The Estimation of the Variogram in Geostatistical Data with Outliers

MARCH 2015

Sujung Kim

Graduate School of Environmental and Life Science
(Doctor Course)
OKAYAMA UNIVERSITY

Contents

1. Introduction

Recently, researchers of the various fields where the spatial analysis is needed have demonstrated more interest in spatial statistics. Spatial data, also termed random field data consist of observations measured at known specific locations or within specific regions. Because there are innumerable situations in which data are collected at various locations in space, application fields of spatial statistics are extensive. For example, the application fields include geology, soil science, image processing, epidemiology, crop science, forestry, astronomy, atmospheric science, and environmental science. Many studies have been carried out in these fields. A representative example of how to use geostatistics in environmental problems is given by Journel (1984). Istok and Cooper (1988) demonstrated how to predict ground contaminant concentrations using geostatistics, and Myers (1989) implemented it to assess the movement of a multi-pollutant plume. Furthermore, Webster (1985) investigated soil characteristics and Piazza et al. (1981) analyzed gene frequencies.

Geostatistics emerged in the early 1980s as a hybrid discipline of mining engineering, geology, mathematics, and statistics. Its strength over more classical approaches to ore-reserve estimation is that it recognizes spatial variability at both the large scale and the small scale, or in statistical parlance it models both spatial trend and spatial correlation. Watson (1972) compares the two approaches and points out that most geological problem have a small-scale variation, typically exhibiting strong positive correlation between data at nearby spatial locations. One of the most important problems in geostatistics is to predict the ore grade in a mining block from observed samples. Matheron (1963) has called this process of prediction kriging (Cressie, 1993).

An important problem in geostatistical data is to predict the unobserved value $z(s_0)$ based on the information for n observations $z(s_\alpha)$, $\alpha = 1, \ldots, n$. It can be achieved in three stages of (1) estimation of the variogram, (2) fitting the theoretical variogram models to the sample variogram, and (3) predicting the value at a specified location using the fitted variogram model (kriging). It is very important to detect

influential observations that could affect the result of analysis when geostatistical data set is analyzed. Observed variables often contain outliers that have unusually large or small values when compared with others in a data set. Moreover, because variogram modeling is significantly affected by outliers, methods to detect and clean outliers from data sets are critical for proper variogram modeling.

On the one hand, these influential of outliers might give rise to the wrong result of kriging that is one of the major purposes in analyzing the geostatistical data. On account of these, the problem of detecting influential observations is embossed as a subject of interest in spatial statistics and many studies for this field have been progressing for sensitivity functions. Therefore, this thesis also put emphasis on the method to detect influential observations in the geostatistical statistics. For this purpose, sample influence functions (SIF) are derived as a tool to detect influential observations in stage above assuming that the underlying process of the observed geostatistical data is second-order stationary. Through the studies of the simulation and the real numerical example, we show the performance of the proposed method based on the sample influence functions.

We conduct a simulation study to demonstrate our procedure. For simplicity, we assume that the underlying process of the observed geostatistical data is stationary and isotropic. In all data analyses, we used the environment of R.

We describe the general geostatistical statistics approach in Chapter 2. This chapter consists of the contents as 1) types of spatial data, 2) spatial prediction, 3) estimation of the variogram, 4) fitting the theoretical variogram models to the sample variogram, and 5) predicting the value at a specified location using the fitted variogram model (kriging). In Chapter 3, here we address the problem of fitting a theoretical variogram model to various variogram estimators. In this chapter, we propose a method for choosing the optimal number of lags based on leave-one-out cross-validation (LOOCV) and the Akaike information criterion (AIC). Moreover, we compare the fitting a theoretical variogram model based on ordinary least square method with those based on maximum likelihood estimation. Chapter 4 deals with influence analysis for observations in the geostatistical analysis above. In this chapter,

we propose a procedure to detect outliers for geostatistical data analysis. Here, to detect outliers, we use the sample influence function (SIF) for the Akaike information criterion (AIC) and the maximum likelihood method. We present the simulation results to show the performance of our proposed procedure. In Chapter 5, by applying our approach to an empirical example with rainfall data in Chugoku district, Japan, we show the performance and usefulness of our proposed method. Finally, we give our concluding remarks in Chapter 6.

2. Outline of Spatial Statistics

2.1 Introduction

This chapter provides some introductory materials on spatial data analysis including some definitions and an overview of the basic ideas for spatial statistics.

2.2 Types of Spatial Data

Spatial data consist of observations or measurements measured at specific locations or within specific regions. In addition to values for various attributes of interest, spatial data sets also include the locations or relative positions of the data values. Locations may be point or region. For example, point referenced data are observations recorded at specific fixed locations and might be referenced by latitude and longitude. Areal referenced data are observations specific to a region. (Kaluzny et al., 1996)

Spatial data are largely classified into three types (Cressie, 1993); geostatistical data, lattice data, and spatial point patterns. Geostatistical data are measurements taken at fixed locations. The locations are generally continuous. Example of continuous geostatistical data include mineral concentrations measured at test sites within a mine, rainfall recorded at weather stations, concentrations of pollutants at monitoring stations, and soil permeabilities at sampling locations within a watershed. An example of discrete geostatistical data is cont data, such as the number of scallops at a series of fixed sampling sites along the coast. Lattice data are observations associated with spatial regions, where the regions can be regularly or irregularly spaced. The spatial regions can be any spatial collection, and are not limited to a grid. Generally, neighborhood information for the spatial regions is available. An example of regular lattice data is information obtained by remote sensing from satellites, and an example of irregular lattice data is cancer rates corresponding to county in a state. And spatial point patterns consist of a finite number of locations observed in a spatial region. Identification of spatial randomness, clustering, or regularity is often the first analysis performed when looking at point patterns. Examples of point pattern data

include locations of a species of tree in a forested region, and locations of earthquake epicenters. (Kaluzny et al., 1996)

In this paper, we focus on geostatistical data among three spatial data types that are dealt with in spatial statistics.

2.3 Spatial Prediction

Spatial data can be considered to be a realization of a stochastic process $Z(\mathbf{s})$, i.e.,

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

where $\mathbf{s}$ indicates a location in D and $\mathbf{R}^d$ $(d = 1, 2, 3)$ is a d-dimensional Euclidean space. The basic form of spatial data is expressed as $(\mathbf{z}_i, \mathbf{s}_i)$, $i = 1, \ldots, n$, where $\mathbf{z}_i$ is the i-th observation of a phenomenon of interest at location $\mathbf{s}_i$.

Assume that this process satisfies the hypothesis of intrinsic stationarity:

$$\begin{aligned}
&(a)\ E(Z(\mathbf{s})) = \mu, \quad \text{for all } \mathbf{s} \in D, \\
&(b)\ Cov(Z(\mathbf{s}_i), Z(\mathbf{s}_j)) = C(\mathbf{h}) = C(\mathbf{s}_i - \mathbf{s}_j) < \infty, \quad \text{for all } \mathbf{s}_i, \mathbf{s}_j \in D, \\
&(c)\ Var(Z(\mathbf{s}_i) - Z(\mathbf{s}_j)) = 2\gamma(\mathbf{s}_i - \mathbf{s}_j) = 2\gamma(\mathbf{h}), \quad \text{for all } \mathbf{s}_i, \mathbf{s}_j \in D,
\end{aligned}$$

where $2\gamma(\mathbf{h})$ is the variogram, and $C(\mathbf{h})$ is the covariance for pairs of points separated by Euclidean distance (the covariogram). In this paper, we suppose that $2\hat{\gamma}(\mathbf{h})$ is a variogram estimator for a given lag $\mathbf{h}$, based on a sample $\{Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)\}$ of the spatial process; let $\mathbf{h}_1, \ldots, \mathbf{h}_k$ be the vector lags defined by $\mathbf{h}_i = i\mathbf{h}/\|\mathbf{h}\|$, $i = 1, \ldots, k$, where $1 \le k \le K$, and K is the maximum possible distance between data in the direction $\mathbf{h}$ (Genton, 1998).

2.3.1 Estimation of the Variogram

We measure the variability of a regionalized variable $z(\mathbf{s})$ at different scales by computing the dissimilarity between pairs of data values, $z(\mathbf{s}_i)$ and $z(\mathbf{s}_j)$ say, located at points $\mathbf{s}_i$ and $\mathbf{s}_j$ in a saptial domain D. The measure for the dissimilarity of two values, labeled γ_{ij}^*, is

$$\gamma_{ij}^{*} = \frac{(z(\mathbf{s}_i) - z(\mathbf{s}_j))^2}{2},$$

i.e. half of the square of the difference between the two values.

We let the dissimilarity γ^* depend on the spacing and on the orientation of the point pair described by the vector $\mathbf{h}$,

$$\gamma^*(\mathbf{h}) = \frac{1}{2}(z(\mathbf{s}_i + \mathbf{h}) - z(\mathbf{s}_i))^2,$$

Using all sample pairs in a data set (up to a distance of half the diameter of the region), a plot of the dissimilarities γ^* against the spatial separation $\mathbf{h}$ is produced which is called the variogram cloud. A schematic example is given on Figure 2.1.

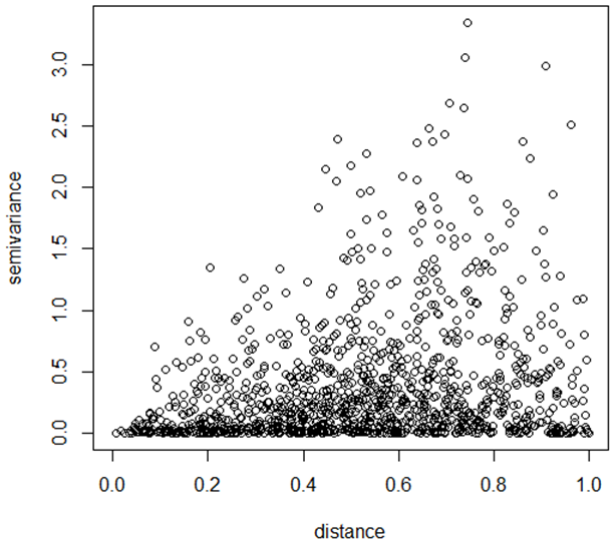


Figure 2.1: Plot of the dissimilarities γ^* against the spatial separation $\mathbf{h}$ of sample pairs; a variogram cloud.

The first step in geostatistical data analysis is estimating the variogram $\gamma(\mathbf{h})$ using the observed data. When we assume the variogram to be isotropic, we can calculate an estimator for the variogram, called the sample variogram (Matheron, 1962), using

$$\begin{aligned}\hat{\gamma}(\mathbf{h}_k) &= \frac{1}{2|N_k|}\sum_{N(\mathbf{h})}(z(\mathbf{s}_i)-z(\mathbf{s}_j))^2 \\ &= \frac{1}{2|N_k|}\sum_{N(\mathbf{h})}(z(\mathbf{s}_i+\mathbf{h})-z(\mathbf{s}_i))^2,\end{aligned} \tag{2.1}$$

where $N(\mathbf{h})=\{(\mathbf{s}_i,\mathbf{s}_j):\mathbf{s}_i-\mathbf{s}_j=\mathbf{h};i,j=1,\ldots,n\}$ and $|N_k|$ is the number of the distinct pairs in $N(\mathbf{h})$. $z(\mathbf{s}_i)$ and $z(\mathbf{s}_j)$ are the data values at spatial locations $\mathbf{s}_i$ and $\mathbf{s}_j$, respectively.

In this formulation, h represents a distance measure with only magnitude. When the variogram is isotropic, we can compute the directional sample variogram using the same formula by replacing h with vector $\mathbf{h}$. In practice, to calculate the variogram values using Eq. (2.1), we first select the lag distances $\mathbf{h}$, then calculate the variogram values by regarding pairs with distance within $\mathbf{h}\pm lag$ tolerance as the pairs in $N(\mathbf{h})$. The lag tolerance, which establishes distance bins for the lag increments, accommodates for unevenly spaced observations. The lag increment defines the distances at which the variogram is calculated, and the number of lags in conjunction with the size of the lag increment will define the total distance over which the variogram is being calculated. To estimate the variogram, we next have to choose the lag increment or the number of lags.

More formally, if N_k denotes the set of distance pairs, $(\mathbf{s}_i,\mathbf{s}_j)$, in bin k, (with the size (number of pairs) in N_k denoted by $|N_k|$), and if the distance between each such pair is denoted by $\mathbf{h}_{ij}=\|\mathbf{s}_i-\mathbf{s}_j\|$, then the lag distance, $\mathbf{h}_k$, for bin k is defined to be

$$\mathbf{h}_k = \frac{1}{|N_k|}\sum_{(\mathbf{s}_i,\mathbf{s}_j)\in N_k}\mathbf{h}_{ij}.$$

2.3.2 Fitting Theoretical Variogram Models to Sample Variogram

The next stage in geostatistical data analysis is fitting a model that gives the best dependence (auto-correlation structure) in the underlying stochastic process. Most variogram models contain three parameters which are sill, range, and nugget (or nugget effect). These parameters are depicted on the generic variogram shown in Figure 2.2 and are defined as follows.

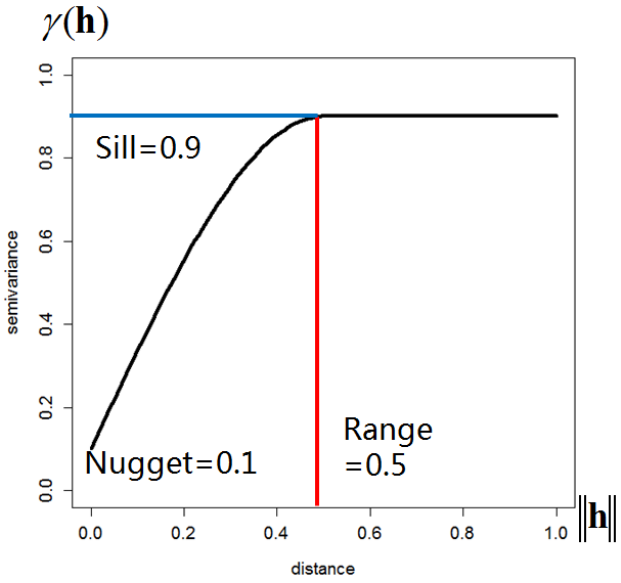


Figure 2.2: A generic variogram showing the sill, and range parameters along with a nugget effect.

Sill is a variogram threshold for lag distances. Range is the lag distance at which the variogram reaches the sill value. The nugget represents the variability at distances smaller than the typical sample spacing, including the measurement error. Thus far, several variogram models have been proposed according to their forms; for example, Gaussian, exponential, and spherical models as bounded variogram models, and power, linear and nugget effect models as unbounded variogram models (Figure 2.3). The selected model influences the prediction of the unknown values, particularly when the shape of the curve near the origin differs significantly. The steeper the curve near the

origin, the more influence the closest neighbors will have on the prediction. Each model is designed to fit different types of phenomena more accurately. These models are defined as follows:

Gaussian model is

$$\gamma_{gau}(\mathbf{h};\boldsymbol{\theta}) = \begin{cases} c_n + c_s\left\{1-\exp\left(-\dfrac{\|\mathbf{h}\|^2}{c_r^{\,2}}\right)\right\}, & \|\mathbf{h}\| > 0, \\ 0, & \|\mathbf{h}\| = 0, \end{cases}$$

for $\boldsymbol{\theta} = (c_n, c_s, c_r)'$, $c_n \geq 0, c_s \geq 0$, and $c_r \geq 0$.

Exponential model is

$$\gamma_{\exp}(\mathbf{h};\boldsymbol{\theta}) = \begin{cases} c_n + c_s\left\{1-\exp\left(-\dfrac{\|\mathbf{h}\|}{c_r}\right)\right\}, & \|\mathbf{h}\| > 0, \\ 0, & \|\mathbf{h}\| = 0, \end{cases}$$

for $\boldsymbol{\theta} = (c_n, c_s, c_r)'$, $c_n \geq 0, c_s \geq 0$, and $c_r \geq 0$.

Spherical model is

$$\gamma_{sph}(\mathbf{h};\boldsymbol{\theta}) = \begin{cases} c_n + c_s, & \|\mathbf{h}\| > c_r, \\ c_n + c_s\left\{\dfrac{3}{2}\left(\dfrac{\|\mathbf{h}\|}{c_r}\right) - \dfrac{1}{2}\left(\dfrac{\|\mathbf{h}\|}{c_r}\right)^3\right\}, & 0 < \|\mathbf{h}\| \leq c_r, \\ 0, & \|\mathbf{h}\| = 0, \end{cases}$$

for $\boldsymbol{\theta} = (c_n, c_s, c_r)'$, $c_n \geq 0, c_s \geq 0$, and $c_r \geq 0$.

Power model is

$$\gamma_{pow}(\mathbf{h};\boldsymbol{\theta}) = \begin{cases} c_n + c_s\|\mathbf{h}\|^{c_r}, & \|\mathbf{h}\| > 0, \\ 0, & \|\mathbf{h}\| = 0, \end{cases}$$

for $\boldsymbol{\theta} = (c_n, c_s, c_r)'$, $c_n \geq 0, c_s \geq 0$, and $c_r \geq 0$.

