

1  
2

3  
4

5

6

8

10

## 12

13

14

15

- 16  
17  
18  
19  
20

## 21

22  
23  
24  
25  
26  
27  
28  
29  
30  
31

32  
33  
34  
35  
36

\*Corresponding Author

*Email addresses: Dumelle.Michael@epa.gov (In alphabetical order Michael Dumelle),  
 Preprint submitted to Mathematical Economics journal August 12, 2021*

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.

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 data are 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 independent random sampling (IRS), stratified random sampling, and cluster sampling. The goal is to use the sampling design and the sampled data to estimate population parameters like means and totals. These population parameters are typically assumed to be fixed but unknown.

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. Means and totals, for example, are asymptotically normally distributed by the Central Limit Theorem. 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 process. Randomness is often incorporated through distributional assumptions on this process. Instead of estimating fixed but unknown parameters (as in the design-based approach), the goal of model-based inference

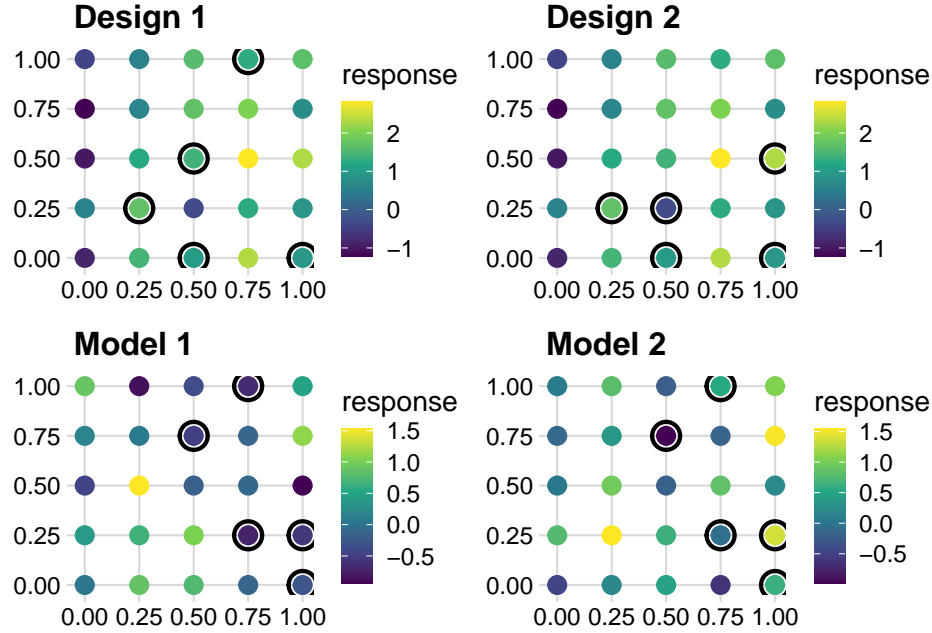


Figure 1: A comparison of sampling under the design-based and model-based frameworks. In the top row, we have one fixed population, and two random samples. In the bottom row, we have two realizations of the same spatial process sampled at the same locations.

in the spatial context is often *prediction* of an unknown quantity. For example, suppose the realized mean of all population units is the quantity of interest. Instead of *estimating* a fixed unknown mean, we are *predicting* the value of the mean, a random variable. We know that if we sampled all population units, we would have an exact prediction for the mean of our one realized process, without any uncertainty. But we are typically not interested in the true, unknown mean of the underlying process.

Assuming the data is a realization of a specific data-generating process yields predictors that are linked to distributional assumptions. These distributional assumptions are used to derive prediction intervals. The distributional assumptions allow the prediction intervals to be more precise. Cressie (1993) and Schabenberger and Gotway (2017) provide reviews of model-based approaches for spatial data.

Description of Figure 1 goes here.

