The dramatic effect of preferential sampling of spatial data on variance estimates [1]

David Clifford, Petra Kuhnert, Melissa Dobbie, Jeff Baldock, Neil McKenzie, Bronwyn Harch
CSIRO, David.Clifford@csiro.au

Ichsani Wheeler, Alex McBratney
University of Sydney

Abstract: Classic probability-based designs are widely used for spatial sampling in environmental research. When sampling over large regions researchers may wish to preferentially sample some sites due to ease of access. If such non-standard probability designs are implemented, Horvitz-Thompson analysis provides unbiased estimates for spatial means and variances provided first and second order inclusion probabilities can be evaluated. However, even with minor departures from standard designs the effect of preferential sampling on the sampling variance can be dramatic. We find significant increases in sampling variance as sampling becomes more and more preferential. We conclude that some non-standard designs can result in significantly weaker sampling performance and recommend they be examined by simulation prior to implementation.

Keywords: Probability design, GRTS, Horvitz-Thompson Estimators

1. Introduction

There are two broad categories of approaches available for surveying soil organic carbon (SOC) across space - model-based approaches and design-based approaches. The former set of approaches is very useful for mapping and prediction but is based on strong assumptions on the distributional properties of SOC. The latter is based entirely on how sites are chosen for sampling and can produce unbiased estimates of mean SOC as well as unbiased estimates of sampling variance. We examine design-based approaches here as they have been garnering more and more attention in recent years.

Probability-based designs have not been implemented on a national scale within Australia. A major challenge to establishing a national monitoring scheme is the large distances one would need to travel to collect data. For many regional sampling schemes there is anecdotal evidence that sample sites tend to be “just inside the gate, along the fence 50m from the road” which indicates a preference for sites that are easy to access. This is a defining feature of what we term the Australian context and this feature can bias the results of an otherwise well designed experiment when the true manner in which sites are chosen is not incorporated into the analysis.

In this report we explore designs that are compatible with the Australian context, i.e. designs that preferentially sample sites that are easy to access over remote sites. Using classic statistical design methodology coupled with modern computer simulation

[1] Financial support for this research comes from the CSIRO Sustainable Agriculture Flagship.

strategies we explore the effects of such preferential sampling on sampling variance. While stratified sampling generally improves sampling variance relative to simple random sampling, preferential sampling negates this benefit. However, we find that the modern technique of generalised random tessellation stratification (GRTS) sampling can incorporate preferential sampling quite well. In all our examples preferential sampling leads to increases in sampling variance but for GRTS this increase is not fatal

2. Preferential Sampling Probability Designs

We compare the performance of simple random sampling (SRS), stratified random sampling (STR) and GRTS using two spatial datasets. Cochran (1977) provides a detailed summary of many classic sampling designs and analysis results including the work of Horvitz and Thompson (1952) for computing unbiased estimators of mean and sampling variance using first and second order inclusion probabilities. GRTS was developed for sampling streams and stream networks (Stevens and Olsen 2003) and can readily handle any set of first order inclusion probabilities. GRTS has been used extensively in the U.S. by the Environmental Protection Agency for water-based monitoring (e.g. Schweiger et al 2005, Wardrop et al 2007) and can also be used for monitoring natural resources in terrestrial applications (Fancy, Gross and Carter 2009) though to the best of our knowledge it has not been used for soil carbon monitoring.

We venture away from classic designs by specifying inclusion probabilities in a manner that preferentially samples sites that are closer to roads that span the space of interest. We parameterise a linear relationship between inclusion probability and distance to road using a single term α that ranges from 0 to 1. When $\alpha = 0$ the linear relationship is flat, i.e. all inclusion probabilities are equal and we have classical non-preferential sampling. When $\alpha=1$ the inclusion probabilities for the sites furthest from the roads are zero. This boundary case is not considered since a design-based approach is no longer applicable to the whole region of interest. For values of α between 0 and 1 the inclusion probabilities decrease with distance, and the rate of decrease increases with α.

We use the work of Hartley and Rao (1962) to sample a specific number of sites according to our pre-specified first order inclusion probabilities as well as for computing approximations for our second order inclusion probabilities, simplified further by Stehman and Overton (1994). These can be used to compute Horvitz-Thompson estimators from implementations of non-standard designs.

3. Data

We use two spatial datasets to evaluate these probability designs. The first is a simulated non-stationary, non-isotropic process from fixed rank kriging of a spatial random effects model (Cressie & Johannesson, 2008). Values for this process are evaluated at 4 million pixels and we draw samples of size n=27. A grid of nine square strata is used for STR for this dataset. The second is a dataset of over 2.5 million predictions of percentage SOC across a large part (150,000 squared-km) of New South Wales in Australia (Wheeler et al, 2010). These predictions come from a Cubist-based data-mining model of legacy %SOC data from the Australian Soil Resource Information System (ASRIS, McKenzie et al 2005). We draw samples of size n=150 from the SOC dataset. We define 16 strata for this dataset based around the major towns of the region with each site allocated to the stratum associated with the closest town.