Linear model is

$$\gamma_{lin}(\mathbf{h};\boldsymbol{\theta}) = \begin{cases} c_n + c_s\|\mathbf{h}\|, & \|\mathbf{h}\| > 0, \\ 0, & \|\mathbf{h}\| = 0, \end{cases}$$

for $\boldsymbol{\theta} = (c_n, c_s)'$, $c_n \geq 0, c_s \geq 0$.

Nugget effect model is

$$\gamma_{nug}(\mathbf{h};\boldsymbol{\theta}) = \begin{cases} c_n, & \|\mathbf{h}\| > 0, \\ 0, & \|\mathbf{h}\| = 0, \end{cases}$$

for $\boldsymbol{\theta} = (c_n)'$, $c_n \geq 0$.

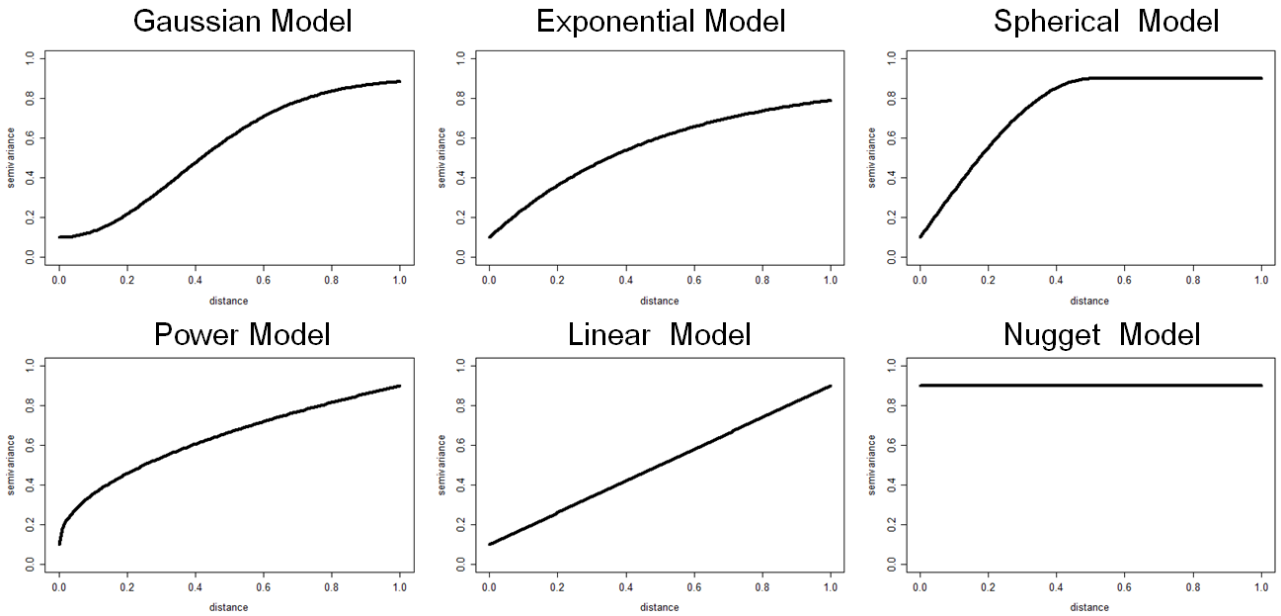


Figure 2.3: Theoretical variogram models.

2.3.3 Kriging

Kriging is a linear interpolation method that allows predictions of unknown values in a random function from observations at known locations (Figure 2.4). There are a few type of kriging for spatial prediction problems in spatial statistics, including simple kriging, ordinary kriging, and universal kriging. In our simulation, we perform only ordinary kriging, which is often associated with the best linear unbiased estimator (BLUE). Ordinary kriging is based on a random function model of spatial correlation for calculating a weighted linear combination of available samples to predict a nearby unsampled location. Weights are chosen to ensure zero average error for the model and to minimize the model's error variance (Isaaks and Srivastava, 1989). Ordinary kriging (Matheron, 1971; Journel and Huijbregts, 1978) refers to spatial prediction under the following two assumptions. First, the model assumption is as follows:

$$Z(\mathbf{s}) = \mu + \delta(\mathbf{s}), \quad \mathbf{s} \in D, \mu \in \mathbf{R},$$

where μ is unknown.

The second is the predictor assumption:

$$Z^*_{OK}(\mathbf{s}_0) = \sum_{\alpha=1}^{n} w_\alpha Z(\mathbf{s}_\alpha) \ .$$

To minimize the error variance under the constraint $\sum_{\alpha=1}^{n} w_\alpha = 1$, we set up a system that minimizes Q, comprising the error variance and an additional term involving the Lagrange parameter, m_{OK}:

$$Q = \mathrm{E}\left[\left(Z^*_{OK}(\mathbf{s}_0) - Z(\mathbf{s}_0)\right)^2\right] + 2m_{OK}\left(1 - \sum_{\alpha=1}^{n} w_\alpha\right).$$

This minimization with respect to the Lagrange parameter forces the constraint to be obeyed:

$$\begin{cases} \dfrac{\partial Q}{\partial w_\beta} = -2\sum_{\alpha=1}^{n} w_\alpha^{OK}\gamma(\mathbf{s}_\alpha - \mathbf{s}_\beta) + 2\gamma(\mathbf{s}_\beta - \mathbf{s}_0) - 2m_{OK} = 0, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \beta = 1,\ldots,n, \\ \dfrac{\partial Q}{\partial m_{OK}} = 1 - \sum_{\alpha=1}^{n} w_\alpha^{OK} = 0. \end{cases}$$

In this case, the system of equations for the kriging weights is

$$\begin{cases} \sum_{\beta=1}^{n} w_\beta^{OK}\gamma(\mathbf{s}_\alpha - \mathbf{s}_\beta) + m_{OK} = \gamma(\mathbf{s}_\alpha - \mathbf{s}_0), \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \alpha = 1,\ldots,n, \\ \sum_{\beta=1}^{n} w_\beta^{OK} = 1, \end{cases}$$

where $\gamma(\cdot)$ is the covariance function for the residual component of the variable. Once the kriging weights (and Lagrange parameter) are obtained, the error variance of the ordinary kriging is given by

$$\sigma_{OK}^2 = m_{OK} - \gamma(0) + \sum_{\alpha=1}^{n} w_\alpha^{OK}\gamma(\mathbf{s}_\alpha - \mathbf{s}_0).$$

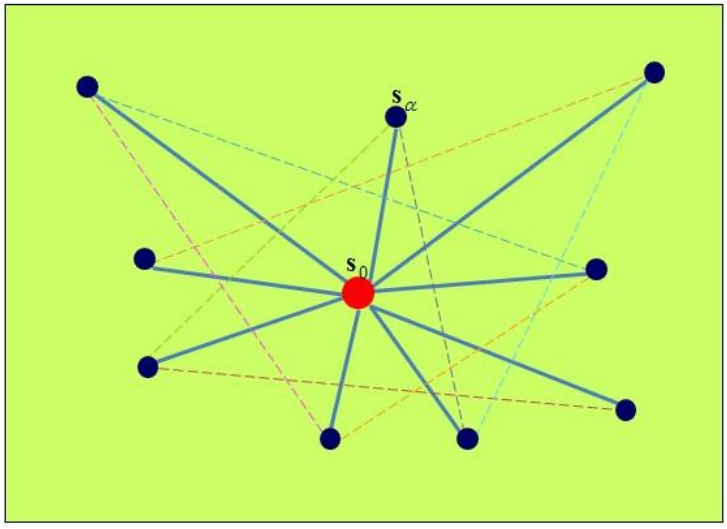


Figure 2.4: The estimation of a value at a point $\mathbf{s}_0$ using information at point $\mathbf{s}_\alpha$, $\alpha = 1,\ldots,n$.

The flow of geostatistical data analysis, from estimating variograms to kriging, is expressed in Figure 2.5.

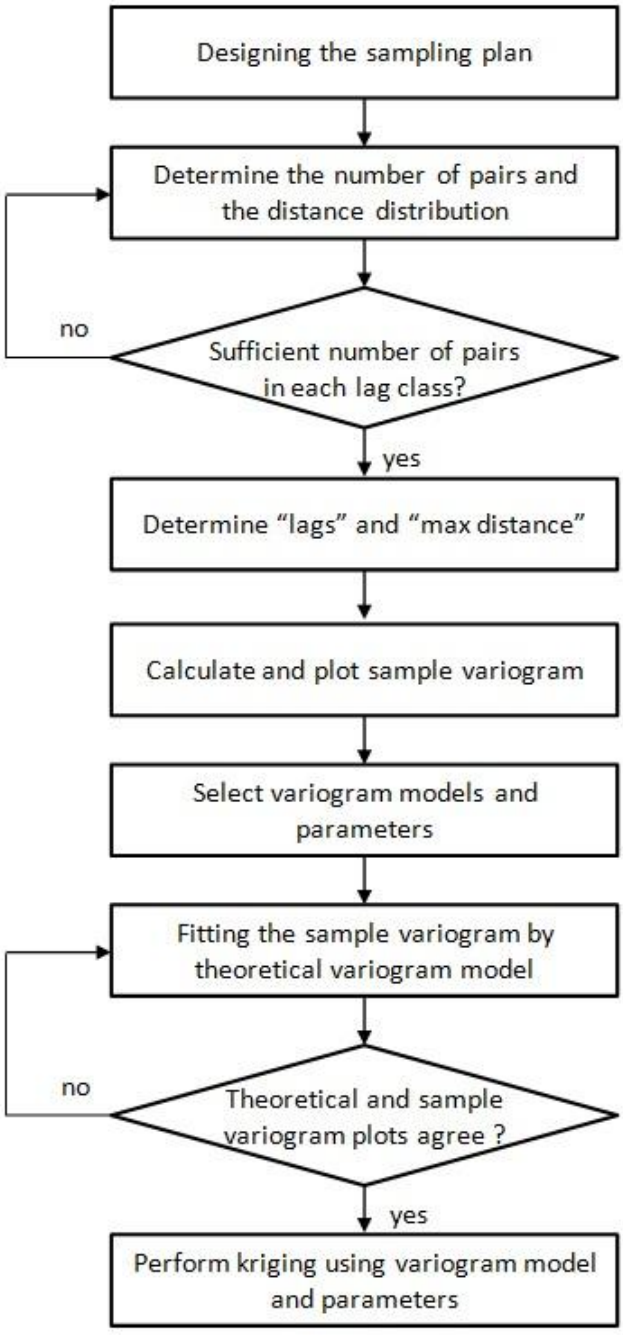


Figure 2.5: Flowchart for geostatistical data analysis.

3. Variogram Model Fitting

3.1 Introduction

In geostatistical data analysis, methods by variogram parameters are based on the fitting of a sample variogram calculated from the data to a theoretical variogram model. Until recently, the most common methods of fitting variogram models to sample variograms were by eye or by least squares. The advantage of the ordinary least square method is the applicability for the analysis without specifying the parameters for distributions. In addition, its method also has the benefit of the computation cost for large amounts of geostatistical data. However, when we use the method in geostatistical data analyses, we have to estimate indirectly the variogram parameters by dividing the group for lag. Variogram estimation is strongly influenced by number of lags k, which serves as a smoothing parameter. This means that k could significantly influence the least square estimator and kriging predictor. However, there is no established rule for selecting the number of lags when estimating variograms.

The selection of a proper k value is important, so few studies have been done in this regard. Kim et al. (2013b, 2014b) proposed a method for selecting the optimal number for the estimator using LOOCV and AIC in the geostatistical data analysis. Lamorey and Jacobson (1995) says the sensitivity of the variogram fit to small changes in the lag increment is used to evaluate if there are enough data to define accurately the sample variogram. Choi et al. (2010) proposed a method for finding the optimal lag using the predicted residual sum of square (PRESS). Hong and Kim (2004) proposed the selection of k in nonparametric variogram estimation in the sense of minimizing the limit of mean integrated squared error (MISE) under infill asymptotics and mixed-increasing domain asymptotics. In this research they have shown that under infill asymptotics small value of k given best results even for large number of sample size. In this chapter, we propose a method for choosing the optimal number of lags based on leave-one-out cross-validation (LOOCV) and the Akaike information criterion (AIC). Moreover, besides the ordinary least square method, we generally use variogram models based on maximum likelihood estimation. We compare the

estimated parameters of the variogram models based on the ordinary least square method with those based on maximum likelihood estimation.

3.2 Least Squares Method

In many cases, variogram parameters are often estimated by using the approach of ordinary least squares. Cressie (1985) introduced three mathematical techniques for fitting the parameter values; ordinary least squares (OLS), weighted least squares (WLS), and generalized least squares (GLS).

The method of OLS is purely a numerical procedure that has an attractive geometric interpretation. The WLS method of variogram model fitting can be implemented through any number of nonlinear estimation algorithms. Model fitting by GLS requires calculating the variances of the estimates of the sample variogram at each lag and covariances between them, which is very complicated.

3.2.1 Ordinary Least Squares

The method of ordinary least square specifies $\boldsymbol{\theta}$ that is estimated by minimizing

$$\sum_{k=1}^{K}\left(\hat{\gamma}^{o}(h_k)-\gamma^{o}(h_k;\boldsymbol{\theta})\right)^2, \tag{3.1}$$

for some direction k. Here k is the number of lags. Eventually, an ordinary least square estimator of $\boldsymbol{\theta}$ is obtained. Although Eq. (3.1) has geometric appeal, it does not contain the information for the distributional variation and covariation of the generic estimator $\hat{\gamma}^{o}$. In Figure 3.1, we represent the sample variogram for each the number of lags at same dataset by the ordinary least squares. We can see that the shape and parameter estimation is influenced by number of lags.

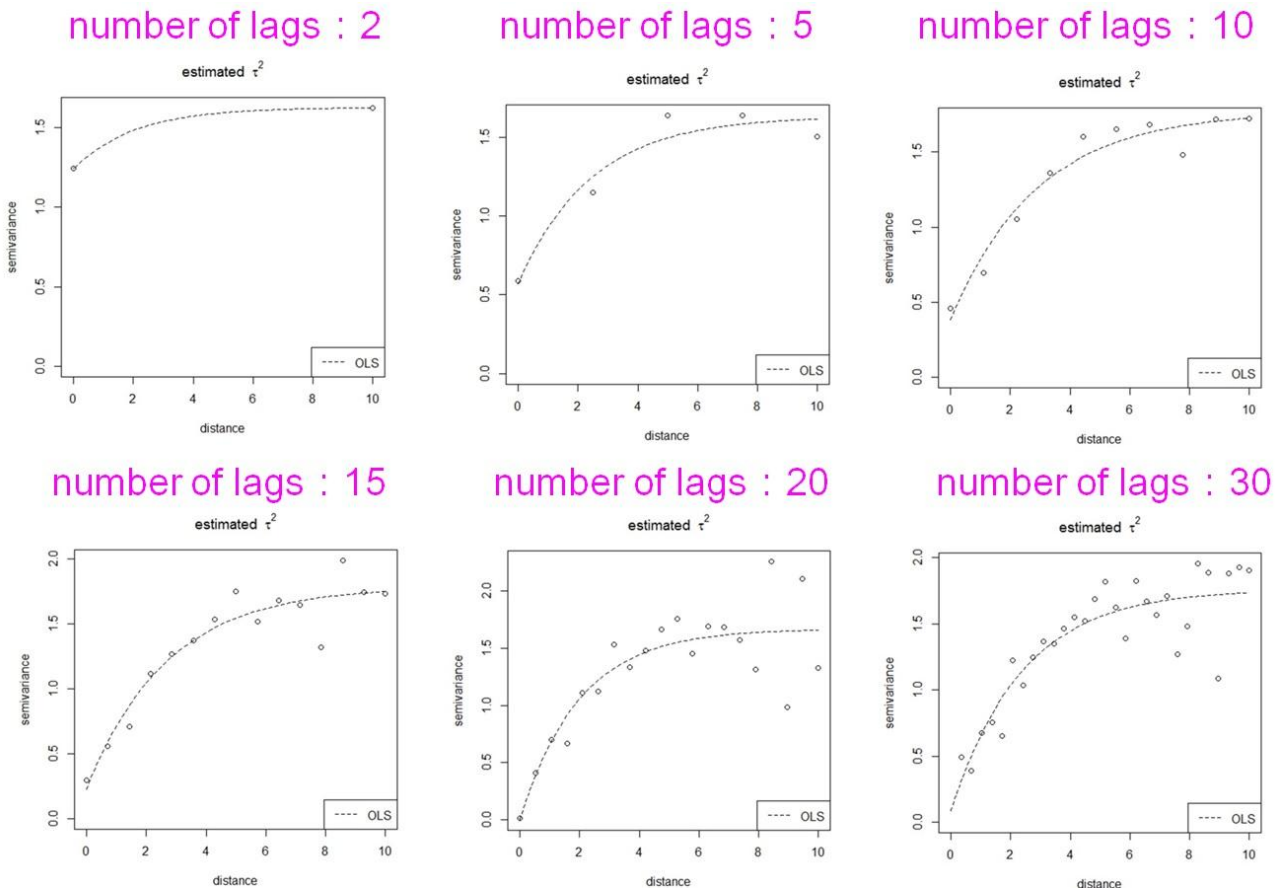


Figure 3.1: Sample variogram for number of lags by the ordinary least squares.

