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

Michael Dumelle[*,a], Matthew Higham[*,b], Lisa Madsen[c], Anthony R. Olsen[a], Jay M. Ver Hoef[d]

[a] *United States Environmental Protection Agency, 200 SW 35th St, Corvallis, Oregon, 97333*
[b] *Saint Lawrence University Department of Math, Computer Science, and Statistics, 23 Romoda Drive, Canton, New York, 13617*
[c] *Oregon State University Department of Statistics, 239 Weniger Hall, Corvallis, Oregon, 97331*
[d] *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.

2. Background

2.1. Design-Based Philosphy

Design-Based Overview

Design-based inference uses characteristics of the sampling design to to estimate parameters of interest Typically, there are few assumptions involved because intervals are derived using the sampling design itself……….

2.2. Model-Based Philosphy

Model-Based Overview

On the other hand, model-based inference imposes additional assumptions on the data with a potential to provide more precise estimates 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

[*] Corresponding Author

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

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.

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

Design-Based and Model-Based Comparisons

There have been many comparisons between the two paradigms.......

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 an 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)). 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 frames 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).

accomodates stratification, unequal selection probabilities, oversample sites It has also recently been updated to accommodate legacy (historical) sites, minimum distance between sites, and a nearest neighbor replacement sites comp efficient software

2.5. Finite Population Block Kriging

Finite Population Block Kriging (FPBK) is an alternative to samipling-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,

$$\begin{pmatrix} \mathbf{z}_s \\ \mathbf{z}_u \end{pmatrix} = \begin{pmatrix} \mathbf{X}_s \\ \mathbf{X}_u \end{pmatrix} \boldsymbol{\beta} + \begin{pmatrix} \boldsymbol{\delta}_s \\ \boldsymbol{\delta}_u \end{pmatrix}, \tag{1}$$

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 var($\boldsymbol{\delta}$) is

$$\text{cov}(\delta_i, \delta_j) = \theta_3 + \theta_1 \exp(-h_{i,j}/\theta_2), \tag{2}$$

$$\sigma^2[(1 - v) \exp(-3\mathbf{h}_{i,j}/\phi) + v\mathbb{1}\{\mathbf{h}_{i,j} = 0\}], \tag{3}$$

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 $\tau(\mathbf{b}'\mathbf{z})$ Did you mean to give the form of the BLUP here? $\tau(\mathbf{b}'\mathbf{z})$ 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

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.

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

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

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.

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.

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.

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.