

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

3 Michael Dumelle^{*,a}, Matt Higham^b, Jay M. Ver Hoef^c, Anthony R. Olsen^a, Lisa
4 Madsen^d

5 ^a*United States Environmental Protection Agency, 200 SW 35th St, Corvallis, Oregon, 97333*

6 ^b*Saint Lawrence University Department of Mathematics, Computer Science, and Statistics,
7 23 Romoda Drive, Canton, New York, 13617*

8 ^c*Marine Mammal Laboratory, Alaska Fisheries Science Center, National Oceanic and
9 Atmospheric Administration, Seattle, Washington, 98115*

10 ^d*Oregon State University Department of Statistics, 239 Weniger Hall, Corvallis, Oregon,
11 97331*

12 **Abstract**

- 13 1. The design-based and model-based approaches to frequentist statistical
14 inference lie on fundamentally different foundations. In the design-based
15 approach, inference depends on random sampling. In the model-based
16 approach, inference depends on distributional assumptions. We compare
17 the approaches for finite population spatial data.
- 18 2. We provide relevant background for the design-based and model-based
19 approaches and then study their performance using simulations and an
20 analysis of real mercury concentration data. In the simulations, a variety of
21 sample sizes, location layouts, dependence structures, and response types
22 are considered. In the simulations and real data analysis, the population
23 mean is the parameter of interest and performance is measured using
24 statistics like bias, squared error, and interval coverage.
- 25 3. When studying the simulations and mercury concentration data, we found
26 that regardless of the strength of spatial dependence in the data, sampling
27 plans that incorporate spatial locations (spatially balanced samples) gener-
28 ally outperform sampling plans that ignore spatial locations (non-spatially
29 balanced samples). We also found that model-based approaches tend to

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

Preprint submitted to *Methods in Ecology and Evolution*

December 21, 2021

outperform design-based approaches, even when the data are skewed (and by consequence, the model-based distributional assumptions violated). The performance gap between these approaches is small when spatially balanced samples are used but large when non-spatially balanced samples are used. This suggests that the sampling choice (whether to select a sample that is spatially balanced) 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; Model-based inference; Spatially balanced sampling; Spatial covariance

1. Introduction

There are two general approaches for using data to make frequentist statistical inferences about a population: design-based and model-based. 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 of population units is called a sample. In the design-based approach, inferences about the underlying population are informed via a probabilistic process that randomly assigns some population units to be in the sample. Alternatively, in the model-based approach, inferences

are made from specific assumptions about the underlying process generating the data. Each paradigm has a deep historical context (Sterba, 2009) and its own set of benefits and drawbacks (Hansen et al., 1983).

Though the design-based and model-based approaches apply to statistical inference in a broad sense, we focus on comparing these approaches for spatial data. We define spatial data as data that incorporates the specific locations of the population units into either the sampling or estimation process. De Gruijter and Ter Braak (1990) give an early comparison 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; Wang et al., 2012). 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 and model-based principles (see e.g., Sterba (2009) and Cicchitelli and Montanari (2012), and see Chan-Golston et al. (2020) for a Bayesian approach).

Certainly comparisons between design-based and model-based approaches to spatial data have been studied. But no numerical comparison has been made between design-based approaches that incorporate spatial information and model-based approaches. In this manuscript, we compare design-based approaches that incorporate spatial information to model-based approaches for finite population spatial data. A finite population contains a finite number of population units (we assume the finite number is known); an example is lakes (treated as a whole with the lake centroid representing location) in the contiguous United States. Though we focus on finite populations, these comparisons generalize to

82 infinite populations as well. An infinite population contains an infinite number
83 of population units; an example is locations within a single lake.

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

92 *1.1. Background*

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

97 *1.1.1. Comparing Design-Based and Model-Based Approaches*

98 The design-based approach assumes the population is fixed. Randomness
99 is incorporated via the selection of population units according to a sampling
100 design. A sampling design assigns a positive probability of inclusion (inclusion
101 probability) in the sample to each population unit. These inclusion probabilities
102 are later used to analyze data. Some examples of commonly used sampling
103 designs include simple random sampling, stratified random sampling, and cluster
104 sampling.

