12-2003

Determination of Spatial Strata for Environmental Regulatory Purposes

John Edward Daniels
Western Michigan University

DETERMINATION OF SPATIAL STRATA FOR ENVIRONMENTAL REGULATORY PURPOSES

by

John Edward Daniels

A Dissertation
Submitted to the
Faculty of The Graduate College
in partial fulfillment of the
requirements for the
Degree of Doctor of Philosophy
Department of Statistics

Western Michigan University
Kalamazoo, Michigan
December 2003

DETERMINATION OF SPATIAL STRATA FOR ENVIRONMENTAL REGULATORY PURPOSES

John Edward Daniels, Ph.D.

Western Michigan University, 2003

This dissertation introduces spatial strata modelling, a methodology that combines spatial statistics, cluster analysis, and geographic information system theories to analyze the background level of naturally occurring contaminants of concern (COCs). The objective of spatial strata modelling is to divide a geographic area of interest into mutually exclusive geographic zones (spatial strata); with each stratum representing a different level of COC concentration. An estimate of each stratum's COC concentration level, representing an upper regulatory limit, will also be provided. Data provided by the Michigan Department of Environmental Quality describing the spatial location and arsenic concentrations of 211 Michigan sites (arsenic data) is analyzed using the spatial strata modelling method introduced here. In addition, various infill sampling strategies will be investigated using the arsenic data as an example. An optimal infill sampling algorithm is recommended in order to improve the accuracy of the spatial strata modelling method.

ACKNOWLEDGMENTS

I wish to express my gratitude to my advisor Dr. Michael R. Stoline whose continued guidance and support made this possible. I also appreciate the timely advice provide to me by Dr. Robert Buck, Dr. Alan Kehew, Dr. Charles Emerson, and Dr. Daniel Mihalko. My thanks go out to David Slayton and Sarah Hession of the Michigan Department of Environmental Quality who provided me with the data and additional insight necessary to complete this research. I appreciate the patience of my loving wife, Jennifer, who gave me the time I needed. Finally, a special thanks to Katharine Elisabeth who spent many an hour sleeping on Daddy's shoulder so he could keep typing.

John E. Daniels

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

BACKGROUND

1.1 U.S. Environmental Regulatory Agencies and Hazardous Waste Management

Environmental regulatory agencies exist at both the federal (United States Environmental Protection Agency) and state (Michigan Department of Environmental Quality) levels of government. A primary goal of these environmental regulatory agencies is to ensure that risks to human health and the environment are reduced to a minimal level. These environmental risks are regulated in several media including: soil, air, ground water, surface water, etc. A complete review of the responsibilities and activities of these agencies is beyond the scope of this dissertation. However, the topic of hazardous waste management will be discussed.

The State of Michigan (Michigan 1994) defines hazardous waste as:

> ... a combination of waste and other discarded material including solid, liquid, semisolid, or contained gaseous material that because of its quantity, quality, concentration, or physical, chemical, or infectious characteristics may cause or significantly contribute to an increase in

> mortality or an increase in serious irreversible illness or serious incapacitating but reversible illness, or may pose a substantial present or potential hazard to human health or the environment if improperly treated, stored, transported, disposed of, or otherwise managed.

Hazardous waste can contain some compounds that are also a naturally occurring substance (arsenic, lead, mercury, etc.) in addition to man-made substances (PCBs, dioxin, etc.). Hazardous waste may be found in the soil, air, or water media. The list of known hazardous wastes is exhaustive. However, the United States Environmental Protection Agency (USEPA, 1995) provides a good summary of known substances that above a certain concentration are considered to be hazardous to human health and the environment. In the soil media, these concentrations are typically expressed in parts per million (ppm), which is one milligram of contaminant per kilogram of soil.

In managing these hazardous wastes, environmental regulatory agencies have two primary responsibilities:

1. Working with business and industry to maintain compliance with existing statutes that regulate the treatment, storage, or disposal of hazardous wastes.

2. Investigating potentially hazardous waste sites where improper past practices may have created environmental contamination.

1.2 Contaminants of Concern, Closure, and Testing Procedures

Substances that are found in soil, but not yet determined to be hazardous wastes, will be collectively described as "Contaminants of Concern" (COCs). In order to assess if a COC may be considered a health or environmental risk, it is necessary to develop regulatory testing procedures for analyzing the concentration of COCs at a geographic area of interest (GAI). The overall goal of these regulatory testing procedures is to determine whether or not "closure" should be granted to the GAI by the environmental regulatory agency. Closure, in the world of environmental regulation, is an important concept and requires further explanation.

Closure is granted when a GAI is tested and determined to be suitable for a particular use. A closure document specifies the use and any limitations. Closure may be specific to a facility, a property, a regulated unit, or a specific area within a property. Closure does not necessarily mean that the area in closure status is free from risk to human health and the environment with regard to any possible COC. Rather, closure is limited to the specific COCs addressed during the GAI investigation and testing procedure. If certain contaminants, geographical areas, or environmental media were not specifically evaluated under this process, closure will not apply to them.

The procedure for closure evaluation is as follows: (1) determine the default standard for the COC, (2) identify the GAI, (3) collect the sample data from the

GAI, and (4) perform the statistical comparison between the GAI data and the default standard.

A default standard is defined as an upper regulatory limit of a COC's concentration level. This default standard is used throughout a regulatory agency's jurisdiction. A default standard may be categorized as either risk-based (based on the toxicity of the COC to humans), technology-based (based on the limit of remediation), or background-based (based upon the natural level of the COC). This dissertation will focus on background based default standards.

When making a comparison between the GAI data and the default standard, statistical methods are usually utilized. This is because there is a need to make decisions regarding whether or not closure should be granted in the face of uncertainty. This uncertainty occurs because investigators are limited to analyzing only a sample of all surface soil at GAI. However, any conclusions made regarding closure must be made for the GAI (the population), not just for the sample. Statistical methods are designed to facilitate these conclusions by making inferences about a population characteristic (COC concentration level) using only the sample data.

1.3 MDEQ: Guidelines and Testing Procedures

1.3.1 Determination of a Single Statewide Default Standard

In the state of Michigan, soil background based default standards for some naturally occurring COCs were established by the Michigan Department of Environmental Quality (MDEQ 1993). This document calculated the default background standard for 17 different COCs by analyzing data from 70 locations in Michigan (WMD 1991) . These locations are shown in Figure 1. Some of these 70

Figure 1. Michigan Background Soil Survey Locations

locations were sampled more than once. When more than one sample was taken from a location, the mean of the samples was used to represent the COC concentration level for that specific location. For each COC analyzed, the mean and standard deviation from the 70 sample locations were determined. The MDEQ default background standard is defined as the mean plus one standard deviation. For arsenic, the mean (3.0 ppm) plus one standard deviation (2.8 ppm) equals

a default background standard of 5.8 ppm. Partial results from the Michigan Background Soil Survey are provided in Table 1.

Table 1. Michigan Soil Default Background Standards

COC	Default Standard (ppm)
Arsenic	5.8
Copper	32.0
Lead	21.0
Mercury	.13

1.3.2 Testing of Site Data

To evaluate a closure plan, the MDEQ (2002) discusses the use of a confidence interval procedure as follows:

> The 95% upper confidence limit (UCL) of the mean is calculated for each constituent [contaminant] of concern and compared to the regulatory threshold (RT)[default standard].

The 95% UCL is calculated as:

$$95\%\ \text{UCL} = \overline{\text{x}} + t_{(.95,n-1)} \times s/\sqrt{n} \tag{1.1}$$

where:

$\overline{\text{x}}$ is the sample mean, n is the sample size, $s = \sum(x_i - \overline{x})^2/(n-1)$ is the sam-

ple standard deviation, and $t_{(.95,n-1)}$ is the upper critical value of Student's t-distribution with $(n-1)$ degrees of freedom for an area of $\alpha = .05$ in the upper tail.

For example, assume that a GAI in Michigan is sampled to evaluate closure for arsenic. If the 95% UCL of this arsenic sample exceeds 5.8 ppm (from Table 1), one of two alternatives may be chosen: (1) a default standard specific only to the GAI is determined and used or (2) a variety of remediation (cleanup) techniques may be utilized.

With regards to a GAI specific default standard, the MDEQ states in Operational Memorandum #15 (MDEQ, 1993):

> It is acceptable to establish site-specific background concentration higher than the default values. Such sampling should be conducted according to requirements in existence before the issuance of this memorandum. Comparison of site values is made against the mean plus three standard deviations calculated from background samples as provided for in existing ERD guidance regarding verification of soil remediation.

If all concentration levels sampled from the GAI are less than the GAI specific default standard (mean plus three standard deviations), closure may be granted. If closure is still not granted, a remediation technique may be considered. These techniques include, but are not limited to (1) chemical modification of

the soil, (2) removal of contaminants from the soil, (3) soil incineration, (4) soil removal, and (5) site capping. The intent of these remediation techniques is to restore the level of the COC concentration to a level at or below the site specific default standard or to prevent exposure. After remediation, the GAI may be retested.

1.3.3 Arsenic in Michigan Soil

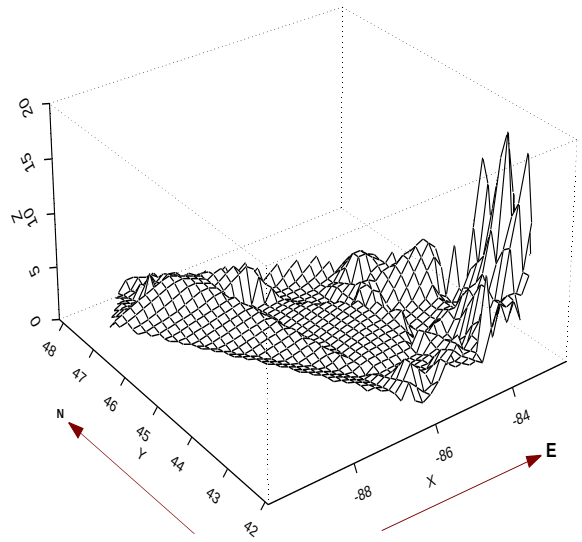
Arsenic is an element that is widely distributed in the earth's crust and cannot be destroyed in the environment. Arsenic can only change its form (inorganic or organic) by reacting with oxygen or other molecules present in air, water, or soil, or by the action of bacteria that live in soil or sediment. Jacobs et al., (1970) determined that the ability of soil to sorb arsenic is related to the free iron oxide content, with arsenic sorption increasing as the free iron oxide content increased, but that the amount of organic matter contributes little to arsenic sorption. Hounslow (1980) reported that inorganic arsenic $(AsO_3)^{-3}$ is 60 times more toxic to humans than organic arsenic $(AsO_4)^{-3}$ and the National Research Council has determined that inorganic arsenic is a carcinogen (NRC, 1999). Analytical methods used by scientists to determine the levels of arsenic in the environment generally do not determine the specific form of arsenic present. Therefore, it is not always known the form of arsenic a person may be exposed to. Similarly, it is often not known what forms of arsenic are present at hazardous waste sites. The

concentration of arsenic in soil varies widely, generally ranging from about 1 to 40 parts of arsenic to a million parts of soil (ppm) with an average level of 5 ppm (ATSDR, 2000). However soils in the vicinity of arsenic-rich geological deposits some mining and smelting sites, or agricultural areas where arsenic pesticides had been applied in the past may contain much higher levels of arsenic (USGS 1999).

This potential for a wide variation in arsenic concentration is demonstrated by Figure 2. Using data provided by the Michigan Department of Environmental Quality, Figure 2 is an estimate (for illustration purposes only) of the naturally occurring arsenic concentration in Michigan's soil. As shown in Figure 2, the naturally occurring arsenic concentration levels are indicated by the elevation of the surface across the longitude (X) and latitude (Y) of Michigan. This plot was produced by linearly interpolating between actual observations and displaying these estimates over a grid. The plot does not indicate the boundaries of the state, but the direction arrows provide a reference for the spatial trend that exists in the naturally occurring arsenic concentrations in soil within the state of Michigan.

Recall the default background standard for arsenic (5.8 ppm) was determined from the Michigan Background Soil Survey (WMD 1991) to be the sum of mean (3.0 ppm) and standard deviation (2.8 ppm) Clearly there are areas in Figure 2, particularly in the lower right corner of the box (southeastern lower peninsula), where the arsenic concentration levels appear to be much higher than the default background standard. The MDEQ has provided the following state-

Figure 2. Surface Plot of Soil Arsenic in Michigan



ment with regards to arsenic concentration in soil:

> Staff of the MDEQ responsible for background soil data compiled and developed the arsenic information in response to numerous inquiries regarding naturally occurring levels of arsenic in soils. Background is defined in Part 201 of Act 451 as "The concentration or level of a hazardous substance which exists in the environment at or regionally proximate to a site that is not attributable to any Release at or regionally proximate to the site". Many instances have occurred where the site-specific background concentration of arsenic was found to be greater than the default background level of 5.8 mg/kg (MERA Op. Memo #15) and the Part 201 Direct Contact criteria of 7.6 mg/kg.

The default background levels for metals are used only to determine if a facility meets a background concentration that is acceptable to the DEQ anywhere in Michigan. If below the default background value, there is no need to take site-specific background samples. In terms of actual cleanup levels, the background concentrations, if properly obtained and approved by the DEQ, become the Part 201 cleanup criteria when they are greater than the corresponding risk-based criteria. A majority of the interest is focused on the thumb region of Michigan, in relation to remediation and dredging projects. Therefore, the inquiries have centered on whether or not soils from the southeast part of Michigan may have slightly higher concentrations of naturally occurring arsenic. In the case of arsenic, it is very important for the DEQ to determine if there are major spatial variations in metals concentrations across the state. Staff of the DEQ need to know if a proposed "background" concentration is indeed legitimate, particularly when it exceeds the current risk-based criteria. In the case of arsenic, it is possible that the natural concentrations in soil in some parts of Michigan are higher than the residential direct contact criteria found in Part 201 of Act 451. Knowing an area's true natural background soil values is important for ensuring thorough cleanups, yet not excessively remediating soil beyond what is natural. In the case of arsenic, and possible

> area associated with a higher concentration, the direct contact criteria (which is very low for this particular metal) is another level of concern to deal with regarding this possible natural variation. The DEQ obviously needs to be certain that this is indeed a true natural variation in concentration associated with specific geographic areas.

By stating, "Therefore, the inquiries have centered on whether or not soils from the southeast part of Michigan may have slightly higher concentrations of naturally occurring arsenic", the MDEQ supports the idea that spatial location may influence higher arsenic concentration levels. In addition, Figure 2 supports this MDEQ statement by providing evidence that spatial location may play a role in the level of arsenic concentration in Michigan soil. In contrast, section 1.3.1 discusses the MDEQ procedure used in determining a single statewide default background standard for arsenic. This procedure does not allow spatial location to be considered as a predictor of an arsenic default background standard. Further, by stating "... yet not excessively remediating soil beyond what is natural", the MDEQ infers that excessive remediation of soil beyond the natural background level should be avoided.

Given this frame of reference, the estimation of several default background standards, as opposed to a single default background standard, is proposed as a research topic. This research would include the development of a new analysis methodology. This methodology would recognize spatial location as a significant

predictor of default background standards. Such a methodology may be able to identify two or more geographic zones of arsenic default background standards in Michigan soil. Each of these zones would be defined by its spatial boundaries and each would have its own estimate of an arsenic default background standard. For example, assume the methodology indicates that southeastern Michigan has a higher arsenic default background standard than the rest of the state. Further assume the methodology indicates that there are two zones of arsenic default background standards; one zone in southeastern Michigan and one zone in the rest of the state. The methodology should estimate the spatial boundaries and an estimate of default background standard for each of these two zones. By including spatial location in this methodology, incidents of excessive soil remediation may be reduced. This reduction would occur since the default background standard for southeastern Michigan would be probably be greater than the present statewide default background standard of 5.8 ppm.

1.4 Spatial Analysis of COCs

If a naturally occurring phenomena (COCs) exhibit characteristics of spatial correlation, then this spatial information should be considered for improved statistical modelling and decision making. As Isaaks and Srivastava (1989) state:

> Unfortunately, most classical statistical methods make no use of the spatial information in earth science data sets. Geostatistics offers a

way of describing the spatial continuity that is an essential feature of many natural phenomena and provides adaptations of classical regression techniques to take advantage of this continuity.

Recall that the MDEQ single statewide default background standard for arsenic, 5.8 ppm, was determined by taking samples of soil from the state of Michigan, determining the arsenic concentration for each sample, then adding the mean and standard deviation of the sample. The spatial location of the data was not used in the determination of the arsenic default background standard.

However, within the realm of environmental statistics, past research has employed various techniques to analyze the concentration level of a COC. For example, the Ontario Ministry of the Environment (OME 2001) used contour plots to estimate the spatial distribution of several COCs at the site of a closed refinery. The spatial distributions depicted contained estimates of the mean, but not of the variance of the estimation error. Utilizing geographic information systems (GIS) and Voronoi diagrams, Burmaster and Thompson (1997) estimated a spatially-averaged mean concentration for a COC on a hypothetical property. GIS is a collection of computer software tools that facilitate, through georeferencing, the integrations of spatial, non-spatial, qualitative, and quantitative data into a database that can be managed under a one system environment (e.g., Burrough, 1986) In this dissertation, the software package Arc/Info© is used for all GIS applications. Voronoi diagrams, also known as Delauney triangles, divide the GAI

into a subsets of points. Let the GAI contain the locations of the sampled data s_i. A Delauney triangle D_i is a subset of the GAI. Within the GAI, let D_i contain all points that are physically closer to sample data point s_i than to any other sample data point in the GAI.

This use of a Voronoi diagram also introduced an important new concept; spatially defining subsets of similar COC concentrations. However, this research appears to be more focused on assessing existing contamination issues, not on the evaluation of default background standards.

1.5 Consideration of Multiple Default Background Standards

If spatial location is indeed a significant predictor of arsenic concentration, then default background standards may be based on spatial location. This idea gives rise to a methodology that can consider multiple default background standards, rather than a single default background standard for an entity such as the state of Michigan. Specifically, the entity is divided into two or more non-overlapping zones such that the zones encompass the entire GAI. Each zone within the state will have its own default background standard. The division of the GAI into separate zones is based on spatial considerations.

To develop this methodology, it is proposed to combine and then extend the research of Burmaster and Thompson (1997) and OME (2001). Let a GAI be broken up into a grid of b contiguous, but non-overlapping spatial "blocks" B_i for

i=1 to b. In this dissertation, a square block B is to be considered a homogeneous experimental unit. It is assumed that the same concentration level of arsenic concentration exists throughout B. Let the B_i units be the basic building units of a zone that contains similar default background standards.

By analyzing spatial modelling estimates of the mean and the variance/covariance of the estimation error for these blocks, the methodology should be able to cluster the B_i based on a chosen criterion (statistically similar COC concentrations). Hence, the blocks are grouped into clusters that become spatial subsets of the GAI. The proposed methodology should also be able to estimate a COC default background standard for each of these clusters. Such a methodology would then allow for the determination and estimation of multiple default background standards, rather than calculating just a single statewide default background standard.

If multiple default background standards are to be determined from a GAI, it is necessary that the proposed methodology answer the following questions: (1) How many unique spatial subsets are there with statistically similar COC concentrations?, (2) What are the spatial boundaries for each spatial subset?, and (3) What is the estimate of default background standard within each unique spatial subset?

In this dissertation, the proposed methodology (known as spatial strata modelling or SSM) to accomplish these topics will be discussed. To illustrate the SSM method, this dissertation will analyze arsenic in Michigan residential surface

soil. The MDEQ has provided a total of 219 arsenic concentrations from spatial locations throughout the state. This data, which will henceforth be called the "arsenic data", has three variables. Let x = the "easting location" or longitude location (in Michigan GeoRef coordinates). Let y = the "northing location" or the latitude location (in Michigan GeoRef coordinates). Let $z(s)$ = the arsenic concentration at the $(x, y) \in s$ location (in ppm).

1.6 Dissertation Outline

A cluster with an estimate of default background standard will be known as a stratum. SSM attempts to determine the quantity, spatial boundaries, and an estimate of the default background standard for the strata within the GAI. The COCs analyzed by SSM will be naturally occurring substances.

Chapter II of this dissertation begins with a review of spatial statistic theory and methods relevant to SSM. Let $z(s) = g(s) + \epsilon(s) : s \in G$ be the realization of a random process where $z(s)$ may be considered a COC concentration level at spatial location $(x, y) \in s$, $g(s)$ is a smooth deterministic function describing any spatial trend in the data, $\epsilon(s)$ is a zero-mean intrinsically stationary random process, and G may be considered as the GAI. Under the assumption that $z(s)$ (or $\epsilon(s)$) is stationary over G, spatial modelling and prediction may be carried out. Briefly, a stationary random process has a constant mean μ, a positive definite covariance function, and a covariance function that depends only on the spatial

distance between two observations and not on the direction between them. Chapter II will contain a discussion of the statistical concepts of stationary random processes and the modelling of such processes.

Assuming $z(s)$ is a stationary random process, the spatial modelling of $z(s)$ begins with the estimation of the empirical semivariogram function $\hat{\gamma}(h)$. For purposes of this dissertation, the empirical semivariogram function will be estimated as:

$$\hat{\gamma}(h) = \frac{1}{2[N(h)]} \sum_{(\epsilon_i, \epsilon_j) \in S(h)} (\epsilon_i - \epsilon_j)^2 \tag{1.2}$$

where:

$\hat{\gamma}(h)$ is the variogram value at separation distance h,

$\epsilon_i \in (x_i, y_i)$ and $\epsilon_j \in (x_j, y_j)$ represent the residuals at the respective spatial locations and are separated by distance h,

h is a vector and

$$h_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{1.3}$$

which is the euclidean distance between two observations,

ϵ_i and ϵ_j at spatial locations (x_i, y_i) and (x_j, y_j),

S(h)= the set of all pairwise $(x_i, y_i), (x_j, y_j)$ spatial locations separated by h, and

N(h)=the number of distinct pairs in $S(h)$.

These residuals exist after fitting $g(s)$ with a function to model any spatial data trends. It should be noted that in the absence of trend, $g(s) = 0$ and e_i and e_j revert to z_i and z_j (COC concentrations) respectively.

The variogram has a wide variety of uses in spatial modelling. These uses will be described in more detail in Chapter II but are briefly mentioned here. The variogram is used to estimate the key spatial model parameters: the nugget, sill, and range. The nugget represents the microscale variation of the random process, or the variogram value at h=0. The sill plus nugget is an estimate of the variance of the random process and is the variogram value at which the variogram function reaches a plateau. The range is the distance h at which this plateau is reached and the random variables are no longer considered to be spatially correlated. In addition, the variogram is used to verify that any estimated trend function $g(s)$ has been properly specified; hence to indicate that $\epsilon(s)$ is indeed a stationary random process. The variogram may also verify that $\epsilon(s)$ is isotropic, which is a requirement of stationarity and means that $\epsilon(s)$ changes only with distance h and not with the direction of vector $\mathbf{h}$. The variogram may also be analyzed for outliers (via the variogram cloud) that may strongly influence the estimation of the spatial model parameters; namely the nugget, sill, and range.

The nugget, sill, and range of the variogram function are estimated by an optimization technique, typically some form of nonlinear model-fitting like weighted least squares. Once these three parameters have been estimated, the process of *kriging* may be used to predict unknown $z(s)$ at known spatial locations $s \in (x, y)$.

Kriging (Matheron, 1963) is a linear interpolation method that can be

considered as a B.L.U.E. (best linear unbiased estimator). When an estimation is made over a square area, comprised of contiguous points, rather than estimating at a single location, kriging becomes "block kriging". Block kriging provides an estimate of block mean ($\hat{z}_B$) and the variance of the estimation error ($\hat{\sigma}^2_B$) for any block within the GAI. A more detailed discussion of kriging and block kriging, including the incorporation of trend into the spatial model, is contained in Chapter II.

After block kriging has been performed, the spatial model should be verified for adequacy of the model fit. Cross-validation is a tool used in this model adequacy determination process. Cross-validation is a technique in which an observed random variable point is removed from the data set for one spatial location. The spatial model then predicts the removed random variable using the remaining data. Let this predicted value be $\hat{z}(s)$. This procedure is done for all z(s) one at a time. All residuals $\epsilon_i = z(s_i) - \hat{z}(s_i)$ resulting from the cross-validation procedure are then analyzed. This analysis is done using various methods that will be discussed in more detail in Section 2.1.11 of Chapter II. Cross-validation cannot prove the spatial model is correct, but it can indicate model mispecification. Chapter II will provide a more detailed discussion of the interpretation of cross-validation results.

Chapter II also contains a discussion of infill sampling theory as it pertains to SSM. An infill sample will supplement the original data. Using the spatial

model, this is accomplished by first identifying locations within the GAI with the highest levels of $\sqrt{\hat{\sigma}_B^2}$, the standard error of the block kriging estimate. An infill sampling strategy then recommends the combination of spatial locations at which additional samples should be taken. These additional samples will serve to lower $\sqrt{\hat{\sigma}_B^2}$ within the GAI; some blocks will have their $\sqrt{\hat{\sigma}_B^2}$ lowered more than other blocks. The intent is to select the combination of additional samples in order to "most effectively" decrease those blocks with the highest $\sqrt{\hat{\sigma}_B^2}$. This dissertation's definition and use of the phrase "most effectively" will be discussed in further detail in Chapter II. In order to eliminate an exhaustive sampling process, which would evaluate all possible combinations of infill samples, Chapter II contains a discussion of the *greedy/sequential exchange* algorithm. This algorithm was introduced by Aspie and Barnes (1990) and is designed to replace an exhaustive sampling process with a more time efficient procedure.

The block kriging procedure will produce an estimate of block mean $\hat{z}_B$ and an estimate of variance of the block estimation error $\hat{\sigma}_B^2$ for each spatial block in the GAI. Because these blocks are spatially correlated, caution must be exercised by avoiding any parametric or non-parametric statistical testing or estimation procedure that requires independence among the random variables measured across the blocks. Hence, a measurement that can quantify the statistical distance between the spatial blocks must be determined. Chapter II contains a discussion of dissimilarity coefficients (DCs) which will be used for this purpose.

Using these DCs, SSM will employ cluster analysis to group the spatial blocks with similar COC concentrations. Before beginning SSM, the actual number of clusters was not known. Chapter II contains a discussion of the issues pertaining to the selection of an appropriate clustering technique. These issues include the selection of: (1) a partitioning or hierarchical clustering method, (2) an agglomerative or divisive clustering method, and (3) a linkage method for defining the distance between two clusters.

A partitioning cluster method assumes the number of clusters is known before the cluster analysis is performed. A hierarchical cluster method has no *apriori* knowledge as to the number of clusters in the data. Let there be n objects (blocks) to be analyzed.

An agglomerative method starts with each block as its own cluster. Hence there are n clusters. The algorithm then combines the most similar blocks in a sequential fashion. This process continues until all blocks belong to the same cluster with n objects in this cluster.

A divisive cluster method takes the opposite approach. All blocks start in the same cluster. The divisive clustering method then splits off the most dissimilar blocks until there are finally n clusters. Each cluster contains only one block.

A linkage method defines the distance between two clusters. Let cluster A and cluster B be considered as two arbitrary clusters. The distance between cluster A and cluster B may be defined by the minimum dissimilarity between cluster A

and cluster B. This is determined by selecting the block belonging to cluster A and the block belonging to cluster B that provides the smallest dissimilarity between all possible combinations. This linkage method is known as single linkage. Similarly, linkage may defined by other commonly used methods such as complete linkage (maximum dissimilarity), centroid method (Sokal and Michener, 1958), and weighted average linkage (Sokal and Sneath, 1963), etc.

Chapter II also contains a review of theory pertaining to the estimation of the 95^{th} percentile for lognormally distributed and the 95% upper prediction limit for normally distributed and non-parametrically distributed random variables. These statistics will later serve as the basis for estimating multiple default background standards.

Chapter III contains a discussion of the development of SSM by first establishing the size and orientation of the spatial blocks that will be overlayed on the GAI. These blocks will define the spatial boundaries for each cluster. Issues to be explored include: (1) a justification for the use of spatial blocks (as opposed to points), and (2) the determination of the physical size of the spatial blocks, and (3), the orientation of the spatial blocks of the GAI. Using the selected spatial model, each spatial block will have an estimate of block mean $\hat{z}_B$ and variance of the block estimation error $\hat{\sigma}_B^2$ determined.

Chapter III contains a discussion of four new infill sampling algorithms (*MinMean, MinMed, MinMax, MinVar)* that will use the results of the spatial

model. For a given infill sample size (n), the *MinMean* algorithm determines the spatial locations of the n samples that will minimize the mean of all $\sqrt{\hat{\sigma}_B^2}$. Similarly, the *MinMed, MinMax,* and *MinVar* algorithms will determine the spatial locations of the n infill samples that will minimize the median, maximum, and variance respectively of all $\sqrt{\hat{\sigma}_B^2}$. A ranking based evaluation method will be introduced in Chapter III. This method will be used to compare and rank each algorithm against its three competing algorithms based on minimizing the mean, median, maximum and variance of all $\sqrt{\hat{\sigma}_B^2}$. The infill sample sizes evaluated will be $n = 10, n = 20$, and $n = 30$. For each of these sample sizes, the algorithm with the lowest total rank will be chosen as the "most effective" algorithm.