3.2.2 Optimal Number of Lags

In general, the variogram is estimated with a method of moment estimator (Matheron, 1962), and the lag increment or number of lags must be chosen as it is being estimated. In practical simulation analysis, a data analyst estimates the variogram using several different numbers of lags, and then selects the best number of lags value among them. This method is subjective and can sometimes result in preposterous variogram estimation values. This section proposes a method for choosing the optimal number of lags when estimating variograms based on the given geostatistical data.

Since in any finite sample there will generally be at most one pair that are separated by a given distance $\mathbf{h}$, one must necessarily aggregate point pairs $(\mathbf{s}_i, \mathbf{s}_j)$ with similar distances and hence estimate $\gamma(\mathbf{h})$ at only a small number of representative distances for each aggregate. The simplest way to do so is to partition distances into intervals, called bins, and take the average distance, $\mathbf{h}$, in each bin k to be the appropriate representative distances, called lag distances, as shown in following Figure 3.2.

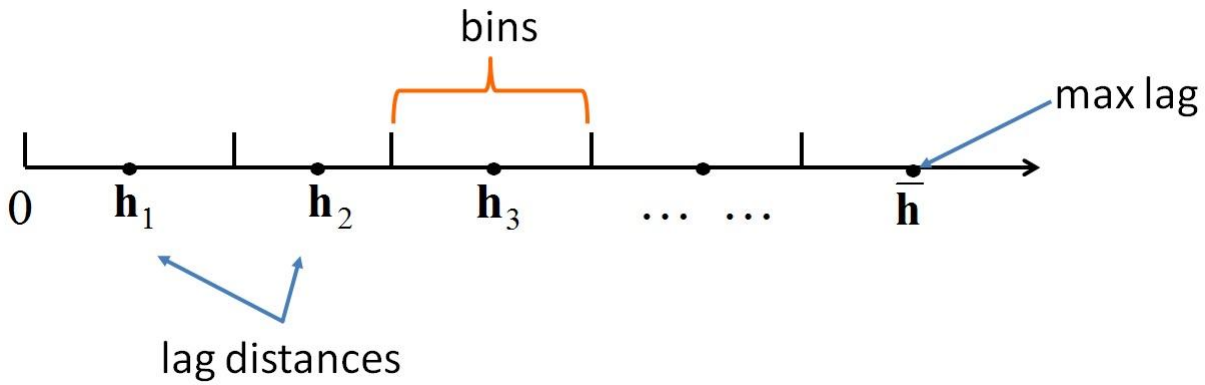


Figure 3.2: Lag distance and bins.

This set of estimates at each lag distance is designated as the sample variogram. An schematic example of sample variogram construction is given in Figure 3.3. The vertical lines separate the bins, as shown for bins k and $k+1$. The red dot in the middle of these points denotes the pair of average values, $(\mathbf{h}_k, \hat{\gamma}_k)$, representing all points in that bin. Hence the sample variogram consists of all these average points, one for each bin of points.

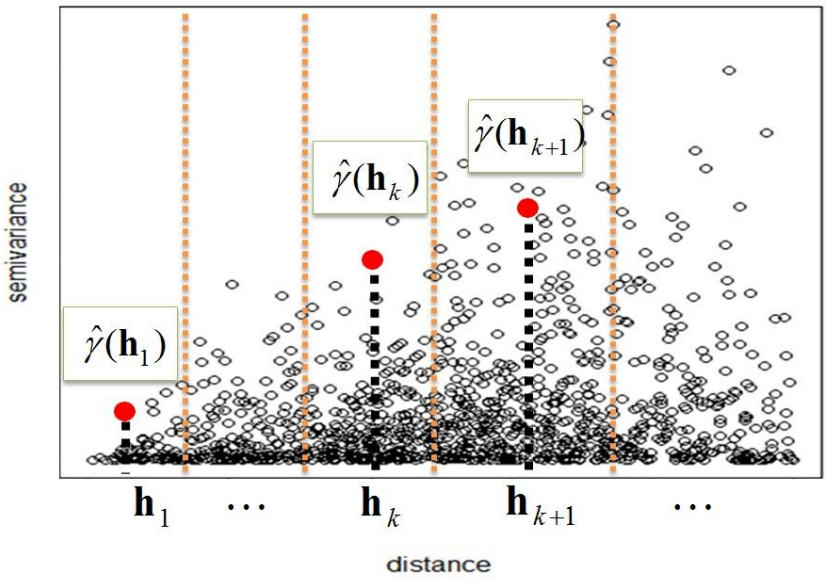


Figure 3.3: Sample variogram.

To determine the size of each bin, the most common approach is to make all bins the same size, in order to insure a uniform approximation of lag distances within each bin. Therefore, distances are subdivided into a number of intervals called lags as illustrated in the following Figure 3.4. The lag intervals are defined in the sample variogram dialog by entering a total number of lags, a unit lag separation distance, and a lag tolerance.

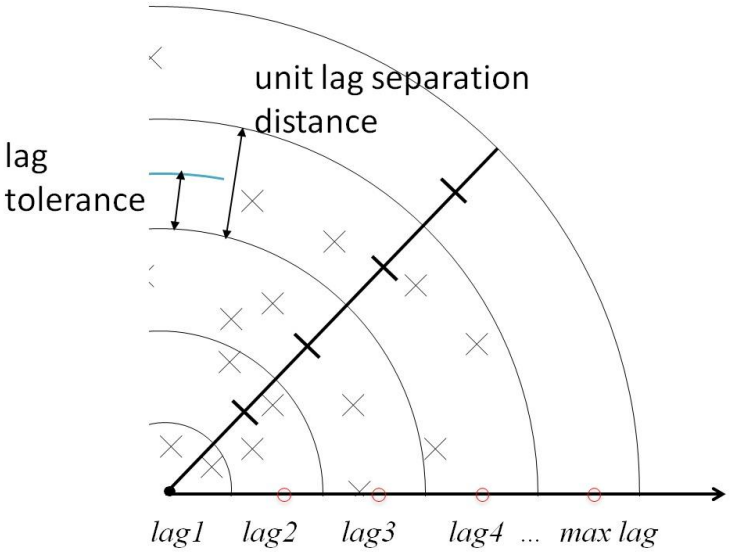


Figure 3.4: Lag separation.

The selection of lag size has significant effects on the sample variogram. For example, if the lag size is too large, short-range autocorrelation may be masked. If the lag size is too small, there may be many empty bins, and sample sizes within the bins will be too small to determine the bins representative averages. However, if the data are acquired using an irregular or random sampling scheme, a suitable lag size selection is not at all straightforward.

There is an implicit tradeoff here between approximation of lag distances and the number of point pairs used to estimate the variogram at each lag distance. Journel and Huijbregts (1978) suggest the following two practical rules in choosing the lag increment and number of lags: (i) the sample variogram should only be considered for distances $\mathbf{h}$ for which the number of pairs is greater than 30, and (ii) the distance of reliability for a sample variogram is $\mathbf{h} < D/2$, where D is the maximum distance over the field of data. However, in practice, these rules are ambiguous when choosing the number of lags or the lag increment. In this section, we rules on the number of lags denoted by symbol k because the above two rules are mutually reciprocal. Our main interest thus becomes finding the optimal number of lags among possible k values.

3.2.2.1 Optimal Number of Lags for Leave-One-Out Cross-Validation

We carried out a simulation study to select of the optimal number of lags. In this section, we consider the exponential and spherical models, which each contain three parameters (sill, range, and nugget), and we restrict the scope of the number of lags to be from 2 to 20 when selecting the optimal k.

As mentioned above, the simulation data are fixed in the two models and their three parameters, and the generated datasets (with sample sizes of 100, 200, and 300) include positions as well as the data values. When a theoretical variogram model is fitted to the number of lags k from 2 to 20, the optimal k can be selected on the basis of leave-one-out cross-validation (LOOCV).

The LOOCV (Devijver and Kittler, 1982) values in Eq. (3.2), are calculated as follows. The LOOCV uses a single observation from the original sample as the validation data, and the remaining observations as the training data. This is repeated such that each observation in the sample is used once as the validation data:

$$\frac{1}{n}\sum_{i=1}^{n}\left|Z(s_i)-\hat{Z}(s_{-i})\right|^2 \tag{3.2}$$

where $Z(s_i)$ and $\hat{Z}(s_{-i})$ represent the observed and predicted values, respectively. The selection of the optimal k can be explained as below.

Step 1. For a fixed lag k $(2 \le k \le 20)$, estimate the variogram using $n-1$ observations excepting the i-th one and obtain the predicted value $\hat{Z}(s_{-i})$ at the i-th location based on the estimated variogram.

Step 2. For every i $(i=1,\ldots,n)$, calculate $Z(s_i)-\hat{Z}(s_{-i})$ based on the Step 1 from 1 to n ($n=100$, 200 and 300).

Step 3. For the fixed lag k $(2 \le k \le 20)$, calculate the LOOCV value $\frac{1}{n}\sum_{i=1}^{n}\left|Z(s_i)-\hat{Z}(s_{-i})\right|^2$.

Step 4. Calculate the LOOCV for every k $(2 \le k \le 20)$, and select the optimal k which minimizes the LOOCV.

The LOOCV results for given numbers of lags is presented in Tables 3.1 and 3.2. From Tables 3.1 and 3.2, we can see that the LOOCV value becomes smaller as the number of lags increased.

Table 3.1: Results of using LOOCV for choosing the optimal number of lags (Exponential model).

Number of lags	Sample size		
	100	200	300
2	1.0801	1.0188	0.9065
3	1.0193	0.993	0.9105
4	0.897	0.8377	0.7439
5	0.7079	0.6566	0.7001
6	0.6611	0.601	0.4275
7	0.6497	0.572	0.4102
8	0.6322	0.5529	0.3983
9	0.6297	0.539	0.3892
10	0.6174	0.5322	0.3829
11	0.6113	0.508	0.3777
12	0.6107	0.5024	0.3733
13	0.6097	0.4972	0.3698
14	0.6062	0.4928	0.3672
15	0.6064	0.4867	0.3652
16	0.6044	0.4883	0.3629
17	0.6043	0.4851	0.3622
18	0.6043	0.4838	0.3618
19	0.6041	0.4821	0.3614
20	0.604	0.4819	0.3607

Table 3.2: Results of using LOOCV for choosing the optimal number of lags (Spherical model).

Number of lags	Sample size		
	100	200	300
2	2.3746	1.7157	1.7523
3	2.2646	1.5523	1.4381
4	2.1245	1.3444	1.1612
5	1.9845	1.1996	1.0182
6	1.9038	1.1152	0.7815
7	1.8287	1.0637	0.7447
8	1.806	1.0291	0.7147
9	1.7799	0.9971	0.6954
10	1.7733	0.978	0.6815
11	1.4172	0.7637	0.5648
12	1.4109	0.7489	0.5521
13	1.3893	0.739	0.5435
14	1.4171	0.7361	0.5383
15	1.3775	0.7329	0.5324
16	1.3841	0.7234	0.5257
17	1.3824	0.72	0.5229
18	1.3724	0.7145	0.521
19	1.3581	0.7158	0.5203
20	1.3511	0.712	0.5181

3.2.2.2 Optimal Number of Lags for Akaike Information Criterion

A satisfactory compromise between goodness of fit and complexity of the model can be achieved based on the Akaike information criterion (AIC). For a given set of data, the variable part of the AIC is estimated by

$$\hat{A} = -2n\ln\hat{R} + 2p$$

where n is the number of sample points on the variogram, $\hat{R}$ is the value of R which maximizes the likelihood (R is a vector of m parameters of covariogram model), and p is the number of parameters in the variogram model. The model to choose is the one for which $\hat{A}$ is least.

Similarly, when applying the AIC, the simulation data are fixed in the two models and their three parameters, and the generated datasets (with sample sizes of 100, 200, and 300) include positions and the data values. When a theoretical variogram model is fitted to the number of lags k from 2 to 20, the optimal number of lags k can be selected on the basis of the AIC. The optimal k is defined to be the value that minimizes AIC. The selection of the optimal k can be explained as below.

Step 1. Calculate the $\hat{R}$ with the given data $\mathbf{z}$ and parameters of covariogram model R.

Step 2. Calculate the AIC for variogram model for every lag k $(2 \le k \le 20)$.

Step 3. Select the optimal k which minimizes the AIC.

From Table 3.3, for the sample size of 100 in the exponential variogram model, the minimum AIC value is achieved at k = 5; for sample sizes of 200 and 300, the minimum AIC values are achieved at k = 7. In addition, from Table 3.4, for the sample size of 100 in the spherical variogram model, the minimum value of AIC is achieved at k = 5; for sample sizes of 200 and 300, the minimum values of AIC are achieved at k = 6 and k = 7, respectively.

Table 3.3: Results of applying the AIC for choosing the optimal number of lags (Exponential model).

Number of lags	Sample size		
	100	200	300
2	871.93	1504.93	2148.32
3	867.5	1479.62	2108.39
4	865.26	1476.76	2100.05
5	**865.13**	1475.28	2094.52
6	866.57	1474.24	2094.13
7	868.43	**1473.09**	**2092.02**
8	868.83	1475.09	2092.87
9	870.19	1477.29	2093.34
10	871.16	1477.72	2093.41
11	874.27	1479.66	2093.56
12	875.17	1481.74	2095.99
13	875.76	1482.48	2096.34
14	877.4	1483.9	2098.17
15	878.95	1483.82	2099.03
16	881.18	1485.92	2101.32
17	881.78	1488.6	2102.84
18	882.28	1489.51	2104.42
19	882.47	1490.4	2105.34
20	887.59	1491.95	2107.46

Table 3.4: Results of applying the AIC for choosing the optimal number of lags (Spherical model).

Number of lags	Sample size		
	100	200	300
2	871.79	1500.28	2152.07
3	862.63	1474.49	2110.97
4	860.67	1470.3	2094.99
5	**859.89**	1469.23	2096.2
6	861.22	**1468.11**	2095.4
7	862.79	1468.42	**2092.13**
8	864.56	1468.15	2093.3
9	865.9	1470.65	2096.14
10	866.52	1472.52	2095.69
11	868.29	1471.98	2097.3
12	869.97	1474.22	2098.08
13	871.39	1476.4	2099.62
14	872.74	1476.99	2099.43
15	874.47	1478.59	2102.89
16	874.3	1480.46	2102.77
17	878.09	1481.32	2105.35
18	878.13	1482.73	2105.4
19	879.42	1484.28	2107.99
20	880.95	1486.11	2109.07

3.3 Maximum Likelihood Method

Estimation procedures that rely crucially on the Gaussian assumption are maximum likelihood (ML) and restricted maximum likelihood (REML) estimation of $\boldsymbol{\theta}$ in

$$P = \{2\gamma : 2\gamma(\cdot) = 2\gamma(\cdot;\boldsymbol{\theta}); \boldsymbol{\theta} \in \boldsymbol{\Theta}\}, \tag{3.3}$$

where P is parametric subset of valid variograms.

The problem with ML estimation is that the estimators of $\boldsymbol{\theta}$ are biased, often prohibitively so in small to moderate samples (Matheron, 1971; Mardia and Marshall, 1984). Maximum likelihood estimation for a model of the spatial covariance of a random variable was proposed by Kitanidis (1983, 1987) for geostatistical purposes. The simple case when the data $\mathbf{Z}$ are in fact independent multivariate Gaussian, $Gau(X\boldsymbol{\beta}, \sigma^2 I)$, yields just one small scale variogram parameter $\theta = \sigma^2$. The ML estimator is, $\hat{\sigma}^2 = \sum_{i=1}^{n}(Z(\mathbf{s}_i) - X\hat{\boldsymbol{\beta}})^2 / n$ where $\hat{\boldsymbol{\beta}}$ is the ordinary least squares estimator of the $q \times 1$ vector $\boldsymbol{\beta}$. It is well known that $\hat{\sigma}^2$ is biased and that $(n/(n-q))\hat{\sigma}^2$ is unbiased; the bias-correction factor $(n/(n-q))$ can be appreciable when q is large relative to n (Cressie, 1993). Suppose that the data $\mathbf{Z}$ are multivariate Gaussian $Gau(X\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$, where X is an $n \times q$ matrix of rank $q < n$, and that the $n \times n$ matrix $\sum(\boldsymbol{\theta}) = (\text{cov}(Z(\mathbf{s}_i), Z(\mathbf{s}_j)))$ depends on $\boldsymbol{\theta}$ through Eq. (3.3). Then the negative loglikelihood is

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{n}{2}\log(2\pi) + \frac{1}{2}\log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| + \frac{1}{2}(\mathbf{Z} - X\boldsymbol{\beta})' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})(\mathbf{Z} - X\boldsymbol{\beta}),$$

and the maximum likelihood estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ satisfy

$$\hat{L} = L(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) = \inf\{L(\boldsymbol{\beta}, \boldsymbol{\theta}) : \boldsymbol{\beta} \in \mathbf{R}^{\mathbf{q}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}.$$

The restricted maximum likelihood method is a particular form of maximum likelihood estimation which is based on all the information target data, but instead uses a likelihood function calculated from a transformed set of data. In the case of a one-dimensional with equally spaced data, Kitanidis (1983) proposed the approach based on the likelihood function by the data $\mathbf{W} \equiv (Z(1) - Z(2), Z(2) - Z(3), \ldots, Z(n-1) - Z(n))'$.

