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

This is the abstract.

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 (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 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 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, 2020; Ver Hoef, 2002; Ver Hoef, 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).

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

Though comparisons between design-based and model-based approaches to spatial data have been studied, no numerical comparison has been made between design-based approaches that incorporate spatial locations and model-based approaches. In this manuscript, we compare design-based approaches that incorporate spatial locations 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 compare several sampling and estimation procedures of the design-based and model-based approaches for finite population spatial data. In Section 3, we use a simulation approach to study the behavior and performance of both approaches. In Section 4, we use both approaches to analyze real data consisting of mercury concentration from lakes in 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 and its effects on subsequent inferences. We then discuss specific inference methods of the approaches 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 smapling frame according to a sampling design. A sampling frame is the set of all units available to be sampled. A sampling design assigns a positive probability of inclusion (inclusion probability) to each unit in the sampling frame. Some examples of commonly used sampling designs include simple random sampling, stratified random sampling, and cluster sampling. If a sampling design selects units from the sampling frame while ignoring their spatial locations, we call them "Independent Random Sampling" (IRS) designs. If a sampling design selects units from the sampling frame while incorporating their spatial locations, we call them spatially balanced designs. Spatially balanced designs can be obtained using the Generalized Random Tessellation Stratified (GRTS) algorithm (Stevens and Olsen, 2004), which we discuss in more detail in Section 2.2. The design-based approach combines the randomness of the sampling design and the data collected via the sample to estimate fixed, unknown parameters (e.g., means and totals) of a population.

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 randomizations of sampling units selected via the sampling design. Means and totals, for example, are asymptotically normally distributed (normal) by the Central Limit Theorem (under some assumptions). If we repeatedly sample the surface, 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 random sampling.

Instead of estimating fixed but unknown parameters like a mean or total (as in the design-based approach), the goal of model-based inference in the spatial context is often 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 process. If we repeatedly generate the response values from the same spatial process and sample, 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 reviews of model-based approaches for spatial data. A visual comparison of the design-based and model-based assumptions is provided in Figure 1 (Ver Hoef (2002) and Brus (2020) provide similar figures).

2.2. Spatially Balanced Design and Analysis

The design-based approach can be used to select samples that are "well-spread" in space, or spatially balanced. Spatially balanced samples are useful because parameter estimates from these samples tend to vary less than parameter estimates from samples that are not spatially balanced (Barabesi and Franceschi, 2011; Benedetti et al., 2017; Grafström and Lundström, 2013; Robertson et al., 2013; Stevens and Olsen, 2004; Wang et al., 2013). The first spatially balanced sampling algorithm that saw 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 Voroni polygons. After the GRTS algorithm was developed, several other spatially balanced sampling algorithms have emerged, including 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 use the Generalized Random Tessellation Stratified (GRTS) algorithm to select spatially balanced samples sampling because the