Two new statistics are proposed in Chapter III. The first statistic is an estimate of the covariance of the block estimation error $\hat{\gamma}(B_i, B_j)$ for two arbitrary spatial blocks B_i and B_j. $\hat{\gamma}(B_i, B_j)$ will account for the spatial correlation that exists between two arbitrary blocks B_i and B_j and is a component in the second new statistic: the spatial dissimilarity coefficient d_{ij}. The proposed SSM algorithm utilizes this spatial dissimilarity coefficient d_{ij} for all pairwise spatial blocks B_i and B_j as follows:

$$
\begin{aligned}
d_{ij} &= \quad 0 \quad \text{for } i = j \\
d_{ij} &= \left|(\hat{z}_{B_i} - \hat{z}_{B_j})\right| \Big/ \sqrt{\hat{\sigma}_{B_i}^2 + \hat{\sigma}_{B_j}^2 - 2\hat{\gamma}(B_i, B_j)} \quad \text{for } i \neq j
\end{aligned} \tag{1.4}
$$

where:

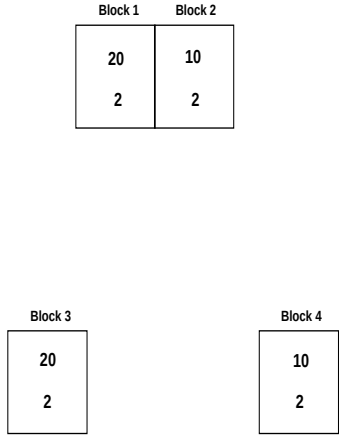
$\hat{z}_{B_i}, \hat{z}_{B_j}$ are the estimates of block mean for block B_i and block B_j.

$\hat{\sigma}^2_{B_i}, \hat{\sigma}^2_{B_j}$ are the variance of the block estimation error for block B_i and block B_j.

$\hat{\gamma}(B_i, B_j)$ is the covariance of the block estimation error for block B_i and block B_j.

The need to include $\hat{\gamma}(B_i, B_j)$ in d_{ij} can be illustrated as follows.

Figure 3. Four Spatial Blocks



Let Figure 3 depict four spatial blocks (Block 1, Block 2, Block 3, and Block 4). The top value in each block is the spatial model's estimate of mean COC concentration level ($\hat{z}_B$). The bottom value in the block is ($\hat{\sigma}^2_B$) resulting from the spatial model. Chapter III will discuss $\hat{\gamma}(B_i, B_j)$ in more detail and show that as the distance between two blocks increases, $\hat{\gamma}(B_i, B_j)$ will monotonically decrease until the distance between the blocks is equal to or greater than the range. At this point the blocks are considered statistically independent and $\hat{\gamma}(B_i, B_j) = 0$. For

illustration purposes let Block 3 and Block 4 be separated at a distance greater than the range so that $\hat{\gamma}(B_3, B_4) = 0$. Let $\hat{\gamma}(B_1, B_2) = 1$.

Without including the expression $\hat{\gamma}(B_i, B_j)$ in the dissimilarity coefficient:

$$d_{12} = d_{34} = |20 - 10| \Big/ \sqrt{2 + 2} = 5$$

However, when $\hat{\gamma}(B_i, B_j)$ is included:

$$d_{12} = |20 - 10| \Big/ \sqrt{2 + 2 - 1} = 7.5$$

$$d_{34} = |20 - 10| \Big/ \sqrt{2 + 2 - 0} = 5$$

As stated by Stephan (1934):

> Data of geographic units are tied together, like bunches of grapes, not separate, like balls in an urn.

Hence, blocks that are closer together in space should be more similar than blocks that are spatially far apart. As shown in Figure 3, Block 1 and Block 2 are closer together in space and should be less dissimilar than Block 3 and Block 4. However, in this example $\hat{z}_{B_1} - \hat{z}_{B_2} = \hat{z}_{B_3} - \hat{z}_{B_4}$ and $\hat{\sigma}^2_{B_1} = \hat{\sigma}^2_{B_2} = \hat{\sigma}^2_{B_3} = \hat{\sigma}^2_{B_4}$. The blocks that are closer together (Block 1 and Block 2) are not more similar than the blocks that are farther apart (Block 3 and Block 4). Thus, the dissimilarity coefficient includes the covariance of the block estimation error $\hat{\gamma}(B_i, B_j)$ in order to account for the spatial distance between the blocks when calculating the dissimilarity measurement. Without including $\hat{\gamma}(B_i, B_j)$, potentially important spatial information is ignored in the estimation of the d_{ij}.

Using a cluster analysis method, Chapter III continues discussion of the SSM method, by describing how the d_{ij} will be grouped together in order to identify blocks of similar COC concentration. Each individual block then belongs to one of an undetermined number of clusters. The spatial boundaries for each cluster are defined by its block membership. Because there is no *apriori* knowledge of the correct number of clusters, a discussion of the criterion proposed by Calinski and Harabaz (1974) will be included in Chapter III. This criterion seeks to identify the correct number of clusters by attempting to maximize the following:

$$C = \Big(B/(g-1)\Big) \Big/ \Big(W/(b-g)\Big) \tag{1.5}$$

in which:

$B = \sum_{i=1}^{g} a_i(\hat{z}_{B_{i.}} - \hat{z}_{B_{..}})^2$ = sum of squares between clusters in which $\hat{z}_{Bi.}$ represents the average of the block estimates ($\hat{z}_{B_j}$) for all blocks $B_j \in$ Cluster C_i, $\hat{z}_{B_{..}}$ represents the average of the block estimates over the entire GAI, a_i represents the total number of blocks $B_j \in$ Cluster C_i, and

$W = \sum_{i=1}^{g} \sum_{j=1}^{a_i}(\hat{z}_{B_{ij}} - \hat{z}_{B_{i.}})^2 =$ sum of squares within clusters in which $\hat{z}_{B_{ij}}$ represents the individual blocks $B_j \in$ Cluster C_i, $g =$ total number of clusters, and $b =$ the total number of blocks.

Once the correct number of clusters has been determined, the block membership of each cluster will define that cluster's spatial boundaries.

In the next step, using GIS methods (Arc/Info©), the total area of each cluster may be determined. In addition, using GIS the percent contribution of

each member block to its total cluster area may also be determined. This percent contribution will be assigned to each block and be known as a *spatial weight*. These spatial weights will then be used to estimate the mean and variance of the estimation error of the COC concentration level within each cluster. Once a cluster has been assigned an estimate of mean and variance of estimation error, the cluster becomes a *stratum*.

To complete the creation of a stratum, it is also necessary to estimate each stratum's default background standard. Depending upon the distribution of the observed data, estimates of the 95^{th} percentile or 95% upper prediction limit are determined for each stratum. These statistics will be used to represent the estimates of multiple default background standards within the GAI.

Chapter IV contains an illustration of the SSM method applied to the Michigan residential surface soil arsenic data. This data set specifically contains 219 spatial locations (x,y) and the corresponding arsenic concentration levels (ppm) which are currently utilized by the MDEQ for regulatory purposes.

Using the SSM method, four different configurations describing spatial boundaries of the arsenic strata within the state of Michigan are described. Estimates of default background standards for each strata configuration are also presented.

Chapter IV also contains details describing the execution of the competing infill sampling algorithms *MinMean, MinMed, MinMax,* and *MinVar* using the

arsenic data. Each algorithm is evaluated using the method described in Chapter III. These four infill methods are then compared utilizing the arsenic data. Using the recommended infill strategy, infill sample locations for $n = 10, n = 20$, and $n = 30$ additional samples in Michigan will be given for the arsenic data.

Chapter V includes a summary of the SSM method. The results obtained by performing SSM on the arsenic data are also discussed. Chapter V also includes a discussion of additional new research topics that could potentially enhance the existing SSM method. These ideas include: (1) estimation of an upper confidence limit of the mean (UCL) using spatial estimates, (2) sequential testing of spatial blocks, (3) a discussion of additional infill sampling algorithms, and (4) development of a polygon kriging procedure for the SSM method.

CHAPTER II

Existing Theory Relevant to Spatial Strata Modelling

2.1 Review of Spatial Statistical Theory and Methods

2.1.1 Stationary Random Processes

Let Z(s) represent the realization of a random process in two dimensional space with spatial location $(x, y) \in s$. In order to accurately model spatially correlated data, the random process $Z(s)$ must be considered a stationary random process. A stationary random process has three main properties (stationarity, positive definite covariance structure, and isotropy). These properties are outlined by Cressie (1993) and Isaaks and Srivastava (1989) and are repeated here:

The first property is that Z(s) must be at least second order stationary. (2.1)

Let $Z(s) : s \in G$ represent the realization of a random process where G represents a geographic area of interest (GAI). Let $(x_i, y_i) \in s_i$ and let $(x_j, y_j) \in s_j$. Let h_{ij} represent the separation distance between s_i and s_j such that $h_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$. Then $Z(s)$ is second order stationary over G if (1) $E[Z(s)] = \mu$, for all $s \in G$, meaning that the mean is constant over G and (2) $E[Z(s_i) - \mu][Z(s_j) - \mu] = \text{Cov}(h) = C(h)$, meaning that the covariance function $C(h)$ depends only on h, the distance between the two spatial locations as defined

by (1.3).

The second property is that the covariance function of Z(s) must be positive definite.

(2.2)

Let $C(h) = C(s_i - s_j)$, the covariance function between two spatial points $(x_i, y_i) \in s_i$ and $(x_j, y_j) \in s_j$ separated by h as defined by (1.3). As stated by Cressie (1993), to be considered a positive definite function:

$$\sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j C(s_i - s_j) \geq 0 \tag{2.3}$$

for any finite number of spatial locations $(x_i, y_i) \in s_i : \{i = 1, ..., m\}$ and any set of m real numbers $a_i : \{i = 1, ...m\}$.

The third property is that the covariance function of Z(s) must be isotropic.

(2.4)

If $C(h)$ is a function of distance h only and not the direction between $(x_i, y_i) \in s_i$ and $(x_j, y_j) \in s_j$ then $C(h)$ may be considered an isotropic covariance function.

2.1.2 The Variogram

To determine if the properties necessary for a stationary random process are satisfied, the spatial correlation of $Z(s)$ will be examined as a function of $\mathbf{h}$, a vector describing distance and direction between points (x_i, y_i) and (x_j, y_j). This

dissertation will estimate and model the empirical omnidirectional semivariogram, $\gamma(h)$, a function which describes how the observed data $z(s)$ are correlated with vector $\mathbf{h}$.

If (1.3) represents the euclidean distance between locations (x_i, y_i) and (x_j, y_j) the estimated empirical semivariogram function $\hat{\gamma}(h)$, originally defined by Matheron (1962), is half the average of the squared difference between all observations separated by a distance h. Let

$$\hat{\gamma}(h) = \frac{1}{2[N(h)]} \sum_{S(h)} (z_i - z_j)^2 \tag{2.5}$$

where:

$\hat{\gamma}(h)$ = variogram value at distance h.

$S(h)$ = the set of all pairwise $(x_i, y_i), (x_j, y_j)$ spatial locations separated by distance h.

$N(h)$ = the number of distinct pairs in $S(h)$

$z_i, z_j =$ observed random variables measured at spatial locations $(x_i, y_i)$and (x_j, y_j) respectively separated by distance h.

By (2.1), a stationary random process has a covariance that remains constant at a given distance h. The covariance of a random variable with itself is the variance. Thus, a stationary random process has a constant variance at which $\mathrm{Var}\big(z(s)\big) = \sigma^2$. Assuming the random process is second order stationary, the

expected value of (2.5) can be written as:

$$E[\hat{\gamma}(h)] = E\left[\frac{1}{2N(h)}\sum_{S(h)}(z_i - z_j)^2\right]$$
$$= \sigma^2 - \sum_{S(h)}\frac{1}{N(h)}\left[\mathrm{Cov}(z_i, z_j)\right] \tag{2.6}$$

For comparison purposes, let z_i and z_j be spatially correlated random variables from a second order stationary random process separated by h. By examining the equation $\frac{1}{2}\,\mathrm{Var}(z_i - z_j)$:

$$\frac{1}{2}\,\mathrm{Var}(z_i - z_j) = \frac{1}{2}[\mathrm{Var}(z_i) + \mathrm{Var}(z_j) - 2\mathrm{Cov}(z_i, z_j)]$$
$$= \sigma^2 - \mathrm{Cov}(z_i, z_j), \tag{2.7}$$

it can seen that the expectation of the semi-variogram is the mean of one half the variance of the difference between two spatially correlated random variables in a second order stationary random process separated by h.

Because of (2.1), σ^2 is nonstochastic and a constant. Hence, there exists a one-to-one relationship between the variogram estimate $\widehat{\gamma}(h)$ and $\mathrm{Cov}(z_i, z_j)$ where:

$$\mathrm{Cov}(z_i, z_j) = \sigma^2 - \hat{\gamma}(h) \tag{2.8}$$

Thus, the covariance between two random variables may be determined from the variogram function.

The semi-variogram graph describes the relationship between the difference in sample values z_i and z_j versus their separated distance h. This is, effectively,

an approximation to the distance function based on the sample data. The prefix *semi* comes from the $\frac{1}{2}$ in the equation. However, it has become common to refer to the semivariogram simply as the variogram. Throughout this dissertation, the term variogram will be used even though the term is theoretically incorrect.

2.1.3 EDA - Data Transformation and Trend Removal

Exploratory Data Analysis (EDA) of the random process $Z(s)$ usually begins with the empirical variogram function as given in (2.5). If the variogram is an increasing function, this may be an indication of a random process that is not second order stationary. If this situation occurs, a transformation of the observed data, a removal of any spatial trend from the data, or a combination of both may be considered.

With regards to a transformation of the observed data $z(s)$, contaminants of concern (COCs) are often observed to have a positively skewed distribution (Gilbert 1987). It is not necessary to transform positively skewed data before analyzing the variogram. However, if the z(s) are positively skewed (whether or not the data follow a lognormal distribution), taking the natural logarithm (log_e) of the sample data tends to provide a more stable variogram. As stated by Clark and Harper (2000):

> By at least approximately normalising the data, the calculation of 'semi-variance' [sic: variogram] becomes a lot more sensible and more

stable.

To continue EDA, the z(s), transformed or untransformed, should be examined for any spatial trends in the data that may contribute to an increasing variogram function. With regards to spatial trends, let the observed random variable of interest $z(s)$ be considered as:

$$z(s) = g(s) + \epsilon(s) \tag{2.9}$$

in which $g(s)$ is a smooth deterministic function describing the systematic aspect of the process, called the trend, or drift, and $\epsilon(s)$ is the residual after fitting $g(s)$. In this situation, the requirement for second order stationarity falls on $\epsilon(s)$. The three mathematical requirements for second order stationary from Section 2.1.1 are the same for $\epsilon(s)$ as was for $z(s)$, with the exception being that $E[\epsilon(s)] = 0$ instead of $E[z(s)] = \mu$.

If $z(s) = g(s) + \epsilon(s)$ is the chosen model, the residuals $\epsilon(s)$ of the fitted trend model will be used in the variogram function. Now modify (2.5) such that:

$$\hat{\gamma}(h) = \Big[\frac{1}{2N(h)} \sum_{S(h)} (\epsilon_i - \ \epsilon_j)^2\Big] \tag{2.10}$$

The focus is then on selecting a trend model that provides residuals that are non-increasing when analyzed via the empirical variogram.

The selection of an appropriate trend model is subjective. Some options include, but are not limited to, ordinary or generalized least squares regression, generalized additive models, local regression models (loess), etc. The trend model

should be selected with parsimony in mind. Select the simplest model that provides residuals that satisfy the necessary variogram requirements (nonincreasing, isotropic, etc.)

Because the residuals are not independent, the usual regression diagnostics (residual plots, etc.) will not provide valid information. However, residual plots versus each component defining spatial location (x and y) should still be examined to verify that any apparent spatial trend has been removed from the model. In addition, Kaluzny, etc. (1998) recommend including a loess smoothing curve to the residual plot to assist in the trend identification.

The loess method, which stands for locally weighted regression scatter plot smoothing, was developed by Cleveland (1979). The loess smoothing curve is obtained by fitting successive linear regression functions in local neighborhoods. The method is similar to the moving average and running median methods in that it uses a neighborhood around each X value to obtain a smoothed Y value corresponding to that X value. If a trend model is fit and the smoothing curve indicates a spatial trend, then this suggests that perhaps a higher order regression model or data transformation may be necessary.

2.1.4 The Variogram Cloud

The variogram cloud (Gandin, 1963) is a diagnostic tool that can be used to look for potential outliers and to assess random process variability with increas-

ing distance h. The variogram cloud plots the function $(z_i - z_j)^2$ on the y-axis versus the separation distance h between all observations z_i and z_j. Cressie (1993) demonstrates the usefulness of considering the square-root-differences cloud as an x-y plot with $\sqrt{|z_i - z_j|}$ on the y-axis and distance h on the horizontal axis. Intuitively, observations that are physically close might be expected to have similar values, but this is not always the case. Typically, the variogram cloud can identify observations that are spatially close but have a large differences in the random variable values.

2.1.5 Anisotropy

Anisotropy is present when the spatial correlation of the random process changes with both magnitude and direction of vector **h**. Clearly, this situation is a violation of the requirements for stationarity as stated by (2.4). Anisotropy is an indication of the correlation of the random process evolving differently in different directions in space. This phenomena can be detected by comparing variogram functions over different spatial directions. Typically, the directions examined are 0 degrees - north/south, 45 degrees - northeast/southwest, 90 degrees - east/west, and 135 degrees southeast/northwest. However, smaller increments of direction may be used depending upon the context of the data.

One type of anisotropy, zonal anisotropy, exists when the variance of the random process changes with direction. This situation may be corrected by de-

trending the data. Another type of anisotropy, geometric anisotropy, occurs when the separation distance h at which the $z(s)$ are no longer correlated changes with direction. Geometric anisotropy is generally corrected by a linear transformation of the separation distance h to an equivalent isotropic model. Detailed discussions of anisotropy are beyond the scope of this dissertation. The reader is referred to Isaaks and Srivastava (1989) and Kaluzny et al (1998) for a more detailed discussion of anisotropy and adjustments necessary to build an isotropic model.

2.1.6 Variogram Parameters and Models

To build the spatial model, it is necessary to estimate the type of variogram function and to estimate three variogram parameters (nugget, range, and sill). These three parameters will further define the variogram function and together with the variogram function will define the covariance structure of the data.

Nugget

At distance $h = 0$, the value of the variogram $\gamma(h)$ is theoretically $\gamma(0) = 0$. However, several factors, including measurement error etc., may interfere with this relationship. These factors may cause sample values separated by extremely small distances to be quite dissimilar. This situation would contribute to discontinuity at the origin of the variogram. In order to address this situation, the nugget c_o is included as a variogram parameter. The nugget is estimated from the variogram

function as the value:

$$c_0 = \hat{\gamma}(h), \qquad h = 0, \tag{2.11}$$

where $\hat{\gamma}(h)$ is the variogram of a second order stationary random process.

Sill

The sill c_s is part of $\hat{\gamma}(h)$ when the variogram reaches a plateau. This plateau may be expressed as:

$$\lim_{h \to \infty} \hat{\gamma}(h) = c_o + c_s = \sigma^2 \tag{2.12}$$

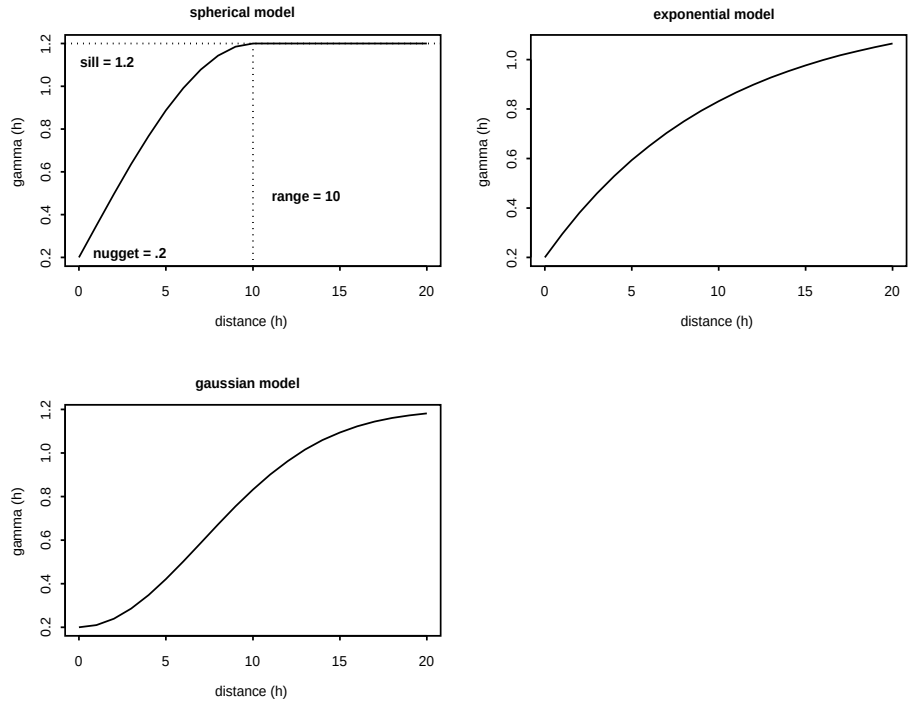
where c_o is the nugget, c_s is the sill, and $\sigma^2 = \text{Var}[z(s)]$ is the variance of the random process.

Range

The range a is the distance at which the variogram estimate reaches its plateau such that $\hat{\gamma}(a) = c_o + c_s = \sigma^2$. Using (2.8), at $h = a$, $\text{Cov}(z_i, z_j) = \sigma^2 - \sigma^2 = 0$. Hence, the range is the distance $h = a$ at which $\text{Cov}(z_i, z_j) = 0$ and the data are no longer considered correlated.

There are three variogram models commonly used to define a variogram function. These models (spherical, exponential, and gaussian) are shown in Figure 4 with the spherical variogram illustrating the nugget, range, and sill parameters.

Figure 4. Theoretical Variogram Models



These three variogram models, along with other variogram models, are discussed in Cressie (1993) and in Isaaks and Srivastava (1989) and are defined as follows:

Spherical Variogram Model:

$$\gamma(h;\theta) = \begin{cases} c_0, & h = 0 \\ c_0 + c_s\big((1.5)(h/a) - .5(h/a)^3\big), & h \le a \\ c_0 + c_s, & h \ge a, \end{cases} \tag{2.13}$$

Exponential Model:

$$\gamma(h;\theta) = \begin{cases} c_0, & h = 0 \\ c_0 + c_s\big(1 - \exp\left(-3h/a\right)\big), & h \neq 0, \end{cases} \tag{2.14}$$

Gaussian Model:

$$\gamma(h;\theta) = \begin{cases} c_0, & h = 0 \\ c_0 + c_s\big(1 - \exp\left(-3h^2/a^2\right)\big), & h \neq 0. \end{cases} \tag{2.15}$$

where c_0 is the nugget, c_s is the sill, and a is the range.

The separation distance h is a continuous variable with an infinite number of possible measurements. As stated by Kaluzny et al (1998), the construction of the variogram requires consideration of an appropriate lag increment for h, a tolerance for the lag increment, and the number of lags (L) over which the variogram will be calculated. The lag increment defines the distances at which the variogram is calculated. The tolerance establishes distance bins for the lag increments, to accommodate unevenly spaced observations. The number of lags in conjunction with the size of the lag increment will define the total distance over which a variogram is calculated. There are two practical rules (Journel and Huijbregts, 1978) that should be considered when choosing the lag increment and number of lags. One rule is that the experimental variogram should only be considered for separation distances h in which the number of pairs is greater than 30. The second rule is that the maximum distance for an experimental variogram should not exceed $D/2$, where D is the maximum distance h over the field of data.

Each of the three variogram models shown in Figure 4 satisfy the positive definiteness requirement (2.2) for the covariance function of the random process $Z(s)$. It should be noted that other variogram models (linear, power, hole effect, etc.) exist but are not strictly positive definite. Within these models, there are combinations of nugget, range, and sill that could produce a non-positive definite covariance matrix. This condition does not exclude these variogram functions from consideration, but caution should be exercised when any of these variogram models are being considered.

2.1.7 Estimation of the Variogram Function

A variogram function can be modelled by an optimization technique, typically some form of nonlinear model fitting. Cressie (1985) describes weighted least squares and generalized least squares approaches for variogram fitting. Zimmerman and Zimmerman (1991) compare several estimation methods and conclude that some form of weighted nonlinear least squares or ordinary nonlinear least squares is usually as good as many of the more complicated and computationally intensive methods.

In addition, it should be noted that because $\hat{\gamma}(h)$ is a mean, $\hat{\gamma}(h)$ is sensitive to outliers (Cressie, 1993). Robust variogram estimators have also been proposed by Cressie and Hawkins (1980), Dowd (1982), and Armstrong and Delfiner (1980). The use of these robust estimators is a matter of personal preference and are not

utilized in this dissertation.

The process of modeling a variogram, particularly in regards to detecting anisotropy, relies greatly on visual interpretation of the variogram, rather than on any procedural methods. It is difficult to construct statistical tests in geostatistics because of the spatial dependence between observations. The points of the sample variogram are correlated, as noted by Jowett (1955). Switzer (1984) proposes a few tests and confidence limits for variogram parameters. The idea is to linearly transform the data to uncorrelated quantities of constant variance and then consider certain rank orderings. For example, to test for the range, select a subset of data points whose interpoint distances all exceed the range; then test for randomness based on the rank correlation between $|z(s_1) - z(s_2)|$ and $|s_1 - s_2|$ based on the Spearman rank correlation statistic.

Cressie (1993) states that at a given distance h, $[z_i - z_j]^2 \Big/ 2\gamma^*(h)$ follows a χ^2 distribution with 1 degree of freedom where $\gamma^*(h)$ is value obtained from the fitted variogram model.

When evaluating the fit of the variogram model, there are several competing goodness-of-fit tests. The residual sum of squares (SSR) sums the differences between the empirical variogram value and the fitted variogram model:

$$\sum_{i=1}^{L} \left(\hat{\gamma}(h)_i - \gamma^*(h)_i \right)^2 \tag{2.16}$$

where L is the number of discrete distance lags over which the variogram will be calculated, $\hat{\gamma}(h)_i$ is a value from (2.5) at distance lag "i" and $\hat{\gamma}^*(h)_i$ is the

variogram value from the fitted variogram model at distance lag "i".

Cressie (1989) recommends the following function:

$$N_k \sum \left(\frac{(\hat{\gamma}(h) - \gamma^*(h))^2}{\gamma^*(h)} \right) \tag{2.17}$$

where N_k is the number of pairs used for each variogram distance h.

Clark and Harper (2000) support Cressie's function, but as a weighted average where each component is divided by the total number of pairs N used in computing the entire empirical variogram:

$$N_k \Big/ N \left[\sum \left((\hat{\gamma}(h) - \gamma^*(h))^2 \Big/ \gamma^*(h) \right) \right] \tag{2.18}$$

2.1.8 Ordinary Kriging

Kriging (Matheron, 1963) is a linear interpolation method that allows prediction of the random variable $z_k : k = n + 1$ and its estimation error at an un-sampled new spatial location $(x_k, y_k) : k = n + 1$. The linear estimate $\hat{z}_k$ of z_k is determined by weighting the existing data $z_j : j = 1$ to n at the observed known spatial locations $(x_j, y_j) : j = 1$ to n, respectively. For example, to determine the estimated value $\hat{z}_k$ at the new known location (x_k, y_k), the linear kriging estimate of z_k would be:

$$\hat{z}_k = \sum_{j=1}^{n} w_{kj} z_j \tag{2.19}$$

where z_j are the observed random variables and w_{kj} are the kriging weights. The w_{kj} weights are the solution to the kriging equations which will be shown later in

this section. As stated by Isaaks and Srivastava (1989), kriging can be considered a B.L.U.E. (best linear unbiased estimator) by meeting the following three requirements.

The first requirement is that kriging must be a linear estimator. Since the estimate $\hat{z}_k = \sum w_{kj} z_j$ is a weighted linear combination of the available data, kriging is a linear estimator.

The second requirement is that kriging must produce unbiased estimates. To be an unbiased estimator, the expected value of the estimation error $E[z_k - \hat{z}_k]$ must equal 0. As shown by Isaaks and Srivastava (1989):

$$\sum_{j=1}^{n} w_{kj} = 1 \tag{2.20}$$

is a necessary condition in order for kriging to be considered an unbiased estimator of the random variable z_k.

The third requirement is that kriging must minimize Var $(z_k - \hat{z}_k)$, the variance of the estimation errors. In satisfying this requirement and to estimate a single point z_k, Isaaks and Srivstava (1989) showed the ordinary kriging equations in matrix form as follows:

$$\begin{matrix} & C_O & & & \times & w_O & = & D_O \end{matrix}$$

$$\begin{bmatrix} \mathrm{Cov}(z_1, z_1) & \dots & \mathrm{Cov}(z_1, z_n) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \mathrm{Cov}(z_n, z_1) & \dots & \mathrm{Cov}(z_n, z_n) & 1 \\ 1 & \dots & 1 & 0 \end{bmatrix} \times \begin{bmatrix} w_{k1} \\ \vdots \\ w_{kn} \\ \lambda_0 \end{bmatrix} = \begin{bmatrix} \mathrm{Cov}(z_k, z_1) \\ \vdots \\ \mathrm{Cov}(z_k, z_n) \\ 1 \end{bmatrix} \tag{2.21}$$

where:

C_O is an (n+1) x (n+1) matrix in which $\mathrm{Cov}(z_i, z_j)$ represents the covariance between the existing z_i, z_j observations as defined by (2.8).

w_O is a (n+1) x 1 matrix in which $(w_{k1}, ...w_{kn})$ represent the kriging weights and λ_0 represents a LaGrange parameter, which follows from the linear condition (2.20).

