1991

Statistical analysis of spatial pattern in ecological data

Jay Michael Ver Hoef
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/rtd
Part of the Biostatistics Commons, and the Ecology and Evolutionary Biology Commons

Recommended Citation

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.
Statistical analysis of spatial pattern in ecological data

Ver Hoef, Jay Michael, Ph.D.
Iowa State University, 1991
Statistical analysis of spatial pattern in ecological data

by

Jay Michael Ver Hoef

A Dissertation Submitted to the
Graduate Faculty in Partial Fulfillment of the
Requirements for the Degree of
DOCTOR OF PHILOSOPHY

Departments:  Botany
               Statistics
Co-majors:   Ecology and Evolutionary Biology
               Statistics

Approved:

Signature was redacted for privacy.

In Charge of Major Work
Signature was redacted for privacy.

Signature was redacted for privacy.

For the Major Departments
Signature was redacted for privacy.

For the Graduate College
Iowa State University
Ames, Iowa
1991
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENERAL INTRODUCTION</td>
<td></td>
</tr>
<tr>
<td>1. EXPLANATION OF DISSERTATION FORMAT</td>
<td>9</td>
</tr>
<tr>
<td>PART I. NESTED ANOVA AND VARIOGRAM ANALYSIS FOR SPATIAL PATTERN</td>
<td>11</td>
</tr>
<tr>
<td>1. INTRODUCTION</td>
<td>14</td>
</tr>
<tr>
<td>2. NESTED ANOVA AND TTLV</td>
<td>17</td>
</tr>
<tr>
<td>3. THE VARIOGRAM</td>
<td>21</td>
</tr>
<tr>
<td>4. RELATIONSHIPS BETWEEN THE VARIOGRAM, NESTED ANOVA AND TTLV</td>
<td>25</td>
</tr>
<tr>
<td>5. COMPARISON OF $2^\alpha_l (1:m)$ AND $2^\alpha_{1:m}$</td>
<td>27</td>
</tr>
<tr>
<td>6. EXAMPLES</td>
<td>29</td>
</tr>
<tr>
<td>7. DISCUSSION AND CONCLUSIONS</td>
<td>32</td>
</tr>
<tr>
<td>8. ACKNOWLEDGMENTS</td>
<td>35</td>
</tr>
<tr>
<td>9. LITERATURE CITED</td>
<td>36</td>
</tr>
<tr>
<td>10. APPENDIX</td>
<td>40</td>
</tr>
<tr>
<td>PART II. STATISTICAL TECHNIQUES FOR AVERAGE PATCH SIZE IN A TRANSECT</td>
<td></td>
</tr>
<tr>
<td>1. INTRODUCTION</td>
<td>53</td>
</tr>
<tr>
<td>2. METHODS</td>
<td>56</td>
</tr>
<tr>
<td>2.1. Two term local variance (TTLV)</td>
<td>56</td>
</tr>
<tr>
<td>2.2. Moving two-sample t-test (MT)</td>
<td>56</td>
</tr>
<tr>
<td>2.3. A Bayesian approach with simulated annealing (BSA)</td>
<td>58</td>
</tr>
<tr>
<td>2.4. A factorial experiment of simulated transects</td>
<td>60</td>
</tr>
<tr>
<td>3. SIMULATION RESULTS</td>
<td>63</td>
</tr>
<tr>
<td>3.1. The 3x2x3 factorial experiment</td>
<td>63</td>
</tr>
<tr>
<td>3.2. Two term local variance</td>
<td>64</td>
</tr>
</tbody>
</table>
4.4. Prediction with variograms and cross-variorams;  
Multivariable kriging  

4.5. Relationship to cokriging one variable at a time  

5. EXAMPLES  

5.1. Optimal linear predictors and prediction intervals  

5.2. Comparison of the cross-variorams $2\gamma(*)$ with the 
quantities $2\nu(*)$  

6. DISCUSSION AND CONCLUSIONS  

7. ACKNOWLEDGMENTS  

8. LITERATURE CITED  

9. APPENDIX  

GENERAL SUMMARY  

LITERATURE CITED  

ACKNOWLEDGMENTS
GENERAL INTRODUCTION

Pattern, its definition, and the consequences of its analysis, have a confused and controversial history in ecology. Greig-Smith (1979) states that "one of the few generalizations that we can make about vegetation is that it is spatially heterogeneous." How this knowledge has contributed to our understanding of organisms, though, is unclear (Hill 1973). For example, McIntosh (1985, p. 237) says, "It is striking that ecosystem studies and population studies have both run into the same phenomenon, pattern, which confounds easy generalization." So, while pattern is one of the few generalizations of ecology, it confounds easy generalization!

Ecologists often define spatial pattern as the nonrandom horizontal spatial abundance of organisms (Greig-Smith 1979, McIntosh 1985, Kershaw and Looney 1985). In ecology, beginning with the work of Watt (1947), the concepts of pattern and process have been closely related. Process is the ecological factor or mechanism that produces a pattern; pattern is the spatial abundance in response to the process.

In statistics, there are also definitions for pattern and process. Let a collection of spatially explicit random variables \( Z(\mathbf{s}) \), where \( \mathbf{s} \) is a vector of spatial coordinates, be contained in a set,

\[
\{Z(\mathbf{s}): \mathbf{s} \in D \subset \mathbb{R}^d\}; \tag{1}
\]

in (1), \( D \) is some subset of \( \mathbb{R}^d \), which is Euclidean space with dimension \( d=1, 2, \) or 3. The set of random variables (1) is the spatial process. Then the set of the realized values \( \{z(\mathbf{s})\} \) of each random variable in \( \{Z(\mathbf{s})\} \) is the spatial
pattern.

The ecological and statistical definitions of process and pattern may be unified by simply assuming that the ecological process is a collection of spatially explicit random variables, as in (1). This unification provides a useful framework for evaluation of some current methods of pattern analysis in ecology, and development new ones. The overall goal of this dissertation is to use (1) to examine estimators of the variogram as a function of aggregation, to examine estimators of average patch size for transect data, and to develop multivariable spatial prediction.

In this dissertation, Part I begins with (1), and then considers a more special case where (1) is a linear model composed of fixed and random effects,

$$Z(s) = \sum_{i=1}^{p_1} \alpha_i \mu_i(s) + \sum_{j=1}^{p_2} \sigma_j \delta_j(s); \ s \in \mathbb{R}^d,$$

where \(\{\mu_i(s); i=1,\ldots,p_1\}\) are fixed or nonrandom effects (parameters) in the mean structure with coefficients \(\{\alpha_i; i=1,\ldots,p_1\}\), and \(\{\delta_j; j=1,\ldots,p_2\}\) are zero mean, unit variance random variables with coefficients \(\{\sigma_j; j=1,\ldots,p_2\}\). An early method of pattern analysis was to assume a model of nested effects in (2) and to perform a nested analysis of variance (nested ANOVA, Greig-Smith 1952). In (2), the nested effects are simply assumed to be the effects due to aggregating the data along a line transect. The usual tests of hypotheses for nested ANOVA in this spatial setting were quickly criticized by Thompson (1955, 1958) and others due to probable lack of independence. Nevertheless, the plotting of mean squared error against the effects of aggregation continued to be popular as a graphical method to explore spatial data.
Nested ANOVA was subject to several modifications, and many other statistical pattern analysis methods have been proposed for ecological data (e.g., Usher 1975, Ripley 1978, Galiano 1983, Dale and MacIsaac 1989, and Orłóci and Orłóci 1990). Comparisons of different methods on artificial and real data sets (e.g., Ludwig 1979, Carpenter and Chaney 1983) recommended two methods. One was a modification of nested ANOVA, namely, two term local variance (TTLV, Hill 1973), which continues to be a popular method (e.g., Dale and MacIsaac 1989, Ver Hoef et al. 1989, Dale and Blundon 1990). Another method, paired quadrat variance (PQV, Ludwig and Goodall 1978), was also recommended. Nested ANOVA and TTLV yield statistics that aggregate or block data from contiguous plots; PQV is based on the distance between plots. Is there some mathematical relationship between the aggregation and distance methods?

The purpose of Part I is to show some mathematical relationships between nested ANOVA, TTLV, and PQV. The geostatistical quantity, the variogram, which captures the spatial dependency among random variables, allows the mathematical relationships to be developed. Equations (1) and (2) provide the theoretical framework for these relationships.

Assessing the performance of a pattern-analysis method requires some criterion on which it is evaluated. Often, data are simulated, and the criterion is the ability of the method to recover information on some known feature of the simulated data. Most often for transect data the feature of interest is average patch size (APS), or "scale of pattern" (e.g., Usher 1975, Ludwig and Goodall 1978, Dale and MacIsaac 1989, Dale and Blundon 1991).
Again, consider a special case of (1),

\[ (Z(s); \ Z(s) = \mu(s) + \delta(s); \ s \in D \subset R^1), \]

where \( D=(1,2,\ldots,n) \). Call \( (\mu(s); \ s=1,\ldots,n) \) the mean structure. Now suppose \( \mu(s)\in(a,b) \) for all \( s \), implying a two-phase mean structure composed of patches, and suppose that the random errors \( (\delta(s)) \) are mutually independent. Then a patch of size \( \ell \) may be defined as a set \( (\mu(s); \ \mu(s)=\mu(s+1)=\ldots=\mu(s+\ell)) \). A change point \( c(s) \) is any location \( s \) along the transect where \( \mu(s)-\mu(s+1)\neq0 \), which indicates crossing from one patch to another. If we denote \( N(c) \) as the number of change points, then average patch size (APS) can be defined as,

\[ APS = \frac{n}{N(c)+1}. \]

Part II evaluates several methods for estimating APS. Although Part I shows that TTLV can be used to estimate the variogram under aggregation, in Part II TTLV is used to estimate APS since a "peak" in TTLV as a function of aggregation should indicate APS (e.g., Dale and MacIsaac 1989). Several new methods, a moving two-sample t-test (MT), and a Bayesian approach with simulated annealing (BSA), are also proposed for estimating APS. These methods are described in Part II.

Of the many simulation studies in the literature, some were designed where the models (3) contained fixed patch sizes without random error (Errington 1973, Usher 1969, Dale and MacIsaac 1989), variable patch sizes without random error (Errington 1973, Carpenter and Chaney 1983), and fixed patch sizes with random error (Usher 1969, 1975). Only a few have combined variable patch size and random error (Ludwig and Goodall 1978, Ludwig 1979). From these studies it
appears that three factors have important effects on the performance of pattern methods: 1) the signal to noise ratio (R factor); 2) the expected size of the patches relative to the plot size (S factor); and 3) the distribution of patch sizes (D factor), which are described in Part II.

One objective of Part II is to examine the effects of R, S, and D on TTLV, MT, and BSA by simulating transects. Good simulation studies should be statistically well-designed, just as any good field experiment (Kleijnen 1987, Sacks et al. 1989). To examine the effects of factors R, S, and D on TTLV, MT, and BSA, a computer-simulation experiment was analyzed using a 3x2x3 factorial design. The second objective of Part II is to demonstrate each of the methods, especially BSA, on real vegetation, since several "nuisance" parameters must first be estimated for BSA.

Parts I and II are concerned with estimation of a parameter, or set of parameters, for models (2) or (3). Part I deals with the estimation of the variogram under aggregation from transect data, and Part II is concerned with the estimation of APS from transect data. In contrast, the topic of Part III is the prediction of pattern for a spatial process.

In sciences such as ecology, it is often desirable to predict variables (such as biomass, species counts, etc.) at unsampled spatial locations, based on data observed at nearby locations. Geostatistical methods, which use the variogram, have had a major role in spatial description, modeling, and prediction for the geological sciences, but they have only recently been adopted by the ecological sciences for analyzing spatial data (e.g., Robertson 1987, Robertson et al. 1988, Legendre and Fortin 1989).
For spatial prediction, it has been usual to predict one variable at a
time, with a predictor based on either the same type of variable (spatial best
linear unbiased prediction or kriging) or using concomitant variables
(cokriging). Kriging, which is the name used in geostatistics for optimal
spatial prediction for a scalar-valued process, was proposed by Matheron (1963).
Similar formulations for optimal spatial prediction arose in several other
fields (e.g., Gandin, 1963, in meteorology), but Matheron's theories and
terminology seem to have the widest use. Kriging uses the variogram for spatial
prediction. Spatial prediction is also possible by using covariances. Part III
reviews optimal spatial prediction, and in particular best linear unbiased
prediction, using variograms and covariances.

Part III begins with a linear model, which is similar to (2) and (3),
except the response is vector-valued, rather than scalar-valued,
\[
(z(s); z(s) = u(s) + \delta(s); s \in D \subset \mathbb{R}^d).
\]
(4)
The term cokriging has been used when predicting only one variable type at some
spatial location for a vector-valued process. One difficulty is that there has
been no general formulation of cokriging in terms of the traditionally-used
cross-variogram,
\[
2 \nu_{ij}(s_k, s_l) = \text{cov}([z_i(s_k) - z_i(s_l)], [z_j(s_k) - z_j(s_l)]),
\]
except under restrictive covariance conditions (Journel and Huijbregts, 1978).
Nevertheless, there are several papers which use \(2 \nu_{ij}((\cdot))\) for cokriging,
apparently assuming that the covariance conditions are satisfied (e.g., Carr and
1988, Mulla 1988, and Hoeksema et al. 1989). A more recent definition of the
cross-variogram (Clark et al. 1989) is
\[ 2\gamma_{ij}(s_i, s_j) = \text{var}[z_i(s_i) - z_j(s_j)]. \]

Part III investigates the question of which is the proper cross-variogram, \( 2\nu_{ij}(\cdot, \cdot) \) or \( 2\gamma_{ij}(\cdot, \cdot) \), for cokriging.

Spatial prediction of several variables simultaneously is also of interest. For example, in ecology, a community is defined as the co-occurrence and abundance of several species at the same spatial locale. Ecologists, then, are concerned with the joint spatial patterns of these species, and it is desirable to predict the joint abundance of species at unsampled spatial locations.

There are several immediate problems when considering multivariable spatial prediction. First, as in cokriging, there is the question of which cross-variogram should be used, \( 2\nu_{ij}(\cdot, \cdot) \) or \( 2\gamma_{ij}(\cdot, \cdot) \)? Second, what is the criterion to be minimized? For scalar prediction, we minimize
\[ E[(p(z; s_0) - z(s_0))^2], \] (5)
where \( p(z; s_0) \) is the predictor, based on all available data \( z \), for \( z_i(s_0) \). Then two candidates for the multivariate problem are to minimize
\[ E[(p(z; s_0) - z(s))^T(p(z; s_0) - z(s))], \] (6)
or to find \( g(z; s_0) \) such that,
\[ E[(g(z; s_0) - z(s))^T[g(z; s_0) - z(s)]], \] (7)
is nonnegative-definite for all \( g(z; s_0) \). Notice that both (6) and (7) reduce to (5) in the univariate case. Which is to be preferred, (6) or (7)? Myers (1982) suggested (6), which was developed further by Quimby et al. (1986). The objectives of Part III are to show that (7) is to be preferred, and that \( 2\gamma_{ij}(\cdot, \cdot) \) is the proper cross-variogram for multivariable spatial prediction. Part III
also develops optimal vector prediction using covariances, and shows some relationships to generalized least squares and prediction of one variable at a time.
1. EXPLANATION OF THE DISSERTATION FORMAT

This dissertation is organized in the alternative format for submission to the Graduate Faculty, with each part a separate paper which will be submitted for publication to a refereed journal in statistics or ecology. Each paper's style follows that of the journal to which it has been submitted, or to which it will be submitted. The dissertation consists of three parts which deal with estimation and prediction for spatial processes, especially for ecological data. In Part I, the variogram under aggregation is shown to be a mathematical link between several pattern statistics proposed in the ecological literature, and estimation of the variogram under aggregation is also considered. In Part II, estimation of average patch size for transect data is considered, where three methods are compared in a computer-simulation experiment and an ecological example is given. Part III considers the simultaneous prediction of several variable types for a vector-valued process, and extends results from scalar prediction of variables.

This dissertation has been completely written by me, I wrote all of the computer programs, and I made all of the figures. The questions and objectives of the first two papers were conceived by me. Dr. Cressie suggested the problem of Part III after we discussed my particular needs and interests for ecological data that I have collected. It is difficult to catalog the exact contributions to results and ideas for each part of the dissertation. In some cases I used established results in new ways on ecological data, in other cases I obtained new results when needed. Several of the earlier results in Part III are due to
Dr. Cressie, and I extended them to the multivariate case. All authors, when listed, played an integral role in the conceptual progress of each of these papers, and many of these ideas evolved through the many discussions I have had with my advisors, Dr. Glenn-Lewin and Dr. Cressie.
PART I.

NESTED ANOVA AND VARIOGRAM ANALYSIS FOR SPATIAL PATTERN
Nested ANOVA and variogram analysis for spatial pattern

authors:

Jay M. Ver Hoef$^{1,2}$
Noel A. C. Cressie$^1$
David C. Glenn-Lewin$^2$

1 Department of Statistics, Iowa State University, Ames, Iowa, 50011
2 Department of Botany, Iowa State University, Ames, Iowa, 50011
ABSTRACT

A general statistical framework is proposed for comparing linear models of spatial process and pattern. A spatial linear model for nested analysis of variance can be based on either fixed effects or random effects. Greig-Smith (1952) originally used a fixed effects model, but there are examples of random effects models in the soil science literature. Assuming intrinsic stationarity for a linear model, the expectations of a spatial nested ANOVA and two term local variance (TTLV, Hill 1973) are functions of the variogram, and several examples are given. Paired quadrat variance (PQV, Ludwig and Goodall 1978) is a variogram estimator which can be used to approximate TTLV, and we provide an example from ecological data. Both nested ANOVA and TTLV can be seen as weighted lag-1 variogram estimators that are functions of support, rather than distance. We show that there are two unbiased estimators for the variogram under aggregation, and computer simulation shows that the estimator with smaller variance depends on the process autocorrelation.

Keywords: pattern analysis, two term local variance, paired quadrat variance, first-order autoregressive model, autocorrelation
1. INTRODUCTION

Ecologists concerned with spatial pattern historically first centered their interest on testing the hypothesis that a species exhibited a complete spatially random pattern. Many indices have been proposed (for a review, see Goodall and West 1979, Cressie 1991). It soon became apparent that very few organisms exhibit complete spatial randomness at all scales, so attention has turned to the description of nonrandom pattern.

Ecologists often define spatial pattern as the nonrandom horizontal spatial abundance of organisms (Greig-Smith 1979, McIntosh 1985, Kershaw and Looney 1985), although Pielou (1977) allows for random pattern. There are several features of pattern. Grain is the scale of pattern in which differences in abundance occur—the size of patches, for instance. Intensity is the extent to which abundance differs over an area (Pielou 1977). Interest in inference about intensity and scale of pattern has generated many statistical methods (e.g., Greig-Smith 1952, Hill 1973, Usher 1975, Ludwig and Goodall 1978, Ripley 1978, Galiano 1983, Dale and MacIsaac 1989, and Orloci and Orloci 1990). Comparisons of some of the earlier methods (Ludwig 1979, Carpenter and Chaney 1983) recommended two term local variance (TTLV, Hill 1973) and paired quadrat variance (PQV, Ludwig and Goodall 1978) for estimating grain and intensity. In particular, TTLV remains one of the most popular methods (e.g., Dale and MacIsaac 1989, Ver Hoef et al. 1989, Dale and Blundon 1990).

In this paper, we show some relationships between nested analysis of variance (nested ANOVA), first used for pattern analysis by Greig-Smith (1952),
TTLV (a modification of nested ANOVA), and PQV. Nested ANOVA and TTLV are statistics that aggregate or block data from contiguous plots; PQV is based on the distance between plots. Is there some mathematical relationship between aggregation and distance based methods? To answer this requires a mathematical framework, so we begin with definitions of process and pattern by placing them in the context of a statistical model. Let the ecological measure of interest (e.g., biomass, abundance, etc.) be a random variable. Then a spatial process is a collection of these random variables, $Z(\mathbf{s})$, indexed by spatial location vectors $\mathbf{s}$ (e.g., $\mathbf{s} = (x,y)'$ coordinates), which we shall denote as,

$$
(Z(\mathbf{s}); \mathbf{s} \in D \subset \mathbb{R}^d);
$$

(1)

here $D$ is some subset of $\mathbb{R}^d$, which is Euclidean space with dimension $d = 1, 2,$ or $3$. The model (1) is the generating mechanism, or process; a real outcome (also called a realization or data) from (1) is the pattern.

The process (1) is very general. More specifically, consider a process $Z(\cdot)$ consisting of a linear model of fixed and random effects,

$$
(Z(\mathbf{s}); Z(\mathbf{s}) = \sum_{i=1}^{p_1} \alpha_i \mu_i(\mathbf{s}) + \sum_{j=1}^{p_2} \sigma_j \delta_j(\mathbf{s}); \mathbf{s} \in \mathbb{R}^d);
$$

(2)

where $\{\mu_i(\mathbf{s}); i=1,\ldots,p_1\}$ are fixed or nonrandom effects (parameters) in the mean structure with coefficients $\{\alpha_i; i=1,\ldots,p_1\}$; these fixed effects are the deterministic, ecological effects that ecologists often associate with the word "process." Randomness is contained in the set $\{\delta_j(\mathbf{s}); j=1,\ldots,p_2\}$, which are zero mean, unit variance random variables with coefficients $\{\sigma_j; j=1,\ldots,p_2\}$. Equation (2) generalizes the model of Morris (1987) by making it spatially explicit, and it allows for more combinations of fixed and random effects.
The model (2) has several distinct advantages over a verbal definition of pattern as, say, nonrandom spatial abundance. First, the model makes the definition and concept of pattern mathematically explicit. Second, as a special case of (2), where,

\[ \sum_{1=1}^{p_1} \alpha_i \mu_i(s) = \mu \quad \text{and} \quad \sum_{j=1}^{p_2} \sigma_j \delta_j(s) = \sigma \delta(s), \]

with \( \delta(*) \) a standard white noise process, pattern may indeed be spatially random. Finally, this model allows for the description of pattern in a positive sense. That is, rather than describing pattern as nonrandomness, it is seen as some combination of mean structure and spatially dependent random error.

