



Copyright Statement

The digital copy of this thesis is protected by the Copyright Act 1994 (New Zealand). This thesis may be consulted by you, provided you comply with the provisions of the Act and the following conditions of use:

- Any use you make of these documents or images must be for research or private study purposes only, and you may not make them available to any other person.
- Authors control the copyright of their thesis. You will recognise the author's right to be identified as the author of this thesis, and due acknowledgement will be made to the author where appropriate.
- You will obtain the author's permission before publishing any material from their thesis.

To request permissions please use the Feedback form on our webpage.

<http://researchspace.auckland.ac.nz/feedback>

General copyright and disclaimer

In addition to the above conditions, authors give their consent for the digital copy of their work to be used subject to the conditions specified on the Library

[Thesis Consent Form](#)

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, $\hat{\theta}_{SR}$). 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 ‘Kriging’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 ‘Kriging’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

ABSTRACT	II
PREFACE	IV
LIST OF TABLES.....	X
LIST OF FIGURES.....	XIII
BASIC NOTATION.....	XXII
CHAPTER 1: INTRODUCTION	1
1.1 BACKGROUND.....	1
1.1.1 <i>Scope of thesis</i>	3
CHAPTER 2: SYSTEMATIC SAMPLING AND GEOSTATISTICS: A REVIEW.....	6
2.1 INTRODUCTION.....	6
2.1.1 <i>Spatial autocorrelation</i>	6
2.2 SYSTEMATIC SAMPLING	8
2.2.1 <i>The method of single random start systematic sampling (SYS)</i>	8
2.2.2 <i>The efficiency of SYS</i>	13
2.2.3 <i>The problem with SYS</i>	18
2.3 GEOSTATISTICS	26
2.3.1 <i>Assumptions used in geostatistics</i>	27
2.3.2 <i>Semi-variogram (variogram) models</i>	31
2.3.3 <i>Working with the variogram instead of covariances</i>	36
2.3.4 <i>Choosing the correct theoretical model</i>	38
2.4 THE KRIGING ALGORITHM.....	39
2.4.1 <i>Ordinary Kriging</i>	40
2.4.2 <i>The kriging system of equations</i>	49
2.5 USING GEOSTATISTICS TO ESTIMATE THE MEAN OF AN AREA	56
2.5.1 <i>Kriging the (local) mean</i>	56
2.5.2 <i>Block Kriging</i>	59

CHAPTER 3: SPATIAL AUTOCORRELATION: BANE OR BONUS?.....	67
3.1 INTRODUCTION.....	67
3.2 SPATIAL AUTOCORRELATION	68
3.3 TEMPORAL VS. SPATIAL AUTOCORRELATION	70
3.4 EXPLOITING SPATIAL AUTOCORRELATION.....	77
3.4.1 <i>Design-based methods of exploiting autocorrelation</i>	78
3.4.2 <i>Model-based (geostatistical) methods of exploiting autocorrelation</i>	84
3.4.3 <i>Characterizing autocorrelation</i>	89
3.4.4 <i>Hypothesis testing</i>	90
3.4.5 <i>Regression</i>	92
3.5 DISCUSSION	96
CHAPTER 4: SIMULATING ECOLOGICAL SPATIAL DISPERSION	102
4.1 SPATIAL PATTERNS OF ABUNDANCE FOUND IN ECOLOGY	102
4.1.1 <i>Characteristics common in ecological abundance data</i>	103
4.1.2 <i>The importance of scale</i>	113
4.2 SIMULATION PARAMETER VALUES	115
4.2.2 <i>The simulation algorithm</i>	122
4.3 ESTIMATION OF THE THEORETICAL VARIOGRAM	126
4.3.1 <i>The estimation of θ</i>	126
4.3.2 <i>The choice of estimation model</i>	126
4.3.3 <i>The search neighbourhood distance</i>	127
4.3.4 <i>Fitting criteria (Maximum likelihood method or WLS)</i>	127
4.3.5 <i>likfit vs variofit</i>	128
4.3.6 <i>The discretization grid</i>	130
CHAPTER 5: THE USE OF VARIOGRAPHY IN THE ESTIMATION OF θ_{sys}.....	132
5.1 INTRODUCTION.....	132
5.1.1 <i>Sample support & dispersion variance</i>	133
5.1.2 <i>Krige's Additivity Relationship</i>	134
5.1.3 <i>An example of KAR</i>	136

