

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

Michael Dumelle^{*,a}, Matthew 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 Math, 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.

It consists of two paragraphs.

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

1. Introduction

finish spatial introduction revise section 2 reread brus and wang write down potential data sets

Please leave comments in your color: Michael, Matt, Lisa, Tony, Jay.

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 (known as 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 made 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).

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 variables measured at specific geographic locations. De Gruijter and Ter Braak (1990) give an early comparison of design-based and

*Corresponding Author

Email addresses: Dumelle.Michael@epa.gov (Michael Dumelle), mhigham@stlaw.edu (Matthew Higham)

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 location in sampling to model-based approaches (Brus and De Gruijter, 1997; Ver Hoef, 2002, 2008). 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 location in sampling 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 properties of parameter estimates from both approaches. And in Section 4, 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

2.1.1. Design-Based Philosophy

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.

2.1.2. Model-Based Philosophy

The model-based approach assumes the data are a random realization of a process. Randomness is often incorporated through distributional assumptions on the data-generating process. Instead of estimating fixed but unknown parameters (as in the design-based approach), the goal of model-based inference 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 the true mean of the spatial process that generated our realized data is still not known. When predicting the realized mean, we typically are not interested in the underlying process' true mean.

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 can be more precise. Cressie (1993) and Schabenberger and Gotway (2017) provide reviews of model-based approaches for spatial data.

Several comparisons between the model-based paradigm and the design-based paradigm have been made in many different contexts. Sterba (2009) give some history of the two paradigms as well as some applications in psychological contexts. Cooper (2006) review the two approaches in an ecological context before introducing a “model-assisted” variance estimator that combines aspects from each approach. We note that while 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, CITE some more of these), our goal is not to expand upon or improve these methods.

Figure 1a. Brus (2020): Data is fixed. In a finite population example, show a 3d surface that can be generated by anything. If we repeatedly sample the surface, then 95% of all 95% CIs will contain the true mean, which never changes.

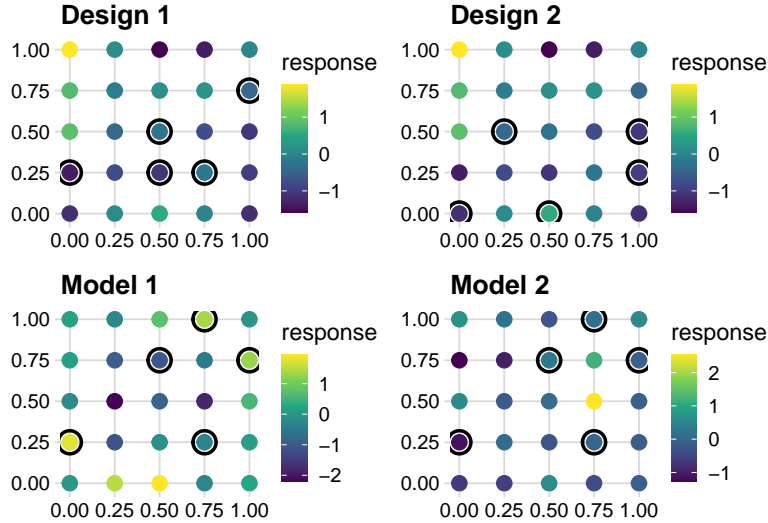


Figure 1b. Spatial process is fixed. In a finite population example, show 10 3d surfaces that are generated from some model. If we repeatedly generate the

surface and obtain a sample, then 95% of all 95% PIs will contain the realized means. The realized mean changes from surface to surface and it's not necessarily the case that 95% of all 95% PIs will contain the true, underlying mean.

2.2. Spatially Balanced Design and Analysis

Sampling designs can incorporate spatial locations to obtain samples that are spatially balanced with respect to the population (Stevens Jr and Olsen, 2004). A sample is spatially 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 population units mirrors the density of all population units. Spatial balance with respect to the population is different than spatial balance with respect to geography. A sample that is spatially balanced with respect to geography is spread out in some type of equidistant manner over geographical space and are not meant to be miniatures of the population. When we refer to spatial balance henceforth, we mean spatial balance with respect to the population.

