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

The design-based and model-based approaches to frequentist statistical inference lie on fundamentally different foundations. In the design-based approach, inference depends on random sampling. In the model-based approach, inference depends on distributional assumptions. In this manuscript, we compare the approaches for finite population spatial data. We first provide relevant background for the approaches and then use a simulation study and an analysis of real mercury concentration data to numerically compare them. We find that sampling plans that incorporate spatial locations (spatially balanced samples) perform better than sampling plans ignoring spatial locations (non-spatially balanced samples), regardless of whether design-based or model-based approaches were used to analyze the data. We also find that within sampling plans, the model-based approaches often outperform design-based approaches, even for skewed data. This gap in performance is small when spatially balanced samples are used but large when non-spatially balanced samples are used.

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 is called a sample. In the design-based approach, inferences about the underlying population are informed via a probabilistic process assigning some population units to be part of 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

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

data. We define spatial data as data that incorporates the specific locations of the population units into either the design 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), Cicchitelli and Montanari (2012), 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 incorporating spatial information and design-based approaches. In this manuscript, we compare design-based approaches incorporating spatial information to model-based approaches for spatial data. We focus on finite populations, but these comparisons generalize to infinite populations as well. A finite population contains a finite number of population units; an example is lakes (treated as a whole with the lake centroid representing location) in the contiguous United States. An infinite population contains an infinite number of population units; an example is locations within a single lake.

The rest of the manuscript is organized as follows. In Section 2, we introduce and provide relevant background for the design-based and model-based approaches to finite population spatial data. In Section 3, we use a simulation study to compare the performance of the approaches in a variety of scenarios. In Section 4, we compare the performance of the approaches on real data that contains mercury concentration in lakes from the contiguous United States. And in Section 5, we end with a discussion and provide directions for future research.

2. 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.

2.1. Comparing Design-Based and Model-Based Approaches

The design-based approach assumes the population is fixed. Randomness is incorporated via the selection of units in a sampling frame. A sampling frame is the set of all units available to be sampled. Units from the sampling frame are selected as part of the sample according to a sampling design, which assigns a positive probability of inclusion (inclusion probability) to each unit from the sampling frame. Some examples of commonly used sampling designs include simple random sampling, stratified random sampling, and cluster sampling. When sampling designs incorporate spatial locations into sampling, we call the resulting samples "spatially balanced." One approach to selecting spatially balanced samples is the Generalized Random Tessellation Stratified (GRTS) algorithm (Stevens and Olsen, 2004), which we discuss in more detail in Section 2.2. 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.

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, or value. For example, suppose the realized mean of all population units is the value of interest. Instead of *estimating* a fixed, unknown mean, we are *predicting* the value of the mean, a random variable. Prediction intervals are then derived using assumptions of the data-generating stochastic process. If we repeatedly generate response values from the same data-generating stochastic 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 Figure 1, we provide a visual comparison of the design-based and model-based approaches (Ver Hoef (2002) and Brus (2021) provide similar figures).

2.2. Spatially Balanced Design and Analysis

We previously mentioned that the design-based approach can be used to select spatially balanced samples (samples that incorporate spatial locations of the population units and are "well-spread" is space). Spatially balanced samples are useful because parameter estimates from these samples tend to