5.1.4	<i>Variance Components</i>	138
5.1.5	<i>Estimating the average within-sample variance, (σ_s^2)</i>	139
5.1.6	<i>Estimating the total variance (σ^2)</i>	144
5.1.7	<i>Ensuring a positive variance</i>	147
5.2	METHODS.....	149
5.2.1	<i>The stochastic realization (raw and $\log[x+1]$ scale)</i>	149
5.2.2	<i>Reference variograms</i>	151
5.2.3	<i>Systematic sample variograms</i>	154
5.2.4	<i>Estimating the variance of sample means (θ)</i>	154
5.3	RESULTS	157
5.3.1	<i>Raw data: examining the distribution of $(\hat{\theta} - \theta)$</i>	157
5.3.2	<i>Log(x+1) data: examining the distribution of $(\hat{\theta} - \theta)$</i>	175
5.4	DISCUSSION	191
5.4.1	<i>The use of geostatistics with SYS</i>	191
5.4.2	<i>Confusion with regards to SYS and KAR</i>	193
5.4.3	<i>Inference using $\hat{\theta}_{KAR}$</i>	195
5.4.4	<i>Inference using the reference variogram</i>	196
5.4.5	<i>Inference using SYS-based variograms</i>	199
5.5	APPENDIX A: DERIVATION OF KAR	201
5.6	APPENDIX B: THE ADDITIVITY OF VARIANCES.....	202
5.7	APPENDIX C: $E[\Gamma\text{-CLOUD}] = S^2$	204
CHAPTER 6: A COMPARISON OF SYSTEMATIC SAMPLING ESTIMATORS		206
6.1	INTRODUCTION.....	206
6.1.1	<i>θ_{SYS} estimation</i>	206
6.2	METHODS.....	208
6.2.1	<i>Design-based methods of estimating θ</i>	208
6.2.2	<i>Model-based (and pseudo model-based) methods</i>	215
6.2.3	<i>Computer surface simulations</i>	218
6.2.4	<i>Sampling methodology</i>	218
6.3	RESULTS	221

6.3.1	<i>Untransformed data</i>	221
6.3.2	<i>Log(x+1) transformed data</i>	237
6.4	DISCUSSION	249
6.4.1	<i>Recommendations</i>	251
6.4.2	<i>The estimators</i>	252
CHAPTER 7: OBSERVATIONS ABOUT SYSTEMATIC SAMPLING.....		265
7.1	SYS AND INDEPENDENCE	265
7.1.1	<i>Introduction</i>	265
7.1.2	<i>Methods</i>	267
7.1.3	<i>Results</i>	273
7.1.4	<i>Discussion</i>	277
7.2	THE EFFECT OF THE SKIP INTERVAL ON MODEL MISSPECIFICATION	278
7.2.1	<i>Introduction</i>	278
7.2.2	<i>Methods</i>	280
7.2.3	<i>Results</i>	281
7.2.4	<i>Discussion</i>	288
CHAPTER 8: CONCLUSIONS.....		289
8.1	A SUMMARY OF RESULTS	289
8.2	DISCUSSION	293
8.2.1	<i>Inference space</i>	293
8.2.2	<i>Limitations of thesis simulations</i>	293
8.2.3	<i>Variogram Automation</i>	295
8.2.4	<i>Future work</i>	300
BIBLIOGRAPHY		289

List of Tables

Table I : Sampling Notation	xxii
Table II: Geostatistical Notation.....	xxiv
Table 2.1: The k possible systematic samples of $n = 4$, using data from the regionalized variable shown in Figure 2.1. <i>SYS</i> can be thought of as a special case of cluster sampling (consider rows as a cluster), with only one randomly chosen cluster being sampled.....	10
Table 2.2: Two situations are depicted. Situation 1 describes a situation where the values are random. Situation 2 describes the sampling situation where there is order in the data.	16
Table 2.3: This table shows a variety of different variance estimates that can be obtained. The systematic sample variance is given by the expected value of the 3rd column – this is the variance given by the full block kriging equation.	64
Table 3.1: Using <i>SRS</i> ($n = 100$), 10,000 samples were taken from each surface A and B (shown in Figure 3.2). These were used to calculate 95% coverage probabilities for the (i) two surface means and (ii) process mean (surface A and B have the same generating process mean).....	75
Table 3.2: Inference of model-based methods is based on multiple-realizations (J) of the same stochastic process. The complexity of model-based inference means that a number of different variance estimates can be obtained. The variance of three different columns (shown by the shaded squares) refers to the variance of \bar{X} , μ and the block kriging error respectively.....	86
Table 4.1: Parameter of the random field, ζ on the log ($x+1$) transformed scale. (Variable names used in <i>R</i> -code are shown in brackets)	122
Table 5.1: The k possible systematic samples of $n = 4$, using data from the regionalized variable shown in Figure 5.2. <i>SYS</i> can be thought of as a special case of cluster sampling (consider rows as a cluster), with only one randomly chosen cluster being sampled.....	137
Table 5.2: Bias, variance and MSE [$E(\hat{\theta}-\theta)^2$] for the 17 different estimators on the untransformed data. $n = 36$. Estimators are ranked by MSE (best to worst).	170
Table 5.3: Bias, variance and MSE [$E(\hat{\theta}-\theta)^2$] for the 17 different estimators on the untransformed data. $n = 64$. Estimators are ranked by MSE (best to worst).	171
Table 5.4: Bias, variance and MSE [$E(\hat{\theta}-\theta)^2$] for the 17 different estimators on the untransformed data. $n = 100$. Estimators are ranked by MSE (best to worst).....	172