D_O is a (n+1) x 1 matrix in which $\mathrm{Cov}(z_k, z_j)$ represents the covariance between the point to be estimated (z_k) and the n existing observations $(z_j : j = 1 \text{ to } n)$, as defined by (2.8).

As indicated by the matrix subscript "O" , this type of kriging is known as ordinary kriging. Ordinary kriging is performed when the random process does not require any trend removal as described in Section 2.1.3. The kriging weights may be determined by solving for matrix w_O such that:

$$w_O = C_O^{-1} \times D_O \tag{2.22}$$

allows for both the estimate of $\hat{z}_k$ and the variance of the estimation error $\hat{\sigma}_k^2$ to be determined. In developing these kriging equations, Isaaks and Srivstata (1989) also showed that the variance of the estimation error is:

$$\hat{\sigma}_k^2 = \mathrm{Var}(z_k - \hat{z}_k) = \mathrm{Var}(z_k) + \mathrm{Var}(\hat{z}_k) - 2\ \mathrm{Cov}(z_k, \hat{z}_k) \tag{2.23}$$

The first term to the right of the equal sign in (2.23) $\mathrm{Var}(z_k)$ is the variance of the random process (σ^2), which is estimated by the nugget and sill of the variogram as defined in (2.12). The second term is the variance of $\hat{z}_k$. Since $\hat{z}_k = \sum_j w_{kj} z_j$ is a linear combination of random variables, the variance can be expressed as $\sum_i \sum_j w_{ki} w_{kj}\ \mathrm{Cov}(z_i, z_j)$ where Cov (z_i, z_j) is defined by (2.8) using the distance h_{ij} between the observed data z_i and z_j. The third term is the covariance between the point to be estimated z_k and its estimate $\hat{z}_k = \sum_j w_{kj} z_j$ and can be expressed as $\sum_j w_{kj} \mathrm{Cov}(z_k, z_j)$ where $\mathrm{Cov}(z_k, z_j)$ is defined by (2.8) using the distance h_{kj} between z_k, the location to be estimated, and z_j, the observed data. By summing these terms, the kriging variance of the estimation error $\hat{\sigma}_k^2$ may be expressed as:

$$\hat{\sigma}_k^2 = \mathrm{Var}(z_k - \hat{z}_k) = \sigma^2 + \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ki} w_{kj} \mathrm{Cov}(z_i, z_j) - 2 \sum_{j=1}^{n} w_{kj} \mathrm{Cov}(z_k, z_j) \quad (2.24)$$

2.1.9 Universal Kriging

Recall in Section 2.1.3 in which an increasing variogram function may be an indication of a non-stationary random process. An increasing variogram function may be modified by the elimination of any spatial trend present in the random process. Spatial trends can be modelled by a variety of regression models. For many spatial trends, the linear, quadratic, cubic, or other higher order polynomials may be used with the x and y location parameters as independent variables.

This type of kriging, in which the residuals from a trend model are used to estimate the variogram function, is called universal kriging. Universal kriging

without the trend component will reduce to ordinary kriging, which was described in Section 2.1.8. When performing universal kriging, the estimate of the trend parameters must also be unbiased. This inclusion of unbiased trend parameters in the universal kriging equations will now be discussed. Here, the focus is on two of the possible spatial trend models: linear and quadratic. The rationale for discussing the linear and quadratic spatial models is that it is desirable to use the simplest models, if possible. Details for the cubic and higher order models can be easily deduced from those presented here for the linear and quadratic models.

Linear Spatial Model

Let the random process $Z(s)$ have a spatial trend which can be fit by a first order linear regression model. Let the trend model be $z_j = \beta_0 + \beta_1 x_j + \beta_2 y_j$, where z_j is the observed random variable of interest at spatial location (x_j, y_j). Similarly, the estimate at an un-sampled location $\hat{z}_k : k = n+1$ at spatial location (x_k, y_k) may also be considered as $\hat{z}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k + \hat{\beta}_2 y_k + \hat{\epsilon}_k$. As shown by Clark and Harper (2000), the spatial estimates may be expressed as:

$$E[\hat{z}_k] = E\Big[\sum_{j=1}^{n} w_{kj} z_j\Big]$$

$$E[\hat{\beta}_0 + \hat{\beta}_1 x_k + \hat{\beta}_2 y_k + \hat{\epsilon}_k] = E\Big[\sum_{j=1}^{n} w_{kj}(\beta_0 + \beta_1 x_j + \beta_2 y_j)\Big] \qquad (2.25)$$

By combining like terms, it can be seen that in order to provide an unbiased estimator of z_k:

$$E[\hat{\beta}_0] = E\Big[\sum_{j=1}^{n} w_{kj}\beta_0\Big]$$
$$\beta_0 = \sum_{j=1}^{n} w_{kj}\beta_0$$
$$\sum_{j=1}^{n} w_{kj} = 1. \tag{2.26}$$

Similarly:

$$\sum_{j=1}^{n} w_{kj}x_j = x_k \tag{2.27}$$
$$\sum_{j=1}^{n} w_{kj}y_j = y_k. \tag{2.28}$$

To solve for the kriging weights with a linear spatial trend, the ordinary kriging matrices in (2.21) may be appended as follows:

$$\underset{C_{UL}}{\begin{bmatrix} \mathrm{Cov}(z_1,z_1) & \dots & \mathrm{Cov}(z_1,z_n) & 1 & x_1 & y_1 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \mathrm{Cov}(z_n,z_1) & \dots & \mathrm{Cov}(z_n,z_n) & 1 & x_n & y_n \\ 1 & \dots & 1 & 0 & 0 & 0 \\ x_1 & \dots & x_n & \vdots & \vdots & \vdots \\ y_1 & \dots & y_n & 0 & 0 & 0 \end{bmatrix}} \times \underset{w_{UL}}{\begin{bmatrix} w_{k1} \\ \vdots \\ w_{kn} \\ \lambda_0 \\ \lambda_1 \\ \lambda_2 \end{bmatrix}} = \underset{D_{UL}}{\begin{bmatrix} \mathrm{Cov}(z_k,z_1) \\ \vdots \\ \mathrm{Cov}(z_k,z_n) \\ 1 \\ x_k \\ y_k \end{bmatrix}} \tag{2.29}$$

where:

C_{UL} is a (n+3) x (n+3) matrix in which $\text{Cov}(z_i, z_j)$ represent the covariance between existing z_i, z_j observations as defined by (2.8). In addition, $\Big((x_1, y_1), ...(x_n, y_n)\Big)$ represent the spatial locations for known observations $(z_1, ...z_n)$.

w_{UL} is a (n+3) x 1 matrix in which $(w_{k1}, ...w_{kn})$ represent the kriging weights and λ_0, λ_1, and λ_2 represent LaGrange parameters associated with constraints (2.26), (2.27), and (2.28).

D_{UL} is a (n+3) x 1 matrix in which $\text{Cov}(z_k, z_j) : k = n+1, j = 1$ to n represent the covariance between the point to be estimated (z_k) and the existing observations $(z_j : j = 1 \text{ to } n)$, as defined by (2.8). In addition, $(x_k, y_k) \in z_k$ represent the spatial location of the point to be estimated for $k = n + 1$.

As indicated by the matrix subscript "UL" , this type of kriging is known as universal kriging with a linear trend. The universal kriging weights may be determined by solving for matrix w_{UL} such that:

$$w_{UL} = C_{UL}^{-1} \times D_{UL} \tag{2.30}$$

Quadratic Spatial Model

Assume now that a quadratic spatial trend model is fit. Let the trend model for the observed data (z_j) be: $z_j = \beta_0 + \beta_1 x_j + \beta_2 y_j + \beta_3 x_j^2 + \beta_4 y_j^2 + \beta_5 x_j y_j$. In addition, the estimate at an un-sampled location $\hat{z}_k : k = n + 1$ at spatial location (x_k, y_k) may also be considered as $\hat{z}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k + \hat{\beta}_2 y_k + \hat{\beta}_3 x_k^2 + \hat{\beta}_4 y_k^2 + \hat{\beta}_5 x_k y_k + \hat{\epsilon}_k$. Taking expectations as was done in (2.25), the three conditions shown to maintain unbiased linear trend parameter estimates (2.26), (2.27), and (2.28) may be duplicated and included here. In addition, three conditions resulting from modeling a quadratic trend are as follows:

$$\sum_{j=1}^{n} w_{kj} x_j^2 = x_k^2 \tag{2.31}$$

$$\sum_{j=1}^{n} w_{kj} y_j^2 = y_k^2 \tag{2.32}$$

$$\sum_{j=1}^{n} w_{kj} x_j y_j = x_k y_k \tag{2.33}$$

For modeling a quadratic spatial trend, all six of these conditions must be incorporated into the universal kriging matrices. Using matrix (2.29) as a guide, the universal kriging matrices with a quadratic trend can be represented as follows:

$$
\begin{array}{ccccc}
C_{UQ} & & & \times \; w_{UQ} & = \; D_{UQ}
\end{array}
$$

$$
\begin{bmatrix}
\mathrm{Cov}(z_1,z_1) & \dots & \mathrm{Cov}(z_1,z_n) & 1 & x_1 & y_1 & x_1^2 & y_1^2 & x_1y_1 \\
\vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\mathrm{Cov}(z_n,z_1) & \dots & \mathrm{Cov}(z_n,z_n) & 1 & x_n & y_n & x_n^2 & y_n^2 & x_ny_n \\
1 & \dots & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
x_1 & \dots & x_n & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
y_1 & \dots & y_n & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
x_1^2 & \dots & x_n^2 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
y_1^2 & \dots & y_n^2 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
x_1y_1 & \dots & x_ny_n & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
\times
\begin{bmatrix} w_{k1} \\ \vdots \\ w_{kn} \\ \lambda_0 \\ \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \lambda_5 \end{bmatrix}
=
\begin{bmatrix} \mathrm{Cov}(z_k,z_1) \\ \vdots \\ \mathrm{Cov}(z_k,z_n) \\ 1 \\ x_k \\ y_k \\ x_k^2 \\ y_k^2 \\ x_ky_k \end{bmatrix}
\tag{2.34}
$$

where:

C_{UQ} is a (n+6) x (n+6) matrix in which $\mathrm{Cov}(z_i, z_j)$ represent the covariance between existing z_i, z_j observations as defined by (2.8). In addition, $\Big((x_1, y_1), ...(x_n, y_n)\Big)$ represent the spatial locations for known observations $(z_1, ...z_n)$.

w_{UQ} is a (n+6) x 1 matrix in which $(w_{k1}, ...w_{kn})$ represent the kriging weights and $\lambda_0, \lambda_1,$ and λ_2 represent LaGrange parameters associated with (2.26), (2.27), and (2.28) respectively. In addition, λ_3, λ_4, and λ_5 represent LaGrange parameters associated with constraints (2.31), (2.32), and (2.33), respectively.

D_{UQ} is a (n+6) x 1 matrix in which $\mathrm{Cov}(z_k, z_j) : k = n+1$ represent the covariance between the point to be estimated (z_k) and the existing observations $(z_j : j =$

1 to n) as defined by (2.8). In addition, (x_k, y_k) represents the spatial location of the point to be estimated.

As indicated by the matrix subscript "UQ" , this type of kriging is known as universal kriging with a quadratic trend. The universal kriging weights may be determined by solving for matrix w_{UQ} such that:

$$w_{UQ} = C_{UQ}^{-1} \times D_{UQ} \tag{2.35}$$

2.1.10 Block Kriging

Up until this section, the topic of kriging has been concerned with estimation of the random variable z_k at a single new un-sampled point $(x_k, y_k) : k = n+1$. In this dissertation, it is desirable to obtain an estimate of the mean of z for a prescribed local area within the GAI. If a square block is chosen as this local area, then block kriging may be used to obtain this mean estimate $\hat{z}_B$. Block kriging determines an estimate of mean that represents the entire square block. In this dissertation, such a block B is a subset of the GAI and has a kriged linear estimate expressed as:

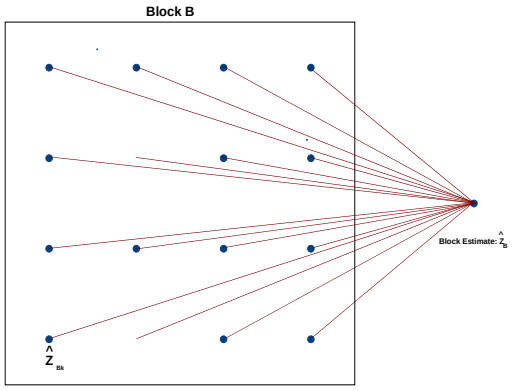
$$\hat{z}_B = \sum_{j=1}^{n} w_{Bj} z_j \tag{2.36}$$

where w_{Bj} represents the Block B kriging weight "j" for the observed random variables $z_j : j = 1$ to n.

The estimate of block mean $\hat{z}_B$ may also be considered as the average

value of an infinite number points contained within the block. One method for obtaining $\hat{z}_B$ is to discretize the square block B into a finite number of points $\hat{z}_{Bk} : (k = 1 \text{ to } m)$ and then take the average of these point estimates $\hat{z}_{Bk}$. Though conceptually simple, this procedure may be computationally inefficient depending on the number of points chosen to represent the block B. Isaaks and Srivasatava (1989) suggest that a 4x4 grid of symmetrically placed points (m=16) is sufficiently accurate to estimate $\hat{z}_B$. Figure 5 depicts this type of block kriging from a graphical perspective.

Figure 5. Block Kriging Estimate with m=16 points



To investigate the statistical properties of the block kriging estimates, Isaaks and Srivastava (1989) demonstrate that the true block mean z_B, and its corresponding estimate $\hat{z}_B$ may indeed be determined by taking the average of the points that are contained within the block. Hence z_B may be expressed as:

$$z_B = \frac{1}{m} \sum_{k=1}^{m} z_{Bk} \qquad z_{Bk} \in B \tag{2.37}$$

where z_B is the block mean, m is the number of points selected within the block,

and z_{Bk} is a random variable at the k^{th} point contained within block B, for $k = 1, ...m$.

Similarly, $\hat{z}_B$ may be represented as:

$$\hat{z}_B = \frac{1}{m} \sum_{k=1}^{m} \hat{z}_{Bk} \qquad \hat{z}_{Bk} \in B \tag{2.38}$$

where $\hat{z}_B$ is an estimate block mean, m is the number of points selected within the block, and $\hat{z}_{Bk}$ is an estimate of the random variable at the k^{th} point z_{Bk} contained within block B, for $k = 1, ...m$. Because of the stationarity conditions discussed in Section 2.1.1,

$$E[\hat{z}_B] = E[z_B] = \mu \tag{2.39}$$

and hence is an unbiased estimator.

Recall the three matrix systems (2.21), (2.29), and (2.34) given for ordinary and the two cases of universal kriging. Looking at these matrix systems, only the D-type matrices (D_O, D_{UL}, D_{UQ}) contain any information about un-sampled points z_k, which may now be considered as components of block B given in (2.37) and (2.38).

To perform block kriging, it is necessary to modify the D-type matrices given in (2.21), (2.29), and (2.34) from a point estimation formulation to a block estimation formulation. The point covariance expressions $\text{Cov}(z_k, z_j) : j = 1$ to n in these D-type matrices will be replaced by $\text{Cov}(z_B, z_j) : j = 1$ to n for use in the block kriging equations, where for convenience, z_B is defined by (2.37). A further

examination of $\mathrm{Cov}(z_B, z_j)$ shows that:

$$\begin{aligned}
\mathrm{Cov}(z_B, z_j) &= E[z_B z_j] - E[z_B]E[z_j] \\
&= E\left[\frac{1}{m}\sum_{k=1}^{m} z_{Bk} z_j\right] - E\left[\frac{1}{m}\sum_{k=1}^{m} z_{Bk}\right] E[z_j] \\
&= \frac{1}{m}\sum_{k=1}^{m}\Big(E[z_{Bk} z_j] - E[z_{Bk}]E[z_j]\Big) \\
&= \frac{1}{m}\sum_{k=1}^{m} \mathrm{Cov}(z_{Bk}, z_j) \\
&= \overline{\mathrm{Cov}}(z_{Bk}, z_j). \qquad (2.40)
\end{aligned}$$

Equation 2.40 is the average of the covariances between the $z_{Bk} \in B$ and z_j, which are the observed random variables.

To estimate z_B, the matrix system given for ordinary kriging (2.21), may now be modified as follows:

$$\begin{array}{ccccc}
 & C_B & & \times\ w_B\ = & D_B \\
\begin{bmatrix}
\mathrm{Cov}(z_1, z_1) & \dots & \mathrm{Cov}(z_1, z_n) & 1 \\
\vdots & \ddots & \vdots & \vdots \\
\mathrm{Cov}(z_n, z_1) & \dots & \mathrm{Cov}(z_n, z_n) & 1 \\
1 & \dots & 1 & 0
\end{bmatrix} & \times & \begin{bmatrix} w_{B1} \\ \vdots \\ w_{Bn} \\ \lambda_0 \end{bmatrix} & = & \begin{bmatrix} \mathrm{Cov}(z_B, z_1) \\ \vdots \\ \mathrm{Cov}(z_B, z_n) \\ 1 \end{bmatrix}
\end{array} \qquad (2.41)$$

where:

C_B is an (n+1) x (n+1) matrix in which $\mathrm{Cov}(z_i, z_j)$ represents the covariance

between the existing z_i, z_j observations as defined by (2.8).

w_B is a (n+1) x 1 matrix in which $(w_{B1}, ...w_{Bn})$ represent the block kriging weights and λ_0 represents a LaGrange parameter, which follows from the linear condition (2.20).

D_B is a (n+1) x 1 matrix in which $\mathrm{Cov}(z_B, z_j)$ represents the covariance between the block to be estimated (z_B) and the n existing observations $(z_j : j = 1 \text{ to } n)$, as defined by (2.40).

As indicated by the matrix subscript "B" , this type of kriging is known as block kriging. The block kriging weights may be obtained by solving for matrix w_B such that:

$$w_B = C_B^{-1} \times D_B \tag{2.42}$$

If a spatial trend is present in the data, universal block kriging may also be performed. Similar to universal kriging for points (Section 2.1.9), unbiased estimates of the trend parameters must also be converted from a point kriging formulation to a block kriging formulation. This inclusion of unbiased trend parameters in the universal kriging equations will now be examined. As previously discussed, the focus is on two of the possible spatial trend models: linear and quadratic. The rationale for discussing the linear and quadratic spatial models is again that it is desirable to use the simplest models, if possible. Details for the cubic and higher order models can be easily deduced from those presented here

for the linear and quadratic models.

Block Kriging With Linear Trend

Let the random process $Z(s)$ have a spatial trend which can be fit by a linear regression model. Let the trend model be $z_j = \beta_0 + \beta_1 x_j + \beta_2 y_j$, where z_j is the observed random variable of interest at spatial location (x_j, y_j). Similarly, the estimate at an un-sampled block location $z_{Bk} : k = 1, ..., m$ at spatial location (x_{Bk}, y_{Bk}) may also be considered as:

$$\hat{z}_{Bk} = \hat{\beta}_0 + \hat{\beta}_1 x_{Bk} + \hat{\beta}_2 y_{Bk} \tag{2.43}$$

Using (2.36), (2.38), and (2.43), an unbiased spatial estimate for z_B using block kriging with a linear trend may be expressed as:

$$\begin{aligned} \hat{z}_B &= \frac{1}{m}\sum_{k=1}^{m} \hat{z}_{Bk} \\ \Big[\sum_{j=1}^{n} w_{Bj} z_j\Big] &= \Big[\frac{1}{m}\sum_{k=1}^{m}(\hat{\beta}_0 + \hat{\beta}_1 x_{Bk} + \hat{\beta}_2 y_{Bk})\Big] \\ \Big[\sum_{j=1}^{n} w_{Bj}(\beta_0 + \beta_1 x_j + \beta_2 y_j + \epsilon_j)\Big] &= \Big[\frac{1}{m}\sum_{k=1}^{m}(\hat{\beta}_0 + \hat{\beta}_1 x_{Bk} + \hat{\beta}_2 y_{Bk})\Big] \end{aligned}$$

By combining like terms and taking expectations, it can be seen that:

$$\begin{aligned} E\Big[\sum_{j=1}^{n} w_{Bj}\beta_0\Big] &= E\Big[\frac{1}{m}\sum_{k=1}^{m}\hat{\beta}_0\Big] \\ \beta_0 \sum_{j=1}^{n} w_{Bj} &= \frac{1}{m}\sum_{k=1}^{m}\beta_0 \\ \sum_{j=1}^{n} w_{Bj} &= 1 \end{aligned} \tag{2.44}$$

Similarly, it can be shown that:

$$\sum_{j=1}^{n} w_{Bj}x_j = \frac{1}{m}\sum_{k=1}^{m} x_{Bk} = \overline{x}_B \tag{2.45}$$

where $\overline{x}_B$ represents the average of the "x" spatial locations for points $x_{Bk} \in B$. In addition,

$$\sum_{j=1}^{n} w_{Bj}y_j = \overline{y}_B \tag{2.46}$$

where $\overline{y}_B$ represents the average of the "y" spatial locations for points $y_{Bk} \in B$.

To estimate z_B using universal block kriging with a linear spatial trend, the universal kriging matrices in (2.29) may be appended as follows:

$$
\begin{array}{ccc} C_{BL} & \times \; w_{BL} \; = & D_{BL} \end{array}
$$

$$
\begin{bmatrix}
\mathrm{Cov}(z_1,z_1) & \dots & \mathrm{Cov}(z_1,z_n) & 1 & x_1 & y_1 \\
\vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\
\mathrm{Cov}(z_n,z_1) & \dots & \mathrm{Cov}(z_n,z_n) & 1 & x_n & y_n \\
1 & \dots & 1 & 0 & 0 & 0 \\
x_1 & \dots & x_n & \vdots & \vdots & \vdots \\
y_1 & \dots & y_n & 0 & 0 & 0
\end{bmatrix}
\times
\begin{bmatrix} w_{B1} \\ \vdots \\ w_{Bn} \\ \lambda_0 \\ \lambda_1 \\ \lambda_2 \end{bmatrix}
=
\begin{bmatrix} \mathrm{Cov}(z_B,z_1) \\ \vdots \\ \mathrm{Cov}(z_B,z_n) \\ 1 \\ \overline{x}_B \\ \overline{y}_B \end{bmatrix}
\tag{2.47}
$$

where:

C_{BL} is a (n+3) x (n+3) matrix in which $\mathrm{Cov}(z_i, z_j)$ represent the covariance between existing i, j observations as defined by (2.8). In addition, $\Big((x_1, y_1), ...(x_n, y_n)\Big)$ represent the spatial locations for known observations $(z_1, ...z_n)$.

w_{BL} is a (n+3) x 1 matrix in which $(w_{B1}, ..., w_{Bn})$ represent the kriging weights and λ_0, λ_1, and λ_2 represent LaGrange parameters associated with constraints (2.44), (2.45), and (2.46).

D_{BL} is a (n+3) x 1 matrix in which $\mathrm{Cov}(z_B, z_j)$ is defined by (2.40) and represents the average of $z_{Bk} \in B$ with the observed data z_j (j=1 to n) and $\overline{x}_B$ represents the average x value of the points contained within block B etc.

As indicated by the matrix subscript "BL", this kriging is block kriging with a linear trend. The block kriging weights may be obtained by solving for matrix w_{BL} such that:

$$
w_{BL} = C_{BL}^{-1} \times D_{BL} \tag{2.48}
$$

Block Kriging With Quadratic Trend

Let the random process $Z(s)$ have a spatial trend which can be fit by a quadratic regression model. Let the trend model be $z_j = \beta_0 + \beta_1 x_j + \beta_2 y_j + \beta_3 x_k^2 + \beta_4 y_k^2 + \beta_5 x_k y_k$, where z_j is the observed random variable of interest at spatial location (x_j, y_j). Similarly, the estimate at an un-sampled block location $z_{Bk} : k = 1, ..., m$ at spatial location (x_{Bk}, y_{Bk}) may also be considered as:

$$\hat{z}_{Bk} = \hat{\beta}_0 + \hat{\beta}_1 x_{Bk} + \hat{\beta}_2 y_{Bk} + \hat{\beta}_3 x_{Bk}^2 + \hat{\beta}_4 y_{Bk}^2 + \hat{\beta}_5 x_{Bk} y_{Bk} \tag{2.49}$$

Using (2.36), (2.38), and (2.49), unbiased spatial estimates for block kriging with a quadratic trend may be expressed as:

$$\hat{z}_B = \frac{1}{m} \sum_{k=1}^{m} \hat{z}_{Bk}$$

$$\Big[\sum_{j=1}^{n} w_{Bj} z_j \Big] = \frac{1}{m} \sum_{k=1}^{m} \hat{z}_{Bk}$$

$$\Big[\sum_{j=1}^{n} w_{Bj} (\beta_0 + \beta_1 x_j + \beta_2 y_j + \beta_3 x_j^2 + \beta_4 y_j^2 + \beta_5 x_j y_j + \epsilon_j) \Big] = \Big[\frac{1}{m} \sum_{k=1}^{m} (\hat{\beta}_0 + \hat{\beta}_1 x_{Bk} + \hat{\beta}_2 y_{Bk} + \hat{\beta}_3 x_{Bk}^2 + \hat{\beta}_4 y_{Bk}^2 + \hat{\beta}_5 x_{Bk} y_{Bk}) \Big]$$

By combining like terms and taking expectations, it can be seen that in addition to (2.44), (2.45), and (2.46) that:

$$\sum_{j=1}^{n} w_{Bj} x_j^2 = \frac{1}{m} \sum_{k=1}^{m} x_{Bk}^2 = \overline{x}_B^2 \tag{2.50}$$

where $\overline{x}_B^2$ represents the average of the "x^2" spatial locations for points $x_{Bk} \in B$. In addition,

$$\sum_{j=1}^{n} w_{Bj} y_j^2 = \overline{y}_B^2 \tag{2.51}$$

where $\overline{y}_B^2$ represents the average of the "y^2" spatial locations for points $y_{Bk} \in B$, and

$$\sum_{j=1}^{n} w_{Bj} x_j y_j = \overline{x}_B \overline{y}_B \tag{2.52}$$

where $\overline{x}_B \overline{y}_B$ represents the average of the "xy" interaction spatial locations for points $x_{Bk} \in B$ and $y_{Bk} \in B$.

To estimate z_B using universal block kriging with a quadratic spatial trend, the universal kriging matrices in (2.34) may be appended as follows:

$$
\begin{array}{ccc} C_{BQ} & \times & w_{BQ} = D_{BQ} \end{array}
$$

$$
\begin{bmatrix}
\mathrm{Cov}(z_1,z_1) & \dots & \mathrm{Cov}(z_1,z_n) & 1 & x_1 & y_1 & x_1^2 & y_1^2 & x_1y_1 \\
\vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\mathrm{Cov}(z_n,z_1) & \dots & \mathrm{Cov}(z_n,z_n) & 1 & x_n & y_n & x_n^2 & y_n^2 & x_ny_n \\
1 & \dots & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
x_1 & \dots & x_n & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
y_1 & \dots & y_n & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
x_1^2 & \dots & x_n^2 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
y_1^2 & \dots & y_n^2 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
x_1y_1 & \dots & x_ny_n & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
\times
\begin{bmatrix} w_{B1} \\ \vdots \\ w_{Bn} \\ \lambda_0 \\ \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \lambda_5 \end{bmatrix}
=
\begin{bmatrix} \mathrm{Cov}(z_B,z_1) \\ \vdots \\ \mathrm{Cov}(z_B,z_n) \\ 1 \\ \overline{x}_B \\ \overline{y}_B \\ \overline{x}_B^2 \\ \overline{y}_B^2 \\ \overline{x}_B\overline{y}_B \end{bmatrix}
\tag{2.53}
$$

where:

C_{BQ} is a (n+6) x (n+6) matrix in which $\mathrm{Cov}(z_i, z_j)$ represent the covariance between existing i, j observations as defined by (2.8). In addition, $\Big((x_1, y_1), ...(x_n, y_n)\Big)$ represent the spatial locations for known observations $(z_1, ...z_n)$.

w_{BQ} is a (n+6) x 1 matrix in which $(w_{B1}, ..., w_{Bn})$ represent the kriging weights and λ_0, λ_1, λ_2, λ_3, λ_4, and λ_5 represent LaGrange parameters associated with constraints (2.44), (2.45), (2.46), (2.50), (2.51), and (2.52), respectively.

D_{BQ} is a (n+6) x 1 matrix in which $\mathrm{Cov}(z_B, z_j)$ is defined by (2.40) and represents the average of $z_{Bk} \in B$ with the observed data z_j (j=1 to n) and $\overline{x}_B$ represents the average x value of the points contained within block B etc.