Equivalently, minimize

$$L_w(\boldsymbol{\beta},\boldsymbol{\theta}) = \frac{n-1}{2}\log(2\pi) + \frac{1}{2}\log|A'\boldsymbol{\Sigma}(\boldsymbol{\theta})A| + \frac{1}{2}(\mathbf{W} - A'X\boldsymbol{\beta})'(A'\boldsymbol{\Sigma}(\boldsymbol{\theta})A)^{-1}(\mathbf{W} - A'X\boldsymbol{\beta}),$$

where $A = (a_{ij})$ is an $(n-1)\times n$ matrix whose elements are

$$a_{ij} = \begin{cases} 1, & \text{for } i = j,\ j = 1,\ldots,n-1, \\ -1, & \text{for } i = j+1,\ j = 1,\ldots,n-1, \\ 0, & \text{elsewhere}. \end{cases}$$

In this section, we present a simulation study to validate the proposed estimation method. We compare the estimated parameters of the variogram models based on the ordinary least square method with those based on maximum likelihood estimation. As calculated above, we selected the number of lags based on LOOCV when estimating variogram based on ordinary least square method. The $\mathbf{z}_i (i = 1,\ldots,n)$ observation with the specified parameter the $(\mathbf{s}_i, \mathbf{s}_j)$ positions can be generated using the function 'grf' in spatial module geoR of R. The approach of parameter estimation can be explained as below.

Step 1. Fix the three models and the values of parameters (sill = 2, range = 2, and nugget = 0.1), and generate datasets (100, 200, and 300) including positions as well as data values.

Step 2. Estimate parameters in two ways (OLS and ML).

Step 3. Predict the prediction point by using kriging.

Step 4. Compare the predicted values with observed values.

The procedure is repeated 30 times. The simulation results for Gaussian, exponential, and spherical models, obtained by the above procedures are presented in Table 3.5 to Table 3.10. Table 3.5, Table 3.7, and Table 3.9 show the parameters for the ordinary least squares (OLS) and the maximum likelihood (ML) estimation method, respectively. Table 3.6, Table 3.8, and Table 3.10 show the results in terms of the leave-one-out cross-validation for the ordinary least squares and the maximum likelihood estimation method, when we used the evaluation measures were used in Eq. (3.2). From Table 3.5 to Table 3.10, the parameter estimation methods based on maximum likelihood estimation gave a better performance than OLS method from the point of view of LOOCV.

Table 3.5: Parameters for the OLS and the ML estimation method ($n=100$).

Estimation method	Models	Parameters		
		Nugget	Sill	Range
ordinary least squares (OLS)	Gaussian	0.627	0.748	1.183
	Exponential	0.628	0.753	0.747
	Spherical	0.692	0.734	1.013
maximum likelihood (ML)	Gaussian	0.179	1.283	1.180
	Exponential	0.000	1.448	1.356
	Spherical	0.021	1.299	2.364

Table 3.6: LOOCV for the OLS and the ML estimation method ($n=100$).

Estimation method	Models	LOOCV
ordinary least squares (OLS)	Gaussian	0.881
	Exponential	0.999
	Spherical	1.032
maximum likelihood (ML)	Gaussian	0.574
	Exponential	0.586
	Spherical	0.566

Table 3.7: Parameters for the OLS and the ML estimation method ($n = 200$).

Estimation method	Models	Parameters		
		Nugget	Sill	Range
ordinary least squares (OLS)	Gaussian	0.666	1.342	1.122
	Exponential	0.612	1.618	1.631
	Spherical	0.613	1.161	1.034
maximum likelihood (ML)	Gaussian	0.222	1.158	3.780
	Exponential	0.000	1.590	4.787
	Spherical	0.027	1.595	2.880

Table 3.8: LOOCV for the OLS and the ML estimation method ($n = 200$).

Estimation method	Models	LOOCV
ordinary least squares (OLS)	Gaussian	0.634
	Exponential	0.494
	Spherical	0.728
maximum likelihood (ML)	Gaussian	0.402
	Exponential	0.387
	Spherical	0.398

Table 3.9: Parameters for the OLS and the ML estimation method ($n = 300$).

Estimation method	Models	Parameters		
		Nugget	Sill	Range
ordinary least squares (OLS)	Gaussian	0.567	1.440	1.194
	Exponential	0.474	2.016	1.852
	Spherical	0.468	1.208	1.024
maximum likelihood (ML)	Gaussian	0.255	0.865	3.269
	Exponential	0.022	1.622	4.901
	Spherical	0.026	1.534	2.805

Table 3.10: LOOCV for the OLS and the ML estimation method ($n = 300$).

Estimation method	Models	LOOCV
ordinary least squares (OLS)	Gaussian	0.581
	Exponential	0.375
	Spherical	0.518
maximum likelihood (ML)	Gaussian	0.324
	Exponential	0.326
	Spherical	0.324

4. Geostatistical Data Analysis with Outlier Detection

4.1 Introduction

Many researchers have used the variogram method to reduce the effect of outliers in spatial data analysis. Different approaches have been proposed to detect outliers. It is known that estimating the variogram after replacing outliers is more efficient. Nirel et al. (1998) proposed a method for removing spatial data, in which outliers are first detected and then replaced by values calculated from the remaining data. Two different methods (distributional inference method and deletion method) for detection of spatial outliers were proposed by Yoo and Um (1999). Sensitivity analysis, based on the influence functions for auto-and cross-variogram, was proposed by Choi et al. (2000). Kim and Jung (2005) proposed the outlier detection method in multivariate regression. Based on the sign of the influence function, Hayashi et al. (2013) proposed a new framework of statistical sensitivity analysis for linear discriminant analysis. Kim et al. (2013a, 2014a) focused on an estimation approach based on maximum likelihood method, and detected outliers with the sample influence function. On the other hand, geostatistical data analysis is sensitive to outliers. Figures 4.1 and 4.2 show the geostatistical data analysis based on the presence of outliers. We can see that the shape and parameter estimation are influenced by outliers. Moreover, this variogram cloud (Figure 4.1, Top) plot shows the bias and trend.

On the other hand, the maximum likelihood method based on the likelihood method is sensitive to outliers, and the parameters estimated by this method are affected by them. In this chapter, to achieve a stable analysis in variogram models based on the maximum likelihood method, we propose a procedure for stable geostatistical data analysis. Here, we detect outliers on the target dataset for geostatistical analysis with the sample influence function (SIF) for the Akaike information criterion (AIC) and the maximum likelihood method, and estimate the parameters by deleting them. We conduct a simulation study to demonstrate our procedure. For simplicity, we assume that the underlying process of the observed geostatistical data is stationary and isotropic.

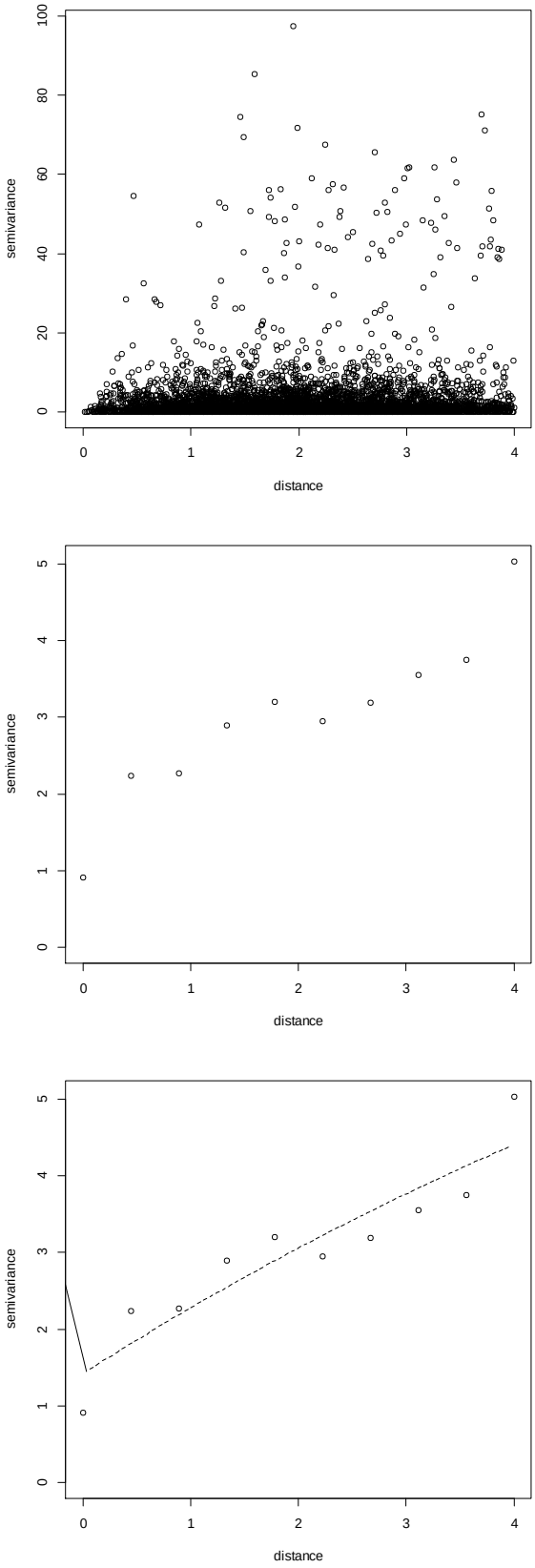


Figure 4.1: Outliers in geostatistical data; variogram cloud (Top), sample variogram (Center), fitting the theoretical variogram model (Bottom).

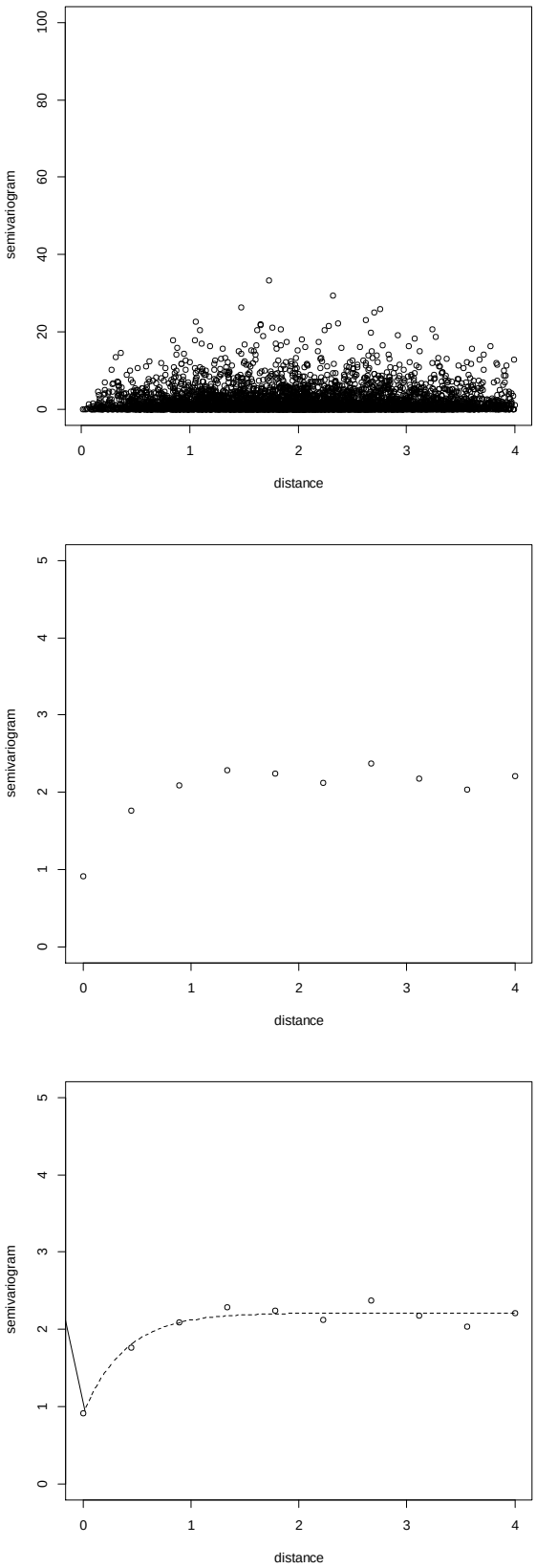


Figure 4.2: Non-outliers in geostatistical data; variogram cloud (Top), sample variogram (Center), fitting the theoretical variogram model (Bottom).

4.2 Sample Influence Functions for the Maximum Likelihood with the Akaike Information Criteria

Studies on detecting influential observations in spatial statistics have actively progressed in recent years. Gunst and Hartfield (1997) suggest influence function to quantify the effects of influential data values on the sample and robust variogram estimators. The influence function (IF) is a representative function for detecting outliers, introduced by Hampel (1974). From the definition of influence functions (Hampel, 1974; Hampel et al., 1986; Tanaka, 1994), the theoretical influence function (TIF) is given by

$$TIF(Z(\mathbf{s});\theta) = \lim_{\varepsilon \to 0} \frac{[\theta((1-\varepsilon)F + \varepsilon\delta_{Z(\mathbf{s})}) - \theta(F)]}{\varepsilon},$$

where $\delta_{Z(\mathbf{s})}$ is the cdf of a unit point mass at $Z(\mathbf{s})$ and $\theta = \theta(F)$ is a parameter which is expressed as a functional of the cumulative distribution function (cdf) F of random variables $Z(\mathbf{s})$. The TIF for θ is the derivative of the function $\theta(\varepsilon) \equiv \theta((1-\varepsilon)F + \varepsilon\delta_{Z(\mathbf{s})})$ with respect to ε evaluated at $\varepsilon = 0$. The empirical influence function (EIF) is obtained by replacing cdf $\hat{F}$ for F in the definition of the TIF. The EIF at the $Z(\mathbf{s}) = Z(\mathbf{s}_i)(i = 1, \cdots, n)$ is given by

$$EIF(Z(\mathbf{s}_i);\hat{\theta}) = \lim_{\varepsilon \to 0} \frac{[\theta((1-\varepsilon)\hat{F} + \varepsilon\delta_{Z(\mathbf{s}_i)}) - \theta(\hat{F})]}{\varepsilon}.$$

The sample influence function (SIF), which is obtained by omitting "lim" and setting $\varepsilon = 1/(n-1)$ in EIF, is expressed as

$$SIF(Z(\mathbf{s}_i);\hat{\theta}) = -(n-1)(\hat{\theta}_{(i)} - \hat{\theta}),$$

where the subscript (i) means the omission of the i-th individual.

The maximum likelihood method based on likelihood function is sensitive to outliers. And the goodness of fit can be achieved based on the maximum likelihood. A satisfactory compromise between goodness of fit and complexity of the model can be

achieved based on the AIC. Then, in this study, we use the SIF for the maximum likelihood method (Section 3.3) and the AIC.

The SIF for the maximum likelihood method is calculated as follows:

$$SIF\left(De;\hat{L}\right)=-\left(\frac{n-t}{t}\right)\left(\hat{L}_{(De)}-\hat{L}\right), \tag{4.1}$$

where $\hat{L}_{(De)}$ is the maximized log-likelihood $\hat{L}$ in the case of deleting De. Here, n is the number of all the target data. t is the number of the data that belong to De. De means the subset of target observations to be evaluated. AIC penalizes minus twice log-likelihood by twice the number of parameters (Akaike, 1974). For a given set of data the variable part of the AIC is estimated by

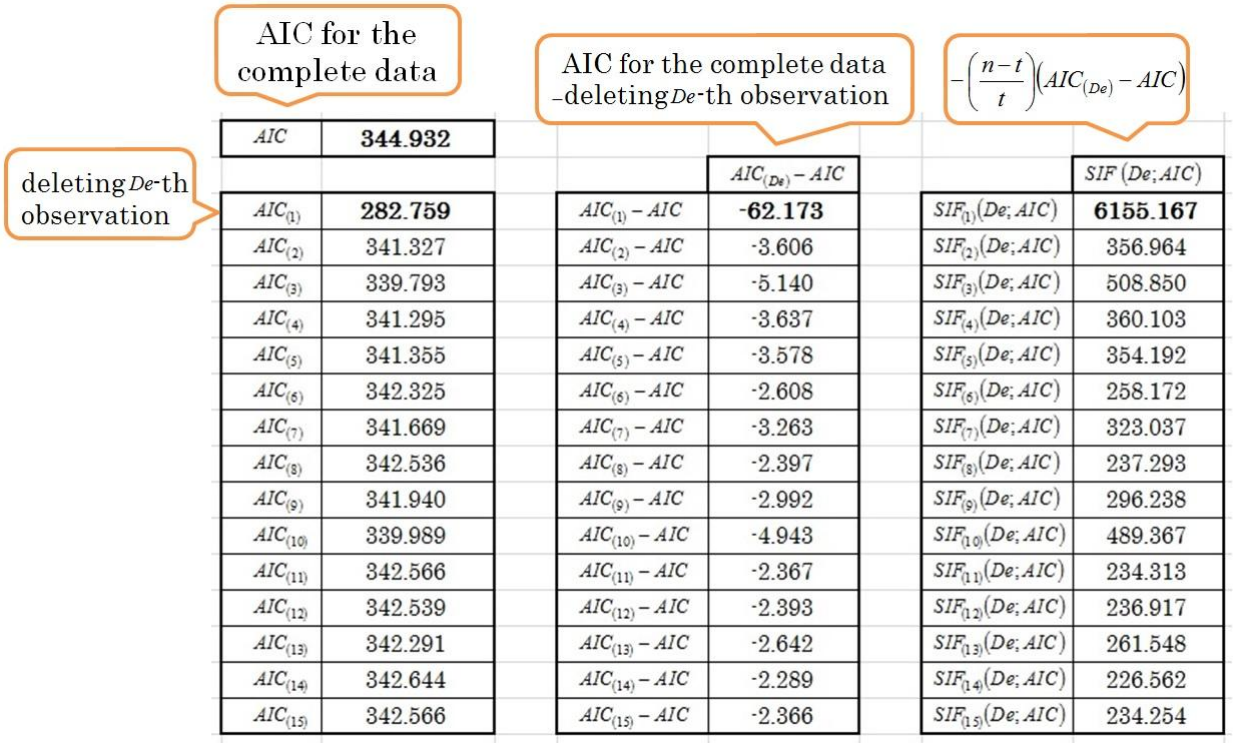
$$AIC:=-2\hat{L}+2p,$$

where AIC is the value of the Akaike Information Criteria, $\hat{L}$ is the maximized log-likelihood and p is the number of parameters in the model. The model with minimum AIC value is chosen as the best model to fit the data. In AIC, the compromise takes place between the maximized log-likelihood, i.e., $-2\hat{L}$ (the lack of fit component) and p, the number of free parameters estimated within the model (the penalty component) which is a measure of complexity or the compensation for the bias in the lack of fit when the maximum likelihood estimators are used.