Table 5.5: Bias, variance and MSE $[E(\hat{\theta}-\theta)^2]$ for the 17 different estimators on the untransformed data. $n = 144$. Estimators are ranked by MSE (best to worst).....	173
Table 5.6: Bias, variance and MSE $[E(\hat{\theta}-\theta)^2]$ for the 17 different estimators on the $\log(x+1)$ transformed data, $n = 36$. Estimators are ranked by MSE (best to worst).....	186
Table 5.7: Bias, variance and MSE $[E(\hat{\theta}-\theta)^2]$ for the 17 different estimators on the $\log(x+1)$ transformed data, $n = 64$. Estimators are ranked by MSE (best to worst).....	187
Table 5.8: Bias, variance and MSE $[E(\hat{\theta}-\theta)^2]$ for the 17 different estimators on the $\log(x+1)$ transformed data, $n = 100$. Estimators are ranked by MSE (best to worst).....	188
Table 5.9: Bias, variance and MSE $[E(\hat{\theta}-\theta)^2]$ for the 17 different estimators on the $\log(x+1)$ transformed data, $n = 144$. Estimators are ranked by MSE (best to worst).....	189
Table 5.10: An example population showing all k possible systematic samples of $n = 4$	195
Table 6.1: Coefficients assigned to the differencing window, d_w	214
Table 6.2: Summary statistics for the estimator distribution of differences – untransformed data, $n = 36$. Estimators are sorted by MSE (best to worst).....	228
Table 6.3: Summary statistics for the estimator distribution of differences – untransformed data, $n = 64$. Estimators are sorted by MSE (best to worst).....	229
Table 6.4: Summary statistics for the estimator distribution of differences – untransformed data, $n =$ 100 . Estimators are sorted by MSE (best to worst).....	230
Table 6.5: Summary statistics for the estimator distribution of differences – untransformed data, $n =$ 144 . Estimators are sorted by MSE (best to worst).....	231
Table 6.6: Percentage of times that $\hat{\theta}$ was more accurate (i.e. closer to θ) than <i>SRS</i> (using untransformed data).	235
Table 6.7: Summary statistics for the estimator distribution of differences – $\log(x+1)$ transformed data, $n = 36$. Estimators are sorted by MSE (best to worst).....	242
Table 6.8: Summary statistics for the estimator distribution of differences – $\log(x+1)$ transformed data, $n = 64$. Estimators are sorted by MSE (best to worst).....	243
Table 6.9: Summary statistics for the estimator distribution of differences – $\log(x+1)$ transformed data, $n = 100$. Estimators are sorted by MSE (best to worst).....	244
Table 6.10: Summary statistics for the estimator distribution of differences – $\log(x+1)$ transformed data, $n = 144$. Estimators are sorted by MSE (best to worst).	245

Table 6.11: Percentage of times that $\hat{\theta}$ was more efficient (i.e. closer to the θ) than <i>SRS</i> (using $\log[x+1]$ transformed data)	248
Table 7.1: A list of the θ_{SYS} and θ_{SRS} variances for different sample sizes. The shaded rows have a sample resolution with neighbouring points distant enough to be effectively independent (i.e. the skip interval >24)	273
Table 7.2: Three different surface generating processes (Exponential, Gaussian and Spherical) were used to each create 1000 realizations. θ_{SYS} and θ_{SRS} were calculated for each surface and the average value (of the 1000 realizations) is shown. The third column shows the relative efficiency of θ_{SYS} compared to θ_{SRS} (as a percentage of the variance)	274
Table 7.3: A comparison of <i>SYS</i> and <i>RandSYS</i> using bias and MSE on the untransformed data.	281
Table 7.4: A comparison of <i>SYS</i> and <i>RandSYS</i> using bias and MSE on the $\log(x+1)$ transformed data.	285