## 2.2. Spatially Balanced Design and Analysis

The design-based approach can use spatial locations to obtain spatially balanced samples. First we discuss spatial balance with respect to the population (Stevens Jr and Olsen, 2004). A sample is spatially balanced with respect to the population if the sampled population units are a miniature of the population units. A sample is a miniature of the population if the distribution of the sampled

99 population units mirrors the density of all population units. Spatial balance  
100 with respect to the population is different than spatial balance with respect to  
101 geography. A sample that is spatially balanced with respect to geography is  
102 spread out in some type of equidistant manner over geographical space and is  
103 not meant to be miniatures of the population. When we refer to spatial balance  
104 henceforth, we mean spatial balance with respect to the population.

105 Spatially balanced samples are useful because they tend to yield estimates that  
106 have lower variance than estimates constructed from sampling designs lacking  
107 spatial balance (Barabesi and Franceschi, 2011; Benedetti et al., 2017; Grafström  
108 and Lundström, 2013; Robertson et al., 2013; Stevens Jr and Olsen, 2004; Wang  
109 et al., 2013). To quantify spatial balance, Stevens Jr and Olsen (2004) proposed  
110 loss functions based on Voroni polygons. The first spatially balanced sampling  
111 algorithm that saw widespread use was the Generalized Random Tessellation  
112 Stratified (Stevens Jr and Olsen, 2004). Since GRTS was developed, several  
113 other spatially balanced sampling algorithms have emerged, including the Local  
114 Pivotal Method (Grafström et al., 2012; Grafström and Matei, 2018), Spatially  
115 Correlated Poisson Sampling (Grafström, 2012), Balanced Acceptance Sampling  
116 (Robertson et al., 2013), Within-Sample-Distance (Benedetti and Piersimoni,  
117 2017), and Halton Iterative Partitioning (Robertson et al., 2018). We focus  
118 on the Generalized Random Tessellation Stratified (GRTS) algorithm to select  
119 spatially balanced sampling because it has several attractive properties detailed  
120 by Stevens Jr and Olsen (2004) and Dumelle et al. (2021).

121 The GRTS algorithm is used to sample from finite and infinite populations  
122 and works by utilizing a mapping between two-dimensional and one-dimensional  
123 space. The population units in two-dimensional space are divided into cells using  
124 a hierarchical index. Population units are then mapped to a one-dimensional  
125 line via the hierarchical indexing. The line length of each population unit equals  
126 its inclusion probability. A systematic sample is conducted on the line and these  
127 samples are linked to a population unit in two-dimensional space, which results  
128 in the desired sample. Stevens Jr and Olsen (2004) provide and Dumelle et al.  
129 (2021) 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  $\tau$  is a population total, then the Horvitz-Thompson estimator of  $\tau$   
(denoted by  $\hat{\tau}_{ht}$ ), is given by

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

130 where  $Z_i$  and  $\pi_i$  are the observed value and inclusion probability of the  $i$ th  
131 population unit selected in the sample. A similar formula exists for estimating  
132 the mean,  $\mu$ . Horvitz and Thompson (1952) and Sen (1953) provide variance  
133 estimators for  $\hat{\tau}_{ht}$ , but they have two drawbacks. First, they rely on calculating  
134  $\pi_{ij}$ , the probability that population unit  $i$  and population unit  $j$  are included in  
135 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 Jr and Olsen (2003) proposed a local neighborhood variance estimator. The local neighborhood variance estimator does not rely on  $\pi_{ij}$ , and it incorporates spatial locations by assigning higher weights to nearby observations. Stevens Jr and Olsen (2003) show this variance estimator tends to reduce the estimated standard error of  $\hat{\tau}$ , yielding narrower confidence intervals for  $\tau$ .

### 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 variable 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.

Typically, however, we 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  $u$ ,

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

where  $\mathbf{X}_s$  and  $\mathbf{X}_u$  are the design matrices for the sampled and unsampled population units, respectively;  $\boldsymbol{\beta}$  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 typically 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  $\theta_1$ , the range  $\theta_2$ , and the nugget  $\theta_3$ , and  $\mathbb{1}$  is an indicator function. 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).

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.