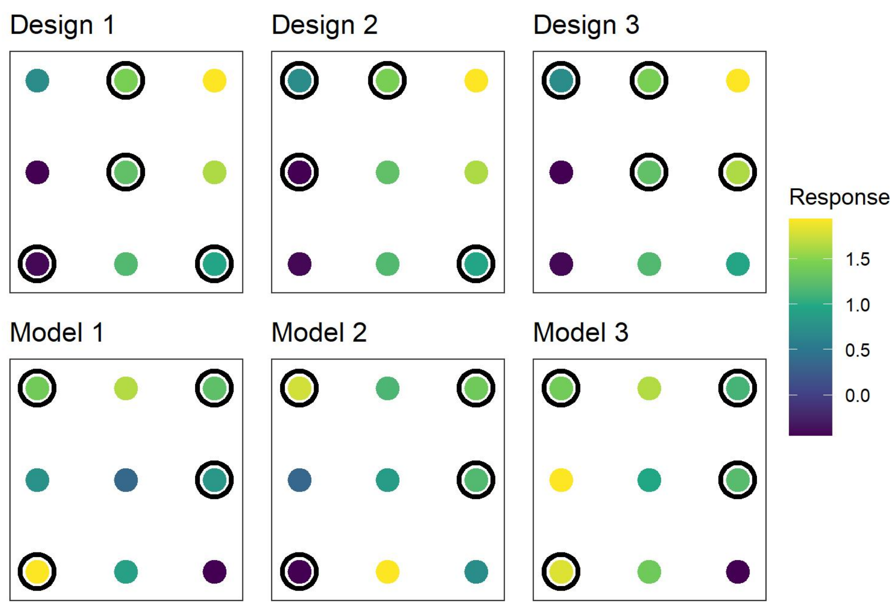

Figure 1: A visual comparison of the design-based and model-based approaches. In the top row, there is one fixed population with nine population units and three random samples of size four (points circled are those sampled). The response values at each site are fixed, but we obtain different estimates for the mean response in each random sample. In the bottom row, 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

vary less than parameter estimates from samples that are not spatially balanced (Barabesi and Franceschi, 2011; Benedetti et al., 2017; Grafström and Lundström, 2013; Robertson et al., 2013; Stevens and Olsen, 2004; Wang et al., 2013). The first spatially balanced sampling algorithm seeing widespread use is 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 (Dirichlet Tessellations). After the GRTS algorithm was developed, several other spatially balanced sampling algorithms emerged, such as the Local Pivotal Method (Grafström et al., 2012; Grafström and Matei, 2018), Spatially Correlated Poisson Sampling (Grafström, 2012), Balanced Acceptance Sampling (Robertson et al., 2013), Within-Sample-Distance Sampling (Benedetti and Piersimoni, 2017), and Halton Iterative Partitioning Sampling (Robertson et al., 2018). In this manuscript, we select spatially balanced samples using the Generalized Random Tessellation Stratified (GRTS) algorithm because it has several attractive properties. More specifically, the GRTS algorithm accommodates finite and infinite sampling frames, equal, unequal, and proportional (to size) inclusion probabilities, legacy (historical) sampling (Foster et al., 2017), a minimum distance between units in a sample, and replacement units (replacement units are population units that can be sampled when a population unit originally selected can no longer be sampled). The GRTS algorithm selects samples by utilizing a particular mapping between two-dimensional and one-dimensional space that preserves proximity relationships. Via this mapping, units in two-dimensional space are partitioned using a hierarchical address. This hierarchical address is used to map population units to a one-dimensional line. On the one dimensional line, each population unit's line length equals its inclusion probability. Then, a systematic sample of population units is selected on the line, yielding desired sample. Stevens and Olsen (2004) provides 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 estimate of τ, denoted by $\hat{\tau}_{ht}$, is is given by

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

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

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

2.3. Finite Population Block Kriging

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

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

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

FBPK assumes $\boldsymbol{\delta}$ in Equation 2 has mean-zero and a spatial correlation structure that can be modeled using a covariance function. This covariance function is commonly assumed to be non-negative (between zero and one), 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 both. A variety of flexible covariance functions can be used to model $\boldsymbol{\delta}$ (Cressie, 1993); one example is the exponential covariance function (for a thorough list of spatial covariance functions, see Cressie (1993). The i, jth element of the exponential covariance matrix, $\mathrm{cov}(\boldsymbol{\delta})$, is

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

where σ_1^2 is the variance parameter quantifying the variability that is dependent (coarse-scale), σ_2^2 is the variance parameter quantifying the variability that is independent (fine-scale), ϕ is the range parameter measuring the distance-decay rate of the covariance, and $h_{i,j}$ is the Euclidean distance between population 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), random forests (Breiman, 2001), Bayesian models (Chan-Golston et al., 2020), among others, could also be used to obtain predictions for a mean or total from spatially correlated responses of a finite population. 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 forests in a variety of spatial data contexts, and FBPK tended to perform best. Compared to the Bayesian approach, we prefer FPBK mostly because it is much more computationally efficient.