List of Figures

Figure 2.1: Hypothetical values for a regionalized variable (e.g. counts of an organism).....	9
Figure 2.2: Representation of an area that can be split into 16 quadrats. A-D show the possible starting points for a systematic sample of size $n = 4$. If A was the (randomly) chosen starting point, then the (shaded) quadrats would represent the sample positions for the 2×2 systematic sample.	10
Figure 2.3: Some common sample schemes ($n = 64$) for a regionalized variable. Sample position is depicted by hollow circles. Using <i>SYS</i> , the placement of all sample points within the strata is determined by the random position of one point. <i>CSYS</i> chooses sample points to lie at the centre of the strata. <i>RandSYS</i> places an independently random data point within each stratum. <i>SRS</i> randomly positions all sample points.	12
Figure 2.4: A 1-D representation of a cyclical population with a regular period. The four <i>SYS</i> (represented by: hollow circles, solid circles, triangles and square points) have sample means that are highly variable. The inefficiency of the sample means is due to the sample spacing being a multiple of the population periodicity.....	13
Figure 2.5: An example of post stratification of a systematic sample. Circles represent <i>SYS</i> sample points, and the dashed lines show a possible stratification of neighbouring pairs (in this instance stratum pairing was oriented ‘down’ rather than ‘across’)......	23
Figure 2.6: A variogram cloud for a random sample ($n = 64$). The rectangles show distance-classes. The width of the rectangle covers the x -range of the smoothing bin, the top of each rectangle is the average γ -value for those points lying in their respective bin.	33
Figure 2.7: A plot showing the experimental variogram data points (using data and distance-bins shown in Figure 2.6). The dashed line shows a theoretical variogram model fit to the bin-smoothed data (an exponential function was used as the theoretical variogram model). The solid lines indicate positions of the sill and range for the model.	35
Figure 2.8: The variation of the semi-variance $[\gamma_h]$ and the covariance plotted by distance, h , for a transition model (a second-order stationary process with a sill).	37
Figure 2.9: A survey extent with random samples (hollow dots), and a location of interest, x_0 (depicted by a black dot).....	40
Figure 2.10: The position of three sample points (x_1, x_2, x_3), and a third estimated point, (x_0).....	44

Figure 2.11: The shape of some negative-definite models (Exponential, Spherical and Gaussian model). All models shown are transitional (i.e. they have a sill)	54
Figure 3.1: An autocorrelated time series, with recorded data randomly sampled in the extent denoted by the dashed box (sample times indicated by vertical lines)	72
Figure 3.2: Two equal-sized surfaces simulated using an isotropic, stationary process. The surfaces were generated using the <i>GaussRF</i> function within the <i>RandomFields</i> library of 'R' (Schlather 2001). The range of autocorrelation relative to the extent of the surface is small in B compared to A.	74
Figure 3.3: A surface comprised of three independently distributed regionalized variables. Considered without stratification, the regionalized variable across the entire extent is autocorrelated.	79
Figure 3.4: Stratification of a surface, and allocating samples proportional to the size of the strata (using <i>SRS</i> within strata). If this data was analyzed as simply completely <i>SRS</i> then $s^2_{\bar{x}_{SRS}} = 0.63$; when taking the strata into account, $s^2_{\bar{x}_{STRAT}} = 0.12$	81
Figure 4.1: A contour plot of a simulated random field (on the log scale) exhibiting the patchy nature of the spherical autocorrelation function.	106
Figure 4.2: The distribution of the regionalized variable values (using the data shown in Figure 4.9)	108
Figure 4.3: Taylor's Power Plot [$\log(\text{mean})$ vs $\log(\text{variance})$] with statistics evaluated using a 5×5 adjacent points with random starting position from each of the 1000 realizations. A regression line through the data has intercept = -0.,14 and slope = 1.54.	111
Figure 4.4: The distribution of realization means, $E(\zeta)$, and realization variances, $var(\zeta)$, for the 1000 stochastic realizations that were generated (used in Chapter six).	117
Figure 4.5: The generated distributions of the least variable realization (out of the 1000 generated realizations).....	118
Figure 4.6: The generated distributions of the most variable realization (out of the 1000 generated realizations).....	119
Figure 4.7: The distribution of simulated counts for the least variable realization and most variable realization.....	120
Figure 4.8: An example of the regionalized variable on the log scale [generated using the parameters outlined in step (1)].....	123