### 3. Numerical Study

There were several variables to alter in the simulations, and the names of the scenarios in future plots mirror these variables

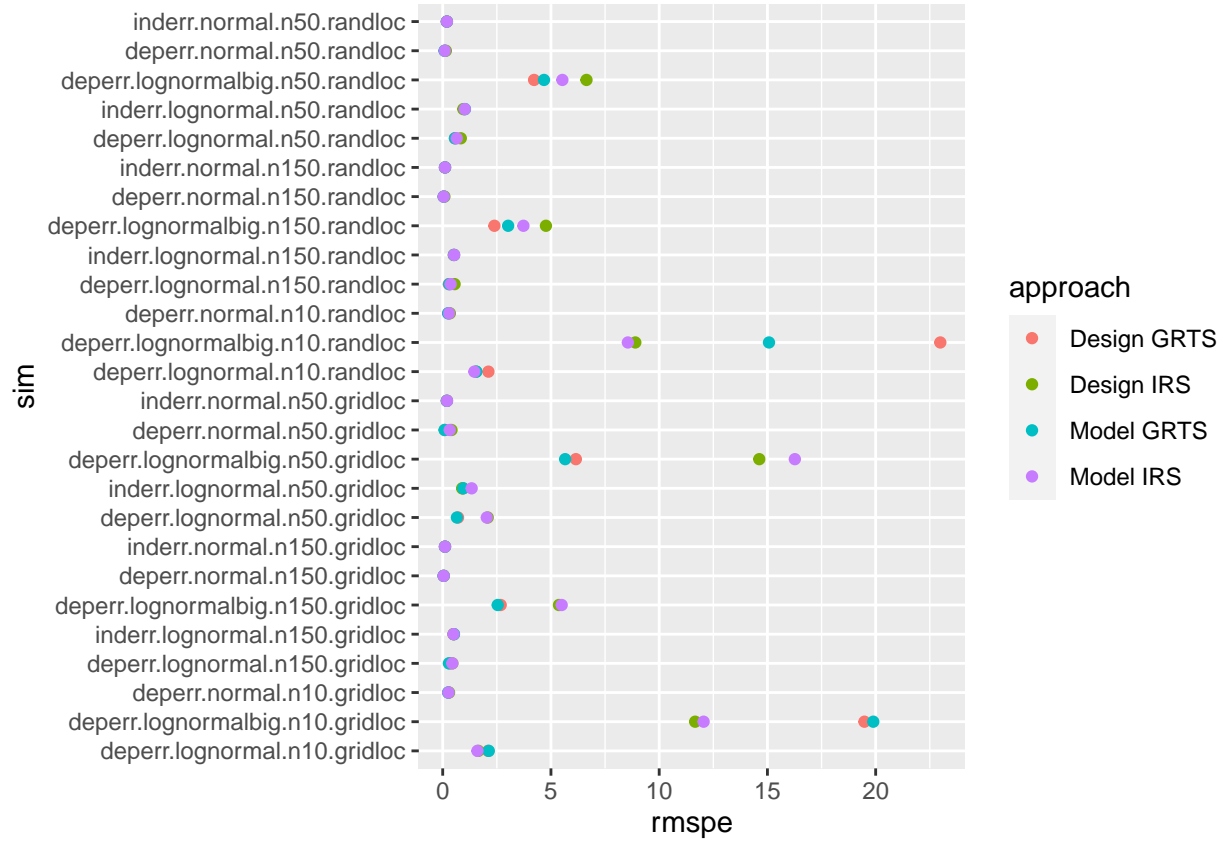
- correlation type: dependent errors or independent errors
- error type:
  - normal: mean 0, variance 2
  - lognormal: log scale mean 0, log scale variance 2
  - lognormalbig: log scale mean 0, log scale variance 4
- sample sizes:  $n = 10, 50, 150$ ;  $N = 900$
- layout: gridded vs random population locations

So for example, the `inderror.normal.n50.randloc` is the simulation having independent random errors that are normal, a sample size of 50, and random population locations.