3. Simulation Study

We used a simulation study to investigate performance of four sampling-analysis combinations: IRS with a design-based analysis, called "IRS-Design"; IRS with a model-based analysis, called "IRS-Model"; GRTS sampling with a design-based analysis, called "GRTS-Design"; 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, 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 (of the population units) 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 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 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 simualtion scenario, the total variance was two.

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 confidence (design-based) or 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 **R** package `spsurvey` (Dumelle et al., 2021). FPBK 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).

The average bias was nearly zero for all four combinations in all 36 scenarios, so we omit a more detailed summary of those results here. Tables for average bias in all 36 simulation scenarios are provided in the supplementary material.

Figure 2 shows the relative rMS(P)E of the four approaches from Table 1

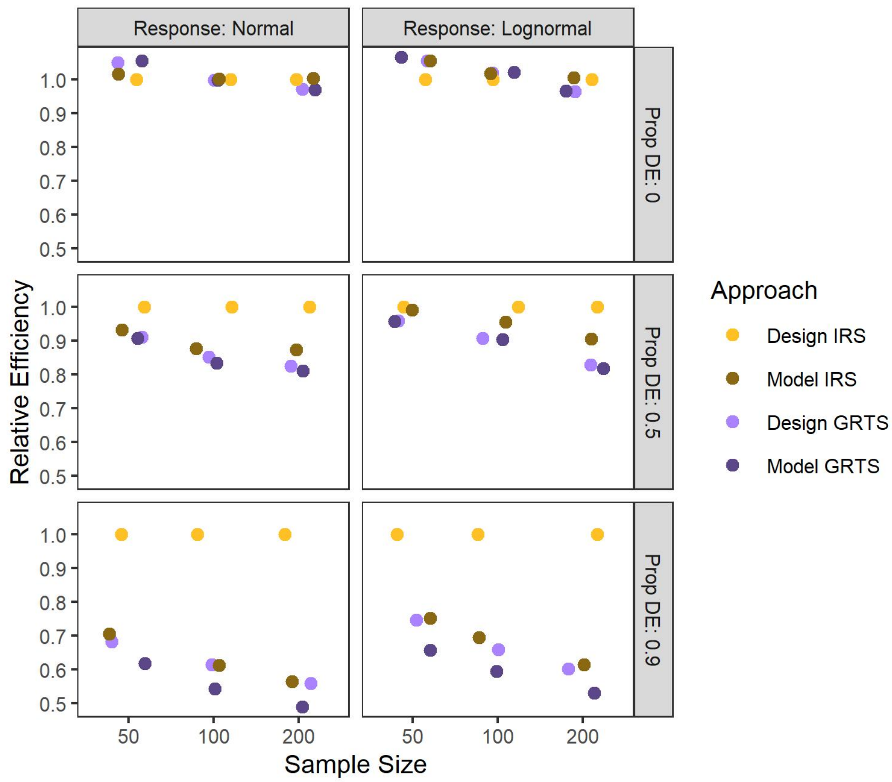

Figure 2: Relative rMS(P)E for the four sampling-analysis combinations. The rows indicate the proportion of dependent error and the columns indicate the response type.

using the random location layout with "IRS-Design" as the baseline.