4. Methods

We repeatedly apply the probability designs to our two datasets changing the strength of preferential sampling through the parameter α. For each design and α value we examine the distribution of our estimates of sampling variance. Effective sample sizes are found by matching the median sampling variance with sampling variance estimates based on non-preferential SRS. As is well known, when sampling a spatial process, switching from SRS to STR or GRTS leads to an immediate large jump in effective sample size. We wish to investigate what happens to sampling variance and effective sample size as α changes from 0 to just under 1 for each design.

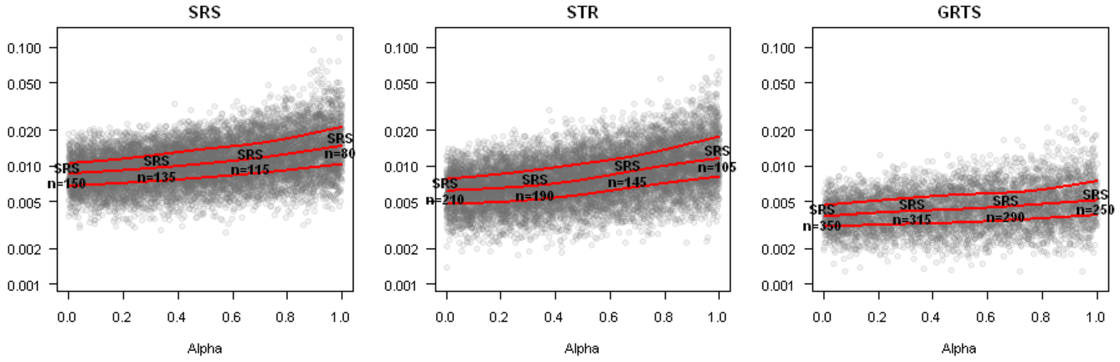

Figure 2: Effect of preferential sampling on the sampling variance of mean estimates for the SOC dataset under SRS, STR and GRTS designs. Red lines within each plot indicate 1st, 2nd and 3rd quartiles of the variance estimates. The text indicates selected *approximate* effective sample sizes under non preferential SRS designs based on smooth quantile regression.

5. Results and Discussion

Preferential sampling results in larger sampling variances in all cases. The gains in effective sample size one attains by switching to STR can be all but wiped out when preferential sampling is employed. Preferential sampling of n=27 sites under STR is routinely found to be worse than SRS based on far fewer sites. For GRTS the effect of preferential sampling is not as dramatic. Figure 2 plots our estimates of sampling variance for many values of α for each design for the SOC dataset. Each panel includes red lines based on smooth robust regression of the data to estimate the 1st, 2nd and 3rd quartiles as functions of α. The text written over each plot indicates effective sample sizes required to achieve similar sampling variances under non-preferential SRS.

This research indicates that continental-scale sampling schemes can be designed and implemented in a manner that better reflects how they are used in practice. While preferential sampling designs more accurately reflect practical concerns, we demonstrate that they can have dramatic inflationary effects on sampling variance. As such, we recommend a thorough evaluation of any sampling approach prior to implementation. In the examples explored here we found that estimates of sampling variance from GRTS are least affected by preferential sampling. This suggests that GRTS is a viable approach for designing spatial sampling schemes at large scales. The success of GRTS is due partly to its use of a neighbourhood variance estimator (Stevens and Olsen 2004) and partly to the fact that GRTS achieves much better spatial balance compared to STR.

References

Cochran, W. G. (1977) *Sampling Techniques*, 3rd Edition Wiley

Cressie, N. & Johannesson, G. (2008) Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B*, 70, 209-226

Fancy, S.; Gross, J. & Carter, S. (2009) Monitoring the condition of natural resources in US national parks *Environmental Monitoring and Assessment*, 151, 161-174

Hartley, H. O. & Rao, J. N. K. (1962) Sampling with Unequal Probabilities and without Replacement *The Annals of Mathematical Statistics,* 33, 350-374

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

McKenzie, N.; Jacquier, D.; Maschmedt, D.; Griffin, E. & Brough, D. (2005) Australian Soil Resource Information System (ASRIS) *National Committee on Soil and Terrain Information*

Schweiger, E.; Bolgrien, D.; Angradi, T. & Kelly, J. (2005) Environmental Monitoring and Assessment of a Great River Ecosystem: The Upper Missouri River Pilot *Environmental Monitoring and Assessment*, 103, 21-40

Stehman, S. V. & Overton, W. S. (1994) Comparison of Variance Estimators of the Horvitz-Thompson Estimator for Randomized Variable Probability Systematic Sampling *Journal of the American Statistical Association,* 89, 30-43

Stevens, D. L. & Olsen, A. R. (2003) Variance estimation for spatially balanced samples of environmental resources *Environmetrics*, 14, 593-610

Stevens, D. L. J. & Olsen, A. R. (2004) Spatially Balanced Sampling of Natural Resources *Journal of the American Statistical Association*, 99, 262-278

Wardrop, D.; Kentula, M.; Stevens, D.; Jensen, S. & Brooks, R. (2007) Assessment of wetland condition: An example from the Upper Juniata watershed in Pennsylvania, USA *Wetlands*, 27, 416-431

Wheeler, I.; Minasny, B.; McBratney, A. & Bui, E. (2010) A regional soil organic carbon prediction function for south-eastern Australia *19th World Congress of Soil Science*