SYSTEMATIC SAMPLING IN ECOLOGY

by

Matthew David McDonald Pawley

Supervisor: **Brian McArdle**

A thesis

submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy,

The University of Auckland, Departments of Statistics and Biology

March 2006.

Abstract

Random start systematic sampling (*SYS*) is a survey design that is simple (it selects the whole sample with one random start), easy to implement and that can, in theory, give precise estimates of ecological abundance in the presence of positive spatial autocorrelation. However, *SYS* suffers from a serious defect, namely, that it is not possible to obtain an unbiased estimator of sampling variance (θ_{SYS}) on the basis of a single sample. A variety of approximations have been suggested, and unbiased model-based methods have been calculated, but validation of these estimators has been limited within the ecological literature.

The heart of any spatial problem is how to deal with spatial autocorrelation. We show that the scale of spatial inference gives a framework that unifies the commonly reported, discordant views about autocorrelation (i.e. 'autocorrelation increases the power of the analysis' vs 'autocorrelation decreases the power of the analysis'). The scale of spatial inference is rarely discussed or considered, but we suggest that it should be the first step in any (spatial) analysis.

The thesis then uses computer simulation to compare the performance of eleven previously proposed *SYS* estimators (including simple random sampling, $\hat{\theta}_{SRS}$). The computer simulations are designed to recreate the spatial distribution characteristics that are common within ecological abundances. We also develop and test a novel method of estimating θ_{SYS} based on 'Krige's Additivity Relationship' and variography (geostatistics). This estimator was labelled $\hat{\theta}_{KAR}$.

We found that if the right spatial model (i.e. a reference theoretical variogram) is used, then $\hat{\theta}_{KAR}$ appears to be an unbiased estimator of θ. Without *a priori* knowledge about the spatial structure (so the theoretical variogram is constructed solely from *SYS* data), $\hat{\theta}_{KAR}$ was generally one of the least biased and most stable estimators out of those examined. The other estimator that fared well, $\hat{\theta}_{r1}$, was also model-based; it used an estimate of the first order autocorrelation in its estimate of θ. $\hat{\theta}_{SRS}$ performed comparatively well on untransformed ecological simulations, but was the worst performing estimator after a log(x+1) transformation.

Preface

This thesis has its genesis in discussions between me and my supervisor, Brian McArdle, about the nature of spatial autocorrelation in ecology. In earlier fieldwork, I had used systematic sampling extensively and we debated about the best manner in which to analyze the results. Brian suggested that I examine the geostatistical method of Block Kriging (*BK*) as a viable method to estimate θ_{SYS}.

I discovered certain theoretical problems with using *BK*. These problems were further compounded by the fact that after considerable research in the matter other authors independently published a field example (in the journal *Ecology*) of what I had been working on, i.e. using *BK* to estimate of θ_{SYS}.

However, during my foray into geostatistics I had stumbled across 'Krige's Additivity Relationship' (*KAR*). *KAR* basically states that the variance of a 'spatial block' as it is moved around an area equals the variance of (smaller) units within the area minus the average variance of those units within the block. It struck me that as long as this 'spatial block' could systematically cover the entire area, it didn't matter if the block consisted of disjoint units, i.e. a systematic sample could be considered a 'block'. In theory then *KAR* could be used to estimate θ_{SYS}.

Initial investigations failed dismally because I thought the elements of *KAR* (the total variance and the average within-sample variance) might be well approximated using the variogram sill and s^2. Unfortunately the use of either approximation made for an extremely unstable estimator. However after noticing the average γ-value in the

variogram cloud equalled s^2 I felt that using some sort of 'expectation of the variogram cloud' might give a more stable estimate of the within sample variance. Using the theoretical variogram seemed like a solution, and after some further fine-tuning (e.g. inclusion of 'zero distances' in the variogram calculations to make it unbiased), $\hat{\theta}_{KAR}$ was established.

I wish to thank the Departments of Statistics and Biology for their financial support, forbearance and guidance. I'd also like to my supervisor, Brian McArdle for his trenchant comments, guidance, leeway and (most of all) patience.

I also wish to thank my colleagues and friends: Carl Donovan, Monique Mackenzie, Cameron Walker, little Michael O'Sullivan and Giant Matt. Their discussions and arguments in the office and at coffee stimulated a variety of research.

Finally I'd like to thank my partner Catherine and my daughter Sofia. You two may have delayed arrival at the destination, but you certainly enhanced the journey – and that's the important thing.

Contents

List of Tables

List of Figures

Basic Notation

There is no standard notation in texts on sampling theory.

With respect to sampling, I will use a notation similar to that of Thompson (1992). Geostatistical notation may slightly differ to sampling notation (see Table II).

Table I : Sampling Notation

Symbol	**Meaning**
N	The finite population size
n	The sample size
μ	The finite population mean
μ_P	The process mean
A	The survey extent, a defined area (region) of interest
a	Sample unit of particular 'support' (unit size/orientation)
α	A 'block' (of space). Note: With respect to the use of *KAR* with *SYS*, 'block' refers to the systematic sample
X_i	Sample locations, $X_1, X_2, .. X_n$ within A
$Z(X_i)$	Sample value of a regionalized variable at location, X_i (see also Y_i below)
Y_i	Sample values $Y_i = y_1, y_2 .. y_n$. (Typically at locations $X_1, X_2 .. X_n$)
σ_P^2	The process variance
k	The *SYS* skip interval (distance between neighbouring points on the *SYS* grid). $k = \frac{N}{n}$, an integer that equals the number of possible *SYS* (of size n)
$\bar{y}$ ($\bar{x}$)	The sample mean of a variable Y (or X):

	$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
σ^2	The 'finite population' variance of *Y*: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)^2$
s^2	The sample variance of *Y*: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$
σ_s^2	The expected 'within-sample' variance (of *SYS*) (population parameter): $\sigma^2 = E\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y_k})^2\right]$
$\hat{\sigma}_s^2$	The estimate of σ_s^2 (from a single sample): $\hat{\sigma}_s^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$
θ	The variance of sample means. If a design-based estimator is used then: $\theta = k^{-1}\sum_{m=1}^{k}(\bar{y}_m - \mu)^2$
π_i	The 'inclusion probability' of unit i (the probability that it is sampled)
SRS	Simple Random sampling
SYS	Random start systematic sampling
RandSYS	Stratified random (systematic) sample
KAR	Krige's Additivity Relationship
$\hat{\theta}_{KAR}$	Systematic sample variance estimate using *KAR* and a variogram derived from the systematic sample data
ρ_k	Either: 'Cochran's within-sample correlation co-efficient', or The autocorrelation at distance, *k*.

Table II: Geostatistical Notation

Symbol	Meaning
h (or d)	Euclidean distance
ξ	A regionalized variable (variable distributed over a finite area, A)
x_i	Sample location $x_1, x_2, \ldots, x_n$ in the area, A
$Z(x_i)$	The value of a regionalized variable ξ at position x
γ_h or $\gamma(x_i, x_j)$	The variogram function value (at distance h), $\gamma_h = \frac{1}{2} [Z(x_i) - Z(x_{i+h})]^2$ or $\gamma_h = \frac{1}{2} [Z(x_i) - Z(x_j)]^2$
C (or cov)	The covariance function of the regionalized variable
C_0	The nugget variance
$C_0 + C_1$	The sill of the theoretical variogram
a_r	Range of autocorrelation in the variogram
m	Discretization grid sample size
M	The realization mean
$\hat{m}$	An estimate of the realization mean