105 When sampling designs incorporate spatial locations into sampling, we call
106 the resulting samples “spatially balanced.” One approach to selecting spatially
107 balanced samples is the Generalized Random Tessellation Stratified (GRTS)

algorithm (Stevens and Olsen, 2004), which we discuss in more detail in Section 1.1.2. When sampling designs do not incorporate spatial locations into sampling, we call the resulting samples “non-spatially balanced.”

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

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

Instead of estimating fixed, unknown population parameters, as in the design-

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

147 1.1.2. *Spatially Balanced Design and Analysis*

148 We previously mentioned that the design-based approach can be used to
 149 select spatially balanced samples (samples that incorporate spatial locations of
 150 the population units). Spatially balanced samples are useful because parameter
 151 estimates from these samples tend to vary less than parameter estimates from
 152 samples that are not spatially balanced (Barabesi and Franceschi, 2011; Benedetti
 153 et al., 2017; Grafström and Lundström, 2013; Robertson et al., 2013; Stevens and
 154 Olsen, 2004; Wang et al., 2013). The first spatially balanced sampling algorithm
 155 to see widespread use was the Generalized Random Tessellation Stratified (GRTS)
 156 algorithm (Stevens and Olsen, 2004). To quantify the spatial balance of a
 157 sample, Stevens and Olsen (2004) proposed loss metrics based on Voronoi
 158 polygons (Dirichlet Tessellations). After the GRTS algorithm was developed,
 159 several other spatially balanced sampling algorithms emerged, including the
 160 Local Pivotal Method (Grafström et al., 2012; Grafström and Matei, 2018),
 161 Spatially Correlated Poisson Sampling (Grafström, 2012), Balanced Acceptance

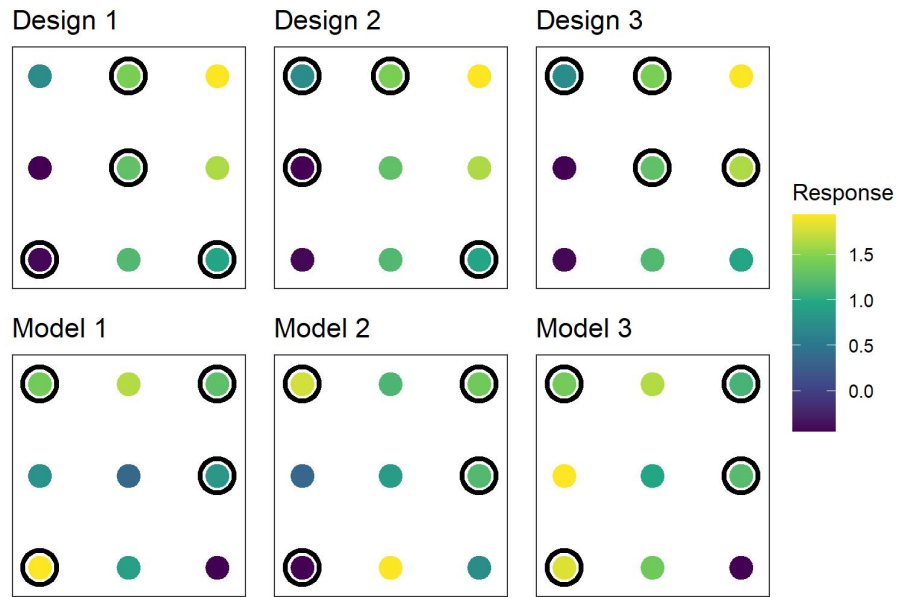


Figure 1: A visual comparison of the design-based and model-based approaches. In the top row, the design-based approach is highlighted. There is one fixed population with nine population units and three random samples of size four (points circled are those sampled). The response values at each site are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

162 Sampling (Robertson et al., 2013), Within-Sample-Distance Sampling (Benedetti
 163 and Piersimoni, 2017), and Halton Iterative Partitioning Sampling (Robertson
 164 et al., 2018). In this manuscript, we select spatially balanced samples using
 165 the Generalized Random Tessellation Stratified (GRTS) algorithm because it
 166 has several attractive properties: the GRTS algorithm accommodates finite and
 167 infinite sampling frames, equal, unequal, and proportional (to size) inclusion
 168 probabilities, legacy (historical) sampling (Foster et al., 2017), a minimum
 169 distance between units in a sample, and replacement units (replacement units are
 170 population units that can be sampled when a population unit originally selected
 171 can no longer be sampled). The GRTS algorithm selects samples by utilizing a
 172 particular mapping between two-dimensional and one-dimensional space that
 173 preserves proximity relationships. Via this mapping, units in two-dimensional
 174 space are partitioned using a hierarchical address. This hierarchical address is
 175 used to map population units to a one-dimensional line. On the one dimensional
 176 line, each population unit's line length equals its inclusion probability. Then, a
 177 systematic sample of population units is selected on the line and mapped back
 178 to two-dimensional space, yielding the desired sample. Stevens and Olsen (2004)
 179 provide more technical details.

