

A comparison of design-based and model-based approaches for spatial data.

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

^aUnited States Environmental Protection Agency, 200 SW 35th St, Corvallis, Oregon, 97333

^bSaint Lawrence University Department of Mathematics, Computer Science, and Statistics,
23 Romoda Drive, Canton, New York, 13617

^cOregon State University Department of Statistics, 239 Weniger Hall, Corvallis, Oregon,
97331

^dMarine Mammal Laboratory, Alaska Fisheries Science Center, National Oceanic and
Atmospheric Administration, Seattle, Washington, 98115

Abstract

This is the abstract.

Text based on elsarticle sample manuscript, see <http://www.elsevier.com/author-schemas/latex-instructions#elsarticle>

Potential Journals:

- Ecological Applications
- Methods in Ecology and Evolution
- Journal of Applied Ecology
- Environmetrics
- Environmental and Ecological Statistics

1. Introduction

There are two general approaches for using data to make statistical inferences about a population: design-based approaches and model-based approaches. When data cannot be obtained for all units in a population (population units), data on a subset of the population units is collected in a sample. In the design-based approach, inferences about the underlying population are informed from a probabilistic process in which population units are selected to be in the sample. Alternatively, in the model-based approach, inferences are made from specific assumptions about the underlying process that generated the data. Each paradigm has a deep historical context (Sterba, 2009) and its own set of general advantages (Hansen et al., 1983).

Tony O.: Should this paragraph address that spatial information can be incorporated in the design stage or in the analysis stage (or both). In general, it's not clear whether we are referring to site selection process or the estimation process

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
Preprint submitted to *An awesome journal*
Email addresses: Dumelle.Michael@epa.gov (In alphabetical order Michael Dumelle),
mhigham@stlaw.edu (Matt Higham) October 10, 2021

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. Thereafter, several comparisons between design-based and model-based for spatial data have been considered, but they tend to compare design-based approaches that ignore spatial locations to model-based approaches (Brus and De Gruijter, 1997; Ver Hoef, 2002; Ver Hoef, 2008). Cooper (2006) review 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. Cicchitelli and Montanari (2012), Chan-Golston et al. (2020) for a Bayesian approach, and Sterba (2009)). More recent overviews include Brus (2020) and Wang et al. (2012), but no numerical comparison has been made between design-based approaches that incorporate spatial locations and model-based approaches.

Lisa M.: Add paragraph describing contribution of manuscript.

The rest of this paper is organized as follows. In Section 2, we compare sampling and estimation procedures between the design-based approach and the model-based approach. In Section 3, we use simulated and real data to study the behavior of both approaches. 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 for the design-based and model-based approaches for spatial data.

2.1. Comparing Design-Based vs. Model-Based

The design-based approach assumes the population is fixed. Randomness is incorporated in the selection of population units according to a sampling design. A sampling design assigns a positive probability of inclusion in the sample (inclusion probability) to each population unit. Some examples of commonly used sampling designs include simple random sampling, stratified random sample, and cluster sampling, which we refer to as Independent Random Sampling (IRS) survey designs. The goal is to use the sampling design and the sampled data to estimate population parameters like means and totals. These population parameters are traditionally assumed to be fixed but unknown.

Treating the data as fixed and incorporating randomness through the sampling design (top row of Figure 1 ((cite Brus 2021 here since our figure is similar?))) yields estimators having very few other assumptions. Confidence intervals for these types of estimators are typically derived using limiting arguments. Means

80 and totals, for example, are asymptotically normally distributed by the Central
 81 Limit Theorem. If we repeatedly sample the surface, then 95% of all 95%
 82 confidence intervals constructed from a procedure with appropriate coverage will
 83 contain the true, fixed mean. Särndal et al. (2003) and Lohr (2009) provide
 84 thorough reviews of the design-based approach.

85 **Jay VH:** I think it is important to stress that the limiting distribution is
 86 over all possible randomizations, constrained by whatever design is used.

87 **Jay VH:** quantity is vague. We should stick with variables, or realized
 88 variables (we might also call these values, but we should define and establish a
 89 consistent terminology early on.) **Matt H:** I think, though this comment is for
 90 this paragraph, we should establish the terminology earlier.

91 The model-based approach assumes the data are a random realization of a
 92 data-generating process. Randomness is often incorporated through distribu-
 93 tional assumptions on this process and need not be incorporated through random
 94 sampling (bottom row of Figure 1). Instead of estimating fixed but unknown
 95 parameters (as in the design-based approach), the goal of model-based inference
 96 in the spatial context is often *prediction* of an unknown quantity. For example,
 97 suppose the realized mean of all population units is the quantity of interest.
 98 Instead of *estimating* a fixed unknown mean, we are *predicting* the value of the
 99 mean, a random variable. We know that if we sampled all population units, we
 100 would have an exact prediction for the mean of our one realized process, without
 101 any uncertainty.