As indicated by the matrix subscript "BQ", this type of kriging is block kriging with a quadratic trend. The block kriging weights may be obtained by solving for matrix w_{BQ} such that:

$$w_{BQ} = C_{BQ}^{-1} \times D_{BQ} \tag{2.54}$$

The variance of the block kriging estimation error was discussed by Isaaks and Srivstava (1989) and were shown to be:

$$\begin{aligned}\hat{\sigma}_B^2 &= E[z_B - \hat{z_B}]^2 \\ &= \frac{1}{m^2}\sum_{k=1}^{m}\sum_{l=1}^{m}\mathrm{Cov}(z_{Bk}, z_{Bl}) - \frac{2}{m}\sum_{k=1}^{m}\sum_{j=1}^{n} w_{Bj}\ \mathrm{Cov}(z_{Bk}, z_j) + \sum_{i=1}^{n}\sum_{j=1}^{n} w_{Bi}w_{Bj}\mathrm{Cov}(z_i, z_j)\end{aligned} \tag{2.55}$$

The first term, containing $\mathrm{Cov}(z_{Bk}, z_{Bl})$ terms, represents the average of all covariances between the m points contained within block B as defined by the variogram function. The second term, containing $\mathrm{Cov}(z_{Bk}, z_j)$ terms, represents the weighted average of the covariance between each point within block B ($z_{Bk} \in B$), and the observed random variable of interest (z_j) as defined by the variogram function. The third term, containing $\mathrm{Cov}(z_i, z_j)$ terms, represents the weighted covariances between all observed random variables as defined by the variogram function .

2.1.11 Cross-Validation

Cross-validation (Stone, 1974) is a technique used to evaluate the accuracy of a spatial model. Cross-validation cannot prove that the selected spatial model

is correct, but it can provide indication of model mispecification. Cross-validation may also be used to compare competing spatial models (Isaaks and Srivastava, 1989). The process of cross-validation is as follows: starting with the sample data set, one of the observed random variables z_j is removed from the data set. Then, using the selected spatial model, the removed variable is predicted at this vacated sample location as $\hat{z}_{-j}$. The difference between the observed value and the estimate is defined as the estimation error: $e_{-j} = (z_j - \hat{z}_{-j})$. In addition, let the standardized estimation error be defined as

$$E_{-j} = e_{-j}/\hat{\sigma}_{-j} \tag{2.56}$$

in which $\hat{\sigma}^2_{-j}$ is the kriging variance of the estimation error for $\hat{z}_{-j}$ and may be determined using equation 2.24 and considering z_j as z_k, $\hat{\sigma}^2_{-j}$ as $\hat{\sigma}^2_k$, and $\hat{z}_{-j}$ as $\hat{z}_k$ respectively.

Although these estimation errors are not independent, Chiles and Delfiner (1999) recommend the estimation errors be inspected for indications of bias, outliers, and model mispecification using (1) a scatterplot of e_{-j} versus spatial location x, (2) a scatterplot of e_{-j} versus spatial location y, (3) a histogram of the standardized errors E_{-j}, (4)a q-q plot of E_{-j}, (5) a scatterplot of $\hat{e}_{-j}$ versus $\hat{z}_{-j}$, and (6) a scatterplot of $\hat{z}_{-j}$ versus z_j.

Additional research has recommended other applications using these diagnostics. Cressie (1993) states (2.56) should have a mean of 0 with the root-mean-square approximately 1. Meyers (1997) recommends that the e_{-j} should follow a

symmetric distribution.

2.2 Infill Sampling

Infill sampling is defined as adding additional observations to the existing data set in an attempt to improve the accuracy of the spatial model. The addition of n observations to the data set will be made from a pool of N predetermined infill candidates. It should be noted that the infill research performed thus far has been in reference to a point kriging formulation (Gao, Wang, Zhao (1996)). The infill sampling strategy explored in this dissertation will be in reference to a block kriging formulation.

To establish an infill sampling strategy, it is important to recognize how the inclusion of these additional observations into the data set influences the kriging estimates of block mean $\hat{z}_B$ and variance of the block estimation errors $\hat{\sigma}_B^2$. As stated by Kacewicz (1991), the kriging estimates of mean will be updated and the variances of the block estimation errors will decrease when additional samples are incorporated into the kriging system. As discussed in Aspie and Barnes (1990), updated estimates of the $\hat{\sigma}_B^2$ can be calculated without knowing the actual measurement of the random variable of interest. This is because the $\hat{\sigma}_B^2$ depend only on the location of the additional observations and the variogram, which are known, and not on any measurements of the observed random variables. Hence, it is unnecessary to predict the random variable of interest at any location

within the GAI. Given a desired infill sample size (n), different combinations of these n infill candidates will be added to the data and the standard error of the block estimation error $\sqrt{\hat{\sigma}_B^2}$ will evaluated over the GAI. The $\hat{z}_B$ for all blocks will remain unchanged because the variogram used to predict the block means has not been modified. However, the standard error of the block estimate $\sqrt{\hat{\sigma}_B^2}$ will decrease in varying degrees based on the chosen combination of infill candidates.

The selection of an optimal infill sample can be a time consuming and computationally exhausting process. For example, if there is a pool of $N = 100$ infill candidates and n= 10 infill samples are desired, there are 1.73×10^{13} (100 choose 10) possible infill sample combinations. Hence, an exhaustive search is not practical and a shorter search algorithm is required.

As an alternative to exhaustive sampling, the *Greedy/Sequential Exchange Algorithm* (Aspie and Barnes, 1990) will be used in this dissertation. This algorithm is designed to find a reasonable, if not optimal, solution when time and computer memory do not allow for an exhaustive search. Consider a situation where n additional samples are needed and there are N available candidate locations. Let $n < N$. The *greedy* algorithm searches for n locations one at a time. The algorithm first searches for the one sample location that best meets the decision criterion. Then, the location that makes the best pair with the first one is chosen. Next, the sample location that makes the best trio with the first two is chosen. The greedy algorithm continues in this manner until all n sample

locations have been selected. While this algorithm does not guarantee an optimal solution from the set of N candidate locations, it does provide a good starting point.

The *sequential exchange* algorithm searches through $k*n*N$ combinations where k is a small integer. This algorithm starts with the results of the *greedy* algorithm. The sequential algorithm proceeds by sequentially optimizing each of the n candidates while the other $n-1$ are fixed at their current locations. In other words, all $(N+1)-n$ remaining candidate locations are considered as the first sample while the 2^{nd} through n^{th} samples are held constant. The first sample is then set to the location that is found to be optimal in combination with the current settings for the 2^{nd} through n^{th} samples. Next, the second sample runs through the $(N+1)-n$ remaining possible locations while the 1^{st} and $3rd$ through n samples are held constant. The second sample is then set to its new optimal location. This procedure is performed for each of the n samples. After the sequential algorithm has checked each of the N samples for an improvement, the new set of optimal locations now becomes the current solution and the process begins again. In this way, the solution iteratively converges to an answer. After k iterations, the solution will not be improved upon. The last solution is then the final answer. The last solution is not necessarily optimal, but it is often quite close, and is much more efficient than an exhaustive search. If the results of the *greedy* algorithm are reasonable, k should be less than 5.

2.3 Dissimilarity Coefficient

Regardless of the underlying distribution of a data set, the estimates of two or more block means resulting from the block kriging procedure may not be considered as independent statistics.

The assumed spatial correlation that exists within the random variables z(s) violates the statistical independence assumption that is an important requirement for many statistical testing procedures. Hence, before the number and spatial boundaries of the strata can be determined, it is necessary to develop a measurement that can quantify the statistical distance and account for the spatial correlation between the blocks.

Dissimilarities are defined as nonnegative numbers $d(i, j)$ that are small (close to 0) when objects i and j are "near" to each other and become large when objects i and j are very "distant". As discussed by Seber (1984) the most common dissimilarity measure for measuring the distance between two objects (d_{ij}) is a metric that maps multi-dimensional information onto the real number line ($R^d \times R^d$: R^1) and satisfies the following axioms: (1) $d(i, j) \geq 0$, (2) $d(i, j) = 0$ if and only if i=j, (3) $d(i, j) = d(j, i)$ for all i, j, and (4) $d(i, j) \leq d(i, k) + d(k, j)$, for all i, j, and k.

To be considered as a metric, axioms (1) and (2) imply that the dissimilarity function is positive definite, axiom (3) implies symmetry, and axiom (4) is the triangle inequality.

Jading and Sibson (1971) first used the term "dissimilarity coefficient" (DC) to describe a dissimilarity function that does not satisfy axioms (2) and (4). Sibson (1972) states:

> A DC thus looks rather like a distance function, or metric. It does not necessarily have the property that $d(x, y) = 0 \implies x = y$. This is simply a reflection of the fact that two differently labelled objects might coincide in their descriptions. The other omission is a much more significant one: $d(x, z) + d(z, y) \geq d(x, y)$.

2.4 Cluster Analysis

In order to select an appropriate clustering technique, it is necessary to first examine the types of available clustering methods within the context of the data being analyzed. In this dissertation, a clustering method will be selected by answering three questions.

The first question examined is determining whether a partitioning or hierarchical clustering method more appropriate for the data. A partitioning method is used when the number of clusters has been predetermined. In contrast, a hierarchical method does not predetermine the number of clusters. Rather, the data is examined at a varying number of clusters and the "best" clustering arrangement is selected from all available clusters. The use of cluster diagnostics to select the "best" clustering arrangement is used. Such cluster diagnostics will be discussed

later in this dissertation.

The second question examined is a determination of whether an agglomerative or divisive method is more appropriate for the data. Within hierarchical clustering models, there are two basic types: agglomerative and divisive. An agglomerative clustering method starts with each of n objects (blocks) in its own cluster. Hence, the algorithm begins with n clusters. The algorithm then groups together the most similar pair of blocks. There are now $n - 1$ clusters. This grouping process repeats itself until all clusters are grouped into one large cluster. In contrast, a divisive method starts with one large cluster of all available objects (blocks). The most dissimilar cluster is then split off so that there are now two clusters. This splitting process continues until every object is in its own separate cluster. Hence, the algorithm ends with n clusters.

The third question examined is determining what type of linkage method would be most appropriate for the data. Linkage defines the distance between two clusters that is used in either the grouping or splitting process previously described. When clusters are split (or grouped), it is the linkage method that relates these two cluster to each other. For example, a commonly used clustering method involves a technique known as single linkage (nearest neighbor). If C_1 and C_2 are two clusters, then the distance between them is defined to the smallest dissimilarity between a member of C_1 and a member of C_2 (Sneath [1957], Sokal and Sneath [1963], Johnson [1967]), namely,

$$d_{(C_1)(C_2)} = \min(d_{ij}) : B_i \in C_1, B_j \in C_2 \tag{2.57}$$

In addition to single linkage, other commonly used linkage methods such as group average (average dissimilarity between two clusters), complete linkage (maximum dissimilarity between two clusters) can be classified as either space conserving or space distorting.

Some clustering algorithms do not have any options when choosing the linkage method. Others algorithms have several linkage methods to choose from. As previously stated, an important issue is to select a linkage method based upon the context of the data being analyzed. This issue will be examined in more detail in Section 3.5 of this dissertation.

2.5 Upper Prediction Limits/Upper Percentiles of Kriging Estimates

The 95% upper confidence limit for the mean (UCL) is a statistic that is commonly used by environmental regulatory agencies (MDEQ 2002). However, present research has not yet provided a theoretically complete estimation of a UCL based upon kriging statistics. Instead, an estimate of the 95% upper prediction limit will be provided in this dissertation for normally distributed and non-parametric random variables. An estimate of the 95^{th} percentile will be provided for lognormally distributed random variables.

2.5.1 Normally Distributed Random Variables

Let the observed second order stationary random variables $z_j : j = 1$ to n have a multivariate normal distribution $\mathbf{Z}$ where μ is an (n x 1) vector of identical means and $\mathbf{\Sigma}$ is an (n x n) variance/covariance matrix described as follows:

$$\begin{bmatrix} \text{Cov}(z_1, z_1) & \dots & \text{Cov}(z_1, z_n) \\ \vdots & \ddots & \vdots \\ \text{Cov}(z_n, z_1) & \dots & \text{Cov}(z_n, z_n) \end{bmatrix} \tag{2.58}$$

where $\mathbf{\Sigma}$ is symmetric and $\text{Cov}(z_i, z_j)$ represent the covariance between existing z_i, z_j observations as defined by (2.8).

Recall from (2.36) that the block kriging estimate of a block mean $\hat{z}_B = \sum_{j=1}^{n} w_{Bj} z_j$ is a linear function of the n random variables in the vector $\mathbf{Z}' = (z_1, ..., z_n)$, where $\mathbf{Z}$ has a multivariate normal distribution. Hence, an estimate of the block mean $\mathbf{w'_B Z} = \sum w_{Bj} z_j$, is normally distributed with mean $\mathbf{w'_B}\ \mu$ and variance $\mathbf{w'_B \Sigma w_B}$, where $\mathbf{w'_B} = (w_{B_1}, ..., w_{B_n})$.

Assuming the observed data are normally distributed, Cressie (1993, p. 122) discusses the estimation of the 95% upper prediction limit based on kriging estimates. Using this discussion, the 95% upper prediction limit for z_B, the block mean under the assumption of normality, may be written as:

$$UP_{BN} = \hat{z}_B + 1.645 \sqrt{\hat{\sigma}_B^2} \tag{2.59}$$

where $\hat{z}_B$ is the estimate of block mean as defined by (2.36) and $\hat{\sigma}_B^2$ is the variance of the estimation error as defined by (2.55).

2.5.2 Lognormally Distributed Random Variables

Recall in Section 2.1.3 that $z_j : j = 1$ to n may follow a lognormal distribution and a log_e transformation of z_j may be a necessary step in the spatial modelling process. If a log_e transform of the z_j is taken, estimates of mean provided by the ordinary, universal, or block kriging procedures will be based on log_e units, not on original units. This type of estimation procedure is known as lognormal kriging. As stated by Vann and Guibel (1998):

> If the data are truly lognormal, then it is possible, by taking the log, and assuming that the resulting values are multigaussian, to perform a lognormal kriging.

Hence, by taking a log_e transform of $z_j : j = 1$ to n the data may be assumed to be multivariate normally distributed and the properties discussed in Section 2.5.1 are repeated here. Let $\mathbf{Z}^* = log_e\mathbf{Z}$ such that $\mathbf{Z}^*$ is multivariate normally distributed with mean μ^* and variance/covariance matrix Σ^* as defined in Section 2.5.1.

Thus, an estimate of block mean $\mathbf{w}'_\mathbf{B}\mathbf{Z}^* = \sum w_{Bj} z_j^*$, is normally distributed with mean $\mathbf{w}'_\mathbf{B}\ \mu^*$ and variance $\mathbf{w}'_\mathbf{B}\mathbf{\Sigma}^*\mathbf{w}_\mathbf{B}$, where $\mathbf{w}'_\mathbf{B} = (w_{B_1}, ..., w_{B_n})$.

In order to express the lognormal kriging block estimates in original units, a back-transformation of these (log_e) based kriging estimates is necessary. A 95% upper prediction limit for z_B, the block mean, under the assumption of

lognormality may be written as:

$$UP_{BL} = \exp\left(\hat{z}_B^* + 1.645\sqrt{\hat{\sigma}_B^{2*}}\right) \tag{2.60}$$

where $\hat{z}_B^*$ is the estimate of block mean as defined by (2.36) resulting from the lognormal kriging procedure and $\hat{\sigma}_B^{2*}$ is the variance of the estimation error as defined by (2.55) resulting from the lognormal kriging procedure.

2.5.3 Non-Parametric Random Variables

If the observed second order stationary data z_j do not follow either a normal or lognormal distribution, a non-parametric prediction interval may be considered.

Helsel and Hirsch (1993) describe a one-sided $100(1-\alpha)\%$ nonparametric upper prediction interval:

$$UP_{NP} = Z_{\left([1-\alpha]\times(n+1)\right)} \tag{2.61}$$

which is calculated from the ordered array $(Z_1, ..., Z_n)$ of the observed data Z, where α represents the error risk, and n is the sample size.

CHAPTER III

Spatial Strata Modelling Method

3.1 Block Design

The first step in the Spatial Strata Modelling (SSM) method is to construct a square grid that will overlay the Geographic Area of Interest (GAI). In some cases, the GAI may be irregularly shaped. The grid must be of sufficient size to completely cover the GAI. The grid will consist of non-overlapping blocks square (B) of uniform size ($L \times L$).

The decision to use blocks in SSM, rather than individual points, is based on considering the context of the data. In most situations, it is not practical to assume the data being analyzed will come from a preconceived sampling plan. For this reason, SSM is designed to be used on observational study data. Such data may not be from an *apriori* sampling plan.

Chiles and Delfiner (1999) demonstrate that among a gridded, stratified, or random sample plan that $\sigma^2_{grid} \leq \sigma^2_{strat} \leq \sigma^2_{rand}$ where σ^2 refers to the variance of the estimation error resulting from a spatial model. The subscripts respectively refer to: (1) a gridded sampling plan in which the samples are taken at the nodes of a regular grid with square cells, (2) a stratified sampling plan in which the GAI is divided into N similar disjoint zones of influence. Within each zone of influence,

a sample is selected at random, and (3) a random sampling pattern over the GAI.

Given that a random sample may result from observational study data, other alternatives are considered to lower a spatial model's variance of estimation error. The use of block averages, rather than individual points, has been demonstrated to provide lower error variances and is verifiable from models discussed by Whittle (1962) and Modjeska and Rawlings (1983). Further, the development of an infill sampling design, as discussed in Section 3.2, may be considered as a method for lowering the model's variance of estimation error.

The determination of a block size ($L \times L$) is a non-trivial problem. For example, the default prediction grid used by the software S-Plus© is the range of the spatial locations ($x_{max} - x_{min}$) and ($y_{max} - y_{min}$) divided across 30 evenly spaced points in both x and y directions. In the case of a designed experiment, the size and spacing of the prediction grid could be determined *apriori* based on the context of the investigation (Chiles and Delfiner, 1999, Channel Tunnel). However, in the case of an observational study, the literature tends to ignore the issue of grid spacing for estimation purposes.

In order for SSM to be a repeatable methodology, a criterion to determine block size (L) is necessary. The block width L may be considered like a histogram bin width, since all of the square blocks (B) are of uniform size and non-overlapping. A data set that is to be spatially modelled will contain the spatial location of the samples (x,y)$\in$ s and the value of the random variable z(s) for

that specific spatial location. Since this spatial location is two dimensional and the blocks are two dimensional and square, two estimates of histogram bin width (Freedman and Diaconis, 1981) are calculated as follows:

$$L_x = 2R_x n^{-1/3}$$

$$L_y = 2R_y n^{-1/3}$$

In which R_x and R_y are the interquartile ranges of x and y respectively and n is the sample size of the data set. The minimum of L_x and L_y will be selected as the block width L.

$$L = min(L_x, L_y) \tag{3.1}$$

3.2 Infill Sampling Strategy

Let N be considered as all possible infill points (candidates) within the GAI. Let n be the number of infill samples chosen from the N candidates. In order to recommend an in-fill sampling strategy, it is necessary to choose an evaluation method to compare the *MinMean, MinMed, MinMax,* and *MinVar* sampling algorithms.

Recall that by using the spatial model and original data set, the variance of the block estimation error ($\sqrt{\hat{\sigma}_B^2}$) was determined for each block in the GAI. Four summary statistics (mean, median, maximum, variance) for all $\hat{\sigma}_B^2$ are determined. Further recall that $\sqrt{\hat{\sigma}_B^2}$ will decrease when additional samples are incorporated into the spatial model. The amount of decrease in $\sqrt{\hat{\sigma}_B^2}$ is specific to each block,

depending upon the spatial locations of the additional infill samples. Using the *greedy sequential exchange* algorithm, four different combinations of n infill samples will be chosen. Each of these four infill samples will be chosen based on one of the following criteria: (1) minimize the mean of all $\sqrt{\hat{\sigma}_B^2}$(*MinMean* algorithm), (2) minimize the median of all $\sqrt{\hat{\sigma}_B^2}$ (*MinMed* algorithm), (3) minimize the maximum of all the $\sqrt{\hat{\sigma}_B^2}$ (*MinMax* algorithm), and (4) minimize the variance of all $\sqrt{\hat{\sigma}_B^2}$ (*MinVar* algorithm).

Using each algorithm, the infill samples will be added to the original data and updated summary statistics (mean, median, maximum, variance) for all $\sqrt{\hat{\sigma}_B^2}$ will be calculated. Each of these revised summary statistics will be ranked against the same summary statistic obtained by the other algorithms. The ranking will range from 1 (lowest) to 4 (highest) for each summary statistic.

For example, let n additional observations to be added to the original data set from a candidate pool of N. Using a chosen spatial model, the n additional sample locations that provide the smallest mean of all the $\sqrt{\hat{\sigma}_B^2}$ will be added to the original data. The updated mean of the $\sqrt{\hat{\sigma}_B^2}$ is determined, as well as the median, maximum, and variance of the updated $\sqrt{\hat{\sigma}_B^2}$. Returning these n samples to the original data set, the *MinMed* algorithm is performed and the n additional sample locations that provide the smallest median of $\sqrt{\hat{\sigma}_B^2}$ will be added to the original data. The updated mean of $\sqrt{\hat{\sigma}_B^2}$ is determined, as well as the median, maximum, and variance of the updated $\sqrt{\hat{\sigma}_B^2}$. The *MinMax* and *MinVar* algo-

rithms are also similarly executed with the updated summary statistics of $\sqrt{\hat{\sigma}_B^2}$ recorded. The results of four decision algorithms, each with four summary statistics of the updated $\sqrt{\hat{\sigma}_B^2}$ are now compared to each other. The algorithms will be ranked 1 (lowest) through 4 (highest) for each of the four summary statistics. The algorithm with the lowest total rank will be chosen as the optimal decision algorithm for n infill samples. An example of this ranking procedure is given in Table 2. The results in Table 2 show the mean algorithm (*MinMean*) as having

Table 2. Algorithm Ranking Example

Algorithm ↓	$\sqrt{\hat{\sigma}_B^2}$ Mean	Rank	$\sqrt{\hat{\sigma}_B^2}$ Med.	Rank	$\sqrt{\hat{\sigma}_B^2}$ Max.	Rank	$\sqrt{\hat{\sigma}_B^2}$ Var.	Rank	Rank Total
MinMean	10	1	10	2	20	2	2	2	7
MinMed	12	2	8	1	40	4	3	3	10
MinMax	15	3	15	4	10	1	10	4	12
MinVar	20	4	12	3	30	3	1	1	11

the lowest mean, the second lowest median, the second lowest maximum, and the second lowest variance. With a rank total of 7, the *MinMean* algorithm would be selected as the optimal decision algorithm for this particular sample size. The *MinMed* algorithm, with a total rank of 10 would be next. The *MinVar* algorithm would be third best, followed by the *MinMax* algorithm.

Using the Greedy/Sequential Exchange Algorithm, the *MinMean, Min-*

Med, MinMax, and *MinVar* infill algorithms will be repeated over different infill sample sizes. For each sample size, the infill algorithm with the lowest rank total will be chosen as the optimal algorithm. In this dissertation, approximately 5%, 10%, and 15% of the original data points will be added to determine not only the optimal decision algorithm, but to determine if a universally optimal algorithm exists. A universally optimally decision algorithm would be one that is optimal for all evaluated infill sample sizes.

3.3 Block Covariance

Using techniques outlined in Section 2.1.10, the weights necessary to estimate the block means $\hat{z}_B$ and estimation of the variance of the block estimation error $\hat{\sigma}_B^2$ are determined. An additional statistic that will be used in calculating the SSM dissimilarity coefficient (Section 3.4) is an estimation of the covariance of the block estimation error $\hat{\gamma}(z_{B_i}, z_{B_j})$. Since the covariance of a random variable with itself is the variance, the derivation of $\hat{\gamma}(z_{B_i}, z_{B_j})$ is a generalized version of $\hat{\sigma}_B^2$. However, the literature does not derive this statistic so $\hat{\gamma}(z_{B_i}, z_{B_j})$ will be introduced in this dissertation.

Theorem 3.3.1. *Under the conditions of a stationary random process as stated in (2.1), (2.2), and (2.4), the estimated covariance of a block B_1 estimation error with any other distinct arbitrary block B_2 estimation error is:*

$$\hat{\gamma}(z_{B_1}, z_{B_2}) = \frac{1}{m_1 m_2}\sum_{k=1}^{m_1}\sum_{j=1}^{m_2}\text{Cov}(z_{B_{1k}}, z_{B_{2j}}) - \frac{1}{m_1}\sum_{k=1}^{m_1}\sum_{i=1}^{n} w_{B_{1i}}\ \text{Cov}(z_{B_{1k}}, z_i)$$
$$- \frac{1}{m_2}\sum_{j=1}^{m_2}\sum_{i=1}^{n} w_{B_{2i}}\ \text{Cov}(z_{B_{2j}}, z_i) + \sum_{i=1}^{n}\sum_{l=1}^{n} w_{B_{1i}} w_{B_{2l}}\ \ \text{Cov}(z_i, z_l) \quad (3.2)$$

where $z_{B_{1k}} : k = 1, ..., m_1$ and $z_{B_{2j}} : j = 1, ..., m_2$ refer to the m_1 and m_2 points contained within Blocks B_1 and B_2 respectively, and $\text{Cov}(z_{B_{1k}}, z_{B_{2j}})$ is the covariance between the point $z_{B_{1k}}$ in Block B_1 and the point $z_{B_{2j}}$ in Block B_2 as defined by the variogram function. The weights $w_{B_{1i}}$ and $w_{B_{2i}}$ refer to the block kriging weights assigned to observed data point z_i in order to estimate z_{B_1} and z_{B_2}, respectively. $\text{Cov}(z_{B_{1k}}, z_i)$ and $\text{Cov}(z_{B_{2j}}, z_i)$ are the covariances between observed data points z_i and the point $z_{B_{1k}}$ in Block B_1 and $z_{B_{2j}}$ in Block B_2 respectively, as defined by (2.40). $\text{Cov}(z_i, z_l)$ is the covariance between the observed data z_i and z_l as defined by the variogram function.

Proof:

The covariance of the block estimation error with any other block estimation error for a second order stationary random process can be written as:

$$\begin{aligned}
\hat{\gamma}(z_{B_1}, z_{B_2}) &= E[z_{B_1} - \hat{z}_{B_1}][z_{B_2} - \hat{z}_{B_2}] \\
&= E[z_{B_1} z_{B_2}] - E[z_{B_1}\hat{z}_{B_2}] - E[z_{B_2}\hat{z}_{B_1}] + E[\hat{z}_{B_1}\hat{z}_{B_2}] \\
&= \text{Cov}(z_{B_1}, z_{B_2}) + E[z_{B_1}]E[z_{B_2}] - \text{Cov}(z_{B_1}, \hat{z}_{B_2}) - E[z_{B_1}]E[\hat{z}_{B_2}] \\
&\quad - \text{Cov}(z_{B_2}, \hat{z}_{B_1}) - E[z_{B_2}]E[\hat{z}_{B_1}] + \text{Cov}(\hat{z}_{B_1}, \hat{z}_{B_2}) + E[\hat{z}_{B_1}]E[\hat{z}_{B_2}]
\end{aligned} \quad (3.3)$$

By (2.39), $E[z_{B_1}] = E[\hat{z}_{B_1}] = E[z_{B_2}] = E[\hat{z}_{B_2}] = \mu$. Hence,

$$\begin{aligned}\hat{\gamma}(z_{B_1}, z_{B_2}) &= \mathrm{Cov}(z_{B_1}, z_{B_2}) + \mu^2 - \mathrm{Cov}(z_{B_1}, \hat{z}_{B_2}) - \mu^2 \\ &\quad - \mathrm{Cov}(z_{B_2}, \hat{z}_{B_1}) - \mu^2 + \mathrm{Cov}(\hat{z}_{B_1}, \hat{z}_{B_2}) + \mu^2 \\ &= \mathrm{Cov}(z_{B_1}, z_{B_2}) - \mathrm{Cov}(z_{B_1}, \hat{z}_{B_2}) - \mathrm{Cov}(z_{B_2}, \hat{z}_{B_1}) + \mathrm{Cov}(\hat{z}_{B_1}, \hat{z}_{B_2})\end{aligned} \tag{3.4}$$

Using (2.37), the first term of (3.4) can be written as:

$$\begin{aligned}\mathrm{Cov}(z_{B_1}, z_{B_2}) &= \mathrm{Cov}\left[\frac{1}{m_1}\sum_{k=1}^{m_1} z_{B_{1k}}, \frac{1}{m_2}\sum_{j=1}^{m_2} z_{B_{2j}}\right] \\ &= E\left[\frac{1}{m_1}\sum_{k=1}^{m_1} z_{B_{1k}} \frac{1}{m_2}\sum_{j=1}^{m_2} z_{B_{2j}}\right] - E\left[\frac{1}{m_1}\sum_{k=1}^{m_1} z_{B_{1k}}\right] E\left[\frac{1}{m_2}\sum_{j=1}^{m_2} z_{B_{2j}}\right] \\ &= \frac{1}{m_1 m_2}\sum_{k=1}^{m_1}\sum_{j=1}^{m_2}\Big(E[z_{B_{1k}} z_{B_{2j}}] - E[z_{B_{1k}}]E[z_{B_{2j}}]\Big) \\ &= \frac{1}{m_1 m_2}\sum_{k=1}^{m_1}\sum_{j=1}^{m_2}\mathrm{Cov}(z_{B_{1k}}, z_{B_{2j}})\end{aligned} \tag{3.5}$$

which is the average of all covariances between the points $z_{B_{1k}} \in$ Block B_1 and $z_{B_{2j}} \in$ Block B_2 as defined by the variogram.

