ESTIMATION OF SPATIAL VARIABILITY
PART I: NONPARAMETRIC VARIOGRAM ESTIMATION

by

Michael Stein
Stanford University

TECHNICAL REPORT NO. 73
AUGUST 1984

STUDY ON STATISTICS AND ENVIRONMENTAL
FACTORS IN HEALTH (SIMS)

PREPARED UNDER A GRANT TO SIMS FROM

ENVIRONMENTAL PROTECTION AGENCY (EPA)
SLOAN FOUNDATION
NATIONAL SCIENCE FOUNDATION (NSF)

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
ESTIMATION OF SPATIAL VARIABILITY
PART I: NONPARAMETRIC VARIOGRAM ESTIMATION

by

Michael Stein
Stanford University

TECHNICAL REPORT NO. 73
AUGUST 1984

STUDY ON STATISTICS AND ENVIRONMENTAL
FACTORS IN HEALTH (SIMS)

PREPARED UNDER A GRANT TO SIMS FROM

ENVIRONMENTAL PROTECTION AGENCY (EPA)
SLOAN FOUNDATION
NATIONAL SCIENCE FOUNDATION (NSF)

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
# TABLE OF CONTENTS

I.0 **INTRODUCTION**

I.1 Spatial and Spatio-Temporal Processes ........................................... 1
  I.1.1 Some problems in spatial analysis ........................................ 1
  I.1.2 The basic model .................................................................. 1
  I.1.3 More about variograms and interpolation .......................... 4

I.2 Other Approaches to Interpolation and Analysis of Spatial Data. ... 8

I.3 Summary of Contents ................................................................. 12
  I.3.1 Estimation of variograms .................................................. 12
  I.3.2 Regression and co-kriging ............................................... 18
  I.3.3 A model for spatial processes observed repeatedly
    in time ........................................................................... 19
  I.3.4 Sequential siting .......................................................... 21
  I.3.5 Other results ............................................................... 21

I.4 Notation ..................................................................................... 23

II.0 **NON-PARAMETRIC VARIOGRAM ESTIMATION**

II.1 Statement of Problem ................................................................ 25

II.2 A Generalization of Matheron's Estimator ............................. 26

II.3 Conditions for Existence of Solution to Unbiasedness
  Constraints ............................................................................. 30
  II.3.1 Necessary and sufficient condition for a solution .......... 30
  II.3.2 The case of linear drift .................................................. 31

II.4 The Case of Uncorrelated Pairs .............................................. 32
  II.4.1 An $s^2$-like estimator of $Y(d)$ .................................. 32
  II.4.2 Using estimators based on residuals when pairs are
    correlated ........................................................................ 34
  II.4.3 Repeated observations in time ...................................... 34

II.5 Irregularly Spaced Data ........................................................... 37

II.6 Choosing a Form for the Mean Function ................................... 38

II.7 Misspecification of the Relationship Between Covariates and
  the Dependent Variable ............................................................. 40
  II.7.1 Alternatives to the regression model .............................. 40
  II.7.2 Misspecification of regression equation ....................... 41

II.8 Application to Landsat Data ..................................................... 44

III.0 **PARAMETRIC VARIOGRAM ESTIMATION**

III.1 Introduction ............................................................................ 67

III.2 Single Scalar Parameter .......................................................... 70
  III.2.1 Maximum likelihood and modified maximum
    likelihood estimators ....................................................... 71
### III.2.2 Confidence intervals ........................................... 71
### III.2.3 Implications of the previous section ....................... 73

#### III.3 Estimation of Parameters: General Case .................... 76

- III.3.1 Maximum likelihood and modified maximum likelihood estimators ........................................... 76
- III.3.2 Generalized minimum norm quadratic estimators ........... 79
- III.3.3 Another parametric estimator ................................. 83

#### III.4 Asymptotics ....................................................... 85

- III.4.1 A consistency result which does not depend on the locations of the observations ................. 85
- III.4.2 A discussion of asymptotic properties with multi-parameter models ............................. 89

#### III.5 Effects of Misspecification of the Mean Function .......... 93

#### III.6 Local Parametric Estimators ................................ 95

#### III.7 Choosing a Parametric Model for the Variogram .......... 100

#### III.8 Analysis of Landsat Data ...................................... 104

### IV.0 A COMPARISON OF CO-KRIGING AND REGRESSION MODELS

#### IV.1 Introduction .................................................... 115

#### IV.2 Comparison of Probabilistic Properties of the Two Models ... 116

#### IV.3 Asymptotic Behavior of Estimators Based on the Regression Model When it is Not Valid .............. 120

- IV.3.1 Model with repeated observations in time ............... 120
- IV.3.2 Mean changing in time ..................................... 122
- IV.3.3 Mean fixed in time ........................................ 125
- IV.3.4 Interpolation based on the regression model ............ 129

### V.0 KRIGING WITH REPEATED OBSERVATIONS IN TIME AND AN UNKNOWN, VARYING SPATIAL MEAN FUNCTION

#### V.1 Description of Model ........................................... 132

#### V.2 Interpolation .................................................... 134

- V.2.1 Mean changing in time ..................................... 134
- V.2.2 Mean fixed in time ........................................ 138

#### V.3 Estimation of Time Trends ..................................... 142

#### V.4 Estimation of the Variograms ................................ 145

### VI.0 THE EFFECT OF SEQUENTIAL SITING ON ESTIMATION

#### VI.1 Description of the Problem ................................... 148

#### VI.2 A Theorem on the Effect of Sequential Siting .............. 148

### APPENDIX A: An Asymptotic Result for Estimation of the Variogram in a One-Dimensional Process .......... 154
| APPENDIX B: Use of the Estimated Covariance Structure Instead of the True Covariance Structure | 160 |
| APPENDIX C: Proof of Theorem III.1 | 168 |
| APPENDIX D: An Example Where the Maximum Likelihood Estimator of the Covariance Structure is Not Determined by the Variogram Model | 170 |
| APPENDIX E: Proofs of Results from Section III.4 | 172 |
| APPENDIX F: Some Notes on Spatio-Temporal Variograms | 176 |
| References | 181 |
TABLE OF FIGURES

II.1: Plot of visible frequencies at Site 1 .......... 45
II.2: Plot of visible frequencies at Site 2 .......... 46
II.3: Plot of infrared frequencies at Site 1 .......... 47
II.4: Plot of infrared frequencies at Site 2 .......... 48
II.5: Unbiased non-parametric semi-variogram estimates for
      visible frequencies .......................... 50
II.6: Unbiased non-parametric semi-variogram estimates for
      infrared frequencies .......................... 51
II.7: Unbiased non-parametric semi-variogram estimates by
      subregion - Site 1 - Visible frequencies ......... 53
II.8: Unbiased non-parametric semi-variogram estimates by
      subregion - Site 1 - Infrared frequencies ......... 54
II.9: Unbiased non-parametric semi-variogram estimates by
      subregion - Site 1 - Visible frequencies ......... 55
II.10: Unbiased non-parametric semi-variogram estimates by
      subregion - Site 1 - Infrared frequencies ......... 56
II.11: Interpolators based on four nearest neighbors ..... 59
II.12: Interpolators based on eight nearest neighbors .... 59
II.13: Prediction of interpolation error of eight nearest
      neighbor "full model" interpolator .............. 61
II.14: Comparison of semi-variogram estimators .......... 64

III.1: Set of sites in which second pair of points provides little
       additional information about $\gamma(d)$ ........... 75
III.2: Set of sites in which second pair of points "doubles"
       information about $\gamma(d)$ .................... 75
III.3: Pairs of covariance functions in $\mathbb{R}^2$ which do and do not
       satisfy Equation (41) .......................... 87
III.4: Parametric estimates of $\theta$ by subregion .......... 106
III.5: Comparison of semi-variogram estimators .......... 106
III.6: Performance of non-parametric semi-variogram estimator ... 108
III.7: Rescaled non-parametric estimates of $\gamma^*_V(1)$ by subregion .. 108
III.8: Performance of non-parametric semi-variogram estimator ... 110
III.9: Performance of non-parametric semi-variogram estimator ... 110
III.10: Parametric estimates of $\theta$ by subregion .......... 111
III.11: Comparison of semi-variogram estimators .......... 113

vi
Note: For the purpose of making technical reports of manageable proportions, this thesis has been divided into three parts. In each part, the entire introductory chapter is included. Part I contains Chapter II, "Non-Parametric Variogram Estimation." Part II contains Chapter III, "Parametric Variogram Estimation," and Appendices A through E. Part III, entitled "Some Models for Spatial and Spatio-Temporal Processes," contains Chapters IV-VI and Appendix F.
CHAPTER I: Introduction

I.1 Spatial and Spatio-Temporal Processes

I.1.1 Some problems in spatial analysis

Analysis of processes which take on values in space or in space and time is a rapidly growing field in statistics. Statistical models of spatial phenomenon are used in many fields, including meteorology, air and soil pollution, forestry, agriculture, geology, and satellite imagery. For example, we may be interested in estimating mineral concentrations in the ground based on some set of core samples. Or we may be interested in identifying ground features based on photographs from satellites. Or we may be interested in changes in acid deposition in a region from one year to the next. We will consider the properties and uses of some possible stochastic models of spatial and spatio-temporal phenomena.

I.1.2 The basic model

In this work, we will focus mainly on problems which relate to doing point interpolation of a random field. That is, having observed some quantity \( z(\cdot) \) at \( x_1, x_2, \ldots, x_n \), we want to estimate \( z(x_0) \). While \( x \) will usually just indicate a place, in theory, it could also indicate a time, or a place and a time. The basic model for \( z(\cdot) \) will be

\[
z(x) = m(x) + e(x).
\]

The mean function, \( m(\cdot) \), will have some specific structure, depending on the problem. It is often just a constant. The disturbance function, \( e(\cdot) \), will be a stochastic process with some sort of stationarity condition. The usual assumption about \( e(\cdot) \) is
\[
E(e(x)) = 0 \quad \text{for all } \ x, \ \text{and}
\]
\[
\frac{1}{2} E(e(x) - e(x'))^2 = \gamma(x - x').
\]

The function \( \gamma(\cdot) \) is known as the semi-variogram; the "semi" referring to the "semi" in Equation (2). The disturbance function is stationary in the sense that its mean is independent of \( x \), and \( \gamma(\cdot) \) depends only on \( x - x' \). In some situations, we may also assume that \( e(\cdot) \) is an isotropic process; that is, that \( \gamma(\cdot) \) depends only on \( |x - x'| \), the distance between the two points.

The model described by (1) and (2) is commonly used in geostatistics. Basic references in geostatistics are Matheron (1971) and Journel and Huijbregts (1978). This model has also been used to describe spatial distributions of acid deposition (Eynon and Switzer (1983)), petroleum reserves (Davis (1981)), and forest characteristics (Matern (1960)). One of the notable features of this model is the use of the semi-variogram instead of the covariance function. If the covariance function, defined as
\[
E(e(x) e(x')) = C(x - x')
\]
exists, then we have the relationship
\[
\gamma(r) = C(0) - C(r).
\]
The simplicity of this relationship indicates why the semi-variogram, \( \gamma(\cdot) \), is commonly used instead of the variogram, \( 2\gamma(\cdot) \). Matern (1960) contains an extensive discussion on the properties of stationary and isotropic covariance functions. One reason for using variograms instead of covariances is that it is possible to have situations in which \( \gamma(\cdot) \) is finite but covariances are infinite. For example, a possible form for \( \gamma(\cdot) \) is \( \gamma(r) = b|r| \) for all vectors \( r \). In this case, \( C(r) = \infty \) for all
r. Even when the covariance function is finite, the semi-vario-
gram may be easier to estimate, as we shall see in I.1.3. In particular, we do
not need to estimate $\text{Var } z(x) = C(0)$ to estimate $\gamma(\cdot)$.

One model for $z(\cdot)$ that will be used extensively throughout this
work is

$$z(x) = a_0 + \alpha'_f(x) + e(x), \quad (5)$$

where the disturbance term, $e(\cdot)$, has mean zero and a stationary vario-
gram (Equation (2)). The mean function is $a_0 + \alpha'_f(x)$, where $f(x)$ is a
known, vector-valued function and $a_0$ and $\alpha$ are unknown coefficients.
Because of the form of the mean function, Equation (5) will be referred
to as the "regression model." In this work, $f(\cdot)$ will usually be con-
sidered to be a vector of observed covariates. For example, if $z(\cdot)$
represents the level of some pollutant, $f(\cdot)$ may represent weather char-
acteristics. Or if $z(\cdot)$ represents average annual temperature, $f(\cdot)$
may represent altitude. In what is known as "universal kriging," the
mean function is assumed to be a smooth unknown function that can be
modelled, at least locally, as a low order polynomial (Delfiner (1975)).
This situation creates different problems than when $f(\cdot)$ is made up of
observed covariates, and for the most part, we will not discuss universal
kriging in this work.

Now, let us consider the problem of estimating $z(x_0)$ based on
$z(x_1), \ldots, z(x_n)$ and the $f(\cdot)$ field. Suppose we restrict ourselves to
linear unbiased estimators of $z(x_0)$. Such estimators are of the form

$$\hat{z}(x_0) = \sum_{\alpha=1}^{n} \lambda_\alpha z(x_\alpha), \quad (6)$$

where the unbiasedness condition, $\text{E}z(x_0) = \text{E}\hat{z}(x_0)$, implies that
\[
\sum_{\alpha=1}^{n} \lambda_{\alpha} = 1 , \quad \text{and} \\
\sum_{\alpha=1}^{n} \lambda_{\alpha} f(x_{\alpha}) = f(x_{0}).
\]

(7)

We can evaluate the mean squared error of a linear unbiased interpolator in terms of the semi-variogram. We obtain (Matheron (1971)),

\[
E\left(\hat{z}(x_{0}) - z(x_{0})\right)^2 = 2 \sum_{\alpha=1}^{n} \lambda_{\alpha} \gamma(x_{\alpha} - x_{0}) - \sum_{\alpha=1}^{n} \sum_{\beta=1}^{n} \lambda_{\alpha} \lambda_{\beta} \gamma(x_{\alpha} - x_{\beta}).
\]

(8)

If the variogram is specified, then we can use this equation to find the linear unbiased estimator of \( z(x_{0}) \) with minimum variance, also known as the best linear unbiased estimator (BLUE). This problem can be easily solved using Lagrange multipliers (Journel and Huijbregts (1978), pp. 316-320) to obtain the BLUE of \( z(x_{0}) \). This method of interpolation is known as "kriging" in geostatistics. One of the desirable properties of kriging is that we obtain an expression for the variance of the error of our interpolator along with the interpolator. Goldberger (1962) considered essentially the same problem with applications to econometrics. If the disturbance field, \( e(\cdot) \), is Gaussian, then the BLUE of \( z(x_{0}) \) will also be the uniformly minimum variance unbiased estimator (UMVUE) of \( z(x_{0}) \). Of course, there may be biased and/or non-linear estimators with smaller mean squared error than the BLUE. However, to evaluate biased linear estimators we need to specify \( \text{Var} \, z(x) = \gamma(0) \), and to evaluate non-linear estimators, we need to know more about the distribution of the disturbance field than its covariance structure. Therefore, in this work, we will only consider linear unbiased estimators.