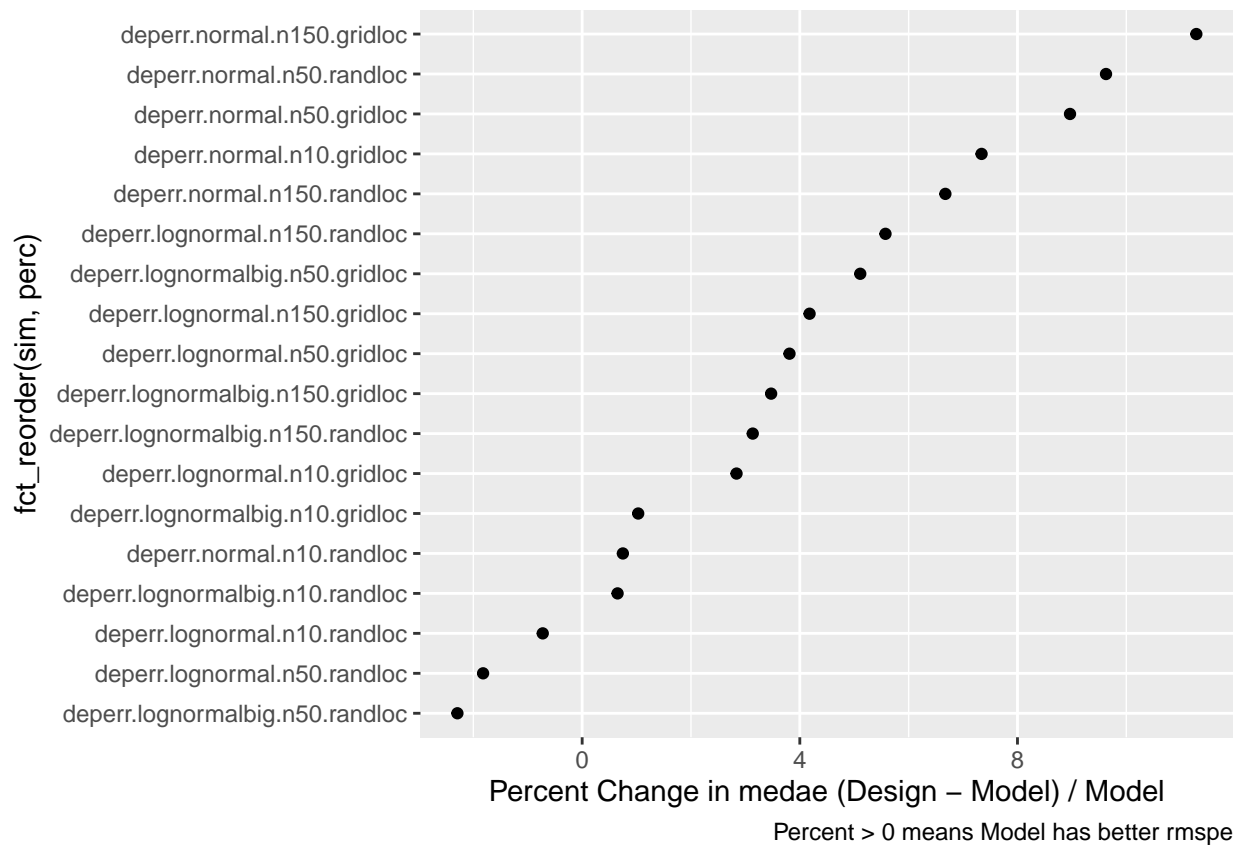
In each simulation, a GRTS sample and an IRS sample were selected. Then for the GRTS sample, the design-based approach using the local neighborhood variance (Design GRTS) and a model-based approach were applied (Model GRTS). Then for the IRS sample, the design-based approach using the simple random sample variance (Design IRS) and a model-based approach were applied (Model IRS).