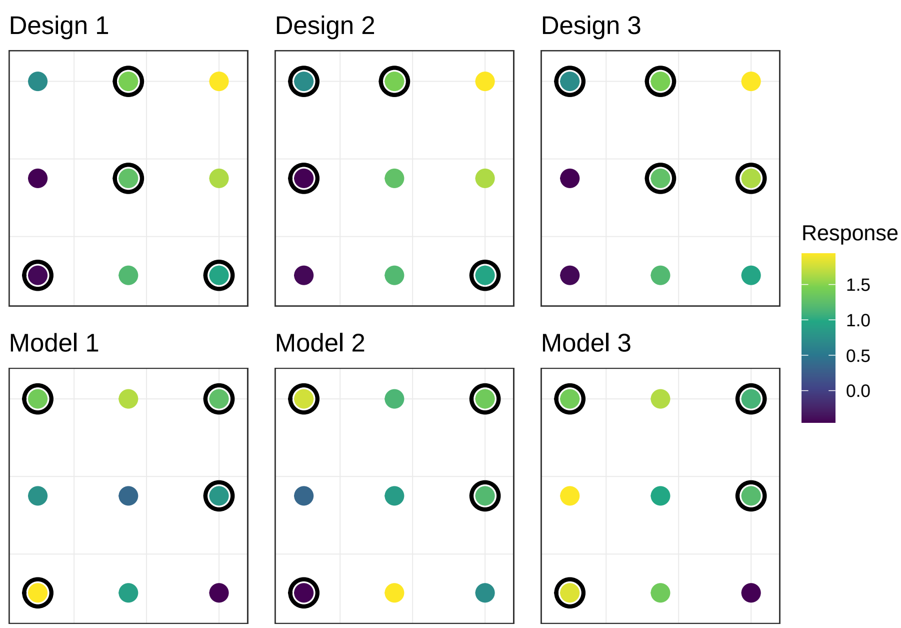


Figure 1: A comparison of sampling under the design-based and model-based frameworks. Points circled are those that are sampled. In the top row, we have one fixed population, and three random samples of size four. The response values at each site are fixed, but we obtain different estimates for the mean response because the randomly sampled sites vary from sample to sample. In the bottom row, we have three realizations of the same spatial process sampled at the same locations. The spatial process generating the response values has a single mean, but the realized mean is different in each of the three panels.

algorithm has several attractive properties. It accommodates finite and infinite sampling frames. It accommodates equal, unequal, and proportional (to size) inclusion probabilities. It accommodates legacy (historical) sampling (Foster et al., 2017). It accommodates a minimum distance between units in a sample. Lastly, it accommodates replacement units in a sample, which are units that can be sampled in place of an original unit that can no longer be sampled. The GRTS algorithm samples from finite and infinite populations by utilizing a mapping between two-dimensional and one-dimensional space. The units in the two-dimensional sampling frame are divided into cells using a hierarchical address. This hierarchical address is then used to map the units from two-dimensional space to a one-dimensional line where each unit's line length equals its inclusion probability. A systematic sample is conducted on the line and linked back to a unit in two-dimensional space, which results in the desired sample. Stevens and Olsen (2004) provides further details.

After selecting a spatially balanced sample using the GRTS algorithm (i.e., a GRTS sample), data are collected and used to estimate population parameters. To unbiasedly estimate population means and totals from sample data, one can use 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 unit in the sample and π_i is the inclusion probability of the ith unit in the sample. An estimate of the population mean can be obtained by dividing $\hat{\tau}_{ht}$ by number of population units.

While the Horvitz-Thompson estimator is unbiased for population means and totals, it is also important to quantify the uncertainty in these estimates. 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 unit i and 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 units in the sampling frame. 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 $\hat{\tau}$ compared to variance estimators ignoring spatial locations, yielding narrower confidence intervals for τ.

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 process. Ver Hoef (2008) gives details on the theory of FPBK, but some of the basic principles are summarized below. 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 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. For example, if we want to predict the population total across all population units, then we would use a vector of 1's for the weights.

We often only have a sample of the N population units. 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, and $\boldsymbol{\beta}$ is the parameter vector of fixed effects.

Let $\boldsymbol{\delta} \equiv [\boldsymbol{\delta}_s \ \boldsymbol{\delta}_u]'$, where $\boldsymbol{\delta}_s$ and $\boldsymbol{\delta}_u$ are random errors for teh sampled and unsampled population units, respectively. We assume $E(\boldsymbol{\delta} = \mathbf{0}$ and that there is spatial correlation in $\boldsymbol{\delta}$ that can be modeled using a covariance function. It is common to assume the covariance function is second-order stationary and isotropic (Cressie, 1993), and that the spatial covariance decreases as the separation between population units increases. Many spatial covariance functions exist, but the primary function we use throughout the simulations and applications in this manuscript is the exponential covariance function: the i, jth elment of hte 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 dependent random error variance measuring coarse-scale (correlated) variability, σ_2^2 is the independent random error variance measuring fine-scale (independent) variability, ϕ is the range parameter measuring the distance-decay rate of the correlation, and $h_{i,j}$ is the Euclidean distance between population units i and j. Often σ_1^2 and σ_2^2 are called the partial sill and nugget, respectively. Any spatial covariance function could be used in the place of the exponential, including functions that allow for non-stationarity or anisotropy (Chiles and Delfiner, 1999, pp. 80–93).

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.

We note that we only use FPBK in this paper in order to focus more on comparing the design-based and model-based approaches. Other methods, such as k-nearest-neighbors (Fix and Hodges, 1951; Ver Hoef and Temesgen, 2013), random forest (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. We choose to use FPBK because it is faster than a Bayesian approach and it was developed with theoretically-based variance estimators of means and totals for spatial data, whereas random forests and k-nearest-neighbors use ad-hoc variance estimators in most cases (Ver Hoef and Temesgen, 2013); additionally, FBPK outperformed the other methods in most scenarios.