102 Assuming the data is a realization of a specific data-generating process yields
 103 predictors that are linked to distributional assumptions. These distributional as-
 104 sumptions are used to derive prediction intervals. The distributional assumptions
 105 allow the prediction intervals to be more precise. If we repeatedly generate the
 106 response values from a fixed spatial process and obtain a sample, then 95% of all
 107 95% prediction intervals constructed from a procedure with appropriate coverage
 108 will contain their respective realized means. Cressie (1993) and Schabenberger
 109 and Gotway (2017) provide reviews of model-based approaches for spatial data.

110 **Tony O.:** Before this section is it useful to have a section that lays out the
 111 general site selection and general analysis options. Thinking about site selection
 112 as design-based IRS, design-based GRTS, Arbitrary set of sites, selection for
 113 model-based. Then general analysis options as design-based no spatial, design-
 114 based spatial, model-based. This four by three table would show that model-based
 115 analyses are possible for all selection options. Design-based options with no
 116 spatial info possible for IRS-based and GRTS-based. Design-based options with
 117 spatial info possible for GRTS-based.

118 **Jay VH:** What about the design for model-based inference? Strictly speaking,
 119 it is fixed – there is no probabilistic use of a randomized design. However, we
 120 are going to have to deal with Diggle et al. (2010).

121 2.2. Spatially Balanced Design and Analysis

122 **Lisa M.:** Need a more precise definition of “miniature” in this context, and
 123 need an example.

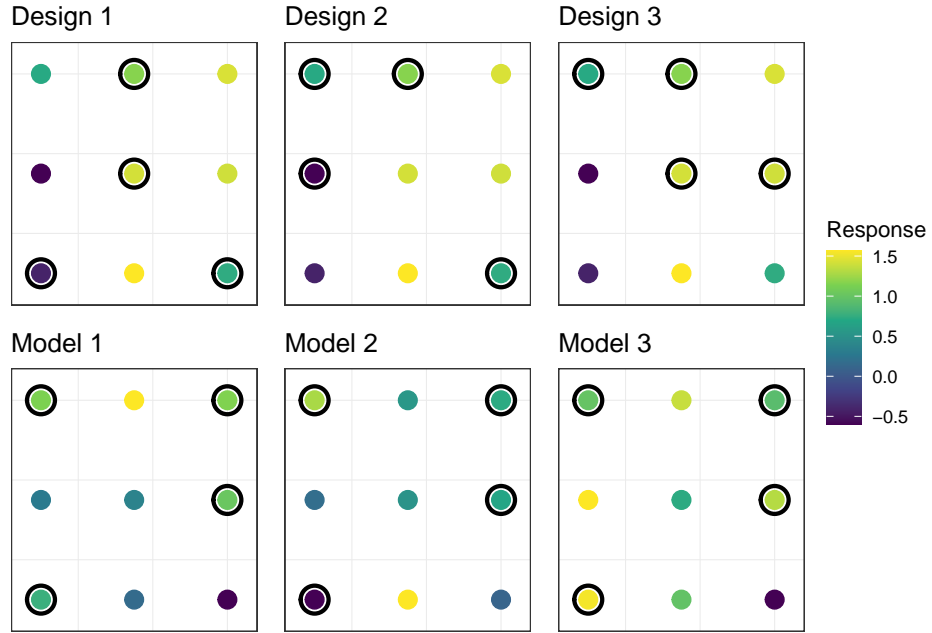


Figure 1: A comparison of sampling under the design-based and model-based frameworks. Points circled are those that are sampled. In the top row, we have one fixed population, and three random samples of $n = 4$. The response values at each site are fixed, but we obtain different estimates for the mean 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.

124 **Jay VH:** Saying “the distribution of the sampled population units mirrors
125 the density of...” is confusing to me. Are these formal statistical definitions of
126 distribution (cumulative distribution function) and density (probability density
127 function)? Wouldn’t IRS sample be a miniature, as it should, on average, mirror
128 a population?

129 The design-based approach can use spatial locations to obtain spatially
130 balanced samples. First we discuss spatial balance with respect to the population
131 (Stevens and Olsen, 2004). A sample is spatially balanced with respect to the
132 population if the sampled population units are a miniature of the population
133 units. A sample is a miniature of the population if the distribution of the sampled
134 population units mirrors the density of all population units. Spatial balance
135 with respect to the population is different than spatial balance with respect to
136 geography. A sample that is spatially balanced with respect to geography is
137 spread out in some type of equidistant manner over geographical space and is
138 not meant to be miniatures of the population. When we refer to spatial balance
139 henceforth, we mean spatial balance with respect to the population.