Major Points from August 3 Simulations:

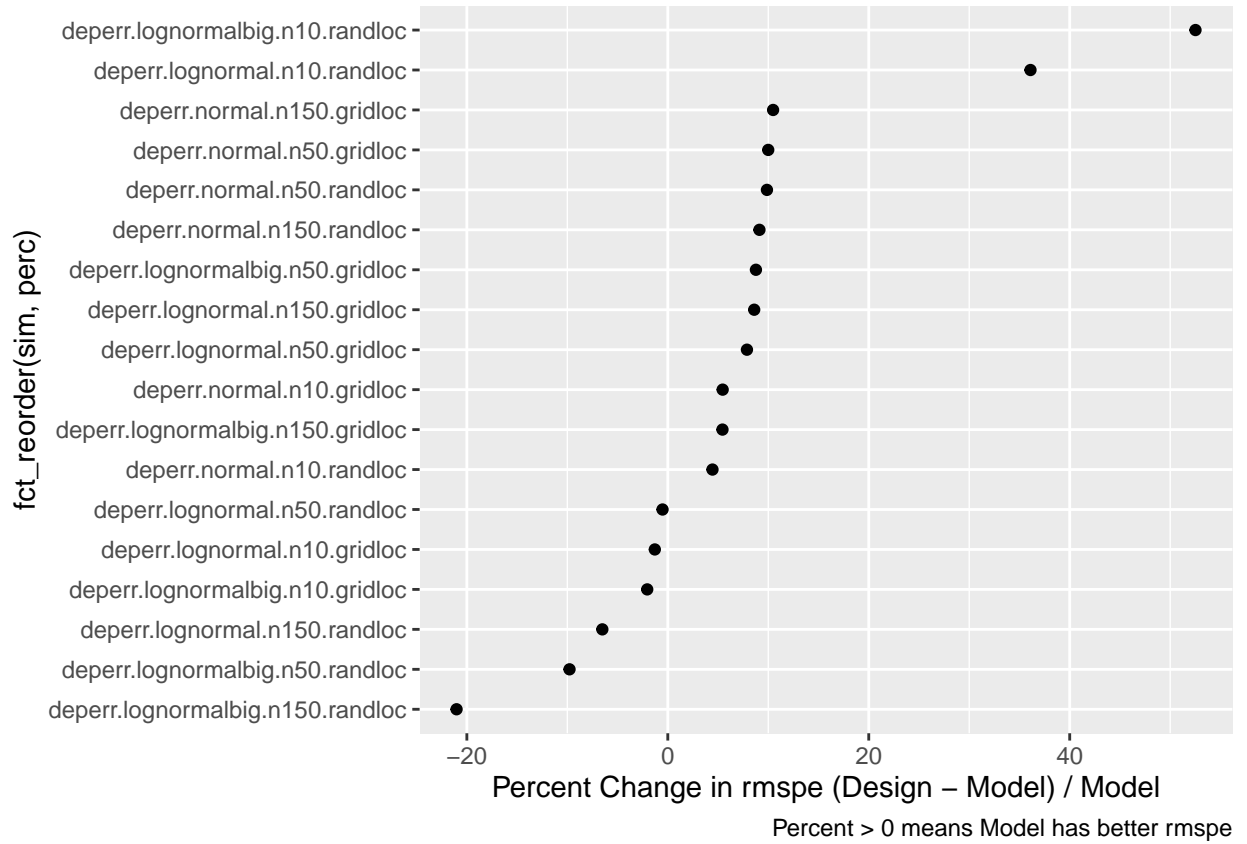
1. In most of the dependent error simulation settings, either all four approaches (IRS-Design, IRS-Model, GRTS-Design, and GRTS-Model) perform equally or GRTS-Design and GRTS-Model outperform IRS-Design and IRS-Model. Exceptions to this are a couple of the settings with very small sample sizes ( $n = 10$ ), 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.



2. We will now focus in a bit more on comparing Design-GRTS to Model-GRTS, the two best approaches for any reasonable sample size. In the independent error settings, the two approaches perform very similarly, so those results are omitted in the following graph. In the dependent error settings, using rmspe as the performance criterion, Model-GRTS outperforms Design-GRTS in 12 of the 18 settings, the two approaches perform very similarly in 3 settings, and Design-GRTS outperforms Model-GRTS in 3 settings.