3. Numerical Study

We used a simulation study to investigate performance of four sampling-analysis combinations: IRS-Design, IRS with a design-based analysis; IRS-Model, IRS with a model-based analysis; GRTS-Design, GRTS sampling with a design-based analysis; and GRTS-Model, GRTS sampling with a model-based analysis. 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 of the four sampling-analysis combinations were 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 were random and gridded. Locations in the random layout were selected randomly from the unit square ($[0, 1] \times [0, 1]$). Locations in the gridded layout were selected randomly on a fixed grid from 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 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 decays 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 total variance of 2 on the exponentiated scale. The lognormal responses were used to evaluate performance of the sampling-analysis approaches for data that were skewed.

a crossed design with the simulation parameters given in Table 2 for a total of 36 scenarios. All scenarios used exponential correlation with a $\sqrt{4}/3$ for $N = 900$

response values simulated on the unit square in either random locations (Layout = Random) or gridded locations (Layout = Gridded). The mean for the spatial process generating the response was set to zero.

For the lognormal scenarios, the response values were simulated using the specified correlation parameters using a normal distribution and were subsequently exponentiated. A total variance of 0.6931 and a mean of 0 on the normal scale yielded a total variance of 2 and a mean of 1.414 after exponentiation. Therefore, when the model-based methods were used for lognormal response, the correlation was mis-specified. We chose to simulate values with a lognormal distribution so that we could test the model-based analysis approach with a mis-specified model and so that we could test both analysis approaches on data that exhibits a large amount of skewness.

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 mean was zero and the variance was two.

There were 2000 simulation trials for each of the 36 parameter combinations. In each trial, response values were generated from a spatial process with the specified parameters, and a GRTS sample and an IRS sample were selected. For the GRTS sample, the design-based approach using the local neighborhood variance (GRTS-Design) and a model-based approach were applied (GRTS-Model). For the IRS sample, the design-based approach using the simple random sample variance (IRS-Design) and a model-based approach were applied (IRS-Model).

The GRTS algorithm and the local neighborhood variance estimator are available in the **R** package spsurvey (Dumelle et al., 2021). FPBK can be readily performed using the `sptotal` **R** package (Higham et al., 2021). We use sptotal for both the simulation analysis and the application, estimating covariance parameters with Restricted Maximum Likelihood (REML).

We define rMS(P)E as the root-mean-squared error (design-based) or the root-mean-squared-prediction error (model-based).Figure 2 shows the relative rMS(P)E of the four approaches from Table 1 using the random location layout with "IRS-Design" is the baseline. More formally, 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, top row), the four approaches have approximately equal rMSPE, even when the assumptions of the model-based approaches are violated. So, using GRTS or using a spatial model does not result in much, if any, loss in efficiency even if the response variable is not spatially