The SIF for the AIC is calculated as follows:

$$SIF(De;AIC)=-\left(\frac{n-t}{t}\right)\left(AIC_{(De)}-AIC\right). \tag{4.2}$$

By using $SIF\left(De;\hat{L}\right)$, we can evaluate the influence of each observation for the fitting of all observed data through variogram models. On the other hand, with $SIF(De;AIC)$, we can assess the influence of each observation for the prediction on variogram estimations.



AIC	**344.932**
$AIC_{(1)}$	**282.759**
$AIC_{(2)}$	341.327
$AIC_{(3)}$	339.793
$AIC_{(4)}$	341.295
$AIC_{(5)}$	341.355
$AIC_{(6)}$	342.325
$AIC_{(7)}$	341.669
$AIC_{(8)}$	342.536
$AIC_{(9)}$	341.940
$AIC_{(10)}$	339.989
$AIC_{(11)}$	342.566
$AIC_{(12)}$	342.539
$AIC_{(13)}$	342.291
$AIC_{(14)}$	342.644
$AIC_{(15)}$	342.566

	$AIC_{(De)}-AIC$
$AIC_{(1)}-AIC$	**-62.173**
$AIC_{(2)}-AIC$	-3.606
$AIC_{(3)}-AIC$	-5.140
$AIC_{(4)}-AIC$	-3.637
$AIC_{(5)}-AIC$	-3.578
$AIC_{(6)}-AIC$	-2.608
$AIC_{(7)}-AIC$	-3.263
$AIC_{(8)}-AIC$	-2.397
$AIC_{(9)}-AIC$	-2.992
$AIC_{(10)}-AIC$	-4.943
$AIC_{(11)}-AIC$	-2.367
$AIC_{(12)}-AIC$	-2.393
$AIC_{(13)}-AIC$	-2.642
$AIC_{(14)}-AIC$	-2.289
$AIC_{(15)}-AIC$	-2.366

	$SIF\ (De;AIC)$
$SIF_{(1)}(De;AIC)$	**6155.167**
$SIF_{(2)}(De;AIC)$	356.964
$SIF_{(3)}(De;AIC)$	508.850
$SIF_{(4)}(De;AIC)$	360.103
$SIF_{(5)}(De;AIC)$	354.192
$SIF_{(6)}(De;AIC)$	258.172
$SIF_{(7)}(De;AIC)$	323.037
$SIF_{(8)}(De;AIC)$	237.293
$SIF_{(9)}(De;AIC)$	296.238
$SIF_{(10)}(De;AIC)$	489.367
$SIF_{(11)}(De;AIC)$	234.313
$SIF_{(12)}(De;AIC)$	236.917
$SIF_{(13)}(De;AIC)$	261.548
$SIF_{(14)}(De;AIC)$	226.562
$SIF_{(15)}(De;AIC)$	234.254

Figure 4.3: Explanation for the SIF statistic.

In general, the model with minimum AIC value is chosen as the best model to fit the data. In case of the difference between "AIC for the complete data (include the outlier) and deleting $De-th$ observation (exclude the outlier)" is relatively large. The outlier is defined to be the largest absolute SIF. We can see that the 1^{st} observation is an outlier from Figure 4.3.

4.3 Simulation Study

In this section, we present a simulation study to validate the proposed outlier detection method. We considered the Gaussian, exponential, and spherical models, which each model contains three parameters (sill, range, and nugget). We present results for the cases of including 1, 3, and 5 outliers. For each setting, we generate $n=100$ samples. We generated artificial outliers based on range of $\pm 3\sigma$ (Figure 4.4). The approach of outlier detection can be explained as below.

Step 1. Fix the three models (Gaussian, exponential, and spherical) and parameters (sill = 7, range = 2, and nugget = 1), and generate datasets including positions as well as data values.

Step 2. Generate outliers in the generated datasets.

Step 3. Calculate all possible combination subsets Des based on all data.

Step 4. For each De, calculate $-\left(\frac{n-t}{t}\right)\left(\hat{L}_{(De)}-\hat{L}\right)$ and $-\left(\frac{n-t}{t}\right)\left(AIC_{(De)}-AIC\right)$.

Step 5. Detect outliers based on the magnitude of $-\left(\frac{n-t}{t}\right)\left(\hat{L}_{(De)}-\hat{L}\right)$ and $-\left(\frac{n-t}{t}\right)\left(AIC_{(De)}-AIC\right)$.

The procedure is repeated 30 times. The simulation results for Gaussian, exponential, and spherical models, obtained by the above procedures are presented in Table 4.1 to Table 4.12. $SIF\left(De;\hat{L}\right)$ and $SIF(De;AIC)$ are SIFs for the maximized log-likelihood and AIC, respectively. Table 4.1 to Table 4.6 show the results of a single influential observation, respectively, when we used the evaluation measures were used in (4.1) and (4.2).

Table 4.7 to Table 4.12 shows the results in terms of the influence of multiple influential observations (in the case of 3 and 5 outliers). We can see that there is a substantial difference between the SIF in the case of containing outliers and that of containing non-outliers.

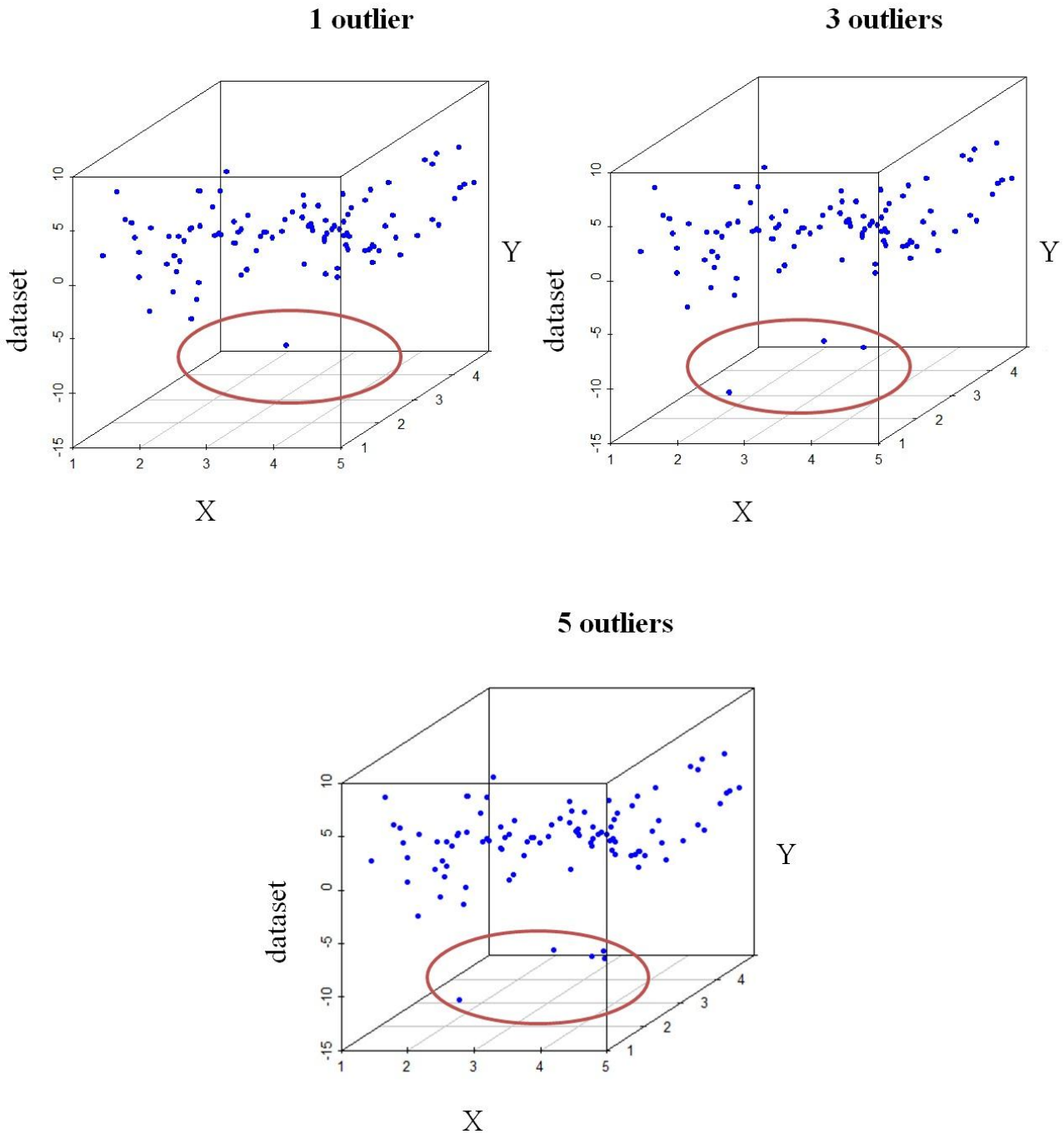


Figure 4.4: A geostatistical data set. Objects are located in the X − Y plane. The height of each vertical line segment represents the attribute value of each object.

Table 4.1: Results for $SIF\left(De;\hat{L}\right)$ in the case of the outlier and non-outlier (Gaussian model).

	1 Outlier	3 Outliers	5 Outliers
	$SIF\left(De;\hat{L}\right)$	$SIF\left(De;\hat{L}\right)$	$SIF\left(De;\hat{L}\right)$
Outlier	-1550.246	-1103.709 -992.690 -945.047	-947.016 -805.913 -767.517 -737.669 -723.356
Non Outlier	-207.660	-256.078	-246.994
	-204.465	-249.797	-238.029
	-196.926	-246.656	-236.382
	-155.034	-241.240	-230.085
	-191.329	-236.297	-226.869
	-197.667	-231.916	-222.665
	-194.580	-230.395	-221.569
	-138.446	-225.595	-217.950
	-176.744	-221.406	-215.449
	-173.056	-218.423	-212.358
	⋮	⋮	⋮
	-148.318	-251.791	-248.093
	-199.806	-259.146	-253.726
	-194.838	-263.678	-258.384
	-211.521	-268.644	-265.113
	-217.929	-277.112	-269.639
	-216.369	-282.477	-274.323
	-236.823	-294.675	-283.874
	-253.822	-313.781	-301.404
	-267.097	-330.142	-315.696
	-266.085	-366.129	-354.280

Table 4.2: Results for $SIF(De; AIC)$ in the case of the outlier and non-outlier (Gaussian model).

	1 Outlier	3 Outliers	5 Outliers
	$SIF(De; AIC)$	$SIF(De; AIC)$	$SIF(De; AIC)$
Outlier	3100.492	2207.419 1985.381 1890.096	1894.032 1611.827 1535.034 1475.339 1446.712
Non Outlier	415.318	512.156	493.990
	408.931	499.597	476.058
	393.852	493.310	472.765
	310.070	482.481	460.171
	382.658	472.594	453.739
	395.335	463.828	445.331
	389.160	460.789	443.140
	276.895	451.191	435.902
	353.488	442.811	430.899
	346.113	436.848	424.715
	⋮	⋮	⋮
	296.638	503.581	496.185
	399.614	518.291	507.450
	389.677	527.355	516.770
	423.044	537.288	530.227
	435.860	554.224	539.280
	432.739	564.953	548.645
	473.646	589.350	567.747
	507.644	627.561	602.809
	534.194	660.283	631.392
	532.170	732.260	708.561

Table 4.3: Results for $SIF(De;\hat{L})$ in the case of the outlier and non-outlier (Exponential model).

	1 Outlier	3 Outliers	5 Outliers
	$SIF(De;\hat{L})$	$SIF(De;\hat{L})$	$SIF(De;\hat{L})$
Outlier	-2327.431	-1101.638 -1129.866 -887.597	-798.507 -692.876 -723.429 -726.259 -671.705
Non Outlier	-573.663	-205.343	-225.909
	-578.364	-210.023	-216.817
	-574.626	-216.416	-235.635
	-562.442	-195.917	-201.167
	-572.470	-198.438	-207.059
	-570.208	-201.062	-229.982
	-589.079	-225.658	-231.092
	-569.979	-202.715	-206.217
	-570.123	-207.545	-240.823
	-562.781	-196.253	-204.780
	⋮	⋮	⋮
	-561.072	-193.451	-193.804
	-568.189	-194.224	-197.526
	-565.110	-199.571	-198.020
	-572.212	-200.049	-200.824
	-566.273	-197.374	-197.240
	-591.722	-208.614	-213.802
	-595.709	-221.195	-216.391
	-602.550	-220.079	-224.521
	-633.645	-247.609	-241.396
	-704.408	-303.931	-276.661

Table 4.4: Results for $SIF(De; AIC)$ in the case of the outlier and non-outlier (Exponential model).

	1 Outlier	3 Outliers	5 Outliers
	$SIF(De; AIC)$	$SIF(De; AIC)$	$SIF(De; AIC)$
Outlier	4654.860	2203.275 2259.731 1775.194	1597.015 1385.754 1446.858 1452.522 1343.413
Non Outlier	1147.324	410.682	451.819
	1156.727	420.046	433.635
	1149.251	432.828	471.273
	1124.883	391.831	402.334
	1144.939	396.876	414.120
	1140.414	402.123	459.965
	1178.158	451.316	462.189
	1139.958	405.428	412.435
	1140.245	415.089	481.648
	1125.562	392.504	409.561
	⋮	⋮	⋮
	1122.143	386.898	387.610
	1136.377	388.445	395.055
	1130.218	399.143	396.042
	1144.422	400.097	401.651
	1132.543	394.745	394.483
	1183.441	417.225	427.607
	1191.415	442.389	432.783
	1205.098	440.158	449.045
	1267.289	495.216	482.794
	1408.814	607.861	553.323

Table 4.5: Results for $SIF(De;\hat{L})$ in the case of the outlier and non-outlier (Spherical model).

	1 Outlier	3 Outliers	5 Outliers
	$SIF(De;\hat{L})$	$SIF(De;\hat{L})$	$SIF(De;\hat{L})$
Outlier	-1958.926	-1040.371 -982.605 -1039.146	-876.300 -745.267 -720.075 -626.589 -693.806
Non Outlier	-209.252	-263.882	-265.525
	-222.505	-222.999	-252.489
	-212.984	-224.582	-238.484
	-215.135	-272.985	-223.680
	-201.808	-214.853	-255.451
	-202.845	-234.939	-219.861
	-212.050	-224.854	-279.177
	-209.727	-210.315	-231.191
	-198.865	-205.723	-236.865
	-195.021	-211.764	-255.030
	⋮	⋮	⋮
	-191.960	-204.348	-197.839
	-197.245	-208.816	-202.360
	-200.160	-200.005	-213.463
	-205.149	-199.508	-219.184
	-198.632	-212.202	-206.626
	-213.813	-217.874	-216.825
	-228.823	-221.583	-231.347
	-225.182	-222.956	-227.483
	-260.644	-250.771	-247.148
	-305.593	-288.540	-282.782

Table 4.6: Results for $SIF(De; AIC)$ in the case of the outlier and non-outlier (Spherical model).