Figure 4.9: The data in Figure 4.8 after steps 2-3 of the algorithm. This is an example of the final regionalized variable (on the untransformed scale).....	124
Figure 4.10: The data shown in Figure 4.9 after a $\log(x+1)$ transformation.....	125
Figure 4.11: A comparison of <i>likfit</i> and <i>variofit</i> parameter estimation within the <i>geoR</i> library. This particular fit was one of the worst for <i>likfit</i>	129
Figure 4.12: The effect of discretization grid size on the estimation of the total variance. The dashed blue line is the total variance of the realization (σ^2). The circles show the estimated total variance using discretization grids of various sizes.....	131
Figure 5.1: Representation of a finite area that can be partitioned into 16 quadrats. The shaded area (depicting a systematic sample within the area) can represent a block (a) in <i>KAR</i>	135
Figure 5.2: Values for a regionalized variable (e.g. counts of an organism).	136
Figure 5.3: A variogram cloud for a random sample ($n = 64$). The rectangles show distance-classes. The width of the rectangle covers the x -range of the smoothing bin, the top of each rectangle is the average γ -value for those points lying in their respective bin.	139
Figure 5.4: The theoretical variogram model (shown by the dashed curve) is based on the variogram cloud shown in Figure 5.3. The circles equate to the binned distance-class averages. The theoretical variogram model fit (a spherical function) is shown by the dashed line.	142
Figure 5.5: The surface realization depicting untransformed organism density. The realization consisted of 120 x 120 (i.e. $N = 14,400$) different points.	149
Figure 5.6: The surface realization depicting organism density after a $\log(x+1)$ transformation. The realization contains 120 x 120 ($N = 14,400$) different points.....	150
Figure 5.7: The reference variogram for the untransformed data. The reference variogram is based on a random sample ($n = 2500$). The solid line represents the automated fit (using maximum likelihood) – a spherical model was subjectively chosen for the theoretical model structure.....	152
Figure 5.8: The reference variogram for the $\log(x+1)$ data. The reference variogram was based on a random sample of $n = 2500$ rather than the complete dataset. The solid line represents the automated fit (using maximum likelihood) – a spherical model was subjectively chosen for the theoretical model structure.....	153
Figure 5.9: The distribution of differences ($\hat{\theta} - \theta$) for the <i>KAR</i> estimator ($\hat{\theta}_{1a}$) and variants (including <i>SRS</i>) using <i>SYS</i> variogram model (untransformed data) when $n = 36$. The dashed horizontal lines	

- indicate where there is no difference between the estimator ($\hat{\theta}$) and the true variance of sample means (θ). Note the change in the scale of the y-axis between the different plots. 157
- Figure 5.10: The distribution of differences ($\hat{\theta} - \theta$) for the *KAR* estimator ($\hat{\theta}_{1a}$) and variants (including *SRS*) using *SYS* variogram model (untransformed data) when $n = 64$. The dashed horizontal lines indicate where there is no difference between the estimator ($\hat{\theta}$) and the true variance of sample means (θ). Note the change in the scale of the y-axis between the different plots. 158
- Figure 5.11: The distribution of differences ($\hat{\theta} - \theta$) for the *KAR* estimator ($\hat{\theta}_{1a}$) and variants (including *SRS*) using *SYS* variogram model (untransformed data) when $n = 100$. The dashed horizontal lines indicate where there is no difference between the estimator ($\hat{\theta}$) and the true variance of sample means (θ). Note the change in the scale of the y-axis between the different plots. 159
- Figure 5.12: The distribution of differences ($\hat{\theta} - \theta$) for the *KAR* estimator ($\hat{\theta}_{1a}$) and variants (including *SRS*) using *SYS* variogram model (untransformed data) when $n = 144$. The dashed horizontal lines indicate where there is no difference between the estimator ($\hat{\theta}$) and the true variance of sample means (θ). Note the change in the scale of the y-axis between the different plots. 160
- Figure 5.13: A comparison of estimator variance (untransformed data) for the *SYS*-based variogram estimators (including *SRS*) plotted on the log scale (y-axis). 161
- Figure 5.14: A comparison of estimator bias (untransformed data) for the *SYS*-based variogram estimators (including *SRS*). The line graph is truncated for some estimators due to the extreme positive bias when $n = 36$. Truncation for these estimators is signified by a hollow circle (\circ). The solid horizontal line represents when $E(\hat{\theta}) = \theta$ 162
- Figure 5.15: A comparison of estimator MSE (untransformed data) for the *SYS*-based variogram estimators (including *SRS*) plotted on the log scale (y-axis). 164
- Figure 5.16: The distribution of differences ($\hat{\theta} - \theta$) for the *KAR* estimator ($\hat{\theta}_{1a}$) and variants (including *SRS*) using the reference variogram model (untransformed data) when $n = 36$. The dashed horizontal lines indicate where there is no difference between the estimator ($\hat{\theta}$) and the true variance of sample means (θ). Note the change in the scale of the y-axis between the different plots. 166
- Figure 5.17: The distribution of differences ($\hat{\theta} - \theta$) for the *KAR* estimator ($\hat{\theta}_{1a}$) and variants (including $\hat{\theta}_{SRS}$) using the reference variogram model (untransformed data) when $n = 64$. The dashed horizontal lines indicate where there is no difference between the estimator ($\hat{\theta}$) and the true