1.1.3 More about variograms and interpolation

In this section, we will briefly review some of the basic facts about variograms and their relationship to kriging interpolators.
It is well known that a covariance function must be positive definite (Matern (1960)); i.e., for any set of sites $x_1, \ldots, x_n$ and weights $\lambda_1, \ldots, \lambda_n$, the covariance function,
\[
C(x, x') = E[(z(x) - Ez(x))(z(x') - Ez(x'))],
\]
must satisfy
\[
\sum_{\alpha=1}^{n} \sum_{\beta=1}^{n} \lambda_\alpha \lambda_\beta C(x_\alpha, x_\beta) \geq 0.
\]  
(9)

Similarly, a semi-variogram
\[
\gamma(x, x') = \frac{1}{2} E\left[ (z(x) - Ez(x)) - (z(x') - Ez(x')) \right]^2
\]
must satisfy the slightly weaker condition
\[
\text{if } \sum_{\alpha=1}^{n} \lambda_\alpha = 0, \text{ then } \sum_{\alpha=1}^{n} \sum_{\beta=1}^{n} \lambda_\alpha \lambda_\beta \gamma(x_\alpha, x_\beta) \geq 0
\]
(10)

(Matheron (1971)). The function $-\gamma(\cdot, \cdot)$ is said to be "conditionally positive definite."

In order to introduce some of the properties of variograms and kriging, we will consider the simple case of an isotropic variogram and a mean function which is an unknown constant, or $m(x) \equiv m$. First, we should note that in many instances, while the local average of a process may remain relatively constant over short distances, it may vary considerably over longer distances. To protect against this possibility, we can make our interpolator of $z(x_0)$ depend only on $z(x_\alpha)$'s for which $|x_\alpha - x_0|$ is relatively small. We should note that in some circumstances, it may not matter whether slow variations in the $z(\cdot)$ field are modelled as changes in the mean field or as variations in the disturbance field over long distances. However, one of our main principles
of point interpolation will be that our procedure should try to be independent of any modelling of the process over long distances. The reasoning behind this principle is that interpolation is essentially a problem of modelling the local variations in a spatial process; thus, we should focus our modelling efforts on understanding the local behavior of the process, and not worry about what happens over long distances. This reasoning may not be valid if there are strong periodicities in the process. In this case, observations far away from \( x_0 \) may give us substantial information about \( z(x_0) \) which is not contained in nearby observations.

The semi-variogram is a measure of local variation, and thus its use fits in with the above principle. To estimate the semi-variogram over a short distance \( d \), we can take all \( n_d \) pairs of points separated by \( d \) and use

\[
\hat{\gamma}(d) = (2n_d)^{-1} \sum_{|x_i - x'_i| = d} (z(x_i) - z(x'_i))^2 .
\]  

(11)

This simple, unbiased estimator was suggested by Matheron (1971). Note that this estimator does not depend on the unknown constant mean, \( m \), in any way. Also, if \( m \) actually varies slowly in space, but does not vary significantly over the distance \( d \), then \( \hat{\gamma}(d) \) will not be greatly affected. On the other hand, in order to estimate \( C(d) \), the covariance at distance \( d \), we need to have an estimate of \( m \), since

\[
C(d) = E[z(x) - m][z(x') - m], \text{ where } |x - x'| = d.
\]

Therefore, to estimate \( C(d) \), we need to estimate the global mean, which is a quantity we do not need to know in order to compute the variance of linear unbiased estimators. So we see that \( \gamma(\cdot) \), and not \( C(\cdot) \), is the appropriate function.
simple calculation shows that the BLUE is given by

$$\hat{z}(0) = \frac{1}{2}rz(h) + (1 - \frac{1}{2}r)z(2h).$$  \hspace{1cm} (12)$$

If \( r=1 \), which will occur if \( \gamma(\cdot) \) exhibits a pure nugget effect, then we get

$$\hat{z}(0) = \frac{1}{2}(z(h) + z(2h)).$$

If \( 1 < r < 2 \), then \( z(2h) \) will get a smaller, but still positive, weight than \( z(h) \). For \( h \) small, this behavior for \( \gamma(\cdot) \) will occur if \( \gamma'(0^+) > 0 \) and \( \gamma''(0^+) < 0 \). A commonly used example is the "exponential" semi-variogram,

$$\gamma(d) = c_1(1 - e^{-c_2d^2}).$$

If \( r=2 \), then \( \hat{z}(0) = z(h) \); that is, \( z(2h) \) gets no weight in estimating \( z(0) \). This phenomenon will occur if \( \gamma(d) = c_1d \). If \( 2 < r < 4 \), then \( z(2h) \) will get a negative weight. Examples of when \( 2 < r < 4 \) for small \( h \) are

$$\gamma(d) = c_1(1 - e^{-c_2d^2}), \text{ and } \gamma(d) = c_1d^\alpha, \quad 1 < \alpha < 2.$$  

In these cases, there is no nugget effect, \( \gamma'(0^+) = 0 \), and \( z(\cdot) \) is differentiable. Therefore, it is not surprising that we get a negative weight for \( z(2h) \), since if \( z(2h) - z(h) \) is positive, we would guess that \( z(h) - z(0) \) is also positive. If \( r = 4 \), then \( z(0) = 2z(h) - z(2h) \); that is, we can estimate \( z(0) \) without error. This situation will occur if \( \gamma(d) = c_1d^2 \). A quadratic variogram is an extreme case, as it is equivalent to all of the observations lying on a single line with probability one. It is not possible to have \( r > 4 \), as \( -\gamma(\cdot) \) will then not be conditionally positive definite (Equation (10)).

The behavior of \( \gamma(d) \) as \( d \) approaches infinity is another basic feature of a semi-variogram. As we expect the relationship between \( z(x) \)
and $z(x')$ to weaken as $|x-x'|$ increases, we will usually assume that
$\gamma(\cdot)$ is monotonically increasing in $d$. If
$$\lim_{d \to \infty} \gamma(d) = \infty,$$
then we also have $C(d) = \infty$ for all $d$. If
$$\lim_{d \to \infty} \gamma(d) = \gamma(\infty) < \infty,$$ \hfill (13)
then $\gamma(\infty)$ is known as the "sill value;" it just equals $\text{Var}(z(x)) = C(0)$. The distance at which $\gamma(\cdot)$ stops increasing is known as the "range" of the semi-variogram. Observations at sites farther apart than the range
are uncorrelated.

I.2 Other Approaches to Interpolation and Analysis of Spatial Data

While the model described in I.1.2 is extensively used in geological
applications, it is not used so much in other fields where statistical
models of spatial processes are needed. Spatial statistics have been
used in a variety of fields other than geology, including geography
(Cliff and Ord (1975)), archaeology (Hodder and Orton (1976)), and agri-
culture (Bartlett (1975)). Recent books on spatial statistics include
those by Ripley (1981), Bennett (1979), Bartlett (1975) and Cliff and Ord
(1981). We will briefly discuss two often used models for spatial
processes.

A model which has been frequently used in geography and ecology is
one in which the conditional distribution of each observation given all
other observations is considered (Besag (1974)). If our process is
assumed to be Gaussian, then it is sufficient to specify the conditional
mean and variance of each observation. If $x_1, \ldots, x_n$ represent the ran-
dom values of a spatial process at $n$ sites and $x_1, \ldots, x_n$ their
to estimate if all we want to do is evaluate the variance of linear unbiased interpolators.

Since we are interested in the local behavior of the \( z(*) \) field, it follows that we are most interested in \( \gamma(*) \) over short distances. We see, by definition, that \( \gamma(0) = 0 \). However, we may have

\[
\gamma(0^+) = \lim_{d \to 0} \gamma(d) > 0.
\]

In geostatistical parlance, the size of this discontinuity at the origin is known as the "nugget effect." It may be due to microstructure in the \( z(*) \) field, that is, variations in the process over very short distances, or due to measurement error, or both. If \( \gamma(d) \) is some constant for all distances greater than zero, then we have what is known as a "pure nugget effect." Alternatively, we could say \( z(*) \) is a white noise process. In this case, the process exhibits no spatial structure, and our interpolator will be the same at all places at which there is not an observation: the average of all observations.

Another important feature of the semi-variogram over short distances is \( \gamma'(0^+) \). In models used in practice, \( \gamma'(0^+) \) is almost always well-defined and finite. Let us consider a process \( z(*) \) with constant mean and an isotropic semi-variogram which has no nugget effect. Then, if \( \gamma'(0^+) > 0 \), \( z(*) \) will be continuous but not differentiable. If \( \gamma'(0^+) = 0 \), then \( z(*) \) will be differentiable.

The following simple example will illustrate the impact of the behavior of \( \gamma(*) \) at the origin on kriging interpolators. Suppose \( z(*) \) is a process in \( \mathbb{R}^1 \) with fixed, unknown mean and stationary variogram. We observe \( z(h) \) and \( z(2h) \), where \( h \) is "small," and wish to estimate \( z(0) \). We will find the BLUE of \( z(0) \) as a function of \( r = \gamma(2h)/\gamma(h) \). A
respective observed values, then one model for the conditional means and variances is

\[ E(X_i | x_j, j \neq i) = EX_i + \sum_{j \neq i} \beta_{i,j} (x_j - EX_j), \quad \text{and} \]

\[ \text{Var}(X_i | x_j, j \neq i) = \sigma_i^2 \]  \hfill (14, 15)

(Besag (1975)). If we did not actually observe \( x_n \), then we could, in theory, use this model to estimate it (by the modelled conditional expectation) and obtain an expression for the error of this estimator (by the modelled conditional variance). Thus, this model can be used to produce point interpolators along with estimates of the variance of the interpolation error. If we are only interested in values of the process on a regular grid, then there are natural ways to model the \( \beta_{i,j} \)'s and \( \sigma_i \)'s. For example, if the data sites form a lattice, then it is reasonable to assume all of the \( \sigma_i \)'s are equal except for those sites near the outer edge of the lattice. However, in this work, we are interested in modelling processes which are not necessarily observed at sites on a grid, and even if the observation sites are on a grid, we want to be able to do point interpolation at sites not on the grid. Some work has been done on applying the model in (14) and (15) to irregularly spaced data (Besag (1975)), but the models for the \( \beta_{i,j} \)'s and \( \sigma_i \)'s, while qualitatively reasonable, are somewhat arbitrary. The kriging formulation, in which we try to measure the degree of association in a process at all distances (through the variogram), seems more natural when we are interested in the value of the process at all sites, and not just those on a regular grid.

Another model which can be used to do point interpolation with spatial data is splining. Wegman and Wright (1983) contains a review of
the uses of splines in statistics. "Smoothing" splines try to find a smooth surface which pass near the observations. For example, suppose we observe \( z(x_1), \ldots, z(x_n) \), where \( x = (x^1, x^2) \) is in \( \mathbb{R}^2 \). Then the "thin plate" smoothing spline will be a function \( f(\cdot) \) which minimizes

\[
\frac{1}{n} \sum_{j=1}^{n} (f(x_j) - z(x_j))^2 + \lambda J_m(f), \text{ where}
\]

\[
J_m(f) = \sum_{k, \ell=1}^{m} \int_{\mathbb{R}^2} \left( \frac{\partial^m f}{\partial x^k \partial x^\ell} \right)^2 \, dx^1 \, dx^2.
\]

The parameters \( \lambda \) and \( m \) are usually chosen by a cross-validation procedure (Wahba (1979)). As splining is usually used, no attempt is made to predict the variance of \( f(x_0) - z(x_0) \) at a specific place \( x_0 \), which is something the kriging formulation allows us to do. While kriging and splining appear to be rather different approaches to the problem of point interpolation, they in fact can produce equivalent interpolators. It turns out that some splines give the same estimation surfaces as particular cases of "universal kriging," which is described by Delfiner (1975). A specific example of such an equivalence is given by Dubrule (1983); Wegman and Wright (1983) also mention some equivalences between splining and universal kriging.

We could also consider doing point interpolation using methods of non-parametric regression not specifically aimed at spatial data. Such methods include nearest neighbor estimators, local linear estimators (see Stone (1977) and the references therein), and projection pursuit (Friedman and Stuetzle (1981)). While these methods may produce reasonable interpolators, they do not, in general, allow us to estimate the variance of the interpolation error at a particular place.
I.3 Summary of Contents

I.3.1 Estimation of variograms

To calculate the variance of the error of a linear unbiased interpolator of $z(x_0)$, it is necessary to specify the variogram (see Equation (8)). However, variograms usually cannot be specified a priori, but must be estimated form the data. The problem of estimating the variogram is addressed in Chapters II and III. We will use the regression model given in Equation (5). To avoid confusion, we will usually assume the variogram is isotropic, although the procedures considered can be easily modified to apply to stationary but anisotropic variograms.

In Chapter II, we will consider non-parametric estimators of the variogram. The estimators will be non-parametric in the sense that $\gamma(\cdot)$ is estimated at a specific distance $d$ in a way that does not require any assumptions about $\gamma(\cdot)$ at distances other than $d$. Also, since we do not need to estimate the regression coefficients directly to use unbiased linear interpolators, we will seek variogram estimators which do not depend on explicitly estimating the regression coefficients. Matheron's estimator of $\gamma(d)$, given in (11), satisfies these properties when the mean is a constant. However, it cannot be used if the mean function is not a constant, as $z(x)-z(x')$ will no longer have mean zero. When the mean is given by a regression equation, an unbiased estimator of $\gamma(d)$ which does not depend on estimating the regression coefficients is given in II.2. It is based on a weighted sum of squares of all observed values of $z(x)-z(x')$ which satisfy $|x-x'| = d$. This estimator cannot be used when the mean function includes an arbitrary polynomial in $x$ of degree greater than zero (see II.3); thus it is not
applicable to what is commonly called "universal kriging" by geostatisticians (Delfiner (1975)). In II.4, we consider how we can take advantage of an assumption that some pairs of points separated by $d$ are uncorrelated with other pairs. This situation can occur with a space-time process with repeated observations in time which are sufficiently separated in time so that temporal uncorrelatedness can be assumed.