	1 Outlier	3 Outliers	5 Outliers
	$SIF(De; AIC)$	$SIF(De; AIC)$	$SIF(De; AIC)$
Outlier	3917.853	2080.743 1965.211 2078.291	1752.602 1490.537 1440.151 1253.179 1387.613
Non Outlier	418.504	527.763	531.050
	445.012	445.998	504.979
	425.968	449.164	476.967
	430.271	545.969	447.360
	403.613	429.705	510.906
	405.690	469.878	439.722
	424.100	449.707	558.353
	419.456	420.629	462.383
	397.731	411.446	473.731
	390.041	423.529	510.062
	⋮	⋮	⋮
	383.921	408.697	395.680
	394.490	417.634	404.722
	400.319	400.010	426.927
	410.297	399.016	438.368
	397.262	424.405	413.255
	427.626	435.745	433.650
	457.645	443.169	462.695
	450.366	445.912	454.968
	521.288	501.541	494.296
	611.185	577.080	565.564

Gaussian model

Table 4.7: *SIF* s in the case of the evaluation of multiple influential observations.

case of 3 outliers	$SIF\left(De;\hat{L}\right)$	$SIF(De;AIC)$
{1, 2}	-3095.229	6190.458
{1, 3}	-3081.245	6162.490
{2, 3}	-2964.817	5939.981
{1, 2, 3}	-3309.833	6635.346

Table 4.8: *SIF* s in the case of the evaluation of multiple influential observations.

case of 5 outliers	$SIF\left(De;\hat{L}\right)$	$SIF(De;AIC)$
{1, 2}	-599.506	1199.012
{1, 3}	-584.876	1169.752
{1, 4}	-549.517	1099.034
{1, 5}	-642.737	1285.475
{2, 3}	-627.181	1254.363
{2, 4}	-612.377	1224.756
{2, 5}	-510.083	1020.168
{3, 4}	-453.429	906.860
{3, 5}	-507.128	1014.257
{4, 5}	-320.269	640.539
{1, 2, 3}	-739.811	1479.624
{1, 2, 4}	-737.631	1475.264
{1, 2, 5}	-771.211	1542.425
{2, 3, 4}	-721.406	1442.812
{2, 3, 5}	-763.303	1526.606
⋮	⋮	⋮
{1, 2, 3, 4}	-830.237	1660.475
{1, 2, 3, 5}	-792.444	1584.890
{1, 2, 4, 5}	-780.469	1560.939
{1, 3, 4, 5}	-793.228	1586.457
{2, 3, 4, 5}	-805.083	1610.167
{1, 2, 3, 4, 5}	-888.203	1776.407

Exponential model

Table 4.9: *SIF* s in the case of the evaluation of multiple influential observations.

case of 3 outliers	$SIF\left(De;\hat{L}\right)$	$SIF(De;AIC)$
{1, 2}	-1262.089	2524.177
{1, 3}	-1147.409	2294.817
{2, 3}	-1110.905	2221.809
{1, 2, 3}	-1414.418	2828.834

Table 4.10: *SIF* s in the case of the evaluation of multiple influential observations.

case of 5 outliers	$SIF\left(De;\hat{L}\right)$	$SIF(De;AIC)$
{1, 2}	-819.641	1639.283
{1, 3}	-804.195	1608.391
{1, 4}	-813.437	1626.875
{1, 5}	-790.238	1580.477
{2, 3}	-792.844	1585.689
{2, 4}	-776.822	1553.643
{2, 5}	-775.938	1551.877
{3, 4}	-733.646	1467.292
{3, 5}	-729.219	1458.440
{4, 5}	-730.569	1461.138
{1, 2, 3}	-905.583	1811.167
{1, 2, 4}	-896.455	1792.910
{1, 2, 5}	-879.090	1758.180
{2, 3, 4}	-871.668	1743.335
{2, 3, 5}	-851.052	1702.104
⋮	⋮	⋮
{1, 2, 3, 4}	-988.596	1977.193
{1, 2, 3, 5}	-962.909	1925.817
{1, 2, 4, 5}	-974.337	1948.674
{1, 3, 4, 5}	-924.012	1848.024
{2, 3, 4, 5}	-923.020	1846.040
{1, 2, 3, 4, 5}	-1128.507	2257.013

Spherical model

Table 4.11: *SIF* s in the case of the evaluation of multiple influential observations.

case of 3 outliers	$SIF\left(De;\hat{L}\right)$	$SIF(De;AIC)$
{1, 2}	-1177.918	2355.836
{1, 3}	-1165.855	2331.711
{2, 3}	-1136.960	2273.921
{1, 2, 3}	-1403.828	2807.656

Table 4.12: *SIF* s in the case of the evaluation of multiple influential observations.

case of 5 outliers	$SIF\left(De;\hat{L}\right)$	$SIF(De;AIC)$
{1, 2}	-832.872	1665.744
{1, 3}	-793.049	1586.098
{1, 4}	-784.540	1569.081
{1, 5}	-819.103	1638.207
{2, 3}	-806.339	1612.680
{2, 4}	-780.239	1560.479
{2, 5}	-796.703	1593.406
{3, 4}	-800.198	1600.397
{3, 5}	-808.757	1617.515
{4, 5}	-794.171	1588.342
{1, 2, 3}	-899.321	1798.644
{1, 2, 4}	-882.913	1765.827
{1, 2, 5}	-906.797	1813.594
{2, 3, 4}	-862.908	1725.816
{2, 3, 5}	-867.170	1734.340
⋮	⋮	⋮
{1, 2, 3, 4}	-998.206	1996.411
{1, 2, 3, 5}	-991.303	1982.605
{1, 2, 4, 5}	-971.552	1943.104
{1, 3, 4, 5}	-958.872	1917.744
{2, 3, 4, 5}	-942.635	1885.271
{1, 2, 3, 4, 5}	-1141.654	2283.309

5. Real Data Analysis with Outlier Detection

5.1 Introduction

In this chapter, we apply the proposed method to real data based on the sample influence functions. These sample influence functions are derived for the geostatistical data analysis. A real numerical example is analyzed to show the validity or usefulness of the proposed sample influence functions.

5.2 Rainfall Data Analysis with Outlier Detection

5.2.1 Data

We applied the proposed method to real data. We focused on the daily maximum rainfall data from January 2010 to December 2012. We used a public dataset from the Japan Meteorological Agency website. In this chapter, we particularly considered 119 areas of Chugoku, Japan for data selection. The data contained 119 daily maximum rainfall observations collected by latitude and longitude planar coordinates (see Figures 5.1, 5.2 and Table 5.1).

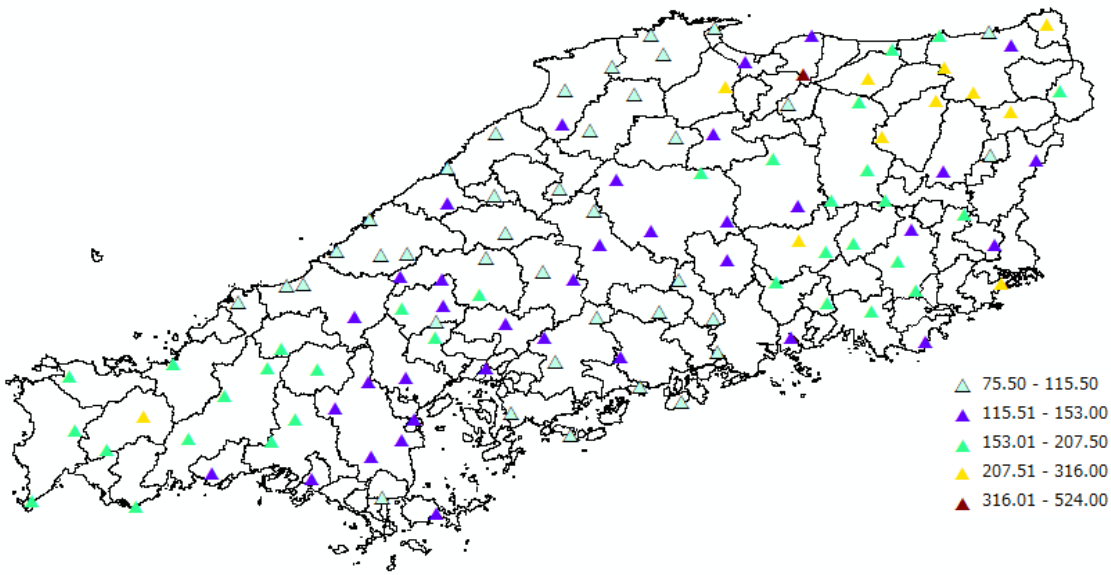


Figure 5.1: Map of Japan (Chugoku) showing 119 rainfall recording locations.

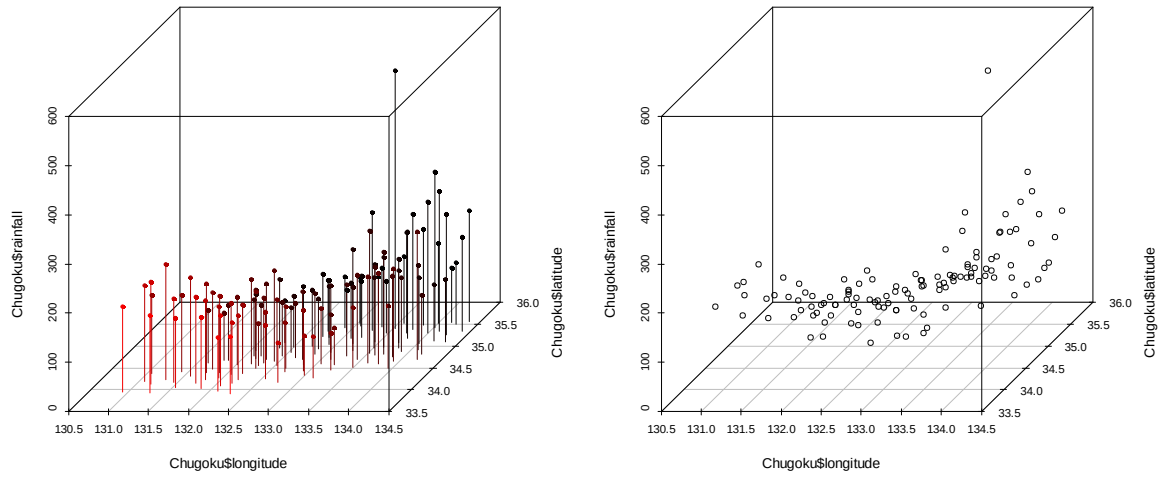


Figure 5.2: Plot of Japan (Chugoku) showing 119 rainfall recording locations.

Table 5.1: Daily maximum rainfall for the 119 areas of the Japan (Chugoku).

NO.	Station	Latitude	Longitude	Daily maximum rainfall (mm)
1	Imaoka	35.098	134.325	126.5
2	Kuse	35.068	133.753	170
3	Tsuyama	35.063	134.008	132.5
4	Niimi	34.943	133.518	144.5
5	Akaiwa	34.918	134.082	170.5
6	Jinyama	34.828	133.523	249.5
7	Fukuwatari	34.867	133.903	150
⋮	⋮	⋮	⋮	⋮
113	Iwakuni	34.155	132.178	123.5
114	Yanai	33.958	132.113	109
115	Rakanzan	34.350	132.063	145.5
116	Wada	34.148	131.735	176
117	Shinobu	34.303	131.577	201.5
118	Kano	34.225	131.815	160.5
119	Higashiatsu	34.118	131.182	207.5

5.2.2 Variogram Estimation

The resulting squared-differences variogram cloud is shown in Figure 5.3. Table 5.2 shows the values of variogram parameters in Gaussian, exponential, and spherical model for daily maximum rainfall data.

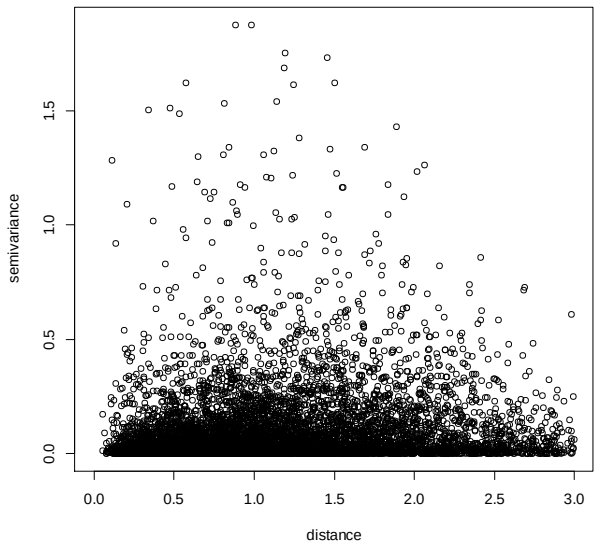


Figure 5.3: Variogram cloud for daily maximum rainfall data (119 areas).

A variogram cloud is the distribution of the variance between all pairs of point at all possible distances $\mathbf{h}$. The variogram cloud is a diagnostic tool that can be used in conjunction with boxplots to look for potential outliers or trends, and to assess variability with increasing distance (Kaluzny et al., 1996). This variogram cloud plot (Figure 5.3) and four plots (Figure 5.4) show the bias of the rainfall data. Therefore, with this plot, we can detect the potential outliers.

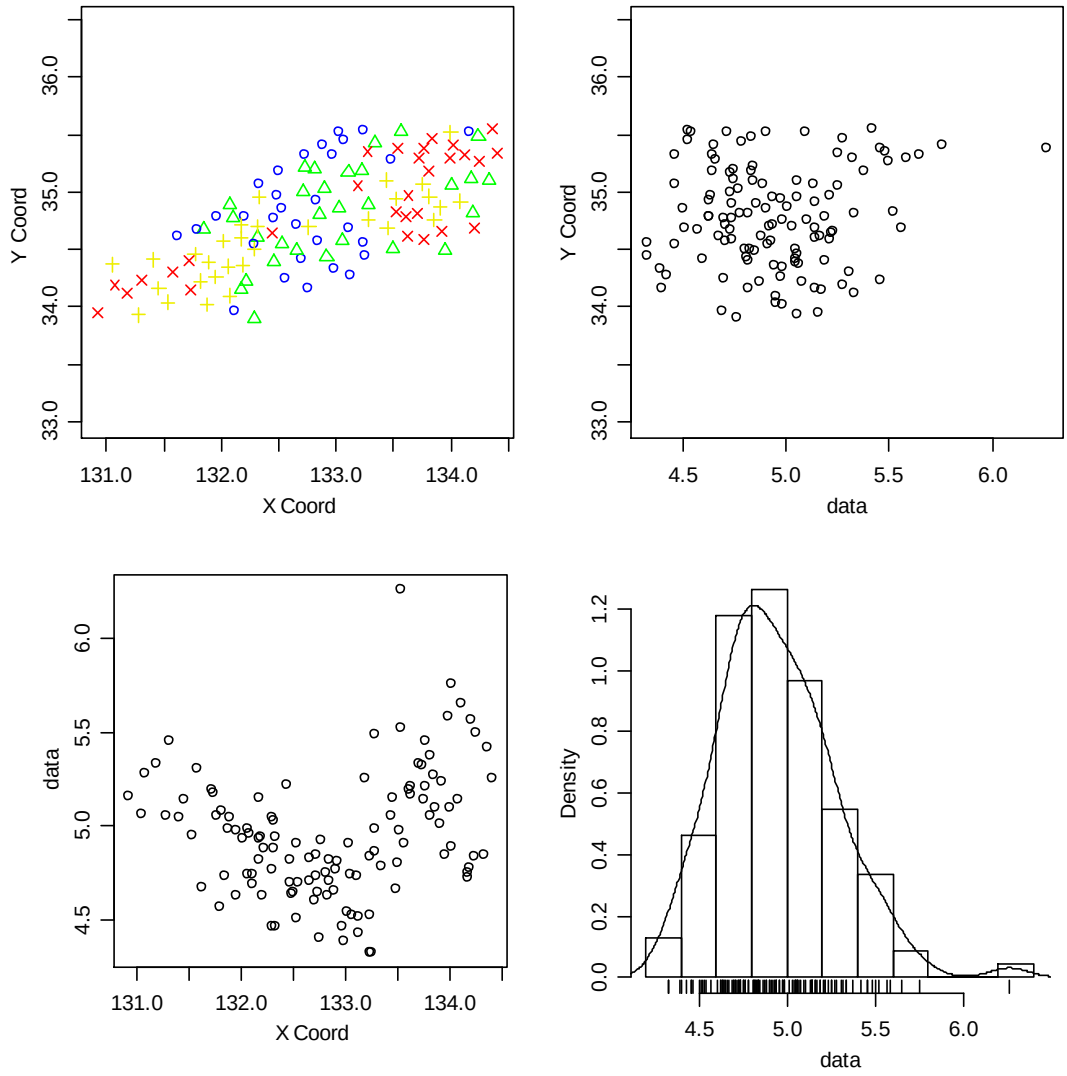


Figure 5.4: Plot for daily maximum rainfall data (119 areas); the observation locations assign different colors to data in different quartiles (Top left), the data against the Y coordinates (Top right), the data against the X coordinates (Bottom left), the histogram of the observation values (Bottom right).

Table 5.2: Variogram parameters for daily maximum rainfall data (119 areas).

Models / Parameters	Gaussian	Exponential	Spherical
Nugget	0.051	0.041	0.043
Sill	0.091	0.129	0.099
Range	0.842	1.181	1.762

5.2.3 Outlier Detection Using the Sample Influence Functions

Table 5.3 and Table 5.4 shows the influence of a single large influential observation for the maximized log-likelihood and AIC calculated from observed data based on SIF, respectively. From Table 5.3 and Table 5.4, we could regard the $12^{th}, 32^{nd}, 39^{th}, 57^{th}$, and 81^{st} observations as potential outliers. Table 5.5 and Table 5.6 shows the results in terms of the influence of multiple influential observations on the five observations (the $12^{th}, 32^{nd}, 39^{th}, 57^{th}$, and 81^{st} observations).

Table 5.3: $SIF(De;\hat{L})$ statistic for large influential data.