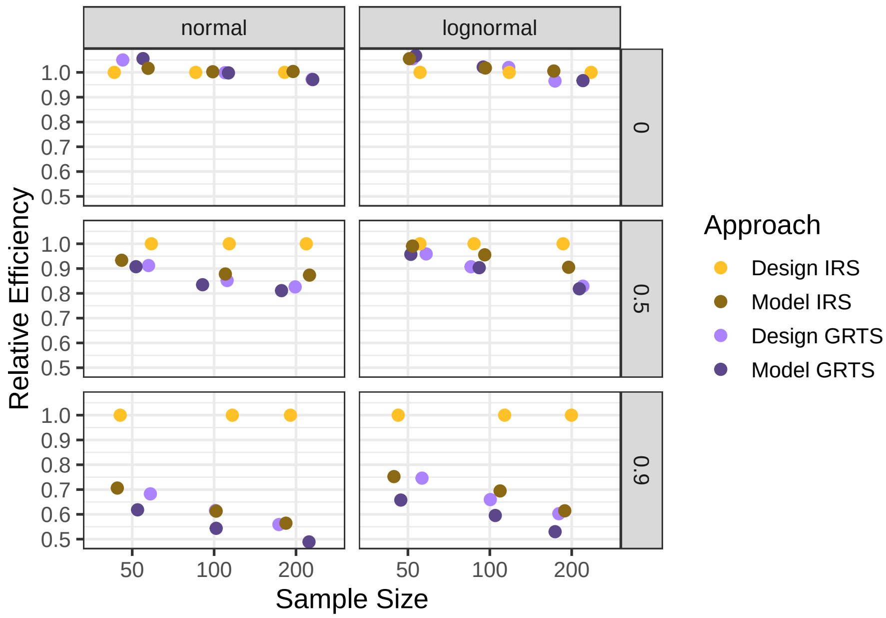


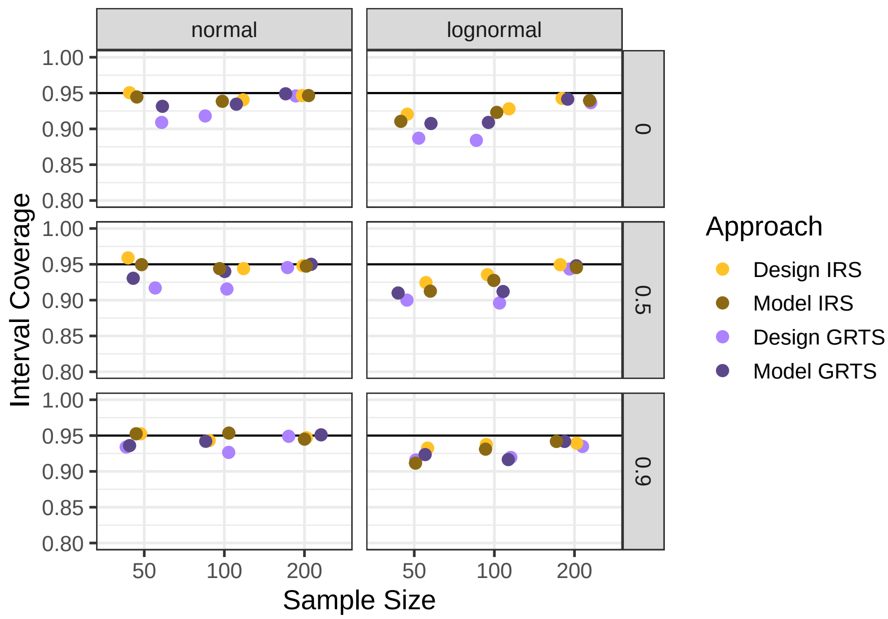
Figure 2: Relative rMS(P)E for the four sampling-analysis combinations. The rows indicate the proportion of dependent error and the columns indicate the response type.

correlated. When there is high spatial correlation (Figure 2, bottom row), the GRTS-Model approach tends to perform best, but difference in relative efficiency between GRTS-Model and GRTS-Design is small. In the lognormal, high partial sill settings (Figure 2, bottom row and right column), IRS-Model outperforms IRS-Design by a large margin, suggesting that the poor design properties of IRS are largely mitigated by the model-based analysis.

Unsurprisingly, Figure 2 also shows that, when the assumptions for GRTS-Model are satisfied, the approach outperforms GRTS-Design. However, even when the model that generates the data is different than the model used to fit the data, as in the lognormal response, the model-based approach still outperforms the design-based approach when there is a high amount of spatial correlation. Additionally, as the sample size increases, IRS-Design performs relatively worse compared to the other approaches. These conclusions were similar to those observed when the data were gridded.

We also studied 95% interval coverage among the approaches. The design-based 95% confidence intervals and model-based 95% prediction intervals were constructed using the normal distribution. Justification for the design-based and model-based intervals comes from the asymptotic normality of totals via the Central Limit Theorem (under some assumptions).

Figure 3 shows the 95% interval coverage for each of the four approaches in the random location layout. All four approaches have somewhat similar interval coverage in all settings, with GRTS-Design having slightly lower coverage when the response is normal. Coverage in the normal response settings tended to be near 95%, slightly higher than coverage in the lognormal settings, which still generally exceeded 90%. Coverage for all approaches increased with the sample size for both the normal and lognormal scenarios. At a sample size of 200, all four sampling-design combinations had approximately 95% interval coverage in both response scenarios. These conclusions were similar to those observed when the data were gridded.



\begin{figure}
\caption{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.} \end{figure}

In addition to rMS(P)E and interval coverage, we also recorded average bias. The average bias is nearly zero for all approaches in all scenarios, so we omit a visualization of the results here. The supplementary material contains tables with mean bias, rMS(P)E, and interval coverage for all 36 simulation scenarios.