variance of sample means (θ). Note the change in the scale of the y-axis between the different plots.....	167
Figure 5.18: The distribution of differences ($\hat{\theta} - \theta$) for the <i>KAR</i> estimator ($\hat{\theta}_{Ia}$) and variants (including <i>SRS</i>) using the reference variogram model (untransformed data) when $n= 100$. The dashed horizontal lines indicate where there is no difference between the estimator ($\hat{\theta}$) and the true variance of sample means (θ). Note the change in the scale of the y-axis between the different plots.....	168
Figure 5.19: The distribution of differences ($\hat{\theta} - \theta$) for the <i>KAR</i> estimator ($\hat{\theta}_{Ia}$) and variants (including <i>SRS</i>) using the reference variogram model (untransformed data) when $n= 144$. The dashed horizontal lines indicate where there is no difference between the estimator ($\hat{\theta}$) and the true variance of sample means (θ). Note the change in the scale of the y-axis between the different plots.....	169
Figure 5.20: The distribution of differences ($\hat{\theta} - \theta$) for the <i>KAR</i> estimator ($\hat{\theta}_{Ia}$) and variants (including <i>SRS</i>) using <i>SYS</i> variogram models [$\log(x+1)$ transformed data] when $n= 36$. The dashed horizontal lines indicate where there is no difference between the estimator ($\hat{\theta}$) and the true variance of sample means (θ). Note the change in the scale of the y-axis between the different plots.....	175
Figure 5.21: The distribution of differences ($\hat{\theta} - \theta$) for the <i>KAR</i> estimator ($\hat{\theta}_{Ia}$) and variants (including <i>SRS</i>) using <i>SYS</i> variogram models [$\log(x+1)$ transformed data] when $n= 64$. The dashed horizontal lines indicate where there is no difference between the estimator ($\hat{\theta}$) and the true variance of sample means (θ). Note the change in the scale of the y-axis between the different plots.....	176
Figure 5.22: The distribution of differences ($\hat{\theta} - \theta$) for the <i>KAR</i> estimator ($\hat{\theta}_{Ia}$) and variants (including $\hat{\theta}_{SRS}$) using <i>SYS</i> variogram models [$\log(x+1)$ transformed data] when $n= 100$. The dashed horizontal lines indicate where there is no difference between the estimator ($\hat{\theta}$) and the true variance of sample means (θ). Note the change in the scale of the y-axis between the different plots.....	177
Figure 5.23: The distribution of differences ($\hat{\theta} - \theta$) for the <i>KAR</i> estimator ($\hat{\theta}_{Ia}$) and variants (including $\hat{\theta}_{SRS}$) using <i>SYS</i> variogram models [$\log(x+1)$ transformed data] when $n= 144$. The dashed horizontal lines indicate where there is no difference between the estimator ($\hat{\theta}$) and the true	