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

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

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

183 of population units.

184 It is also important to quantify the uncertainty in $\hat{\tau}_{ht}$. Horvitz and Thompson
185 (1952) and Sen (1953) provide variance estimators for $\hat{\tau}_{ht}$, but these estimators
186 have two drawbacks. First, they rely on calculating π_{ij} , the probability that
187 population unit i and population unit j are both in the sample – this quantity
188 can be challenging if not impossible to calculate analytically. Second, these
189 estimators ignore the spatial locations of the population units. To address these
190 two drawbacks simultaneously, Stevens and Olsen (2003) proposed the local
191 neighborhood variance estimator. The local neighborhood variance estimator
192 does not rely on π_{ij} and incorporates spatial locations – for technical details see
193 Stevens and Olsen (2003). Stevens and Olsen (2003) show the local neighborhood
194 variance estimator tends to reduce the estimated variance of $\hat{\tau}$ and yield narrower
195 confidence intervals compared to variance estimators that ignore spatial locations.

196 1.1.3. Finite Population Block Kriging

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

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

210 where \mathbf{X}_s and \mathbf{X}_u are the design matrices for the sampled and unsampled
 211 population units, respectively, $\boldsymbol{\beta}$ is the parameter vector of fixed effects, and
 212 $\boldsymbol{\delta} \equiv [\boldsymbol{\delta}_s \ \boldsymbol{\delta}_u]'$, where $\boldsymbol{\delta}_s$ and $\boldsymbol{\delta}_u$ are random errors for the sampled and unsampled
 213 population units, respectively.

FBPK 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 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

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

214 where σ_1^2 is the variance parameter quantifying the variability that is dependent
 215 (coarse-scale), σ_2^2 is the variance parameter quantifying the variability that is
 216 independent (fine-scale), ϕ is the range parameter measuring the distance-decay
 217 rate of the covariance, and $h_{i,j}$ is the Euclidean distance between population
 218 units i and j . The proportion of variability attributable to dependent random

error is $\sigma_1^2/(\sigma_1^2 + \sigma_2^2)$. Similarly, the proportion of variability attributable to independent random error is $\sigma_2^2/(\sigma_1^2 + \sigma_2^2)$. Finally we note that σ_1^2 and σ_2^2 are often called the partial sill and nugget, respectively.

With the above model formulation, 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.

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 FBPK 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) studied compared FBPK, k-nearest-neighbors, and random forest in a variety of spatial data contexts, and FBPK tended to perform best.

2. Materials and Methods

2.1. Simulation Study

We used a simulation study to investigate performance of four sampling-analysis combinations: Independent Random Sampling (IRS) with a design-based analysis, called “IRS-Design” (IRS samples are not spatially balanced because population units are selected independent of their spatial locations); IRS with a model-based analysis, called “IRS-Model”; GRTS sampling with a design-based analysis, called “GRTS-Design”; and GRTS sampling with a model-based

analysis, called “GRTS-Model”. These combinations are also provided in Table 1.

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

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

Performance for the four sampling-analysis combinations was evaluated in 36 different simulation scenarios. The 36 scenarios resulted from the crossing of three sample sizes, two location layouts (of the population units), two response types, and three proportions of dependent random error. The three sample sizes (n) were $n = 50, n = 100$, and $n = 200$. Samples were always selected from a population size (N) of $N = 900$. The two location layouts were random and gridded. Locations in the random layout were randomly generated inside the unit square $([0, 1] \times [0, 1])$. Locations in the gridded layout were placed on a fixed, equally spaced grid inside the unit square. The two response types were normal and lognormal. For the normal response type, the response was simulated using mean-zero random errors with the exponential covariance (Equation 3) for varying proportions of dependent random error. The proportion of dependent random error is represented by $\sigma_1^2 / (\sigma_1^2 + \sigma_2^2)$, where σ_1^2 and σ_2^2 are the dependent random error variance (partial sill) and independent random error variance (nugget), respectively, from Equation 3. The total variance, $\sigma_1^2 + \sigma_2^2$, was always 2. The range was always $\sqrt{2}/3$, which means that the correlation in the dependent random error decayed to nearly zero at the largest possible distance between two population units in the domain. For the lognormal response type, the response was first simulated using the same approach as for the normal response type, except that the total variance was 0.6931 instead of 2. The response was then exponentiated, yielding a lognormal random variable whose