140 Spatially balanced samples are useful because they tend to yield estimates
141 that have lower variance than estimates constructed from sampling designs
142 lacking spatial balance (Barabesi and Franceschi, 2011; Benedetti et al., 2017;
143 Grafström and Lundström, 2013; Robertson et al., 2013; Stevens and Olsen,
144 2004; Wang et al., 2013). To quantify spatial balance, Stevens and Olsen (2004)
145 proposed loss functions based on Voroni polygons. The first spatially balanced
146 sampling algorithm that saw widespread use was the Generalized Random
147 Tessellation Stratified (Stevens and Olsen, 2004). Since GRTS was developed,
148 several other spatially balanced sampling algorithms have emerged, including
149 the Local Pivotal Method (Grafström et al., 2012; Grafström and Matei, 2018),
150 Spatially Correlated Poisson Sampling (Grafström, 2012), Balanced Acceptance
151 Sampling (Robertson et al., 2013), Within-Sample-Distance (Benedetti and
152 Piersimoni, 2017), and Halton Iterative Partitioning (Robertson et al., 2018).
153 We focus on the Generalized Random Tessellation Stratified (GRTS) algorithm
154 to select spatially balanced sampling because it has several attractive properties,
155 including **Lisa M.:** List major attractive properties, and detailed by Stevens
156 and Olsen (2004) and Dumelle et al. (2021).

157 The GRTS algorithm is used to sample from finite and infinite populations
158 and works by utilizing a mapping between two-dimensional and one-dimensional
159 space. The population units in two-dimensional space are divided into cells using
160 a hierarchical index. Population units are then mapped to a one-dimensional
161 line via the hierarchical indexing. The line length of each population unit equals
162 its inclusion probability. A systematic sample is conducted on the line and these
163 samples are linked to a population unit in two-dimensional space, which results
164 in the desired sample. Stevens and Olsen (2004) and Dumelle et al. (2021)
165 provide further details.

After collecting a sample using the GRTS algorithm, the data are used to
estimate population parameters. The Horvitz-Thompson estimator (Horvitz and
Thompson, 1952) yields unbiased estimates of population means and totals. For
example, if τ is a population total, then the Horvitz-Thompson estimator of τ

(denoted by $\hat{\tau}_{ht}$), is given by

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

where Z_i and π_i are the observed value and inclusion probability of the i th population unit selected in the sample. A similar formula exists for estimating the mean, μ . Horvitz and Thompson (1952) and Sen (1953) provide variance estimators for $\hat{\tau}_{ht}$, but they have two drawbacks. First, they rely on calculating π_{ij} , the probability that population unit i and population unit j are included in the sample, and this can be very difficult to calculate. Second, they ignore the spatial locations of the population units. To address these drawbacks, Stevens and Olsen (2003) proposed a local neighborhood variance estimator. The local neighborhood variance estimator does not rely on π_{ij} , and it incorporates spatial locations by assigning higher weights to nearby observations. Stevens and Olsen (2003) show this variance estimator tends to reduce the estimated standard error of $\hat{\tau}$, 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 basing inference off of 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 $\mathbf{z} \equiv \{z(s_1), z(s_2), \dots, z(s_N)\}$ be a response vector at locations s_1, s_2, \dots, s_N that can be measured at the N population units and is represented as an $N \times 1$ vector. 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.

However, 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 a subscript u ,

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

where \mathbf{X}_s and \mathbf{X}_u are the design matrices for the sampled and unsampled population units, respectively; β is the parameter vector of fixed effects; and $\boldsymbol{\delta}_s$ and $\boldsymbol{\delta}_u$ are random errors for the sampled and unsampled population units, respectively. Denoting $\boldsymbol{\delta} \equiv [\boldsymbol{\delta}_s \ \boldsymbol{\delta}_u]'$, we assume the expectation of $\boldsymbol{\delta}$ equals $\mathbf{0}$.

We also assume that there is spatial correlation in $\boldsymbol{\delta}$, which 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, j^{th} entry for $\text{cov}(\boldsymbol{\delta})$ is

$$\text{cov}(\delta_i, \delta_j) = \theta_1 \exp(-3h_{i,j}/\theta_2) + \theta_3 \mathbb{1}\{\mathbf{h}_{i,j} = 0\}, \quad (3)$$

where $h_{i,j}$ is the distance between population units i and j , and $\boldsymbol{\theta}$ is a vector of spatial covariance parameters of the partial sill θ_1 , the range θ_2 , and the nugget θ_3 ; and, $\mathbb{1}$ is equal to 1 when distance $h_{i,j}$ is equal to 0, and equal to 0 otherwise. However, 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).

Lisa M. : Include formulas. Perhaps, but, these are very heavy in notation and matrix algebra. We might consider, however, adding the formulas to an Appendix.

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.