NO.	Models / Station	Gaussian	Exponential	Spherical
		$SIF(De;\hat{L})$	$SIF(De;\hat{L})$	$SIF(De;\hat{L})$
12	Mushiake	-332.23	-278.76	-369.79
32	Daisen	-1385.35	-1667.31	-1667.13
39	Ebi	-269.83	-404.26	-425.4
57	Hakuta	-284.51	-351.49	-386.38
81	Kurahashi	-147.55	-207.22	-246.18

Table 5.4: $SIF(De;AIC)$ statistic for large influential data.

NO.	Models / Station	Gaussian	Exponential	Spherical
		$SIF(De;AIC)$	$SIF(De;AIC)$	$SIF(De;AIC)$
12	Mushiake	664.47	557.52	739.58
32	Daisen	2770.69	3334.62	3334.26
39	Ebi	539.66	808.51	850.8
57	Hakuta	569.02	702.98	772.75
81	Kurahashi	295.1	414.44	492.36

Table 5.5: $SIF(De;\hat{L})$ in the case of the evaluation of multiple influential observations.

Models / Subset of De	Gaussian	Exponential	Spherical
	$SIF(De;\hat{L})$	$SIF(De;\hat{L})$	$SIF(De;\hat{L})$
{32, 39}	-782.11	-905.4	-916.59
{32, 57}	-1116.24	-1167.59	-1164.04
{32, 12}	-939.84	-1028.93	-1050.51
{32, 81}	-859.33	-1032.22	-1010.73
{39, 57}	-282.35	-370.41	-369.34
{39, 12}	-321.14	-362.78	-399.2
{39, 81}	-258.32	-348.34	-342.56
{57, 12}	-328.73	-326.18	-367.08
{57, 81}	-270.04	-295.37	-305.08
{12, 81}	-297.22	-251.34	-294.21
{32, 39, 57}	-760.87	-800.06	-805
{32, 39, 12}	-686.74	-736.11	-761.96
{32, 39, 81}	-633.27	-737.89	-733.68
{32, 57, 12}	-896.12	-926.76	-944.48
{32, 57, 81}	-835.08	-950.93	-927.26
{32, 12, 81}	-733.8	-835.57	-831.77
{39, 57, 12}	-316.32	-361.76	-379.37
{39, 57, 81}	-267	-365.97	-343.81
{39, 12, 81}	-302.17	-349.33	-362.82
{57, 12, 81}	-310.44	-304.28	-326.97
{32, 39, 57, 12}	-687.08	-711.87	-732.95
{32, 39, 57, 81}	-657.65	-728.79	-729.19
{32, 39, 12, 81}	-614.71	-665.97	-686.27
{32, 57, 12, 81}	-768.75	-843.12	-851.34
{39, 57, 12, 81}	-317.05	-371.41	-378.48
{32, 39, 57, 12, 81}	-618.88	-687.69	-694.49

Table 5.6: $SIF(De; AIC)$ in the case of the evaluation of multiple influential observations.

Models / Subset of De	Gaussian	Exponential	Spherical
	$SIF(De; AIC)$	$SIF(De; AIC)$	$SIF(De; AIC)$
{32, 39}	1564.23	1810.8	1833.18
{32, 57}	2232.49	2335.19	2328.08
{32, 12}	1879.67	2057.87	2101.01
{32, 81}	1718.67	2064.44	2021.47
{39, 57}	564.7	740.83	738.68
{39, 12}	642.28	725.56	798.4
{39, 81}	516.64	696.68	685.12
{57, 12}	657.47	652.36	734.16
{57, 81}	540.08	590.74	610.16
{12, 81}	594.44	502.67	588.42
{32, 39, 57}	1521.75	1600.13	1610
{32, 39, 12}	1373.48	1472.22	1523.93
{32, 39, 81}	1266.54	1475.77	1467.36
{32, 57, 12}	1792.24	1853.52	1888.97
{32, 57, 81}	1670.16	1901.86	1854.51
{32, 12, 81}	1467.6	1671.15	1663.53
{39, 57, 12}	632.64	723.52	758.75
{39, 57, 81}	534.01	731.94	687.63
{39, 12, 81}	604.34	698.66	725.64
{57, 12, 81}	620.88	608.55	653.94
{32, 39, 57, 12}	1374.17	1423.74	1465.91
{32, 39, 57, 81}	1285.28	1457.57	1437.45
{32, 39, 12, 81}	1199.39	1331.94	1351.6
{32, 57, 12, 81}	1507.47	1686.24	1681.75
{39, 57, 12, 81}	604.07	742.83	736.03
{32, 39, 57, 12, 81}	1237.75	1375.38	1388.98

5.2.4 Variogram Estimation and Outlier

Based on the value of the *SIF* s in Table 5.3 to Table 5.6, we can see that the 32^{nd} observation corresponds to a large influential outlier. We show the variogram cloud with four plots by removing the 32^{nd} observation that is an outlier (see Figures 5.5, 5.6).

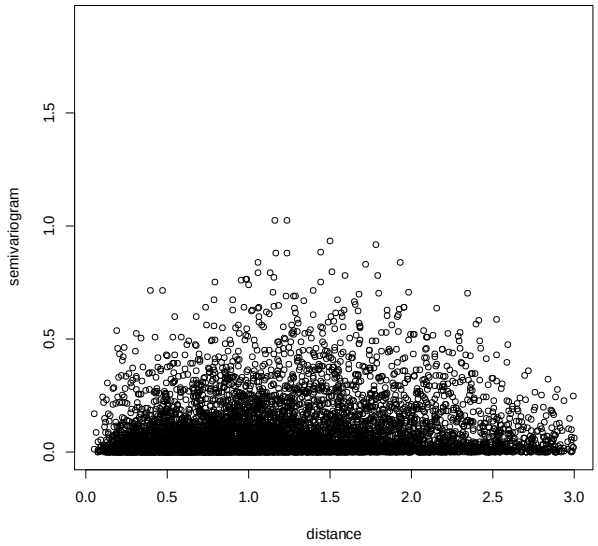


Figure 5.5: Variogram cloud for daily maximum rainfall data (118 areas).

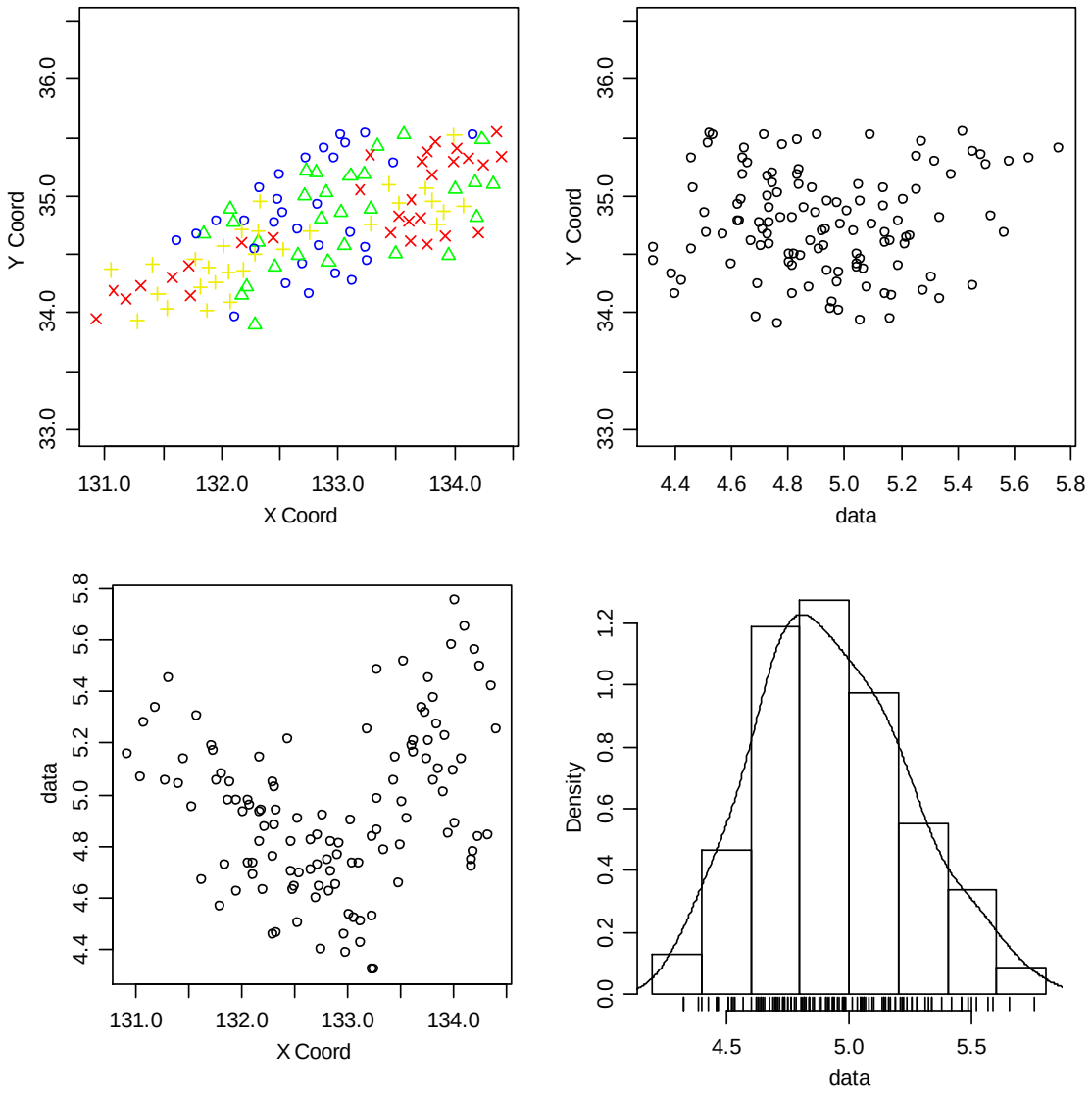


Figure 5.6: Plot for daily maximum rainfall data (118 areas); the observation locations assign different colors to data in different quartiles (Top left), the data against the Y coordinates (Top right), the data against the X coordinates (Bottom left), the histogram of the observation values (Bottom right).

When we deleted the 32^{nd} observation for the target dataset, we got results of the variogram parameters in the three models as Table 5.7.

Table 5.7: Variogram parameters for daily maximum rainfall data (118 areas).

Models / Parameters	Gaussian	Exponential	Spherical
Nugget	0.045	0.039	0.038
Sill	0.081	0.127	0.087
Range	0.906	1.551	1.854

5.2.5 Results of Kriging

We carried out a kriging to investigate the influence of the outliers. To perform the comparison of the kriging prediction before the deletion of the 32^{nd} observation and that of the kriging after deletion of the observation, we used the mean squared errors (MSE).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{Z}_i^* - Z_i)^2,$$

where $\hat{Z}_i^*$ and Z_i represent the predicted and observed values, respectively.

The results of the three comparisons are shown in Table 5.8. Based on these results, we removed the 32^{nd} observation and enhanced the performance of the prediction in terms of the kriging method from the point of view of MSE.

Table 5.8: Results of MSE for kriging.

Areas / Models	Before removed outliers (119 areas)	After removed outliers (118 areas)
Gaussian	0.0508	0.0492
Exponential	0.0525	0.0487
Spherical	0.0508	0.0462

6. Conclusions

The variogram plays a central role when analyzing spatial data. A valid variogram model must first be selected, and the parameters of the model estimated before kriging (spatial prediction) is performed. In this paper, we focused on the number of lag and the outlier detection.

The firstly, we examined the performance of a variogram estimator in spatial models, focusing on a piecewise constant estimator for an isotropic variogram. We proposed a method for selecting the optimal number for the estimator using leave-one-out cross-validation (LOOCV) and Akaike information criterion (AIC) in the geostatistical data analysis. The usefulness of the proposed method was illustrated through a simulation study. Moreover, we compared the estimated parameters of the variogram models based on ordinary least square method with that based on maximum likelihood estimation.

Each estimation method, we investigated prediction performance with exponential and spherical models. As a result, the parameter estimation methods based on maximum likelihood estimation gave a better performance than ordinary least square method from the point of view of leave-one-out cross-validation (LOOCV). In the future, we have to apply our method for finding the optimal number of lag to many real geostatistical data analysis.

Secondly, we focused on an outlier detection approach based on the maximum likelihood method and the Akaike information criterion (AIC) with the sample influence functions (SIF). In the simulation study, we artificially generated a few of large influential data as outliers (single and multiple influential observations). Under this condition, we could detect outliers based on our proposed procedure. Moreover, in the case study of the daily maximum rainfall data, we could also detect outliers through our method.

In both studies, by comparing the value of mean squared errors (MSE) before deleting outliers with that of mean squared errors (MSE) after deleting outliers, we investigated the performance of the prediction in Gaussian, exponential, and spherical

models. We then gave the zero weight to the detected outliers and confirmed that the performance of the prediction on the kriging method was improved from the point of view of mean squared errors (MSE). Through a simulation study and case study, we were able to confirm the usefulness of our proposed method. In the future, to validate the performance of our method in details, we will have to perform additional simulation studies. Also, we need to apply our approach for detecting outliers to other real geostatistical data analyses.

Appendix

A. Data

: Daily maximum rainfall for the 119 areas of the Japan (Chugoku district).

NO.	Station	Latitude	Longitude	Daily maximum rainfall (mm)
1	Imaoka	35.098	134.325	126.5
2	Kuse	35.068	133.753	170
3	Tsuyama	35.063	134.008	132.5
4	Niimi	34.943	133.518	144.5
5	Akaiwa	34.918	134.082	170.5
6	Jinyama	34.828	133.523	249.5
7	Fukuwatari	34.867	133.903	150
8	Wake	34.815	134.183	119
9	Saya	34.685	133.445	171.5
10	Yakage	34.617	133.618	174.5
11	Okayama	34.660	133.917	187
12	Mushiake	34.682	134.207	260.5
13	Kurashiki	34.590	133.768	183.5
14	Tamano	34.487	133.950	127.5
15	Kasaoka	34.502	133.495	122
16	Shimoazae	34.965	133.628	183
17	Takaha	34.792	133.610	179.5
18	Nagi	35.112	134.170	115.5
19	Kaminagata	35.297	133.725	204.5
20	Chiya	35.103	133.435	156.5
21	Onbara	35.300	133.987	266
22	Nichioji	34.757	133.855	164
23	Tomi	35.178	133.805	216
24	Kibichuo	34.817	133.705	207.5

25	Asahinishi	34.962	133.812	157
26	Sakai	35.543	133.235	92.5
27	Aoya	35.520	133.997	163
28	Tottori	35.487	134.238	126
29	Iwai	35.558	134.360	225.5
30	Yonago	35.433	133.338	119.5
31	Kurayoshi	35.473	133.838	195
32	Daisen	35.388	133.537	524
33	Chizu	35.263	134.240	244.5
34	Sekigane	35.378	133.757	234
35	Wakasa	35.333	134.405	191
36	Shiotsu	35.523	133.567	135
37	Chaya	35.187	133.230	126
38	Saji	35.328	134.113	284.5
39	Ebi	35.288	133.483	105.5
40	Shikano	35.413	134.017	316
41	Koyama	35.530	134.165	112
42	Kashima	35.520	133.022	93.5
43	Hikawa	35.413	132.890	104.5
44	Matsue	35.457	133.065	92
45	Daito	35.318	132.965	86.5
46	Ota	35.190	132.497	104
47	Kakeya	35.197	132.815	115.5
48	Yokota	35.173	133.103	113.5
49	Kawamoto	34.977	132.492	103
50	Hamada	34.897	132.070	114
51	Misumi	34.788	131.958	102
52	Masuda	34.677	131.843	113
53	Tsuwano	34.462	131.770	157
54	Sada	35.222	132.723	126.5
55	Sakurae	34.953	132.333	139.5
56	Mizuho	34.853	132.530	90.5

57	Hakuta	35.350	133.273	241
58	Haza	34.780	132.197	102.5
59	Hikimi	34.572	132.017	138.5
60	Izumo	35.332	132.730	104
61	Akana	35.002	132.712	113
62	Yasaka	34.777	132.108	114
63	Fukumitsu	35.070	132.333	87
64	Takatsu	34.675	131.790	96.5
65	Yoshika	34.392	131.893	155.5
66	Dogoyama	35.057	133.188	191
67	Miyoshi	34.812	132.850	123.5
68	Shobara	34.860	133.023	134.5
69	Oasa	34.768	132.463	110
70	Kake	34.610	132.320	132
71	Joge	34.693	133.117	91
72	Uchiguroyama	34.597	132.177	171.5
73	Sera	34.583	133.050	114
74	Higashihiroshima	34.417	132.700	99.5
75	Fukuyama	34.447	133.247	75.5
76	Hiroshima	34.398	132.462	123.5
77	Takehara	34.330	132.982	80.5
78	Ikuchishima	34.278	133.123	83.5
79	Otake	34.222	132.220	131
80	Kure	34.240	132.550	109.5
81	Kurahashi	34.550	132.293	86.5
82	Takano	35.033	132.902	117.5
83	Dongcheng	34.895	133.277	129
84	Koda	34.695	132.760	137
85	Miiri	34.545	132.530	135.5
86	Fuchu	34.562	133.232	75.5
87	Shiwa	34.498	132.660	124.5
88	Odomari	34.698	132.312	153

89	Yawata	34.708	132.173	139
90	Yuki	34.763	133.278	146.5
91	Hatsukaichitsuda	34.365	132.190	140
92	Hongo	34.435	132.918	122.5
93	Saekiyuki	34.498	132.290	155.5
94	Tsushimi	34.647	132.440	184.5
95	Kimita	34.928	132.830	102
96	Midori	34.722	132.655	111
97	Asuka	34.567	132.838	110.5
98	Kuresikamagari	34.165	132.748	81.5
99	Susa	34.615	131.623	107
100	Hagi	34.410	131.405	155
101	Tokusa	34.398	131.725	179.5
102	Akiyoshidai	34.235	131.307	234
103	Hirose	34.262	131.952	145
104	Toyota	34.187	131.073	196
105	Yamaguchi	34.160	131.455	171
106	Hofu	34.040	131.533	141.5
107	Kudamatsu	34.020	131.873	145.5
108	Shimonoseki	33.948	130.925	174
109	Ube	33.930	131.278	157
110	Agenosho	33.903	132.293	117
111	Kuga	34.095	132.075	142
112	Yuya	34.370	131.055	158.5
113	Iwakuni	34.155	132.178	123.5
114	Yanai	33.958	132.113	109
115	Rakanzan	34.350	132.063	145.5
116	Wada	34.148	131.735	176
117	Shinobu	34.303	131.577	201.5
118	Kano	34.225	131.815	160.5
119	Higashiatsu	34.118	131.182	207.5

B. Results of the SIF statistic.

B.1 Sample influence function for the maximum likelihood method.