The relative rMS(P)E is defined as

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

When there is no spatial correlation (Figure 2, "Prop DE: 0" row), the four sampling-analysis combinations have approximately equal rMS(P)E. So using the GRTS sampling plan or a model-based analysis does not result in much, if any, loss in efficiency compared to IRS-Design when there is no spatial correlation. When there is spatial correlation (Figure 2, "Prop DE: 0.5" and "Prop DE: 0.9" rows), GRTS-Model tends to perform best, 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 correlation increases, the gap in rMS(P)E between IRS-Design and the other sampling-analysis combinations widens. Finally we note that when there is spatial correlation, IRS-Model outperforms IRS-Design by a large margin, suggesting that the poor design properties of IRS are largely mitigated by the

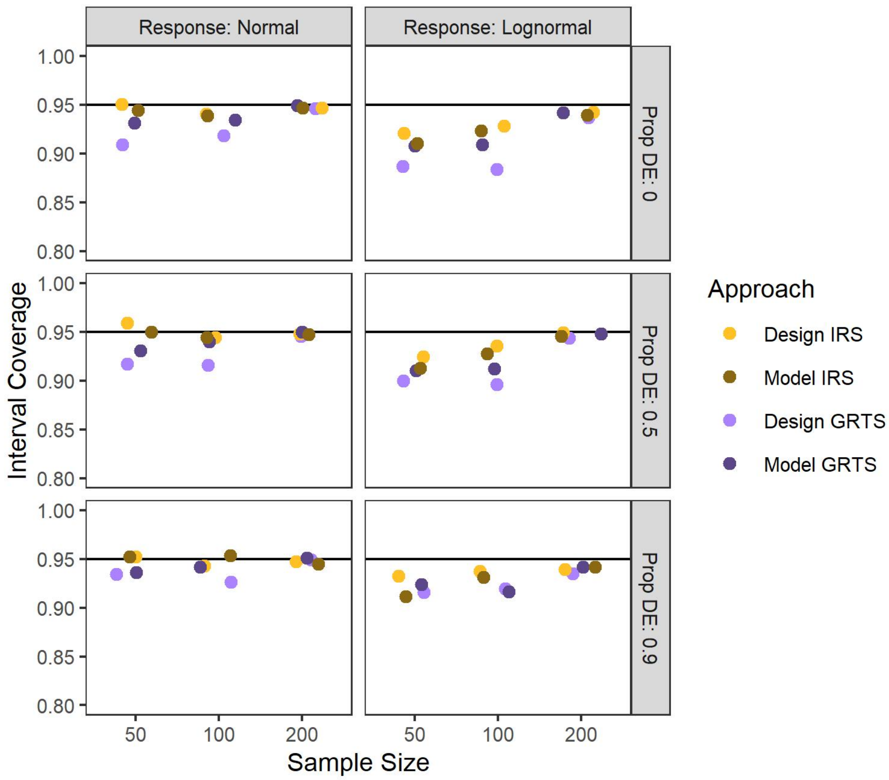

Figure 3: Interval coverage for the four sampling-analysis combinations. The rows indicate the proportion of dependent error and the columns indicate the response type. The solid, black line in each plot represents 95% coverage.

model-based analysis. These 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 supplementary material.

We also studied 95% interval coverage among the sampling-analysis combinations. The design-based confidence intervals and model-based prediction intervals were constructed using the normal distribution. Justification for this comes from the asymptotic normality of means via the Central Limit Theorem.

Figure 3 shows the 95% interval coverage for each of the four sampling-analysis combinations in the random location layout.

Within each scenario, the sampling-analysis combinations tend to have fairly similar interval coverage. Coverage in the normal response scenarios was usually near 95%, while coverage in the lognormal response scenarios varied from from 90% to 95%. Coverage tended to always increase with the sample size. At a sample size of 200, all four sampling-analysis combinations had approximately 95% interval coverage in both response scenarios for all dependent error proportions. These conclusions are similar to those observed in the grid location layout, so we omit a grid location layout figure here. Tables for interval coverage in all 36 simulation scenarios are provided in the supplementary material.