The model (2) is the theoretical framework for the comparison of nested ANOVA, TTLV, and PQV. In particular, we show that the expected values of nested ANOVA and TTLV are obtained from the variogram, and PQV is a variogram estimator. Through the variogram, we shall show that the aggregation approach and the distance approach are related. Geostatistical methods, which use the variogram, have had a major role in spatial description, modeling, and prediction for the geological sciences, but they have only recently been adopted by the ecological sciences for analyzing spatial data (e.g., Robertson 1987, Robertson et al. 1988, Legendre and Fortin 1989).
2. NESTED ANOVA AND TTLV

Greig-Smith (1952) initially applied nested ANOVA to data in 2-dimensional space. Kershaw (1957) modified nested ANOVA for line-transect data of contiguous quadrats; we shall consider the latter. Suppose there are \( n = 2^k \) random variables \( Z(s) \) in a line transect consisting of contiguous sample rectangles of equal size (henceforth called quadrats),

\[
[Z(1)\mid Z(2)\mid Z(3)\mid \ldots \mid Z(n)].
\]

It has been more usual in the ecological literature to use subscripts on \( Z \), but since \( Z \) is a function of location, we retain the \( Z(\cdot) \) notation.

A useful alternative notation is: Let \( Y(ab\ldots cd) = Z(s) \), where \( ab\ldots cd \) denotes the binary representation of \( s-1; \ s=1,\ldots,n=2^k \). That is,

\[
a = \begin{cases} 
0 & \text{if } Z \text{ is in } 1^{st} \text{ half of transect} \\
1 & \text{if } Z \text{ is in } 2^{nd} \text{ half of transect} 
\end{cases}
\]

\[
b = \begin{cases} 
0 & \text{if } Z \text{ is in } 1^{st} \text{ half of } a^{th} \text{ half} \\
1 & \text{if } Z \text{ is in } 2^{nd} \text{ half of } a^{th} \text{ half} 
\end{cases}
\]

etc.

For example, take \( n=8 \) and \( Z(1), Z(2), \ldots, Z(8) \) in a transect of contiguous quadrats. Then a nested spatial structure denoted \( BS(\ell) \), \( \ell=1,2,4,8 \), can be imposed with the equivalent \( Y(\cdot) \) notation,

<table>
<thead>
<tr>
<th>( BS(8) )</th>
<th>( Y(AAA) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( BS(4) )</td>
<td>( Y(O0A) )</td>
</tr>
<tr>
<td>( BS(2) )</td>
<td>( Y(000) )</td>
</tr>
<tr>
<td>( BS(1) )</td>
<td>( Z(1) )</td>
</tr>
</tbody>
</table>
where \( \overline{Y}(aA\ldots A) \) is the average over all \( Y \) values for which \( a \) is given; \( \overline{Y}(abA\ldots A) \) is the average over all \( Y \) values for some \( ab \) combination, etc. Under the \( Y(\cdot) \) notation, calculation of nested ANOVA is given in Table 1.

In plant ecology, Greig-Smith (1952) originally made use of nested ANOVA for spatial data. In terms of the linear model (2), Greig-Smith’s analysis was based on the following model:

\[
Y(ab\ldots cd) = \mu + \mu_{k-1} (a) + \mu_{k-2} (ab) + \ldots + \mu_2 (ab\ldots c) + \sigma \delta (ab\ldots cd).
\]  

(3)

Each \( \mu_k \) is a fixed effect in the mean structure due to the block size \( 2^k \), and \( \delta_1 \) is random error. Table 1 may be used in performing an F-test of the equality of mean effects in model (3) with the following assumptions: All \( \delta_1 \) are independent Gaussian random variables with unit variance. Due to the probable violation of these assumptions, Greig-Smith’s original usage was quickly criticized by Thompson (1955, 1958) and others.

A purely random effects model could also be chosen, where the model, in terms of (2), now becomes,

\[
Y(ab\ldots cd) = \mu + \sigma \delta_1 (a) + \sigma \delta_2 (ab) + \ldots + \sigma \delta (ab\ldots cd).
\]  

(4)

In (4), \( \mu_0 \) is some overall mean and each \( \delta_\ell \) is a random effect due to the block size indicated by its subscript \( \ell \). The expectations for mean squares in Table 1, based on model (4), are given in Table 2.

Here it is assumed that all random variables \( \{\delta_\ell\} \) are independent with zero mean...
and unit variance. Under this model, Table 2 can be used to partition total variance by solving for each spatially nested random effect, $\sigma^2_e$, as has been done in several soil studies (Youden and Mehlich 1937, Webster and Butler 1976, Nortcliff 1978).

However, due to possible correlation between nearby quadrats, nested ANOVA is likely to be inappropriate for testing hypotheses for contiguous quadrats. Despite this limitation, it was quickly adopted as a graphical description of vegetation, based on the following statistic:

$$MSBS(2^r) = \frac{1}{2(2^r)} \left\{ \sum_{j=2^r}^{j+2^{r+1}} \sum_{l=2^{r+1}}^{j+2^{r+1}} Z(i) - \sum_{l=j+2^{r+1}}^{j+2^{r+1}} Z(i) \right\}^2$$

r=0,1,...,k-1; where $MSBS(2^r)$ denotes mean square at block size $2^r$ and $\text{ave}[]$ denotes the average of the terms within the brackets. This is a re-expression (in the Z(•) notation) of each mean square due to each source (block size) in a spatially nested ANOVA (Table 1), and is established in result A1 in the appendix.

The practice of using nested ANOVA for testing was quickly dropped in favor of plotting the block size mean square $MSBS(2^r)$ against block size $2^r$ or against $r$, and looking for a "peak". The location of the peak is reported to indicate the "grain", or scale of pattern, while its height indicates "intensity" (Ludwig and Goodall 1978).

As a description of pattern, there still appears to be difficulties with the nested ANOVA approach; for a review see Goodall (1974) and Pielou (1977).
One problem is that the geometrical progression of $z^r$ in (5) causes larger and larger gaps where pattern may occur. Secondly, results depend on the starting position relative to the pattern (for examples, see Usher 1969 and Errington 1973).

Hill (1973) noticed that, in (5), choosing a fixed starting position $j$ omitted many similar terms, so he proposed the following two term local variance statistic:

$$
TTLV(m) = \frac{1}{2m} \text{ave} \left( \left[ \sum_{i=1}^{j+m-1} Z(i) - \sum_{i=j+m}^{j+2m-1} Z(i) \right]^2 : j=1,2,...,n-2m+1 \right); m=\left[ \frac{n}{2} \right].
$$

(Recall that $[x]$ denotes the integer part of $x$.) This statistic is interpreted in the same way as nested ANOVA; i.e., one is looking for a "peak" when plotting TTLV$(m)$ against $m$. 

20
3. THE VARIOGRAM

Return to the definition of process and pattern in (1). In the case of a univariate distribution, a statistical distribution is characterized by parameters, and it is often the estimation of these parameters that concerns the ecologist or statistician. In the case of many random variables in a process such as (1), parameters specifying the dependence among the random variables are typically needed (e.g., all pairwise correlations). Another set of parameters that captures the spatial dependence is given by the variogram.

The variogram \(2\gamma(\cdot, \cdot)\) is defined as,

\[
2\gamma(u, y) = \text{var}(Z(u) - Z(y)),
\]

for all \(u, y \in D\), where \(D \in \mathbb{R}^d\) and \(\text{var}(\cdot)\) denotes the variance. It is assumed that \(\text{var}(Z(s)) < \infty\) for all \(s \in \mathbb{R}^d\). The process \(Z(\cdot)\) is said to be intrinsically stationary if it satisfies,

(i) \(E[Z(s)] = \mu\), for all \(s \in D\), and,

(ii) \(\text{Var}[Z(u) - Z(y)] = 2\gamma(u - y)\), for all \(u, y \in D\);

where \(E(\cdot)\) denotes expectation. Now, \(2\gamma(h)\), where \(h = u - y\), is a function only of the direction and distance separating any two points \(u\) and \(y\) in \(D\). Also notice that \(2\gamma(0) = 0\). The variogram is called isotropic if \(\text{Var}[Z(u) - Z(y)] = 2\gamma(\|u - y\|)\) for all \(u, y \in D\); that is, \(2\gamma(\|h\|)\) now has domain in \(\mathbb{R}^1\) and is a function only of the distance separating any two points in \(D\). If only (i) holds, the process is called mean stationary.

Under intrinsic stationarity, the linear model (2) can be written as,

\[
Z(s) = \mu + \delta(s),
\]

(7)
where $\mu$ is some overall mean, and $\delta(\cdot)$ has mean zero and variogram $2\gamma(\cdot)$.

(Independence of $Z$'s is included as a special case.)

Probably more familiar to ecologists are the autocovariance and autocorrelation functions (e.g., Sokal and Oden 1978, Cliff and Ord 1981), which are also used extensively in time series. Define the autocovariance function $C(\tau, \tau) = \text{cov}(Z(\tau), Z(\tau))$, where $\text{cov}(\cdot, \cdot)$ denotes the covariance. If $C(\tau, \tau) \neq 0$, then we do not have independence between data at sites $\tau$ and $\nu$. Along with mean stationarity, assuming $C(\tau, \tau) = C(\tau - \tau) = C(h)$ for all $\tau$ and $\nu$ is termed second-order stationarity. The autocorrelation function then is $C(h)/C(0); h \in \mathbb{R}^d$. Notice that, in terms of variances and covariances,

$$2\gamma(\tau, \nu) = \text{var}(Z(\tau) - Z(\nu)) = \text{var}(Z(\tau)) + \text{var}(Z(\nu)) - 2\text{cov}(Z(\tau), Z(\nu)).$$

Thus, the relationship between the variogram and autocovariance function is

$$2\gamma(\tau, \nu) = C(\tau, \nu) + C(\nu, \tau) - 2C(\tau, \nu).$$

Assuming second-order stationarity, we obtain

$$\gamma(h) = \sigma^2 - C(0),$$

where $\sigma^2 = C(0)$.

Now consider the case of a belt transect of contiguous quadrats. Usually, the variogram is defined for $Z(\cdot)$ with point support, but quadrats have nonzero length. Thus, the distance between any two quadrats is ambiguous. We can, however, adopt the following conventions.

Assume that any two adjacent $Z(s)$ and $Z(s+1)$ are 1 unit apart, $Z(s)$ and $Z(s+2)$ are 2 units apart, and, in general, $Z(s)$ and $Z(t)$ are $|s-t|$ units apart. Intrinsic stationarity for the $Z(\cdot)$ random variables over the discretized regions (contiguous quadrats) then is,

(i) $E[Z(s)] = \mu; s = 1, 2, \ldots$ ;

(ii) $\text{var}[Z(s) - Z(t)] = 2\gamma(|s-t|); s = 1, 2, \ldots n ; t = 1, 2, \ldots n.$
Notice that, when defined for quadrats rather than for continuous space, $2\gamma^b(h)$, where $h=|i-j|$, has the nonnegative integers as its domain.

Now we wish to aggregate the $Z(*)$ by averaging them into groups of size $m$. Let

$$Z(s;m) = \frac{1}{m} \sum_{i=s}^{s+m-1} Z(i),$$

where $s$ is the starting position. Then, assuming intrinsic stationarity, the variogram under aggregation is also defined; it is,

$$\text{var}[Z(s;m)-Z(t;m)] = 2\gamma^a(s,t;m) = \mathbb{E} \left[ \frac{1}{m} \sum_{i=s}^{s+m-1} Z(i) - \frac{1}{m} \sum_{l=t}^{t+m-1} Z(l) \right]^2. \quad (8)$$

Then, for the case of the transect of contiguous quadrats that we are considering, the following expression of $2\gamma^a(s,t;m)$ in terms of $2\gamma^b$ can be easily derived:

$$2\gamma^a(s,t;m) = \frac{1}{m^2} \sum_{i=s}^{s+m-1} \sum_{j=s}^{s+m-1} \gamma^b(|i-j|) - \frac{1}{m^2} \sum_{l=t}^{t+m-1} \sum_{j=t}^{t+m-1} \gamma^b(|i-j|) + \frac{2}{m^2} \sum_{i=s}^{s+m-1} \sum_{j=t}^{t+m-1} \gamma^b(|i-j|). \quad (9)$$

Now, after aggregation to size $m$, define the lag between $Z(s+hm;m)$ and $Z(s;m)$ as $h$. That is, we are emphasizing that no matter what the scale of aggregation $m$, two adjacent $Z(s+m;m)$ and $Z(s;m)$ are one unit apart, and the size of that unit depends on $m$. Assuming intrinsic stationarity, from (9) we can express the variogram $2\gamma^a(s,t;m)$ as both a function of aggregation $m$ and distance $h$ conditional on the level of aggregation: Define $2\gamma^a(h;m) = 2\gamma^a(s,s+hm;m)$, and, from (9), we obtain,
2\gamma^a(h;m) = \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{m} 2\gamma^b(|i-j|) + \frac{1}{m} \sum_{i=1}^{m} \sum_{j=h+1}^{m+1} 2\gamma^b(|i-j|),

where \( m \) is the number of adjacent quadrats which are averaged, and \( h \) is the distance (in units of length \( m \)) after aggregation. After some algebra, this can be simplified to,

\[
2\gamma^a(h;m) = \frac{2\gamma^b(hm)}{m} + \frac{2}{m^2} \sum_{i=1}^{m-1} (i-m)2\gamma^b(i)+(m-i)[\gamma^b(hm+i)+\gamma^b(hm-i)].
\]
4. RELATIONSHIPS BETWEEN THE VARIOGRAM, NESTED ANOVA AND TTLV

Now that a spatially discrete version of the variogram is available for any aggregation level and distance, it is possible to derive the expected values of nested ANOVA and TTLV in terms of the variogram. Recall the terms being averaged in (5):

\[
\left\{ \begin{array}{c}
\sum_{i=1}^{j+2^r} Z(i) - \sum_{i=1}^{j+2^{r+1}} Z(i) \\
\int_{1=1}^{j+2^r} Z(i) - \int_{1=1}^{j+2^{r+1}} Z(i)
\end{array} \right\}^2
\]

But, these are just like terms in (8), from which (10) was derived. So,

\[
E[\text{MSBS}(2^r)] = \frac{2^r}{2} \langle 2y^2(1;2^r) \rangle
\]

is obtained under intrinsic stationarity. Likewise,

\[
E[\text{TTLV}(m)] = \frac{m}{2} \langle 2y^2(1;m) \rangle.
\]

In fact, from (8) it can be seen that an unbiased method of moments estimator of (10) is,

\[
2 \hat{y}^2(h;m) = \frac{1}{m^2} \text{ave} \left[ \sum_{i=s}^{s+m-1} Z(i) - \sum_{i=s}^{s+(h+1)m-1} Z(i) : 1 \leq s \leq n-(h+1)m+1 \right]^2
\]

Then we see that Hill’s (1973) statistic (6) has the following relationship to (13):

\[
\text{TTLV}(m) = \frac{m}{2} \langle 2y^2(1;m) \rangle,
\]

which is a weighted lag-1 method of moments estimator of the aggregated variogram expressed as a function of support, or aggregation.
Matheron (1963) proposed an estimator for the variogram $2\gamma^b(h)$ which, for the transect case, is,

$$2\gamma^b(h) = \frac{1}{n-h} \sum_{i=1}^{n-h} (Z(i) - Z(i+h))^2.$$  \hspace{1cm} (15)

In 1974, Goodall introduced randomly-paired quadrats variance, which was subsequently modified to all-paired quadrats variance PQV (Ludwig and Goodall 1978). But PQV is exactly the variogram estimator in (15); i.e.,

$$\text{PQV}(h) = 2\gamma^b(h).$$

Now, replace $2\gamma^b(h)$ in (10) with (15) to obtain,

$$2\gamma^a(1;m) = \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} 2\gamma^b(|i-j|) + \frac{1}{2} \sum_{i=1}^{m} \sum_{j=m+1}^{2m} 2\gamma^b(|i-j|).$$ \hspace{1cm} (16)

Thus, $2\gamma^a(1;m)$ given by (13) and $2\gamma^a(1;m)$ given by (16) both estimate $2\gamma^a(1;m)$.

Since $\text{TTLV}(m) = \frac{m}{2} 2\gamma^a(1;m)$, it is natural to consider an approximation to TTLV based on (16), namely, $\frac{m}{2} 2\gamma^a(1;m)$. 
5. COMPARISON OF $\tilde{\sigma}^2(l;m)$ and $\tilde{\gamma}^2(l;m)$

Due to the difficulties of computing the exact variances, a simulation experiment was conducted to examine the efficiency of the two estimators of $\tilde{\sigma}^2(l;m)$, namely, $\tilde{\gamma}^2(l;m)$ in (13) and $\tilde{\gamma}^2(l;m)$ in (16). The efficiency is defined as,

$$\text{eff}(m,\rho) = \frac{\text{var}[\tilde{\gamma}^2(l;m)]}{\text{var}[\tilde{\gamma}^2(l;m)]} ; \quad \rho = -0.9, ..., 0.9.$$

Suppose we have a set of normally-distributed random variables $Z(i)$; $i=1,2,...,40$, where $i$ denotes transect location. From the following first-order autoregressive model: $E[Z(i)]=0$, $\text{var}[Z(i)]=1$, and $\text{cov}[Z(i),Z(j)]=\rho^{|i-j|}$; $i,j=1,2,...,40$, 1000 independent transect realizations were generated on a computer for each $\rho$ value; $\rho = -0.9, -0.8, ..., 0.8, 0.9$. Both $\tilde{\gamma}^2(l;m)$ and $\tilde{\gamma}^2(l;m)$ were calculated for each level of aggregation, $m=1,2,...,20$, for each transect. Both estimators are unbiased for $\tilde{\sigma}^2(l;m)$. The variance of each statistic for each $m,\rho$ combination was estimated by taking the average (over the 1000 transects) squared difference between the calculated value and the true value.

Fig. 1 shows the estimated efficiency for all $m,\rho$ combinations. It can be seen that, over positive values of $\rho$, the variance of $\tilde{\gamma}^2(l;m)$ is as much as 1.5 times the variance of $\tilde{\gamma}^2(l;m)$, with the efficiency of $\tilde{\gamma}^2(l;m)$ increasing with $m$. However, over negative values of $\rho$, the variance of $\tilde{\gamma}^2(l;m)$ is as much as
10 times the variance of \( 2\hat{\gamma}^2(l;m) \), and \( 2\hat{\gamma}^\oplus(l;m) \) also does relatively better than \( 2\hat{\gamma}^\odot(l;m) \) with increasing \( m \).
6. EXAMPLES

Consider the example of spatial independence. In this case the variogram looks like Fig. 2a; \( \sigma^2 = 5 \) was arbitrarily chosen as the variance of all the \( Z(i) \).

Figure 2

(The variance is one-half the "sill," the constant height that the variogram attains.) Under this variogram model, \( 2\gamma^s(l;m) \) in (10) and the expectation of \( \text{TTLV}(m) \) in (12) can be seen in Fig. 2b. This shows that the weighting factor \( (m/2) \) keeps \( \text{TTLV}(m) \) invariant to aggregation under the model of spatial independence.

Several common variogram models are often fit to actual data, a few of which are shown here; for the mathematical formulae and fuller discussion, see Journel and Huijbregts (1978). Fig. 3a shows a spherical variogram model.

Figure 3

Here, variables close together have a relatively high correlation, but reach zero correlation at higher lags (arbitrarily chosen to be \( h = 5 \) in this example), where the variogram flattens out to the value \( 2\sigma^2 = 10 \). (Again, the variance for all \( Z(i) \) was arbitrarily chosen to be \( \sigma^2 = 5 \).) The expectations, \( 2\gamma^s(l;m) \) and \( \mathbb{E}[\text{TTLV}(m)] \), for the spherical variogram model are given in Fig. 3b. Fig. 4a

Figure 4

shows the "hole-effect," or wave variogram model, which often occurs under
periodic correlation structure. Fig. 4b shows the corresponding values of 
$2\gamma^a(l;m)$ and $E[TTLV(m)]$.

Notice that for both the spherical and wave models, $E[TTLV(m)]$ increases to 
values above the true variance of $\sigma^2=5$. Also, although not plotted on the 
figures, $E[MSBS(2^r)]$ is the same as $E[TTLV(m)]$ for those abscissa values where 
m=2^r; r=0,1,...,k-1.

For real data, we took a set which appeared to have no trend (Fig. 5a). 
These data consist of percent vertical cover in an igneous glade (Nelson 1985) 
in the Ozarks of southeast Missouri (Shannon County). Igneous glades in this 
area occur on a bedrock of rhyolite. These glades are grassy openings in a 
mixed oak (*Quercus* spp.) and hickory (*Carya texana*, primarily) forest matrix. 
The herbaceous nature of glades in this area is due to the shallow and extremely 
droughty nature of the soils. The glade flora is dominated by *Andropogon* 
*scoparius* and other grasses in areas with deep soil (greater than 15 cm) and by 
*Crotonopsis elliptica* and other annuals in areas with shallow soil (less than 15 
cm).

Along a 30 m transect, each 50 cm segment was photographed from ground 
level through a 10 cm wide strip of vegetation. Each segment was backed by a 
white screen with a vertical scale on it. The 60 photographs were each 
digitized, and image analysis was used to compute the percent of nonwhite pixels 
in a vertical column 1.5 m tall by 10 cm wide, making a total of $n=300$ estimates 
of percent vertical cover along the transect. This method was adapted from 
Roebertson et al. (1988) and used by Ver Hoef et al. (1989). The raw data are 
given in Fig. 5a.
We calculated the estimator \( \hat{\gamma}^b(h) \), given in (15), of the variogram \( \gamma^b(h) \) (Fig. 5b). Next we estimated \( \gamma^a(1;m) \) with both estimators: \( \hat{\gamma}^a(1;m) \) in (13) and \( \hat{\gamma}^b(1;m) \) in (16) (Fig. 5c). After scaling each by \((m/2)\), TTLV(m) and \((m/2)\hat{\gamma}^a(1;m)\) are obtained (Fig. 5d). It can be seen that the approximation using \((m/2)\hat{\gamma}^a(1;m)\) is very close to TTLV(m), and at least retains all of the features of TTLV(m).
7. DISCUSSION AND CONCLUSIONS

The overall goal of this paper was to show some mathematical relationships between nested ANOVA, TTLV, and PQV. In order to make these relationships, the statistical model (2) was necessary. With (2), we showed typical model assumptions for nested ANOVA and variogram analysis of spatial data.