One of the main reasons for wanting to estimate $\gamma(\cdot)$ non-parametrically is that it allows us to plot $\hat{\gamma}(d)$ as a function of $d$. These plots can serve as guides to the structure of $\gamma(\cdot)$ without making any a priori assumptions (beyond stationarity or isotropy) as to what that structure should be. Therefore, such plots are very useful in choosing a parametric model for the variogram which fits the data. In Section II.6, we discuss how these plots can also be used to help choose a subset of regressors from a set of potential regressors.

One of the assumptions of the regression model is that the parametric form of the mean function is correctly specified. In some circumstances, there may be physical reasons for believing a specific parametric model for the mean function. However, parametric regression functions are often used even when the relationship between the variables is assumed to be only approximately linear. In spatial processes, there is the added problem that the relationship between the covariates and the variable of interest may change in space. In II.7, the effects of misspecification of the mean function on both linear unbiased interpolators and on non-parametric variogram estimators are considered.

To illustrate some of the ideas and methods of this chapter, some computations are done using Landsat data in II.8. In these data, integrated surface reflectance in four different frequency bands is obtained
at each of 400 different pixels in a 16 x 25 rectangular region. The analysis focuses on two frequencies at a time, as the frequencies can be naturally separated into pairs of visible and infrared. The results indicate that much of the spatial variation in the energies at one frequency can be explained by the other frequency, and that if the first frequency had a missing value at a pixel, estimation of the missing value can be improved by using the second frequency rather than just taking an average of nearby values of the first frequency. The unbiased non-parametric estimator of the semi-variogram given in II.2 appears to be a quite noisy estimator. With these data, it seems that we are better off using an estimator based on residuals which will be slightly biased, but has lower variance.

In Chapter III, parametric methods of estimation of the variogram are considered. One reason parametric estimators are needed is to obtain an estimator of $\gamma(\cdot)$ which yields a permissible semi-variogram, i.e., we want $\hat{\gamma}(\cdot)$ to be conditionally positive definite (see Equation (10)). If the plot of a non-parametric estimator of $\gamma(d)$ v. d is smoothed, then the resulting function will not, in general, be a permissible semi-variogram. It is difficult to specify a class of variograms non-parametrically, but easy to specify a parametric class of variograms. Hence, the plot of $\hat{\gamma}(d)$ v. d is almost always fit by choosing a parametric class of semi-variograms which seems to be appropriate based on the plot, and then estimating the parameters by some means. Within the geostatistical literature, the parameters are usually estimated by ad hoc procedures, or by fitting the plot of $\hat{\gamma}(d)$ v. d by (weighted) least squares. The question of when such procedures provide adequate estimators of the
variogram is quite unclear, although there must be situations in which more efficient estimators of the parameters would be of practical value. Therefore, in Chapter III, we will consider applying some of the traditional methods of estimation to the problem of estimating the parameters in a variogram model. A substantial literature exists concerning the estimation of parameters in a covariance matrix; Harville (1977) contains a review of this problem. Methods discussed in this chapter include maximum likelihood (ML), modified or restricted maximum likelihood (MML), and minimum norm quadratic estimators. Such methods can be much more efficient than ad hoc or least squares fits to the non-parametric estimator.

In Section III.2, the problem of estimating a semi-variogram which is specified up to a scalar multiple is considered. That is, we assume

\[ \gamma(*) = \theta \gamma_0(*) , \]

where \( \gamma_0(*) \) is specified and \( \theta \) is to be estimated. When the disturbance field is Gaussian, the modified maximum likelihood estimator (MMLE) is the UMVUE (III.2.1). Also, it is shown to be the same as the maximum likelihood estimator (MLE) times a constant in this special case (III.2.1). Based on the MMLE of \( \theta \), we can obtain exact confidence intervals for \( z(x) \), where \( x \) is a place at which \( z(*) \) is not observed (III.2.2). The distribution of the MMLE of \( \theta \) does not depend on the locations of the observation sites as long as they are distinct. Thus, the modified maximum likelihood estimator will converge to the correct value (assuming the model in (18) is correct) as the number of observations gets large, even if the observations are all contained in a finite region. In contrast, the efficiency of non-parametric estimators is highly
dependent on the locations of the observation sites (see III.2.3) and are much harder to use when the sites are irregular in space (see II.5).

In Section III.3, parametric estimators of the semi-variogram are considered for more general parametric models. In III.3.1, the MLE and MMLE are given. In III.3.2, what Rao (1979) calls the MINQE(U,I) (minimum) norm quadratic estimator which is unbiased and invariant) is reviewed. Minimum norm estimators are only defined for models where the covariance matrix is linear in the parameters. In this section, a natural generalization of the MINQE(U,I) is given for general parametric models of covariance structures. This generalized MINQE(U,I), while no longer having an explicit solution as does the MINQE(U,I), is still easier to compute than the MLE or MMLE. Both the MINQE(U,I) and its generalization depend on some a priori estimate of the parameters. This a priori estimate can be obtained by using an eyeball fit to the plot of the non-parametric estimator of the semi-variogram versus distance.

In Section III.4, asymptotic properties of these estimators are considered. The asymptotics are difficult, since the joint distribution of the observations depends not only on their number but also on their locations. However, in III.4.1, a consistency result, independent of the locations (as long as they are distinct), is shown for a very special case of the generalized MINQE(U,I). We assume

$$\gamma(\cdot) = \gamma_1(\cdot) + \theta \gamma_2(\cdot),$$  \hspace{1cm} (19)

where $\theta$ is the parameter to be estimated and $\gamma_1(\cdot)$ and $\gamma_2(\cdot)$ are specified and satisfy the condition that $\gamma_2 - \alpha \gamma_1(\cdot)$ is a permissible semi-variogram for some $\alpha > 0$. In III.4.2, heuristic arguments are given for asymptotic covariances of the MMLE and MINQE(U,I). When
there is only one parameter, we show that the hypothesized asymptotic variance of the MMLE is less than the hypothesized asymptotic variance of the generalized MINQE(U,I), with equality if and only if the a priori estimate of the parameter used to compute the generalized MINQE(U,I) is equal to the true value of the parameter.

In III.5, the effects of misspecification of the mean function on parametric estimators are considered. We see that, in an important way, the parametric estimators are more sensitive to misspecifications of the mean function than the non-parametric estimators. However, perhaps the main reason ML, MML, and minimum norm quadratic estimators have largely been ignored by geostatisticians is the computational and numerical problems that occur when these methods are used with large data sets. One way to alleviate these problems is to compute the parametric estimator in a set of subregions, and then obtain a global estimator of the parameters by taking some sort of average of the local estimators (III.6). This method has the additional benefit of protecting against the effect of some types of misspecification of the mean function. In particular, by doing estimation locally, we protect ourselves against slow spatial variations of the regression coefficients. Local estimators can also be used as a way to detect non-stationarities in the variogram. The greater efficiency of parametric estimators should allow for greater sensitivity in picking up non-stationarities than non-parametric estimators. In III.7, strategies for picking parametric models for the variogram are considered. In III.8, the local parametric estimator of a variogram specified up to a scalar multiple (Equation (18)) is tested using the Landsat data. Specifically, the parameter is first estimated within each of 20 4x5 pixel subregions, and these local estimates are averaged to
obtain a global estimate of the parameter. These estimates are easy and quick to compute, as the congruence of the subregions greatly reduces the calculations. The degree to which the parametric estimator provides information about the variogram not contained in non-parametric estimators depends on the strength of local correlations and on the locations of the observations. In any case, computing the parametric estimator does no harm, and if it gives a much different variogram than an eyeball fit to a non-parametric estimator, we would at least know that the non-parametric estimator cannot be trusted.

I.3.2 Regression and co-kriging

In Chapters II and III, covariates are considered as fixed regressors of the variable of interest. An alternative method of using covariates is to consider them as a jointly stationary stochastic process. That is, we consider \((\mathbf{f}(\cdot),\mathbf{z}(\cdot))\) to be a stationary, vector-valued stochastic process. We could then estimate all covariances between the variables, and then obtain an interpolator by using the BLUE based on these estimated covariances. This technique of interpolation is known as co-kriging, see Journel and Huijbregts (1978), pp. 324–26. In Chapter IV, the regression and co-kriging models are compared. While the co-kriging model will be inappropriate for situations where the covariates cannot reasonably be modelled as a stationary stochastic process, the co-kriging model does not require that the covariates be observed at specific sites and will produce better interpolators than the regression model in some circumstances. The main result of this chapter is that if the co-kriging model is correct, then the regression model will only strictly make sense if there exists a vector \(\mathbf{a} \) such that
\[
\text{Cov}(\hat{f}(x), z(x') - \hat{a}'\hat{f}(x')) = 0 \quad \text{for all } x, x'.
\] (20)

This condition can be interpreted as meaning that the covariate and disturbance fields are uncorrelated. If (20) holds, then we roughly have that all of the information about \(z(x)\) contained in the \(\hat{f}(\cdot)\) field can be obtained from \(\hat{f}(x)\), which is an assumption the regression model implicitly makes.

Because of the simplicity and familiarity of the regression model, we may decide to use it even when there is no \(\hat{a}\) which satisfies (20). Suppose we observe many independent replications of \(z(\cdot)\) and \(\hat{f}\) at a fixed set of sites. For example, the replications may be observations at different times. Suppose we use the regression model, and estimate the regression coefficients by a generalized least squares procedure. Then, in general, what our estimators of the regression coefficients converge to will depend on both the weighting matrix used in the generalized least squares procedure and on the locations of the fixed sites (IV.3.2 and IV.3.3). However, if there is an \(\hat{a}\) which satisfies (20), then the estimators of the regression coefficients will converge to that \(\hat{a}\) for any weighting matrix and set of sites. Even when there is no such \(\hat{a}\), we may still be able to produce reasonable (although not optimal) interpolators and accurate estimates of the variance of the interpolation error using the regression model (IV.3.4).

1.3.3 A model for spatial processes observed repeatedly in time

In Chapter V, we consider spatio-temporal processes which can be modelled as

\[
z(x,t) = m(x) + \mu(t) + e(x,t).
\] (21)

We assume that this process is observed at a fixed set of sites at times
The sequence $\mu(t_1), \ldots, \mu(t_k)$ is taken to be an arbitrary, deterministic time series and $e(x,t)$ is assumed to be uncorrelated from one time to another. Two problems related to this model are considered. In V.2.1, we consider estimation of $z(x_0, t_0)$, where $\alpha$ is in $\{1, \ldots, k\}$ and $x_0$ is not one of the fixed observation sites. Clearly, we need to impose some structure on $m(x)$. The approach taken here is to assume $m(x)$ has a fixed mean and a stationary variogram. Using this model, the BLUE of $z(x_0, t_0)$ is given. An asymptotic result is given as $k$, the number of observations in time, tends to infinity. This asymptotic expression can be easily interpreted as having a term which interpolates the $m(*)$ field and a second term which interpolates the $e(*, t_0)$ field. In V.2.2, similar results are developed when $\mu(t)$ is assumed to be fixed in time.

The second problem considered in Chapter V is the estimation of the average change in the $z(*, *)$ field from time $t_1$ to $t_2$ in some specified region $R$. The main idea of this section is that we should estimate this quantity by computing $z(x_i, t_1) - z(x_i, t_2)$ for $i=1, \ldots, k$, and then estimate the average value of this "difference" field in the region $R$. By using this procedure, we only have to model the difference field, which is independent of $m(x)$, so that the problem of modelling $m(x)$ is eliminated. Alternatively, we could separately estimate the average of $z(*, t_1)$ and $z(*, t_2)$ in $R$, take the difference of these estimators to obtain our estimator of the average difference, and assume the estimators of the average in $R$ within each time are uncorrelated to obtain the variance of the estimator of the average difference. This procedure essentially assumes that $m(x)$ is a constant, which will cause us to overestimate the variance of the estimator of the average difference if $m(x)$, in
fact, varies in $x$. The situation is similar to that of doing a paired 
t-test rather than a two-sample t-test.

I.3.4 Sequential siting

In some situations, the locations of all of the observations of a 
spatial process are not determined before any observations are taken. 
Instead, an initial sample may be taken, and further locations are chosen 
based on the previous observations. A common example occurs when it is 
important to estimate $z(\cdot)$ well in regions where it takes on large 
values, a problem which is crucial in pollution monitoring. After an 
initial set of observations are taken, further observations are taken 
near those sites with relatively high values of $z(\cdot)$. Using the regres-
sion model in Equation (5), let us define a contrast as a linear combina-
tion of the data which has expectation zero for all possible values of 
the regression coefficients. We will assume the disturbance term is 
Gaussian and the variogram has been specified. In Chapter VI, we show 
that as long as the locations of future observations are functions of 
the contrasts of previous observations, the distribution of the error of 
a kriging interpolator will be unaffected by the fact that the locations 
of the observations were not all selected before any observations were 
taken.

I.3.5 Other results

In Appendix A, we consider the problem of estimating $\gamma'(0)$ for a 
Gaussian process in an interval of $\mathbb{R}^1$ when $\gamma(\cdot)$ exhibits linear be-
behavior at zero. In particular, a simple, non-parametric estimator of 
$\gamma'(0)$ is compared to the UMVUE when $\gamma(\cdot)$ can be specified up to a 
multiplicative constant. As the observations in the interval get dense, 
the first two terms in an asymptotic expansion of the mean squared error
of the non-parametric estimator are given. As a corollary, we obtain
that the non-parametric estimator has the same asymptotic efficiency as
the optimal parametric estimator.

In Appendix B, we consider the effect of using estimated covariances
instead of actual covariances to produce interpolators and estimators of
regression coefficients. In particular, consider the regression model
given in Equation (5), and assume the disturbance term is Gaussian. Let
us assume the covariance structure is estimated based only on the con-
trasts of the data, which is true for the MLE, MMLE, and MINQE(U,I).
If we are doing interpolation or estimating a regression coefficient,
then the variance of a "pseudo-UMVUE" which uses the estimated covariance
structure as if it were the actual covariance will be greater than the
variance of the true UMVUE using the correct covariance structure. This
result is not trivial since the pseudo-UMVUE need not be unbiased. In
fact, what is actually proved in this appendix is the stronger result
that the conditional variance of the pseudo-UMVUE given all contrasts is
equal to the variance of the UMVUE.