variance of sample means (θ). Note the change in the scale of the y-axis between the different plots.....	178
Figure 5.24: A comparison of estimator variance [$\log(x+1)$ transformed data] for the <i>SYS</i> -based variogram estimators (including <i>SRS</i>) plotted on the log scale (y-axis).....	179
Figure 5.25: A comparison of estimator bias ($\log[x+1]$ transformed data) for the <i>SYS</i> -based variogram estimators (including <i>SRS</i>). Due to the difference in scales these are presented as separate graphs. The solid horizontal lines on each graph represents where $E(\hat{\theta}) = \theta$	180
Figure 5.26: A comparison of estimator MSE [$\log(x+1)$ transformed data] for the <i>SYS</i> -based variogram estimators (including <i>SRS</i>) plotted on the log scale (y-axis).	181
Figure 5.27: The distribution of differences ($\hat{\theta} - \theta$) for the <i>KAR</i> estimator (<i>Ia</i>) and variants (including <i>SRS</i>) using the reference variogram model [$\log(x+1)$ transformed data] when $n= 36$. The dashed horizontal lines indicate where there is no difference between the estimator ($\hat{\theta}$) and the true variance of sample means (θ). Note the change in the scale of the y-axis between the different plots.....	182
Figure 5.28: The distribution of differences ($\hat{\theta} - \theta$) for the <i>KAR</i> estimator (<i>Ia</i>) and variants (including <i>SRS</i>) using the reference variogram model [$\log(x+1)$ transformed data] when $n= 64$. The dashed horizontal lines indicate where there is no difference between the estimator ($\hat{\theta}$) and the true variance of sample means (θ). Note the change in the scale of the y-axis between the different plots.....	183
Figure 5.29: The distribution of differences ($\hat{\theta} - \theta$) for the <i>KAR</i> estimator (<i>Ia</i>) and variants (including <i>SRS</i>) using the reference variogram model [$\log(x+1)$ transformed data] when $n= 100$. The dashed horizontal lines indicate where there is no difference between the estimator ($\hat{\theta}$) and the true variance of sample means (θ). Note the change in the scale of the y-axis between the different plots.....	184
Figure 5.30: The distribution of differences ($\hat{\theta} - \theta$) for the <i>KAR</i> estimator (<i>Ia</i>) and variants (including <i>SRS</i>) using the reference variogram model [$\log(x+1)$ transformed data] when $n= 144$. The dashed horizontal lines indicate where there is no difference between the estimator ($\hat{\theta}$) and the true variance of sample means (θ). Note the change in the scale of the y-axis between the different plots.....	185

Figure 5.31: The relationship between the (exponential) variogram function and the realization variance (for a finite regionalized variable). The realization variance is depicted as the lower horizontal line. The shaded areas (A and B) are equal because the area under the curve also equates to the realization variance (σ^2). 192

Figure 6.1: An example of the 2×2 overlapping strata used by Millar (1995). The arrows indicate that strata are formed around every possible group of (square) 2×2 systematic sample points. 210

Figure 6.2: An example of post stratification of a systematic sample. Circles represent *SYS* sample points, and the dashed lines show a possible stratification of neighbouring pairs (in the figure the chosen orientation was ‘across’ rather than ‘down’). 211

Figure 6.3: An example of two randomly chosen realizations (using the untransformed data) from the set of 1000 realizations that were generated. 218

Figure 6.4: The two randomly chosen realizations shown in Figure 6.3 (after using a $\log[x+1]$ transformation). 220

Figure 6.5: The distribution of θ across multiple realizations using the untransformed data. 221

Figure 6.6: The distribution of $\log(\hat{\theta}/\theta)$ i.e. $\log(\text{estimated variance of sample means}) / (\text{‘true’ [design-based] variance of sample means})$ when $n = 36$ (on untransformed data). The horizontal dashed line shows when $\hat{\theta} = \theta$ 223

Figure 6.7: The distribution of $\log(\hat{\theta}/\theta)$ i.e. $\log(\text{estimated variance of sample means}) / (\text{‘true’ [design-based] variance of sample means})$ when $n = 64$ (on untransformed data). The horizontal dashed line shows when $\hat{\theta} = \theta$ 224

Figure 6.8: The distribution of $\log(\hat{\theta}/\theta)$ i.e. $\log(\text{estimated variance of sample means}) / (\text{‘true’ [design-based] variance of sample means})$ when $n = 100$ (on untransformed data). The horizontal dashed line shows when $\hat{\theta} = \theta$ 225

Figure 6.9: The distribution of $\log(\hat{\theta}/\theta)$ i.e. $\log(\text{estimated variance of sample means}) / (\text{‘true’ [design-based] variance of sample means})$ when $n = 144$ (on untransformed data). The horizontal dashed line shows when $\hat{\theta} = \theta$ 226

Figure 6.10: Interaction plot of bias for the 11 examined estimators. The horizontal line shows where $E(\hat{\theta}) = \theta$ 232

Figure 6.11: Interaction plot of MSE (shown on the log scale) for the 11 examined estimators. 233

Figure 6.12: The distribution of θ across multiple realizations using the $\log(x+1)$ transformed data. . 237