We note that we only use FPBK in this paper in order to focus more on comparing the design-based and model-based approaches. However, 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, can also be used to obtain predictions for a mean or total from spatially correlated responses in a finite population setting. We choose to use FPBK because it is faster than a Bayesian approach and random forest and because Ver Hoef and Temesgen (2013) showed that the method outperforms k-nearest-neighbors in many scenarios.

3. Numerical Study

We used a numerical simulation study to investigate performance of four design-analysis combinations, summarized in Table 1.

Table 1: Types of Sampling Design and Analysis combinations considered in the simulation study. The columns give the two types of sampling designs while the rows give the two types of analyses.

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

We used a crossed design with the simulation parameters given in Table 2 for a total of 36 scenarios. All scenarios used exponential correlation with an effective range of $\sqrt{2}$ for $N = 900$ response values simulated on the unit square in either random locations (Site Locations = Random) or gridded locations (Site

Locations = 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 2 and a mean of 0 on the normal scale is equivalent to a total variance of 47 and a mean of 2.72 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.

Table 2: Simulation parameters. Total variability for all scenarios was 2 so that the partial sill was 0, 1, or 1.8.

Sample Size (n)	50	100	200
Site Locations	Random	Gridded	
Partial Sill / Total Variance	0	0.5	0.9
Response Type	Normal	Lognormal	

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 in **R** with the `sptotal` package (Higham et al., 2020). We use `sptotal` for both the simulation analysis and the application, estimating parameters with Restricted Maximum Likelihood (REML).

Figure 2 shows the relative efficiency of the four approaches from Table 1 with “IRS-Design” as the baseline:

$$E = \frac{\text{rMSPE of approach}}{\text{rMSPE of IRS-Design}},$$

where rMSPE is the root-Mean-Squared-Prediction-Error. When there is no spatial correlation (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 loss in efficiency even if the response variable is not spatially correlated. When there is high spatial correlation (bottom row), the GRTS-Model approach tends to perform

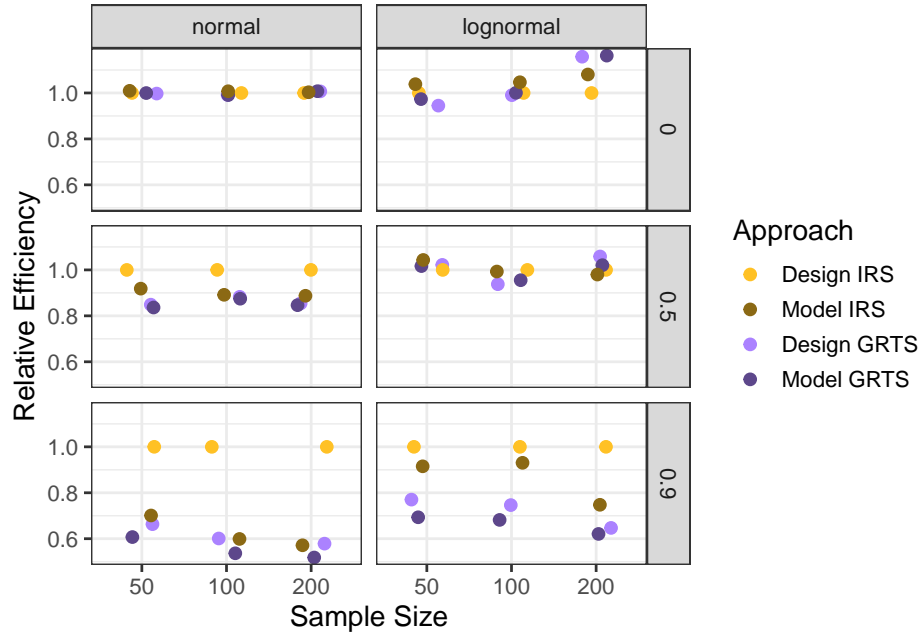


Figure 2: Relative Efficiency of the four design-analysis approaches. The plot is faceted by the type of response on the columns and the partial-sill to total-variance ratio on the rows.

best, but difference in relative efficiency between GRTS-Model and GRTS-Design is not very big. In the lognormal, high partial sill settings (bottom-right facet), GRTS-Design outperforms IRS-Model by a large margin, suggesting that the design decision (whether to use IRS or GRTS) is more important than the analysis decision (whether to analyze using model assumptions or not).

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.