For nested ANOVA, it has been usual to assume spatially nested mean effects, (3), (e.g., Greig-Smith, 1952); or spatially nested random effects, (4), (e.g., Youden and Mehlich 1937), with independence among random variables. In contrast, the usual assumptions for a variogram analysis are intrinsic stationarity, namely, a constant mean and spatial dependence among random variables, or a variogram, which is a function only of the displacement $h$ between the spatial locations under consideration.

Assuming the variogram model, or intrinsic stationarity, the aggregation of data by TTLV led naturally to a definition of the statistical quantity $(m/2)\gamma^A(l;m)$, which is a variogram as a function of aggregation. Given intrinsic stationarity, the expected values of nested ANOVA (11) and TTLV (12) are $(m/2)\gamma^A(l;m)$. Miesch (1975) has shown the opposite. That is, he assumed nested random effects, and then showed that the classical variogram estimator (15) is a function of the mean square estimates from nested ANOVA. However, the result of Miesch is only true for the rather special sampling design he specifies, and would not hold for a transect of contiguous quadrats.

Although TTLV has been used previously to estimate the "scale of pattern," a definition of "scale of pattern", and moreover a statistical quantity that
embodies it, are difficult to find in the ecological literature. On the other hand, TTLV estimates the statistical quantity \((m/2)2\gamma^\theta(l;m)\), which is of interest since it is invariant to aggregation \(m\) when there is spatial independence among the random variables. We have shown that there are two estimators of \((m/2)2\gamma^\theta(l;m)\), namely, \((m/2)2\gamma^\theta(l;m) (=\text{TTLV}(m))\) and \((m/2)2\gamma^\theta(l;m)\). Since \((m/2)2\gamma^\theta(l;m)\) is a function of \(\text{PQV} (=2\gamma^b(h))\), TTLV can be approximated from knowledge of \(\text{PQV}\). As Fig. 5d shows, they may often yield similar results.

From our simulation study, in order to estimate \((m/2)2\gamma^\theta(l;m)\), TTLV(m) can be recommended when there is negative autocorrelation, and \((m/2)2\gamma^\theta(l;m)\) is recommended when there is positive autocorrelation. The variance of TTLV(m) is much smaller than \((m/2)2\gamma^\theta(l;m)\) for negative values of \(\rho\), while the variance of \((m/2)2\gamma^\theta(l;m)\) is only moderately smaller than TTLV(m) over positive values of \(\rho\). Therefore, the "safe" estimator to use is TTLV(m). However, when positive spatial dependence is thought to be present (which is often the case), \((m/2)2\gamma^\theta(l;m)\) has some slight advantage. The direction (and strength) of dependence can be checked by estimating the variogram (e.g., with equation 15). (Negative autocorrelation will give a wave-type model, and positive autocorrelation will give an exponential-type model.) However, note that the efficiencies which we calculated depend on the first-order autoregressive model from which the data were generated, and may not hold true for other cases.

Of practical concern to the ecologist is the choice of a variogram or a nested ANOVA approach for analyzing spatial data. This decision should be based on which model assumptions are most realistic. However, the two methods are not mutually exclusive. It is possible to consider models such as spatially nested
random effects, where the random variables are independent among blocks, but are correlated within blocks. Intuitively, the spatially nested effects could be removed with nested ANOVA, and the residuals analyzed by a variogram analysis. The properties of such approaches need further study.
8. ACKNOWLEDGMENTS

This work was supported by funds from the Whitehall Foundation, the North Atlantic Treaty Organization (NATO), the National Park Service, and the National Science Foundation, grant number DMS-9001862.
9. LITERATURE CITED


10. APPENDIX

The mean square for each block (aggregation) effect in a spatially nested ANOVA of a transect of contiguous quadrats can be written as:

\[ MSBS(2^r) = \frac{1}{2^{2^r}} \text{ave} \left[ \sum_{i=j+1}^{j+2^r-1} Z(i) - \sum_{i=1}^{j} Z(i) \right]^2 : j=(q-1)2^r+1; \ q=1,\ldots,2^{k-r}-1; \]

\( r=0,1,\ldots,k-1. \)

Proof

First, consider the algebra for calculating the mean square of block size 1, MSBS(1). From Table 1, change from Y(•) notation to Z(•) notation to obtain,

\[ MSBS(1) = \frac{1}{2^{k-1}} \left[ \left( \frac{Z(1)-Z(2)}{2} \right)^2 + \left( \frac{Z(3)-Z(4)}{2} \right)^2 + \left( \frac{Z(5)-Z(6)}{2} \right)^2 + \cdots \right] \]

\[ = \frac{1}{2^{k-1}} \frac{1}{2} \left[ (Z(1)-Z(2))^2 + (Z(3)-Z(4))^2 + \cdots + (Z(n-1)-Z(n))^2 \right]. \quad (A1.1) \]

Notice that there are exactly \( 2^{k-1} \) terms summed in (A1.1), so write,

\[ MSBS(1) = \frac{1}{2} \text{ave} \left[ \left( \frac{Z(1)-Z(2)}{2} \right)^2, \left( \frac{Z(3)-Z(4)}{2} \right)^2, \ldots, \left( \frac{Z(n-1)-Z(n)}{2} \right)^2 \right]. \quad (A1.2) \]

For block size 2, let,

\[ W(1) = \frac{Z(1)+Z(2)}{2}, \ W(2) = \frac{Z(3)+Z(4)}{2}, \ldots, \text{etc.} \]

Then, calculate the mean square for block size 2, MSBS(2), as was done for MSBS(1) in (A1.1):
\[
\text{MSBS}(2) = \frac{2}{2^{k-2}} \left[ \left( W(1) - W(2) \right)^2 + \left( W(3) - W(4) \right)^2 + \ldots + \left( W \left( \frac{n}{2} - 1 \right) - W \left( \frac{n}{2} \right) \right)^2 \right].
\]

(A1.3)

In (A1.3), there are exactly \(2^{k-2}\) terms summed, so write,

\[
\text{MSBS}(2) = \text{ave} \left[ \left( W(1) - W(2) \right)^2, \left( W(3) - W(4) \right)^2, \ldots, \left( W \left( \frac{n}{2} - 1 \right) - W \left( \frac{n}{2} \right) \right)^2 \right]
\]

\[
= \text{ave} \left[ \left( \frac{Z(1)+Z(2)}{2} - \frac{Z(3)+Z(4)}{2} \right)^2, \ldots, \left( \frac{Z(n-3)+Z(n-2)}{2} - \frac{Z(n-1)+Z(n)}{2} \right)^2 \right]
\]

\[
= \frac{1}{4} \text{ave} \left[ \left( \frac{Z(1)+Z(2)-Z(3)-Z(4)}{2} \right)^2, \ldots, \left( \frac{Z(n-3)+Z(n-2)-Z(n-1)+Z(n)}{2} \right)^2 \right].
\]

(A1.4)

Generalizing from (A1.1) through (A1.4), we obtain,

\[
\text{MSBS}(2^r) = \frac{2^{2^r}}{2} \text{ave} \left[ \frac{1}{(2^r)^2} \left( Z(1)+\ldots+Z(2^r)-Z(2^r+1)-Z(2^{r+1}) \right)^2, \ldots \right]
\]

\[
= \frac{1}{2(2^r)} \text{ave} \left[ \left( Z(1)+\ldots+Z(2^r)-Z(2^r+1)-Z(2^{r+1}) \right)^2, \right.
\]

\[
\left. \left( Z(2(2^r)+1)+\ldots+Z(3(2^r))-Z(3(2^r)+1)\ldots-Z(4(2^r)) \right)^2, \text{etc.} \right]\]

so,

\[
\text{MSBS}(2^r) = \frac{1}{2(2^r)} \text{ave} \left[ \sum_{i=1}^{2^r} Z(i) - \sum_{j=2^{r+1}}^{2^{r+1}} Z(i) ; j=(q-1)2^{r+1}; q=1,\ldots,2^{k-r-1} \right]
\]

\(r=0,1,\ldots,k-1.\)
Table 1. Nested ANOVA table for spatial data along a line transect.

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>BS($2^{k-1}$)</td>
<td>1</td>
<td>$2^{k-1} \sum {\bar{Y}(aA...A) - \bar{Y}(AA...A)}^2$</td>
<td>S.S.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$a=1$</td>
<td>d.f.</td>
</tr>
<tr>
<td>BS($2^{k-2}$)</td>
<td>2</td>
<td>$2^{k-2} \sum \sum {\bar{Y}(abA...A) - \bar{Y}(aA...A)}^2$</td>
<td>S.S.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$a=1$ $b=1$</td>
<td>d.f.</td>
</tr>
<tr>
<td>BS(2)</td>
<td>$2^{k-2}$</td>
<td>$2 \sum \sum \sum {\bar{Y}(ab...cA) - \bar{Y}(ab...AA)}^2$</td>
<td>S.S.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$a=1$ $b=1$ $c=1$</td>
<td>d.f.</td>
</tr>
<tr>
<td>BS(1)</td>
<td>$2^{k-1}$</td>
<td>$\sum \sum \sum {\bar{Y}(ab...cd) - \bar{Y}(ab...cA)}^2$</td>
<td>S.S.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$a=1$ $b=1$ $d=1$</td>
<td>d.f.</td>
</tr>
<tr>
<td>Total</td>
<td>$2^{k-1}$</td>
<td>$\sum \sum \sum {\bar{Y}(ab...cd) - \bar{Y}(AA...AA)}^2$</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Expected mean square column for nested ANOVA (Table 1) under the random effects model in equation (4).

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>Mean Square</th>
<th>Expected Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>BS($2^{k-1}$)</td>
<td>1</td>
<td>S.S.</td>
<td>$\sigma_1^2 + 2\sigma_2^2 + \ldots + 2^{k-2}\sigma_{k-2}^2$</td>
</tr>
<tr>
<td>BS($2^{k-2}$)</td>
<td>2</td>
<td>S.S.</td>
<td>$\sigma_1^2 + 2\sigma_2^2$</td>
</tr>
<tr>
<td>BS(2)</td>
<td>$2^{k-2}$</td>
<td>d.f.</td>
<td>$\sigma_1^2 + 2\sigma_2^2$</td>
</tr>
<tr>
<td>BS(1)</td>
<td>$2^{k-1}$</td>
<td>d.f.</td>
<td>$\sigma_1^2$</td>
</tr>
</tbody>
</table>
Figure 1. Isopleth lines for the efficiency of $\hat{2}^\gamma(1;m)$ versus $2^\gamma(1;m)$, where $m$ is the level of aggregation and $\rho$ is autocorrelation in the first-order autoregressive model given in the text.
Figure 2. (a) Variogram model under complete spatial randomness, where the variance was arbitrarily chosen as 5. (b) Aggregated variogram and $E(TTLV(m))$ under complete spatial randomness.
Figure 3. (a) Spherical model for the variogram, where the variance was arbitrarily chosen as 5. (b) Aggregated variogram and \( E(TTLV(m)) \) for the spherical variogram model.
Figure 4. (a) Wave model for the variogram, where the variance was arbitrarily chosen as 5. (b) Aggregated variogram and E[TTLV(m)] for the wave variogram model.
Figure 5. (a) Percent vertical cover data for igneous glade vegetation along a line transect of 300 contiguous plots. (b) The empirical variogram (15) calculated from the data.
Figure 5. (c) The aggregated variogram (10) estimated by blocking directly, and estimated from (16). (d) TTLV (5) calculated by blocking directly, and approximated from the empirical variogram (15) used in (16).
PART II.

STATISTICAL TECHNIQUES FOR AVERAGE PATCH SIZE IN A TRANSECT
Statistical techniques for average patch size in a transect

Authors:

Jay M. Ver Hoef\textsuperscript{1,2}
Noel A.C. Cressie\textsuperscript{1}
David C. Glenn-Lewin\textsuperscript{2}

1 Department of Statistics, Iowa State University, Ames, Iowa, 50011
2 Department of Botany, Iowa State University, Ames, Iowa, 50011
ABSTRACT

For determining average patch size in a line transect of contiguous plots, we investigate three methods: 1) two term local variance (Hill 1973), 2) moving two-sample t-tests, and 3) a Bayesian approach using simulated annealing. In the line transect, we assume that true patch sizes vary and are hidden by random noise. To assess the performance of the three methods, we used a 3x2x3 factorial design in a computer experiment to determine the effects of three factors: 1) the signal-to-noise ratio, 2) the expected sizes of the patches relative to the plot size, and 3) the distribution of patch sizes. The results indicate that all methods estimate average patch size better when the signal-to-noise ratio is high, when the average size of patches is large, and when the patch size variation is small. The strengths and weaknesses of each method are discussed, but the Bayesian approach seems best suited for estimating average patch size. An example from grassland vegetation is included.

Keywords: Bayesian statistics, pattern analysis, change-points, edge detection, two-term local variance, simulated annealing, Markov random fields, image analysis
1. INTRODUCTION

There are many different spatial models which might be assumed for observed spatial patterns in ecological data. A very general mathematical structure which encompasses many of these models is the random process,

$$(Z(s): s \in \mathbb{D} \subset \mathbb{R}^d),$$

where $Z(s)$ is a random variable indexed by spatial location vector $s$, and $\mathbb{D}$ is some domain which is a subset of Euclidean space $\mathbb{R}^d$ with dimension $d=1$, 2, or 3. More specifically, consider the following model:

$$(Z(s): Z(s) = \mu(s) + \delta(s); s \in \mathbb{D} \subset \mathbb{R}^d), \quad (1)$$

where $\mu(s)$ is a fixed mean parameter and $\delta(s)$ is a zero mean random variable at location $s$.

Now, suppose that $(Z(i): i=1,2,...,n)$ are arranged in a line transect of contiguous plots, or quadrats,

$$Z(1) Z(2) \ldots Z(n),$$

and are modeled according to (1) where $\mathbb{D}=(1,2,...,n)$ and $d=1$. For example, $\mu(i)$ might consist of smooth functions; e.g., $\mu(i)=\lambda_1 + \lambda_2 \sin(\lambda_1 i)$, or $\mu(i)$ might consist of patches marked by discontinuities; e.g., $\mu(i) \in P$, where $P$ is a set with a finite number of real values. The random errors $\{\delta(i)\}$ may be mutually independent or contain spatial dependence.

We define a patch of size $\ell$ as a set $\{\mu(i): \mu(i)=\mu(i+1)=\ldots=\mu(i+\ell)\}$, and we will call $\{\mu(i): i=1,2,...,n\}$ the mean structure for the transect. In all that is to follow, suppose $\mu(i) \in (a,b)$ for all $i$, implying a two-phase mean structure composed of patches, and suppose that the random errors $\{\delta(i)\}$ (or noise) are
mutually independent. Fig. 1 gives three examples. Some authors distinguish the higher value of a or b as the "patch," and the lower value as the "gap" (e.g., Carpenter and Chaney 1983, Galiano 1983, Dale and MacIsaac 1989).

In the statistical literature, a change-point c(i) is any point along a transect in time or space where one mean changes to another, or \( \mu(i) - \mu(i+1) \neq 0 \), which indicates crossing from one patch to another (e.g., Lombard 1989). This has also been called an "edge" (Ludwig and Cornelius 1987, Orloci and Orloci 1990). Let \( N(c) \) denote the number of change-points along the transect, or,

\[
N(c) = \sum_{i=1}^{n-1} I(i), \quad I(i) = \begin{cases} 
1 & \text{if } \mu(i) - \mu(i+1) \neq 0 \\
0 & \text{if } \mu(i) - \mu(i+1) = 0 
\end{cases}
\]  

(2)

For the transect we are considering, we define average patch size (APS) as the total length of the transect \( n \) divided by the number of patches, or,

\[
\text{APS} = \frac{n}{N(c)+1} .
\]  

(3)

Ecologists often reduce the complex nature of species and communities to a single index, such as the various indices of diversity (for a review, see Magurran 1988), so that comparisons may be made among different areas or through time; APS is such a measure. The index APS has also been called the "scale of pattern" (Usher 1975, Ludwig and Goodall 1978, Dale and MacIsaac 1989, Dale and Blundon 1991).

For transect data, two term local variance (TTLV, Hill 1973) is reported to have a peak at APS (e.g., Dale and MacIsaac 1989). There have been many investigations of the efficacy of TTLV and other pattern statistics. Results
have been obtained for simple artificial patterns that contain fixed patch sizes $\ell$ and no noise $\delta(i)$ (e.g., Errington 1973, Usher 1969, Dale and MacIsaac 1989). Some studies allow for variable patch sizes but no noise (Errington 1973, Carpenter and Chaney 1983), some allow noise but fixed patch sizes (Usher 1969, 1975), while just a few allow for variable patch sizes and noise (Ludwig and Goodall 1978, Ludwig 1979).

From these studies and our own observations, three factors appear to have important effects on the performance of pattern statistics: 1) the signal-to-noise ratio (R factor), 2) the expected size of patches relative to the plot size (S factor), and 3) the distribution of patch sizes (D factor). These three factors are described further in the methods section.

For estimating APS, the objectives of this paper are to examine the effects of R, S, and D on TTLV, as well as on a moving two-sample t-test (MT) and a Bayesian approach using simulated annealing (BSA). Good simulation studies should be statistically well-designed, just as any good field experiment (Kleijnen 1987, Sacks et al. 1989). To examine the effects of factors R, S, and D on TTLV, MT, and BSA, we analyzed a computer simulation experiment using a 3x2x3 factorial design. After assessing the performance of each method, we conclude with a real example on grassland vegetation.
2. METHODS

2.1. **Two term local variance (TTLV)**

Two term local variance (Hill 1973) is given by,

\[
\text{TTLV}(m) = \frac{1}{2m} \text{ave} \left( \sum_{i=1}^{j+m-1} z(i) - \sum_{i=1}^{j+m} z(i) \right)^2 : j=1,2,...,n-2m+1; m \leq \lfloor n/2 \rfloor,
\]

where data \( Z = z = (z(1),...,z(n))' \) are observed, \( \text{ave}[\cdot] \) denotes the average of terms within the brackets, and \( \lfloor n/2 \rfloor \) denotes the integer part of \( n/2 \). The quantity \( \text{TTLV}(m) \) is usually plotted (as a function of \( m \)) against \( m \). Then the "peak", or maximum of \( \text{TTLV}(m) \), is reported to be at the "scale of pattern," or APS. Notice that when \( m=\lfloor n/2 \rfloor \) in (4), there is only one term within \( \text{ave}[\cdot] \); due to a lack of averaging and the noisy nature of the transect, any maximum at \( m=\lfloor n/2 \rfloor \) may be spurious. This is generally true for larger values of \( m \), so we constrained the largest \( m \) to \( \lfloor n/3 \rfloor \). Then define

\[
\text{TTLV} = \max \{ \text{TTLV}(m) : m \in M=\{1,2,\ldots,\lfloor n/3 \rfloor \}, \}
\]
as an estimator of APS.

2.2. **Moving two-sample t-test (MT)**

Next we consider a method of edge detection described by Webster (1973). Ludwig and Cornelius (1987) provide a concise review of the multivariate method, but here we consider the univariate case that we call the moving two-sample t-test. Denote a two sample t-test of means for the data beginning with \( z(i) \) as:
MTT(i;w) = \frac{\bar{z}_1(i) - \bar{z}_2(i)}{[2S^2_p(i)/w]^{1/2}} \quad ; \quad i=1,2,\ldots,n-2w+1, \quad (6)

where,

$$\bar{z}_1(i) = \frac{\sum_{j=1}^{1+w-1} z(j)}{w}, \quad \bar{z}_2(i) = \frac{\sum_{j=1+w}^{1+2w-1} z(j)}{w},$$

and,

$$S^2_p(i) = \frac{\sum_{j=1}^{1+w-1} (z(j) - \bar{z}_1(i))^2 + \sum_{j=1+w}^{1+2w-1} (z(j) - \bar{z}_2(i))^2}{2(w-1)},$$

and w is the "window" size, the number in each mean. The quantity $S^2_p(i)$ is the familiar "pooled" estimate of variance (Snedecor and Cochran 1980, p. 91), which depends on the starting position i, as do $\bar{z}_1(i)$ and $\bar{z}_2(i)$. Now, if $|MTT(i;w)|$ becomes large, say greater than some $t^\alpha_\alpha$, we declare a change-point at $i+w-1$. We call this method a moving two-sample t-test since, using (6), we begin at plot i=1 and simply slide along the transect one plot at a time, calculating the two-sample t-value at each location. To estimate the number of change-points (2) in the transect, for some w and cutoff value $t^\alpha_\alpha$, use

$$MT(\alpha,w) = \sum_{i=1}^{n-2w+1} I(|MTT(i;w)| > t^\alpha_\alpha \cdot (1-I(|MTT(i-1;w)| > t^\alpha_\alpha)), \quad (7)$$

where $I(\cdot)$ is the indicator function ($I(\cdot)=1$ if $\cdot$ is true, $I(\cdot)=0$ otherwise)

and $t^\alpha_\alpha$ is the appropriate two-sample t-value for a given $\alpha$-level and sample size w. When the value of the two-sample t-test crosses the threshold $\pm t^\alpha_\alpha$, the first term in the summand of (7) indicates that a change-point has occurred. The
second term simply ensures that another change-point is not counted until MTT(i;w) is within the threshold bounds again. Since (7) estimates the number of change-points (2), from (3) APS is estimated by,

\[ \frac{n}{MT(\alpha,w)+1}. \]  

The change-points themselves are estimated to be at i+w-1 (the last location in \( \tilde{z}^i_{1} \) before \( \tilde{z}^i_{2} \)) for i wherever the maximum of |MTT(i;w)| occurs once it crosses the threshold \( t^\alpha \). MT can then be used to estimate each \( \mu(i) \) by defining \( \hat{\mu}(i) \) to be the average of all \( z(i) \)'s between two change-points (including the right change-point but not the left). For MT, it is likely that \( \hat{\mu}(i) \in (a,b) \), but \( \hat{\mu}(i) \) can be modified to be either a or b, depending on which value is nearest.