total variance is 2. The lognormal responses were used to evaluate performance of the sampling-analysis approaches for data that were skewed (i.e., not normal).

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

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

In each of the 36 simulation scenarios, there were 2000 independent simulation trials. In each trial, IRS and GRTS samples were selected and then design-based and model-based analyses were used to estimate (design-based) or predict (model-based) the mean and construct 95% confidence (design-based) or 95% prediction (model-based) intervals. Then we recorded the bias, squared error, and interval coverage for all sampling-analysis combinations. After all 2000 trials, we summarized the long-run performance of the combinations by calculating average bias, rMS(P)E (root-mean-squared error for the design-based approaches and root-mean-squared-prediction error for the model-based approaches), and the proportion of times the true mean is contained in its 95% interval. The GRTS algorithm and the local neighborhood variance estimator are available in the **spsurvey R** package (Dumelle et al., 2021). FPKB is available in the **sptotal R** package (Higham et al., 2021) and covariance parameters were estimated using Restricted Maximum Likelihood (Harville, 1977; Patterson and Thompson, 1971; Wolfinger et al., 1994).

2.2. Application

The United States Environmental Protection Agency (USEPA), states, and tribes periodically conduct National Aquatic Research Surveys (NARS) to assess the water quality of various bodies of water in the contiguous United States. We will use data from the 2012 National Lakes Assessment (NLA), which measures

290 various aspects of lake health and water quality (USEPA, 2012). Specifically,
 291 we will analyze mercury concentration in lakes. Although we can calculate the
 292 true mean mercury concentration values for the 986 lakes from the 2012 NLA,
 293 we will explore whether or not we obtain an adequately precise estimate for the
 294 realized mean mercury concentration if we sample only 100 of the 986 lakes.
 295 For each of the four familiar sampling-analysis combinations (IRS-Design, IRS-
 296 Model, GRTS-Design, and GRTS-Model), we estimate (design-based) or predict
 297 (model-based) the mean mercury concentration and construct 95% confidence
 298 (design-based) or 95% prediction (model-based) intervals from this sample of
 299 100 lakes, which we compare to the true mean mercury concentration from all
 300 986 lakes.

301 **3. Results**

302 *3.1. Simulation Study*

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

Fig. 2 shows the relative rMS(P)E of the four sampling analysis combinations
 using the random location layout with “IRS-Design” as the baseline. The relative
 rMS(P)E is defined as

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

307 When there is no spatial covariance (Fig. 2, “Prop DE: 0” row), the four
 308 sampling-analysis combinations have approximately equal rMS(P)E. So using
 309 the GRTS sampling plan or a model-based analysis does not result in much, if
 310 any, loss in efficiency compared to IRS-Design when there is no spatial covariance.

311 When there is spatial covariance (Fig. 2, “Prop DE: 0.5” and “Prop DE: 0.9”
 312 rows), GRTS-Model tends to perform best, followed by GRTS-Design, IRS-
 313 Model, and finally IRS-Design, though the difference in relative rMS(P)E among
 314 GRTS-Model, GRTS-Design, and IRS-Model is relatively small. As the strength
 315 of spatial covariance increases, the gap in rMS(P)E between IRS-Design and the
 316 other sampling-analysis combinations widens. Finally we note that when there
 317 is spatial covariance, IRS-Model outperforms IRS-Design by a large margin,
 318 suggesting that the poor design properties of IRS are largely mitigated by the
 319 model-based analysis. These conclusions are similar to those observed in the grid
 320 location layout, so we omit a grid location layout figure here. Tables for rMS(P)E
 321 in all 36 simulation scenarios are provided in the supporting information.