In Appendix F, we show some results about spatio-temporal vari-
ograms. One result, Theorem F.3, is a basic result about the relation-
ship between variograms and covariance functions which does not appear
to have ever been proved before. This theorem says that if a spatial
semi-variogram, \( \gamma(\cdot) \), has a finite sill, \( L \), then \( L - \gamma(\cdot) \) is a per-
missible spatial covariance function; that is, it is a positive definite
function (see Equation (9)).
I.4 Notation

We now review some of the notational conventions to be used in this work.

In denoting vectors and matrices, the subscripts \( i \) and \( j \) will have special meanings. That is, \((a_i)\) will denote a column vector whose \( i \)'th element is \( a_i \). And, \((a_{ij})\) will denote a matrix whose element in the \( i \)'th row and \( j \)'th column will be \( a_{ij} \). This notation is used extensively in Chapter III.

For a matrix \( B \), we will define \( \mathcal{N}(B) \) to be the set of all matrices of full rank whose rows form a basis for the left null-space of \( B \).

Thus, if \( B \) has \( n \) rows and is of rank \( k < n \), then \( A \in \mathcal{N}(B) \) means \( AB = 0 \), and \( A \) has \( n-k \) rows and rank \( n-k \). This notation is frequently used in Chapters III and VI where it is important to consider the contrasts of a random vector. If \( Z \) is a random vector with mean \( F'\alpha \), where \( F \) is a known matrix of regressors and \( \alpha \) a vector of unknown regression coefficients, then a contrast is a linear combination of the components of \( Z \) that has an expected value of zero for all possible values of \( \alpha \). If \( C \in \mathcal{N}(F') \), then \( CZ \) forms a basis for all contrasts. That is, if \( y'Z \) is a contrast, then \( y'Z = \omega' CZ \) for some \( \omega \).

Other notation:

\[
\text{BLUE} = \text{best (minimum variance) linear unbiased estimator.}
\]

\[
\text{Cov}(\bar{x}, y') = E[(x - E(x))(y - E(y))']
\]

\( e_n \) or \( \varepsilon_n \) = a vector of 1's of length \( n \). The \( n \) will be suppressed if it is clear from context.

\( I_n \) = the identity matrix of order \( n \).
MINQE(U, I) = minimum norm quadratic estimator which is unbiased ("U") and invariant ("I") with respect to changes in the regression coefficients. See III.3.2.

MLE = maximum likelihood estimator.

MMLE = modified maximum likelihood estimator. Also known as restricted maximum likelihood estimator. See III.3.1.

\( N(\mu, \Sigma) \) = multivariate normal distribution with mean vector \( \mu \) and covariance matrix \( \Sigma \).

0 = a matrix of all zeroes.

Positive definite (p.d.): A square, symmetric matrix \( B \) is said to be positive definite if \( x' B x \geq 0 \) for all \( x \) with equality only if \( x = 0 \). A function \( f(x, x') \) is said to be positive definite if

\[
\sum_{i,j=1}^{n} \omega_i \omega_j f(x_i, x_j) \geq 0
\]

for all \( x_1, \ldots, x_n \), all \( \omega_1, \ldots, \omega_n \), and all \( n \).

Positive semi-definite (p.s.d.): A square, symmetric matrix is positive semi-definite if \( x' B x \geq 0 \) for all \( x \). We also define, for square, symmetric matrices \( P \) and \( Q \),

\( P \preceq Q \) if \( P - Q \) is p.s.d.

UMVUE= uniformly minimum variance unbiased estimator.

\( \text{Var} \, X = \text{Variance of} \, X = E(X - EX)^2 \).

Finally, we will use the following standard definitions:

\( x_n = o(y_n) \) means \( x_n / y_n \to 0 \) as \( n \to \infty \), and

\( x_n = O(y_n) \) means \( x_n / y_n \) remains bounded as \( n \to \infty \).
Chapter II: Non-Parametric Variogram Estimation

II.1 Statement of Problem

In this chapter and in the next, we will discuss the problem of estimating a variogram when the mean of a random field is given by an unknown linear combination of known functions. That is, suppose

\[ z(x) = a_0 + \mathbf{x}^T \mathbf{f}(x) + e(x) \]  

(1)

for all \( x \) in some region, where

\[ a = (a_1 \ldots a_p)' \] is a vector of unknown parameters,

\[ \mathbf{f}(x) = (f_1(x) \ldots f_p(x))' \] is known for all \( x \) (or at least all \( x \) at which we observe or wish to estimate \( z(\cdot) \)), and \( e(\cdot) \) is a random field with

\[ \mathbb{E}(e(x)) = 0 \quad \text{for all } x, \quad \text{and} \]

\[ \frac{1}{2} \mathbb{E}(e(x) - e(x'))^2 = \gamma(x-x'). \]

The problem, then, is to estimate \( \gamma(\cdot) \) based on \( z(x_1), z(x_2), \ldots, z(x_n) \). In this work, we will usually assume \( \gamma(\cdot) \) is isotropic. However, unless some explicit comment is made, the results herein can be easily modified to the case where \( \gamma(\cdot) \) is only assumed to be stationary. Then, as noted in I.1.3, if the mean function is a constant (\( p = 0 \)), \( \gamma(d) \) is usually estimated by the unbiased estimator (Matheron (1971))

\[ \hat{\gamma}(d) = \frac{1}{2n_d} \sum_{\|x_i-x_i'\|=d} (z(x_i)-z(x_i'))^2, \]  

(2)

where the sum is over all pairs of points separated by the distance \( d \), and \( n_d \) is the number of pairs of points separated by \( d \).

Two important features of this estimator is that it allows us to estimate \( \gamma(d) \) without estimating \( a_0 \), and more importantly, without making any assumptions about the value of \( \gamma(\cdot) \) at distances other than \( d \).
Based on estimates of $\gamma(\cdot)$ at various distances, we can then plot $\hat{\gamma}(\cdot)$ as a function of $d$. This plot can then be used to pick a parametric class of variograms, which is perhaps the main use of the non-parametric estimator. The parameters can be estimated by means discussed in Chapter III.

In this chapter, we consider estimation of $\gamma(d)$ when the mean is not a constant. We seek estimators which are simple, do not depend on estimating $a_0$ or $a$, and are unbiased no matter what $\gamma(\cdot)$ is at distances other than $d$. We should emphasize that it is not the unbiasedness, \textit{per se}, in which we are so interested, but in the fact that a reasonable estimator of $\gamma(d)$ is obtained without having to estimate the regression coefficients and without having to model $\gamma(\cdot)$ at other distances. By obtaining estimators with these properties, we can, as in the case of a constant mean, obtain a plot of $\hat{\gamma}(d)$ v. $d$. As these plots will be used mainly as a way of obtaining a reasonable parametric class of variograms, we will not be overly concerned with the efficiency of these estimators.

II.2 A Generalization of Matheron's Estimator

Matheron's estimator, equation (2), cannot be used when the mean is not a constant, since

$$E_k(z(x) - z(x'))^2 = \gamma(|x - x'|) + \frac{1}{2}a^2(\tilde{f}(x) - \tilde{f}(x')).$$

Therefore, we consider a slightly more complicated estimator of $\gamma(d)$.

Assume there are $m$ pairs of points separated by $d$, and let $(x_i(j), x_i'(j))$ be the $j$'th pair. Thus $i(\cdot)$ and $i'(\cdot)$ are functions from $(1, 2, \ldots, m)$ to $(1, 2, \ldots, n)$. Define
\[ y_j = z(x_{i(j)}) - z(x_{i'}(j)), \text{ and} \]
\[ g_j = (g_j^1 \ldots g_j^p)' = f(x_{i(j)}) - f(x_{i'}(j)). \]

We will consider estimators of the form
\[ \frac{1}{2} \sum_{j=1}^{m} w_j y_j^2. \quad (3) \]

Now,
\[ E\left[ \frac{1}{2} \sum_{j=1}^{m} w_j y_j^2 \right] = \gamma(d) \sum_{j=1}^{m} w_j + \frac{1}{2} \sum_{j=1}^{m} w_j (a' \cdot g_j)^2. \]

We want this expectation to equal \( \gamma(d) \) for all values of \( \underline{a} \). Thus, we need
\[ \sum_{j=1}^{m} w_j = 1, \quad \text{and} \]
\[ \sum_{j=1}^{m} w_j (a' \cdot g_j)^2 = 0 \quad \text{for all} \quad \underline{a}. \quad (4) \]

These conditions can be easily shown to be equivalent to
\[ \sum_{j=1}^{m} w_j = 1, \quad \text{and} \]
\[ \sum_{j=1}^{m} w_j g_j g_j' = 0. \quad (4') \]

Any set of \( w_j \)'s which satisfy these constraints will yield an unbiased estimator for \( \gamma(d) \). Since \( \sum_{j} w_j g_j g_j' \) is a symmetric matrix, the second equation in \( (4') \) includes at most \( p(p+1)/2 \) independent constraints on the \( w_j \)'s. We can write the set of constraints in matrix form:
\[ A'w = (1 \ 0 \ ... \ 0)', \quad \text{where} \]
\[ w = (w_1 \ ... \ w_m)', \quad (5) \]

and
\[
A' = \begin{bmatrix}
1 & \cdots & 1 \\
(g_1^2) & \cdots & (g_m^2) \\
\vdots & \ddots & \vdots \\
g_1^p & \cdots & g_m^p \\
(g_1^2) & \cdots & (g_m^2) \\
\vdots & \ddots & \vdots \\
g_1^p & \cdots & g_m^p
\end{bmatrix},
\]
so that \( A' \) has \( 1 + p(p+1)/2 \) rows.

For the moment, let us assume that (5) has a solution. Usually, it will have an infinite number of solutions. Therefore, we must have a criterion to decide which solution to use. The obvious criterion would be to minimize the variance. However, the variance depends on the fourth moments of the \( z(\cdot) \) field. Qualitatively, we know that we do not want any one term in the summation in (3) to contribute too much to the total. That is, we want the \( w_j \)'s to be small in magnitude. Therefore, finding the solution to (5) which minimizes \( y'w = \sum_{j=1}^{m} w_j^2 \) is at least a qualitatively reasonable criterion for choosing \( y \). This minimum norm solution, which we call \( y^* \), will be the unique solution of (5) which is orthogonal to the right nullspace of \( A' \). That is, if \( A'y = 0 \), then \( y'y^* = 0 \). Thus we see that
\[
y^* = A(A'A)^+(1 \ 0 \ \cdots \ 0)', 
\]
where \( X^+ \) is the Moore-Penrose generalized inverse of \( X \) (Rao (1973)), which just equals \( X^{-1} \) if \( X \) is non-singular. If \( p = 0 \), then Matheron's estimator, given by Equation (2), is recovered. Also,
\[
y^*y^* = (1 \ 0 \ \cdots \ 0)(A'A)^+(1 \ 0 \ \cdots \ 0)'.
\]
We will define the estimator based on \( \tilde{y}^* = (w_1^* \cdots w_m^*) \) as
\[ \hat{\gamma}(d) = \frac{1}{2} \sum_{j=1}^{m} w_j^* y_j^2. \] (8)

Given the fourth moments of the \( z(\cdot) \) field, we can write down the variance of an estimator of the form (3) which satisfies (5), but the expression would be quite complicated and not very useful. If \( e(\cdot) \) is a Gaussian field, then straightforward calculations yield

\[
\text{Var} \frac{1}{2} \sum_{j=1}^{m} w_j y_j^2 = \frac{1}{2} \sum_{j,k} w_j w_k [\gamma(\|x_i(j)-x_i(k)\|) + \gamma(\|x_i'(j)-x_i'(k)\|) - \gamma(\|x_i(j)-x_i'(k)\|) - \gamma(\|x_i'(j)-x_i(k)\|)]^2 
+ \sum_{j,k} w_j w_k (a'g_j)(a'g_k) [\gamma(\|x_i(j)-x_i'(k)\|) + \gamma(\|x_i'(j)-x_i(k)\|) - \gamma(\|x_i(j)-x_i'(k)\|) - \gamma(\|x_i'(j)-x_i(k)\|)].
\] (9)

If the \( y_j \)'s are independent, which might occur if the pairs are sufficiently separated in space or time, then only those terms in (9) with \( j=k \) will be non-zero. Thus, we have

\[
\text{Var} \frac{1}{2} \sum_{j=1}^{m} w_j y_j^2 = 2\gamma(d)^2 \sum_{j} w_j^2 + 2\gamma(d) \sum_{j} w_j^2 (a'g_j)^2. \] (10)

We see that minimizing \( \sum w_j^2 \) minimizes the first term in (10), and, while not minimizing the second, does qualitatively constrain it. However, as we shall see in II.4, \( \hat{\gamma}(d) \) is a poor choice for an estimator of \( \gamma(d) \) when \( e(\cdot) \) is Gaussian and the \( y_j \)'s are independent.

While minimizing \( \sum w_j^2 \) does provide a criterion for selecting \( w \) which yields a simple solution, the resulting estimator, while unbiased, may have too high a variance to be used even as a guide for selecting a parametric class of variograms. This problem can occur if one or more of the following conditions exist: \( m \) is not sufficiently large, the
locations of the observations are highly irregular (see III.2.3), or
Equation (5) almost has a singularity (see II.3.1). In these cases, it
may be necessary to pick \( \hat{\gamma} \) using a criterion which comes closer to
minimizing the variance than minimizing \( \Sigma w_j^2 \). For example, we may try
to minimize the right hand side of (10) using some simple estimate of \( \hat{\gamma} \)
in place of \( \gamma \). However, we should note that if \( \hat{\gamma} \) depends on
\( z(x_1), \ldots, z(x_n) \), our estimator of \( \gamma(d) \) may no longer be unbiased, even
if \( \hat{\gamma} \) satisfies (5).

II.3 Conditions for Existence of Solution to Unbiasedness Constraints

II.3.1 Necessary and sufficient condition for a solution

Let us write \( A \) as \( (e \ A_2) \). Then (5) will have a solution if and
only if

\[
e \text{ is not in the column space of } A_2, \quad \text{and} \quad (11a)
\]

\[
A_2' \text{ has a non-trivial right null-space;} \quad (11b)
\]

i.e., \( A_2' \gamma = 0 \) has a non-zero solution.