2.3. A Bayesian approach with simulated annealing (BSA)

Assume the following model: Conditional on the mean, all random variables \( Z(i) \) are independent and identically distributed as normal random variables:

\[ g[z(i)|\mu(i)] \sim \text{i.i.d. } N[\mu(i),\sigma^2]. \]  

Then the joint conditional density is

\[ f(z|\mu) = \prod_{i=1}^{n} g[z(i)|\mu(i)]. \]

Now let \( \mu \), the vector of all \( \mu(i) \), have a prior distribution:

\[ \pi(\mu;a,b,\beta) \propto \exp \left( \sum_{i=1}^{n} \sum_{j=1}^{n} \beta_{ij} \left[ \frac{(a-\mu(i))(a-\mu(j))}{(a-b)^2} + \frac{(b-\mu(i))(b-\mu(j))}{(a-b)^2} \right] \right), \]  

where
The prior distribution (10) is a generalization of that used by Greig et al. (1989), where they use \( a = 0 \) and \( b = 1 \). Notice that \( \mu \) has a distribution which is a Markov random field; that is, conditional probabilities depend only on local neighborhoods:

\[
\Pr(\mu(i) | \{ \mu(j) : j \neq i \}) = \Pr(\mu(i) | \mu(i-1), \mu(i+1)).
\]

For example, let \( n = 4, a = 0, \) and \( b = 1 \); then the complete parameter space \( \Theta \) for \( \mu \) along with prior probabilities is given in Table 1, where \( c = 2(e^{2\beta}+1)^3 \) is a normalizing constant so that

\[
\sum_{\mu \in \Theta} \pi(\mu) = 1.
\]

Thus, for example, \( \Pr(\mu(2)=0|\mu(1)=\mu(3)=\mu(4)=1) = 1/(1+\exp(4\beta)) = \Pr(\mu(2)=0|\mu(1)=\mu(3)=1) \). In general,

\[
\pi(\mu) = \frac{e^{2(n-N(c)-1)\beta}}{2(e^{2\beta}+1)^{n-1}},
\]

where \( N(c) \) is the number of change-points in \( \mu \) given by (2). From (11), \( \pi(\mu) \) is a decreasing exponential function of \( N(c) \). The parameter \( \beta \) may be thought of as a smoothing parameter of the mean structure since, as \( \beta \) increases, it puts relatively lower probability on many change-points in \( \mu \).

With the model (9) and prior (10), the posterior distribution is,

\[
p(\mu | z) = (1/\tau) f(z | \mu) \pi(\mu),
\]

where \( \tau \) is a normalizing constant which does not depend on \( \mu \). Now, given the data \( z = z \), we want to find the \( \mu \) that has the overall maximum probability from the posterior distribution \( p(\mu | z) \); this is called the maximum a posteriori (MAP)
estimate of $\mu$ (Besag 1986). A straightforward procedure would be to maximize over all $2^n$ possibilities for $\mu$, but this may be computationally prohibitive if $n$ is even moderately large.

To perform the maximization, we used simulated annealing instead (Geman and Geman, 1984); see the appendix for a description. Other maximization procedures have been proposed, such as iterated conditional modes (Besag 1986) and extensions of the Ford-Fulkerson algorithm (Greig et al. 1986).

The BSA method gives an estimate $\hat{\mu}(i)$ equal to either $a$ or $b$; $i=1,2,...,n$. These estimates are then substituted directly into (2) and (3) to estimate APS.

2.4. *A factorial experiment of simulated transects*

The factors we considered were 1) the signal-to-noise ratio (R factor), 2) the expected size of patches relative to plot size (S factor), and 3) the distribution of patch sizes (D factor).

For the R factor, let the true mean values $\mu(i)$ be only one of two values, $a$ or $b$. Conditional on $\mu(i)$, let each $Z(i)$ be an independently and normally distributed random variable with variance $\sigma^2$. Then the signal-to-noise ratio $R$ for the transect is $R=|a-b|/\sigma$. We took $R = 2^r$, $r = 0, 1, 2$, and $\sigma^2=1$. For example, fix $a=0$; then $b$ is 1, 2, or 4 corresponding to $R = 1, 2, \text{ or } 4$, respectively.

For the S factor, let individual patch sizes take on one of five values, contained in the vector $\xi = (\ell_1,...,\ell_5)'$. For all distributions chosen, the middle entry of the vector is the expected patch size. We chose the following two levels of the factor: $\xi_3=(1,2,3,4,5)'$ ($S=3$) and $\xi_8=(6,7,8,9,10)'$ ($S=8$).
For the D factor, let $\ell_1$ take on values with probability $k_1$. The five probabilities are contained in the vector $k = (k_1, \ldots, k_5)'$. We chose probabilities $k_U = (.45, .04, .02, .04, .45)'$ ($D=U$), $k_H = (.2, .2, .2, .2)'$ ($D=H$), and $k_A = (.01, .04, .90, .04, .01)'$ ($D=A$) for patch sizes.

We then simulated 100 transects for each combination of the R, S, and D factors in a 3x2x3 factorial design. For S=3, each transect was assigned 16 patches of each mean level, a and b, according to $k_3$ and $k$. For S=8, each transect was assigned 6 patches of each mean level, a and b, according to $k_8$ and $k$. Thus, for both S=3 and S=8, the expected transect length was 96, and hence the expected average patch size was 3 or 8, respectively. But, since patch size was a random variable, transect length varied, and so did the average patch size for each simulated transect.

An example of a simulated transect with R=1, S=3, and D=U, is given in Fig. 1a. Notice that most patches are of size 5 or 1, with only the third from last patch of size 4. This is because most of the probability is assigned to patch sizes of 1 and 5 in $k_3$ by the probability vector $k_U$. For this simulation, the transect length was 99, so the true APS was 3.09. An example of a simulated transect with R=2, S=3, and D=H is given in Fig. 1b. Here, patch sizes of $k_3$ had equal probability, $k_H$. For this simulation, the transect length was 89, so APS was 2.78. An example of a simulated transect with R=4, S=8, and D=A is given in Fig. 1c. Here, patches of lengths $k_8$ occurred with probability $k_A$, and all patches ended up with length 8 except the third, which was of length 7. For this simulation, the transect length was 95, so APS was 7.92.

For the simulated transects, we kept track of the true $y$ values so we knew
the true APS, and then calculated TTLV, MT, and BSA on each transect. The MT
and BSA methods have additional parameters; for MT denote,

\[ MT_{21} = MT(\alpha=0.1,w=2), \]
\[ MT_{25} = MT(\alpha=0.05,w=2), \]
\[ MT_{61} = MT(\alpha=0.1,w=6), \]
\[ MT_{65} = MT(\alpha=0.05,w=6). \]

For BSA, we took \( \sigma, a, \) and \( b \) to be the values used in the simulated transects,
and,

\[ BSA_1 \text{ used } \beta=0.1, \]
\[ BSA_2 \text{ used } \beta=0.25, \]
\[ BSA_3 \text{ used } \beta=0.5, \]
\[ BSA_4 \text{ used } \beta=1.0, \]

where recall \( \beta \) is the smoothing parameter in (11). For all simulated transects
and each method, we calculated the deviation \( \Delta_{\text{method}} \) from the true APS; e.g.,

\[ \Delta_{\text{method}} = APS - \hat{APS}_{\text{method}}, \]

(13)

where \( \hat{APS}_{\text{method}} \) is the estimated APS using method=TTLV,MT21,...,BSA4. The BSA
and MT methods also estimate each \( \mu(i) \), so for these methods the percent
misclassified were also calculated.
3. SIMULATION RESULTS

An example of a simulated transect and the results of TTLV, MT, and BSA are given in Fig. 2. Fig. 2a shows a simulated transect with R=2, S=8, and D=H.

Fig. 2b shows the true µ, as well as ̂µ for both BSA4 and MT61. Both BSA4 and MT61 correctly estimated the true APS=7.75 (=93/12). BSA misclassified 7.53% and MT61 misclassified 6.59% of the µ(i)'s. Fig. 2c shows the values of MTT(i,w), where w=6, as it moves along the transect; also shown is the threshold t̂. Notice that the first minimum of MTT(i,w), once it crosses the threshold, is at i=3, so the change-point is estimated to be at i+w-1=8. Fig. 2d shows TTLV(m) for m up to [n/3]=31, with a peak at 7. In this example, the deviations (13) are ΔBSA1=ΔMT61=0 and ΔTTLV=-0.75. This type of analysis was repeated for all 1800 simulated transects.

3.1. The 3x2x3 factorial experiment

For all simulated transects in the factorial experiment, the deviations (13) for each method were calculated. The results are given as "box and whisker" plots based on 100 simulated transects per R,S,D-combination (Fig. 3).

The lower end of the "box" represents the 25th percentile of the deviations Δ (13), and the upper end is the 75th percentile. The "whisker," or the long...
vertical line, represents the range of $\Delta$. The horizontal line is the median for $\Delta$, and the solid circle is the mean for $\Delta$. Since $\Delta$ is the deviation from the true APS, 0 is the ideal for all statistics.

We could have analyzed the results for each method with an ANOVA for the factorial design, or, since all methods were applied to each transect, the whole experiment could be analyzed as a split-plot design with the factorial treatments as whole plot effects. However, examination of Fig. 3 shows that the residuals from the cell means of $\Delta$ were not normally distributed, nor did they have homogeneous variances across factor combinations. In fact, some of the interest lies in the different variances among the methods. For these reasons, we simply present the results graphically through box and whisker plots (Fig. 3).

3.2. **Two term local variance**

TTLV is affected by all three factors. As $R$ increases, TTLV performs better (Fig. 3). There is also some interaction with the $D$ and $S$ factors. When $R=4$, $S=3$ and $D=A$, TTLV performs very well, but there is still a large range in $\Delta$ when $R=4$, $S=3$ and $D=U$, where TTLV overestimates APS. The range decreases and overestimation of APS disappears when $S=8$, $R=4$ and $D=U$. Overall, TTLV does better as $R$ increases, as $D$ goes from $U$ to $H$ to $A$, and when $S=8$.

For TTLV, $\Delta$ has some very skewed distributions, such as when $R=4$, $S=3$ and $D=H$, where the mean is above the 75th percentile. In fact, for all $R=1$, the mean is well above the median. This indicates that TTLV occasionally takes on very large positive deviations from the true APS.
For S=8, the medians (Fig. 3) show that TTLV consistently underestimates APS. This agrees with Dale and Blundon (1990), who obtained the same result in transects of fixed patch size without noise. Underestimation decreases for S=3 compared to S=8. In their paper, Dale and Blundon (1990) provide a correction to TTLV for underestimating APS. Our results indicate that when noise is present, and the signal-to-noise ratio is moderate (R=2 or less), on the average TTLV tends to overestimate APS; note this is in contrast to indications from the medians. Again, this is due to the skewed distribution of A, but suggests that it may not always be advisable to correct TTLV for underestimating APS.

3.3. **Moving two-sample t-test**

With MT, there are several choices regarding window size w and cutoff level \(\alpha\), choices that are critical in the performance of MT (Fig. 3). For example, when S=3, the mean structure is changing within a window size w=6, so in this case it is essentially impossible to detect change-points. Hence, MT61 and MT65 grossly overestimated APS for S=3, and were not even plotted. However, it is best to choose a window size as large as possible but still smaller than the smallest patch size in order to decrease the variability in the two sample means, \(\bar{z}_1(i)\) and \(\bar{z}_2(i)\) in (6). This can be seen in Fig. 3, since MT61 and MT65 perform better than MT21 and MT25 when S=8.

For S=3, MT21 performed considerably better than MT25, which suggests choosing an \(\alpha=.1\) cutoff value from the t distribution with 2 degrees of freedom. Likewise, for S=8, MT61 has a smaller range and does not overestimate APS as much as MT65, which again suggests choosing an \(\alpha=.1\) cutoff value from the
t-distribution with 10 degrees of freedom. In estimating average patch size, it is just as bad to estimate wrongly a change-point as to miss one, so the lower \( \alpha \)-level is justified. In fact, these results suggest that even lower \( \alpha \)-levels should be considered. The choice of \( \alpha \) appears to become less important as \( R \) goes to 4 (Fig. 3).

The method MT performs better for \( S=8 \) and increasing \( R \). In contrast to TTLV, the performance of MT changes little with changes in \( D \) for constant \( S \) and \( R \). For \( S=3 \), MT21 performs better than TTLV since it does not overestimate APS as much, and the range of \( \Delta \) is smaller (Fig. 3). For \( S=8 \), the performance of MT61 appears to be about equal to TTLV, as judged by the lengths of the boxes (Fig. 3), but the range of \( \Delta \) for TTLV is larger than that for MT61. Also the median for MT61 is very near zero for the cases where TTLV underestimates APS.

3.4. A Bayesian approach with simulated annealing

As with MT, BSA requires some choices before using the method. Although \( a \), \( b \), and \( \sigma \) can be chosen with a suitable data analysis (see Section 4), \( \beta \) is a smoothing parameter (11) which will affect the results of the method. For instance, notice that

\[
\text{average } \Delta_{\text{BSA}1} < \text{average } \Delta_{\text{BSA}2} < \text{average } \Delta_{\text{BSA}2} < \text{average } \Delta_{\text{BSA}4},
\]

for all combinations of \( S \), \( R \), and \( D \) (Fig. 3), indicating that the smaller the \( \beta \), the lower the estimate of APS. However, as \( R \) increases, the choice of \( \beta \) seems to become less important since BSA for all four \( \beta \) values appears to converge to the true APS (Fig. 3).

BSA does an excellent job in estimating APS, especially in those cases.
which are very difficult for TTLV or MT, such as when R=1 and S=3. For S=3, BSA1 repeatedly estimated the true APS with a very small range and box (Fig. 3), regardless of R or D (Fig. 3). This is more remarkable considering that the mean structure is not a Markov random field, which is the chosen prior, indicating BSA is quite robust. For S=8, it is difficult to say whether BSA3 or BSA4 performs better; an intermediate value like β=0.75 may be best when S=8.

For S=3, BSA with β=0.1 is clearly the best estimator of APS among all methods. For S=8, BSA3 and BSA4 are about equal to MT61 and TTLV, with all methods performing very well compared to S=3.

3.5. Misclassification rates

Both BSA and MT can be used to estimate μ(i). The true values μ(i) were stored while simulating each transect, so that the μ̂(i) could be compared to the true values. For each combination of R, S, and D, the average (over the 100 transects per factor combination) percent μ̂(i) misclassified are given in Table 2. The percent misclassified drops rapidly for both BSA and MT as R increases. Again, BSA1 is the best method when S=3. There seems to be little effect due to S in misclassification rates for BSA. However, for MT, there is a large effect; for S=8, MT61 has the smallest misclassification rate, including the BSA estimators. Notice that while BSA1 does a good job estimating APS when R=1 and S=3, there is still about a 30% misclassification rate. Also, notice that BSA1 always has a lower misclassification rate than BSA2, BSA3, and BSA4, even when S=8, where BSA3 and BSA4 estimated APS better. This indicates that one would not choose the same β to estimate APS as to minimize misclassification rates.
4. A GRASSLAND EXAMPLE

We took a set of data consisting of percent vertical cover in a dolomite glade in the Ozarks of southeast Missouri (Shannon County). These glades are grassy openings in a mixed oak forest matrix, and are dominated by *Andropogon scoparius* with a rich mixture of other grasses and herbs. Along a 30 m transect, each 50 cm segment was photographed from ground level through a 10 cm wide strip of vegetation. Each segment was backed by a white screen with a vertical scale on it. The 60 photographs were each digitized, and image analysis was used to compute the percent of non-white pixels in a vertical column 1.5 m tall by 10 cm wide, making a total of n=300 estimates of percent vertical cover along the transect. This method was adapted from Roebertson et al. (1988) and used by Ver Hoef et al. (1989). The data \{z(i)\} were log-transformed, \(y(i)=\log(z(i)+1)\) (Fig. 4a), due to skewness of the \(z(i)\)'s towards large values.

4.1. Estimating parameters

We chose this example to demonstrate the power of BSA, along with ways to estimate \(a\), \(b\), and \(\sigma\) before performing BSA. For initial estimates of \(a\) and \(b\), we plotted a histogram of the distribution of data \(y(i)\) (Fig. 5c). It appeared to be bimodal, suggesting a mixture of two bell-shaped distributions. From the
histogram (Fig. 5c), we estimated $\hat{a}=0.6$ (the center of the category where the first histogram peak occurred) and $\hat{b}=1.5$ (the value between the two categories, which were about equal, where the second histogram peak occurred). Next, we assumed a mixture of two normal populations, one with 100 observations and a mean at 0.6, and the other with 200 observations and a mean of 1.5, with a common variance $\sigma$. After ranking all of the data, and assuming 100 normally-distributed observations with a mean of 0.6, the average of the second $y_{(2)}$ and third $y_{(3)}$ ranked $y(i)$ values should be about $2\sigma$ from the population mean of 0.6. Hence, one estimate of $\sigma$ is

$$\hat{\sigma}_1 = \frac{0.6-(y_{(2)}+y_{(3)})/2}{2},$$

assuming that there are no extreme outliers from the population with mean 1.5 so that $y_{(1)}$, $y_{(2)}$, and $y_{(3)}$ are from the population with mean 0.6. It is not possible to estimate $\sigma$ from the other tail of the population with a mean of 0.6 since these data are mixed with data from the population with a mean of 1.5. However, the 295$^{th}$ ranked observation $y_{(295)}$ should be at about $2\sigma$ from the population mean 1.5, assuming normality and 200 observations in this population, so

$$\hat{\sigma}_2 = \frac{(y_{(295)}-1.5)/2}{2},$$

again assuming no contamination in the right tail of the population with mean 1.5 from the population with mean 0.6. Then, assuming a common variance, $\hat{\sigma}=(\hat{\sigma}_1+\hat{\sigma}_2)/2$, which was $\hat{\sigma}=0.3$ for these data.

With $\hat{a}=0.6$, $\hat{b}=1.5$, and $\hat{\sigma}=0.3$, we used BSA with $\beta=0.1$ to obtain $\hat{\mu}$. The estimated mean structure $\hat{\mu}(i)$, $i=1,2,\ldots,300$, is shown in Fig. 4b, yielding an APS of 2.94. After estimating $\hat{\mu}$, there were $n_a=118$ with $\hat{\mu}(i)=a$ and $n_b=182$ with
\( \hat{\mu}(i) = b. \) Other summaries of the mean structure are given in Table 3.

4.2. Diagnostics

We also performed some diagnostic checking. Assuming a normal distribution with a mean of 0.6, a standard deviation of 0.3, and \( n_a = 118, \) the expected histogram numbers of data in each category are given in Fig. 5a. Also, assuming a normal distribution with a mean of 1.5, a standard deviation of 0.3, and \( n_b = 182, \) the expected histogram numbers of data in each category are given in Fig. 5b. Fig. 5c gives the sum of Fig. 5a and Fig. 5b, along with the observed numbers. It appears that these data are modeled quite well as a mixture of two normal distributions, with separate means but a common variance.

Next, we calculated residuals \( r(i) = y(i) - \hat{\mu}(i). \) We analyzed these residuals \( r(i) \) using the SAS\(^\circ\) univariate procedure. There is no evidence that the residuals are significantly different from zero (at \( \alpha = .01, \) using a t-test), and although the test for normality (Shapiro and Wilk 1965) led to rejection at \( \alpha = .05, \) the normal probability plot (Fig. 6) is quite linear. The standard deviation of the residuals is \( \hat{\sigma} = .289, \) which is very close to our initial estimate of 0.3. Therefore, assuming each \( Y(i) \) to be normally distributed with a mean \( \hat{\mu}(i) \) (from BSA) and a standard deviation \( \hat{\sigma} = .289 \) is a good approximation for these data.

In order to assess whether the residuals are spatially independent, we estimated the variogram,
\[ 2\tilde{\gamma}(h) = \frac{1}{n-h} \sum_{i=1}^{n-h} (r(i)-r(i+h))^2, \]
on the residuals. We then simulated 1000 transects of 300 spatially independent, normally distributed random variables with mean zero and standard deviation \( \sigma = .289 \). The expected variogram and the 2.5 and 97.5 percentiles for the 1000 simulated transects are shown in Fig. 7. It can be seen that there is no evidence for spatial dependence among residuals. We tried other values for \( \beta \) in BSA, but found that the residuals did not appear spatially independent for higher values of \( \beta \).

In summary, a Bayesian posterior distribution (12) seems reasonable for these data. The \( y(i) \) are conditionally independent and normally distributed random variables as in (9), with \( \sigma = .289 \). The vector of mean parameters, \( \mu \), has a prior distribution (10) with \( a = 0.6, b = 1.5, \beta = 0.1 \) and \( \mu \) as estimated by BSA (Fig. 4b).

Assuming this Bayesian model to be correct, the performance of TTLV and MT might be predicted from Fig. 3, since for the grassland example \( R = 3, S = 3, \) and \( D \) has probabilities concentrated on small patch sizes. Hence, from Fig. 3, we might expect MT21 to estimate APS best among the MT estimators, but still overestimate APS. The APS estimate using MT21 is 4.23, which is indeed higher than 2.94 estimated by BSA1. In all, BSA1 estimated 31 patches of size one in these data, so we might expect MT to do quite poorly, even with \( w = 2 \). As predicted from Fig. 3, TTLV for the data \( y(i) \) overestimates APS (Fig. 8), when
considering either the largest peak (m=13) or the first peak (m=4). It is also interesting to see the effect of noise by calculating TTLV on the estimates $\hat{\mu}$ from BSA (Fig. 8); the overall shape of TTLV remains unchanged, except it has been lowered and shifted slightly left. Now the first peak is at m=3, while the APS for $\hat{\mu}$ is 2.94. That is, noise has tended to cause TTLV to overestimate APS. Also notice that TTLV has many peaks, with the highest peak at 13. As indicated by other authors (Usher 1975, Dale and Blundon 1990), TTLV may exhibit several peaks due to arrangements of the patches themselves, which may also be of interest, but it confounds inference on APS when using TTLV.

Finally, we emphasize that BSA estimates each $\mu(i), i=1,\ldots,n$, which allows estimation of APS. However, other summaries can also be calculated. For instance, Galiano (1983) introduced "New Two Term Local Variance" since he was more interested in the average patch size of, say, the b-phase (the "patch"), than the overall average patch size in combination with the a-phase (the "gap"). The BSA method can be used to estimate average patch sizes separately for both the a-phase (2.31) and the b-phase (3.57) (Table 3). In this paper, our objectives are to demonstrate and compare techniques for the grassland example; ecological interpretations of the results of BSA will appear in a later paper with similar analyses on other transects.
5. DISCUSSION AND CONCLUSIONS

Overall, it can be seen that all three factors (R, S, and D) are important. BSA is primarily affected by R, with better performance when the signal-to-noise ratio is high (R=4). MT performs better when signal-to-noise ratio is high (R=4) and when expected patch sizes are large relative to plot size (S=8). TTLV performs better when signal-to-noise ratio is high (R=4), when expected patch sizes are large relative to plot size (S=8), and when the patch size variation is small (D=A).