322 We also studied 95% interval coverage among the sampling-analysis com-
 323 binations. The design-based confidence intervals and model-based prediction
 324 intervals were constructed using the normal distribution. Justification for this
 325 comes from the asymptotic normality of means via the Central Limit Theorem.
 326 Fig. 3 shows the 95% interval coverage for each of the four sampling-analysis
 327 combinations in the random location layout. Within each scenario, the sampling-
 328 analysis combinations tend to have fairly similar interval coverage. Coverage
 329 in the normal response scenarios was usually near 95%, while coverage in the
 330 lognormal response scenarios varied from from 90% to 95% but increased with
 331 the sample size. At a sample size of 200, all four sampling-analysis combinations
 332 had approximately 95% interval coverage in both response scenarios for all
 333 dependent error proportions. These conclusions are similar to those observed in
 334 the grid location layout, so we omit a grid location layout figure here. Tables for
 335 interval coverage in all 36 simulation scenarios are provided in the supporting
 336 information.

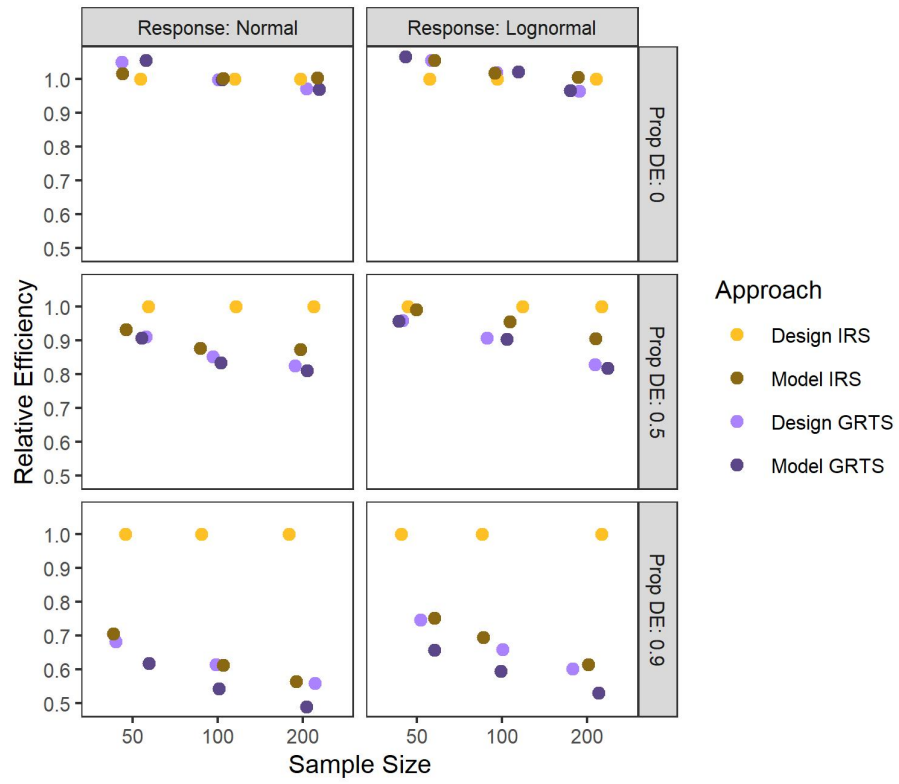


Figure 2: Relative RMS(P)E in the simulation study for the four sampling-analysis combinations. The rows indicate the proportion of dependent error and the columns indicate the response type.