Condition (11b) is equivalent to \( m > \text{rank}(A_2) \). Thus, (11b) is satisfied
if \( m > p(p+1)/2 \). So, (11b) will be satisfied as long as there are a
sufficient number of pairs of points separated by \( d \). However, (11a)
need not be satisfied even if \( m \) is large. Even if it is satisfied, \( e \)
may "almost" be in the column space \( A_2 \) (that is, there exists \( \gamma \) such
that \( A_2 \gamma = e \)). In this case, \( \hat{\gamma}_1 \gamma \) will tend to be large, and \( \hat{\gamma}(d) \)
as defined in (8) will probably have large variance. For example, if
\( f(x) = f_1(x) \), then

\[
\hat{\gamma} = \frac{(u'u)e - (e'u)u}{m(u'u) - (e'u)^2}, \quad (12)
\]
where
\[ u = \begin{bmatrix} (f(x_1'(1)) - f(x_1'(1))^2 \\ \vdots \\ (f(x_1'(m)) - f(x_1'(m))^2 \end{bmatrix}. \]

Then,
\[ \psi^* \psi^* = \frac{u' u}{m(u' u) - (e' u)^2}. \quad (13) \]

If, for example, \( u = (0 \ 1 \ \ldots \ 1)' \), then \( \psi^* \psi^* = 1 \), whereas we would hope that it would be \( O(1/m) \), since if (10) holds, then \( \text{Var}(\gamma(d)) \geq 2 \gamma(d)^2 \psi^* \psi^* \).

So, if \( u \) is nearly a multiple of \( e \), then we pay a high price for demanding unbiasedness.

II.3.2 The case of linear drift

(Note: In this section, it will be necessary to treat stationary and isotropic variograms separately. To avoid confusion, \( r \) will always be a vector, and \( d \) will always be a distance.)

One common case in which (5) does not have a solution is when \( \gamma(*) \) is stationary and there are any linear drift terms in the mean function; that is, if one of the components of \( f(x) \) is of the form \( c' x \) for some fixed vector \( c \). If we then try to estimate \( \gamma(r) \) using an estimator which satisfies (5), there will be a column in \( A_2 \) which will equal \( (c' r)^2 e \). Thus, (11a) is violated, and no solution of (5) will exist.

Therefore, if any of the components of \( f(x) \) is a linear combination of the coordinates, then no unbiased estimators of \( \gamma(r) \) of the form given in (3) will exist.

Even if we assume \( \gamma(*) \) is isotropic, there is still a problem with allowing linear drift. Suppose \( x = (x^1 \ldots x^q)' \). Also suppose \( f_k(x) = x^k \), \( k = 1, \ldots, q \); that is, we want to allow \( f(x) \) to include an arbitrary
first order polynomial. Now consider the problem of estimating $\gamma(d)$. Then we have

$$|x_i(j) - x_i'(j)|^2 = \sum_{k=1}^{q} (x_i^k(j) - x_i'(j))^2 = d^2$$ for $j=1, \ldots, m$,

using the notation of II.2. Now, $A_2$ will have a column

$$\left((x_1^k(1) - x_1'(1))^2 \ldots (x_1^k(m) - x_1'(m))^2\right)'$$ for $k=1, \ldots, q$.

If we add these $q$ columns together, they sum to $d^2 e$. Thus, (11a) is again violated, so (5) will not have a solution. However, assuming $\gamma(\cdot)$ is isotropic, we can allow less than $q$ linear combinations of the coordinates to be elements of $f(\cdot)$ and still have a solution to (5). For example, with two-dimensional geographic data, we could allow a latitude variable in $f(\cdot)$. However, if both latitude and longitude variables are included, then (5) will not have a solution.

Allowing arbitrary linear drift is quite common in spatial statistics, and the fact that there are no unbiased estimators of the form given by (3) even when isotropy is assumed is a major limitation on this technique. In particular, this method is of no use in what is called universal kriging in geostatistics; that is, when the mean function is assumed to be, at least locally, an arbitrary polynomial of degree greater than zero. Where this technique will work is when the $f_k$'s are actually observed covariates, and not polynomials meant to model a smooth, but otherwise unknown mean function.

II.4 The Case of Uncorrelated Pairs

II.4.1 An $s^2$-like estimator of $\gamma(d)$

As was mentioned in II.2, there is a more natural and often much
better estimator of $\gamma(d)$ than those of the type given in (3) if $y_1, \ldots, y_m$ are uncorrelated. In this case, we can write

$$y_j = \beta' g_j + v_j ,$$

where

$$v_j = e(x_i(j)) - e(x_i'(j)).$$

Then we have

$$E v_j = 0 ,$$

$$E v_j v_k = \begin{cases} 2\gamma(d), & j=k, \text{ and} \\ 0 , & j \neq k. \end{cases}$$

We see that we have a standard regression problem with uncorrelated observations, with $2\gamma(d)$ playing the role of the variance of the observations. Thus, we can use the standard "$s^2$" estimator based on these $m$ observations to obtain

$$\hat{\gamma}_s(d) = \frac{1}{2(m-p')} \sum_{j=1}^{m} (y_j - \hat{\beta}' g_j)^2 ,$$

(15)

where $\hat{\beta}$ is the least squares estimator of $\beta$ based on $y_1, \ldots, y_m$, and $p' = \text{rank}(g_1 \ldots g_m)$. This estimator will be unbiased as long as the assumption of uncorrelatedness is correct. If the $e(*)$ field is Gaussian, then

$$\frac{(m-p')\gamma_s(d)}{\gamma(d)} \sim \chi^2_{m-p},$$

and thus

$$\text{Var} \hat{\gamma}_s(d) = \frac{2\gamma(d)^2}{m-p'} .$$

Rao (1973, p. 319) shows that this estimator is the UMVUE under the uncorrelatedness and Gaussian assumptions. Thus, in this case, $\hat{\gamma}_s(d)$ is uniformly better than any estimator of the form $\sum_j y_j^2$ which satisfies the unbiasedness constraints in (5). If the $v_j$'s are independent,
but not necessarily Gaussian, the variances of $\hat{\gamma}_s(r)$ and $\sum_{j} \gamma_j^2$ are determined by the first four moments of the $\gamma_j$'s. Thus, if the $\gamma_j$'s are independent and have zero skewness and kurtosis, $\hat{\gamma}_s(d)$ will be uniformly better than any unbiased estimator of the form $\sum_{j} \gamma_j^2$.

II.4.2 Using estimators based on residuals when pairs are correlated

If $\hat{\gamma}_s(d)$ is used and the $\gamma_j$'s are not uncorrelated, then the estimator will, in general, be biased. The bias will depend on $\gamma(\cdot)$ for distances other than $d$, so that there is no simple way to correct for this bias. However, if $m$ is large and the observations are spread out over a large region, then the least squares estimator of $\sigma$ may be close to $\sigma$ with high probability. In this case, $\hat{\gamma}_s(d)$ will be nearly unbiased. Of course, if $\sigma$ can be estimated well using just the $m$ $\gamma_j$'s, it can certainly be estimated well using all of the data. Thus, we might consider the residuals, $z(x_j) - \hat{\sigma}' \hat{\mathbf{f}}(x_j)$, where $\hat{\sigma}$ is an estimator of $\sigma$ based on all of the data, as close to the true errors, and estimate the variogram using them. However, even when there is a large amount of data, $\sigma$ may not be so easy to estimate, and thus the residuals may not be almost the same as the errors. Problems in estimating $\sigma$ can occur not only because the observations are correlated, but also because $\sigma_0$ may vary slowly in space. We may even have that $\sigma$ varies slowly in space, in which case, it is quite unclear what the residuals are measuring. The effect of $\sigma_0$ and $\sigma$ varying slowly in space on interpolaters and non-parametric variogram estimators will be discussed in II.6.2. The "residual" approach is also considered using Landsat data in II.8.

II.4.3 Repeated observations in time

One likely instance in which some observations will be independent
of others is when there are repeated observations in time which are sufficiently separated in time so that observations from different times can be assumed to be independent. More specifically, assume

\[ z(x, t_\lambda) = m(t_\lambda) + a' \xi(x, t_\lambda) + e(x, t_\lambda), \quad \lambda = 1, \ldots, k, \tag{16} \]

where \( m(t_\lambda) \) is an unknown, deterministic times series and

\[ \mathbb{E}[(e(x, t_\lambda) - e(x', t_\lambda))^2] = \gamma(x, x'), \text{ and} \]

\[ \mathbb{E}[(e(x, t_k) - e(x', t_k))(e(x'', t_k) - e(x'', t_k))] = 0, \quad \text{for } k \neq \lambda. \]

This assumption on the temporal structure is a sort of uncorrelatedness, and is weaker than independence. Its implications will be discussed in Chapter V. Note for the moment we have not assumed that \( \gamma(\cdot, \cdot) \) is isotropic nor stationary. Thus, we write it as a function of both locations.

If \( z(t) \) and \( \xi(t) \) are observed at \( x \) and \( x' \) at times \( t_1, \ldots, t_k \), then

\[ z(x, t_\lambda) - z(x', t_\lambda) = a' (\xi(x, t) - \xi(x', t_\lambda)) + e(x, t_\lambda) - e(x', t_\lambda), \]

\( \lambda = 1, \ldots, k \), form \( k \) uncorrelated observations, and hence an unbiased estimator of \( \gamma(x, x') \) is

\[ \hat{\gamma}_s(x, x') = \frac{1}{2(k-p')} \sum_{\lambda=1}^{k} \left( z(x, t_\lambda) - z(x', t_\lambda) - \hat{a}' (\xi(x, t_\lambda) - \xi(x', t_\lambda)) \right)^2, \tag{17} \]

where \( \hat{a} \) is the least squares estimator of \( a \) based on these \( k \) differences, and

\[ p' = \text{rank} \left[ (\xi(x, t_1) - \xi(x', t_1)) \ldots (\xi(x, t_k) - \xi(x', t_k)) \right], \]

where we assume \( k > p' \). If \( k \) is substantially larger than \( p \), we may consider using (16) to estimate \( \gamma(x, x') \) for each pair of observation sites. For \( k \) large enough, \( \hat{\gamma}_s(x, x') \) will be a decent estimator of \( \gamma(x, x') \). Therefore, we can check our assumptions of isotropy or
stationarity by examining \( \hat{\gamma}_s(x, x') \) for pairs of points separated by the same distance or vector. If we are satisfied that \( \gamma(\cdot) \) is approximately isotropic, we can then estimate \( \gamma(d) \) by

\[
\hat{\gamma}_s(d) = \frac{1}{m} \sum_{j=1}^{m} \hat{\gamma}_s(x_i(j), x_i'(j)),
\]

borrowing the notation from II.2. This estimator, because it takes advantage of the temporal uncorrelatedness, may often be considerably better than unbiased estimates of the form given in (3), which in this case can be written as

\[
\hat{\gamma}(d) = \frac{1}{2} \sum_{k=1}^{k} \sum_{j=1}^{m} \sum_{k} w_{jk} (z(x_i(j), t_k) - z(x_i'(j), t_k))^2
\]

subject to the unbiasedness constraints. Of course, an estimator of this form will be unbiased even if the temporal uncorrelatedness condition does not hold, whereas the same cannot be said of \( \hat{\gamma}_s(d) \), given in (18).

One further point worth noting is that these \( s^2 \)-like estimators of \( \gamma(d) \) work for any \( f(\cdot) \). Thus, there is no problem with \( f(\cdot) \) including linear drift, as there was with the estimators of the form given in (3). Therefore, if the \( v_j \)'s, defined in (14), are uncorrelated, we may think we can use these \( s^2 \)-like estimators of \( \gamma(d) \) when we are doing universal kriging. However, when using universal kriging, we usually only assume that the mean function is locally a polynomial. For example, if we are in one dimension, and we use first order universal kriging, then we assume that, locally, \( m(x) = a_0 + a_1 x \). (The "order" of universal kriging refers to the order of the polynomial used.) We would not usually believe that \( m(x) \) was globally linear. Unfortunately, \( \hat{\gamma}_s(d) \), given in (15), assumes that the regression coefficients other than \( a_0 \)
(see II.6.2) are globally constant. Therefore, in general, it should not be used to estimate the variogram when doing universal kriging even if the \( v_j \)'s are uncorrelated.

II.5 Irregularly Spaced Data

Up until this point, in this chapter we have always assumed that there are a sufficient number of points separated by a distance \( d \) so that \( \gamma(d) \) can be estimated. If the data lie on a regular grid, then multiples of the grid size will serve as distances which satisfy this condition. However, as noted in I.2, one of the advantages of the kriging model is that it is still appropriate when the observations are irregularly spaced. In this case, what is usually done is to estimate \( \gamma(d) \) using pairs of points separated by approximately \( d \). More explicitly, \([0, R]\), where \( R \) is the largest distance at which we wish to estimate \( \gamma(\cdot) \), is partitioned into a set of intervals, called distance classes. Then \( \gamma(d) \) is estimated using pairs of points which are separated by a distance that is in the same distance class as \( \gamma(d) \). Journel and Huijbregts (1978, pp. 210-212) briefly consider the problem of irregularly spaced observations when the mean is a constant.

In general, as the distance classes grow larger and fewer, our estimator of \( \gamma(d) \) will have less variance but more bias. If the mean is a constant, then making the distance classes larger will just mean we are averaging over more pairs of points. If the mean is not a constant, and we use \( \hat{\gamma}(d) \) given in (8), then the situation is more complicated. First, we should note that the bias of \( \hat{\gamma}(d) \) will not be affected by \( \alpha \), even though \( |x_i(j) - x_i'(j)| \) will not be exactly \( d \) for all \( j \). However, the weights in (8) will change in a non-obvious way as the distance classes change. If, in meeting the constraints in (5), some of the
weights are forced to be relatively large in magnitude, we may end up with a considerably noisier estimator of \( \hat{\gamma}(d) \) than we would have expected based on the size of \( m \). More specifically, if \( \hat{\gamma}^* \hat{\gamma}^* \) is much greater than \( 1/m \), which is the least it can possibly be, then satisfying the constraints in (5) may have caused a substantial increase in the variance of \( \hat{\gamma}(d) \). As \( m \) increases and we have more pairs to fit the same number of constraints, this problem may be mitigated. Therefore, there may be more of a need to use larger distance classes when there are regressors in the model than when the mean is just a constant. In II.8, we will see how noisy \( \hat{\gamma}(d) \) as given in (8) can be when \( m \) is not large.

II.6 Choosing a Form for the Mean Function

In this section, we will briefly discuss the problem of choosing the number and form of the components of \( \hat{f}(\cdot) \). Since the number of constraints in (5) grows as the square of the number of components of \( \hat{f}(\cdot) \), there should be an emphasis on keeping the number of regressors to a minimum. Therefore, it is important to use only regressors which explain a substantial amount of the variation of \( z(\cdot) \) which is not explained by other regressors in the model.