The strength of TTLV is that there are relatively few subjective judgments (although the maximum m must be chosen). It has the disadvantage of being quite erratic in behavior, as judged by the usually large range of deviations (13) in Fig. 3. TTLV consistently either overestimates or underestimates APS, depending on the combination of R, S, and D, and TTLV is the only method affected by D. Also, TTLV does not estimate the $\mu(i)$'s, which is useful for other data summaries besides APS. Overall, TTLV performed the poorest among the three methods.

The strength of MT is that it has a smaller range of deviations (13) than TTLV, and for moderate R, it estimates APS well when w and $\alpha$ are properly chosen. However, there are no practical guidelines for choosing w and $\alpha$, and too large a choice of w is a fatal blow to the estimator, regardless of R. Also, if there are patches of size 1 in the data, MT simply will not work well. MT does have the nice property of actually estimating the change-points, and subsequently $\mu(i)$ for all $i$, $i=1,...,n$. 
The strength of BSA is that it estimates APS well, even when the data contain many small patch sizes. Its estimation of APS is very stable, as judged by the small range of deviations (13) in Fig. 3. Although it is necessary to estimate \( a, b, \) and \( \sigma \) when using BSA, we showed how that can be handled easily in the grassland example. Some trial and error may be necessary when selecting \( \beta \), but is possible to check that the parameter estimates are realistic in the Bayesian framework, as was done for the grassland example. As \( R \) increases, the choice of \( \beta \) appears to become less important. The only disadvantage of BSA is that it is computationally the most expensive. Nevertheless, it is the method that we recommend. Additionally, since BSA estimates each \( \mu(i) \), \( i=1,\ldots,n \), it allows mean-structure summaries beyond APS.

There is increasing interest in Bayesian statistics in ecology (e.g., Reckhow 1990). Bayesian statistics allow inferences that might otherwise be very difficult. It might seem surprising that we are estimating 300 parameters \( \{\mu(i)\} \) from 300 data points \( \{y(i)\} \). It is the Bayesian framework which allows this. For analyzing a two-phase pattern along a transect, we adapted the Bayesian approach from image analysis. It has a straightforward generalization to more than two phase types, and to two-dimensional space, and has been used successfully in those contexts (Geman and Geman 1984, Besag 1986).

In practicing spatial pattern analysis, there is no reason not to try all of the methods we have examined, as well as others (e.g., Greig-Smith 1952, Usher 1975, Ludwig and Goodall 1978, Ripley 1978, Galliano 1983, Dale and MacIsaac 1989, and Orłóci and Orłóci 1990). Here, we have tried to decompose the data into a mean structure, composed of patches, and noise. This
decomposition is not unique (Ripley 1978, Cressie 1991, Chapter 3). For example, spectral analysis may also be used to decompose the data into a mean structure, composed of sine waves, and noise. The choice of the form of the mean structure depends on which appears more meaningful biologically, and no method is uniformly best.
6. ACKNOWLEDGMENTS

This work was supported by funds from the Whitehall Foundation, the North Atlantic Treaty Organization (NATO), the National Park Service, and the National Science Foundation, grant number DMS-9001862.
7. LITERATURE CITED


Here, we give the details of simulated annealing for maximum a posteriori (MAP) estimation of $\mu$. Conditional independence of $g(z(i)|\mu(i))$ in (9) and pairwise-only dependence of the $\mu(i)$'s in (10) imply (12) is again a Markov random field (e.g., Cressie 1991, Section 7.4),

$$\Pr[\mu(i)|\mu(j), j \neq i, z] = \Pr[\mu(i)|\mu(i-1), \mu(i+1), z].$$  \hspace{1cm} (A1)

Next, consider the joint probability distribution,

$$p_{\tau}(\mu|z) = [\tau(T)]^{-1}p(\mu|z)^{1/\tau},$$  \hspace{1cm} (A2)

where $\tau(T)$ is a normalizing constant that ensures (A2) is a probability, and $p(\mu|z)$ is given by (12). Now, since $p(\mu|z)$ is a Markov random field (A1), this implies that (A2) is also (Geman and Geman 1984), and, as $T \to 0$, $p_{\tau}(\mu|z)$ becomes concentrated on the MAP estimate $\hat{\mu}$.

Simulating annealing begins by simulating a realization from the joint probability distribution in (A2); as $T \to 0$ the estimate $\hat{\mu}$ is obtained. Let $\hat{\mu}_t$ be a realization for some iteration $t$; $t=1,2,...,m$. The initial estimates $\hat{\mu}_0(i)$, $i=1,2,...,n$, were taken to be either $a$ or $b$, depending on whether $a$ or $b$ is closer to $z(i)$. All $n$ sites are visited repeatedly in a fixed order, $i_t$, $i=1,...,n$, $t=1,...,m$, where for the first $n$ iterations $i=t$. Then, $\hat{\mu}_t(i)$ is chosen to be either $a$ or $b$ with probability,

$$p_{\tau}(\hat{\mu}_t(i) = k|\hat{\mu}_{t-1}(i-1), \hat{\mu}_{t-1}(i+1), z(i))$$

$$= \tau_t(T(t))^{-1}g(z(i))\hat{\mu}_t(i) = k \cdot \pi(\hat{\mu}_t(i) = k|\hat{\mu}_{t-1}(i-1), \hat{\mu}_{t-1}(i+1))^{1/\tau(t)},$$

where $k \in \{a,b\}$ and
\[ \tau_t(T(t))^{-1} = \sum_{k \in \{a, b\}} g(z(t) | \hat{\mu}_t(i) = k) \cdot \pi_t(i) = k | \hat{\mu}_{t-1}(i-1), \hat{\mu}_{t-1}(i+1)]^{1/T(t)}. \]

Finally, let \( \hat{\mu}_t(i) = \hat{\mu}_{t-1}(i) \) for \( i \neq t \). This produces a series of estimates \( \hat{\mu}_0, \hat{\mu}_1, \ldots, \hat{\mu}_m \).

The annealing schedule is simply the functional dependence of \( T(t) \) on \( t \), which determines the rate at which \( T(t) \) approaches zero. With an appropriate annealing schedule, \( \hat{\mu}_t \) converges to the MAP estimator after many iterations \( t \).

There have been several suggestions for the annealing schedule. Geman and Geman (1984) show that theoretically a log annealing schedule is best. Greig et al. (1986) also used a geometric annealing schedule since the log annealing schedule proceeds very slowly. We found that both the log and geometric annealing schedules converged too slowly to be practical for our computer experiment. We used a linear annealing schedule, which risks getting trapped in a local maximum. We tried it several times with different starting points on some practice data sets, and \( \hat{\mu} \) was always the same for a given data set, indicating a satisfactory performance. Our convergence criterion was that \( \hat{\mu} \) remain unchanged for 10 full sweeps through all locations \( i \).
Table 1. Example of all possible mean structure vectors $\mu$ for $n=4$, $a=0$, and $b=1$, with their associated probability listed directly below; $c$ is the normalizing constant $c$.

$$
\begin{align*}
\mu &= \begin{pmatrix}
0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\
0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1
\end{pmatrix} \\
\pi(\mu) &= \frac{e^{6\beta}}{c}, \frac{e^{4\beta}}{c}, \frac{e^{2\beta}}{c}, \frac{e^{4\beta}}{c}, \frac{e^{4\beta}}{c}, \frac{e^{0\beta}}{c}, \frac{e^{2\beta}}{c}, \frac{e^{4\beta}}{c}, \frac{e^{0\beta}}{c}, \frac{e^{4\beta}}{c}, \frac{e^{2\beta}}{c}, \frac{e^{4\beta}}{c}, \frac{e^{6\beta}}{c}
\end{align*}
$$
Table 2. Percent of $\hat{\mu}(i)$ misclassified based on 100 transects per R,S,D-combination.

<table>
<thead>
<tr>
<th>S</th>
<th>D</th>
<th>R</th>
<th>MT21</th>
<th>MT25</th>
<th>MT61</th>
<th>MT65</th>
<th>BSA1</th>
<th>BSA2</th>
<th>BSA3</th>
<th>BSA4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>38.18</td>
<td>42.04</td>
<td>49.43</td>
<td>49.92</td>
<td>29.98</td>
<td>37.31</td>
<td>43.16</td>
<td>46.88</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
<td></td>
<td>24.62</td>
<td>31.95</td>
<td>49.50</td>
<td>49.83</td>
<td>14.73</td>
<td>25.14</td>
<td>29.66</td>
<td>36.31</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td></td>
<td>5.12</td>
<td>11.07</td>
<td>49.61</td>
<td>49.83</td>
<td>1.97</td>
<td>3.85</td>
<td>5.54</td>
<td>6.94</td>
</tr>
<tr>
<td>3</td>
<td>H</td>
<td>2</td>
<td>38.10</td>
<td>40.98</td>
<td>43.43</td>
<td>45.37</td>
<td>30.13</td>
<td>45.43</td>
<td>47.66</td>
<td>49.72</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>25.98</td>
<td>31.05</td>
<td>38.79</td>
<td>42.15</td>
<td>15.01</td>
<td>29.57</td>
<td>34.07</td>
<td>39.18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>11.13</td>
<td>15.93</td>
<td>35.40</td>
<td>39.52</td>
<td>1.94</td>
<td>2.88</td>
<td>3.73</td>
<td>4.82</td>
</tr>
<tr>
<td>U</td>
<td>2</td>
<td></td>
<td>37.42</td>
<td>41.05</td>
<td>39.64</td>
<td>42.06</td>
<td>30.43</td>
<td>47.16</td>
<td>51.24</td>
<td>55.43</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>27.15</td>
<td>31.83</td>
<td>29.10</td>
<td>34.04</td>
<td>14.63</td>
<td>27.56</td>
<td>34.13</td>
<td>37.76</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>18.55</td>
<td>22.34</td>
<td>20.66</td>
<td>26.72</td>
<td>1.98</td>
<td>5.35</td>
<td>8.23</td>
<td>10.32</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>32.52</td>
<td>38.07</td>
<td>25.91</td>
<td>31.36</td>
<td>27.03</td>
<td>28.44</td>
<td>33.57</td>
<td>50.99</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td></td>
<td>2.74</td>
<td>8.01</td>
<td>.92</td>
<td>.90</td>
<td>2.14</td>
<td>2.43</td>
<td>2.05</td>
<td>1.80</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>32.57</td>
<td>36.88</td>
<td>24.83</td>
<td>29.93</td>
<td>26.32</td>
<td>36.48</td>
<td>40.94</td>
<td>41.10</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td></td>
<td>2.33</td>
<td>7.48</td>
<td>.52</td>
<td>.52</td>
<td>1.55</td>
<td>3.19</td>
<td>3.58</td>
<td>3.23</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>32.30</td>
<td>36.80</td>
<td>25.55</td>
<td>31.50</td>
<td>27.58</td>
<td>40.25</td>
<td>45.74</td>
<td>47.05</td>
</tr>
<tr>
<td>U</td>
<td>2</td>
<td></td>
<td>17.54</td>
<td>25.07</td>
<td>6.16</td>
<td>8.32</td>
<td>13.68</td>
<td>23.24</td>
<td>26.97</td>
<td>27.51</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td></td>
<td>2.77</td>
<td>7.69</td>
<td>.52</td>
<td>.52</td>
<td>1.97</td>
<td>4.07</td>
<td>6.11</td>
<td>6.71</td>
</tr>
</tbody>
</table>
Table 3. Summary of patch size results using BSA on a grassland transect of percent vertical cover data. The table entries are the number of patches for each patch size.

<table>
<thead>
<tr>
<th>patch size</th>
<th>number of patches</th>
<th>number of quadrats</th>
<th>patch size mean</th>
<th>patch size median</th>
<th>patch size mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>μ(s)</td>
<td>0.6</td>
<td>1.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>12</td>
<td>8</td>
<td>7</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>200</td>
<td>182</td>
<td>3.57</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>31</td>
<td>22</td>
<td>21</td>
<td>11</td>
<td>6</td>
</tr>
</tbody>
</table>
Figure 1. Examples of simulated transects. (a) \( R=1, S=3, D=U \). (b) \( R=2, S=3, D=H \). (c) \( R=4, S=8, D=A \).
Figure 2. (a) Simulated transect with R=2, S=8, and D=H. (b) True mean structure \( y \) and \( \hat{y} \) estimated by BSA4 and MT61.
Figure 2. (c) MTT(i,6) for all i=n-2w+1 of simulated transect. (d) TTLV for simulated transect.
Figure 3. Box and whisker plots of deviations Δ from true APS for each method of the computer simulation experiment. The "a" indicates that values were beyond the ordinate's limits.
Figure 4. (a) Log-transformed data \( y(i) \) of vertical cover (per cent).  
(b) Estimated mean structure \( \hat{\mu}(i) \) using BSA.
Figure 5. (a) Expected numbers of data in each category assuming $n=118$ normally-distributed random variables, each with a mean of 0.6 and a standard deviation of 0.3. (b) Expected numbers of data in each category assuming $n=182$ normally-distributed random variables, each with a mean of 1.5 and a standard deviation of 0.3. (c) Expected numbers of data in each category assuming a mixed population of from (a) and (b) above. The observed numbers of the $y(i)$ for the grassland transect are also given.
Figure 6. Normal probability plot of the residuals r(i).
Figure 7. Variogram on residuals $r(i)$, along with the expected variogram and 95% confidence intervals for $n=300$ normally and independently distributed random variables with a standard deviation of $\sigma=.289$, based on 1000 transect simulations.
Figure 8. TTLV calculated on the raw data, and calculated on \( \hat{\mu}(l) \) estimated using BSA.
PART III.
MULTIVARIABLE SPATIAL PREDICTION
Multivariable spatial prediction

Authors:

Jay M. Ver Hoef
Noel Cressie

Department of Statistics, Iowa State University, Ames, Iowa, 50011
ABSTRACT

For spatial prediction, it has been usual to predict one variable at a time, with the predictor based on either the same type of variable (spatial BLUP or kriging) or additional concomitant variables (cokriging). Optimal predictors can be expressed in terms of covariance functions or variograms. In earth science applications, it is often desirable to predict the joint spatial abundance of variables. After a brief review of spatial BLUP, kriging, and cokriging, the simultaneous spatial prediction of several variables is developed. It is shown that the multivariable spatial predictor is the same as cokriging one variable at a time. However, multivariable spatial prediction yields the mean-squared-prediction-error matrix, and so allows construction of joint multivariable prediction regions.

Keywords: geostatistics, kriging, cokriging, best linear unbiased prediction, generalized least squares
1. INTRODUCTION

In sciences such as biology, ecology, geology, etc., it is often desirable to predict variables (such as biomass, species counts, soil nitrogen, gold content, etc.) at unsampled spatial locations, based on data observed at nearby locations. For spatial prediction, it has been usual to predict one variable at a time, with a predictor based on either the same type of variable (spatial BLUP or kriging) or using concomitant variables (cokriging). However, the spatial prediction of several variables simultaneously may also be of interest. For example, in ecology, a community is defined as the co-occurrence and abundance of several species at the same spatial locale. Ecologists, then, are concerned with the joint spatial patterns of these species, and it is desirable to predict the joint abundance of species at unsampled spatial locations.

Consider a spatial vector-valued process:

\[ \{z(s): \ s \in D\}, \]

where \( z(s) \in \mathbb{R}^m \) and \( D \subseteq \mathbb{R}^d \). Let the data consist of \( \{z(s_1), ..., z(s_n)\} \) at spatial locations \( \{s_1, ..., s_n\} \) in \( D \) (Fig. 1). The goal is spatial prediction; i.e.,

\begin{equation}
 z(s) = \mu(s) + \delta(s), \quad (1)
\end{equation}

based on the data, we wish to predict \( z(s_0) \) or \( z(s_0) \). Assume

\[ z(s) = \mu(s) + \delta(s), \]

where \( \mu(s) \) is a mean vector composed of fixed effects, and \( \delta(s) \) is a random vector with zero mean. Define \( z_L = [z_{s_1}, ..., z_{s_n}]' \) and \( \delta_L = [\delta_{s_1}, ..., \delta_{s_n}]' \). Also, let \( \mu_L(s) = [\mu_{s_1}(s), ..., \mu_{s_n}(s)]' \beta_L \), where \( \beta_L \) is a \((p_L \times 1)\) vector of parameters, and \([x_L(s)]\) is a vector of "explanatory" variables for the
$i^{th}$ response variable. Finally, let $X_i$ be an $(nxp_i)$ matrix whose $k^{th}$ row is $[x_i(s_k)]'; k=1,...,n; i=1,...,m$, where $m$ is the number of response variables.

Then the linear model (1) can be written as,

$$
\begin{pmatrix}
Z_1 \\
Z_2 \\
\vdots \\
Z_m
\end{pmatrix} =
\begin{pmatrix}
X_1 & 0 & \cdots & 0 \\
0 & X_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & X_m
\end{pmatrix}
\begin{pmatrix}
\beta_1 \\
\beta_2 \\
\vdots \\
\beta_m
\end{pmatrix}
+ 
\begin{pmatrix}
\delta_1 \\
\delta_2 \\
\vdots \\
\delta_m
\end{pmatrix},
$$

or $Z = X\beta + \delta$. Also,

$$
z(s_o) = X(s_o)\beta + \delta(s_o) + z(s_o) + \delta(s_o),
$$

where $w = [0, \ldots, 0]$ for $i=1, \ldots, m$. 

Now, let $E[\delta]=0$ and $E[\delta(s_o)]=0$, and write $\Sigma = \text{var}(\delta)$ and $\Sigma_{s(s_o)} = \text{cov}[\delta(s_o), \delta(s_o)]$, where the entries of

$$
\Sigma =
\begin{pmatrix}
\Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1m} \\
\Sigma_{21} & \Sigma_{22} & \cdots & \Sigma_{2m} \\
\vdots & \vdots & \ddots & \vdots \\
\Sigma_{m1} & \Sigma_{m2} & \cdots & \Sigma_{mm}
\end{pmatrix},
$$

and $\Sigma_{s(s_o)} = \text{cov}(\delta(s_o), \delta(s_o))$. Further notation used is:

$\Sigma_{s(s_o)} = \Sigma_{s(s_o)}$, $\Sigma_{o(s_o)} = \Sigma_{o(s_o)}$, $\Sigma_{s(s_o)} = \Sigma_{s(s_o)}$, and $\Sigma_{o(s_o)} = \Sigma_{o(s_o)}$.

In subsequent sections, best linear unbiased prediction for a spatial process, including kriging and cokriging, will be reviewed and new results obtained. These results will then be generalized to simultaneous spatial prediction for several variables.
2. OPTIMAL PREDICTORS

2.1. *The optimal predictor*

Define a scalar function \( p(z; s_0) \) or vector function \( p(z; s_0) \) of the data \( z \) as a predictor of \( z_{\cdot}(s_0) \) or \( z(s_0) \), respectively. Let

\[
M_p = E\{[E(z; s_0) - Z(s_0)][E(Z; s_0) - Z(s_0)]'\}
\]

be the mean-squared-prediction-error matrix of the predictor \( p(z; s_0) \). Define \( q(z; s_0) \) to be an optimal predictor if \( M_p - M_q \) is nonnegative-definite for every \( p(z; s_0) \). This is just a generalization of the more familiar univariate problem: When predicting \( z_{\cdot}(s_0) \) with \( p_{\cdot}(z; s_0) \), one wishes to minimize \( E[p_{\cdot}(z; s_0) - z_{\cdot}(s_0)]^2 \) (the mean-squared-prediction-error) for \( p_{\cdot}(z; s_0) \); i.e., find \( q_{\cdot}(z; s_0) \) such that

\[
E[p_{\cdot}(z; s_0) - z_{\cdot}(s_0)]^2 - E[q_{\cdot}(z; s_0) - z_{\cdot}(s_0)]^2
\]

is nonnegative for every \( p_{\cdot}(z; s_0) \).

It is proved in the appendix (see result A1) that the optimal \( q(z; s_0) \) is \( E[z(s_0) | z] \). As a consequence, we obtain the well-known univariate result,

\[
q_{\cdot}(z; s_0) = E[z_{\cdot}(s_0) | z].
\]

2.2. *Best linear unbiased predictor*

Consider predicting the random variable \( z_{\cdot}(s_0) \). The linear predictor \( b'z \) is an unbiased predictor of \( z_{\cdot}(s_0) \) if \( E[b'z - z_{\cdot}(s_0)] = 0 \). Next, define \( a'z \) as the best linear unbiased predictor (BLUP) if \( a'z \) is unbiased and

\[
E[b'z - z_{\cdot}(s_0)]^2 - E[a'z - z_{\cdot}(s_0)]^2
\]

is nonnegative for every \( b \). The BLUP is now generalized for the multivariate case.