Using (2.36) and (2.37) the second term of (3.4) can be written as:

$$\begin{aligned}\mathrm{Cov}(z_{B_1}, \hat{z}_{B_2}) &= \mathrm{Cov}\left[\frac{1}{m_1}\sum_{k=1}^{m_1} z_{B_{1k}}, \sum_{i=1}^{n} w_{B_{2i}} z_i\right] \\ &= E\left[\frac{1}{m_1}\sum_{k=1}^{m_1} z_{B_{1k}} \sum_{i=1}^{n} w_{B_{2i}} z_i\right] - E\left[\frac{1}{m_1}\sum_{k=1}^{m_1} z_{B_{1k}}\right] E\left[\sum_{i=1}^{n} w_{B_{2i}} z_i\right] \\ &= \frac{1}{m_1}\sum_{k=1}^{m_1}\sum_{i=1}^{n} w_{B_{2i}}\Big(E[z_{B_{1k}}\ z_k] - E[z_{B_{1k}}]E[z_i]\Big) \\ &= \frac{1}{m_1}\sum_{k=1}^{m_1}\sum_{i=1}^{n} w_{B_{2i}}\ \mathrm{Cov}(z_{B_{1k}}, z_i)\end{aligned} \tag{3.6}$$

which is the average of the weighted covariances between the m_1 points $z_{B_{1k}} \in B_1$ and the observed data values z_i.

The third term of (3.4) is identical to the second term, except for the notation, and can be written as:

$$\text{Cov}(z_{B_{2j}}, \hat{z}_{B_1}) = \frac{1}{m_2} \sum_{j=1}^{m_2} \sum_{i=1}^{n} w_{B_{1i}} \text{ Cov}(z_{B_{2j}}, z_i) \tag{3.7}$$

which is the average of the weighted covariances between the m_2 points $z_{B_{2j}} \in B_2$ and the known values z_i.

Finally, using (2.36), the fourth term of (3.4) can be written as:

$$\begin{aligned}
\text{Cov}(\hat{z}_{B_1}, \hat{z}_{B_2}) &= \text{Cov}\left[\sum_{i=1}^{n} w_{B_{1i}} z_i, \sum_{l=1}^{n} w_{B_{2l}} z_l \right] \\
&= E\left[\sum_{i=1}^{n} w_{B_{1i}} z_i \ \sum_{l=1}^{n} w_{B_{2l}} z_l \right] - E\left[\sum_{i=1}^{n} w_{B_{1i}} z_i \right] E\left[\sum_{l=1}^{n} w_{B_{2l}} z_l \right] \\
&= \sum_{i=1}^{n} \sum_{l=1}^{n} w_{B_{1i}} w_{B_{2l}} \Big(E[z_i z_l] - E[z_i] E[z_l] \Big) \\
&= \sum_{i=1}^{n} \sum_{l=1}^{n} w_{B_{1i}} w_{B_{2l}} \ \text{Cov}(z_i, z_l)
\end{aligned} \tag{3.8}$$

which is the sum of the n^2 weighted covariances between the existing observed data points z_i and z_l.

Combining all four terms, the expression for the estimate of the covariance of the block estimation error given in (3.2) is:

$$\begin{aligned}
&\frac{1}{m_1 m_2} \sum_{k=1}^{m_1} \sum_{j=1}^{m_2} \text{Cov}(z_{B_{1k}}, z_{B_{2j}}) - \frac{1}{m_1} \sum_{k=1}^{m_1} \sum_{i=1}^{n} w_{B1_i} \text{ Cov}(z_{B_{1k}}, z_i) \\
&- \frac{1}{m_2} \sum_{j=1}^{m_2} \sum_{i=1}^{n} w_{B_{2i}} \text{ Cov}(z_{B_{2j}}, z_i) + \sum_{i=1}^{n} \sum_{l=1}^{n} w_{B_{1i}} w_{B_{2l}} \ \text{Cov}(z_i, z_l)
\end{aligned}$$

3.4 Dissimilarity Coefficient

A dissimilarity coefficient (DC) comparing the statistical distance between two spatially estimated blocks selected from the GAI is defined as follows:

Definition 3.4.1. d_{ij} is a dissimilarity coefficient (DC) between Blocks B_i and B_j and is expressed as:

$$\begin{aligned} &\text{for } i = j: \qquad d_{ij} = \ 0 \\ &\text{for } i \neq j: \qquad d_{ij} = \left|(\hat{z}_{B_i} - \hat{z}_{B_j})\right| \Big/ \sqrt{\hat{\sigma}^2_{B_i} + \hat{\sigma}^2_{B_j} - 2\hat{\gamma}(z_{B_i}, z_{B_j})} \end{aligned} \tag{3.9}$$

Recall from (2.36) that $\hat{z}_B = \sum w_{B_j} z_j$ is the estimate of block mean and that the estimate of block error variance $\hat{\sigma}^2_B$ may be determined using (2.55). In addition, $\hat{\gamma}(z_{B_i}, z_{B_j})$ may be determined using (3.2).

A review of the mathematical properties of d_{ij} will now be considered. Recall the axioms stated in Section 2.3 of this dissertation.

Axiom 1: $d_{ij} \geq 0$, all i, j in $\mathbb{R}^d$.

By definition, d_{ij} will always be greater than or equal to 0.

Axiom 1 is satisfied.

Axiom 2: $d_{ij} = 0$ if and only if $i = j$.

Let two different blocks ($i \neq j$) have the same estimated mean. Thus, $\hat{z}_{B_i} = \hat{z}_{B_j}$ and $d_{ij} = 0$. Since $d_{ij} = 0$ when $i \neq j$ the axiom is violated.

Axiom 2 is not satisfied

Axiom 3: $d_{ij} = d_{ji}$, all i, j in $\mathbb{R}^d$.

Recall that an isotropic variogram function is used to determine the components of d_{ij}. Further recall than an isotropic variogram function depends only on the separation distance h and not on the direction of h. Hence, an interchange of i and j does not change the separation distance h and thus will have no effect on d_{ij}

Axiom 3 is satisfied.

Axiom 4: $d_{ij} \leq d_{ik} + d_{kj}$, all i, j, k

In order to contradict $d_{ij} \leq d_{ik} + d_{kj}$ it must be shown that conditions may exist in which $d_{ij} > d_{ik} + d_{kj}$. In other words, the maximum distance measurement between two blocks is greater than the sum of the two smaller distances measurements. Recall from the investigation of infill strategy that Aspie and Barnes (1990) discussed that kriging standard error is determined only by the sample locations and the variogram, which are known, and not on the random variables z(s). Recall the dissimilarity coefficient:

$$\text{for } i = j: \qquad d_{ij} = \ 0$$

$$\text{for } i \neq j: \qquad d_{ij} = \left|(\hat{z}_{B_i} - \hat{z}_{B_j})\right| \Big/ \sqrt{\sigma^2_{B_i} + \sigma^2_{B_j} - 2\hat{\gamma}(z_{B_i}, z_{B_j})}$$

which contains an estimate of block mean ($\hat{z}_B = \sum w_{B_j} z_j$) in the numerator and a function of the estimation of error variance σ_B^2 in the denominator. It can be demonstrated that a kriging model exists in which the triangle equality is violated using the dissimilarity coefficient definition given in (3.9). This violation is illustrated in the following example:

Kriging Model Example: Violation of Triangle Inequality for d_{ij}

Let Figure 6 depict $Z(s) : s \in V_1 \subset \mathbb{R}^2$, the realization of a random process observed on a plot of land V_1 that is being evaluated. Let (B_1, B_2, B_3) be three blocks of interest. To simplify things, let there be only three data points $z(s) : (z_1, z_2, z_3)$ that are shown on Figure 6 with spatial location (x, y) and COC concentration $z(s)$ as indicated in Table 3. Let the variogram for V_1 be chosen to be an exponential model with a range of 3, a sill of 1, and a nugget of 0. Clearly, all of these conditions are possible.

Figure 6. Data Points and Blocks of Interest

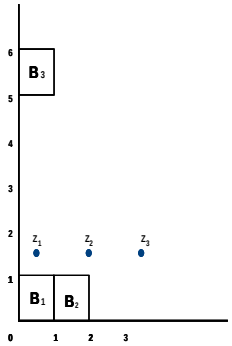


Table 3. Data Points - Spatial Location and COC Concentration

Point	Location (x, y)	COC Concentration $z(s)$
z_1	(.5,1.5)	5
z_2	(2,1.5)	2
z_3	(3.5,1.5)	1

The estimate of block means $(\hat{z}_{B_1}, \hat{z}_{B_2}, \hat{z}_{B_3})$ are given in Table 4. The variance/covariance matrix of the estimation errors for all three blocks is shown in Table 5. From these tables, it can be shown that:

Table 4. Block Means

Block	$\hat{z}_B$
1	3.00
2	1.76
3	2.85

Table 5. Block Covariances

Block	1	2	3
1	.121	-.110	-.0898
2		.121	-.0697
3			1.26

$d_{12} = 1.82,\ d_{13} = .120,$ and $d_{23} = .869.$ Clearly $d_{12} > d_{13} + d_{23}$ and the triangle equality is violated.

Axiom 4 is not satisfied for this model.

In summary, for the dissimilarity coefficient d_{ij} defined in (3.9) axiom (1) and axiom (3) are satisfied; axiom (2) and axiom (4) are not satisfied. Hence, d_{ij} is not a metric and must be given a different classification. Jading and Sibson (1972) first used the term "dissimilarity coefficient" (DC) to describe a dissimilarity function that does not satisfy axioms (2) and (4). Sibson (1972) states:

> A DC thus looks rather like a distance function, or metric. It does not necessarily have the property that
>
> $$d(x, y) = 0 \Longrightarrow x = y;$$
>
> this is simply a reflection of the fact that two differently labelled objects might coincide in their descriptions. The other omission is a much more significant one:
>
> $$d(x, z) + d(z, y) \geq d(x, y).$$

Sibson (1972) further argues that the triangle inequality forces consideration of adding DC values, and it may not be possible to do that. Sibson points out that just because a DC is a number, there is no reason to assume that all operations

performed on any set of numbers may be performed on the DC. Hence, Sibson states that the triangle inequality is not an essential requirement. For these reasons, the term dissimilarity coefficient (DC) was created and the d_{ij} dissimilarity coefficient given in (3.9)will be considered as such.

3.5 Cluster Analysis

Let $d_{ij} = [D]$, a $b \times b$ matrix containing the d_{ij} between all pairwise "b" blocks defined in Section 3.4. It is necessary to determine the number and boundaries of similar levels of COC concentrations within the GAI. This will be accomplished by choosing an appropriate clustering technique on $[D]$ and grouping the blocks into one of $n \geq 1$ cluster(s).

In order to select an appropriate clustering technique, the types of available clustering methods should be considered within the context of the data. Recall the questions posed in Section 2.4:

1. Is a partitioning or hierarchical clustering method more appropriate for the data?

Examining this question in the context of the data, there is clearly no *apriori* knowledge of the number of clusters in the arsenic data. From Figure 1.3.3 the southeast area of Michigan has arsenic levels that are higher than the other areas, but determining how many unique clusters exist is an important issue of this dissertation. Hence, a hierarchical model will be chosen.

2. Is an agglomerative or divisive method more appropriate for the data?

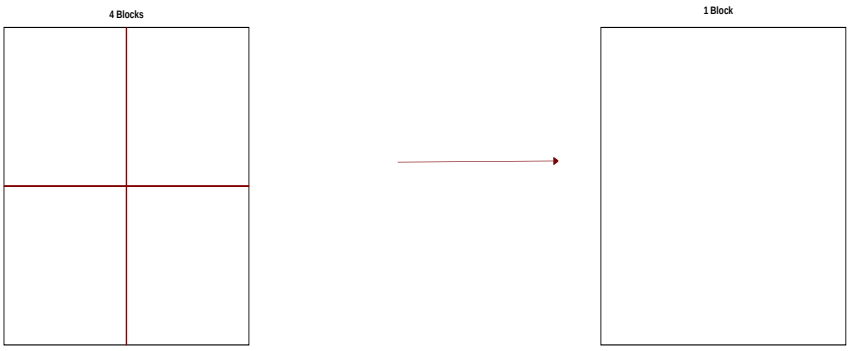
In examining this question within the context of the data, recall Section 2.1.2. It was stated that $Z(s) : s \in G(\text{arsenic data})$ represents the realization of a random process observed at a GAI. Further recall Section 2.1.3 in which COCs often follow a lognormal distribution. Intuitively, it might be concluded that since the random variables come from the same distribution that a divisive method might be more appropriate. The random variables are not being grouped in order to create a distribution, the distribution already exists and the random variables from this particular distribution are being partitioned into groups. If a parametric distribution were partitioned based on its quantiles $(Min - Q_1, Q_1 - Q_2, Q_2 - Q_3, Q3 - Max)$, a type of divisive method would be utilized since an existing distribution is being considered.

3. What type of linkage method would be most appropriate for the data?

Again, within the context of the arsenic data, the concept of a block should be considered. It has already been determined that d_{ij} is a function of the kriging estimates of block mean and variance/covariance of the block estimation errors for block B_i and B_j. Recall that a block is made up of an infinite number of points, but that $m = 16$ points were chosen to represent a block. As shown in (2.38), the average of the individual points within a block were used to determine the estimate of block mean. Similarly, (2.40)

shows that the average of the covariances between all $z_k \in B$ and z_j, was used to represent the covariance between block B and the observed random variables z_j. Further recall in Section 3.1, in which a method was introduced to determine the block size L. The method is reasonable, but still arbitrary. As shown in Figure 7, if block size L were increased, the 4 blocks shown could

Figure 7. Increase in Block Size



be theoretically transformed into the larger block shown. Thus, the spatial model could be reconfigured to calculate $\hat{z}_B$ as the average of all points contained in this single larger block. Hence, it seems reasonable that using an average to represent a cluster's dissimilarity is consistent with utilizing a block to model the random process.

In summary, the appropriate clustering method should meet the following requirements: (1) the clustering method should analyze the dissimilarity between objects, since d_{ij}, as defined in (3.9), is a measurement of dissimilarity, (2) the clustering method should be a hierarchical clustering method, since the number

of clusters is not known *apriori*, (3) the clustering method should use a divisive technique, since the random process Z(s) will very often follow a lognormal distribution, and (4) the clustering method should use an average linkage method, since a block is represented as an average of its individual points.

Kaufman and Rousseeuw (1990) have provided numerous modern techniques for cluster analysis. These stand-alone Fortran programs were collectively known as the library CLUSFIND. Struyf, Hubert, and Rousseeuw (1997) have incorporated these programs into the software S-Plus© as executable functions. Of these S-Plus functions, the clustering algorithm *Diana* appears to meet all four of the stated requirements.

As discussed in Struyf, Hubert, and Rousseeuw (1997), *Diana* is a divisive hierarchical method. The initial clustering (at step 0) consists of one large cluster containing all n objects. In each subsequent step, the cluster C with largest diameter (defined as $\text{diam}(C) := max_{i,j \in C}\, d(i,j)$) is split into two smaller clusters. The method is further described in Struyf, Hubert, and Rousseeuw (1997) and consists of the following five steps:

1. Split up cluster C into two clusters A and B. At first cluster $A := C$ and cluster $B := \emptyset$.

2. For each object $i \in A$, calculate $a(i)$, the average dissimilarity to all other objects of A. The block B of A for which $a(B)$ is the largest is moved to B:

3. Calculate $a(i)$ for all $i \in A$, and the average dissimilarity of i to all objects of B, denoted as $d(i, B)$. Select the object $h \in A$ that $a(h) - d(h, B) = max_{i \in A}(a(i) - d(i, B))$.

4. If $a(h) - d(h, B) > 0 \rightarrow$ move h from A to B and repeat step 2. In other words, the object moved from A to B is more dissimilar to A than to B.

5. If $a(h) - d(h, B) \leq 0 \rightarrow$ the process stops. Keep A and B as they are now.

To assist in the evaluation of different cluster sizes, *Diana* provides a clustering tree and a clustering banner. A clustering tree defines the cluster membership at $g : g = 1$ to n clusters. A clustering banner provides the divisive coefficient (Rousseeuw 1986), which measures the cluster structure of the data set. For each object i, denote by $d(i)$ the average dissimilarity of the last cluster to which it belongs (before being split off as a single object) divided by the average dissimilarity of the whole data set. The divisive coefficient is then defined as the average of all $d(i)$ or the average amount of coverage (blackness) on the banner plot. A very pronounced structure implies that the diameter of the entire data set is much larger than the diameters of the individual clusters, leading to a wide banner. Since the width of the banner reflects the strength of the clustering, a higher divisive coefficient indicates better clustering structure.

Since neither the tree diagram nor clustering banner provides the "best" number of clusters, this issue must be explored further. A criterion proposed by Calinski and Harabasz [1974] seeks to compare the between cluster sum of squares

versus the within cluster sum of squares as follows:

$$C = \Big(B/(g-1)\Big) \Big/ \Big(W/(b-g)\Big) \tag{3.10}$$

in which:

$B = \sum_{i=1}^{g} a_i(\hat{z}_{B_{i.}} - \hat{z}_{B_{..}})^2$ = sum of squares between clusters in which $\hat{z}_{Bi.}$ represents the average of the block estimates ($\hat{z}_{B_j}$) for all blocks $B_j \in$ Cluster C_i, $\hat{z}_{B_{..}}$ represents the average of the block estimates over the entire GAI, a_i represents the total number of blocks $B_j \in$ Cluster C_i, and

$W = \sum_{i=1}^{g} \sum_{j=1}^{a_i} (\hat{z}_{B_{ij}} - \hat{z}_{B_{i.}})^2 =$ sum of squares within clusters in which $\hat{z}_{B_{ij}}$ represents the individual blocks $B_j \in$ Cluster C_i, $g =$ total number of clusters, and $b =$ the total number of blocks.

A value of C increasing monotonically with g suggests no cluster structure, whereas C decreasing monotonically with g suggests a hierarchical structure. When C rises to a maximum at g this suggests the presence of g clusters.

3.6 Strata - Spatially Weighted Clusters

Recall that the grid discussed in Section 3.1 is being overlayed on an irregularly shaped GAI. Hence, if a cluster contains blocks that are not 100% filled with GAI area, those blocks must be spatially weighted by their contribution to their cluster's total land area.

Utilizing Arc/Info©, a GIS software package, the quantity of geographic area contained in each block may be calculated. Knowing the block membership of each cluster allows a determination of the percent contribution by each block contained in a particular cluster. This percentage will define a spatial weight v_i.

Definition 3.6.1. A spatial weight v_i for Block i may be expressed as:

$$v_i = A[B_i] \Big/ \sum_{i=1}^{b} A[B_i] \qquad : B_i \in C \tag{3.11}$$

such that

$$\sum_{i=1}^{b} v_i = 1.$$

The spatial weight v_i is the percent contribution of each Block i's area to that block's total cluster area. In addition, $A[B_i]$ = the area of block B_i, b is the number of blocks contained in Cluster C, and $\sum_{i=1}^{b} A[B_i]$ = the total area of cluster C.

Definition 3.6.2. A stratum S is a cluster (C) resulting from a second order stationary random process with spatially weighted estimates of mean and variance. The spatial weights are defined by Definition 3.6.1. The geographic boundaries of the stratum are defined by the cluster's member blocks $(B_i \in C \equiv S)$.

Definition 3.6.3. A stratum mean (z_S) is expressed as

$$z_S = \sum_{i=1}^{b} v_i z_{B_i} : B_i \in C \equiv S$$

in which the block spatial weights v_i are defined by Definition 3.6.1 and cluster (C) member blocks B_i are defined by (2.38).

Definition 3.6.4. An estimate of stratum mean ($\hat{z}_S$) is expressed as

$$z_S = \sum_{i=1}^{b} v_i \hat{z}_{B_i} = \sum_{i=1}^{b} \sum_{j=1}^{n} v_i w_{B_{ij}} z_j : B_i \in C \equiv S$$

in which spatial weights v_i are defined by Definition 3.6.1, $\hat{z}_{B_i}$ is defined by (2.39), $w_{B_{ij}}$ are the block kriging weights for observed data point z_j used to estimate Block i.

Theorem 3.6.1. *Under the conditions of a stationary random process as stated in Section 2.1.1, the expected value of a stratum* $E[z_S] = \mu$

Proof:

$$E[z_S] = E\Big[\sum_{i=1}^{b} v_i z_{B_i}\Big]$$

It has already been shown by (2.39) that $E[z_B = \mu]$. Thus:

$$E[z_S] = \sum_{i=1}^{b} v_i \mu \ .$$

Since by Definition 3.6.1

$$\sum_{i=1}^{b} v_i = 1,$$

it follows that:

$$E[z_S] = \mu.$$

Theorem 3.6.2. *Under the conditions of a stationary random process as stated in Section 2.1.1, the estimate of stratum mean $\hat{z}_S$ is an unbiased estimator of z_S, the stratum mean.*

Proof:

$$E\Big[\hat{z}_S\Big] = E\Big[\sum_{i=1}^{b} v_i \hat{z}_{B_i}\Big]$$
$$= E\Big[\sum_{i=1}^{b} v_i \sum_{j=1}^{n} w_{B_{ij}} z_j\Big]$$

where v_i is the spatial weight for Block i and $w_{B_{ij}}$ are the block kriging weights for observed data point z_j used to estimate Block i. Hence,

$$E\Big[\hat{z}_S\Big] = \sum_{i=1}^{b} v_i \sum_{j=1}^{n} w_{ij} E[z_j]$$

Because of stationarity (2.1), $E[z_j] = \mu$. Thus:

$$= \sum_{i=1}^{b} v_i \sum_{j=1}^{n} w_{B_{ij}} \mu$$

Since by Definition 3.6.1 and (2.44), $\sum v_i = \sum w_{B_{ij}} = 1$,

$$E\Big[\hat{z}_S\Big] = \mu \tag{3.12}$$

Using these definitions and theorems, it is now possible to derive the estimate of the variance of the stratum estimation error.

Theorem 3.6.3. *Under the conditions of a stationary random process as stated in Section 2.1.1, the variance of a stratum's estimation error $\hat{\sigma}_S^2$ is defined as:*

$$\mathrm{Var}(\hat{z}_S) = \sum_{i=1}^{b}\sum_{j=1}^{b} v_i v_j \mathrm{Cov}(z_{B_i}, z_{B_j}) - 2\sum_{i=1}^{b}\sum_{j=1}^{b} v_i v_j \mathrm{Cov}(z_{B_i}, \hat{z}_{B_j}) + \sum_{i=1}^{b}\sum_{j=1}^{b} v_i v_j \mathrm{Cov}(\hat{z}_{B_i}, \hat{z}_{B_j}) \tag{3.13}$$

where b is the number of blocks used to represent a stratum, $\mathrm{Cov}(z_{B_i}, z_{B_j})$ is shown as (3.5), $\mathrm{Cov}(z_{B_i}, \hat{z}_{B_j})$ is shown as (3.6), and Cov $(\hat{z}_{B_i}, \hat{z}_{B_j})$ is shown as (3.8).

Proof:

Let $\hat{\sigma}_S^2 = E(z_S - \hat{z}_S)^2$ where z_S is the true stratum mean as given by Definition 3.6.3 and $\hat{z}_S$ is an unbiased estimate of z_S as defined by (3.6.4). It can then be shown that:

$$\begin{aligned}
\hat{\sigma}_S^2 &= E[z_S - \hat{z}_S]^2 \\
&= E[z_S]^2 - 2E[z_S\hat{z}_S] + E[\hat{z}_S]^2 \\
&= \mathrm{Var}(z_S) + E[z_S]^2 - 2\mathrm{Cov}(z_S, \hat{z}_S) - 2E[z_s]E[\hat{z}_S] + \mathrm{Var}(\hat{z}_S) + E[\hat{z}_S]^2 \\
&= \mathrm{Var}(z_S) + \mu^2 - 2\mathrm{Cov}(z_S, \hat{z}_S) - 2\mu^2 + \mathrm{Var}(\hat{z}_S) + \mu^2 \\
&= \mathrm{Var}(z_S) - 2\mathrm{Cov}(z_S, \hat{z}_S) + \mathrm{Var}(\hat{z}_S)
\end{aligned} \tag{3.14}$$

The first term to the right of the equal sign in (3.14) is the variance of z_S.

$$\begin{aligned}\text{Var}(z_S) &= \text{Var}\sum v_i z_{B_i} \\ \text{Var}(z_S) &= \sum_{i=1}^{b}\sum_{j=1}^{b} v_i v_j \text{Cov}(z_{B_i}, z_{B_j})\end{aligned}$$

where $\text{Cov}(z_{B_i}, z_{B_j})$ can be evaluated using (3.5) for blocks B_i and B_j.

The second term to the right of the equal sign in (3.14) is the covariance between stratum z_S and its estimate $\hat{z}_S$. Using the stated Definitions (3.6.3) and (3.6.4), it can be shown that:

$$\begin{aligned}\text{Cov}(z_S, \hat{z}_S) &= \text{Cov}\Big(\sum v_i z_{B_i}, \sum v_i \hat{z}_{B_i}\Big) \\ &= \sum_{i=1}^{b}\sum_{j=1}^{b} v_i v_j \text{Cov}(z_{B_i}, \hat{z}_{B_j})\end{aligned} \tag{3.15}$$

where v_i and v_j are the spatial weights assigned to the blocks. The expression $\text{Cov}(z_{B_i}, \hat{z}_{B_j})$ can be evaluated using (3.6) for blocks B_i and B_j.

The third term to the right of the equal sign in (3.14) is the variance of $\hat{z}_s$. Using Definition 3.6.4, it can be shown that:

$$\begin{aligned}\text{Var}(\hat{z}_S) &= \text{Var}\Big(\sum v_i \hat{z}_{B_i}\Big) \\ &= \sum_{i=1}^{b}\sum_{j=1}^{b} v_i v_j \text{Cov}(\hat{z}_{B_i}, \hat{z}_{B_j})\end{aligned} \tag{3.16}$$

where Cov $(\hat{z}_{B_i}, \hat{z}_{B_j})$ can be evaluated using (3.8) for blocks B_i and B_j, replacing blocks B_1 and B_2.

Combining these three terms, the variance of a stratum's estimation error can be expressed as given in (3.13) as:

$$\mathrm{Var}(\hat{z}_S) = \sum_{i=1}^{b}\sum_{j=1}^{b} v_i v_j \mathrm{Cov}(z_{B_i}, z_{B_j}) - 2\sum_{i=1}^{b}\sum_{j=1}^{b} v_i v_j \mathrm{Cov}(z_{B_i}, \hat{z}_{B_j}) + \sum_{i=1}^{b}\sum_{j=1}^{b} v_i v_j \mathrm{Cov}(\hat{z}_{B_i}, \hat{z}_{B_j})$$

3.7 Default Standards of Stratum Estimates

Within the context of this dissertation, there are three types of observed data that may be evaluated by SSM: (1) normally distributed random variables, (2) lognormally distributed random variables, and (3) non-parametric: random variables determined to be neither normally nor lognormally distributed. Using the results of SSM, each of these three types of data will now be examined in order to provide estimates of a stratum's upper default standard.

3.7.1 Normally Distributed Random Variables

Let the observed data z_j follow a multivariate normal distribution as described in Section 2.5.1. Recall in this section that it was shown that an estimate of block mean $\hat{z}_B = \mathbf{w_B'Z} = \sum w_{Bj} z_j$, is normally distributed with mean $\mathbf{w_B'}\mu$ and variance $\mathbf{w_B'\Sigma w_B}$.

Using Definition 3.6.4, an estimate of a stratum mean

$$z_S = \sum_{i=1}^{b} v_i \hat{z}_{B_i} = \sum_{i=1}^{b}\sum_{j=1}^{n} v_i w_{B_{ij}} z_j : B_i \in C \equiv S$$

where v_i is the spatial weight for Block i and $w_{B_{ij}}$ are the block kriging weights for observed data point z_j used to estimate Block i in the stratum. Definition 3.6.4

may also be considered in matrix form as $(\mathbf{v}'\mathbf{W}')\mathbf{Z}$ where $\mathbf{v}$ is a (b x 1) vector of spatial weights, $\mathbf{W}$ is a (n x b) matrix of all b block kriging weights assigned to each of the n observed random variables, and $\mathbf{Z}$ is a (n x 1) vector of the observed random variables. Hence the stratum estimate $\hat{z}_S = (\mathbf{v}'\mathbf{W}')\mathbf{Z}$ is a weighted linear combination of the observed random variables and will be normally distributed with with mean $(\mathbf{v}'\mathbf{W}')\ \mu$ and variance $(\mathbf{v}'\mathbf{W}')\mathbf{\Sigma}(\mathbf{v}'\mathbf{W}')'$.