One approach to the problem would be to use standard model-building techniques, pretending that the observations are independent (see Daniel and Wood (1971) or Chatterjee and Price (1977), for example). In particular, partial residual plots may be useful as a way of suggesting possible transformations of the regressors. However, traditional approaches to selecting a subset of regressors to be used in the final model may not be
appropriate due to the lack of independence of the observations. Roughly speaking, model-building techniques strive to minimize the residual variance; that is, that part of the variation in the observations which cannot be explained by the regressors. However, when there are spatial correlations, there is no longer a single number which can describe the residual spatial structure. The traditional estimator of variance, the residual sum of squares divided by the degrees of freedom, will not consider the spatial structure of the data; its expectation will be nearly $\gamma(\infty)$ if the observations are spread out over distances much larger than the effective range (see II.1.3) of the semi-variogram. What we really want to do is to choose the model which "minimizes" the variogram. Of course, two variograms may cross each other, so which one is smallest may be unclear. For interpolation purposes, we are most interested in the variogram over shorter distances. Thus, an appropriate approach to model-building would be to choose the model which "minimizes" the estimated semi-variogram over small distances. If there are only a small number of potential regressors, we may consider estimating $\gamma(\cdot)$ using (18) for all possible subsets, and then choosing the model which "minimizes" $\hat{\gamma}(\cdot)$. Of course, introducing more regressors will tend to lower $\gamma(\cdot)$. Because of this fact and the fact that $\gamma(\cdot)$ can become much more difficult to estimate as the number of regressors increases, a model with more regressors should only be preferred to a model with fewer regressors if its estimated variogram is substantially below the estimated variogram of the smaller model.

In some instances, it may occur that the variogram of the final model will look like white noise (see II.8). This phenomenon can occur if most of the spatial structure of $z(\cdot)$ can be explained by the spatial
structure of the regressors. When the variogram is white noise, problems of interpolation and estimation of regression coefficients are obviously greatly simplified. Thus, in a sense, we may actually simplify the model and reduce unexplained variation by introducing regressors.

II.7 Misspecification of the Relationship Between Covariates and the Dependent Variable

II.7.1 Alternatives to the regression model

Throughout the preceding sections of this chapter, we have assumed the mean of \( z(x) \) is given by \( a_0 + \mathbf{a}' \mathbf{f}(x) \). There are several implicit assumptions in this model. The first assumption is that the relationship between \( z(x) \) and the \( \mathbf{f}(\cdot) \) field can be described by modelling the mean of \( z(x) \) by some function of \( \mathbf{f}(x) \). One way in which this assumption could be incorrect is if the mean of \( z(x) \) depends on the entire \( \mathbf{f}(\cdot) \) field, and not just \( \mathbf{f}(x) \). For example, if \( \mathbf{f}_q(\cdot) \), a component of \( \mathbf{f}(\cdot) \), is an observation with substantial measurement error, we might want to use a smoothed version of \( \mathbf{f}_q(x) \) instead of \( \mathbf{f}_q(x) \) itself as a predictor of \( z(x) \). In general, the mean of \( z(x) \) may be modelled as some complex functional of the \( \mathbf{f}(\cdot) \) field. Another way to model the relationship between \( z(\cdot) \) and \( \mathbf{f}(\cdot) \) is to assume that they form a jointly stationary vector-valued random field. Interpolation based on this model is known as co-kriging (see Journel and Huijbregts (1978)). The relationship between co-kriging and the "regression model" used in this chapter is investigated in Chapter IV. A rather different way to model the relationship between \( z(\cdot) \) and \( \mathbf{f}(\cdot) \) would be to assume \( z(\cdot) \) has a constant mean, and that the variogram depends on \( \mathbf{f}(\cdot) \). For example, if \( \mathbf{f} \) has a single component, we may assume

\[
\gamma(x, x') = \gamma(|x - x'| + c|f_1(x) - f_1(x')|),
\]
where \( c \) is a constant. To date, this model has not been investigated. However, the regression model, because of its simplicity and familiarity, is a tempting one to use, and, in many cases, it may provide an adequate description of the effect of \( f(\cdot) \) on \( z(x) \).

**II.7.2 Misspecification of regression equation**

Given that the mean of \( z(x) \) is some function of \( f(x) \), Equation (1) still makes two additional assumptions about the relationship between \( z(x) \) and \( f(x) \). That is, we have assumed that \( f(x) \) has its components properly transformed, and that the regression coefficients do not depend on \( x \). As noted in II.5.2, plots of the data may suggest possible transformations of components of \( f \). Also, in many cases, we may have physical reasons for believing a linear relationship exists, such as when \( z(\cdot) \) measures temperature, and \( f_q(\cdot) \) measures altitude. Therefore, in many situations, it may be possible to identify approximately correct parametric models for the mean function. It is because we are more likely to be able to specify the mean function than the variogram parametrically that we are interested in the approach used in this chapter of parametrically specifying the mean function but not the variogram.

The problem of the regression coefficients varying in space presents problems not normally encountered in regression analysis, although an analogous problem can occur with regression on time series data. Therefore, we will investigate, in some detail, its effect on interpolators and variogram estimators.

First, let us consider the case where we model the mean as a constant; that is, we assume \( m(x) \equiv \mu \). In many situations, \( m(x) \) may be almost a constant over short distances, but may vary substantially over
the entire range of observations. Therefore, in estimating \( z(x_0) \), we often only use those observations near \( x_0 \) to protect against bias caused by changes \( m(\cdot) \) over longer distances. A local interpolator of the form
\[
\sum_j \lambda_j z(x_j) \quad \text{with} \quad \sum_j \lambda_j = 1
\]
also has the property that the local mean does not have to be explicitly estimated to estimate \( z(x_0) \). Thus, when using such an interpolator, we want our estimator of the variogram over short distances to have similar properties. That is, we want it to work well when \( m(\cdot) \) varies slowly in space, and we do not want it to need an explicit estimate of \( m(x) \). Matheron's estimator of \( \gamma(d) \) given in (2), satisfies these properties, since it is a function only of differences in the \( z(\cdot) \) field over the distance \( d \), and thus does not depend on an estimate of the mean and is largely unaffected by changes in the mean function, as long as they are small over the distance \( d \).

Let us now consider the more general case where we use the model
\[
m(x) = a_0 + a^\prime \hat{\mathbf{f}}(x).
\]
In general, it is possible that any of the coefficients, \( a_0, a_1, \ldots, a_p \), could vary slowly with \( x \), but in many situations, \( a_0 \) is the most likely one to vary with \( x \). For example, spatial variations in \( a_0 \) may reflect the influence of an unobserved covariate which varies slowly in space.

Recall from I.1.3 that when the regression function is correctly specified, an unbiased linear interpolator of \( z(x_0) \) is of the form
\[
\hat{z}(x_0) = \sum_j \lambda_j z(x_j), \quad \text{where}
\]
\[
\sum_j \lambda_j = 1 \quad \text{and} \quad \sum_j \lambda_j \hat{\mathbf{f}}(x_j) = \hat{\mathbf{f}}(x_0).
\]
Let us consider using an interpolator of this form based only on local observations. Such an interpolator will still be approximately unbiased
if the regression coefficients vary slowly in space. Also, if the components of \( \hat{f}(\cdot) \) do not vary greatly over short distances, then this interpolator will be almost unbiased even if the relationship between \( f(x) \) and \( z(x) \) is only approximately linear over restricted ranges of \( f(\cdot) \). Thus, if the \( f(\cdot) \) field exhibits enough spatial continuity, to a certain extent, we protect ourselves against non-linearities in the relationship between \( f(x) \) and \( z(x) \). As in the case of the mean being modelled as a constant, estimators of the form given in (20) do not depend on explicitly estimating the regression coefficients.

It would be desirable to have a variogram estimator which was also insensitive to slow spatial variations in the regression coefficients and did not depend on estimating these coefficients. We see that the estimator of \( \gamma(d) \) given in (18) does not need an estimate of the regression coefficients. Because it is based on differences, it is also insensitive to \( a_0 \) varying slowly in space, and as we noted earlier, \( a_0 \) is the most likely coefficient to vary in space. However, the unbiasedness of \( \hat{\gamma}(d) \) in (18) depends on \( a_1, \ldots, a_p \) being globally constant. So if at least one of \( a_1, \ldots, a_p \) vary slowly in space, estimators of \( z(x_0) \) of the form given in (20) which are based on local observations are still reasonable, but our estimator of \( \gamma(\cdot) \) may not be. Thus, if we suspect that \( a_1, \ldots, a_p \) vary substantially over the region of interest, we probably should not use (18) to estimate \( \gamma(\cdot) \). A possible modification of this estimator would be to divide the region into sub-regions, estimate \( \gamma(\cdot) \) within each sub-region, and then average the estimates from the various sub-regions. This procedure is used in II.8. The unbiasedness constraints in (5) have to be satisfied in each sub-region, and thus, this procedure may produce a much noisier estimator than one which
only satisfies the constraints globally. However, $a_1, \ldots, a_p$ need only be approximately constant within each sub-region rather than globally for this estimator to be almost unbiased. If there are enough data in each sub-region so that the estimator of $\gamma(\cdot)$ within a single sub-region is somewhat precise, such local estimates could also be useful in discovering non-stationarities in $\gamma(\cdot)$. However, non-stationarities are more likely to be identified by local parametric estimators of $\gamma(\cdot)$, as will be discussed in the next chapter.

II.8 Application to Landsat Data

In this section, we will apply the techniques of this chapter to two sets of Landsat data. Each set is made up of 400 pixels (squares) of about one acre each, arranged in a 16 by 25 rectangular grid. Within each pixel, energy reflected from the surface in four different frequency bands of light are observed. Two of the frequencies are in the visible range, and the other two are in the infrared range. In this analysis, the visible and infrared frequencies will be considered separately. The main goal of this analysis is to illustrate the use of the various procedures; it is not meant to be a complete or appropriate analysis of these data. For example, we will assume all variograms are isotropic without attempting to ascertain the validity of this assumption.

As a first step, plots of the data within each of the two regions were made (Figures 1-4). In each case, there is a strong linear relationship between the two variables. Therefore, it does not appear that the model given in Equation (1) is inappropriate. That is, we will assume

$$freq_1(x) = a_V + b_V \times freq_2(x) + e_V(x),$$

and

(21)
Figure II.1: Visible frequencies at Site 1
Figure II.2: Visible frequencies at Site 2
Figure II.4: Infrared frequencies at Site 2
freq3(x) = a_{IR} + b_{IR} \times freq4(x) + e_{IR}(x), \quad (22)

where "V" and "IR" stand for "visible" and "infrared," respectively, and \( e_V(*) \) and \( e_{IR}(*) \) are assumed to be isotropic disturbances. Specifically, we assume

\[
\frac{1}{2} E(e_V(x) - e_V(x'))^2 = \gamma_V(|x - x'|), \text{ and }
\frac{1}{2} E(e_{IR}(x) - e_{IR}(x'))^2 = \gamma_{IR}(|x - x'|).
\]

These models will be referred to as the "full models." We will compare these models to the simpler models

\[
freq1(x) = u_V + e_V^*(x), \quad (23)
\]

\[
freq3(x) = u_{IR} + e_{IR}^*(x), \quad (24)
\]

where \( e_V^(*) \) and \( e_{IR}^(*) \) are also assumed to be isotropic:

\[
\frac{1}{2} E(e_V^*(x) - e_V^*(x'))^2 = \gamma_V^*(|x - x'|), \text{ and }
\frac{1}{2} E(e_{IR}^*(x) - e_{IR}^*(x'))^2 = \gamma_{IR}^*(|x - x'|).
\]

These models will be referred to as the "simple models." We will call the distance between adjacent pixels one. In Figures 5 and 6, the estimates of the various semi-variograms are given for distances 1-8. We could have also obtained estimates of other distances, such as \( \sqrt{2} \), but the distances used are sufficient for our purposes. The semi-variograms of the simple models were estimated using the standard estimator, given in Equation (2). The semi-variograms of the full models were estimated using the unbiased, non-parametric estimator given in Equation (8), where the \( w_j^* \)'s are defined by (6). In all cases, we see that introducing another frequency as a regressor greatly decreased the semi-variogram of the first frequency. That is, the second frequency explains much of the spatial structure of the first frequency. We should also note the
**Figure II.5** - Unbiased non-parametric semi-variogram estimates for visible frequencies

**Site 1**

<table>
<thead>
<tr>
<th>Distance</th>
<th>Estimates of semi-variogram</th>
<th>Full model&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Simple model&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.26</td>
<td></td>
<td>5.38</td>
</tr>
<tr>
<td>2</td>
<td>3.07</td>
<td></td>
<td>10.07</td>
</tr>
<tr>
<td>3</td>
<td>3.70</td>
<td></td>
<td>12.34</td>
</tr>
<tr>
<td>4</td>
<td>4.70</td>
<td></td>
<td>13.19</td>
</tr>
<tr>
<td>5</td>
<td>4.68</td>
<td></td>
<td>13.12</td>
</tr>
<tr>
<td>6</td>
<td>4.22</td>
<td></td>
<td>12.42</td>
</tr>
<tr>
<td>7</td>
<td>3.78</td>
<td></td>
<td>11.56</td>
</tr>
<tr>
<td>8</td>
<td>2.41</td>
<td></td>
<td>11.78</td>
</tr>
</tbody>
</table>

**Site 2**

<table>
<thead>
<tr>
<th>Distance</th>
<th>Estimates of semi-variogram</th>
<th>Full model</th>
<th>Simple model</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.24</td>
<td></td>
<td>5.22</td>
</tr>
<tr>
<td>2</td>
<td>3.13</td>
<td></td>
<td>8.98</td>
</tr>
<tr>
<td>3</td>
<td>4.32</td>
<td></td>
<td>10.79</td>
</tr>
<tr>
<td>4</td>
<td>3.71</td>
<td></td>
<td>12.80</td>
</tr>
<tr>
<td>5</td>
<td>3.99</td>
<td></td>
<td>14.66</td>
</tr>
<tr>
<td>6</td>
<td>5.21</td>
<td></td>
<td>15.45</td>
</tr>
<tr>
<td>7</td>
<td>5.44</td>
<td></td>
<td>17.11</td>
</tr>
<tr>
<td>8</td>
<td>6.50</td>
<td></td>
<td>18.34</td>
</tr>
</tbody>
</table>

<sup>a</sup> Equation (21)

<sup>b</sup> Equation (23)
**Figure II.6** - Unbiased non-parametric semi-variogram estimates for infrared frequencies

