

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

Michael Dumelle<sup>\*,a</sup>, Matthew Higham<sup>\*,b</sup>, Lisa Madsen<sup>c</sup>, Anthony R. Olsen<sup>a</sup>,  
Jay M. Ver Hoef<sup>d</sup>

<sup>a</sup>United States Environmental Protection Agency, 200 SW 35th St, Corvallis, Oregon, 97333

<sup>b</sup>Saint Lawrence University Department of Math, Computer Science, and Statistics, 23  
Romoda Drive, Canton, New York, 13617

<sup>c</sup>Oregon State University Department of Statistics, 239 Weniger Hall, Corvallis, Oregon,  
97331

<sup>d</sup>Marine 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

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

There are two approaches to statistical inference for spatial data: design-based and model-based. In the design-based approach, statistical inference is based on random sampling. In the model-based approach, statistical inference is based on a set of distributional assumptions for the data.

In either approach, statistical inference requires estimating parameters of interest. Often, these parameters are means or totals. In the design-based approach, parameters correspond to some population. In the model-based approach, parameters correspond to some statistical process, sometimes called a superpopulation.

Spatial data paragraph

define sampling design, define parameters, define model-based, sample, population, super-population?, etc.

introduction of spatial data / definition of spatial and maybe general background?

give some examples of possible spatial data

2-3 sentences about each design

---

\*Corresponding Author

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

goals and outline of paper

De Gruijter and Ter Braak (1990) give an early comparison of the paradigms for spatial data, quashing the belief that classical sampling-based methods could not be used for spatially correlated data. Thereafter, several comparisons between model-based and design-based for spatial data have been considered, but they tend to compare spatial models to IRS (Independent Random Sampling) or SIRS (stratified IRS) (Brus and De Gruijter, 1997). No comparison has been made between spatial models and spatially balanced design-based approaches.

brus2020statistical: most similar to what this paper will be, directly comparing the two approaches. This has a lot of other potentially useful references to go back to. (published in 2021 so likely doesn't have many that have cited it yet).

- hansen1983evaluation - general overview (in non spatial context), talk about population vs superpopulation give a set of guiding principles in the discussion
- verhoef2002sampling - strictly defines classical sampling, compares FPK to classical, BK to classical
- wang2012spatial - sampling and inference steps, thorough overview in terms of spatial context, lots of references

In Section 2

## 2. Background

### 2.1. Design-Based Philosophy

#### Design-Based Overview

The design-based approach to inference uses characteristics of the sampling design to estimate parameters of interest. The design-based approach assumes the data,  $\mathbf{Z}$ , are fixed. Randomness is incorporated through the random selection of sampling units based on probabilities derived using the sampling design itself. Some examples of commonly used sampling designs include independent random sampling (IRS), stratified random sampling, and cluster sampling. Viewing the data as fixed and incorporating randomness through the sampling design yields estimators having very few assumptions. Asymptotic properties of these types of estimators are derived using limiting arguments. Means and totals, for example, are asymptotically Gaussian. Särndal et al. (2003) and Lohr (2009) provide thorough reviews of the design-based approach.

### 2.2. Model-Based Philosophy

#### Model-Based Overview

On the other hand, model-based inference imposes additional assumptions on the data with a potential to provide more precise estimators if the additional assumptions hold. Instead of estimating true but unknown parameters, the goal

of model-based inference in the spatial context is often *prediction* of an unknown quantity. This is a fundamental philosophical difference between sampling-based and model-based approaches. 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 sites, we would have an exact prediction for the mean of our one realized spatial surface, without any uncertainty. But, the true mean of the spatial process that generated our realized data is still not known, and, in the prediction context, we typically do not care much about what value the mean of the underlying process takes.

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.

Figure 1a. 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.

### 2.3. Comparing Design-Based vs. Model-Based

Several comparisons between the two paradigms 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.

### 2.4. Spatially Balanced Design and Analysis

Spatially balanced sampling algorithms use spatial information to obtain samples spread out in space. Spatially balanced samples are useful because they tend to yield estimators that are more precise than estimators constructed from a sampling algorithm that is not spatially balanced (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 statistics based on Voroni polygons. Many spatially balanced sampling algorithms exist, including the Generalized Random Tessellation Stratified (Stevens Jr and Olsen, 2004), 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. Here we focus on the Generalized Random Tessellation Stratified (GRTS) algorithm, which has several attractive properties that we discuss next.