Using (2.59), it follows then that an estimate of the 95% upper prediction limit for a stratum, under the assumption of normality, may be expressed as:

$$\widehat{UP}_{SN} = \hat{z}_S + 1.645\sqrt{\hat{\sigma}_S^2} \tag{3.17}$$

where $\hat{z}_S$ the estimate of stratum mean and is defined by Definition 3.6.4, and $(\hat{\sigma}_S^2)$ is the variance of the stratum estimation error defined by (3.13).

3.7.2 Lognormally Distributed Random Variables

For lognormally distributed data, the estimate of the 95^{th} percentile of a lognormally distributed stratum is expressed as:

$$\widehat{UP}_{SLN} = \exp\left(\hat{z}_S^* + 1.645\sqrt{\hat{\sigma}_S^{2*}}\right) \tag{3.18}$$

where $\widehat{UP}_{SLN}$ is the back-transformed estimate of the 95^{th} percentile from a log-normally distributed stratum, $\hat{z}_S^*$ is the estimate of log_e transformed stratum mean defined by theorem 3.6.2 resulting from the lognormal kriging procedure and $\hat{\sigma}_S^{2*}$) is the variance of the stratum estimation error defined by (3.13), substituting log_e

transformed random variables $\hat{\sigma}_S^{*2}$ for σ_S^2, etc. resulting from the lognormal kriging procedure

3.7.3 Non-Parametric Default Standard

Let the observed random variables z_j follow neither a normal nor a lognormal distribution. The results of the cluster analysis discussed in Section 3.5 will identify the block membership for each cluster. By knowing the block membership for each cluster, by definition 3.6.2, the spatial boundaries for each stratum are determined. In order to estimate a non-parametric default standard for each stratum, the existing data inside each individual stratum will be analyzed. Using (2.61), a one-sided 95% upper prediction interval for each stratum may be expressed as:

$$\widehat{UP}_{SNP} = Z_{.95\times(n+1)} \tag{3.19}$$

where UP_{SNP} is the non-parametric estimate of the 95% upper prediction interval from an ordered array $(Z_1, ..., Z_n)$ of the stratum data, Z represents the observed data for a particular stratum, and n represents the sample size for that particular stratum.

If the calculated value (.95 x (n+1)) from the ordered array is not an integer, then linear interpolation between the nearest two integers from the ordered array should be used.

CHAPTER IV

Spatial Strata Modelling of Arsenic Data

4.1 Block Width

Recall that the first step in the Spatial Strata Modelling (SSM) method is to construct a square grid that will overlay the geographic area of interest (GAI). A review of the arsenic data provides the following summary statistics for the x (easting) and y (northing) spatial locations in Michigan GeoRef coordinates:

Table 6. Summary Statistics of Spatial Locations: Arsenic Data

Statistic	X (Easting)	Y (Northing)
Minimum	192,986	151,975.2
Q1	570,835.1	241,437
Median	633,240	324,095
Q3	714,588.5	472,212.5
Maximum	780,108.3	844,974
Sample Size (n)	219	219

Using the summary statistics in Table 6 and (3.1), the block widths for the arsenic data may be determined as follows:

$$L = min\ [2\ R_x n^{-1/3}, 2\ R_y n^{-1/3}]$$

$$L = min\ [2\ (Q_3 - Q_1)_x\ n^{-1/3},\ 2\ (Q_3 - Q_1)_y\ n^{-1/3}]$$

$$L = min\ [2\ (714588.5 - 570835.1)\ 219^{-1/3},\ 2\ (472212.5 - 241437)\ 219^{-1/3}]$$

$$L = min\ [47698, 76572]$$

$$L = 47698$$

L will be rounded up to 48000 (meters), which is approximately 29 miles.

4.2 Grid Construction

Utilizing Arc/Info 8.1©, a GIS software tool, a grid containing a total of $B = n \times m$ Blocks that are each 48000 by 48000 meters will be created, overlayed, and centered on the State of Michigan. Once completed, the grid will contain "n" blocks along the horizontal axis, "m" blocks along the vertical axis, and completely cover the state. For example, to calculate n and m, Arc/Info 8.1© (**describe** command) provides the following spatial boundaries for the entire State of Michigan (in Michigan GeoRef coordinates).

Table 7. Michigan Spatial Boundaries

Boundary	Michigan GeoRef Coordinate
X_{min}	161,257.906
X_{max}	791,764.375
Y_{min}	128,017.859
Y_{max}	851,605.688

The calculation of n is as follows:

$$n = (X_{max} - X_{min}) \Big/ L$$

$$n = (791764.375 - 161257.906) \Big/ 48000$$

$$n = 630506.469 \Big/ 48000$$

$n = 13.14$ or $n = 14$ to ensure complete coverage.

Similarly, $m = 15.07$ or $m = 16$ to ensure complete coverage.

In order to center the grid, it is necessary to determine the total amount of grid overlap for each axis (O_x, O_y) and divide by 2. Doing this ensures that the grid is centered because the east/west and north/south sides of the Michigan map will then have equal amounts of overlap.

For example:

$$O_x = \Big((n \times L) - (X_{max} - X_{min})\Big) \Big/ 2$$

$$O_x = \Big((14 \times 48000) - (791764.375 - 161257.906)\Big) \Big/ 2$$

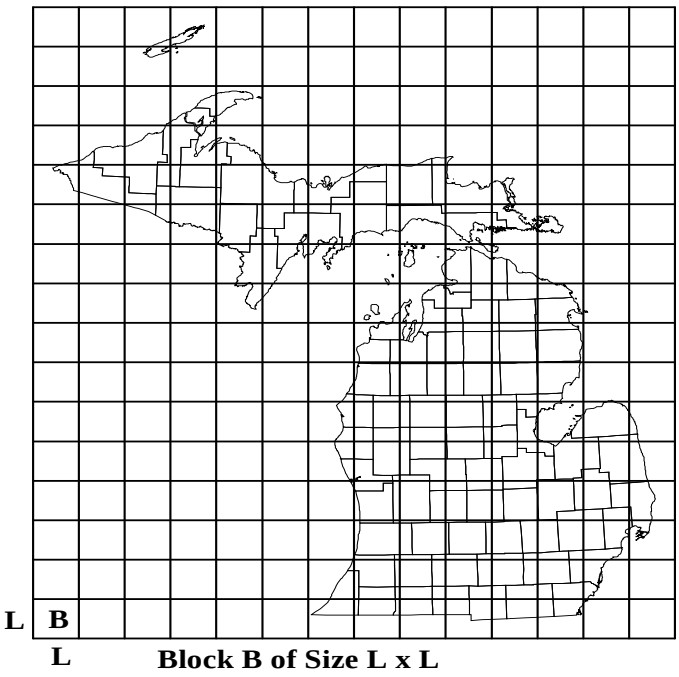
$$O_x = 20747.7655$$

The left side of the first column of blocks will be located at $(X_{min} - O_x)$ or at 140508.1405 Michigan GeoRef coordinates. Each of the 14 blocks will be 48000 meters wide until the right hand side of the last column of blocks is located at $(X_{max} + O_x)$ or at 812511.1405 Michigan GeoRef coordinates. Performing this same procedure for each row of blocks will center the grid in the vertical direction.

As shown in Figure 8, there are total of $B = n \times m = 14 \times 16 = 224$ blocks that are centered over the State of Michigan. Each of these blocks is $L = 48000$

by $L = 48000$ meters in size.

Figure 8. State of Michigan Overlayed with Grid of 224 Blocks



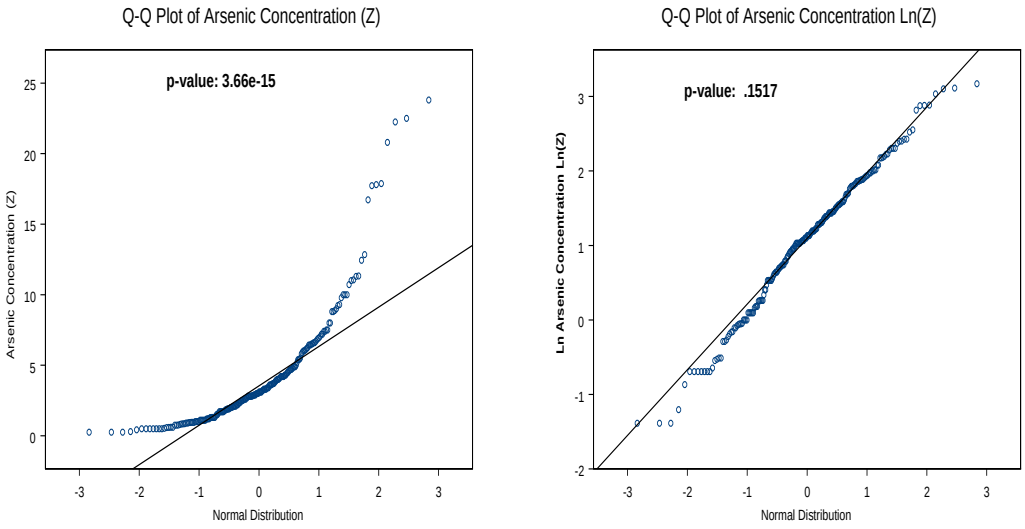
4.3 Data Transformation and Trends

Before the arsenic data variogram is estimated, exploratory data analysis (EDA) is necessary to investigate the arsenic data with respect to the presence of any skewness and/or trends in the data. With a mean of 4.25 ppm, a median of 3.1 ppm, and a skewness (g) of 2.34, the arsenic data appears to be right skewed, so $z^*(s) = log_e[z(s)]$ may be an appropriate transformation.

Figure 9 shows Q-Q plots for the raw $z(s)$ data (left plot) and for $log_e[z(s)]$ (right plot). The p-values given in Figure 9 are from the Shapiro-Francia (1972)

goodness of fit test for normality and for lognormality.

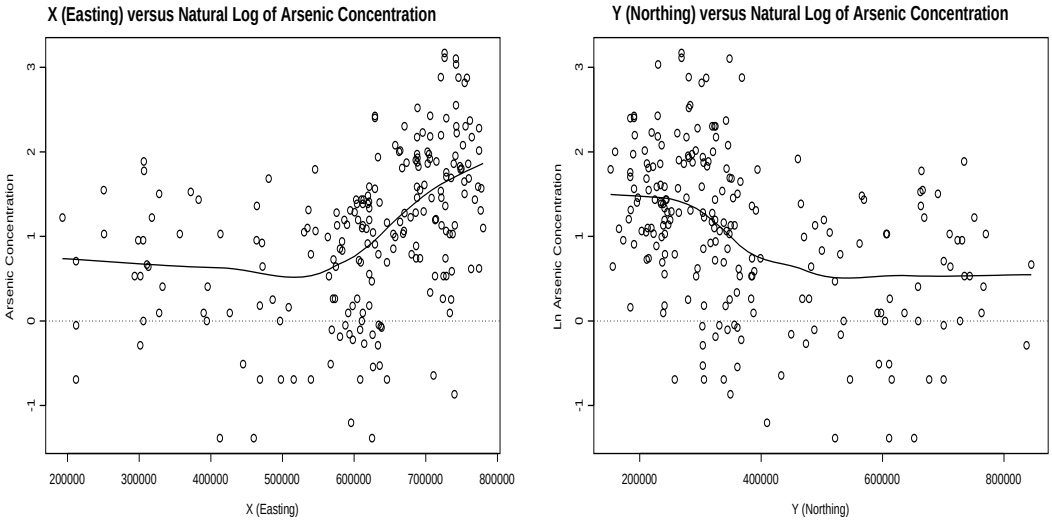
Figure 9. Q-Q plots of Raw and Transformed Arsenic Concentration



From Figure 9, it is clear that $z(s)$ follows a lognormal distribution more closely than a normal distribution. The tails for the lognormal distribution (right plot), particularly the lower tail, appear to be a bit heavy. But since the p-value (.1517) for the Shapiro-Francia goodness-of-fit test is not significant at $\alpha = .05$, sufficient evidence exists to conclude that the arsenic data follows a lognormal distribution. Figure 10 contains scatter plots of $log_e[z(s)]$ versus the easting location x and $log_e[z(s)]$ versus the northing location y. The solid line is a smoothing curve using the loess fitting method (S-Plus© *scatter.smooth* command) to assist in the identification of any trends in the data. The lower dotted line is a reference line.

From Figure 10 it is apparent that $log_e[z(s)]$ is trending higher as x in-

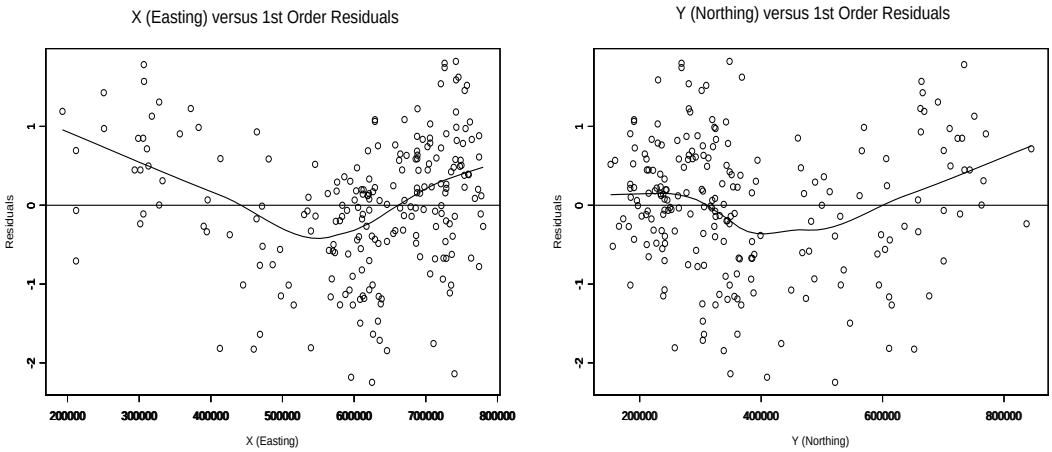
Figure 10. Scatter Plots



creases and y decreases. The order (quadratic, cubic, etc.) of the trend function cannot be determined from this plot, but clearly some trend modeling is required for the arsenic data.

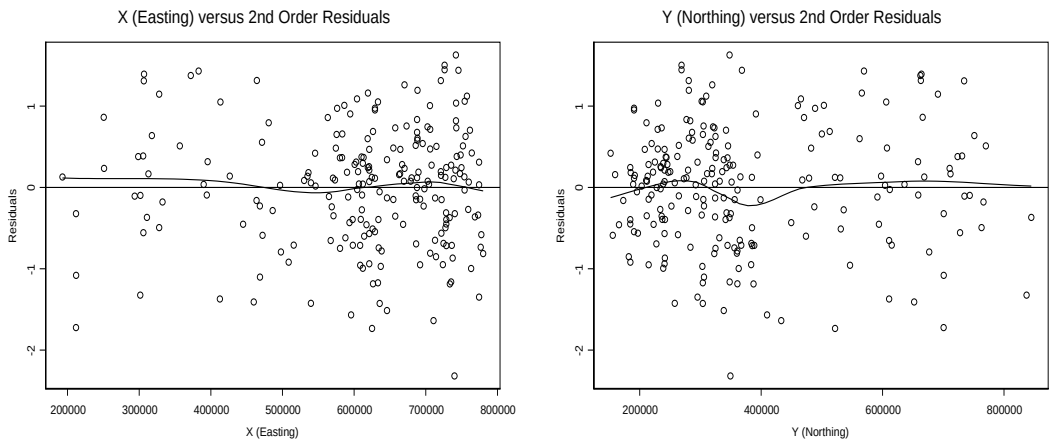
Since a trend in both x and y directions is shown in Figure 10, an attempt will be made to model this trend and analyze the residuals via the variogram. First, the residuals from a linear model (model L) will be examined. A least squares model $log_e[z(s)] = \beta_0 + \beta_1 x + \beta_2 y$ is fit and residual plots versus both x and y are shown in Figure 11.

Figure 11. Linear Trend Model - Residual Plots



In Figure 11 the trend line of the model L residuals, which is a smoothing curve using a loess fitting method, indicates that a trend still exists in both x and y directions. A quadratic least squares model (model Q) $log_e[z(s)] = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 y + \beta_4 y^2 + \beta_5 xy$ is fit and the residual plots are examined in Figure 12.

Figure 12. Quadratic Trend Model - Residual Plots



As shown in Figure 12, it appears that model Q removes the trend better

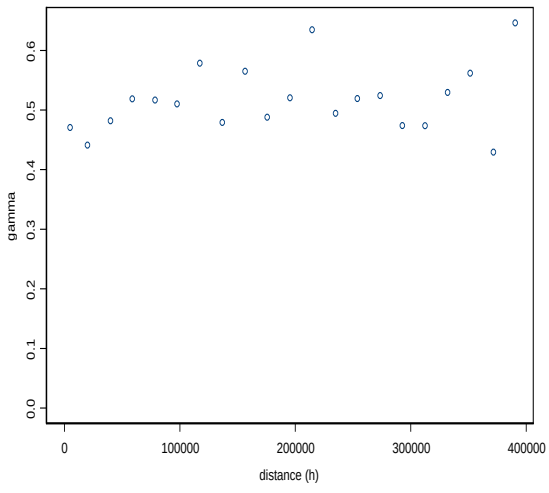
than model L. Model Q, the quadratic trend model, will be selected: $log_e[z(s)] = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 + \hat{\beta}_3 y + \hat{\beta}_4 y^2 + \hat{\beta}_5 xy + \hat{\epsilon}$.

Now that the trend has apparently been removed from the data, a variogram of the residuals $\epsilon(s)$ will be examined via the empirical variogram.

Since the residuals of the trend model are being used, the variogram function (2.10): $\hat{\gamma}(h) = \frac{1}{2[\#N(h)]} \sum(\epsilon_i - \epsilon_j)^2$ is utilized. The residuals ϵ_i and ϵ_j represent the residuals from the quadratic trend model.

Figure 13 is a variogram of the residuals on Model Q. This variogram of the $\epsilon(s)$ appears to be a non-increasing function, indicating that any trend or factors contributing to a nonstationary process have been removed. After a check

Figure 13. Variogram of Residuals



for outliers (Section 4.4) and a check for anisotropy (Section 4.5), the variogram parameters will be estimated.

4.4 The Variogram Cloud

To continue with the analysis of the arsenic variogram, a variogram cloud will be used to analyze the residuals from the second order trend model to identify any outliers.

Figure 14. Variogram Cloud of Trend Residuals (ϵ)

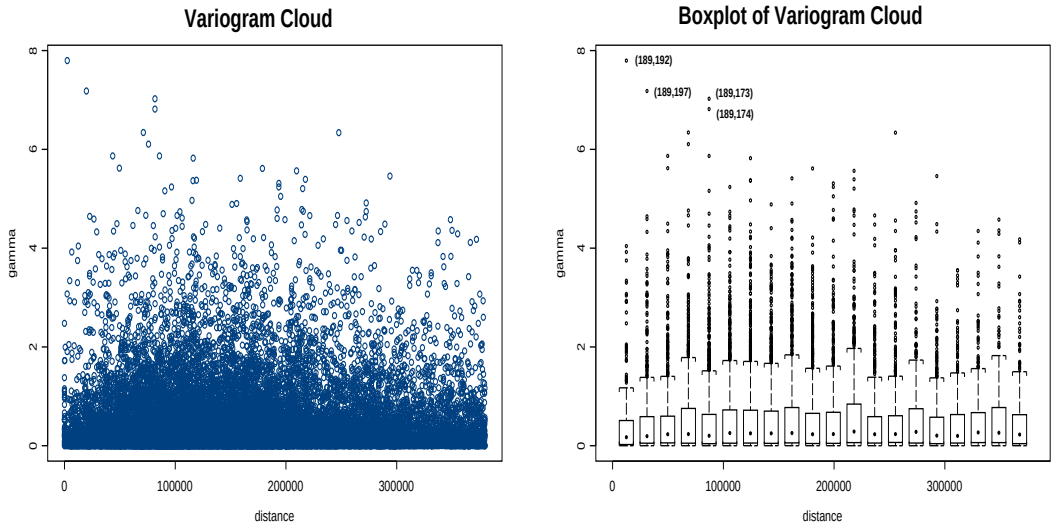


Figure 14 shows a variogram cloud on the left and a boxplot of this variogram cloud on the right of the residuals from the second order trend model (model Q). This variogram cloud reveals the individual $(\epsilon_i - \epsilon_j)^2$ that are used for each distance bin in the variogram shown on the left plot of Figure 13. The variogram cloud indicates several potential outliers, particularly at the shorter distances. Some of these potential outliers have been labelled on the boxplot shown on the right.

The points that are labelled on the boxplot all have the same data point

in common: data point #189. A review of the arsenic data set indicates data point #189 is a very low arsenic concentration level (z=.42 ppm) in an area of very high arsenic concentration levels. (the four points identified with point #189 range from 17.8-23.8 ppm). A discussion with MDEQ concluded that there is not sufficient evidence to remove point #189 from the model. However, this outlier may indicate that factors in addition to spatial location may be influencing arsenic concentration levels in Michigan.

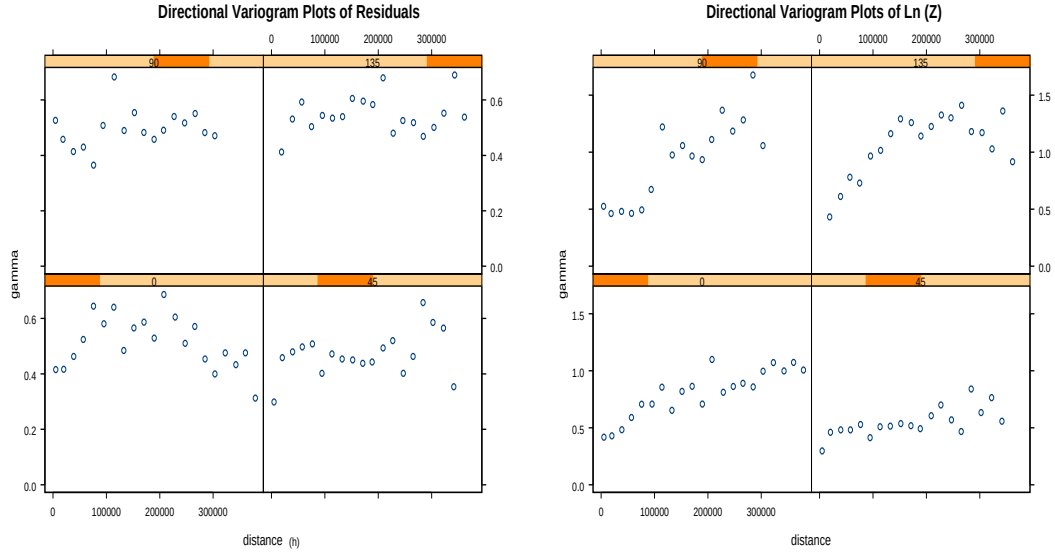
4.5 Anisotropy

To detect the presence of anisotropy, it is necessary to create directional variograms. These directional variograms are subsets of the variograms shown in Figure 13. For comparison purposes, the plots on the left of Figure 15 are directional variograms of ϵ, the residuals of the "detrended" data. The plots on the right are directional variograms of the original data $log_e[z(s)]$. Both variogram plots in Figure 13 are partitioned and shown in Figure 15 by viewing the data in the north/south (0 degree), northeast/southwest (45 degree), east/west (90 degree). Each of these directions will include a tolerance of plus or minus 22.5 degrees, so that all data in similar directions are examined.

Examining the left plot of Figure 15, it appears that all four directions show similar nugget ($\gamma(h) \approx .4$) and sills ($\gamma(h) \approx .5$). There does not appear to be a significant difference between the range (approximately 100000 meters) of

these four directional variograms.

Figure 15. Directional Variograms



The right plot in Figure 15 suggests that there are distinct differences among the four directions. The 0 degree (north/south) and 90 degree (east/west) variograms are generally increasing functions, which could be caused by the presence of trend and/or anisotropy, or some other form of nonstationarity. The 45 degree (northeast/southwest) and 135 degree (northwest/southeast) variograms appear to be non-increasing, indicating that the process is stationary along these axises.

In addition, compare the right plot (original data) with the left plot (detrended data) of Figure 15. The 0 and 90 degree variograms on the left plot are no longer increasing functions indicating that the removal trend in these directions was the source of non-stationarity.

4.6 Estimation of Variogram Parameters

As discussed in section 2.1.7 of this dissertation, weighted least squares may be used to estimate the three variogram parameters (nugget, range, and sill) However, prior to this estimation the variogram model must be selected. To assist in this process, three commonly used variogram models (exponential, spherical, and gaussian) will be evaluated. Each model will have its respective parameters estimated with weighted least squares and evaluated. The three goodness-of-fit tests discussed in section 2.1.7 will be used to evaluate each model. The results of these tests, along with the parameters chosen by weighted least squares, are shown in Table 8.

It should be noted that all three goodness-of-fit (SSR, Cressie, and Clark & Harper) tests discussed in Section 2.1.7 are in agreement with the selection of the gaussian variogram model (nugget=.4417, range=48762.45, sill=.08739). This model is shown in Figure 16.

Now that the covariance structure of the arsenic data has been defined, block kriging will be performed to provide an estimate of mean and variance of estimation error for the blocks overlayed on the state of Michigan.

4.7 Block Kriging of Arsenic Data

Using the spatial model outlined in the previous section, universal block kriging with quadratic spatial trend was performed using the log_e transformed

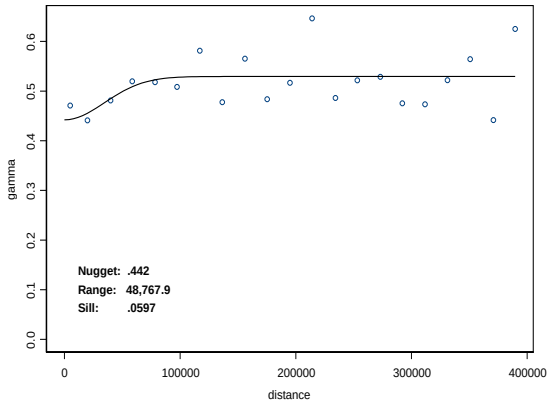
Table 8. Comparison of Variogram Models

		Goodness-of-Fit Tests		
Variogram Model	Parameters	SSR	Cressie	Clark & Harper
Exponential	nugget=.4946 range=189,850.5 sill=.0520	.03897	.0259	.00424
Spherical	nugget=.474 range=189,850.6 sill=.0597	.0234	.0259	.004159
Gaussian	nugget=.442 range=48,767.9 sill=.0874	.00464	.0236	.00407

arsenic concentrations. For reference, Figure 17 depicts the GAI (Michigan) with corresponding block numbers.

The S-Plus© kriging program is capable of performing the block kriging portion of SSM, but the program provides $\hat{z}_B$ and $\sqrt{\hat{\sigma}_B^2}$ only, not estimates of the actual kriging weights. In order to determine the dissimilarity coefficient d_{ij}, the kriging weights must be known. Hence, in order to perform SSM, an alternative kriging program was written using the S-Plus© programming language to obtain the estimated kriging weights for the universal block kriging with quadratic spatial

Figure 16. Chosen Gaussian Variogram Model

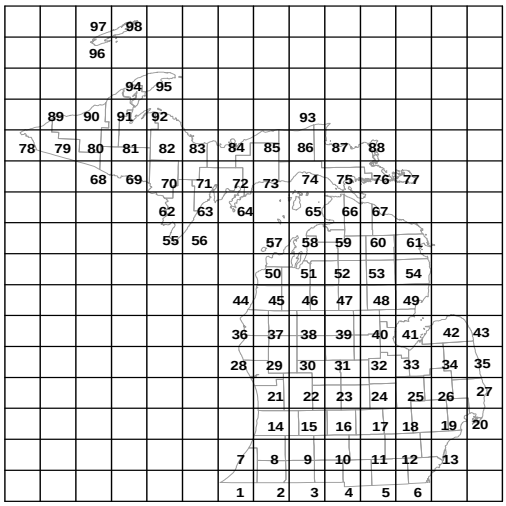


trend model discussed in Section (2.1.10).

This program, known as the dissertation program, creates the C_{BQ} and D_{BQ} matrices, used in (2.35), by inputting the results from the Gaussian variogram function discussed in Section 4.6. Using (2.35), the inverse of C_{BQ} is determined and multiplied by D_{BQ} to obtain the block kriging weights. These weights are then used to complete the SSM method. Except for the infill sampling algorithms (*MinMean, MinMed, MinMax,* and *MinVar*) previously discussed, SSM uses the dissertation program throughout the SSM process.

For comparison purposes, Figure 18 is a scatter plot comparing the dissertation program estimates of block mean $\hat{z}_B$ versus the S-Plus© kriging program estimates of $\hat{z}_B$. The straight line, used for reference purposes, has a slope of 1 and an intercept of 0, which would indicate an exact fit between both programs' estimates. It can be seen from Figure 18 that there appears to be a linear rela-

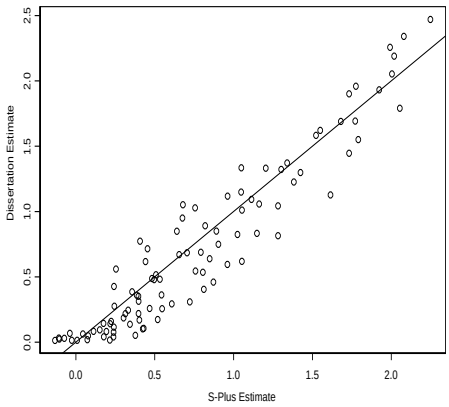
Figure 17. Block Map



tionship between the two kriging programs. The correlation coefficient between the dissertation program block estimate of mean and the S-Plus© program block estimate of mean was calculated to be .9502. The slope of the least squares regression line fit to Figure 18 is 1.0036, indicating that the dissertation program slightly overestimates z_B when compared to S-Plus©. However, the 95% confidence for this slope estimate is (.938, 1.07). Hence, a slope of 1 is significant and their appears to be a strong similarity between the estimates provided by the dissertation and S-Plus© programs.