4. Application

The Environmental Protection Agency (EPA), states, and tribes periodically conduct National Aquatic Research Surveys (NARS) in the United States to assess the water quality of various bodies of water. We will use the 2012 National Lakes Assessment (NLA), which measures various aspects of lake health and water quality for lakes in the contiguous United States. Specifically, we will analyze mercury concentration in lakes. Although we know the true mean mercury concentration values for the 986 lakes from the 2012 NLA, we will explore whether or not we obtain an adequately precise estimate for the realized mean mercury concentration if we sample only 100 of the 986 lakes.

Figure 4 shows that mercury concentration is right-skewed, with most lakes having a low value of mercury concentration but a few having a much higher concentration. Mercury concentration exhibits some spatial correlation, with high mercury concentrations in lakes in the northeast and north central United States. The realized mean mercury concentration in the 986 lakes is 103.2 ng / g. TALK ABOUT EMPIRICAL SEMIVARIOGRAM IN THIS PARAGRAPH.

We selected a single IRS sample and a single GRTS sample and estimated (design-based) or predicted (model-based) the mean mercury concentration and its standard error using using design-based and model-based approaches. For the model-based analyses, the exponential covariance was used. Table 3 shows the results from these analyses. For all four sampling-analysis combinations, the true realized mean mercury concentration is within the bounds of the 95% confidence (design-based) or prediction (model-based) intervals. Though we should not generalize these results to other samples from these data, we do note a couple of patterns. The design-based IRS analysis shows the largest standard error: a likely reason is that this is the only approach that does not incorporate any spatial information regarding mercury concentration. Both analyses using GRTS sampling have lower standard errors than both analyses using IRS sampling. We expect that these patterns are consistent with other samples from these data because mercury concentration exhibits spatial patterning, so a spatially balanced sample should usually yield a lower standard error.

Approach	Estimate	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: Application of design-based and model-based approaches to the NLA data set on mercury concentration. The true mean concentration is 103.2 ng / g.