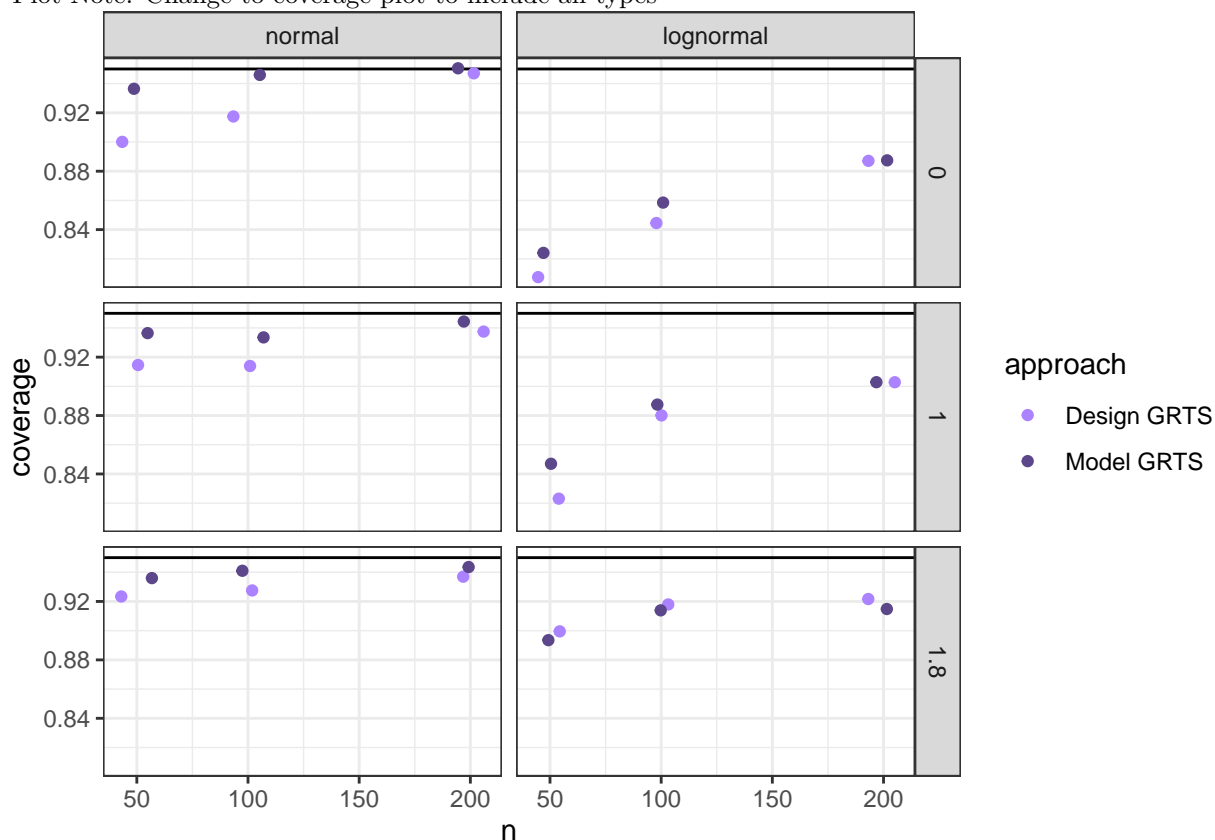
Plot Note: change colours and think about shape think about legend going on graph

MAJOR POINTS for the following Figure:

- coverage for the model-based estimator is slightly higher than coverage for the design-based estimator in the normal settings. Coverages are about equal in the lognormal settings with a slight edge to model-based (this is the point that will be tougher to explain: I would have expected the design-based estimator to have better coverage in the lognormal settings because it has fewer assumptions).
- coverage is at or near the nominal 95% in all of the normal settings, where assumptions for the model approach and the design approach are satisfied.

- for the model-based approach, the more skewed the population is, the higher the sample size needed to satisfy CLT for predicting a mean. The derivation of the BLUP is entirely moment-based (no distribution assumed) but we still need to assume a distribution to estimate spatial parameters and to generate bounds of a prediction interval.
- many confidence intervals generated for design-based approaches also rely on the CLT and the normal distribution to generate the interval. Again, for highly skewed data with a small sample size, this assumption is violated even though all of the assumptions for generating the estimator are valid.