The linear predictor \( b'z \) is an unbiased predictor of \( z(s_0) \) if
E[B'Z] = E[z(z_0)]. Next, $A'Z$ is the best linear unbiased predictor if $A'Z$ is unbiased for $z(z_0)$, and

$$M_B - M_A = E[(B'Z-z(z_0))(B'Z-z(z_0))'] - E[(A'Z-z(z_0))(A'Z-z(z_0))']$$

is nonnegative-definite for every $B$. We call $M_B$ the prediction variance matrix.
Without loss of generality, let us say the goal is to predict \( z(s_0) \) from data \( z(s_1), \ldots, z(s_n) \). Suppose that \( z \) can be written as the linear model (2), and \( g \) is unknown. We shall restrict ourselves to linear unbiased predictors:

If \( m=1 \), use

\[
\sum_{k=1}^{n} b_k z(s_k) = b' z.
\]

If \( m=2 \), use

\[
\sum_{k=1}^{n} b_{1k} z(s_k) + b_{2k} z(s_k) = b'_{11} Z + b'_{21} z = b' z;
\]

and similarly for \( m=3,4, \ldots \). Assume uniform unbiasedness over all \( g \); i.e.,

\[ E[b' z] = E[z(s_0)] \]

for every \( g \). From (2) and (3), it is equivalent to assume

\[ b' X = w' \]

3.1. Prediction with covariances

Consider the case \( m=2 \). Optimal spatial predictors can be expressed in terms of covariances, as is now shown. From (2) and (3), and due to the unbiasedness conditions, we can write,

\[
[b' Z - z(s_0)]^2 = [b' X + b' Z - w' + \delta(s_1)]^2 = [b' \delta - \delta(s_0)]^2
\]

\[
= (b' \delta + b' \delta + b' \delta + \delta^2(s_0)).
\]

Since \( E(\delta) = 0 \) and \( E[\delta(s_0)] = 0 \), taking expectations yields,

\[
M = b' \Sigma b - 2b' c + \Sigma_0 (1,1).
\]

The \( a \) that makes \( M_a - M \) nonnegative for all \( b \) is obtained by minimizing (4) with respect to \( b \), subject to \( b' X = w' \). Then, from result A2 in the appendix, with \( \Sigma = b, \Sigma_0 = \Sigma_0 (1,1), \Sigma = c, \Sigma = \Sigma \), and \( a = -1 \), the optimal \( a \) is given by.
where $\lambda$ is a vector of Lagrange multipliers that guarantees the unbiasedness conditions. The equations (5) can be solved for $\alpha_i$ (see appendix result A2), and hence the BLUP (in terms of covariances) is

$$\hat{\alpha}_i = \lambda_i^X \Sigma^{-1}_X z_i + (w_i^\prime - \Sigma^{-1}_X X_i^\prime \Sigma^{-1}_X z_i)^\prime (X_i^\prime \Sigma^{-1}_X X_i)^{-1} X_i^\prime \Sigma^{-1}_X z_i.$$  

(6)

The prediction variance for the BLUP (6) is

$$M = \Sigma_0 (1,1) - \Sigma_0^X \Sigma_1 + (X_1^\prime \Sigma_1^{-1} X_1 w_i^\prime - w_i^\prime) (X_1^\prime \Sigma_1^{-1} X_1)^{-1} (X_1^\prime \Sigma_1^{-1} X_1 - w_i^\prime).$$  

(7)

In general, to predict $z_i(s_0)$ involves solving the equations,

$$\begin{pmatrix} \Sigma^X & X_i^\prime \\ X_i & 0 \end{pmatrix} \begin{pmatrix} \lambda_i \\ w_i \end{pmatrix} = \begin{pmatrix} \Sigma_i^X \\ w_i \end{pmatrix}; \quad i=1,2,\ldots.$$  

(8)

for $\lambda_i$. These equations (8) are identical to those given by Myers (1982), but, because here the data have been ordered into $z=[z_1^\prime,\ldots,z_m^\prime]$, the equations (8) are in a different order.

3.2. **Relationship to Generalized Least Squares**

The generalized least squares estimator of $\beta$ is $\hat{\beta}_{gL} = (X^\prime \Sigma^{-1}_X X)^{-1} X^\prime \Sigma^{-1}_X z$, so the BLUP (6) can be expressed in terms of $\hat{\beta}_{gL}$:

$$\hat{\alpha}_i = \lambda_i^X \Sigma^{-1}_X z_i + (w_i^\prime - \Sigma^{-1}_X X_i^\prime \Sigma^{-1}_X z_i)^\prime (X_i^\prime \Sigma^{-1}_X X_i)^{-1} X_i^\prime \Sigma^{-1}_X z_i.$$  

(9)

Also notice that, if the data are Gaussian with known means, the optimal predictor among all predictors, linear or otherwise, is

$$q_i(z_i; \xi_0) = E[z_i(s_0)|z] = \Sigma_i^X z_i^{-1} (z_i - X_i \beta_0) + w_i \beta.$$  

(10)

When $\beta$ is unknown, (9) is just (10) with $\beta$ replaced by $\hat{\beta}_{gL}$. 


3.3. **Prediction with variograms; Kriging** (m=1)

The goal is to predict \( z(s_0) \) with \( b'z \) such that \( b'x = [x(s_0)]' \); this latter condition ensures uniform unbiasedness. Since there is only one variable type, the subscript is omitted. As in section 3.1,

\[
[b'z - z(s_0)]^2 = [b'\delta - \delta(s_0)]^2.
\]

Now, provided

\[
\sum_{k=1}^{n} b_k = 1,
\]

the following algebraic identity holds (see appendix result A3):

\[
[b'\delta - \delta(s_0)]^2 = \sum_{k=1}^{n} \sum_{l=1}^{n} b_k b_l [\delta(s_k) - \delta(s_l)]^2 + 2 \sum_{k=1}^{n} b_k [\delta(s_k) - \delta(s_0)]^2.
\]

The condition (11) can be interpreted as requiring a column of all one's (indicating an overall mean effect) in the \( X \) matrix and a one in the first entry of \( x(s_0) \). Now, define the variogram as,

\[
2\gamma(s_k, s_l) = \text{var}[\delta(s_k) - \delta(s_l)],
\]

and since each \( \delta(\cdot) \) has zero expectation, (13) may also be written as

\[
2\gamma(s_k, s_l) = E[\delta(s_k) - \delta(s_l)]^2.
\]

Hence, write the expectation of (12) as,

\[
M_b = -\sum_{k=1}^{n} \sum_{l=1}^{n} b_k b_l \gamma(s_k, s_l) + 2 \sum_{k=1}^{n} b_k \gamma(s_k, s_0).
\]

In matrix notation, (14) is

\[
M_b = -b'\Gamma b + 2b'\chi,
\]

where,

\[
\chi = \begin{pmatrix}
\gamma(s_1, s_0) \\
\gamma(s_2, s_0) \\
\vdots \\
\gamma(s_n, s_0)
\end{pmatrix}
\] and

\[
\Gamma = \begin{pmatrix}
\gamma(s_1, s_1) & \gamma(s_1, s_2) & \ldots & \gamma(s_1, s_n) \\
\gamma(s_2, s_1) & \gamma(s_2, s_2) & \ldots & \gamma(s_2, s_n) \\
\vdots & \vdots & \ddots & \vdots \\
\gamma(s_n, s_1) & \gamma(s_n, s_2) & \ldots & \gamma(s_n, s_n)
\end{pmatrix}.
\]
The \( \alpha \) that makes \( M_a - M_b \) nonnegative for every \( b \) is obtained by minimizing (14) with respect to \( \alpha \), subject to \( b'X = x(s_0)' \). Then, from result A2 in the appendix, with \( B = b \), \( E_{00} = 0 \), \( E_0 = \chi \), \( E = \Gamma \), and \( \alpha = 1 \), the optimal \( \alpha \) is given by,

\[
\begin{pmatrix}
\Gamma \\
X \\
X'
\end{pmatrix}
\begin{pmatrix}
\alpha \\
\lambda
\end{pmatrix}
= \begin{pmatrix}
\chi \\
x(s_0)
\end{pmatrix},
\]

provided (11) holds. Solving (15) for \( \alpha \) gives the BLUP in terms of variograms:

\[
a'z = X' \Gamma^{-1}z + (x(s_0)' - X' \Gamma^{-1}X)(X' \Gamma^{-1}X)^{-1}X' \Gamma^{-1}z;
\]

it is in this form that universal kriging is usually presented (e.g., Journel and Huijbregts, 1978, p. 319). Further, the prediction variance for the BLUP (16) is often called the kriging variance. It is,

\[
M_a = z' \Gamma^{-1}z - [x(s_0)'], (X' \Gamma^{-1}X)^{-1}X' \Gamma^{-1}z,
\]

3.4. **Prediction with variograms; Cokriging (m\( \geq \)2)**

In cokriging, the goal is to predict \( z(s) \) from \( z \) using \( b'z = b_1'z_1 + b_2'z_2 + \ldots + b_m'z_m \). Without loss of generality, consider the case \( m=2 \), and suppose we want to predict \( z(s) \). Again, for uniform unbiasedness, we require \( E[b'z] = E[z(s)] \) for all \( \beta \); equivalent conditions are \( b'X = \omega' \), or \( b_1'X = [x_1(s)]' \)

and \( b_2'X = 0' \). Then,

\[
[b_1'z_1 + b_2'z_2 - z(s)]^2 = [b_1'\delta_1 + b_2'\delta_2 - \delta(s_0)]^2.
\]

Suppose that each \( X' \) has a column of ones, so that \( b_1'1 = 1 \) and \( b_2'1 = 0 \) are part of the unbiasedness conditions. Then, in the same manner that (12) was obtained, we can write the following algebraic identity:
Define the cross-variogram,

\[ 2\gamma_{ij}(s_k, s_{\ell}) = \text{var}[\delta_{k}(s_{ij}) - \delta_{\ell}(s_{ij})], \]

and, since each \( \delta(\cdot) \) has expectation 0, (19) may be written as

\[ 2\gamma_{ij}(s_k, s_{\ell}) = E[\delta_{k}(s_{ij}) - \delta_{\ell}(s_{ij})^2]. \]

Then express the expectation of (18) in terms of the cross-variogram defined in (19):

\[ M_b = \sum_{k=1}^{n} \sum_{\ell=1}^{n} b_{1k} b_{1\ell} \gamma_{11}(s_k, s_{\ell}) \]

\[ - 2 \sum_{k=1}^{n} \sum_{\ell=1}^{n} b_{1k} b_{2\ell} \gamma_{12}(s_k, s_{\ell}) - \sum_{k=1}^{n} \sum_{\ell=1}^{n} b_{2k} b_{2\ell} \gamma_{22}(s_k, s_{\ell}) \]

\[ + 2 \sum_{k=1}^{n} b_{1k} \gamma_{11}(s_k, s_0) + 2 \sum_{k=1}^{n} b_{2k} \gamma_{21}(s_k, s_0). \]

In matrix form, (20) is

\[ M_b = -b_1^T \Gamma b_2 + 2b_1^T \chi_1, \]

where,

\[ \Gamma = \begin{pmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{pmatrix}, \]

\[ \chi_1 = \begin{pmatrix} \chi_{11} \\ \chi_{21} \end{pmatrix}, \]

\[ b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}, \]

\[ \chi_{ij} = \begin{pmatrix} \gamma_{i1}(s_1, s_0) \\ \gamma_{i2}(s_2, s_0) \\ \vdots \\ \gamma_{in}(s_n, s_0) \end{pmatrix}, \]

\[ \Gamma_{ij} = \begin{pmatrix} \gamma_{ij}(s_1, s_1) & \gamma_{ij}(s_1, s_n) & \cdots & \gamma_{ij}(s_1, s_n) \\ \gamma_{ij}(s_2, s_1) & \gamma_{ij}(s_2, s_n) & \cdots & \gamma_{ij}(s_2, s_n) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{ij}(s_n, s_1) & \gamma_{ij}(s_n, s_2) & \cdots & \gamma_{ij}(s_n, s_n) \end{pmatrix}. \]
The \( a \) that makes \( M_b \mathbf{a} \) nonnegative for all \( b \) is obtained by minimizing (20) with respect to \( b \), subject to \( b' \mathbf{x} = w' \). Then, from result A2 in the appendix, with \( \mathbf{B} = \mathbf{b}, \ E_{oo} = 0, \ E_{o} = \mathbf{x}, \ E = \mathbf{r}, \) and \( \alpha = 1 \), the optimal \( a \) is given by,

\[
\begin{pmatrix}
\Gamma_{11} & \Gamma_{12} & \mathbf{x}_1' & 0 \\
\Gamma_{21} & \Gamma_{22} & 0 & \mathbf{x}_2' \\
\mathbf{x}_1 & 0 & \mathbf{1} & 0 \\
0 & \mathbf{x}_2 & 0 & 0
\end{pmatrix}
\begin{pmatrix}
a_1 \\
2_2 \\
\lambda_1 \\
\lambda_2
\end{pmatrix}
= \begin{pmatrix}
\mathbf{z}_1 \\
\mathbf{z}_2 \\
\mathbf{1} \\
0
\end{pmatrix},
\text{ or } \begin{pmatrix}
\mathbf{r} \\
\mathbf{x}' \\
\mathbf{x}_1 & 0 \\
0 & 0
\end{pmatrix}
\begin{pmatrix}
a \\
\lambda
\end{pmatrix}
= \begin{pmatrix}
\mathbf{w}_1 \\
\mathbf{w}_2
\end{pmatrix}.
\tag{21}
\]

provided \( b_1' \mathbf{1} = 1 \) and \( b_2' \mathbf{1} = 0 \). Solving (21) for \( a \) gives the BLUP in terms of cross-variograms; i.e.,

\[
a' \mathbf{z} = \mathbf{z}' \mathbf{1}^{-1} \mathbf{z} + (w'_1 - \mathbf{z}' \mathbf{1}^{-1} \mathbf{z})(\mathbf{x}' \mathbf{1}^{-1} \mathbf{x})^{-1} \mathbf{x}' \mathbf{1}^{-1} \mathbf{z},
\tag{22}
\]

and the cokriging variance of the BLUP (22) is,

\[
M_a = \mathbf{z}' \mathbf{1}^{-1} \mathbf{z} - (w'_1 - \mathbf{z}' \mathbf{1}^{-1} \mathbf{z})(\mathbf{x}' \mathbf{1}^{-1} \mathbf{x})^{-1} \mathbf{x}' \mathbf{1}^{-1} \mathbf{x}.
\tag{23}
\]

Notice, there are many examples in the literature where the cross-variogram has been defined alternatively as,

\[
2\nu_{ij}(s_k, s_\ell) = \text{cov}(z_i(s_k) - z_i(s_\ell), z_j(s_k) - z_j(s_\ell)).
\tag{24}
\]

Further discussion of \( 2\nu_{ij}(s_k, s_\ell) \) and cokriging is given in sections 5 and 6; its relationship to \( 2\gamma_{ij}(s_k, s_\ell) \) is,

\[
2\nu_{ij}(s_k, s_\ell) = -\gamma_{ij}(s_k, s_\ell) + \gamma_{ij}(s_k, s_\ell) + \gamma_{ij}(s_\ell, s_k) - \gamma_{ij}(s_\ell, s_\ell).
\tag{25}
\]
4. PREDICTION OF $z(s_o)$

Now consider the task of multivariable spatial prediction. We wish to predict $z(s_o)$ from data $z(s_1),...z(s_n)$, where $z(s_o) = [z(s_{i_1}),...z(s_{i_j})]'$ and $(i_1,...,i_j) \in (1,...,m)$. Suppose that $z$ and $z(s_o)$ can be written according to the linear models (2) and (3), respectively, where $\beta$ is unknown. We will again restrict ourselves to linear unbiased predictors: $B'z \sim (r \times m)z(m \times 1)$, where $r \times m$.

Without loss of generality, consider the case of predicting the complete vector $z(s_o)$, so that $r=m$. Uniform unbiasedness is expressed as $E[B'z]=E[z(s_o)]$ for all $\beta$; it is equivalent that $B'z \sim X(s_o)$.

4.1. Prediction with covariances

From section 2.2, $M_B = E[(B'z-z(s_o))(B'z-z(s_o))']$ is the matrix of mean-squared-prediction-errors, or prediction variance matrix. The goal is to find the linear predictor $A'X$ such that $M_B - M_A$ is nonnegative-definite for all $B$. The unbiasedness condition allows $M_B$ to be written as,

$$M_B = E[B'\delta-\delta(s_o)](B'\delta-\delta(s_o))'] = B'\Sigma B - B'\Sigma C - C' B \Sigma z_o.$$

The $A$ that makes $M_B - M_A$ nonnegative-definite for all $B$ is contained in the following set of equations (see result A2 in the appendix, with $E_{oo}=\Sigma_o$, $E_o=C$, $E=\Sigma$, and $\alpha=-1$):

$$\begin{pmatrix} \Sigma & X \\ X' & 0 \end{pmatrix} \begin{pmatrix} A \\ \Lambda \end{pmatrix} = \begin{pmatrix} C \\ [X(s_o)]' \end{pmatrix}.$$

(26)

Notice that Myers (1982) and Quimby et al. (1986) minimize

$$E[(B'z-z(s_o))(B'z-z(s_o))],$$

...
which also yields (26). It is easy to see why:

\[
E(\mathbf{B}'\mathbf{z}-\mathbf{z}(s_0))'\mathbf{B}'(\mathbf{z}-\mathbf{z}(s_0)) = tr E(\mathbf{B}'(\mathbf{z}-\mathbf{z}(s_0))'\mathbf{B}'(\mathbf{z}-\mathbf{z}(s_0)))
\]

\[
= E tr(\mathbf{B}'\mathbf{z}-\mathbf{z}(s_0))'\mathbf{B}'(\mathbf{z}-\mathbf{z}(s_0)) = tr \mathbf{M}_B,
\]

where tr is the trace operator. But since differentiation commutes with trace, minimization results in the same set of equations, (26). However, \( \mathbf{M}_B \) is more natural since it gives the prediction variances and covariances of each variable and avoids a criterion that adds together quantities of potentially different units. Solving (26) for \( \mathbf{A} \) gives the BLUP in terms of covariances,

\[
\mathbf{A}' \mathbf{z} = \mathbf{C}' \mathbf{z}^{-1}\mathbf{z}^t((\mathbf{X}(s_0))' - \mathbf{C}' \mathbf{z}^{-1}\mathbf{z})(\mathbf{X}' \mathbf{z}^{-1}\mathbf{z})^{-1}\mathbf{X}' \mathbf{z}^{-1}\mathbf{z},
\]

and the prediction variance matrix of the BLUP (27) is,

\[
\mathbf{M}_A = \mathbf{S}_0 - \mathbf{C}' \mathbf{z}^{-1}\mathbf{z}((\mathbf{X}(s_0))' - \mathbf{C}' \mathbf{z}^{-1}\mathbf{z})(\mathbf{X}' \mathbf{z}^{-1}\mathbf{z})^{-1}(\mathbf{X}' \mathbf{z}^{-1}\mathbf{z} - \mathbf{C}'(\mathbf{X}(s_0))).
\]

Notice that when \( \mathbf{C} = \mathbf{Z} \), \( \mathbf{X}(s_0) = \mathbf{W} \), and \( \mathbf{S}_0(1,1) \) replaces \( \mathbf{S}_0 \), (27) and (28) reduce to (6) and (7), respectively. Since (26) yields the prediction variances and covariances, joint multivariable prediction regions can be calculated (see section 5).

### 4.2. Relationship to generalized least squares

The generalized least squares estimator of \( \mathbf{g} \) is \( \hat{\mathbf{g}}_{\text{GLS}} = (\mathbf{X}' \mathbf{S}^{-1}\mathbf{X})^{-1}\mathbf{X}' \mathbf{S}^{-1}\mathbf{z} \), so the BLUP (27) can be expressed in terms of \( \hat{\mathbf{g}}_{\text{GLS}} \):

\[
\mathbf{A}' \mathbf{z} = \mathbf{C}' \mathbf{z}^{-1}(\mathbf{z}-\mathbf{S}^{-1}\mathbf{X}\hat{\mathbf{g}}_{\text{GLS}}) + [\mathbf{X}(s_0)]\hat{\mathbf{g}}_{\text{GLS}}.
\]

Also notice that, if the data are Gaussian with known mean, the optimal predictor among all predictors, linear or otherwise, is,

\[
g(\mathbf{z}; s_0) = \mathbb{E}[\mathbf{z}(s_0) | \mathbf{z}] = \mathbf{C}' \mathbf{z}^{-1}(\mathbf{z}-\mathbf{X}\hat{\mathbf{g}}_{\text{GLS}}) + [\mathbf{X}(s_0)]\hat{\mathbf{g}}.
\]

When \( \mathbf{g} \) is unknown, (29) is just (30) with \( \mathbf{g} \) replaced by \( \hat{\mathbf{g}}_{\text{GLS}} \).
4.3. **Relationship to prediction of one variable at a time**

For simplicity, consider the case $m=2$. Let $a_1 z$ be the predictor of $z(s_o)$, where $a_1$ is given by (8); and let $a_2 z$ be the predictor of $z(s_o)$, where $a_2$ is also given by (8). Then, the coefficients $a_1$ and $a_2$ are given by,

$$
\begin{bmatrix}
  a_1 \\
  a_2 \\
  \Lambda_1 \\
  \Lambda_2
\end{bmatrix} = 
\begin{bmatrix}
  \Sigma & X \\
  X' & 0
\end{bmatrix}^{-1} 
\begin{bmatrix}
  c_1 \\
  c_2 \\
  w_1 \\
  w_2
\end{bmatrix},
$$

which gives exactly the same solution as obtained by solving (26). Therefore, the predictor for one variable at a time is identical to the predictor of that same variable in the multivariable predictor.

4.4. **Prediction with variograms and cross-variograms; Multivariable kriging**

Recall $M_B = E(B' \delta - \delta(s_o)) (B' \delta - \delta(s_o))'$, which we called the prediction variance matrix. We wish to find $A X$ such that $M_B - M_A$ is nonnegative-definite for all $B$. The unbiasedness condition allows us to write

$$
M_B = E(B' \delta - \delta(s_o)) (B' \delta - \delta(s_o))'.
$$

Without loss of generality, consider the case $m=2$. Now, suppose that $X_1$ and $X_2$ each have a column of ones, so that $z_1(\cdot)$ and $z_2(\cdot)$ each have an overall mean. This implies, as a part of the unbiasedness conditions, that $b' d = I$, where $D_5 = I_2$.