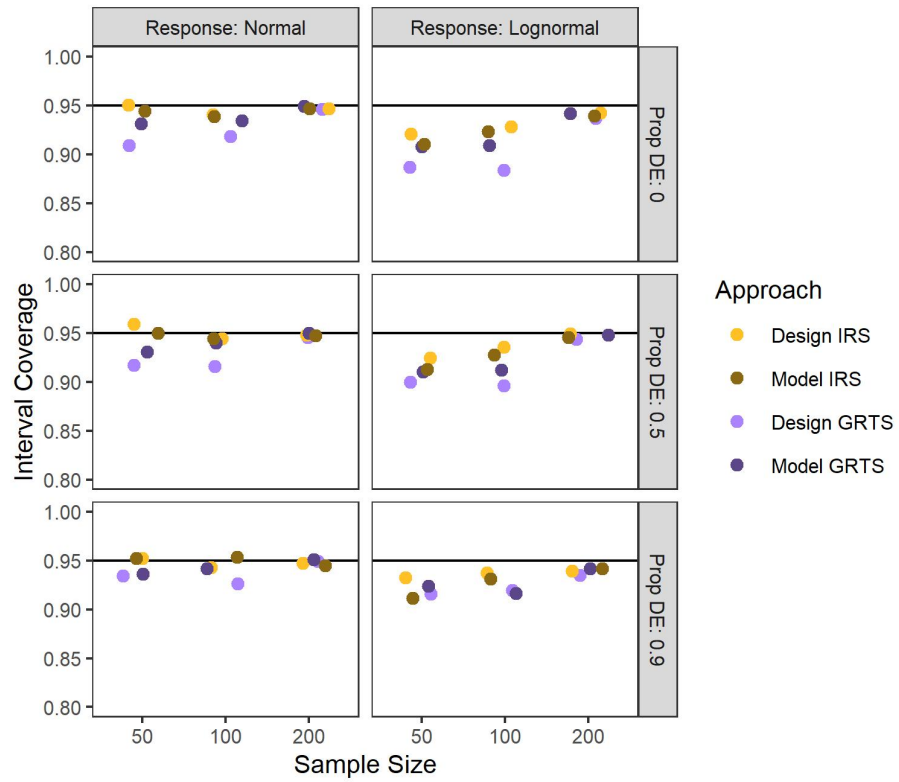


Figure 3: Interval coverage in the simulation study for the four sampling-analysis combinations. The rows indicate the proportion of dependent error and the columns indicate the response type. The solid, black line represents 95% coverage.

337 3.2. Application

338 Fig. 4 shows a map and histogram of mercury concentration in all 986 NLA
 339 lakes. The map shows mercury concentration exhibits some spatial patterning,
 340 with high mercury concentrations in the northeast and north central United
 341 States. The histogram shows that mercury concentration is right-skewed, with
 342 most lakes having a low value of mercury concentration but a few having a
 343 much higher concentration. Fig. 4 also shows mercury concentration's empirical
 344 semivariogram. The empirical semivariogram can be used as a tool to visualize
 345 spatial dependence. It quantifies the halved squared differences (semivariance)
 346 among mercury concentration at different distances apart. When a process
 347 has spatial covariance (exhibits spatial dependence), the semivariance tends
 348 to be smaller at small distances and larger at large distances. The empirical
 349 semivariogram in Fig. 4 suggests that mercury concentration is exhibits spatial
 350 dependence. Lastly we note that the realized mean mercury concentration in
 351 the 986 NLA lakes is 103.2 ng / g.

352 We selected a single IRS sample and a single GRTS sample and estimated
 353 (design-based) or predicted (model-based) the mean mercury concentration and
 354 constructed 95% confidence (design-based) and 95% (model-based) prediction
 355 intervals. For the model-based analyses, the exponential covariance was used.
 356 Table 3 shows the results from these analyses. For all four sampling-analysis
 357 combinations, the true realized mean mercury concentration is within the bounds
 358 of the 95% confidence (design-based) or 95% prediction (model-based) intervals.
 359 Though we should not generalize these results to other samples from these data,
 360 we do note a couple of patterns. The design-based IRS analysis shows the
 361 largest standard error: a likely reason is that this is the only approach that
 362 does not incorporate any spatial information regarding mercury concentration.
 363 Both analyses using GRTS sampling have lower standard errors than both