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 quality in lakes in the contiguous United States, to obtain an interval for mean mercury concentration. Although we know the true mean mercury concentration values for the 986 lakes from the 2012 NLA, we will explore

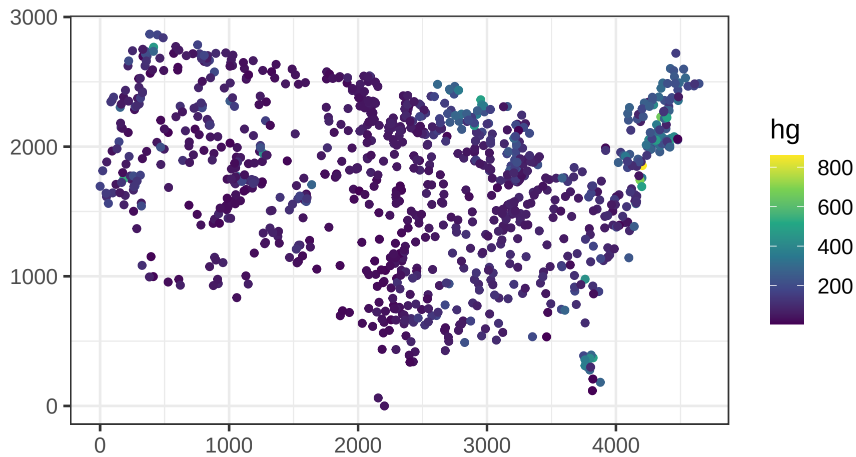


Figure 3: Population distribution of mercury concentration for 986 lakes in the contiguous United States.

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 3 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.

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.

Table 3 shows the application of a design-based analysis of an IRS sample, a model-based analysis of an IRS sample, a design-based analysis of a GRTS sample, and a model-based analysis of a GRTS sample. For all four analyses, the true realized mean mercury concentration is within the bounds of the 95% intervals.

However, we should not generalize the results of this particular realization to any other data set or even to other potential samples of this data set.

But, 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 across the contiguous United States. We also see that both approaches using the GRTS sample have a lower standard error than the both approaches using the IRS sample. We would expect this to be the case for most samples because mercury concentration exhibits spatial patterning, so a spatially balanced sample should usually yield a lower standard error.

5. Discussion

The design-based and model-based approaches to inference are fundamentally different paradigms by which to samples are selected and data are analyzed. The design-based approach incorporates randomness through sampling to estimate a population parameter. The model-based approach incorporates randomness through distributional assumptions to predict the realized value of a random variable. 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 are particularly notable: the design decision (GRTS vs IRS) seems much more important than the analysis decision (design-based vs model-based); independent of the analysis approach, there is no reason to prefer IRS over GRTS for spatial data – GRTS tends to perform at least as well as IRS when there is no spatial correlation and increasingly better than IRS as the strength of spatial correlation increases; the gap in relative efficiency between GRTS-design and GRTS-model widens as the strength of spatial correlation increases; and when the data are skewed, interval coverage for all approaches improves both as the sample size increases and as the strength of correlation increases.

There are several benefits and drawbacks of the design-based and model-based approaches for spatial data, some of which we have not yet discussed but are worthy of consideration in future research. Design-based approaches are often computationally efficient, while model-based estimation of covariance parameters 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 formulation commonly used to handle binary data in a model-based approach. The model-based approach, however, can more naturally quantify the relationship between covariates (predictor variables) and the response variable. The model-based approach also yields estimated spatial covariance parameters, which help better understand the process of study. Model selection is also possible using model-based approaches and criteria such as 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, should be heavily considered when 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 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.

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-based sampling strategies for soil (with discussion). Geoderma 80, 1–44.

Brus, D.J., 2020. Statistical approaches for spatial sample survey: Persistent misconceptions and new developments. European Journal of Soil Science.

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: The official journal of the International Environmetrics Society 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., 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., 1951. Discriminatory analysis, nonparametric discrimination: Consistency properties. USAF School of Aviation Medicine.
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.
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.
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.