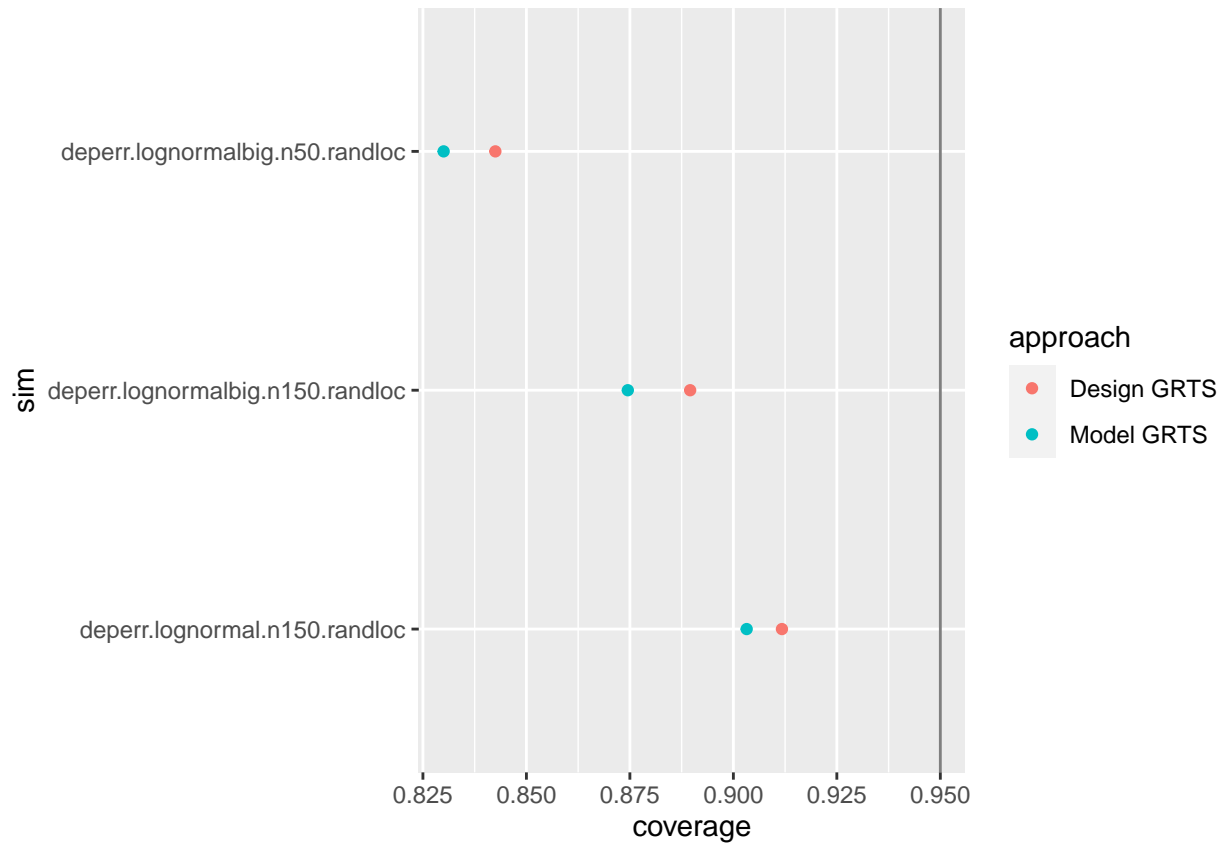






213

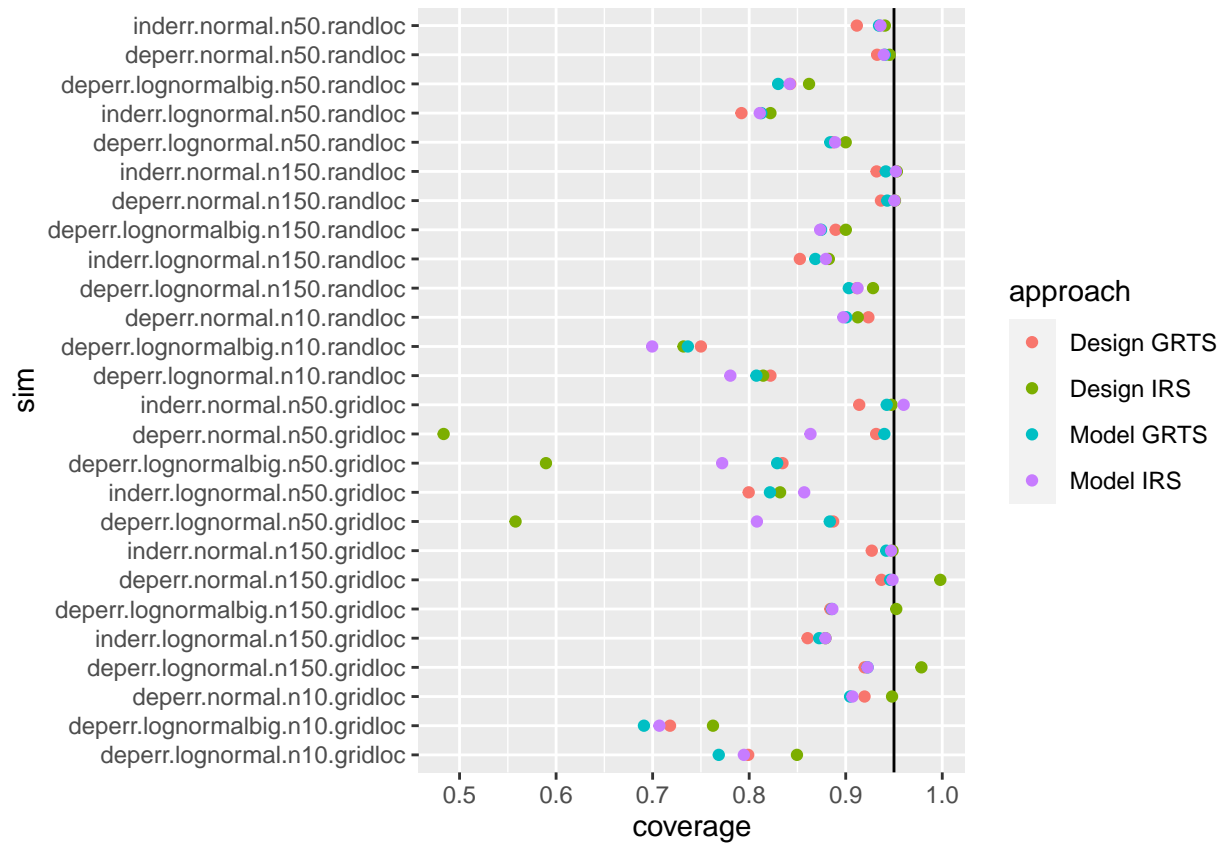
- 214 3. Focusing in on the three settings where Design-GRTS outperforms Model-  
 215 GRTS, we see that, in two of the settings, the log-normal response has a  
 216 large variance, corresponding to a large right-skew after exponentiation.  
 217 All three settings have sites in random locations. However, in only one  
 218 of these settings would we recommend actually using Design-GRTS. In  
 219 the other two settings, the data are sufficiently skewed that a practitioner  
 220 should not use either approach, though it is “safer” to use Design-GRTS.



221

#### 222 4. Coverage

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



228

## 229 5. Take-home messages

- 230
- 231 • In terms of rmspe, a model-based analysis on a GRTS design yields an  
232 rmspe similar to or lower than a design-based analysis on a GRTS design,  
233 as long as the response variable is not “too skewed.”
  - 234 • If the response variable is very skewed, then neither analysis is appropriate,  
235 but, the design-based analysis is quicker.
  - 236 • a spatially balanced GRTS sample outperforms IRS in nearly all dependent  
237 error settings, as expected.
  - 238 • methods that use spatial correlation generally perform better on random  
239 location points than they do on gridded points. This makes some intuitive  
240 sense because (1) on average, the minimum distance between an unobserved  
241 point and its nearest observed neighbor should be lower for random points  
242 and (2) the span of the study area is maximized for a grid based on the  
243 way that we set up the simulations (with the random points being drawn  
244 as uniform random variables within the boundary of the grid).
  - 245 • comparison of Design-GRTS and Model-GRTS between two settings with  
246 different locations of points, but otherwise the same simulation parameters,  
should really be done on the same surface realization. One very strange