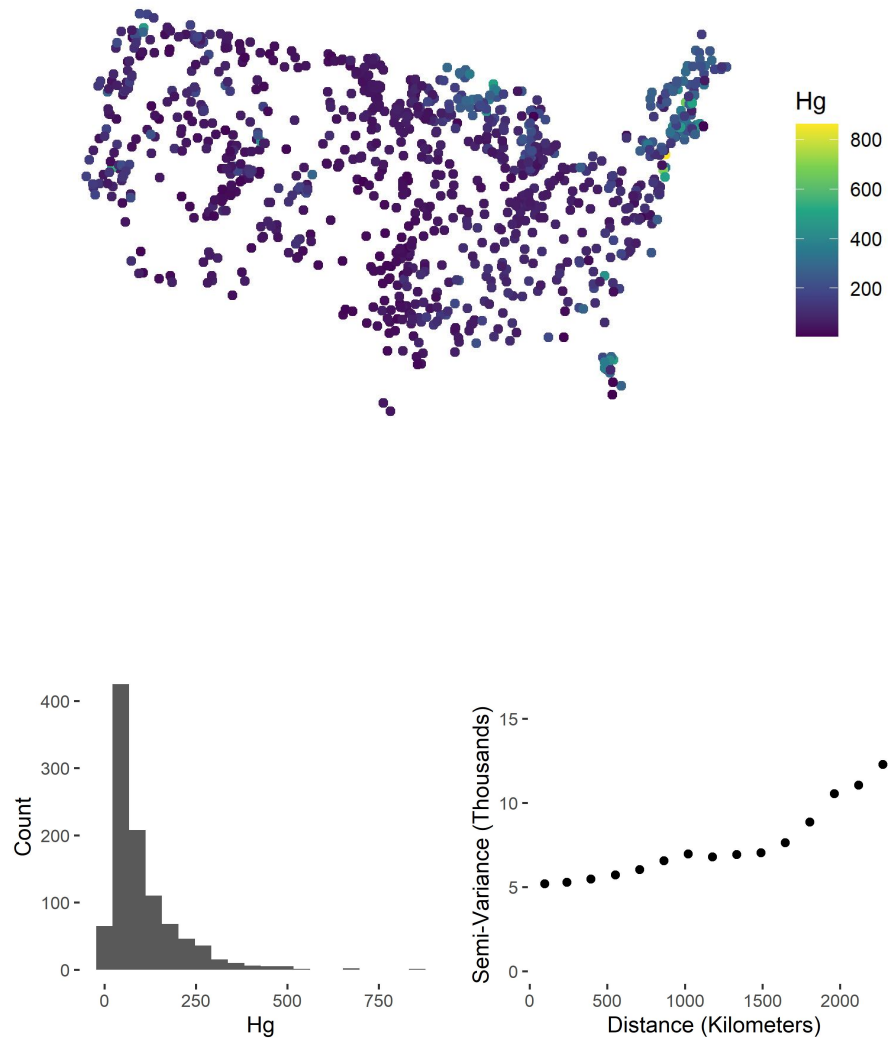


Figure 4: Mercury concentration visualizations for the population (Hg) for 986 lakes in the NLA data. A spatial layout is in the top row, a histogram is in the bottom row and left column, and an empirical semivariogram is in the bottom row and right column.

analyses using IRS sampling. We expect that these patterns are consistent with other samples from these data because mercury concentration exhibits spatial dependence, so a spatially balanced sample should usually yield a lower standard error.

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

Table 3: For each sampling-analysis combination (Approach), 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 true mean concentration of all 986 lakes in the NLA data is 103.2 ng / g.

4. Discussion

The design-based and model-based approaches to statistical inference are fundamentally different paradigms. The design-based approach incorporates randomness through sampling to estimate population parameters. The model-based approach incorporates randomness through distributional assumptions to predict realized values of a stochastic process. Though these 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 local neighborhood variance estimator (in the design-based approach). Aside from the theoretical differences described, a few analytical findings from the simulation study are particularly notable. First, the sampling decision (IRS vs GRTS) is most important when using a design-based analysis. Though GRTS-Model still outperformed IRS-Model, the model-based analysis mitigated much of the inefficiency of the IRS sample. Second, independent of the analysis approach, we found no reason to prefer IRS over GRTS for sampling spatial data – GRTS-Design and GRTS-Model generally

384 performed at least as well as their IRS counterparts when there was no spatial
 385 covariance and noticeably better than their IRS counterparts when there was
 386 spatial covariance. Third, as the strength of spatial covariance increases, the gap
 387 in rMS(P)E between IRS-Design and the other sampling-analysis combinations
 388 also increases. Fourth and finally, when the response was normal, interval
 389 coverage for all sampling-analysis combinations was very close to 95% for all
 390 sample sizes; when the response was lognormal, interval coverage for all sampling
 391 and analysis was between 90% and 95% and closest to 95% when $n = 200$.