Plot Note: Change to coverage plot to include all types

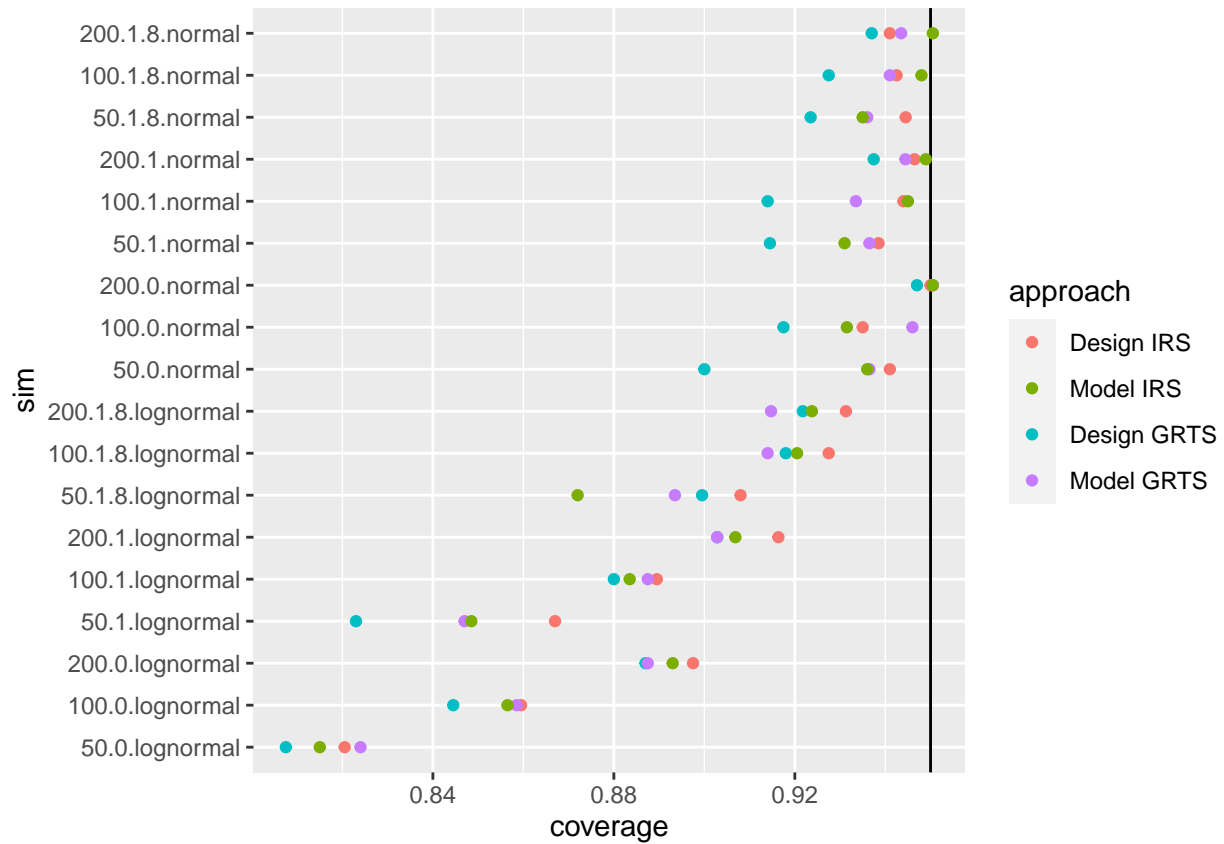


Major Points from August 3 Simulations:

1. In most of the dependent error simulation settings, either all four approaches (IRS-Design, IRS-Model, GRTS-Design, and GRTS-Model) perform equally or GRTS-Design and GRTS-Model outperform IRS-Design and IRS-Model. Exceptions to this are a couple of the settings with very small sample sizes ($n = 10$), in which the IRS does better than GRTS. In the independent error settings, it usually doesn't matter much which approach is used, which makes sense.

- 296 2. We will now focus on comparing Design-GRTS to Model-GRTS, the two
 297 best approaches for any reasonable sample size. In the independent error
 298 settings, the two approaches perform very similarly, so those results are
 299 omitted in the following graph. In the dependent error settings, using
 300 rmspe as the performance criterion, Model-GRTS outperforms Design-
 301 GRTS in 12 of the 18 settings, the two approaches perform very similarly
 302 in 3 settings, and Design-GRTS outperforms Model-GRTS in 3 settings.
- 303 3. Focusing in on the three settings where Design-GRTS outperforms Model-
 304 GRTS, we see that, in two of the settings, the log-normal response has a
 305 large variance, corresponding to a large right-skew after exponentiation.
 306 All three settings have sites in random locations. However, in only one
 307 of these settings would we recommend actually using Design-GRTS. In
 308 the other two settings, the data are sufficiently skewed that a practitioner
 309 should not use either approach, though it is “safer” to use Design-GRTS.
- 310 4. Coverage

311 For Gaussian errors, coverage for all approaches tended to be near 0.95. There
 312 was less between-approach deviation in coverages for random locations compared
 313 to grid locations. Generally, the larger the skew, the worse the coverage, and the
 314 larger the sample size, the better the coverage. GRTS-Design (local neighborhood
 315 variance) tended to slightly undercover, a result Tony was familiar with.