Spatially balanced samples are useful because they tend to yield estimates that have lower variance than estimates constructed from sampling designs lacking spatial balance (Barabesi and Franceschi, 2011; Benedetti et al., 2017; Grafström and Lundström, 2013; Robertson et al., 2013; Stevens Jr and Olsen, 2004; Wang et al., 2013). To quantify spatial balance, Stevens Jr and Olsen (2004) proposed loss functions based on Voroni polygons. The first spatially balanced sampling algorithm that saw widespread use was the Generalized Random Tessellation Stratified (Stevens Jr and Olsen, 2004). Since GRTS 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 (Benedetti and Piersimoni, 2017), and Halton Iterative Partitioning (Robertson et al., 2018) algorithms. We focus on the Generalized Random Tessellation Stratified (GRTS) algorithm to select spatially balanced sampling because the algorithm has several attractive properties detailed by Stevens Jr and Olsen (2004) and Dumelle et al. (2021).

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

After collecting a sample, the data are used to estimate population parameters. The Horvitz-Thompson estimator (Horvitz and Thompson, 1952) and its continuous analog (Cordy, 1993) yield unbiased estimates of population means

and totals. 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. 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, which 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 π_{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 variability associated with estimating τ . This yields confidence intervals for τ that are narrower than confidence intervals constructed from variance estimators ignoring spatial locations.

2.3. Finite Population Block Kriging

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.

Finite Population Block Kriging (FPBK) is an alternative to sampling-based methods (Ver Hoef, 2008). FPBK expands the geostatistical kriging framework to the finite population setting. Instead of basing inference off of a specific sampling design, we assume the data were generated by a spatial process with parameters that can be estimated using the framework of a model.

Ver Hoef (2008) gives details on the theory of FPBK, but some of the basic principles are summarized below. For a response variable \mathbf{z} that can be measured on a finite number of N sites, we want to predict some linear function of the response variable, $\tau(\mathbf{z}) = \mathbf{b}'\mathbf{z}$, where \mathbf{b} is a vector of weights. For example, if we want to predict the total abundance across all sites, then we would use a vector of 1's for the weights.

Typically, however, we only have a sample of the N sites. Denoting quantities that are part of the sampled sites with a subscript s and quantities that are part of the unsampled sites with a subscript u ,

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

where \mathbf{X}_s and \mathbf{X}_u are the design matrices for the sampled and unsampled sites, respectively, and $\boldsymbol{\delta}_s$ and $\boldsymbol{\delta}_u$ are random errors for the sampled and unsampled sites. Denoting $\boldsymbol{\delta} \equiv [\boldsymbol{\delta}_s \ \boldsymbol{\delta}_u]'$, we assume that $E(\boldsymbol{\delta}) = \mathbf{0}$.

We also typically assume that there is spatial correlation in $\boldsymbol{\delta}$, which can be modeled using a covariance function. Many common choices for this function assume that spatial covariance decreases with increasing Euclidean distance between sites. The primary function used throughout the simulations and applications of this manuscript is the Exponential covariance function: the i, j^{th} entry for $\text{var}(\boldsymbol{\delta})$ 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 sites 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 . However, any spatial covariance function could be used in the place of the Exponential, including functions that allow for anisotropy [pg. 80 - 93](Chiles and Delfiner, 1999).

With the above model formulation, the Best Linear Unbiased Predictor (BLUP) for $\tau(\mathbf{b}'\mathbf{z})$ $\tau(\mathbf{b}'\mathbf{z})$ is vague} 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. Neither require a particular distribution for \mathbf{z} .

3. Numerical Study

Approach	Bias	RMSE	MedAE	Coverage	PClose	MedIL
Design	0.0003	0.0353	0.0251	0.9461	0.4889	0.1362
Model	-0.0001	0.0362	0.0253	0.9480	0.5111	0.1430

Table 1: Approach, mean bias (Bias), root-mean-squared error (RMSE), median absolute error (MedAE), 95 percent interval coverage (Coverage), proportion of times the approach was closer to the true value (PClose), and median interval length (MedIL)

Base Simulations

- both good: correctly specified model with high correlation
- break model: non-gaussian errors
- break design: small area estimation
- both good?: misspecified covariance model with high correlation
- break both? non-gaussian areas with smaller sample size
- change n or sampling fraction
- model-based: how should sample be drawn? should locations be fixed?

3.1. Software

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).

3.2. Applied Example

4. Discussion

References

- 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.
- Cordy, C.B., 1993. An extension of the horvitz—thompson theorem to point sampling from a continuous universe. *Statistics & Probability Letters* 18, 353–362.
- 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.
- Dumelle, M., Olsen, A.R., Kincaid, T., Weber, M., 2021. Selecting and analyzing spatial probability samples in r using `spsurvey`. *Manuscript Submitted for Publication*.
- Fix, E., Hodges, J.L., 1951. Discriminatory analysis, nonparametric discrimination: Consistency properties. *USAF School of Aviation Medicine*.

- 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., Bryce, F., 2020. 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 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.