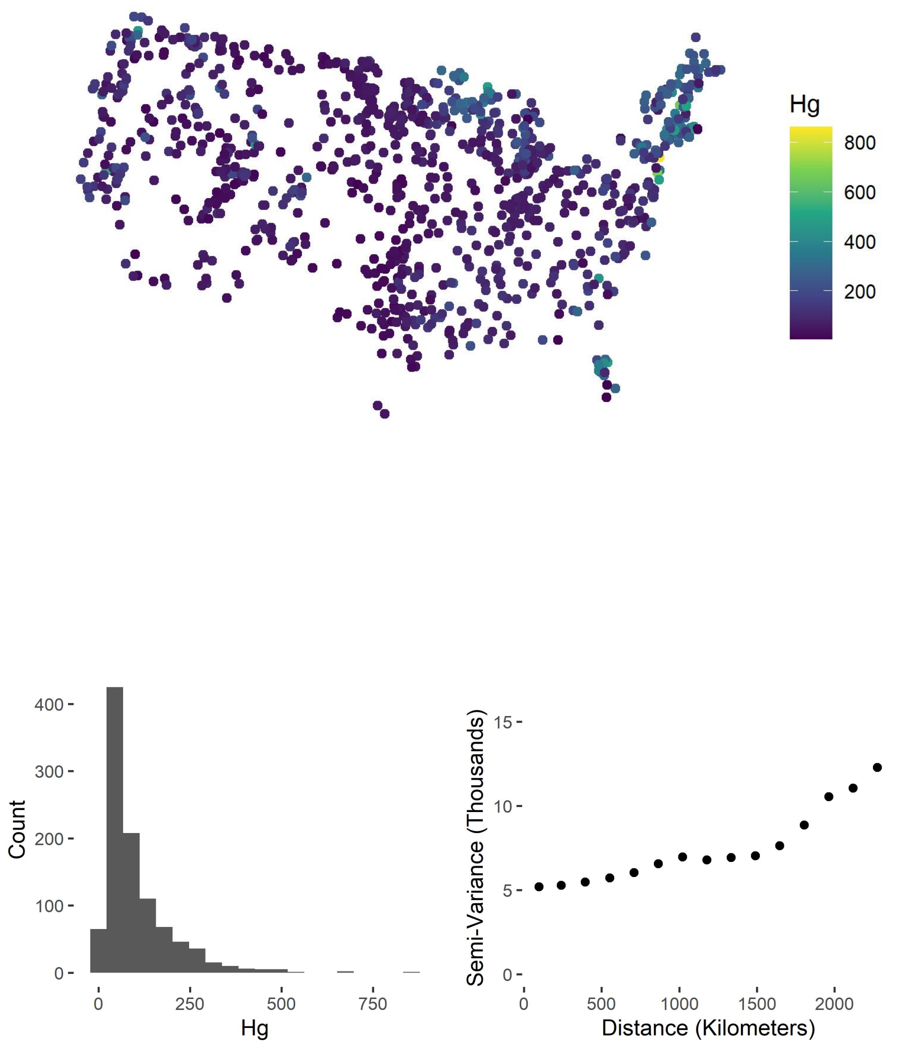

Figure 4: Mercury concentration visualizations for the population (Hg) for 986 lakes in the contiguous United States. A spatial layout is in the top row, a histogram is in the bottom row and left column, and an empirical semivariogram is in the bottom row and right column.

5. Discussion

The design-based and model-based approaches to statistical inference are fundamentally different paradigms by which samples are selected and data are analyzed. 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 random process. Though these approaches have often been compared in the literature both from theoretical and analytical perspectives, our contribution lies in studying them in a spatial context while implementing spatially balanced sampling. Aside from the theoretical differences described, a few analytical findings from the simulation study are particularly notable. First, the sampling decision (GRTS vs IRS) 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 performed at least as well as their IRS counterparts when there was no spatial correlation and noticeably better than their IRS counterparts when there was spatial correlation. Third, as the strength of spatial correlation increases, the gap in rMS(P)E between IRS-Design and the other sampling-analysis combinations also increases. Fourth and finally, when the response was normal, interval coverage for all sampling-analysis combinations was very close to 95% for all sample sizes; when the response was lognormal, interval coverage for all sampling and analysis was between 90% and 95% and closest to 95% when $n = 200$.

There are several benefits and drawbacks of the design-based and model-based approaches for finite population spatial data. Some we have discuss, but others we have not and they are worthy of consideration in future research. Design-based approaches are often computationally efficient, while model-based estimation can be computationally burdensome, especially for likelihood-based methods such as REML that rely on inverting a covariance matrix. The design-based approach also more naturally handles binary data, free from the more complicated logistic regression framework commonly used to analyze binary data in a model-based approach. The model-based approach, however, can more naturally quantify the relationship between covariates (predictor variables) and response variable. The model-based approach also yields estimated spatial covariance parameters, which help better understand the dependence structure in the process of study. Model selection is also possible using model-based approaches and criteria such as cross validation, likelihood ratio tests, or AIC (Akaike, 1974). Model-based approaches are capable of more efficient small-area estimation than design-based approaches by leveraging distributional assumptions in areas with few observed sites. Model-based approaches can also compute site-by-site predictions at unobserved locations and use them to construct informative visualizations. The benefits and drawbacks of both approaches, alongside our theoretical and analytical comparisons, can motive the process of choosing among them. This is especially true from an analysis perspective, as we found that using a spatially balanced sampling algorithm benefits both design-based and model-based analyses.

Data and Code Availability

This manuscript has a supplementary R package that contains all of the data and code used. Instructions for download at available at https://github.com/michaeldumelle/DvMsp.

Supplementary Material

In the supplementary material, we provide tables presenting summary statistics for all 36 simulation scenarios.

Acknowledgments

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

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.

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.

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

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

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

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

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

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.

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

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

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

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

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

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