A Comparison of Design-Based and Model-Based Approaches for Spatial Data

(in alphabetical order) Michael Dumelle and Matthew Higham and Others

Overview

This paper would be a comparison of design-based and model-based frameworks for spatial data with an emphasis on pragmatic applications, like how to choose among the frameworks and how to use them to solve applied problems using open-source software.

This paper could be organized as follows. In Section 1., we outline the design-based and model-based frameworks and then we describe each framework in the context of spatial data. In Section 2., we thoroughly describe and contextualize the assumptions, goals, benefits, and drawbacks of each approach. In Section 3., we compare the two methods by comparing means and variances of spatial data in several simulation scenarios as well as in real data. We also provide reliable software through the use of `spsurvey` and `sptotal`. In Section 4., we end with a discussion and provide directions for future research.

Why This Paper?

We believe the distinction between these approaches is often misunderstood, and there are several ways we could enhance the literature surrounding this topic:

1. Spatial design-based vs spatial model-based: There are no comparisons in the literature between spatial model-based approaches and *spatially balanced* design approaches. These comparisons are typically between spatial model-based approaches and independent or stratified random sample designs. While important to study this behavior, this comparison is no longer fair nor modern. Spatially balanced sampling has exploded in popularity throughout the last decade, and the design-based vs model-based literature needs to reflect this trend.

2. A "fair" comparison: We feel that literature in this area has considered scenarios that are more well-suited for the model-based scenario. For example, if you simulate a Gaussian error with an exponential covariance and then then compare design-based estimates to model-based estimates assuming an exponential covariance, of course the

model-based approach will outperform design-based approaches. These types of comparisons in the literature find model-based approaches generally yield more precise variance estimates. A challenge lies in creating a comparison scenario that is reasonable and intuitive. One thing to consider would be exploring the comparison after estimating a misspecified covariance function using model-based approaches.

3. Finite AND infinite populations: Literature in the area focuses specifically on finite populations or infinite populations; we want to discuss both, with a focus on finite populations.

4. Pragmatic Focus: We see papers in this area tend to be fairly technical. We want the focus to be less on mathematical details and more on discussing the pragmatic questions practitioners will be faced with. For example, a thorough discussion of benefits and drawbacks of each method written for practitioners is warranted.

5. Provide reliable software and code: Much of the current literature presents the ideas but does not provide adequate access to user-friendly software and code. We would clearly explain how to use `spsurvey` to design and analyze spatial probability samples and `sptotal` for FBPK. We would make a GitHub repository / package for the paper walking through code used and containing all manuscript-related documents. Because our code for the simulation study would be easily accessible, practitioners would be able to conduct a small simulation study with parameters that reflect their particular context to compare the two approaches.

Initial Literature

- Design-Based Overview (Särndal et al., 2003; Lohr, 2009)
- Model-Based Overview (Cressie, 2015; Schabenberger and Gotway, 2017)

- Design-Based and Model-Based Comparisons (Hansen et al., 1983; Brus and De Gruijter, 1997; Ver Hoef, 2002; Cooper, 2006; Sterba, 2009; Brus, 2020; Chan-Golston et al., 2020)
- Spatially Balanced Design and Analysis (Stevens Jr and Olsen, 2003, 2004)
- Finite Population Block Kriging (Ver Hoef, 2002, 2008; Higham et al., 2020)

Potential Journals

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

OUTLINE

1. INTRODUCTION

A brief introduction introducing the design-based and model-based frameworks and outlining the rest of the paper.

2. BACKGROUND

A thorough dive into the assumptions, goals, benefits, and drawbacks of the approaches as they relate to *spatial* data. A special focus will be given to pragmatic considerations; we

want this paper to be especially useful for practitioners.

3. NUMERICAL ANALYSIS

We plan to focus on finite populations but also want to accommodate infinite populations.

3.1 Simulation-Based

We would like to keep this section manageable but useful and realistic. Perhaps we start with the following examples:

- simulate via correct model (model outperforms sampling)
- simulate via slightly misspecified model (model still outperforms sampling)
- simulate via very misspecified model (sampling outperforms model)
- simulate via extremely misspecified model (e.g. counts with lots of zeroes and a lot of overdispersion) (neither does well)

We may consider providing an 'R' script that would let practitioners adjust parameters in simulations freely, requiring only that the practitioner know basic 'R' syntax.

3.2 Data-Based

- Use real data set(s) for which we know population values

3.3 Software

`spsurvey` (>= 5.0; we are wrapping up a major update now to 5.0) and `sptotal` (>= 0.1.0)

4. DISCUSSION

We plan to wrap up loose ends, provide directions for future research, and offer any closing thoughts.

References

Brus, D. and De Gruijter, J. (1997). Random sampling or geostatistical modelling? choosing between design-based and model-based sampling strategies for soil (with discussion). *Geoderma*, 80(1-2):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., and Handcock, M. S. (2020). Bayesian inference for finite populations under spatial process settings. *Environmetrics*, 31(3):e2606.

Cooper, C. (2006). Sampling and variance estimation on continuous domains. *Environmetrics: The official journal of the International Environmetrics Society*, 17(6):539–553.

Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons.

Hansen, M. H., Madow, W. G., and Tepping, B. J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78(384):776–793.

Higham, M., Ver Hoef, J., Madsen, L., and Aderman, A. (2020). Adjusting a finite population block kriging estimator for imperfect detection. *Environmetrics*, page e2654.

Lohr, S. L. (2009). *Sampling: design and analysis*. Nelson Education.

Särndal, C.-E., Swensson, B., and Wretman, J. (2003). *Model assisted survey sampling*. Springer Science & Business Media.

Schabenberger, O. and Gotway, C. A. (2017). *Statistical methods for spatial data analysis*. CRC press.

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(6):711–740.

Stevens Jr, D. L. and Olsen, A. R. (2003). Variance estimation for spatially balanced samples of environmental resources. *Environmetrics*, 14(6):593–610.

Stevens Jr, D. L. and Olsen, A. R. (2004). Spatially balanced sampling of natural resources.

Journal of the american Statistical association, 99(465):262–278.

Ver Hoef, J. (2002). Sampling and geostatistics for spatial data. *Ecoscience*, 9(2):152–161.

Ver Hoef, J. M. (2008). Spatial methods for plot-based sampling of wildlife populations. *Environmental and Ecological Statistics*, 15(1):3–13.