316

317 5. Take-home messages

- 318 • In terms of rmspe, a model-based analysis on a GRTS design yields an
319 rmspe similar to or lower than a design-based analysis on a GRTS design,
320 as long as the response variable is not “too skewed.”
- 321 • If the response variable is very skewed, then neither analysis is appropriate,
322 but, the design-based analysis is better.
- 323 • a spatially balanced GRTS sample outperforms IRS in nearly all dependent
324 error settings, as expected.
- 325 • methods that use spatial correlation generally perform better on random
326 location points than they do on gridded points. This makes some intuitive
327 sense because (1) on average, the minimum distance between an unobserved
328 point and its nearest observed neighbor should be lower for random points
329 and (2) the span of the study area is maximized for a grid based on the
330 way that we set up the simulations (with the random points being drawn
331 as uniform random variables within the boundary of the grid).
- 332 • comparison of Design-GRTS and Model-GRTS between two settings with
333 different locations of points, but otherwise the same simulation parameters,
334 should really be done on the same surface realization. One very strange

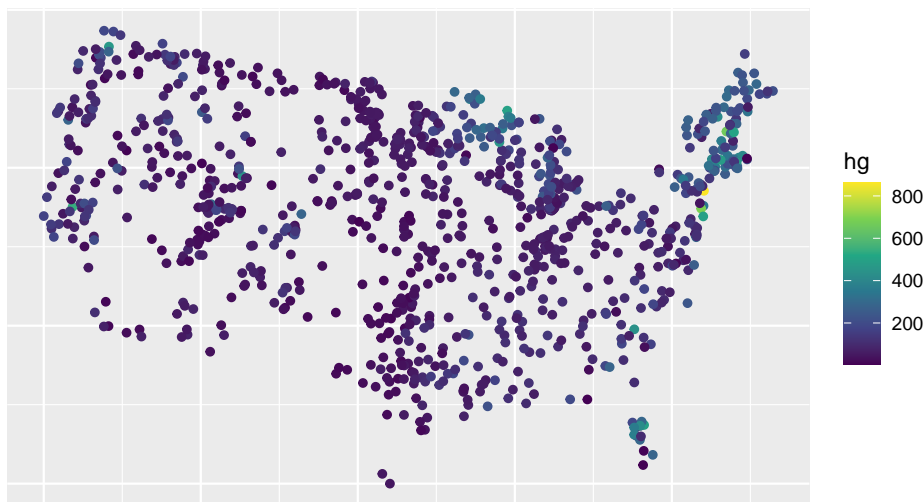


Figure 3: Population distribution of mercury concentration for 986 lakes in the contiguous United States. Thirty-five lakes were dropped from the analysis because they were missing mercury concentration.

335 realized response vector could drastically alter the results, especially on
 336 the exponentiated log data. In the same way that we compare the four
 337 approaches on the same realized data, we should also try to do the same
 338 with the locations, if they are of interest. (The realized mean won't be
 339 exactly the same but should be close).

340 4. Application

341 The Environmental Protection Agency (EPA), states, and tribes periodically
 342 conduct National Aquatic Research Surveys (NARS) in the United States to
 343 assess the water quality of various bodies of water. We will use the 2012 National
 344 Lakes Assessment (NLA), which measures various aspects of lake health and
 345 quality in lakes in the contiguous United States, to obtain an interval for mean
 346 mercury concentration. Although all lakes in the survey were measured in 2012,
 347 there may not always be enough time or money to do so. Therefore, we will
 348 explore whether or not we can still obtain an adequately precise estimate for the
 349 realized mean mercury concentration if we only take a sample of 100 of the 986
 350 lakes.

351 Figure 3 shows that mercury concentration is right-skewed, with most lakes
 352 having a low value of mercury concentration but a few having a much higher
 353 concentration. Mercury concentration exhibits some spatial correlation, with
 354 high mercury concentrations in lakes in the northeast and north central United
 355 States. Because we are considering these lakes to be our entire population, we
 356 know that the realized mean mercury concentration is 103.2 ng / g.

Table 3: Application of design-based and model-based approaches to the NLA data set on mercury concentration. The true mean concentration is 103.2 103.2 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 shows the application of a design-based analysis on an IRS, a model-based analysis on an IRS, a design-based analysis on a GRTS sample, and a model-based analysis on a GRTS sample. We see that, 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 use the spatial correlation in mercury concentration across the contiguous United States. We also see that, for the samples drawn, the both analyses with the GRTS sampling design have a lower standard error than the analyses with the IRS sampling design. We would expect this to be the case for most samples because mercury concentration exhibits spatial correlation so a spatially balanced sample should usually yield a lower standard error. If it is acceptable to have an interval for mean mercury concentration of about 25 ng / g and if we ignore the other variables that the EPA collects information on in these NLA surveys, then the EPA could consider sampling just 50 lakes to save time and money.

5. Discussion

- Pros of Design-Based (items we are not exploring): computationally efficient, few assumptions, more naturally handles binary data,
- Pros of Model-Based (items we are not exploring): covariate inference, more efficient small-area estimation, model selection?, estimated spatial parameters to better understand spatial structure, site-by-site predictions/prediction map

References

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

386 Benedetti, R., Piersimoni, F., 2017. A spatially balanced design with probability
387 function proportional to the within sample distance. *Biometrical Journal* 59,
388 1067–1084.