Let $F_{i,j}(s_k, s_\ell) = [\delta(s_k) - \delta(s_\ell)]^2 = f_{i,j}(s_k, s_\ell)$. Next, define

$$
F_{i,j}(s_k, s_\ell) = [\delta(s_k) - \delta(s_\ell)]^2 = f_{i,j}(s_k, s_\ell).
$$

Let $F_{i,j}$ be the matrix with the $k, \ell$th element $f_{i,j}(s_k, s_\ell)$; $k=1, \ldots, n; \ell=1, \ldots, n.$
and let \( f_{i,j} \) be the vector with the \( k^{th} \) element \( f_{i,j}(s_k^0) \); \( k=1,...,n \). Then, using the relation \( B'\Sigma B = I \), from result A4 in the appendix, we have the algebraic identity,

\[
[B'\Sigma^0-B'\Sigma^0(s_0)][B'\Sigma^0-B'\Sigma^0(s_0)]' = (1/2)B'(-F+DF'+F'D'-DF'D')B. \tag{32}
\]

where,

\[
F = \begin{pmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{pmatrix}, \quad F_{0} = \begin{pmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{pmatrix}, \quad \text{and} \quad F_{oo} = \begin{pmatrix} f_{11}(s_0^0,s_0^0) & f_{12}(s_0^0,s_0^0) \\ f_{21}(s_0^0,s_0^0) & f_{22}(s_0^0,s_0^0) \end{pmatrix}.
\]

Notice that \( F_{oo} \) has zero diagonal elements. Now take the expectation of (32) and use (19) to obtain,

\[
M_B = -B'\Sigma B + G' \Sigma G - G. \tag{33}
\]

where,

\[
G = \begin{pmatrix} \chi_{11} & \chi_{12} \\ \chi_{21} & \chi_{22} \end{pmatrix}, \quad G_0 = \begin{pmatrix} \chi_{11}(s_0^0,s_0^0) & \chi_{12}(s_0^0,s_0^0) \\ \chi_{21}(s_0^0,s_0^0) & \chi_{22}(s_0^0,s_0^0) \end{pmatrix},
\]

and \( \Gamma \) and \( \chi_{ij} \) are defined below (20).

From (33), the \( A \) that makes \( M_B - M_A \) nonnegative-definite for every \( B \) is contained in the following set of equations (see result A2 in the appendix, with \( E_{oo} = G' \), \( E_0 = G \), \( E = \Gamma \), and \( \alpha = 1 \)):

\[
\begin{pmatrix} \Gamma & \chi \\ \chi' & 0 \end{pmatrix} \begin{pmatrix} A \\ \Lambda \end{pmatrix} = \begin{pmatrix} G \\ [X(s_0^0)'] \end{pmatrix}. \tag{34}
\]

provided \( B'\Sigma B = I \). Solving for \( \Lambda \) in (34) gives the BLUP in terms of the cross-variograms:

\[
\Lambda^'_Z = G'\Gamma^{-1}Z + [X(s_0^0)']^{-1}G'\Gamma^{-1}X(X'\Gamma^{-1}X)^{-1}X'\Gamma^{-1}Z. \tag{35}
\]

The multivariable kriging variance matrix of the BLUP (35) is,

\[
M_A = -G' + G'\Gamma^{-1}G - (X'\Gamma^{-1}X)^{-1}(X'\Gamma^{-1}X)^{-1}X'\Gamma^{-1}G - [X(s_0^0)]. \tag{36}
\]
As in (28), (36) can be used to construct joint multivariable prediction regions (see section 5). Also notice that, for the univariate case, $G_0^2 = 0$; cf (17). In (36), $G_0$ has zero diagonals but it is generally a nonzero matrix.

4.5. Relationship to cokriging one variable at a time

Again, for simplicity, consider the case $m=2$. Let $a_1^z$ be the cokriging predictor of $z_1(s_0)$ and $a_2^z$ be the cokriging predictor of $z_2(s_0)$. Then, from (21), the coefficients $a_1$ and $a_2$ are obtained by solving,

$$
\begin{bmatrix}
2_1 & 2_2 \\
\lambda_1 & \lambda_2
\end{bmatrix} =
\begin{bmatrix}
\Gamma & X \\
X' & 0
\end{bmatrix}^{-1}
\begin{bmatrix}
\gamma_1 & \gamma_2 \\
X' & 0
\end{bmatrix}.
$$

But this gives exactly the same solution as obtained by solving (34). This result should not be surprising in light of the result obtained for covariances in Section 4.3.
5. EXAMPLES

Suppose that \( n=5 \) bivariate (\( m=2 \)) random vectors occur at \( s_i = i, \ i=1, \ldots, 5 \) in \( \mathbb{R}^1 \). Let \( E(z) = 0 \) and \( \sigma_i^2 = \sigma_j^2 = 1 \), and define,

\[
\text{cov}(z_i(s_k), z_j(s_k)) \equiv C_{ij}(s_k, s_k) = \begin{cases} 
\rho |k-l| & \text{for } i=j \\
\psi |k-l+\Delta| & \text{for } i \neq j 
\end{cases},
\]

(37)

where \( i, j \in \{1, 2\} \); \( k, \ell \in \{1, \ldots, 5\} \). When \( i=j \), this is a spatial autoregressive process of order 1 (AR(1)). For an AR(1) process, it is well-known that \( \Sigma_{ij} \) is nonnegative-definite for \(-1 \leq \rho \leq 1\). It is not difficult to show that \( \Sigma \) for (37) is also nonnegative-definite for \(-1 \leq \psi \leq 1\) and \(-1 \leq \rho \leq 1\). The parameter \( \Delta \) causes a "shift" in the covariance function \( C_{ij}(s_k, s_k) \) so that \( C_{ij}(s_k, s_k) \neq C_{ij}(s_k, s_k) \), except when \( \Delta = 0 \). Now suppose that \( z(s_k) \) is not observed, and we wish to predict either \( z_i(s_k) \) or \( z_i(s_k) \) from the data collected at the other four locations.

5.1. Optimal linear predictors and prediction intervals

Data were generated from a multivariate Gaussian distribution (\( m=2 \), \( n=5 \)) having \( \Sigma \) with \( \rho=\psi=0.5 \) and \( \Delta=1 \) in the spatial AR(1) process (37). The realization we obtained is given in Table 1. The predicted value \( \hat{z}(s_3) \), using the other four data values and (27), is shown in Fig. 2, along with the actual value \( z(s_3) \). The predictor \( \hat{z}_i(s_3) \) is simply the projection of \( \hat{z}(s_3) \) onto the \( z_i^{\text{th}} \)
axis. The prediction variance matrix $M_A$ is given by (28); it is independent of the data, and for this example is

$$
M_A = \begin{pmatrix}
0.488 & -0.108 \\
-0.108 & 0.488
\end{pmatrix}.
$$

The diagonal elements give the one-at-a-time cokriging prediction variances. For Gaussian data, the 100(1-α)% prediction interval for $\hat{z}(s_0)$ is

$$
\hat{z}(s_0) \pm \phi(\alpha/2)\sqrt{M_{ii}},
$$

where $\phi(\alpha/2)$ is the upper [100(α/2)]th percentile of the standard Gaussian distribution with zero mean and unit variance, and $M_{ii}$ is the $i^{th}$ diagonal element of $M_A$. The 95% prediction intervals for $\hat{z}_1(s_0)$ and $\hat{z}_2(s_0)$ are given in Fig. 2.

Now, when minimizing $E\{(B'\hat{z}-z(s_0))'[B'\hat{z}-z(s_0)]\}$, as suggested by Myers (1982) and Quimby et al. (1986), we do not automatically obtain the prediction covariances of the cokriging predictors. For simultaneous inference on $\hat{z}_i(s_0)$; $i=1,2,...,m$, one reasonable procedure is to use the Bonferroni inequality. Let $R_i$ denote a 100(1-α_i)% prediction interval for $\hat{z}_i(s_0)$. Then,

$$
\Pr[\hat{z}_i(s_0) \in R_i; \ i=1,...,m] \geq 1 - (\alpha_1+\alpha_2+...+\alpha_m),
$$

regardless of the correlation structure. Hence, a conservative 100(1-α)% prediction region for all $z(s_0)$ simultaneously is given by

$$
\hat{z}(s_0) \pm \phi(\alpha/2m)\sqrt{M_{ii}}; \ i=1,2,...,m.
$$

The Bonferroni-corrected 95% prediction region for the example above with $m=2$ is given in Fig. 2.

However, knowledge of prediction covariances, given by $M_A$, allows one to calculate isopleths of constant density from the multivariate normal distribution. The $m$-dimensional ellipsoid of constant density, centered at
\( \hat{z}(s_0) \), and containing 100(1-\( \alpha \))% of probability, is given by all values of \( z \) satisfying,

\[
[z-z(s_0)]'M^{-1}[z-z(s_0)] \leq \chi^2_m(\alpha),
\]

where \( \chi^2_m(\alpha) \) is the upper (100\( \alpha \))th percentile of a chi-squared distribution with \( m \) degrees of freedom. The ellipse for our example is plotted in Fig. 2. It can be seen that, for this particular choice of \( \rho \), \( \psi \), and \( \Delta \), there is negative prediction covariance between \( \hat{z}_1(s_3) \) and \( \hat{z}_2(s_3) \).

5.2. **Comparison of the cross-variograms \( 2\gamma(\cdot) \) with the quantities \( 2\nu(\cdot) \)**

Again, \( m=2 \) and \( n=5 \) data from a multivariate Gaussian distribution were generated from the AR(1) model (37); here \( \rho=\psi=.7 \) and \( \Delta=1 \) were chosen. The realization we obtained is given in Table 2.

| Table 2 |

From (19) and (37), the cross-variogram models,

\[
2\gamma_{ij}(s_k,s_{\ell}) = \begin{cases} 
2\left[1-\rho^{\left|k-\ell \right|}\right] & \text{for } i=j \\
2\left[1-\psi^{\left|k-\ell+\Delta \right|}\right] & \text{for } i\neq j
\end{cases}
\]

are obtained. From (38), the coefficients \( \Delta \) in (35) can be obtained and are given in Table 2.

Further, from (24) and (37), one can show that,
There are many examples in the literature where $\nu_{ij}(s_k, s_\ell)$ have been used in place of $\nu_{ij}(s_k, s_\ell)$, for all $i, j, k$, and $\ell$, in (35) (e.g., Carr and McCallister 1985, Trangmar et al. 1986, Yates and Warrick 1987, Stein et al. 1988, Mulla 1988, and Hoeksema et al. 1989). Let $\mathcal{Y}^\prime$ be the predictor obtained by replacing $\nu_{ij}(s_k, s_\ell)$ with $\nu_{ij}(s_k, s_\ell)$ in (35). From (39) and (35), the coefficients of $\mathcal{Y}$ can be obtained and are also given in Table 2.

Then, from the expression for $M_B^*$ in Section 4.1, we obtain

$$M_Y = \mathcal{V}' \Sigma \mathcal{V} - \mathcal{C}' \mathcal{C} + \mathcal{V}' \Sigma \mathcal{E}_0,$$

and the prediction variance matrix of the optimal predictor,

$$M_A = \mathcal{A}' \Sigma \mathcal{A} - \mathcal{A}' \mathcal{C} \mathcal{C}' \mathcal{A} + \mathcal{A}' \Sigma \mathcal{E}_0,$$

which is equivalent to (36). The predictions $\hat{z}(s_3)$ using both $\mathcal{A}$ and $\mathcal{Y}$, along with their 95% Gaussian prediction regions, are given in Fig. 3. For this realization, predicting $\hat{z}(s_3)$ with $\mathcal{A}$ is closer to the true value than predicting $\hat{z}(s_3)$ with $\mathcal{Y}$, and predictions for $\hat{z}(s_3)$ are about equally close. Also, the 95% prediction region using $\mathcal{A}$ is smaller than that for $\mathcal{Y}$.

One measure of the efficiency of using $2\gamma_{ij}(s_k, s_\ell)$ as opposed to $2\nu_{ij}(s_k, s_\ell)$ is the quantity,

$$\text{Eff}(\rho, \psi; \Delta) = \frac{\det(M_A)}{\det(M_Y)},$$
where $\det(M_A)$ is the determinant of $M_A$ for values of $\rho$ and $\psi$, for a given $\Delta$ in (37). For $\psi=|\rho|$ and $\Delta=1$, $\text{Eff}(\rho,|\rho|;1)$ for various values of $\rho$ is given in Fig. 4. It can be seen that as $|\rho|$ increases, $2\gamma_{ij,k}(s_k,s_{k'})$ becomes more and more efficient as compared to $2\nu_{ij,k}(s_k,s_{k'})$. Journel and Huijbregts (1978, p. 326) note that, for cokriging, $M_b$ given by (4) cannot be minimized using the quantities $2\nu(\cdot)$ except when $C_{ij}(s_k,s_{k'}) = C_{ij}(s_k,s_{k'})$; no such restriction is needed to use the cross-variogram $2\gamma(\cdot)$. For the model given by (37), this restriction only occurs when $\Delta=0$; then both methods give the same results.
6. DISCUSSION AND CONCLUSIONS

In a unified presentation, this paper reviews best linear unbiased spatial prediction of one variable at a time, and extends it to multivariable spatial prediction, based on either covariances or cross-variograms. Relationships to generalized least squares parameter estimation and optimal predictors for Gaussian data are also given. The general linear model, where \( z = X\beta + \delta \), is considered.

Table 3 is offered to clarify the terminology used here. The term "kriging" has been used for spatial prediction with variograms or cross-variograms. Often, \( X_i\mu_i = I \mu_i \), \( i = 1, \ldots, m \) is assumed, which has been called mean stationarity. Kriging, under assumptions of intrinsic stationarity (implying mean stationarity), has been called (ordinary) kriging (Matheron 1971). Kriging under the more general case of \( \mu = X\beta \) has been called universal kriging (Matheron 1969, Huijbregts and Matheron 1971). Here, the same terminology is maintained for (universal) cokriging and multivariable (universal) kriging. Spatial prediction with covariances is called spatial BLUP or multivariable spatial BLUP, regardless of any special form that \( X\beta \) might take (Table 3).

A key result of this paper is that for spatial prediction based on cross-variograms, there must be a vector \( I \) (an overall mean effect) in the column of each \( X_i; i = 1, \ldots, m \), and a 1 in the corresponding column of \( X(s_0) \). This guarantees (11) and \( B'D = I \) (see Section 4.4), which are sufficient conditions to
obtain equations (12), (18), and (32). The expectations of (12), (18), and (32) yields the cross-variograms. Without a vector $\mathbf{l}$ in the column of each $\mathbf{X}_l$, optimal spatial prediction may not be possible with cross-variograms. Notice that intrinsic stationarity for $Z_i(\mathbf{l})$ implies $\mathbf{x}_l^* = \mathbf{l}_2^*.$

We compared the cross-variogram $2\gamma_{ij}(\mathbf{s}_i,\mathbf{s}_j)$ with another candidate, $2\nu_{ij}(\mathbf{s}_i,\mathbf{s}_j).$ In section 5.2, we listed several examples in the literature where $2\nu_{ij}(\mathbf{s}_i,\mathbf{s}_j)$, defined by (24), has been used in the cokriging equations (21). However, $M$ cannot be "minimized" in terms of $2\nu_{ij}(\mathbf{s}_i,\mathbf{s}_j)$ without restrictive assumptions on the symmetry of the covariance structure (e.g., section 5.2). Thus, replacing $2\gamma_{ij}(\mathbf{s}_i,\mathbf{s}_j)$ with $2\nu_{ij}(\mathbf{s}_i,\mathbf{s}_j)$ in the cokriging equations (21) may very well give a nonoptimal linear predictor.

For obtaining the best linear unbiased predictor, we showed the cross-variogram $2\gamma_{ij}(\mathbf{s}_i,\mathbf{s}_j)$, defined by (19), to be the appropriate quantity for the (universal) cokriging equations (21) and the multivariable (universal) kriging equations (34). Although it is clear that $2\gamma_{ij}(\mathbf{s}_i,\mathbf{s}_j)$ is the proper cross-variogram, there has been little practical experience with its use. In the literature, the article by Clark et al. (1989) appears to be the only place where it is used. Clark et al. (1989) mention an additional benefit: each variable does not need to be measured at each location in order to estimate $2\gamma_{ij}(\mathbf{s}_i,\mathbf{s}_j)$, as is required for $2\nu_{ij}(\mathbf{s}_i,\mathbf{s}_j)$.

Note, however, that while Clark et al. (1989) make the usual assumption that $2\gamma_{ij}(\mathbf{s}_i,\mathbf{s}_j)$ is a function depending only on $\mathbf{h}$; i.e., $2\gamma_{ij}(\mathbf{s}_i,\mathbf{s}_j)=2\gamma_{ij}(\mathbf{h})$, where $\mathbf{h}=\mathbf{s}_i-\mathbf{s}_j$, they model it as being symmetric in $\mathbf{h}$. While this is true for $2\gamma_{ij}(\mathbf{h})$, (the usual variogram), it is not necessarily so when $i\neq j$, such as
whenever $\Delta \neq 0$ in (37). In order to implement cokriging and multivariable kriging, valid models for $2y_{ij}(h)$ need to be developed for cases when $i \neq j$; cf. various valid models for $i = j$ (e.g., linear, spherical, exponential, etc.).

It was also shown that the cokriging predictor for one variable at a time is identical to the predictor of that same variable in the multivariable predictor. It is possible to find the spatial BLUP and prediction variance for both $z_1(s_0)$ and $z_2(s_0)$ individually, and then use the Bonferroni inequality to adjust for a simultaneous prediction region (Fig. 2). However, when considering multivariable prediction, the prediction covariances, given in equations (28) and (36), can also be obtained, making construction of the more desirable joint multivariable prediction regions possible (Figures 2 and 3).

Further research is needed into whether using cross-variograms rather than covariances has real advantages for multivariable spatial prediction. In the univariate case, the class of available variogram models is larger than the class of covariance models (Matheron 1971). Similarly, when cross-covariances are defined, an expression for $2y_{ij}(s_k, s_l)$ can always be obtained. Let the cross-covariances be stationary, where $C^{(z)}_{ij}(h) = \text{Cov}[z_i(s), z_j(s)]$ is defined for all $i, j, k,$ and $l$, and $h = s_k - s_l$. Then,

$$2y_{ij}^{(z)}(h) = C^{(z)}_{ii}(h) + C^{(z)}_{jj}(h) - 2C^{(z)}_{ij}(h),$$

so stationary cross-covariances imply stationary cross-variograms. However, now consider a spatial process where $y_i(s_k) = z_i(s_k) + w_i(s_k)$, where $w(\cdot)$ is independent of $z_1(\cdot), z_2(\cdot), \ldots, z_m(\cdot)$; and $w(\cdot)$ is an intrinsically stationary process for which a stationary covariance is undefined. For example, consider a one-dimensional standard Weiner process, where $2y^w(h) = |h|$ and
cov(w(s_i), w(s_j)) = \min(s_i, s_j), which is not a function of h. Then the cross-variogram of y(·) is

$$2\gamma_{ij}(s_i, s_j) = 2\gamma_{ij}(h) + 2\gamma_{ij}(w)(h),$$

which is stationary, but the cross-covariance of y(·) is

$$C_{ij}(y)(s_i, s_j) = C_{ij}(h) + \min(s_i, s_j),$$

which is not stationary. Hence, the class of stationary cross-variogram models is larger than the class of stationary cross-covariance models.

There are also results available from the univariate case that the variogram has better estimation properties than that of covariance function (e.g., Cressie and Grondona 1991). It is not clear whether analogous properties hold for the cross-variogram. For the multivariable case, more work is necessary before one method can clearly be recommended over another.
7. ACKNOWLEDGMENTS

This work was supported by funds from the Whitehall Foundation, the North Atlantic Treaty Organization (NATO), the National Park Service, and the National Science Foundation, grant no. DMS-9001862.
8. LITERATURE CITED


The appendix contains several key results that form the basis of this paper. Consider any m×l vector function \( p(z; s^0) \) of the data \( z \) to be a predictor of the m×l vector \( z(s^0) \). Define, 
\[
\mathcal{M}_p = E [(p(z; s^0) - z(s^0))(p(z; s^0) - z(s^0))'] ,
\]
and define \( q(z; s^0) \) to be an optimal predictor if \( \mathcal{M}_p - \mathcal{M}_q \) is nonnegative-definite for every \( p(z; s^0) \).

**Result A1**

An optimal predictor of \( z(s^0) \) is \( q(z; s^0) = E(z(s^0) | z) \).

**Proof**

For simplicity, write \( p(z; s^0) = p, \ z(s^0) = z^0, \text{ and } E[z(s^0) | z] = E[z | z] \).

Then,
\[
\mathcal{M}_p = E(p - z^0)(p - z^0)' ,
\]
\[
= E(p - z^0 + E(z^0 | z) - z^0)(p - z^0 + E(z^0 | z) - z^0)' ,
\]
\[
= E((E(z^0 | z) - z^0)' + E((E(z^0 | z) - z^0)(E(z^0 | z) - z^0))' + E((p - E(z^0 | z))(p - E(z^0 | z))')) ,
\]
where,
\[
R_1 = E((E(z^0 | z) - z^0)(E(z^0 | z) - z^0))' ,
\]
\[
R_2 = E((E(z^0 | z) - z^0)(p - E(z^0 | z)))' + E((p - E(z^0 | z))(E(z^0 | z) - z^0))' ,
\]
\[
R_3 = E((E(z^0 | z) - E(z^0))(p - E(z^0 | z))') + E((p - E(z^0 | z))(E(z^0 | z) - E(z^0))') .
\]
It is not difficult to show that \( R_1 + R_2 + R_3 = 0 \), so from (A1.1),
\[
\mathcal{M}_p = E(z^0 | z)' E((p - E(z^0 | z))(p - E(z^0 | z)))' ,
\]
where,

\[ \Sigma_{Z_0} = E((z_0 - E(z_0))(z_0 - E(z_0))^\prime) \]

and

\[ \Sigma_{E(z_0 | z)} = E((E(z_0 | z) - E(z_0))(E(z_0 | z) - E(z_0))^\prime). \]

If \( g = E(z_0 | z) \), then,

\[ M_p - M_q = E((p - E(z_0 | z))(p - E(z_0 | z))^\prime), \]

for every \( p \neq q \), which is always nonnegative-definite.

**Result A2**

Define \( M_B = -\alpha B E + \alpha E E \) where \( \alpha \in (-1, 1) \), \( E \) is symmetric, and \( B \) satisfies \( B^\prime X = X_0 \). Then, the matrix \( A \) that makes \( M_B - M_A \) nonnegative-definite for every \( B \) is obtained by solving

\[
\begin{pmatrix}
E & X \\
X' & 0
\end{pmatrix}
\begin{pmatrix}
A \\
A
\end{pmatrix}
= \begin{pmatrix}
E_0' \\
X_0'
\end{pmatrix}.
\]

That is,

\[ A = E^{-1} + E^{-1} X (X' E^{-1} X)^{-1} (X_0' - X' E^{-1} E_0), \]

provided all inverses exist.