Figure 6.13: The distribution of $\hat{\theta} - \theta$ when $n = 36$ (on $\log[x+1]$ transformed data). The horizontal dashed line shows when $\hat{\theta} = \theta$	238
Figure 6.14: The distribution of $\hat{\theta} - \theta$ when $n = 64$ (on $\log[x+1]$ transformed data). The horizontal dashed line shows when $\hat{\theta} = \theta$	239
Figure 6.15: The distribution of $\hat{\theta} - \theta$ when $n = 100$ (on $\log[x+1]$ transformed data). The horizontal dashed line shows when $\hat{\theta} = \theta$	240
Figure 6.16: The distribution of $\hat{\theta} - \theta$ when $n = 144$ (on $\log[x+1]$ transformed data). The horizontal dashed line shows when $\hat{\theta} = \theta$	241
Figure 6.17: Interaction plot of bias for the 11 examined estimators on the $\log(x+1)$ transformed data. The horizontal line shows where $E(\hat{\theta}) = \theta$	246
Figure 6.18: Interaction plot of MSE (shown on the log scale) for the 11 examined estimators on the $\log(x+1)$ transformed data.....	247
Figure 6.19: An empirical variogram generated from a sample of $n = 36$ (the maximum-likelihood parameterized variogram model is shown as the solid line). Note the lack of available information for smaller lags - this is shown by the shaded area. This is a consequence of <i>SYS</i> design.....	250
Figure 6.20: Plots of $\hat{\theta}$ for <i>BK</i> and <i>KAR</i> from all possible <i>SYS</i> (size $n = 36$) on a single surface realization. The invariant reference variogram was used for each sample so differences between the 400 estimates were caused solely by sample unit placement. The red horizontal line shows where $\hat{\theta} = \theta$, the dashed line shows the $E(\hat{\theta}_{BK})$ based on all 400 (shown) estimates.....	257
Figure 6.21: A comparison of the total variation of $\hat{\theta}_{BK}$ and $\Delta\hat{\theta}_{BK}$ using the untransformed data in Chapter five.....	260
Figure 6.22: A plot of $\hat{m} - \bar{x}$ for <i>BK</i> from all possible <i>SYS</i> (size $n = 36$) on a single surface realization. The horizontal line at 0 shows where $\hat{m} = \bar{x}$	261
Figure 7.1: The reference variogram derived from the realization. The variogram model was fit by eye.....	267
Figure 7.2: The shape of the three models examined across multiple realizations (section 7.1.3.2). Each generating function had the same parameters, i.e. nugget = 0, sill = 50 and an effective range of 15.....	271

Figure 7.3: The distribution of ratios ($\theta_{SYS}/\theta_{SRS}$) for the Exponential, Gaussian and Spherical models. Values below the dashed, horizontal line (where the ratio = 1) are those realizations where θ_{SYS} is more efficient than θ_{SRS}	275
Figure 7.4: [shown in the previous chapter (Figure 6.19)] - An empirical variogram generated from a sample of $n = 36$ (the maximum-likelihood parameterized variogram model is shown as the solid line). Note the lack of available information for smaller lags - this is shown by the shaded area. This is a consequence of <i>SYS</i> design.....	278
Figure 7.5: Interaction plot of bias comparing <i>SYS</i> - and <i>RandSYS</i> using untransformed data.	282
Figure 7.6: The distribution of estimated nugget effects when using <i>RandSYS</i>	283
Figure 7.7: Interaction plot of MSE (shown on the log scale) comparing <i>SYS</i> and <i>RandSYS</i> using untransformed data.....	284
Figure 7.8: Interaction plot of bias comparing <i>SYS</i> and <i>RandSYS</i> using the $\log(x+1)$ transformed data.	286
Figure 7.9: Interaction plot of MSE (shown on the log scale) comparing <i>SYS</i> and <i>RandSYS</i> using the $(\log(x+1))$ transformed data.....	287

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 <i>KAR</i> with <i>SYS</i> , 'block' refers to the systematic sample
X_i	Sample locations, X_1, X_2, \dots, 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 $Y_i = y_1, y_2 \dots y_n$. (Typically at locations $X_1, X_2 \dots X_n$)
σ_P^2	The process variance
k	The <i>SYS</i> skip interval (distance between neighbouring points on the <i>SYS</i> grid). $k = \frac{N}{n}$, an integer that equals the number of possible <i>SYS</i> (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 <i>SYS</i>) (population parameter): $\sigma_s^2 = E \left[\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y}_k)^2 \right]$
$\hat{\sigma}_s^2$	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)
<i>SRS</i>	Simple Random sampling
<i>SYS</i>	Random start systematic sampling
<i>RandSYS</i>	Stratified random (systematic) sample
<i>KAR</i>	Krige’s Additivity Relationship
$\hat{\theta}_{KAR}$	Systematic sample variance estimate using <i>KAR</i> 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, \dots, 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 <i>cov</i>)	The covariance function of the regionalized variable
C_0	The nugget variance
$C_0 + C_I$	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