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

### Sample Simulation

For the following simulation results, we simulated 1040 different gridded populations, each of size 900 (on the unit square) with sample size 150. For the design-based approach, population units were selected via GRTS, the Horvitz-Thompson estimator was used, and the local mean variance was used. For the model-based approach (FPBK), population units were selected via Independent Random Sampling (IRS) and the appropriate prediction and prediction variance formulas were used.

The response was normally distributed with an exponential covariance function with partial sill of 0.9, effective range of  $\sqrt{2}$ , and a nugget of 0.1. For model-based, we assumed the correct form of the covariance function (exponential), but estimated the spatial parameters with REML.

#### 3.1. Software

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).

## 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 all lakes in the survey were measured in 2012, there may not always be enough time or money to do so. Therefore, we will explore whether or not we can still obtain a relatively precise estimate for the realized mean mercury concentration if we only take a sample of 100 of the 986 lakes.

Figure 2 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. Because we are considering these lakes to be our entire population, we know that the realized mean mercury concentration is 103.03 ng / g.

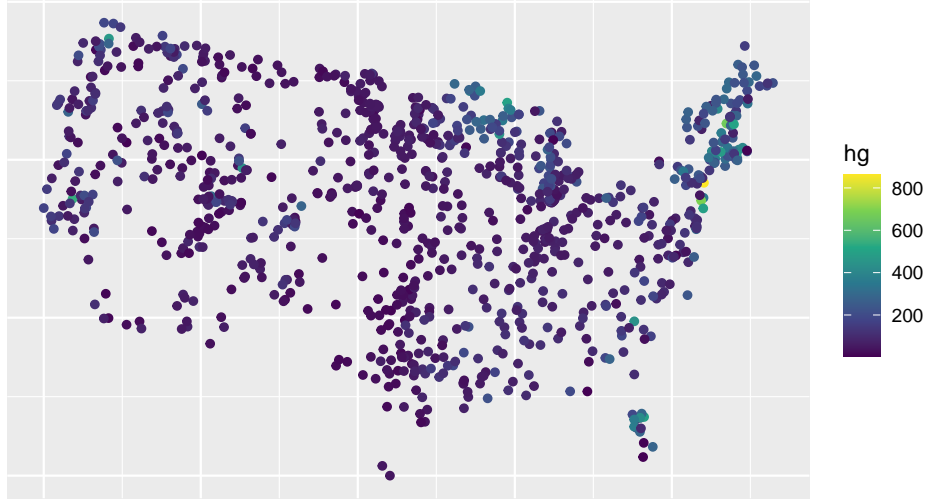


Figure 2: Population distribution of mercury concentration for 986 lakes in the contiguous United States. Thirty-five lakes were dropped from the analysis because they were missing mercury concentration.

Table 1: Table XXX. Application of design-based and model-based approaches to the NLA data set on mercury concentration.

| Approach    | Realized Mean | Estimate | SE  | 95% LB | 95% UB |
|-------------|---------------|----------|-----|--------|--------|
| Design IRS  | 103.2         | 112.7    | 8.8 | 95.4   | 129.9  |
| Model IRS   | 103.2         | 110.5    | 7.9 | 95.0   | 125.9  |
| Design GRTS | 103.2         | 101.8    | 6.1 | 89.8   | 113.7  |
| Model GRTS  | 103.2         | 102.3    | 5.9 | 90.8   | 113.9  |

Table 1 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?

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

361 Robertson, B., McDonald, T., Price, C., Brown, J., 2018. Halton iterative  
362 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 Jr, D.L., Olsen, A.R., 2003. Variance estimation for spatially balanced samples of environmental resources. Environmetrics 14, 593–610.

Stevens Jr, 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.