Figure 19 displays comparison boxplots for the S-Plus© estimates of $\sqrt{\hat{\sigma}_B^2}$ versus the dissertation program estimates of $\sqrt{\hat{\sigma}_B^2}$. It can be seen from Figure 19 that there is a tendency for the dissertation program to calculate $\sqrt{\hat{\sigma}_B^2}$ at a higher level than the S-Plus© program. It is suspected that the S-Plus© program uses a higher degree of precision when estimating C_{UQ}^{-1}, the inverse of the covariance

Figure 18. Comparison of $\hat{z}_B$ Between Dissertation and S-Plus© Kriging Programs

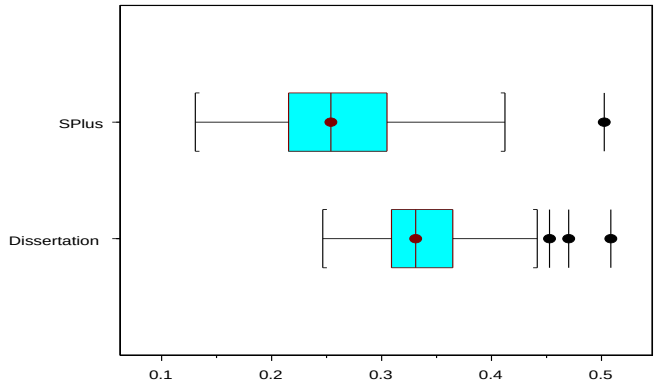


matrix between the existing data. However, the correlation coefficient between the estimates provided by the two models is .9453, indicating that the programs are highly correlated in their respective calculation of $\sqrt{\hat{\sigma}_B^2}$.

4.8 Cross Validation of Arsenic Data

Recall in Section 2.1.11 that the estimation error $e_{-j} = (z_j - \hat{z}_{-j})$ and the standardized estimation error $E_{-j} = e_{-j}/\hat{\sigma}_{-j}$, were defined, where $\hat{\sigma}_{-j}$ is the estimate of error variance. In addition to examining the mean and variance of the standardized estimation errors, the following plots will be examined to verify the spatial model has not been misspecified: (1) scatterplot of e_{-j} versus spatial location (x), (2) scatterplot e_{-j} versus spatial location (y), (3) histogram of the standardized estimation errors E_{-j}, (4) q-q Plot of E_{-j}, (5) scatterplot

Figure 19. Comparison Boxplots of $\sqrt{\hat{\sigma}_B^2}$

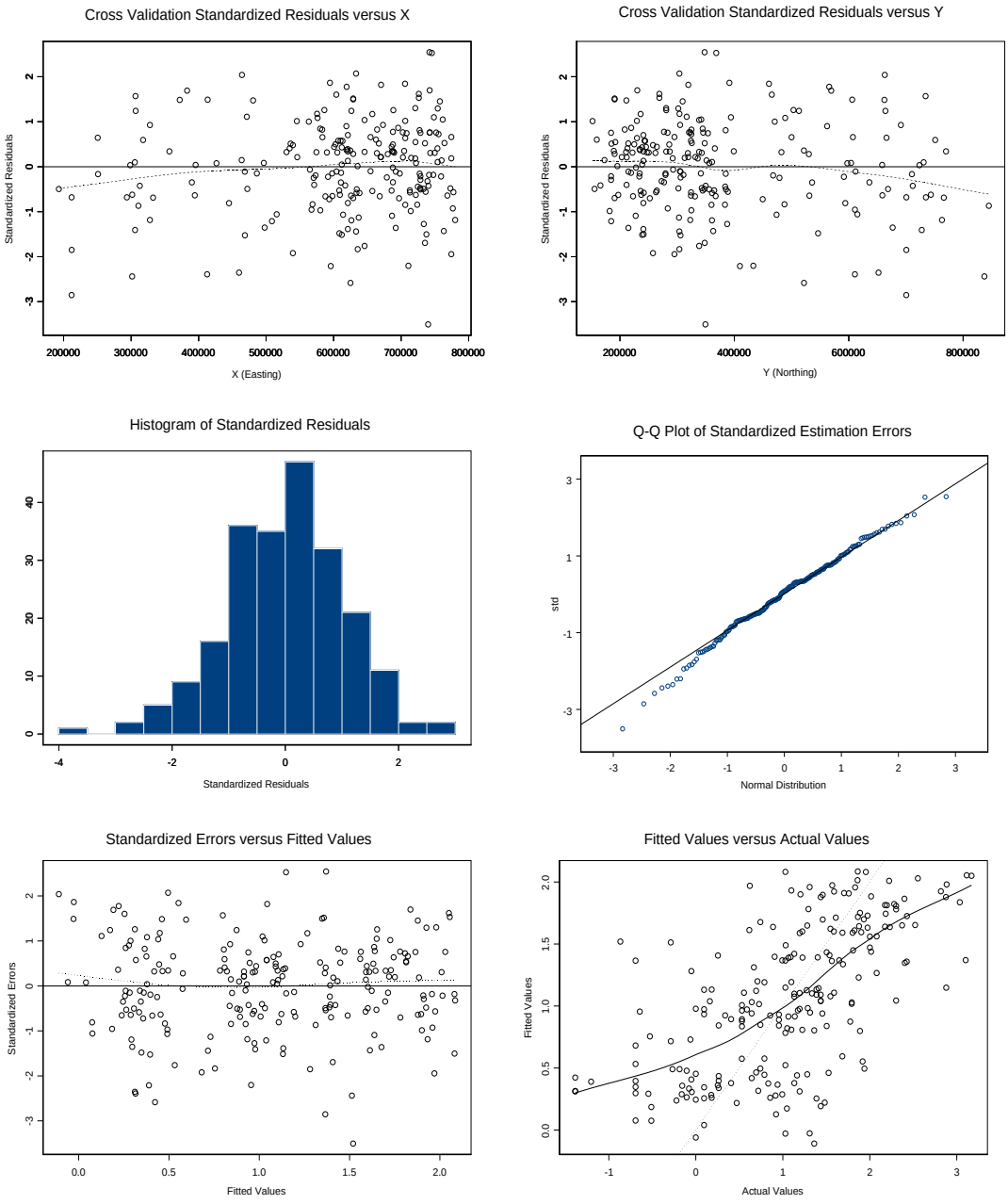


of standardized estimation errors (e_{-j}) versus the fitted values ($\hat{z}_{-j}$), and (6) scatterplot of fitted values ($\hat{z}_{-j}$) versus the actual observed values (z_j).

Figure 20 contains these six plots that will be used to validate the spatial model for the arsenic data. The dotted lines on these plots is a smoothing curve using localized regression (loess) to detect trends or patterns in the plots. Overall, these plots indicate that the spatial model has not been grossly misspecified. The standardized errors versus spatial locations show no evidence of estimation bias based upon location. The histogram of the standardized errors indicate a symmetric distribution that appears normal. In fact, the mean (-.00882) and variance of (1.07) of the standardized errors compare favorably to the expected mean and variance of 0 and 1 respectively. The plots of fitted values versus actual values again support the belief that the model is not biased. Finally, the plot fitted values versus standardized errors illustrate the smoothing characteristics of

kriging. There is a tendency for the low COC values to be overestimated while the high COC values have a tendency to be underestimated.

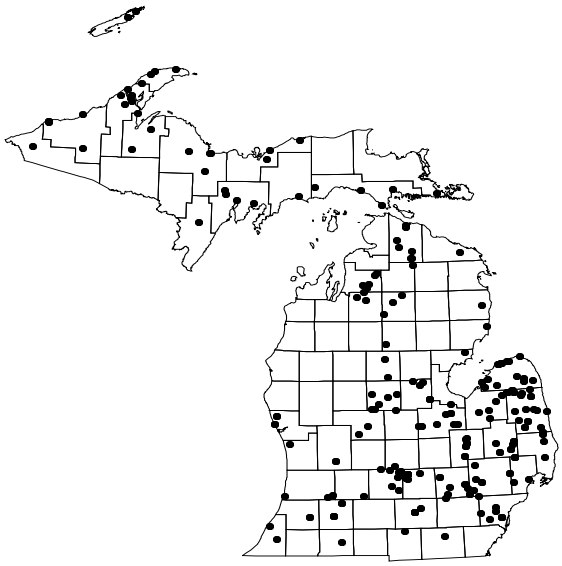
Figure 20. Cross-Validation Plots



4.9 Infill Sampling of Arsenic Data

As with most observational studies, the arsenic data does not uniformly represent the state of Michigan. As shown in Figure 21, the original 219 arsenic data sampling locations are not equally distributed across the state of Michigan. A majority of the arsenic samples are located in the southeastern lower peninsula

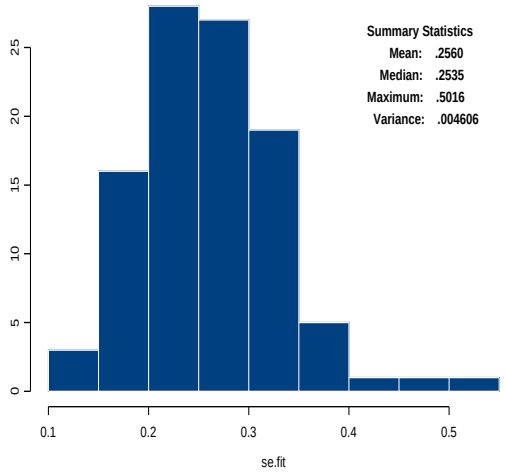
Figure 21. Arsenic Data Sample Locations



(expected since the DEQ's attention would be focused around the high arsenic levels), with sparse sampling in the northern lower peninsula, southern border, and upper peninsula areas. Because there are blocks in areas of both dense and sparse sampling, the standard error of the 98 block estimation errors $\sqrt{\hat{\sigma}_B^2}$ will not be equal. To demonstrate this fact, a histogram of the 98 $\sqrt{\hat{\sigma}_B^2}$, along with four summary statistics of interest, is shown in Figure 22.

It is important to note that in this section, a more efficient kriging algo-

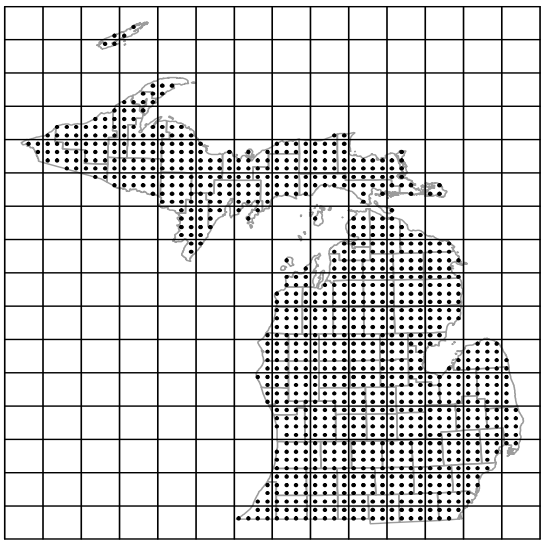
Figure 22. Histogram of $\sqrt{\hat{\sigma}_B^2}$



rithm than the one described in Section 2.1.10 was employed. Due to the long computation times required to perform these simulations, the S-Plus© kriging algorithm was used in the infill simulations. Hence, the $\sqrt{\hat{\sigma}_B^2}$ summary statistics given in Figure 22 will be different from the summary statistics resulting from the SSM procedure. Since this infill simulation in no way effects the results of SSM, it was decided to use the Fortran based S-Plus© algorithm to expedite the simulation. Regardless of the kriging algorithm used, the results here are relative and should remain consistent.

Each of the 98 Michigan blocks contains 16 evenly spaced points in a 4x4 grid. By eliminating those points outside the state boundary, there are $N = 1,062$ candidates available for infill selection and are shown in Figure 23.

Figure 23. Candidates for Infill Sampling



In Section 3.2, the procedure used to evaluate the four infill sampling algorithms (*MinMean, MinMed, MinMax,* and *MinVar*) was discussed. This procedure will now be briefly reviewed.

First, a predetermined number of infill sample points n is chosen. For the *MinMean* algorithm, the combination of n additional sample points, chosen from the candidates, that minimizes the mean of the $\sqrt{\hat{\sigma}_B^2}$ is determined. In addition, the other updated summary statistics of the $\sqrt{\hat{\sigma}_B^2}$ (median, maximum, variance) resulting from the *MinMean* algorithm are recorded. The n additional sample points resulting from the *MinMean* algorithm are then removed from the existing data set and returned to the remaining candidates. Then, the remaining three algorithms are each performed in similar fashion.

Once all four algorithms have been completed, each of the four $\sqrt{\hat{\sigma}_B^2}$ sum-

mary statistics will have four different values; one from each algorithm. For example, the mean of $\sqrt{\hat{\sigma}_B^2}$ will have a value representing *MinMean*, a value representing *MinMed*, a value representing *MinMax*, and a value representing *MinVar*. These four means will then be ranked against each other. The ranking will range from 1 (lowest) to 4 (highest). This ranking is done for each summary statistic. Thus, each algorithm has four ranks; one for each summary statistic. The ranks are totaled, and the algorithm with the lowest total rank is declared the optimal algorithm for the infill sample size n used.

In Table 9, results of $n = 10$ infill samples for the arsenic data are shown. As can be seen, *MinVar* algorithm has a rank total of 7 and is optimal for $n = 10$. The *MinMean* algorithm, with a rank of 9, was next. The *MinMax* algorithm had the third lowest total rank of 11. Finally, the *MinMed* algorithm, with a rank of 13, was the worst.

Table 9. $\sqrt{\hat{\sigma}_B^2}$ Statistics and Algorithm Ranking for n=10

Algorithm ↓	$\sqrt{\hat{\sigma}_B^2}$ Mean	Rank	$\sqrt{\hat{\sigma}_B^2}$ Med.	Rank	$\sqrt{\hat{\sigma}_B^2}$ Max.	Rank	$\sqrt{\hat{\sigma}_B^2}$ Var.	Rank	Rank Total
MinMean	.2190	1	.2262	3	.2624	3	.0009672	2	9
MinMed	.2203	4	.2160	1	.2695	4	.001141	4	13
MinMax	.2195	3	.2263	4	.2552	1	.0009997	3	11
MinVar	.2192	2	.2260	2	.2591	2	.0009547	1	7

In Table 10, results of $n = 20$ infill samples for the arsenic data are shown. As can be seen, *MinVar* algorithm now has a higher rank total of 8, but is still optimal for $n = 20$. This change in total rank was caused by ranking the performance for the median. In ranking the median, the *MinMean* algorithm outperformed the *MinMed* algorithm causing an exchange of rank. However, the performance of the *MinVar, MinMean, MinMax,* and *MinMean* algorithms remains unchanged.

Table 10. $\sqrt{\hat{\sigma}_B^2}$ Statistics and Algorithm Ranking for n=20

Algorithm ↓	$\sqrt{\hat{\sigma}_B^2}$ Mean	Rank	$\sqrt{\hat{\sigma}_B^2}$ Med.	Rank	$\sqrt{\hat{\sigma}_B^2}$ Max.	Rank	$\sqrt{\hat{\sigma}_B^2}$ Var.	Rank	Rank Total
MinMean	.2154	1	.2242	2	.2591	3	.0008337	3	9
MinMed	.2176	4	.2084	1	.2678	4	.001113	4	13
MinMax	.2160	3	.2249	4	.2479	1	.000815	2	10
MinVar	.2158	2	.2251	3	.2545	2	.000805	1	8

In Table 11, the results of $n = 30$ infill samples for the arsenic data are shown. The results are identical to those described for $n = 20$ infill samples. The performance of the algorithms (*MinVar, MinMean, MinMax,* and *MinMean* remains unchanged.

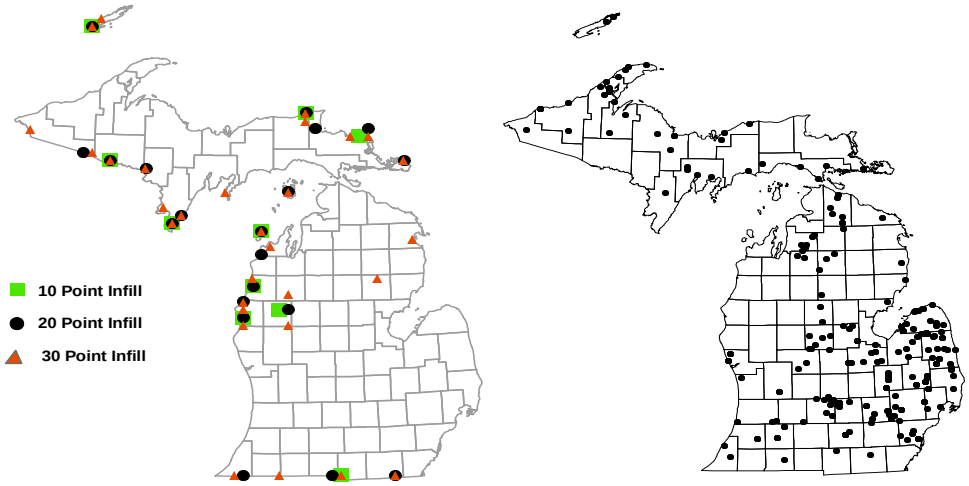
Table 11. $\sqrt{\hat{\sigma}_B^2}$ Statistics and Algorithm Ranking for n=30

Algorithm ↓	$\sqrt{\hat{\sigma}_B^2}$ Mean	Rank	$\sqrt{\hat{\sigma}_B^2}$ Med.	Rank	$\sqrt{\hat{\sigma}_B^2}$ Max.	Rank	$\sqrt{\hat{\sigma}_B^2}$ Var.	Rank	Rank Total
MinMean	.2123	1	.2192	2	.2542	3	.0007359	3	9
MinMed	.2149	4	.2027	1	.2667	4	.00110	4	13
MinMax	.2132	3	.2229	4	.2394	1	.0007144	2	10
MinVar	.2126	2	.2221	3	.2419	2	.0006908	1	8

In summary, the *MinVar* Algorithm has the lowest rank total for all three selected sample sizes and becomes the recommended in-fill sampling strategy algorithm. The *MinMean* algorithm is next, followed by the *MinMax* algorithm. The *MinMed* algorithm performed the worst in all three trials.

Figure 31 shows the *MinVar* infill sampling locations for all three sample sizes analyzed ($n = 10, 20, 30$) compared to the original sampling arrangement. Note that the three areas originally identified in this section as sparsely sampled have been primarily selected by the *MinVar* procedure. It appears that the northwestern lower peninsula, southern border, and upper peninsula areas have all been recommended for additional samples. Looking at the smallest infill sampling size ($n = 10$), there is a tendency to sample at the edges (state borders) of these sparsely sampled areas. As the infill sampling size increases, there is a tendency

Figure 24. Infill Sampling Points and Original Sampling Plan



to either move back into the interior of the state and sample at areas that may be under-represented but were not identified as sparsely sampled.

4.10 Dissimilarity Coefficient

In Section 3.4, the dissimilarity coefficient d_{ij} was defined as:

$$d_{ij} = \quad 0 \quad \text{for } i = j$$

$$d_{ij} = \left|(z_{B_i} - z_{B_j})\right| \Big/ \sqrt{\sigma^2_{B_i} + \sigma^2_{B_j} - 2\hat{\gamma}(z_{B_i}, z_{B_j})} \quad \text{for } i \neq j \tag{4.1}$$

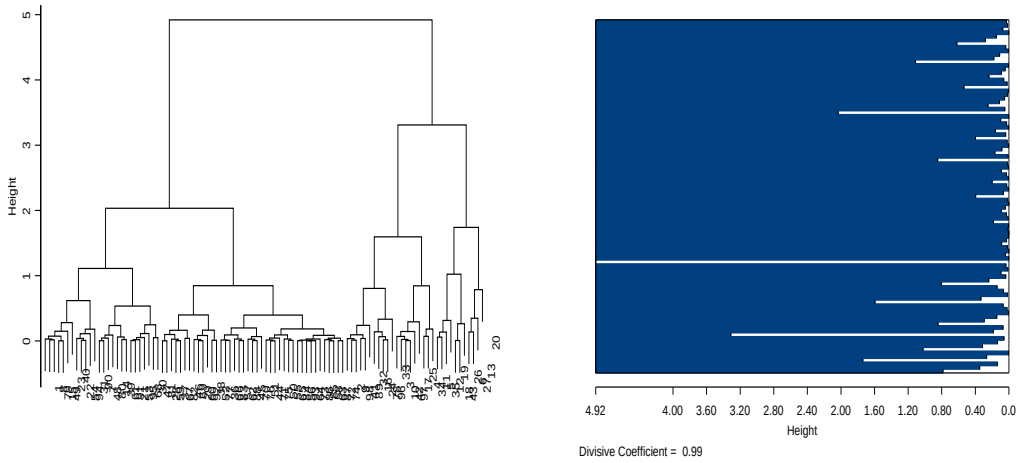
Figure 17 shows Michigan divided into 98 blocks. Let $[D]$ be defined as a 98 x 98 symmetric matrix containing all d_{ij}, in which the diagonal elements of $[D]$=0 and the off diagonal elements are the d_{ij}. The off diagonal elements represent the pairwise DCs between all 98 blocks containing Michigan land area. A program was written in S-Plus© to calculate $[D]$ using the results obtain from the universal

block kriging procedure performed on the arsenic data in Section 4.7.

4.11 Cluster Analysis

Using S-Plus©, a tree diagram and banner plot from performing the *Diana* clustering algorithm on $[D]$ are shown in Figure 25.

Figure 25. Cluster Analysis Results Using *Diana*



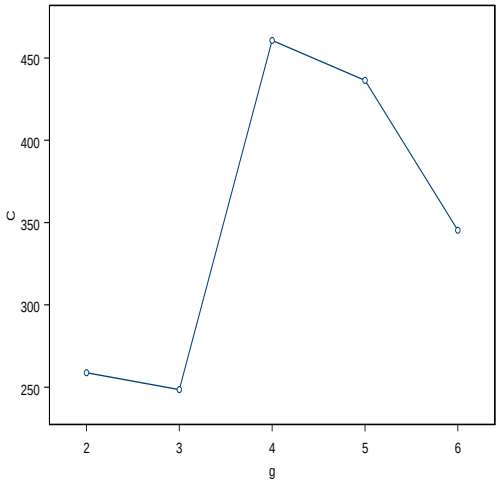
Recall Section 3.5 in which a higher divisive coefficient indicates better clustering structure. Looking at Figure 25, the results of the cluster analysis show a divisive coefficient of 0.98 out of a maximum of 1.0, which indicates $[D]$ has a very strong clustering structure. The tree diagram on the left indicates the sequence of splits that create anywhere from 1 to 98 clusters.

To assist in selecting the "correct" number of a cluster contained in the arsenic data, recall the criterion (3.10) proposed in Section 3.5 by Calinski and Harabasz [1974]:

$$C = \Big(B/(g-1)\Big) \Big/ \Big(W/(b-g)\Big). \tag{4.2}$$

Figure 26 shows the value of C for a range of two through five clusters. Recall

Figure 26. Number of Clusters



Section 3.5, in which it was suggested that a value of C increasing monotonically with g suggests no cluster structure, whereas C decreasing monotonically with g suggests a hierarchical structure. When C rises to a maximum at g this suggests the presence of g clusters. Looking at Figure 26, it appears that four clusters best describe the arsenic data since the Calinski and Harabasz statistic reaches its maximum at this point.

In addition, the between and within cluster sum of squares for two through five clusters are shown in Table 12.

Table 12. Sum of Squares for Cluster Configurations

Number of Clusters	2	3	4	5
Sum of Squares				
Between Cluster	29.28	33.7	37.58	38.11
Within Cluster	10.86	6.44	2.56	2.03
Total	40.14	40.14	40.14	40.14

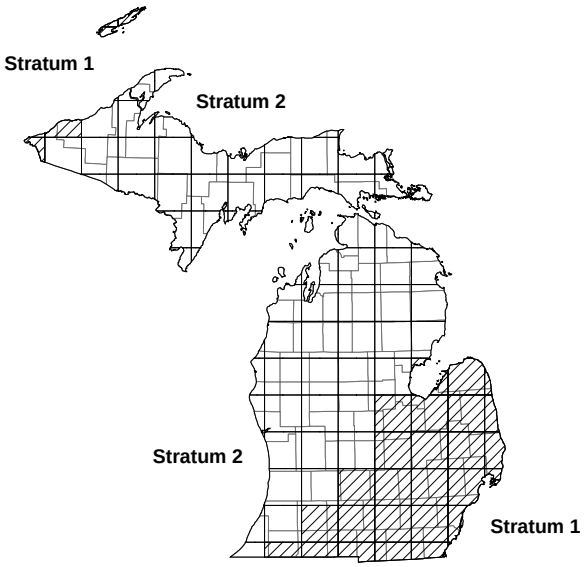
With reference to Figure 17, Table 13 gives the block membership for each of the two clusters.

Table 13. Two Cluster Membership

Cluster	Blocks Contained in Cluster
1	2,3,4,5,6,9,10,11,12,13,16,17,18,19,20,24 25,26,27,32,33,34,35,41,42,43,78,89,96,97,98
2	1,7,8,14,15,21,22,23,28,29,30,31,36,37,38,39,40,44 45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62 63,64,65,66,67,68,69,70,71,72,73,74,75,76,77,79,80 81,82,83,84,85,86,87,88,90,91,92,93,94,95

The spatial boundaries of each cluster are given in Figure 27

Figure 27. Spatial Boundaries of Two Strata



A review of Figure 27 shows that both the southeastern lower peninsula and northwest corner of the upper peninsula are in a cluster of higher arsenic concentration.

With reference to Figure 17, Table 14 gives the block membership for each of three clusters. Using three clusters to represent the arsenic data, the spatial boundaries for each cluster are given in Figure 28.

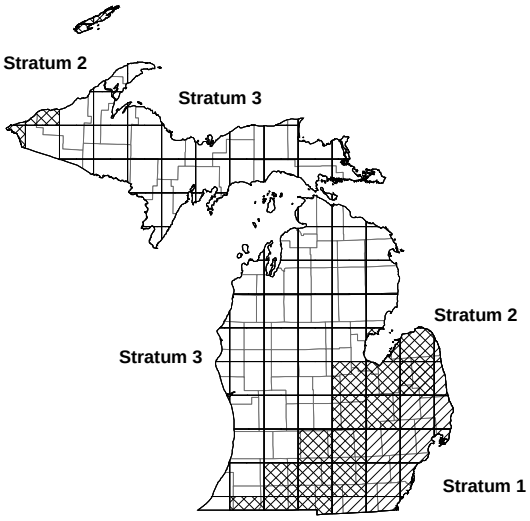
When three clusters are used instead of two, Figure 28 shows that the high arsenic concentration area in the southeastern part of the lower peninsula is split into another cluster. This cluster is the intermediate arsenic level; lower than the remaining southeastern lower peninsula and northwest upper peninsula, but higher than the remainder of the state.

With reference to Figure 17, Table 15 gives the block membership for each of four clusters. Using four clusters to represent the arsenic data, the spatial

Table 14. Three Cluster Membership

Cluster	Blocks Contained in Cluster
1	5,6,12,13,18,19,20,26,27,35,43
2	2,3,4,9,10,11,16,17,24,25,32 33,34,41,42,78,89,96,97,98
3	1,7,8,14,15,21,22,23,28,29,30,31,36,37,38,39,40,44 45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62 63,64,65,66,67,68,69,70,71,72,73,74,75,76,77,79,80 81,82,83,84,85,86,87,88,90,91,92,93,94,95

Figure 28. Spatial Boundaries of Three Strata

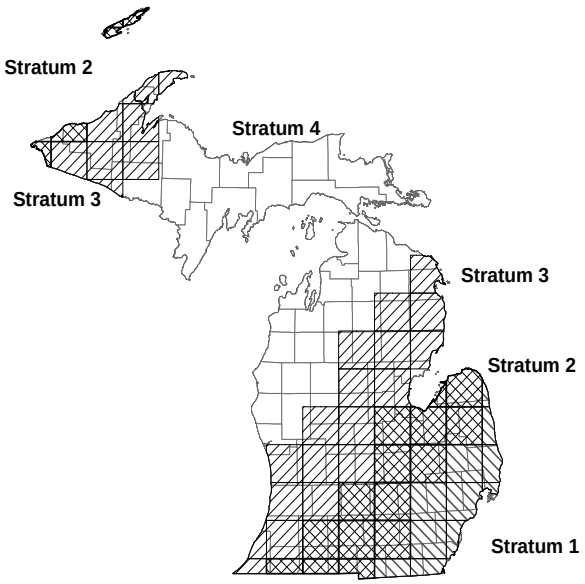


boundaries of each cluster are given in Figure 29. When adding the fourth cluster, the cluster with the lowest arsenic concentration is split into two clusters. The other two clusters remain unchanged.