### Site 1

<table>
<thead>
<tr>
<th>Distance</th>
<th>Estimates of semi-variogram</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Full model$^a$</td>
<td>Simple model$^b$</td>
</tr>
<tr>
<td>1</td>
<td>3.49</td>
<td>9.40</td>
</tr>
<tr>
<td>2</td>
<td>3.94</td>
<td>18.43</td>
</tr>
<tr>
<td>3</td>
<td>3.47</td>
<td>23.64</td>
</tr>
<tr>
<td>4</td>
<td>3.71</td>
<td>26.22</td>
</tr>
<tr>
<td>5</td>
<td>3.01</td>
<td>27.65</td>
</tr>
<tr>
<td>6</td>
<td>2.55</td>
<td>28.34</td>
</tr>
<tr>
<td>7</td>
<td>1.78</td>
<td>27.75</td>
</tr>
<tr>
<td>8</td>
<td>-0.46</td>
<td>27.98</td>
</tr>
</tbody>
</table>

### Site 2

<table>
<thead>
<tr>
<th>Distance</th>
<th>Estimates of semi-variogram</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Full model</td>
<td>Simple model</td>
</tr>
<tr>
<td>1</td>
<td>2.80</td>
<td>8.54</td>
</tr>
<tr>
<td>2</td>
<td>2.30</td>
<td>14.49</td>
</tr>
<tr>
<td>3</td>
<td>3.73</td>
<td>17.17</td>
</tr>
<tr>
<td>4</td>
<td>4.68</td>
<td>21.29</td>
</tr>
<tr>
<td>5</td>
<td>3.32</td>
<td>24.57</td>
</tr>
<tr>
<td>6</td>
<td>4.58</td>
<td>26.32</td>
</tr>
<tr>
<td>7</td>
<td>7.76</td>
<td>27.14</td>
</tr>
<tr>
<td>8</td>
<td>7.13</td>
<td>25.93</td>
</tr>
</tbody>
</table>

$^a$ Equation (22)

$^b$ Equation (24)
qualitative agreement between the estimates at the two sites, which suggests that the large reductions in the variograms caused by introducing a regressor are not "flukes." The relative differences in the estimates between sites is largest for the full model at the larger distances. These results are not surprising, as variogram estimators degrade as the distance increases, and the unbiasedness constraints in (5) cause an increase in uncertainty as the number of regressors in the model increases.

To get a sense of the stability of these estimators with smaller numbers of observations, these estimators were also computed for subregions of the sites. Each site of 16 x 25 pixels was divided into 4 x 5 rectangular blocks or subregions of 4 x 5 pixels. The semi-variograms were estimated within each subregion at distances 1, 2 and 3. The results for Site 1 are given in Figures 7-10, the results for Site 2 being qualitatively similar. In general, the relative noise is greater at larger distances and for the estimates of the variograms in the full model. In particular, note that $\hat{\gamma}_V(d)$ and $\hat{\gamma}_{IR}(d)$ can be negative, and often are for $d = 3$. The possibility of negative estimates is an undesirable feature of $\hat{\gamma}(d)$ as given in (8) which is not shared by Matheron's estimator of the semi-variogram when the mean is a constant, given in (2). When the block estimates are averaged, at distance 1, the estimates based on all of the data (see Figures 5 and 6) are nearly recovered. In the full models, there are large discrepancies at distances greater than one. In the simple models, the only reason there is any difference is that pairs of points from distinct subregions are not included in the average of the block estimates. In the full models, the unbiasedness constraints in (5) must be met within each block, which introduces additional noise. As a result, even the averages of the block estimates are essentially
Figure II.7 - Unbiased non-parametric semi-variogram estimates by subregion

Site 1 - Visible frequencies

Simple model - Equation (23)

<table>
<thead>
<tr>
<th>Distance 1:</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>7.6</td>
<td>4.6</td>
<td>3.9</td>
<td>5.2</td>
<td>3.2</td>
</tr>
<tr>
<td></td>
<td>5.8</td>
<td>7.2</td>
<td>6.6</td>
<td>2.5</td>
<td>6.6</td>
</tr>
<tr>
<td></td>
<td>5.8</td>
<td>7.1</td>
<td>3.2</td>
<td>2.8</td>
<td>8.5</td>
</tr>
<tr>
<td></td>
<td>7.3</td>
<td>10.3</td>
<td>3.4</td>
<td>3.1</td>
<td>1.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Distance 2:</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>19.6</td>
<td>8.9</td>
<td>9.4</td>
<td>9.5</td>
<td>4.8</td>
</tr>
<tr>
<td></td>
<td>11.9</td>
<td>17.7</td>
<td>9.2</td>
<td>4.2</td>
<td>6.3</td>
</tr>
<tr>
<td></td>
<td>8.1</td>
<td>16.5</td>
<td>4.1</td>
<td>4.1</td>
<td>15.7</td>
</tr>
<tr>
<td></td>
<td>8.6</td>
<td>23.1</td>
<td>6.6</td>
<td>5.1</td>
<td>2.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Distance 3:</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>32.8</td>
<td>11.7</td>
<td>9.0</td>
<td>12.3</td>
<td>3.2</td>
</tr>
<tr>
<td></td>
<td>21.6</td>
<td>24.2</td>
<td>6.3</td>
<td>5.2</td>
<td>10.3</td>
</tr>
<tr>
<td></td>
<td>3.7</td>
<td>26.8</td>
<td>2.7</td>
<td>4.3</td>
<td>21.9</td>
</tr>
<tr>
<td></td>
<td>8.3</td>
<td>27.8</td>
<td>9.6</td>
<td>5.6</td>
<td>4.0</td>
</tr>
</tbody>
</table>

Averages of subregion estimates:

<table>
<thead>
<tr>
<th>Distance</th>
<th>Average estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.29</td>
</tr>
<tr>
<td>2</td>
<td>9.82</td>
</tr>
<tr>
<td>3</td>
<td>12.56</td>
</tr>
</tbody>
</table>
**Figure II.8 - Unbiased non-parametric semi-variogram estimates by subregion**

Site 1 - Infrared frequencies

Simple model - Equation (24)

### Distance 1:

<table>
<thead>
<tr>
<th></th>
<th>4.0</th>
<th>5.9</th>
<th>3.3</th>
<th>3.2</th>
<th>7.1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.6</td>
<td>7.5</td>
<td>3.5</td>
<td>5.5</td>
<td>5.0</td>
</tr>
<tr>
<td></td>
<td>4.3</td>
<td>13.4</td>
<td>5.2</td>
<td>3.5</td>
<td>2.4</td>
</tr>
<tr>
<td></td>
<td>3.3</td>
<td>8.7</td>
<td>4.1</td>
<td>3.6</td>
<td>4.1</td>
</tr>
</tbody>
</table>

### Distance 2:

<table>
<thead>
<tr>
<th></th>
<th>6.1</th>
<th>11.3</th>
<th>7.3</th>
<th>4.3</th>
<th>16.5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6.4</td>
<td>15.5</td>
<td>3.0</td>
<td>8.6</td>
<td>6.8</td>
</tr>
<tr>
<td></td>
<td>7.3</td>
<td>17.5</td>
<td>5.6</td>
<td>5.8</td>
<td>3.5</td>
</tr>
<tr>
<td></td>
<td>8.0</td>
<td>13.1</td>
<td>4.4</td>
<td>6.6</td>
<td>3.5</td>
</tr>
</tbody>
</table>

### Distance 3:

<table>
<thead>
<tr>
<th></th>
<th>10.7</th>
<th>13.8</th>
<th>9.5</th>
<th>5.8</th>
<th>15.2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5.6</td>
<td>13.7</td>
<td>4.1</td>
<td>5.2</td>
<td>8.9</td>
</tr>
<tr>
<td></td>
<td>9.7</td>
<td>9.0</td>
<td>1.6</td>
<td>5.4</td>
<td>5.0</td>
</tr>
<tr>
<td></td>
<td>13.0</td>
<td>4.5</td>
<td>0.7</td>
<td>10.8</td>
<td>3.4</td>
</tr>
</tbody>
</table>

### Averages of subregion estimates:

<table>
<thead>
<tr>
<th>Distance</th>
<th>Average estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.04</td>
</tr>
<tr>
<td>2</td>
<td>8.05</td>
</tr>
<tr>
<td>3</td>
<td>7.78</td>
</tr>
</tbody>
</table>
**Figure II.9 - Unbiased non-parametric semi-variogram estimates by subregion**

Site 1 - Visible frequencies

Full model - Equation (21)

Distance 1:

<table>
<thead>
<tr>
<th></th>
<th>3.0</th>
<th>2.0</th>
<th>1.7</th>
<th>1.3</th>
<th>-0.2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.8</td>
<td>2.0</td>
<td>3.9</td>
<td>2.0</td>
<td>3.9</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>4.6</td>
<td>0.6</td>
<td>3.2</td>
<td>3.2</td>
</tr>
<tr>
<td></td>
<td>2.7</td>
<td>2.3</td>
<td>1.8</td>
<td>3.3</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Distance 2:

<table>
<thead>
<tr>
<th></th>
<th>6.8</th>
<th>2.9</th>
<th>3.5</th>
<th>3.3</th>
<th>0.1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.9</td>
<td>0.1</td>
<td>4.9</td>
<td>2.2</td>
<td>2.7</td>
</tr>
<tr>
<td></td>
<td>2.1</td>
<td>8.4</td>
<td>0.6</td>
<td>2.5</td>
<td>5.1</td>
</tr>
<tr>
<td></td>
<td>-2.2</td>
<td>1.4</td>
<td>4.4</td>
<td>5.0</td>
<td>2.2</td>
</tr>
</tbody>
</table>

Distance 3:

<table>
<thead>
<tr>
<th></th>
<th>12.4</th>
<th>5.1</th>
<th>2.3</th>
<th>-0.1</th>
<th>-0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-29.5</td>
<td>-1.6</td>
<td>0.5</td>
<td>-0.1</td>
<td>-2.6</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>3.8</td>
<td>-0.2</td>
<td>3.2</td>
<td>-17.7</td>
</tr>
<tr>
<td></td>
<td>2.5</td>
<td>8.5</td>
<td>6.7</td>
<td>-4.1</td>
<td>3.8</td>
</tr>
</tbody>
</table>

Averages of subregion estimates:

<table>
<thead>
<tr>
<th>Distance</th>
<th>Average estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.33</td>
</tr>
<tr>
<td>2</td>
<td>2.75</td>
</tr>
<tr>
<td>3</td>
<td>-0.39</td>
</tr>
</tbody>
</table>
**Figure II.10** - Unbiased non-parametric semi-variogram estimates by subregion

Site 1 - Infrared frequencies

Full model - Equation (22)

Distance 1:

<table>
<thead>
<tr>
<th>Distance</th>
<th>1.0</th>
<th>3.4</th>
<th>2.0</th>
<th>2.1</th>
<th>2.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.5</td>
<td>2.9</td>
<td>2.7</td>
<td>5.4</td>
<td>4.6</td>
<td></td>
</tr>
<tr>
<td>3.3</td>
<td>1.7</td>
<td>2.2</td>
<td>1.5</td>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td>1.7</td>
<td>0.6</td>
<td>0.3</td>
<td>3.0</td>
<td>3.1</td>
<td></td>
</tr>
</tbody>
</table>

Distance 2:

<table>
<thead>
<tr>
<th>Distance</th>
<th>0.1</th>
<th>3.2</th>
<th>0.0</th>
<th>4.1</th>
<th>3.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.7</td>
<td>2.3</td>
<td>0.9</td>
<td>8.3</td>
<td>6.4</td>
<td></td>
</tr>
<tr>
<td>3.9</td>
<td>-0.8</td>
<td>1.9</td>
<td>1.1</td>
<td>1.3</td>
<td></td>
</tr>
<tr>
<td>2.8</td>
<td>3.5</td>
<td>0.2</td>
<td>2.1</td>
<td>2.8</td>
<td></td>
</tr>
</tbody>
</table>

Distance 3:

<table>
<thead>
<tr>
<th>Distance</th>
<th>1.7</th>
<th>6.5</th>
<th>1.1</th>
<th>7.0</th>
<th>2.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2.5</td>
<td>-1.4</td>
<td>-5.3</td>
<td>3.3</td>
<td>9.4</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>6.8</td>
<td>1.1</td>
<td>4.6</td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td>6.5</td>
<td>1.6</td>
<td>0.6</td>
<td>4.0</td>
<td>1.7</td>
<td></td>
</tr>
</tbody>
</table>

Averages of subregion estimates:

<table>
<thead>
<tr>
<th>Distance</th>
<th>Average estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.26</td>
</tr>
<tr>
<td>2</td>
<td>2.32</td>
</tr>
<tr>
<td>3</td>
<td>2.61</td>
</tr>
</tbody>
</table>
useless for distances greater than one. However, as noted in II.7.2, the fact that the averages of the block estimates at distance one almost reproduce the estimates based on all of the data indicates that the estimates based on all of the data are not being significantly affected by spatial variations in $b_V$ (or $b_{IR}$).

The properties of the simple and full models were also investigated by seeing how they performed as a basis for a local, linear unbiased interpolator. Note that $\gamma_{IR}(\cdot)$, the semi-variogram for the full model using the infrared frequencies, looks approximately like the semi-variogram of a white noise process at Site 1, at least over distances greater than or equal to one. As noted in II.6, in this situation, it is easy to do interpolation when the disturbance term is white noise. Therefore, only the infrared frequencies were considered.

First, let us consider estimating $z(x_o)$ based on the four adjacent pixels. Assuming the variogram is isotropic, by symmetry considerations, the best estimator of $z(x_o)$ is the average of $z(x)$ at the four neighboring points. Using the full model, a local linear unbiased estimator is of the form

\[ \sum_{j=1}^{n} \lambda_j z(x_j), \quad \text{where} \]
\[ \sum_{j=1}^{n} \lambda_j = 1, \]
\[ \sum_{j=1}^{n} \lambda_j \text{freq}4(x_j) = \text{freq}4(x_o). \quad (25) \]

In the present case, $n = 4$. Subject to these unbiasedness constraints, under the assumption that $e_{IR}(\cdot)$ is a white noise process, the minimum variance estimator will minimize
\[ \sum_{j=1}^{n} \lambda_j^2. \] (26)

Of course, it may not be always possible to satisfy the unbiasedness constraints. Even when the constraints can be satisfied, \( \Sigma \lambda_j^2 \) may be very large, and thus the interpolator will have large variance. Thus, let us define

\[ \hat{z}(x_o) = \begin{cases} \sum_{j=1}^{n} \lambda_j z(x_j), & \text{where } \lambda_j \text{'s are defined by (25)} \\ \text{and (26) if constraints in (25)} \\ \text{can be satisfied and } \sum_{j=1}^{n} \lambda_j^2 < K, \\ \frac{1}{n} \sum_{j=1}^{n} z(x_j), & \text{otherwise.} \end{cases} \] (27)

That is, if \( \Sigma \lambda_j^2 \) is larger than some cutoff value, \( K \), we will estimate \( z(x_o) \) by the local average. This interpolator is compared to the local average interpolator in Figure 11. The total number of interpolations per site is \((16-2) \times (25-2) = 322\). At Site 1, the interpolator defined by (27) is better than the local average when there is no cutoff, and it improves slightly when a cutoff of one is used. At Site 2, the interpolator defined by (27) does negligibly worse than the local average when no cutoff is used, and does somewhat better than the local average when a cutoff of one is used.