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

391 Breiman, L., 2001. Random forests. *Machine learning* 45, 5–32.

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

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

397 Chan-Golston, A.M., Banerjee, S., Handcock, M.S., 2020. Bayesian inference for
398 finite populations under spatial process settings. *Environmetrics* 31, e2606.

399 Chiles, J.-P., Delfiner, P., 1999. *Geostatistics: Modeling Spatial Uncertainty*.
400 John Wiley & Sons, New York.

401 Cicchitelli, G., Montanari, G.E., 2012. Model-assisted estimation of a spatial
402 population mean. *International Statistical Review* 80, 111–126.

403 Cooper, C., 2006. Sampling and variance estimation on continuous domains.
404 *Environmetrics: The official journal of the International Environmetrics*
405 *Society* 17, 539–553.

406 Cressie, N., 1993. *Statistics for spatial data*. John Wiley & Sons.

407 De Gruijter, J., Ter Braak, C., 1990. Model-free estimation from spatial samples:
408 A reappraisal of classical sampling theory. *Mathematical geology* 22, 407–415.

409 Dumelle, M., Olsen, A.R., Kincaid, T., Weber, M., 2021. Selecting and analyzing
410 spatial probability samples in r using spsurvey. *Manuscript Submitted for*
411 *Publication*.

412 Fix, E., Hodges, J.L., 1951. Discriminatory analysis, nonparametric discrimina-
413 tion: Consistency properties. *USAF School of Aviation Medicine*.

414 Grafström, A., 2012. Spatially correlated poisson sampling. *Journal of Statistical*
415 *Planning and Inference* 142, 139–147.

416 Grafström, A., Lundström, N.L., 2013. Why well spread probability samples are
417 balanced. *Open Journal of Statistics* 3, 36–41.

418 Grafström, A., Lundström, N.L., Schelin, L., 2012. Spatially balanced sampling
419 through the pivotal method. *Biometrics* 68, 514–520.

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

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

425 Higham, M., Ver Hoef, J., Bryce, F., 2020. Sptotal: Predicting totals and
426 weighted sums from spatial data.

427 Horvitz, D.G., Thompson, D.J., 1952. A generalization of sampling without
428 replacement from a finite universe. *Journal of the American statistical*
429 *Association* 47, 663–685.

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

431 Robertson, B., Brown, J., McDonald, T., Jaksons, P., 2013. BAS: Balanced
 432 acceptance sampling of natural resources. *Biometrics* 69, 776–784.
 433 Robertson, B., McDonald, T., Price, C., Brown, J., 2018. Halton iterative
 434 partitioning: Spatially balanced sampling via partitioning. *Environmental*
 435 *and Ecological Statistics* 25, 305–323.
 436 Särndal, C.-E., Swensson, B., Wretman, J., 2003. *Model assisted survey sampling*.
 437 Springer Science & Business Media.
 438 Schabenberger, O., Gotway, C.A., 2017. *Statistical methods for spatial data*
 439 *analysis*. CRC press.
 440 Sen, A.R., 1953. On the estimate of the variance in sampling with varying
 441 probabilities. *Journal of the Indian Society of Agricultural Statistics* 5, 127.
 442 Sterba, S.K., 2009. Alternative model-based and design-based frameworks for
 443 inference from samples to populations: From polarization to integration.
 444 *Multivariate behavioral research* 44, 711–740.
 445 Stevens, D.L., Olsen, A.R., 2003. Variance estimation for spatially balanced
 446 samples of environmental resources. *Environmetrics* 14, 593–610.
 447 Stevens, D.L., Olsen, A.R., 2004. Spatially balanced sampling of natural re-
 448 sources. *Journal of the american Statistical association* 99, 262–278.
 449 Ver Hoef, J., 2002. Sampling and geostatistics for spatial data. *Ecoscience* 9,
 450 152–161.
 451 Ver Hoef, J.M., 2008. Spatial methods for plot-based sampling of wildlife
 452 populations. *Environmental and Ecological Statistics* 15, 3–13.
 453 Ver Hoef, J.M., Temesgen, H., 2013. A comparison of the spatial linear model
 454 to nearest neighbor (k-NN) methods for forestry applications. *PloS one* 8,
 455 e59129.
 456 Wang, J.-F., Jiang, C.-S., Hu, M.-G., Cao, Z.-D., Guo, Y.-S., Li, L.-F., Liu, T.-J.,
 457 Meng, B., 2013. Design-based spatial sampling: Theory and implementation.
 458 *Environmental modelling & software* 40, 280–288.
 459 Wang, J.-F., Stein, A., Gao, B.-B., Ge, Y., 2012. A review of spatial sampling.
 460 *Spatial Statistics* 2, 1–14.