Table 15. Four Cluster Membership

Cluster	Blocks Contained in Cluster
1	5,6,12,13,18,19,20,26,27,35,43
2	2,3,4,9,10,11,16,17,24,25,32,33,34,41,42,78,89,96,97,98
3	1,7,8,14,15,21,22,23,30,31,39,40,47,48,49,53,54,61,68,79,80,81,90,91,94,95
4	28,29,36,37,38,44,45,46,50,51,52,55,56,57,58,59,60,62,63,64,65,66,67,69,70 71,72,73,74,75,76,77,82,83,84,85,86,87,88,92,93

Figure 29. Spatial Boundaries of Four Strata



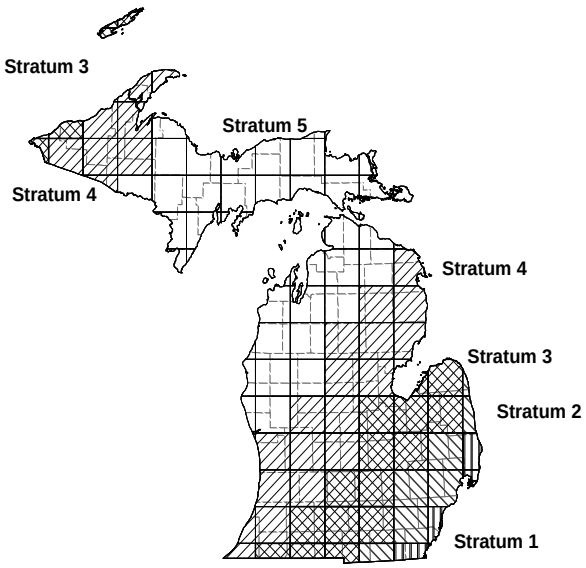
With reference to Figure 17, Table 16 gives the block membership for each of five clusters.

Using five clusters to represent the arsenic data, the spatial boundaries of each cluster are given in Figure 30. When adding the fifth cluster, the cluster

Table 16. Five Cluster Membership

Cluster	Blocks Contained in Cluster
1	13,20,26,27
2	5,6,12,18,19,35,43
3	2,3,4,9,10,11,16,17,24,25,32,33,34,41,42,78,89,96,97,98
4	1,7,8,14,15,21,22,23,30,31,39,40,47,48,49,53,54,61,68,79,80,81,90,91,94,95
5	28,29,36,37,38,44,45,46,50,51,52,55,56,57,58,59,60,62,63,64,65,66,67,69,70 71,72,73,74,75,76,77,82,83,8485,85,87,88,92,93

Figure 30. Spatial Boundaries of Five Strata



with the highest arsenic concentration is split into two clusters. The remaining three clusters remain unchanged.

4.12 Strata - Spatially Weighted Clusters

Using Arc/Info, a map of a particular spatial feature can be thought of as a layer of data describing a geographic area. In Arc/Info, such a layer is called a coverage. By combining the grid of blocks with the map of Michigan, a coverage is created where the amount of Michigan land area contained each block may be determined. When a strata membership for each block is determined, it is possible to determine the total land area for each strata and the contribution from each member block. Utilizing Arc/Info, a coverage of each of the four stratum was defined. Within each polygon coverage, the total land area of each stratum was calculated. The percent land contribution (v_i) of each block (B_i) belonging to a particular stratum was then determined. An estimate of stratum mean and variance was then determined by weighting each block by this percent land contribution and summing the terms. As previously discussed, since $\sum v_i = 1$, the estimates of stratum mean and variance will be unbiased.

4.13 Default Standards of Strata

Recall Section 4.3, in which it was concluded that the arsenic data most closely followed a lognormal distribution. Using equation 3.18 from Section 3.7.2,

default standards for each strata, which are represent here as the 95^{th} percentile of a lognormally distributed random variable is presented here. Each of the strata configurations (Two Strata, Three Strata, Four Strata, and Five Strata) are examined. In addition, for comparison purposes, the non-parametric estimate from equation 3.19 and the Michigan Background Soil Survey (MBSS) default standard (mean plus one standard deviation) described in Section 1.3.1 is included for each strata.

Table 17. Default Standards (in ppm) For Two Arsenic Strata

			Default Standard		
Stratum Number	Number of Blocks	Sample Size	SSM Lognormal	SSM Non-Parametric	MBSS
1	31	133	5.14	17.75	10.33
2	67	86	1.72	4.67	3.43

Table 18. Default Standards (in ppm) For Three Arsenic Strata

			Default Standard		
Stratum Number	Number of Blocks	Sample Size	SSM Lognormal	SSM Non-Parametric	MBSS
1	11	35	9.05	22.56	13.57
2	20	98	4.21	12.65	8.82
3	67	86	1.72	4.67	3.43

Table 19. Default Standards (in ppm) For Four Arsenic Strata

			Default Standard		
Stratum Number	Number of Blocks	Sample Size	SSM Lognormal	SSM Non-Parametric	MBSS
1	11	35	9.05	22.56	13.57
2	20	98	4.21	12.65	8.82
3	26	47	1.94	5.69	3.75
4	41	39	1.22	4.40	3.00

Table 20. Default Standards (in ppm) For Five Arsenic Strata

			Default Standard		
Stratum Number	Number of Blocks	Sample Size	SSM Lognormal	SSM Non-Parametric	MBSS
1	4	17	14.32	23.80	15.95
2	7	18	8.24	20.8	10.86
3	20	98	4.21	12.65	8.82
4	26	47	1.94	5.69	3.75
5	41	39	1.22	4.40	3.00

CHAPTER V

SUMMARY OF SSM, CONCLUSIONS, AND ADDITIONAL RESEARCH

5.1 Summary of SSM

This dissertation introduces the concept of Spatial Strata Modelling (SSM). SSM eliminates reliance on a single default background standard, derived from a random sample, and instead promotes the investigation of multiple default standards for naturally occurring COCs. Each of these multiple default standards exist within the spatial boundaries of a non-overlapping and unique stratum. Each stratum is then defined by its spatial boundary and an estimate of an upper limit of COC concentration level (default standard). SSM does not require a random sample and utilizes the fact that the data may be spatially correlated.

Using the spatial locations of the data, a grid of square blocks is sized, overlayed, and centered on the geographic area of interest (GAI). Employing geostatistic theory, the empirical variogram based on the observational data is estimated. The variogram function and parameters (range, nugget, and sill) are estimated using weighted non-linear least squares. Using the estimated variogram model, block kriging estimates of mean ($\hat{z}_B$), variance of the estimation error ($\hat{\sigma}^2_B$), and covariance of the estimation error $[\hat{\gamma}(B_i, B_j)]$ are determined. Using these estimates, SSM defines a dissimilarity coefficient d_{ij} between all pairwise blocks B_i

and B_j within the GAI in which:

$$d_{ij} = \quad 0 \quad \text{for } i = j$$

$$d_{ij} = \left|(z_{B_i} - z_{B_j})\right| \Big/ \sqrt{\sigma^2_{B_i} + \sigma^2_{B_j} - 2\hat{\gamma}(z_{B_i}, z_{B_j})} \quad \text{for } i \neq j$$

An S-Plus© clustering analysis algorithm (in this dissertation, *Diana*) is performed on the d_{ij} to determine the number and block membership of clusters containing similar COC concentration levels. In order to estimate the correct number of cluster(s), a criterion proposed by Calinski and Harabasz [1974] is utilized. The proposed function seeks to compare the between cluster sum of squares versus the within cluster sum of squares. The cluster size that maximizes this function is chosen as the estimated cluster size. The blocks contained in each of the clusters define the spatial boundaries for that particular cluster.

Spatial weights v_i, which are based on the percent of total cluster land area contained in a member block, are assigned using Arc/Info 8.1©, a GIS software system. Within each cluster, the estimate of stratum mean (Definition 3.6.4) ($\hat{z}_S = \sum v_i \hat{z}_{B_i}$) and variance of the stratum estimation error (3.13)

$$\text{Var}(\hat{z}_S) = \sum_{i=1}^{b}\sum_{j=1}^{b} v_i v_j \text{Cov}(z_{B_i}, z_{B_j}) - 2\sum_{i=1}^{b}\sum_{j=1}^{b} v_i v_j \text{Cov}(z_{B_i}, \hat{z}_{B_j}) + \sum_{i=1}^{b}\sum_{j=1}^{b} v_i v_j \text{Cov}(\hat{z}_{B_i}, \hat{z}_{B_j})$$

are determined.

For comparison purposes, maps describing the spatial boundaries of the strata are shown in Figures 27, 28, 29, and 30. Tables 13, 14, 15,and 16 provide the mean, standard deviaiton, $\widehat{UP}_s$, and $\widehat{UCL}_s$ for GAI with two, three, four, and

five strata for the arsenic data.

In Tables 13, 14, 15,and 16, estimates of the upper 95^{th} percentile (log-normally distributed) and 95% upper prediction limit (normally distributed, and non-parametric) are provided.

Four in-fill sampling strategies were also discussed and evaluated. The algorithms, *MinMean, MinMed, MinMax*, and *MinVar* were analyzed to investigate the existence of a universally optimal algorithm.

5.2 Conclusions

Recall the questions posed in Section 1.5 of this dissertation:

1. How many unique spatial subsets are there with statistically similar COC concentrations?

2 What are the spatial boundaries for each spatial subsets?

3 What is the estimate of default standard within each unique spatial subset?

The results of performing SSM on the arsenic data in Michigan for two, three, four, and five strata are given in Figures (27), (28), (29), and (30) respectively.

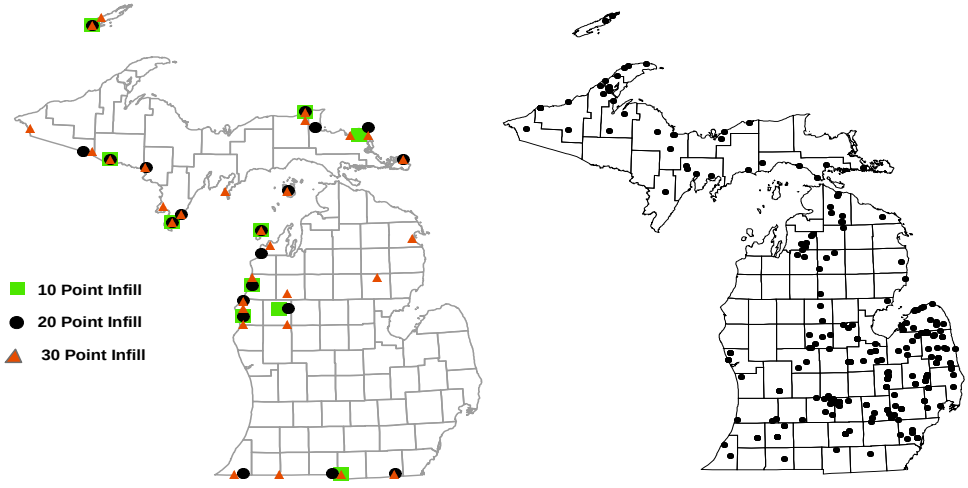
The estimates of the 95^{th} upper percentile, both for lognormal and non-parametric distributions, for each of the various stratum configurations are provided in Tables (17), (18), (19), and (20) respectively. In addition, the estimation

technique used in the Michigan Background Soil Survey (MMSS) ($\overline{x}+s$) is included on each stratum for comparison purposes.

To improve the accuracy of SSM, the development of an infill sampling strategy was discussed. Infill sampling is choosing the optimal combination from a pool of infill candidates to most effectively minimize the estimates of the variance of the ($\sqrt{\hat{\sigma}_B^2}$).

It was determined that minimizing the variance of the $\sqrt{\hat{\sigma}_B^2}$ (*MinVar* algorithm) appears to be the optimal choice for determining the location of the additional infill samples. A map showing the location of the recommended infill sample sizes for $n = 10, n = 20,$ and $n = 30$ using the *MinVar* algorithm has been included in Figure 31 and is repeated here.

Figure 31. Infill Sampling Points and Original Sampling Plan



The SSM methodology as described in this dissertation is unique and repeatable. In addition, SSM is invariant to geographic scale. Whether the GAI is

a small plot of land or an entire planet, this methodology may be utilized as long as the spatial characteristics of the random process are adequately represented by the sample data.

It is anticipated that both federal and state environmental regulatory agencies would consider the SSM method as an alternative to random sampling based procedures. If the spatial distribution of a COC can be accurately modelled and used to define multiple default standards, rather than a single default standard, excessive testing and/or unneeded remediation may be significantly reduced or eliminated. The potential to save both time and money should provide some incentive for fiscally challenged regulatory agencies to explore SSM as a viable alternative in the environmental regulatory process.

5.3 Additional Research

There are several research topics resulting from this dissertation that have the potential to enhance the SSM methodology. These topics are briefly discussed below:

1. Estimation of a UCL for the mean.

It should be noted that an estimate of the UCL for the kriging estimate of a lognormally distributed random variable was proposed by Clark & Harper (2000). Using Sichel's (1966) maximum likelihood estimate of the mean from the lognormal distribution, Clark states the upper confidence limit of

the mean (UCL_p) can be expressed as:

$$UCL_p = t \times \Psi_p \tag{5.1}$$

in which:

t= Sichel's (1966) t-estimator in which:

$t = \exp(\overline{y})\gamma_n(V)$

where $\overline{y}$ is the mean of the log transformed data and $\gamma_n(V)$ is the sample size bias correction factor. $\gamma_n(V)$ may be read from Sichel's tables (see Clark 1987) using sample size n and $V = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2$.

The bias correction factor, $\gamma_n(V)$, is the result of a Taylor Expansion Series for the maximum likelihood estimate of the mean of the lognormal distribution for various samples of size n and is shown by Clark (1987) as:

$$\gamma_n(V) = 1 + \sum_{r=1}^{\infty} \frac{(n-1)^r \;\; V^r}{2r!(n-1)(n+1)...(n+2r-3)} \tag{5.2}$$

where:

n is the sample size and $V = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2$.

In addition,

$$\Psi_p = \exp\left(\frac{1}{2}\sigma_k^2 + N_p\sigma_k\right) \tag{5.3}$$

where:

Ψ_p= Sichel's (1966) psi factor for the p^{th} percentile (see Clark 1987),

σ_k^2= variance of either the ordinary or universal kriging estimation error,

and

N_p= the value corresponding to an area of (1-p)/2 from the center of a standardized normal distribution N(0,1).

It should also be noted that the exact construction of a confidence limit for the mean of a lognormally distributed random variable was also developed by Land (1975). In the realm of environmental statistics, Land's Tables are used extensively and are referenced to in a wide variety of environmental regulatory guidance documents. Using Land's Tables, the $100(1-\alpha)\%$ upper confidence limit for the mean of a lognormally distributed random variable is:

$$UCL_\alpha = \exp\left(\overline{y} + \frac{s_y^2}{2} + \frac{s_y \times (H_{1-\alpha,n})}{\sqrt{n-1}}\right) \tag{5.4}$$

where:

$\overline{y}$ = the mean of the log transformed data.

$s_y = \sqrt{\frac{(\sum y_i - \overline{y})^2}{(n-1)}}$, the standard deviation of the log transformed data

$H_{1-\alpha,n}$ = the value from Land's Tables using s_y and n.

n = the sample size of the data.

Although some similarities between Sichel's and Land's equations are noted, present research deriving the UCL from kriging results has focused on Sichel's theories, but is not complete at this time. In either case, the use of Land's tables or Sichel's theories to express the UCL from kriging estimates requires

further research.

2. Sequential Testing of Blocks

The decision to use cluster analysis, rather than a parametric or nonparametric statistical testing, was not an arbitrary decision. The within and between block correlations that exist in the block kriging estimates $\left(\hat{z}_B, \hat{\sigma}^2_B, \mathrm{Cov}(B_i, B_j)\right)$ restrict the testing procedures to those designed for use on correlated data. Quimby (1986) has developed the correlated two-sample t-test from the theory of generalized linear models. Borgman (1988) discusses the two-sample geostatistical wilcoxon test for correlated samples using the null hypothesis: $H_o : Z = B_i - B_j$ has multivariate symmetry about zero.

where B_i and B_j are the blocks of interest.

These two-sample testing procedures, while useful, do not go far enough to allow for a determination of a stratum. It would be necessary to conduct sequential hypothesis testing of an unknown number of blocks such as $H_o : B_i = B_j = B_k = ..., B_n$. Without a multiple sequential testing procedure, it is not possible to build a stratum of more than two blocks without continuously redefining the blocks. In addition, one must consider a non-subjective procedure for choosing what order to test the blocks. These issues need to be researched before a sequential testing procedure may be developed.

3. Additional In-Fill Sampling Algorithms

With regards to in-fill sampling strategies, the *MinVar* procedure, minimizing the variance of the σ_B^2 was recommended. The *MinVar* had the best overall ranking of minimizing the mean, median, maximum, and variance of the $\hat{\sigma}_B^2$. Additional statistics, such as the median of the pairwise block kriging variance averages (Hodges-Lehmann estimator), trimmed and Winsorized means, qth quantiles, etc. should be investigated and compared to the *MinVar* algorithm.

4. Polygon Kriging

In this dissertation, a block kriging procedure was employed. By definition, block kriging provides estimates over a square area of interest, with block dimensions fixed in all directions. However, it is also possible to perform polygon kriging. Polygon kriging is kriging over an irregular shape. Such a procedure may have useful applications for irregularly shaped GAIs. The SSM procedure weighted the block kriging estimates by the percent of land area contained within the block. However, using polygon kriging, the estimation process assumes the polygon is 100% filled with land area and does not need to be weighted. Polygon kriging estimates are based on the geometry of the polygon and does not assume each block is of homogeneous size and shape.

However, the theoretical problems with an SSM procedure based upon poly-

gon kriging will occur after the estimates have been obtained. The dissimilarity coefficient, as presently defined by 3.9, assumes each block is of equal size and shape. In general, it has been shown by Clark & Harper (2000) that as the size of a geographic area increases, the variance of the block estimation error $\hat{\sigma}_B^2$ will decrease. At present, SSM compares blocks of uniform size and shape. Clearly, an irregularly shaped geographic area would not produce uniformly sized polygons, particularly at the edges of the GAI. The estimates of $\hat{\sigma}_B^2$ and $\hat{\gamma}(B_i, B_j)$, which are used in the calculation of the d_j (equation 3.9) may depend as much on the size and shape of the polygon as on the spatial model. Hence the components of the d_{ij} function must be weighted to reflect difference in polygon size and shape. This weighting would be somewhat analogous to performing a two sample t-test with different sample sizes. Clearly, there are some unexplored issues here and further research into polygon kriging is required.

References

Armstrong, M. and Delfiner, P. (1980), "Towards a more robust variogram: A case study on coal." Internal Note N-671, Centre de Geostatistique, Fontainebleu, France.

Aspie, D. and Barnes, R. (1990), "In-Fill Sampling Design and the Cost of Classification Errors", *Mathematical Geology* , v. 22, no. 8, pp.915-933.

ATSDR (Agency for Toxic Substances and Disease Registry), (2000), "Toxicological profile for arsenic (update), U.S. Department of Health and Human Services, Public Health Service, Atlanta, GA.

Borgman, L.E. (1988), "New Advances in Methodology for Statistical Tests Useful in Geostatistical Studies", *Mathematical Geology.*, Vol 20, No. 4, pp.383-403.

Burmaster D., and Thompson, K., (1997), "Estimating Exposure Point Concentrations for Surface Soils for Use in Deterministic and Probabilistic Risk Assessments". *Human and Ecological Risk Assessment.* Vol 3,No. 3, pp.363-384

Burrough, P.A. (1986), *Principles of Geographic Information Systems for Land Resource Assessment*, Clarendon Press, Oxford.

Calinski, T., and Harabasz, J. (1974), "A dendrite method for cluster analysis. I. robust covariance estimation." *Appl. Stat.*, 29, pp.231-237.

Chiles, J.P., and Delfiner, P. (1999), *Geostatistics: modeling spatial uncertainty*, John Wiley & Sons, Inc., New York, xii+695 pp.

Clark, I., (1987), "Turning the tables - an interactive approach to the traditional estimation of reserves." *J. S. Afr. Inst. Min. Metall*, vol. 87, no. 10. Oct. 1987. pp.293-306.

Clark,I., and Harper, W. (2000), *Practical Geostatistics 2000*, Ecosse North America Llc, Columbus Ohio, USA, 442 p.

Cleveland, W.S. (1979), "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, Vol. 74, pp.829-36.

Cressie, N. (1985), "Fitting variogram models by weighted least squares." *Journal of the International Association for Mathematical Geology*, Vol. 17, pp.563-586.

Cressie, N. (1989), "Geostatistics." *American Statistician*,44, pp.256-258.

Cressie, N. (1993), *Statistics for Spatial Data*, John Wiley & Sons, Inc. New York, xxii+900pp.

Cressie, N. and Hawkins, D.M. (1980), "Robust estimation of the variogram, I." *Journal of the International Association for Mathematical Geology*, Vol. 12, pp.115-125.

Crow, E.L. and Shimizu, K. (1988), *Lognormal Distributions: Theory and Applications*. Marcel-Dekker, New York.

Dowd, P.A. (1982), "Lognormal kriging - The general case." *Journal of the International Association for Mathematical Geology*, Vol. 14, pp.475-499.

Freedman, D., Diaconis, P.(1981), "On the maximum deviation between the histogram and the underlying density." *Z. Wahrsch. Verw. Gebiete*, Vol. 58, No. 2, pp.139-167.

Gandin, L.S. (1963), *"Ob ektivnyi anliz meteorilogicheskikh polei"*. Gidrometeologicheskoe Izdatelstovo, Lenigrad. Translation (1965): *Objective analysis of Meteorological Fields*. Israel Program for Scientific Translations.

Gao, H., Wang, J., and Zhao, P. (1996), "The Updated Kriging Variance and Optimal Sample Design", *Mathematical Geology*, Vol.23, No. 3, 295-312.

Gilbert, R. O., (1987), *Statistical Methods for Environmental Pollution Monitoring.* Van Nostrand Reinhold, New York, NY.

Helsel, D.R., and Hirsch, R.M. (1993), *Statistical Methods in Water Resources*, Elsevier, Amsterdam, The Netherlands.

Hogg, R. and Craig, J., (1995), *Introduction to Mathematical Statistics.* Prentice Hall, Upper Saddle River, NJ.

Hounslow, A.W. (1980), "Ground-water Geochemistry, Arsenic in Landfills", *Ground Water*, Vol. 18, No.4, pp331-333.

IDEM (2001), "Risk Integrated System of Closure (RISC), Technical Guide and Users Guide", Indiana Department of Environmental Management, Office of Land Quality, Indianapolis, IN.

Isaaks, E., and Srivastava, R. (1989), *Applied Geostatistics*, Oxford University Press: New York, xix+561 p.

Jacobs, L.W., Syter, J.K., and Keeney, D.R. (1970)," Arsenic Sorption by Soils", Soil Science Society of American Proceedings, Vol. 34, pp.750-754.

Jardine, N. and Sibson, R. (1971), *Mathematical Taxonomy*, Wiley, New York and London.

Johnson, S.C., (1967), "Heirarchical clustering schemes". *Psychometrika*, Vol. 32, pp.241-254.

Journel, A.G. and Huijbregts, C.J. (1978), *Mining Geostatistics.* Academic Press, London.

Jowett, G.H. (1955), "Sampling properties of local statistics in stationary stochastic series." *Biometrika*, Vol. 42, pp.160-169.

Kacewicz, M., (1991), "Solving the kriging problem by using the Gram-Schmidt orthogonalization", *Mathematical Geology*, Vol. 23, No. 1, pp.111-118.

Kaluzny, S.P., Vega, S.C., Cardoso, T.P., and Shelly, A.A., (1998), *S+ Spatial Stats: User's Manual for Windows© and UNIX©*, Springer-Verlag, New York.

Kaufman, L., and Rousseeuw, P. J. (1990), *Finding groups in data: an introduction to cluster analysis*, John Wiley & Sons, Inc., New York, xi+342 pp.

Land, C.E. (1975), "Tables of Confidence Limits for Linear Functions of the Normal Mean and Variance", *Selected Tables in Mathematical Statistics, Vol III*, American Mathematical Society, Providence, R.I. pp.385-419.

Matheron, G. (1962), "Traite do Geostatistique Appliquee, Tone I". *Memories du Bureau do Recheerches Geologiques et Minieres*, No. 14, Editions Technip, Paris.

Matheron, G. (1963), "Principles of Geostatistics", *Economic Geology*, Vol. 58, p. 1246-1266.

MDEQ (1993), MERA Operational Memorandum #15, Michigan Department of Environmental Quality.

MDEQ (1994), "Guidance Document: Verification of Soil Remediation", Environmental Response Division, Waste Management Division. State of Michigan Department of Environmental Quality, April 1994, Revision 1.

MDEQ (2002), "DEQ Sampling Strategies and Statistics Training Materials for Part 201 Cleanup Criteria", Michigan Department of Environmental Quality, Remediation and Redevelopment Division.

Meyers, J.C. (1997), *Geostatistical Error Management*, Van Nostrand Reinhold, New York.

Michigan (1994), "Natural Resources and Environmental Protection Act, Act 451 of 1994", State of Michigan Legislature, Lansing, MI, Part 111, Hazardous Waste Management.

Modjeska, J.S. and Rawlings, J.O. (1983), "Spatial correlation analysis of uniformity data", *Biometrics*, Vol. 39, pp.373-384.

NRC (National Research Council), (1999), *Arsenic in Drinking Water*, Washington, D.C., National Academy Press, 273 p.

OME (2001), "Soil Investigation and Human Health Risk Assessment for the Rodney Street Community: Port Colbourne", Ontario Ministry of the Environment, Standards Development Branch Report No. SDB-010-3511-2001.

Quimby, W. (1986), "Selected Topics in Spatial Statistical Analysis", Ph.D. thesis: Statistics Department, University of Wyoming, Laramie, Wyoming.

Rousseuw P.J. (1986), "A visual display for hierarchical classification." IN: E. Diday, Y. Escoufier, L. Lebart, J. Pages, Y. Schektman, R. Tomassone (Eds.), *Data Analysis and Informatics*, 4, North-Holland, Amsterdam, pp.743-748.

Seber, G. A. F. (1984), *Multivariate observations.* Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, xx+686 pp.

Shapiro, S.S., and R.S. Francia. (1972), "An Approximate Analysis of Variance Test for Normality". *Journal of the American Statistical Association* 67(337), pp.215-219.

Sibson, R. (1972), "Order invariance methods for data analysis (with discussion)". *J.R. Statistical Society B, 34.* 311-349.

Sichel, H.S. (1966), "The estimation of means and associated confidence limits for small samples from lognormal populations.", Symposium on Mathematical Statistics and Computer Applications in Ore Valuation. Johannesburg, The South African Institute of Mining and Metallurgy, pp.106-122.

Sneath, P. H. A. (1957), "The application of computers to taxonomy." *Journal Gen. Microbiology, 17*, pp.201-226.

Sokal, R.R., and Michener, C.D. (1958), "A statistical method for evaluating systematic relationships", *Univ. Kansas Sci. Bull*, vol. 38, pp.1409-1438.

Sokal, R. R., and Sneath, P. H. A. (1963), *Principles of Numerical Taxonomy*. Freeman: San Francisco.

Stephan, F. (1934), "Sampling errors and interpretations of social data ordered in time and space". In Proceedings of the American Statistical Journal, New Series No. 185A, F. A. Ross, ed. *Journal of the American Statistical Association, 29 Suppl.*, pp.165-166.

Stone, M. (1974), "Cross-validatory choice and assessment of statistical predictions." *Journal of the Royal Statistical society B*,Vol. 36, pp.111-133.

Struyf, A., Hubert, M., and Rousseeuw, P. (1997), "Integrating robust clustering techniques in S-PLUS". *Computational Statistics and Data Analysis*, Vol. 26, pp.17-37.

Switzer (1984), "Inference for spatial autocorrelation functions.", *Geostatistics for Natural Resources Characterization*, G. Verly, M. David, A.G. Journel, and A. Marechal, eds. Reidel, Dordrecht, Holland, Part 1, pp.127-140.

USEPA (1995), The Environmental Protection Agency Code of Federal Regulations (CFR) Title 40 Project, Subchapter I Part 261.

USGS (1999), United States Geological Survey's Mineral Resources Program Activities in the Upper Midwest, USGS Information Handout, April 1999.

Vann, J., and Guibel, D., (1998), "Beyond Ordinary Kriging: An Overview of Non-linear Estimation", *Symposium: Beyond Ordinary Kriging* Geostatistical Association of Australasia.

WMD (1991), "Michigan Background Soil Survey." Technical Support, Hazardous Waste Program, Waste Management Division, Michigan Department of Environmental Quality, Lansing, Michigan.

Whittle, P. (1963), "Stochastic processes in several dimensions", *Bulletin of the International Statistical Institute*, 40, Book 2, pp.974-994.

Zimmerman, D.L. and Zimmerman, M.B. (1991), "A comparison of spatial semivariogram estimators and corresponding ordinary kriging predictors." *Technometrics*, Vol. 33, pp.77-91.