The GRTS algorithm is used to sample from finite and infinite sample frames. A finite sample frame contains a finite number of sampling units and is related to a point geometry. An infinite sample frame contains an infinite number of sampling units and is related to linear and polygon geometries. Examples of point, linear, and polygon resources include lake centroids, stream networks, and wetland areas, respectively. In addition to its applicability for finite and infinite sample frames, the GRTS algorithm naturally accommodates stratified designs and designs with unequal selection probabilities. The algorithm has also been used to select replacement sites using reverse hierarchical ordering (Stevens Jr and Olsen, 2004). Replacement sites are used to replace sites in the original sample that cannot be sampled, often as a result of physical difficulty in reaching the site or landowner denial of access to the sites. More recently, the GRTS algorithm also accommodates legacy (historical) sites, minimum distance between sites, and nearest neighbor replacement sites. The GRTS algorithm is implemented in the **R** package `spsurvey` (Dumelle et al., 2021).

The GRTS algorithm works by mapping two-dimensional space into one-dimensional space while incorporating some elements of randomness. First a square bounding box is placed over the sample frame and divided into four equally sized squares randomly given an level-one address of 0, 1, 2, or 3. Then the inclusion probabilities of the sampling units in each square are summed. If this sum is at least one in any of the cells, a set of four sub-squares are placed within each square. Each set of four sub-squares are randomly given a second address (a level-two address) of 0, 1, 2, 3. Then the inclusion probabilities in each sub-square are summed, and if any of these sums are at least one, a third level of squares are added in the same manner. This hierarchical addressing continues until the inclusion probabilities in the smallest squares are all less than 1. The squares are then mapped to the one-dimensional line in order of their addresses. The length of this line equals  $n$ , the desired sample size (the sum of the inclusion probabilities). A uniform random variable, denoted by  $u_1$  is simulated in  $[0, 1]$  and placed on this line. The location of  $u_1$  on the line can be linked to a sampling unit, denoted by  $s_1$ . Then  $u_2 \equiv u_1 + 1$ , which can be linked to another sampling unit, denoted by  $s_2$ . Because the GRTS algorithm requires the inclusion probabilities of the smallest sub-cells to be less than one,  $s_1$  and  $s_2$  must be distinct. This process continues until the  $n$  sampling units have been selected as part of the sample. Further details are provided by (Stevens Jr and Olsen, 2004) and (Dumelle et al., 2021).

The Horvitz-Thompson estimator (Horvitz and Thompson, 1952) and its continuous analog (Cordy, 1993) yield unbiased estimates of population totals (and means). If  $\tau$  is a population total, then the Horvitz-Thompson estimator of

$\tau$ , denoted by  $\hat{\tau}_{HT}$ , is given by

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

where  $Z_i$  is the observed value of the  $i$ th sampling unit, and  $\pi_i$  is the inclusion probability of the  $i$ th sampling unit. Estimating the variance of  $\hat{\tau}_{HT}$  directly relies on knowing the probability that  $s_i$  and  $s_j$  are both included in the sample (second-order inclusion probability), which can be unfeasible to compute. Furthermore, this variance estimator does not incorporate spatial locations. To address these challenges, (Stevens Jr and Olsen, 2003) proposed an alternative variance estimator: the local neighborhood variance estimator. The local neighborhood variance estimator does not require second-order inclusion probabilities. It also incorporates spatial locations, which tends to reduce the estimated variance of compared  $\hat{\tau}_{HT}$  to a variance estimator ignoring spatial location (Stevens Jr and Olsen, 2003).

## 2.5. Finite Population Block Kriging

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  $u$ ,

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

where  $\mathbf{X}_s$  and  $\mathbf{X}_u$  are the design matrices for the sampled and unsampled 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  $\theta_1$ , the range  $\theta_2$ , and the nugget  $\theta_3$ . 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  $\tau(\mathbf{b}'\mathbf{z})$  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

cases

#### 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?

#### 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).

### 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.
- 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.
- 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.
- 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.
- 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.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.