADD EXTRAS LIKE ANISOTROPY AND UNEQUAL INCLUSION PROBABILITIES

Figure 6: Relative rMS(P)E 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.
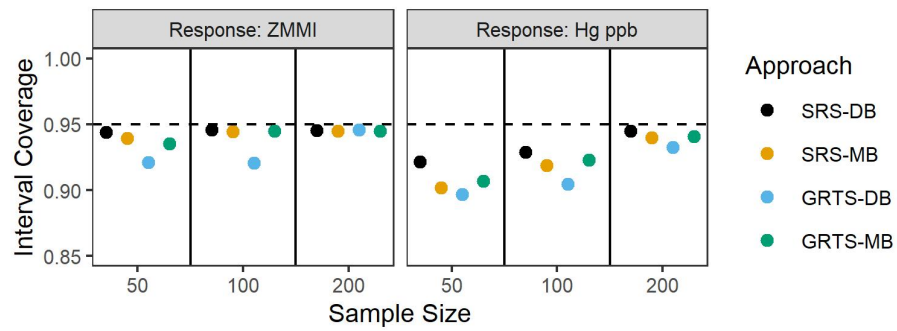

Figure 7: Interval coverage in the data study for the four sampling-inference combinations. The rows indicate the proportion of dependent error and the columns indicate the response type. The solid, black line represents 95% coverage.

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 1: For each sampling-inference 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.

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

472 sampling-inference combinations also increases, likely because IRS-Design is the

473 only combination that ignores spatial locations in sampling and analysis. Fourth

474 and finally, when the response was normal, interval coverage for all sampling-

475 inference combinations was usually close to 95% for all sample sizes; when the

476 response was lognormal, interval coverage for all sampling-inference combinations

477 was usually between 90% and 95% and closest to 95% when $n = 200$.

478   AT LEAST HAVE DISCUSSION ABOUT MODEL BASED ASSUMPTIONS

479 AND MOVE VALIDITY COMMENTS TO RESULTS SECTION.

480   There are several benefits and drawbacks of the design-based and model-

481 based approaches for finite population spatial data. Some we have discussed,

482 but others we have not, and they are worthy of consideration in future research.

483 Design-based approaches are often computationally efficient, while model-based

484 approaches can be computationally burdensome, especially for likelihood-based

485 estimation methods like REML that rely on inverting a covariance matrix. The

486 design-based approach easily handles binary data through a straightforward

487 application of the Horvitz-Thompson estimator. In contrast, analyzing binary

488 data using a model-based approach generally requires a logistic mixed regression

489 model, which can be challenging to estimate and interpret (Bolker et al., 2009).

490 The design-based approach yields valid results because the sampling plan and

491 inclusion probabilities are specified directly by the researcher, while the model-

492 based approach may not yield valid results if the assumptions made do not

493 not accurately capture reality. The model-based approach, however, can more

494 naturally quantify the relationship between covariates (predictor variables) and

495 the response variable. The model-based approach also yields estimated spatial

496 covariance parameters, which help better understand the dependence structure

497 in the process in study. Model selection is also possible using model-based

498 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

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.

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.

624 Ver Hoef, J., 2002. Sampling and geostatistics for spatial data. Ecoscience 9,
625 152–161.

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

628 Ver Hoef, J.M., Temesgen, H., 2013. A comparison of the spatial linear
629 model to nearest neighbor (k-nn) methods for forestry applications. PlOS ONE
630 8, e59129.

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

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

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