This same procedure was repeated using interpolators based on the eight nearest neighbors. In this case, the average of these eight observations is no longer the best interpolator under the simple model, but it will be used anyway for comparison. At both sites, by using eight observations instead of four, the "full model" interpolator improves and the local average interpolator does worse. Thus, we see that when using covariates, there is less of a problem with "oversmoothing" of the data.
### Figure II.11 - Interpolators based on four nearest neighbors

#### Site 1

<table>
<thead>
<tr>
<th>Mean squared error</th>
<th>Local average&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Full model&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Cutoff&lt;sup&gt;c&lt;/sup&gt;</th>
<th># singularities&lt;sup&gt;d&lt;/sup&gt;</th>
<th># &gt; cutoff&lt;sup&gt;e&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>7.47</td>
<td>5.72</td>
<td>∞</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>7.47</td>
<td>5.23</td>
<td>1</td>
<td>2</td>
<td>27</td>
</tr>
</tbody>
</table>

#### Site 2

|                   | 7.69                     | 7.90                  | ∞                 | 2                           | 0                       |
|                   | 7.69                     | 6.41                  | 1                 | 2                           | 34                      |

<sup>a</sup> Average of four nearest neighbors.

<sup>b</sup> Defined by (27).

<sup>c</sup> Only relates to "full model" interpolator.

<sup>d</sup> Number of times unbiasedness constraints in (25) could not be satisfied.

<sup>e</sup> Number of times sum of squared weights exceeded the cutoff.

### Figure II.12 - Interpolators based on eight nearest neighbors

#### Site 1

<table>
<thead>
<tr>
<th>Mean squared error</th>
<th>Local average&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Full model&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Cutoff&lt;sup&gt;c&lt;/sup&gt;</th>
<th># singularities&lt;sup&gt;d&lt;/sup&gt;</th>
<th># &gt; cutoff&lt;sup&gt;e&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8.89</td>
<td>4.48</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>8.89</td>
<td>4.50</td>
<td>1</td>
<td>0</td>
<td>6</td>
</tr>
</tbody>
</table>

#### Site 2

|                   | 9.20                     | 4.95                  | 5                 | 0                           | 0                       |
|                   | 9.20                     | 4.80                  | 1                 | 0                           | 8                       |

<sup>a</sup> Average of eight nearest neighbors.

<sup>b</sup> Defined in (27).

<sup>c</sup> Only relates to "full model" interpolator.

<sup>d</sup> Number of times unbiasedness constraints in (25) could not be satisfied.

<sup>e</sup> Number of times sum of squared weights exceeded the cutoff.
When the cutoff was set at five, the local average smoother was never used as the interpolator based on the full model. Therefore, we can try to predict the mean squared prediction error using our estimate of the semi-variogram \( \gamma_{IR}(\cdot) \). Assuming the disturbance term is white noise, and \( \gamma_{IR}(d) = C_o, \) for \( d \gg 1 \), we have

\[
E(\hat{z}(x_o))^2 = C_o \left( 1 + \sum_{j=1}^{n} \lambda_j^2 \right). \tag{28}
\]

Using \( \hat{\gamma}_{IR}(1) \) as an estimator of \( C_o \), our estimates of the average prediction error are given in Figure 13. We see that the predicted and actual values agree quite well at Site 1, but at Site 2, the actual average squared error is underpredicted by a little over 25%. This underprediction suggests that \( \hat{\gamma}_{IR}(1) \) is too low an estimate of \( \gamma_{IR}(1) \), although the assumption that \( e_{IR}(\cdot) \) is a white noise process appears to be incorrect, which could also account for some of the underprediction of the interpolation error.

Another approach to analyzing the data in the full model is to estimate the regression coefficients by least squares, and then use the residuals as if they were the errors. For example, we can estimate \( a_{IR} \) and \( b_{IR} \) by least squares estimators \( \hat{a}_{IR} \) and \( \hat{b}_{IR} \), and then compute

\[
\hat{e}_{IR}(x) = \text{freq}3(x) - \hat{a}_{IR} - \hat{b}_{IR} \times \text{freq}4(x). \tag{29}
\]

Based on these residuals, we can estimate \( \gamma_{IR}(\cdot) \), and we can interpolate values of \( \text{freq}3(\cdot) \). The semi-variogram can be estimated by

\[
\hat{\gamma}_{IR}(d) = \frac{1}{(2n_d)^{-1}} \sum_{|x_i - x_j| = d} \left( \hat{e}_{IR}(x_i) - \hat{e}_{IR}(x_j) \right)^2, \tag{30}
\]

(see Equation (2) for definitions), which is just an adaptation of Matheron's estimator of the semi-variogram. This estimator is not
Figure II.13 - Prediction of interpolation error of eight nearest neighbor "full model" interpolator

<table>
<thead>
<tr>
<th>Site</th>
<th>( \hat{\gamma}_{IR}(1) )^a</th>
<th>Average sum of squared weights</th>
<th>Predicted^b</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.49</td>
<td>0.25</td>
<td>4.36</td>
<td>4.48</td>
</tr>
<tr>
<td>2</td>
<td>2.80</td>
<td>0.28</td>
<td>3.58</td>
<td>4.89</td>
</tr>
</tbody>
</table>

^a The unbiased non-parametric estimator of \( \gamma_{IR}(1) \).

^b \((1 + \text{Average sum of squared weights}) \times \hat{\gamma}_{IR}(1)\). Assumes \( \gamma_{IR}(d) \) is constant for \( d > 1 \), which appears to be a poor assumption for Site 2 (see Figure II.14).

Figure II.14 - Comparison of Semi-Variogram Estimators

Site 1

<table>
<thead>
<tr>
<th>Distance</th>
<th>Unbiased Estimator^a</th>
<th>Residual Estimator^b</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.49</td>
<td>3.51</td>
</tr>
<tr>
<td>2</td>
<td>3.94</td>
<td>3.79</td>
</tr>
<tr>
<td>3</td>
<td>3.47</td>
<td>3.89</td>
</tr>
<tr>
<td>4</td>
<td>3.71</td>
<td>4.01</td>
</tr>
<tr>
<td>5</td>
<td>3.01</td>
<td>3.74</td>
</tr>
<tr>
<td>6</td>
<td>2.55</td>
<td>3.81</td>
</tr>
<tr>
<td>7</td>
<td>1.78</td>
<td>3.79</td>
</tr>
<tr>
<td>8</td>
<td>-0.46</td>
<td>3.83</td>
</tr>
</tbody>
</table>

Site 2

<table>
<thead>
<tr>
<th>Distance</th>
<th>Unbiased Estimator^a</th>
<th>Residual Estimator^b</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.80</td>
<td>3.90</td>
</tr>
<tr>
<td>2</td>
<td>2.30</td>
<td>4.72</td>
</tr>
<tr>
<td>3</td>
<td>3.73</td>
<td>4.72</td>
</tr>
<tr>
<td>4</td>
<td>4.68</td>
<td>5.20</td>
</tr>
<tr>
<td>5</td>
<td>3.32</td>
<td>5.78</td>
</tr>
<tr>
<td>6</td>
<td>4.68</td>
<td>5.26</td>
</tr>
<tr>
<td>7</td>
<td>7.78</td>
<td>5.80</td>
</tr>
<tr>
<td>8</td>
<td>7.13</td>
<td>5.80</td>
</tr>
</tbody>
</table>

^a Defined in (8).

^b Defined in (30).
unbiased, but if \( \hat{b}_{IR} \approx b_{IR} \), it will be almost unbiased. Since we have 400 observations per site, we can hope that we will have \( \hat{b}_{IR} \approx b \), even if the disturbance terms \( (e_{IR}(x_i), i=1,\ldots,400) \) are not uncorrelated so that \( \hat{b}_{IR} \) is not the BLUE.

The results of this "residual" estimator of \( \gamma_{IR}(\cdot) \) are given in Figure 14, along with the unbiased estimator of \( \gamma_{IR}(\cdot) \) defined in (8). We see, not surprisingly, that the residual estimator is less noisy than the unbiased estimator, especially at longer distances. With the residual estimator, we can see a clear sill value, an asymptotic value of the semi-variogram as distance increases (see I.1.3), at each site. At Site 1, the sill value is about 3.8, and at Site 2, it is about 5.6. The unbiased estimator is too noisy at longer distances to get even an approximate sill value. At shorter distances, the two methods of estimation yield similar estimates at Site 1. However, at Site 2, the unbiased estimator gives much lower values at the shorter distances. In III.8 (Figure III.9), we will see that the unbiased estimator appears to have seriously underestimated \( \gamma_{IR}(\cdot) \) at Site 2. In general, we see that the residual estimator of the semi-variogram may be more helpful in choosing a parametric model for the semi-variogram than the unbiased estimator, since it has considerably lower variance, and its bias may not be too large if the regression coefficients other than the constant term can be estimated well.

We can also use \( \hat{a}_{IR} \) and \( \hat{b}_{IR} \) to produce interpolators of the freq3(\( \cdot \)) field. Let us assume that freq3(x) was not observed, and we wish to estimate it from the other observations. The simplest approach would be to assume the disturbance terms are uncorrelated in
which case, the BLUE of \( \text{freq3}(x) \) is
\[
\text{freq3}(x) = \hat{\mathbf{a}}_{IR} + \hat{\mathbf{b}}_{IR} \times \text{freq4}(x),
\]
where we have ignored the fact that \( \hat{\mathbf{a}}_{IR} \) and \( \hat{\mathbf{b}}_{IR} \) depend on \( \text{freq3}(x) \), since this dependence should be negligible with 400 observations. We will call this estimator the "least squares" interpolator. A somewhat more sophisticated interpolator is given by
\[
\text{freq3}(x) = \hat{\mathbf{b}}_{IR} \times \text{freq4}(x) \\
+ \text{local average of } \{ \text{freq3}(\cdot) - \hat{\mathbf{b}}_{IR} \times \text{freq4}(\cdot) \}.
\]
We will call this estimator the "residual" interpolator. Note that it does not depend on \( \hat{\mathbf{a}}_{IR} \).

In Figure 15, the average squared errors for these two procedures and for the local unbiased estimator defined in (27) (with \( K = \infty \)) are given. Only the \((16-2) \times (25-2) = 322\) pixels not touching the perimeters of the sites were used to compute the average squared error for the least squares estimator so that the result can be compared to the average squared errors of the other two interpolators, which are both based on the 322 "internal" pixels. For the residual estimator in (32), we take the local average of \( \text{freq3}(\cdot) - \hat{\mathbf{b}}_{IR} \times \text{freq4}(\cdot) \) to be the average of this quantity at the eight nearest neighbors of \( x \).

At Site 1, all three interpolators perform equally well. At Site 2, the least squares interpolator does significantly worse than the other two interpolators. The residual estimator of the semi-variogram at Site 1 looks almost like a white noise process, which suggests that the disturbance term, \( e_{IR}(\cdot) \), has almost no spatial structure over distances greater than or equal to one. Thus, the least squares interpolator should do well at Site 1. However, at Site 2, the residual estimator
**Figure II.15 - Comparison of Interpolators**

### Site 1

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Average Squared Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local unbiased(^a)</td>
<td>4.48</td>
</tr>
<tr>
<td>Least squares(^b)</td>
<td>4.42</td>
</tr>
<tr>
<td>Residual(^c)</td>
<td>4.39</td>
</tr>
</tbody>
</table>

### Site 2

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Average Squared Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local unbiased(^a)</td>
<td>4.95</td>
</tr>
<tr>
<td>Least squares(^b)</td>
<td>6.34</td>
</tr>
<tr>
<td>Residual(^c)</td>
<td>4.69</td>
</tr>
</tbody>
</table>

\(^a\) Defined in (27) with \(K = \infty\). Based on eight nearest neighbors.

\(^b\) Defined in (31). Average squared error of the 322 points not on the perimeter of the site.

\(^c\) Defined in (32). Based on eight nearest neighbors.
of the semi-variogram shows a clear increase over the shorter distances. Therefore, values of \( \text{freq3}(\cdot) \) and \( \text{freq4}(\cdot) \) near \( x \) will provide information about \( \text{freq3}(x) \) not contained in the least squares predictor. Thus, it is not surprising that the least squares interpolator is inferior at Site 2. If the regression coefficients were known, then the expected squared error of the interpolator

\[
\hat{\text{freq3}}(x) = a_{IR} + b_{IR} \times \text{freq4}(x)
\]

would be equal to the sill value. If \( \hat{a}_{IR} = a \) and \( \hat{b}_{IR} = b \), then the least squares interpolator should have approximately the same expected squared error. At Sites 1 and 2, the estimated sill value, based on the residual estimators, are approximately 3.8 and 5.6, respectively, and the average squared errors are 4.4 and 6.4, respectively. Thus, at both sites, the estimated sill values slightly underpredict the average squared error. It is not clear whether or not these underpredictions are significant.

An important result is that the residual interpolator did only negligibly better than the local unbiased interpolator using the eight nearest neighbors. The local unbiased interpolator uses only local observations; thus, it is not affected by regression coefficients which vary slowly in space (see II.7.2). However, by using only local observations, some efficiency is lost if the regression coefficients are globally constant. The residual interpolator uses all of the observations to estimate \( b_{IR} \), and also uses local values of the "residual" field to produce an estimator. It is not affected by slow spatial variations in \( a_{IR} \) except to the extent that such variations affect the value of \( \hat{b}_{IR} \). Therefore, if \( b_{IR} \) is globally a constant, we might
expect the residual interpolator to perform better than the local unbiased interpolator. However, with these data, eight nearby observations were able to pick up enough information about $b_{IR}$ to produce nearly as good an interpolator as the residual estimator which uses all of the data to estimate $b_{IR}$. 
References