392 There are several benefits and drawbacks of the design-based and model-
 393 based approaches for finite population spatial data. Some we have discussed,
 394 but others we have not, and they are worthy of consideration in future research.
 395 Design-based approaches are often computationally efficient, while model-based
 396 approaches can be computationally burdensome, especially for likelihood-based
 397 estimation methods like REML that rely on inverting a covariance matrix. The
 398 design-based approach also more naturally handles binary data, free from the
 399 more complicated logistic regression framework commonly used to analyze binary
 400 data in a model-based approach. The model-based approach, however, can more
 401 naturally quantify the relationship between covariates (predictor variables) and
 402 response variable. The model-based approach also yields estimated spatial
 403 covariance parameters, which help better understand the dependence structure
 404 in the stochastic process of study. Model selection is also possible using model-
 405 based approaches and criteria such as cross validation, likelihood ratio tests,
 406 or AIC (Akaike, 1974). Model-based approaches are capable of more efficient
 407 small-area estimation than design-based approaches by leveraging distributional
 408 assumptions in areas with few observed sites. Model-based approaches can
 409 also compute site-by-site predictions at unobserved locations and use them
 410 to construct informative visualizations. In short, when deciding whether the

411 design-based or model-based approach is more appropriate to implement, the
412 benefits and drawbacks of each approach should be considered alongside the
413 particular goals of the study.

414 **Acknowledgments**

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

423 **Conflict of Interest Statement**

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

425 **Author Contribution Statement**

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

430 **Data and Code Availability**

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

434 <https://github.com/michaeldumelle/DvMsp>.

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

436 **Supporting Information**

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

439 **References**

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

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

444 Benedetti, R., Piersimoni, F., 2017. A spatially balanced design with proba-
445 bility function proportional to the within sample distance. Biometrical Journal
446 59, 1067–1084.

447 Benedetti, R., Piersimoni, F., Postiglione, P., 2017. Spatially balanced
448 sampling: A review and a reappraisal. International Statistical Review 85,
449 439–454.

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

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

454 Brus, D.J., 2021. Statistical approaches for spatial sample survey: Persistent
455 misconceptions and new developments. European Journal of Soil Science 72,
456 686–703.

457 Chan-Golston, A.M., Banerjee, S., Handcock, M.S., 2020. Bayesian inference
458 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., 2021. 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.

Foster, S.D., Hosack, G.R., Lawrence, E., Przeslawski, R., Hedge, P., Caley, M.J., Barrett, N.S., Williams, A., Li, J., Lynch, T., others, 2017. Spatially balanced designs that incorporate legacy sites. *Methods in Ecology and Evolution* 8, 1433–1442.

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

486 sampling through the pivotal method. *Biometrics* 68, 514–520.

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

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

492 Harville, D.A., 1977. Maximum likelihood approaches to variance compo-
493 nent estimation and to related problems. *Journal of the American Statistical*
494 *Association* 72, 320–338.

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

497 Horvitz, D.G., Thompson, D.J., 1952. A generalization of sampling with-
498 out replacement from a finite universe. *Journal of the American Statistical*
499 *Association* 47, 663–685.

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

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

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

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

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

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

512 Sen, A.R., 1953. On the estimate of the variance in sampling with varying

513 probabilities. *Journal of the Indian Society of Agricultural Statistics* 5, 127.

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

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

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

521 USEPA, 2012. National lakes assessment 2012. [https://www.epa.gov/national-](https://www.epa.gov/national-aquatic-resource-surveys/national-results-and-regional-highlights-national-lakes-assessment)
522 [aquatic-resource-surveys/national-results-and-regional-highlights-national-lakes-](https://www.epa.gov/national-aquatic-resource-surveys/national-results-and-regional-highlights-national-lakes-assessment)
523 [assessment](https://www.epa.gov/national-aquatic-resource-surveys/national-results-and-regional-highlights-national-lakes-assessment).

524 Ver Hoef, J., 2002. Sampling and geostatistics for spatial data. *Ecoscience* 9,
525 152–161.

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

528 Ver Hoef, J.M., Temesgen, H., 2013. A comparison of the spatial linear
529 model to nearest neighbor (k-nn) methods for forestry applications. *PLoS ONE*
530 8, e59129.

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

534 Wang, J.-F., Stein, A., Gao, B.-B., Ge, Y., 2012. A review of spatial sampling.
535 *Spatial Statistics* 2, 1–14.

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