**Proof**

In order for \( M_B - M_A \) to be nonnegative-definite for every \( B \), \( y' (M_B - M_A) y \geq 0 \) for every \( y \). That is, \( y' M_B y \geq y' M_A y \) for all \( y \). Therefore, minimize \( y' M_B y \) with respect to \( B \), subject to \( B^\prime X = X_0 \). First, fix \( y \), and write

\[ \phi(B) = y' M_B y - 2 \text{tr}(y y' A (X' B - X_0')). \]

The matrix \( A \) consists of Lagrange multipliers for the constraints \( B^\prime X = X_0 \). Since \( (X' B - X_0') = 0 \), we can scale \( A \) as we please, so we premultiply \( A \) by \(-2 y y' \). Then, following the guidelines of Magnus and Neudecker (1988, p. 175), the
The differential of \( \varphi(\mathbf{B}) \) is,
\[
d\varphi(\mathbf{B}) = 2\alpha\mathbf{y}'\mathbf{E}_0' \, d(\mathbf{B})\mathbf{y} - \alpha\mathbf{y}' \, d(\mathbf{B})' \mathbf{E} \mathbf{y} - \alpha\mathbf{y}' \, \mathbf{B}' \mathbf{E} \mathbf{d}(\mathbf{B})\mathbf{y} - 2\alpha tr\left(\mathbf{y} \mathbf{A}' \mathbf{X}' d(\mathbf{B})\mathbf{X}' \right).
\]  
(A2.1)

Next, take the vec of the differential of (A2.1), and commute the vec and differential:
\[
d\text{vec} \varphi(\mathbf{B}) = 2\alpha \left(\mathbf{y} \otimes \mathbf{E}_0\right) \text{vec} \mathbf{B} - \alpha \left(\mathbf{y} \otimes \mathbf{B} \otimes \mathbf{E}_0\right) \text{Kvec} \mathbf{B}
\]
\[
- \alpha \left(\mathbf{y} \otimes \mathbf{B}' \otimes \mathbf{E}\right) \text{dvec} \mathbf{B} - 2\alpha (\text{vec} \mathbf{X}' \mathbf{A} \mathbf{y}') \text{dvec} \mathbf{B},
\]  
(A2.2)

where \( \otimes \) is the Kronecker product (Magnus and Neudecker 1988, p. 27), vec is the vec operator (Magnus and Neudecker 1988, p. 30), and \( \mathbf{K} \) is a commutation matrix (Magnus and Neudecker 1988, p. 46). Simplification of (A2.2) gives,
\[
d\text{vec} \varphi(\mathbf{B}) = 2\alpha \mathbf{y}' \left(\mathbf{y} \otimes \mathbf{E}_0\right) \text{vec} \mathbf{B} - 2\alpha \mathbf{y}' \left(\mathbf{y} \otimes \mathbf{B} \otimes \mathbf{E}_0\right) \text{dvec} \mathbf{B} - 2\alpha (\text{vec} \mathbf{X}' \mathbf{A} \mathbf{y}') \text{dvec} \mathbf{B}.
\]  
(A2.3)

Then,
\[
\frac{\partial \varphi(\mathbf{B})}{\partial (\text{vec} \mathbf{B})} = 2\alpha \mathbf{y} \otimes \mathbf{E}_0 - 2\alpha \mathbf{y} \otimes \mathbf{B} \otimes \mathbf{E}_0 - 2\alpha (\text{vec} \mathbf{X}' \mathbf{A} \mathbf{y}').
\]  
(A2.4)

Removing the vec operator from (A2.4) yields,
\[
\frac{\partial \varphi(\mathbf{B})}{\partial \mathbf{B}} = 2\alpha \mathbf{E}_0 \mathbf{y}' - 2\alpha \mathbf{E} \mathbf{y} \mathbf{y}' - 2\alpha \mathbf{X}' \mathbf{A} \mathbf{y}'.
\]  
(A2.5)

But, (A2.5) implies that \( \mathbf{E} \mathbf{A} + \mathbf{X} \mathbf{A} = \mathbf{E}_0 \), regardless of \( \mathbf{y} \) and \( \alpha \). Add the constraint \( \mathbf{X}' \mathbf{A} = \mathbf{X}' \mathbf{X}_0 \) to obtain,
\[
\begin{bmatrix}
\mathbf{E} & \mathbf{X} \\
\mathbf{X}' & 0
\end{bmatrix}
\begin{bmatrix}
\mathbf{A} \\
\mathbf{X}_0
\end{bmatrix}
= \begin{bmatrix}
\mathbf{E}_0 \\
\mathbf{X}_0
\end{bmatrix}.
\]

Now, solving for \( \mathbf{A} \) in \( \mathbf{E} \mathbf{A} + \mathbf{X} \mathbf{A} = \mathbf{E}_0 \), we obtain,
\[
\mathbf{A} = \mathbf{E}^{-1}(\mathbf{E}_0 - \mathbf{X} \mathbf{A}).
\]  
(A2.6)
Then substituting (A2.6) into the constraints $X'\Delta = X'_0$, we obtain $X'E^{-1}(E_0 - \Delta X) = X'_0$, or,

$$\Delta = (X'E^{-1}X)^{-1}(X'_0 X'E^{-1}E_0 - X'_0).$$  \hspace{1cm} (A2.7)

Finally, substituting $\Delta$ given by (A2.7) into (A2.6) yields,

$$\Delta = E^{-1}E_0 - E^{-1}X(X'E^{-1}X)^{-1}(X'E^{-1}E_0 - X'_0).$$  \hspace{1cm} \qed

**Result A3**

We can write

$$[b'\delta - \delta(s_0)]^2 = \sum_{k=1}^{n} \sum_{\ell=1}^{n} \frac{b_k}{2} [\delta(s_k) - \delta(s_\ell)]^2 + 2 \sum_{k=1}^{n} \frac{b_k}{2} [\delta(s_k) - \delta(s_0)]^2,$$

if,

$$\sum_{k=1}^{n} b_k = 1,$$

and only if,

$$\sum_{k=1}^{n} b_k = 1 \text{ or } \sum_{k=1}^{n} b_k \delta^2(s_k) = \sum_{k=1}^{n} b_k \delta^2(s_k).$$

**Proof**

First, expand the square so,

$$[\delta(s_0) - b'\delta]^2 = \delta^2(s_0) - 2 \sum_{k=1}^{n} b_k \delta(s_0) \delta(s_k) + \sum_{k=1}^{n} \sum_{\ell=1}^{n} b_k b_\ell \delta(s_k) \delta(s_\ell).$$  \hspace{1cm} (A3.1)

By adding and subtracting similar terms, (A3.1) can be written:

$$[\delta(s_0) - b'\delta]^2 = \sum_{k=1}^{n} b_k \delta^2(s_k) - 2 \sum_{k=1}^{n} b_k \delta(s_0) \delta(s_k) + \sum_{k=1}^{n} b_k \delta(s_0)$$
\[
- \frac{1}{2} \left[ \sum_{k=1}^{n} \sum_{\ell=1}^{n} b_k \ell \delta^2(s_k) - 2 \sum_{k=1}^{n} \sum_{\ell=1}^{n} b_k \ell \delta(s_k) \delta(s_\ell) + \sum_{k=1}^{n} \sum_{\ell=1}^{n} b_k \ell \delta^2(s_\ell) \right] \\
+ \delta^2(s_0) - \sum_{k=1}^{n} b_k \delta^2(s_k) - \sum_{k=1}^{n} b_k \delta^2(s_k) + \sum_{k=1}^{n} \sum_{\ell=1}^{n} b_k \ell \delta^2(s_k).
\]

(A3.2)

But (A3.2) is equal to,
\[
2 \sum_{k=1}^{n} \frac{b_k (\delta(s_k) - \delta(s_0))^2}{2} - \sum_{k=1}^{n} \sum_{\ell=1}^{n} \frac{b_k \ell (\delta(s_k) - \delta(s_\ell))^2}{2} + r,
\]

(A3.3)

where,
\[
r = \delta^2(s_0) - \sum_{k=1}^{n} b_k \delta^2(s_k) - \sum_{k=1}^{n} b_k \delta^2(s_k) + \sum_{k=1}^{n} \sum_{\ell=1}^{n} b_k \ell \delta^2(s_k).
\]

Now suppose,
\[
\sum_{k=1}^{n} b_k = c;
\]

then,
\[
r = \delta^2(s_0) - c \delta^2(s_0) - \sum_{k=1}^{n} b_k \delta^2(s_k) + c \sum_{k=1}^{n} b_k \delta^2(s_k).
\]

Thus, \( r = 0 \) if \( c = 1 \) and only if \( c = 1 \) or \( \sum_{k=1}^{n} b_k \delta^2(s_k) = \delta^2(s_0) \).
Result A4.

Let

\[ \mathbf{\delta} = \begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_m \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ \vdots & \vdots & & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nm} \end{pmatrix}, \quad \mathbf{D} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}, \quad \mathbf{f}_{ij} = \begin{pmatrix} f_{ij}(s_i, s_j) \\ f_{ij}(s_j, s_i) \end{pmatrix}, \quad \mathbf{F}_{ij} = \begin{pmatrix} f_{ij}(s_i, s_j) \\ f_{ij}(s_i, s_j) \end{pmatrix}, \quad \mathbf{F}_{00} = \begin{pmatrix} f_{11}(s_0, s_0) & f_{12}(s_0, s_0) & \cdots & f_{1m}(s_0, s_0) \\ f_{21}(s_0, s_0) & f_{22}(s_0, s_0) & \cdots & f_{2m}(s_0, s_0) \\ \vdots & \vdots & & \vdots \\ f_{m1}(s_0, s_0) & f_{m2}(s_0, s_0) & \cdots & f_{mm}(s_0, s_0) \end{pmatrix} \]

where \( \mathbf{\delta}_i \) is given in (2), \( f_{ij}(s_i, s_j) \) is defined in (31), and \( \mathbf{1}_n \) is an \((nx1)\) vector with each element 1. For ease of notation, suppose that all \( m \) variables occur at all \( n \) spatial locations, so \( \mathbf{\delta}_i \) is \((nx1)\) for all \( i; i=1, \ldots, m \). Assume that \( \mathbf{B}'D=\mathbf{I}_m \), where \( \mathbf{I}_m \) is the \((mxm)\) identity matrix. Then,

\[ \mathbf{B}'(\mathbf{\delta}-\mathbf{\delta}(s_0))\mathbf{B}' = (1/2)\mathbf{B}'(\mathbf{\delta}+\mathbf{D}\mathbf{\delta}-\mathbf{\delta}(s_0))\mathbf{B}. \]

Proof

Define \( \mathbf{1}_m \) as an \((mx1)\) vector with each element 1, and \( \mathbf{1}_{mn} \) as an \((mnx1)\) vector with each element 1. Notice that,

\[ \mathbf{F} = \delta_1' \mathbf{1}_{mn} \mathbf{1}_{mn}' - 2\mathbf{\delta}' \mathbf{1}_{mn} - \mathbf{1}_{mn} \mathbf{\delta}' \mathbf{1}_{mn}, \quad \mathbf{F}_{0} = \delta_1' \mathbf{1}_{mn} \mathbf{1}_{mn}' - 2\mathbf{\delta}' \mathbf{1}_{mn} \mathbf{1}_{mn}' + \mathbf{1}_{mn} \mathbf{\delta}' \mathbf{1}_{mn} = (\mathbf{\delta}+\mathbf{D}\mathbf{\delta}-\mathbf{\delta}(s_0))', \quad \mathbf{F}_{00} = \mathbf{\delta}(s_0)' \mathbf{\delta}(s_0) \mathbf{1}_{m} \mathbf{1}_{m}' - 2\mathbf{\delta}(s_0)' \mathbf{\delta}(s_0) \mathbf{1}_{m} \mathbf{1}_{m}' + \mathbf{1}_{m} \mathbf{\delta}(s_0) \mathbf{1}_{m} \mathbf{\delta}(s_0) \mathbf{1}_{m}' \]
where \( \odot \) is the Hadamard product (Magnus and Neudecker 1988, p. 45).

Substituting for the \( F_i \)'s yields,

\[
-F+DF'_{0}+F_{0}D'-DF_{00}D' = 2\dot{\sigma}' - 2D[\dot{\sigma}(s_{0})]D' - 2\dot{\sigma}[\ddot{\sigma}(s_{0})]'D' + 2D[\dot{\sigma}(s_{0})]D[\dot{\sigma}(s_{0})]'D',
\]

where all Hadamard products cancel because \( D_{11} = 1 \). Therefore,

\[
(1/2)B'(-F+DF'_{0}+F_{0}D'-DF_{00}D')B = [B']\dot{\sigma}[\ddot{\sigma}(s_{0})]'B[\dot{\sigma}(s_{0})]',
\]

since \( B'\dot{\sigma} = I_m \). \( \blacksquare \)
Table 1. Realization of 5 random vectors with covariance structure given by (37) in the text with $\rho=\psi=.5$ and $\Delta=1$.

<table>
<thead>
<tr>
<th>$k$</th>
<th>$z_1(s_k)$</th>
<th>$z_2(s_k)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.130</td>
<td>0.475</td>
</tr>
<tr>
<td>2</td>
<td>0.464</td>
<td>0.213</td>
</tr>
<tr>
<td>3</td>
<td>1.340</td>
<td>-0.625</td>
</tr>
<tr>
<td>4</td>
<td>0.783</td>
<td>0.079</td>
</tr>
<tr>
<td>5</td>
<td>2.015</td>
<td>0.098</td>
</tr>
</tbody>
</table>
Table 2. Realization of 5 random vectors with covariance structure given by (37) in the text with $\rho=\psi=.7$ and $\Delta=1$. Also given are the coefficients for $\Lambda$ in (34), using $2\gamma_{ij}(s_k,s_l)$, and $V$, by substituting $2\nu_{ij}(s_k,s_l)$ for $2\gamma_{ij}(s_k,s_l)$ for all $i,j,k,l$ in (34). Predicted values are in brackets at the bottom of the table, along with actual values.

<table>
<thead>
<tr>
<th>random variable</th>
<th>value</th>
<th>$\Lambda$ col. 1</th>
<th>$\Lambda$ col. 2</th>
<th>$V$ col. 1</th>
<th>$V$ col. 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$z_1(s_1)$</td>
<td>1.965</td>
<td>.156</td>
<td>-.040</td>
<td>.012</td>
<td>-.183</td>
</tr>
<tr>
<td>$z_1(s_2)$</td>
<td>1.584</td>
<td>.276</td>
<td>.067</td>
<td>.488</td>
<td>.183</td>
</tr>
<tr>
<td>$z_1(s_3)$</td>
<td>1.186</td>
<td>.555</td>
<td>-.202</td>
<td>.488</td>
<td>.183</td>
</tr>
<tr>
<td>$z_1(s_4)$</td>
<td>1.186</td>
<td>.012</td>
<td>-.005</td>
<td>.012</td>
<td>-.183</td>
</tr>
<tr>
<td>$z_1(s_5)$</td>
<td>0.758</td>
<td>-.005</td>
<td>.012</td>
<td>-.183</td>
<td>.012</td>
</tr>
<tr>
<td>$z_2(s_1)$</td>
<td>1.343</td>
<td>-.202</td>
<td>.555</td>
<td>.183</td>
<td>.488</td>
</tr>
<tr>
<td>$z_2(s_2)$</td>
<td>1.204</td>
<td>.607</td>
<td>.276</td>
<td>.183</td>
<td>.488</td>
</tr>
<tr>
<td>$z_2(s_3)$</td>
<td>1.469</td>
<td>-.400</td>
<td>.156</td>
<td>-.183</td>
<td>.012</td>
</tr>
<tr>
<td>$z_2(s_4)$</td>
<td>1.172</td>
<td>[1.291]</td>
<td>[1.454]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$z_2(s_5)$</td>
<td>0.444</td>
<td>[1.243]</td>
<td>[1.202]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3. Multivariable spatial prediction as covered in this paper. Beneath each $C(h)$ indicates prediction terminology when using covariances; beneath each $\gamma(h)$ indicates prediction terminology when using variograms or cross-variograms.

Multivariable Spatial Prediction

<table>
<thead>
<tr>
<th>Data</th>
<th>Predict</th>
<th>$z(s_0)$</th>
<th>$z(s_0)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m = 1$</td>
<td>$C(h)$</td>
<td>$\gamma(h)$ (Universal) Blup</td>
<td>Multivariable Spatial Blup</td>
</tr>
<tr>
<td></td>
<td>Spatial</td>
<td>kriging</td>
<td>Multivariable Kriging</td>
</tr>
<tr>
<td>m $\geq$ 2</td>
<td>$C(h)$</td>
<td>$\gamma(h)$ (Universal) Cokriging</td>
<td>$C(h)$</td>
</tr>
</tbody>
</table>
Figure 1. Example of spatial locations $s_1$ to $s_3$, at which data are collected, and $s_0$, the spatial location to be predicted, in a two-dimensional spatial domain $D$. 
Figure 2. Vector prediction of a spatial AR(1) process. The predicted value is obtained from (27), using the data in table 1. The confidence intervals and regions are obtained from (28). (36) and (37) give identical results.
true realization of random vector
predicted realization using \( \gamma \)
predicted realization using \( \nu \)
Gaussian simultaneous prediction region using \( \gamma \)
Gaussian simultaneous prediction region using \( \nu \)

Figure 3. Vector prediction of a spatial AR(1) process using \( z(t)'(s_k, s_{t'}) \) versus \( 2\nu_{i,j}(s_k, s_{t'}) \) in equations (34), along with their respective prediction regions. The model is \( \rho = 0.7 \) and \( \Delta = 1 \) in (38) and (39).
Efficiency as a function of $\rho$ for $2\gamma_{ij}(s_{k},s_{l})$ versus $2\nu_{ij}(s_{k},s_{l})$. 

Figure 4. Efficiency as a function of $\rho$ for $2\gamma_{ij}(s_{k},s_{l})$ versus $2\nu_{ij}(s_{k},s_{l})$. 
GENERAL SUMMARY

This dissertation addresses several problems with estimation and prediction for spatial ecological data.

In Part I, using (2), typical model assumptions used for nested ANOVA and variogram analysis are given. The variogram under aggregation, $2\gamma_n(h;m)$, is shown to be a mathematical link between several statistics proposed for spatial pattern analysis of ecological data: nested analysis of variance (nested ANOVA), two term local variance (TTLV), and paired quadrate variance (PQV). Assuming intrinsic stationarity, the expected values of nested ANOVA and TTLV can be written as functions of $2\gamma_n(h;m)$. The quantity $(m/2)2\gamma_n(1;m)$ is of interest since it is invariant to aggregation $m$ when there is spatial independence among the random variables. It was shown that there are two estimators of $(m/2)2\gamma_n(1;m)$, namely, $(m/2)2\gamma_n(1;m)$ [=$TTLV(m)$] and $(m/2)2\gamma_n(1;m)$. Since $(m/2)2\gamma_n(1;m)$ is a function of PQV, it is clear that TTLV can be approximated from PQV.

For estimation of $(m/2)2\gamma_n(1;m)$, a simulation study for a first-order autoregressive model indicates that $TTLV(m)$ can be recommended when there is negative autocorrelation, and $(m/2)2\gamma_n(1;m)$ is recommended when there is positive autocorrelation.

In Part II, estimation of average patch size (APS) for transect data is considered, where the effects of three factors: 1) the signal-to-noise ratio; 2) the expected sizes of the patches relative to the plot size; and 3) the distribution of patch sizes, on three methods: 1) two term local variance
(TTLV); 2) a moving two-sample t-test (MT); and 3) a Bayesian approach with simulated annealing (BSA), are compared in a 3x2x3 factorial simulation experiment. All three factors were important in one or more of the methods. Overall, the signal-to-noise ratio was the only factor to affect all three methods, and seems to be the most important.

Of the three methods, BSA is the method that is recommended. The strength of BSA is that it estimates APS well, even when the data contain many small patch sizes, as judged by the small range of deviations from true APS. Although it is necessary to estimate some nuisance parameters when using BSA, it was shown how that could be handled easily in the grassland example. Some trial and error may be necessary when selecting the value of the smoothing parameter, but it is possible to check that parameter estimates are realistic in the Bayesian framework by examining residuals for normality and spatial independence. Additionally, since BSA estimates each \( \mu(i) \), \( i=1,\ldots,n \), it allows mean structure summaries beyond APS. The BSA method was adapted from image analysis, and has a straightforward generalization to more than two phase types, and to two-dimensional space.

In Part III, the simultaneous prediction of several variable types for a vector-valued process is considered, and results are extended from scalar prediction. Best linear unbiased prediction using variograms and covariances are reviewed. One of the key results is that an overall mean effect is necessary in the model for each variable type; optimal spatial prediction with variograms and cross-variograms may not be possible otherwise.

The cross-variogram \( 2\gamma_{ij}(\cdot) \) was compared with another candidate, \( 2\nu_{ij}(\cdot) \).
It was shown that replacing $2\gamma_{ij}(\bullet)$ with $2\nu_{ij}(\bullet)$ in the cokriging equations may very well give a nonoptimal linear predictor. The natural criterion for developing multivariable spatial prediction was shown to be as follows: find $g(z; s_0)$ such that

$$E\{[E(5::o)-z(s)]\{E(z::o)-z(s)\}']-E\{[g(z; s_0)-z(s)]\{g(z; s_0)-z(s)\}']$$

is nonnegative-definite for all predictors $g(z; s_0)$. This criterion yields the prediction covariances, and so allows construction of simultaneous prediction regions. It was also shown that the cokriging predictor for one variable at a time is identical to the predictor of that same variable in the multivariable predictor.

In conclusion, this dissertation takes the linear model (2), and uses it as a unifying mathematical structure for studying several problems in estimation and prediction for spatial ecological variables.
LITERATURE CITED


ACKNOWLEDGMENTS

This work was made possible by funds from the Whitehall Foundation, the North Atlantic Treaty Organization (NATO), the National Park Service, and the National Science Foundation, grant number DMS-9001862.

I would like to express my sincerest appreciation to my Co-major Professors, Dr. David Glenn-Lewin and Dr. Noel Cressie. From them, I have learned a great deal, and I thank them for their patience and support. Also, I would like to thank my committee members, Dr. Tom Jurik, Dr. Paul Hinz, and Dr. William Clark. Moreover, I would like to thank the faculty and graduate students of the Botany and Statistics departments for making my program at Iowa State University both informative and pleasurable. Finally, I would like to thank my wife, Mary; she tolerated all of this with good humor.