NO.	Station	Gaussian $SIF(De;\hat{L})$	Exponential $SIF(De;\hat{L})$	Spherical $SIF(De;\hat{L})$
1	Imaoka	-16.40	-19.57	-37.44
2	Kuse	31.01	47.84	-0.50
3	Tsuyama	-69.88	-22.34	-33.62
4	Niimi	16.77	13.12	19.16
5	Akaiwa	53.07	40.48	48.87
6	Jinyama	-144.59	-128.32	-137.01
7	Fukuwatari	22.49	40.62	-9.20
8	Wake	-49.32	-75.95	-79.18
9	Saya	10.64	23.48	-22.09
10	Yakage	25.63	41.44	35.24
11	Okayama	32.41	32.52	32.83
12	Mushiake	-332.23	-278.76	-369.79
13	Kurashiki	22.15	31.55	27.54
14	Tamano	-6.51	-12.06	-5.50
15	Kasaoka	48.44	42.17	45.09
16	Shimoazae	54.68	52.90	53.56
17	Takaha	49.10	55.76	55.26
18	Nagi	-135.26	-110.66	-132.14
19	Kaminagata	54.14	53.67	54.13
20	Chiya	44.41	49.58	49.82
21	Onbara	-43.35	-9.94	-16.71
22	Nichioji	52.35	52.55	52.74
23	Tomi	49.82	39.44	41.76
24	Kibichuo	22.28	24.34	22.22
25	Asahinishi	18.47	42.63	36.11
26	Sakai	-61.65	-58.46	-68.03

27	Aoya	17.25	20.41	16.86
28	Tottori	-89.42	-71.21	-86.23
29	Iwai	-71.62	-89.37	-157.32
30	Yonago	6.85	-10.88	-5.01
31	Kurayoshi	51.45	50.10	50.94
32	Daisen	-1385.35	-1667.31	-1667.13
33	Chizu	-81.69	-64.66	-117.31
34	Sekigane	29.64	35.51	31.60
35	Wakasa	16.72	32.74	-25.91
36	Shiotsu	-29.28	-71.65	-54.25
37	Chaya	39.14	33.37	37.04
38	Saji	-140.74	-93.42	-139.03
39	Ebi	-269.83	-404.26	-425.40
40	Shikano	-205.73	-181.48	-182.42
41	Koyama	-238.06	-183.80	-209.12
42	Kashima	29.69	34.34	31.04
43	Hikawa	51.08	51.07	52.19
44	Matsue	8.74	21.81	14.48
45	Daito	-10.17	-10.28	-13.11
46	Ota	49.51	45.85	48.32
47	Kakeya	48.57	51.91	51.77
48	Yokota	40.82	36.49	34.24
49	Kawamoto	43.39	49.24	4.72
50	Hamada	50.52	49.65	50.96
51	Misumi	31.69	31.82	24.37
52	Masuda	41.97	51.67	46.42
53	Tsuwano	46.62	48.88	50.27
54	Sada	13.30	27.08	21.78
55	Sakurae	5.07	-15.55	-11.60
56	Mizuho	-26.51	-8.67	-8.10
57	Hakuta	-284.51	-351.49	-386.38
58	Haza	24.50	15.37	21.65

59	Hikimi	53.24	50.82	52.25
60	Izumo	50.04	48.29	50.24
61	Akana	54.32	51.89	52.98
62	Yasaka	51.66	53.31	7.32
63	Fukumitsu	-8.91	-19.15	-5.47
64	Takatsu	-26.25	-7.55	-25.47
65	Yoshika	50.37	51.15	51.16
66	Dogoyama	-66.94	-83.66	-76.00
67	Miyoshi	49.66	49.38	49.82
68	Shobara	40.60	38.27	39.60
69	Oasa	42.86	43.16	-1.89
70	Kake	55.27	58.03	56.90
71	Joge	-10.49	-4.55	-11.23
72	Uchiguroyama	-21.40	-24.63	-64.61
73	Sera	48.33	40.37	45.27
74	Higashihiroshima	44.95	37.99	41.64
75	Fukuyama	-65.96	-37.47	-49.66
76	Hiroshima	54.13	51.80	52.74
77	Takehara	26.81	11.21	14.48
78	Ikuchishima	37.89	29.59	32.02
79	Otake	53.44	53.30	53.49
80	Kure	50.45	47.63	49.47
81	Kurahashi	-147.55	-207.22	-246.18
82	Takano	54.04	52.46	53.32
83	Dongcheng	43.47	37.17	36.88
84	Koda	20.88	22.69	21.16
85	Miiri	44.29	48.13	44.96
86	Fuchu	-118.33	-77.95	-94.43
87	Shiwa	46.21	48.60	45.09
88	Odomari	15.83	20.41	-24.29
89	Yawata	43.31	46.18	44.01
90	Yuki	42.77	36.11	39.35

91	Hatsukaichitsuda	54.26	52.44	52.65
92	Hongo	-0.16	5.31	3.22
93	Saekiyuki	28.15	13.42	15.15
94	Tsushimi	-104.41	-107.22	-111.42
95	Kimita	36.69	33.74	32.26
96	Midori	50.03	47.67	50.41
97	Asuka	53.81	52.38	53.55
98	Kuresikamagari	5.53	-14.37	-46.59
99	Susa	-11.24	-8.98	-21.79
100	Hagi	46.37	41.88	41.95
101	Tokusa	21.54	28.96	-16.35
102	Akiyoshidai	-33.61	-36.91	-39.93
103	Hirose	54.14	53.23	53.57
104	Toyota	37.63	37.76	-6.08
105	Yamaguchi	51.47	50.91	51.63
106	Hofu	10.52	19.93	13.41
107	Kudamatsu	50.01	47.42	48.33
108	Shimonoseki	35.36	33.74	35.23
109	Ube	33.04	36.36	32.34
110	Agenosho	43.27	41.02	43.35
111	Kuga	51.28	47.58	49.33
112	Yuya	34.85	35.74	35.42
113	Iwakuni	47.02	50.72	48.46
114	Yanai	14.14	15.22	-31.25
115	Rakanzan	54.13	52.85	53.23
116	Wada	41.82	40.61	-2.67
117	Shinobu	7.02	8.47	-33.27
118	Kano	52.49	53.41	53.79
119	Higashiatsu	28.17	28.96	27.94

B.2 Sample influence function for the Akaike information criterion.

NO.	Station	Gaussian $SIF(De; AIC)$	Exponential $SIF(De; AIC)$	Spherical $SIF(De; AIC)$
1	Imaoka	32.80	39.14	74.89
2	Kuse	-62.02	-95.69	1.00
3	Tsuyama	139.77	44.68	67.24
4	Niimi	-33.54	-26.24	-38.32
5	Akaiwa	-106.14	-80.96	-97.75
6	Jinyama	289.19	256.63	274.01
7	Fukuwatari	-44.98	-81.23	18.39
8	Wake	98.65	151.91	158.36
9	Saya	-21.28	-46.96	44.19
10	Yakage	-51.26	-82.87	-70.49
11	Okayama	-64.81	-65.05	-65.66
12	Mushiake	664.47	557.52	739.58
13	Kurashiki	-44.29	-63.10	-55.08
14	Tamano	13.03	24.13	10.99
15	Kasaoka	-96.87	-84.33	-90.17
16	Shimoazae	-109.35	-105.81	-107.12
17	Takaha	-98.20	-111.52	-110.53
18	Nagi	270.52	221.33	264.29
19	Kaminagata	-108.28	-107.34	-108.25
20	Chiya	-88.82	-99.16	-99.64
21	Onbara	86.71	19.87	33.41
22	Nichioji	-104.70	-105.11	-105.48
23	Tomi	-99.63	-78.89	-83.53
24	Kibichuo	-44.56	-48.68	-44.44
25	Asahinishi	-36.94	-85.26	-72.23
26	Sakai	123.30	116.91	136.06
27	Aoya	-34.50	-40.81	-33.72
28	Tottori	178.84	142.41	172.45

29	Iwai	143.23	178.75	314.65
30	Yonago	-13.70	21.77	10.02
31	Kurayoshi	-102.90	-100.19	-101.89
32	Daisen	2770.69	3334.62	3334.26
33	Chizu	163.39	129.31	234.62
34	Sekigane	-59.28	-71.02	-63.20
35	Wakasa	-33.44	-65.48	51.81
36	Shiotsu	58.56	143.30	108.49
37	Chaya	-78.27	-66.73	-74.08
38	Saji	281.48	186.84	278.06
39	Ebi	539.66	808.51	850.80
40	Shikano	411.45	362.97	364.84
41	Koyama	476.12	367.61	418.23
42	Kashima	-59.38	-68.68	-62.09
43	Hikawa	-102.15	-102.14	-104.38
44	Matsue	-17.47	-43.62	-28.96
45	Daito	20.34	20.56	26.21
46	Ota	-99.02	-91.71	-96.64
47	Kakeya	-97.13	-103.81	-103.54
48	Yokota	-81.64	-72.99	-68.48
49	Kawamoto	-86.77	-98.48	-9.44
50	Hamada	-101.04	-99.30	-101.93
51	Misumi	-63.38	-63.64	-48.75
52	Masuda	-83.95	-103.34	-92.83
53	Tsuwano	-93.24	-97.76	-100.53
54	Sada	-26.60	-54.15	-43.55
55	Sakurae	-10.14	31.09	23.20
56	Mizuho	53.02	17.33	16.20
57	Hakuta	569.02	702.98	772.75
58	Haza	-48.99	-30.74	-43.30
59	Hikimi	-106.48	-101.65	-104.50
60	Izumo	-100.08	-96.58	-100.48

61	Akana	-108.64	-103.77	-105.97
62	Yasaka	-103.33	-106.63	-14.64
63	Fukumitsu	17.81	38.30	10.94
64	Takatsu	52.49	15.09	50.94
65	Yoshika	-100.74	-102.30	-102.32
66	Dogoyama	133.88	167.32	152.00
67	Miyoshi	-99.31	-98.75	-99.64
68	Shobara	-81.20	-76.55	-79.21
69	Oasa	-85.71	-86.31	3.79
70	Kake	-110.53	-116.06	-113.80
71	Joge	20.97	9.10	22.45
72	Uchiguroyama	42.80	49.26	129.21
73	Sera	-96.65	-80.74	-90.55
74	Higashihiroshima	-89.90	-75.98	-83.29
75	Fukuyama	131.92	74.94	99.33
76	Hiroshima	-108.26	-103.61	-105.47
77	Takehara	-53.62	-22.41	-28.97
78	Ikuchishima	-75.78	-59.18	-64.04
79	Otake	-106.87	-106.61	-106.98
80	Kure	-100.90	-95.25	-98.94
81	Kurahashi	295.10	414.44	492.36
82	Takano	-108.07	-104.92	-106.65
83	Dongcheng	-86.94	-74.34	-73.76
84	Koda	-41.76	-45.39	-42.32
85	Miiri	-88.58	-96.26	-89.92
86	Fuchu	236.66	155.90	188.85
87	Shiwa	-92.43	-97.19	-90.18
88	Odomari	-31.66	-40.82	48.58
89	Yawata	-86.62	-92.36	-88.01
90	Yuki	-85.54	-72.22	-78.69
91	Hatsukaichitsuda	-108.51	-104.88	-105.31
92	Hongo	0.32	-10.62	-6.44

93	Saekiyuki	-56.29	-26.85	-30.30
94	Tsushimi	208.81	214.44	222.83
95	Kimita	-73.39	-67.47	-64.52
96	Midori	-100.06	-95.34	-100.82
97	Asuka	-107.61	-104.75	-107.10
98	Kuresikamagari	-11.06	28.74	93.18
99	Susa	22.48	17.96	43.57
100	Hagi	-92.74	-83.77	-83.90
101	Tokusa	-43.09	-57.93	32.71
102	Akiyoshidai	67.23	73.83	79.85
103	Hirose	-108.29	-106.47	-107.15
104	Toyota	-75.26	-75.52	12.16
105	Yamaguchi	-102.94	-101.82	-103.27
106	Hofu	-21.04	-39.85	-26.82
107	Kudamatsu	-100.01	-94.83	-96.66
108	Shimonoseki	-70.72	-67.47	-70.46
109	Ube	-66.07	-72.72	-64.69
110	Agenosho	-86.54	-82.04	-86.71
111	Kuga	-102.57	-95.17	-98.65
112	Yuya	-69.69	-71.49	-70.84
113	Iwakuni	-94.03	-101.43	-96.92
114	Yanai	-28.28	-30.44	62.50
115	Rakanzan	-108.26	-105.70	-106.47
116	Wada	-83.64	-81.22	5.34
117	Shinobu	-14.03	-16.94	66.54
118	Kano	-104.97	-106.82	-107.58
119	Higashiatsu	-56.34	-57.93	-55.88

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**(6), 716–723.

Choi, S. B., Kim, K. K. and Tanaka, Y. (2000). Sensitivity analysis in auto– and cross–variogram estimation. *Journal of the Korean Data Analysis Society*, **2**(1), 91–107.

Choi, S. B., Kang, C. W. and Cho, J. S. (2010). Data–dependent choice of optimal number of lags in variogram estimation, *The Korean Journal of Applied Statistics*, **23**(3), 609–619.

Cressie, N. (1985). Fitting variogram models by weighted least square. *Journal of the International Association for Mathematical Geology*, **17**(5), 563–586.

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

Devijver, P. A. and Kittler, J. (1982). Pattern recognition: A statistical approach. Englewood Cliffs, Prentice Hall.

Genton, M. G. (1998). Variogram fitting by generalized least squares using an explicit formula for the covariance structure. *Mathematical Geology*, **30**(4), 323–345.

Gunst, R. F. and Hartfield, M. I. (1997). *Robust Semivariogram Estimation in the Presence of Influential Spatial Data Values.* Springer–Verlag, New York.

Hample, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, **69**(346), 383–393.

Hampel, F. R. and Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Function*. John Wiley & Sons, Inc, New York.

Hayashi, K., Ishioka F., Ueda, T., Suito, H. and Kurihara, K. (2013). A detection of influential individuals for the risk prediction of endoleak formation after TEVAR based on a new framework of statistical sensitivity analysis. *Bulletin of the Computational Statistics of Japan*, **26**(2), 59–77.

Hong, C. G. and Kim, Y. H. (2004). On optimal number of lags in variogram estimation in spatial data analysis. Journal of the Korean Data Analysis Society, **6**(1), 39–69.

Isaaks, E. H. and Srivastava, R. M. (1989). *An Introduction to Applied Geostatistics*. Oxford University Press, Oxford.

Istok, J. D. and Cooper, R. M. (1988). Geostatistics applied to groundwater pollution. III: Global estimates. *Journal of Environmental Engineering*, **114**(4), 915–928.

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

Journel, A. G. (1984). New ways of assessing spatial distributions of pollutants. In *Environmental sampling for Hazardous Waters*, G. Schweitzer(Ed.), American Chemical Society, Washington, 109–118.

Kaluzny, S. P., Vega, S. C., Cardoso, T. P. and Shelly, A. A. (1996). *S + Spatialstats: User's Manual Version 1.0*. MathSoft Inc., Washington.

Kim, M. G. and Jung, K. M. (2005). Detection of outliers in multivariate regression using plug–in method. *Journal of the Korean Data Analysis Society*, **7**(4), 1117–1124.

Kim, S. J., Hayashi, K. and Kurihara, K. (2013a). Parameter estimation for semivariogram of spatial data with outliers. *Proceedings of Joint Meeting of the IASC Satellite Conference for the 59th ISI WSC and the 8th Asian Regional Section of the IASC*, 415–421.

Kim, S. J., Kurihara, K. and Choi, S. B. (2013b). Reliability analysis in variogram estimation of spatial data. *Proceedings of the The 27nd Symposia of Japanese Society of Computational Statistics*, 69–72.

Kim, S. J., Hayashi, K. and Kurihara, K. (2014a). Geostatistical data analysis with outlier detection. *Journal of the Korean Data Analysis Society*, **16**(5), 2285–2297.

Kim, S. J., Hayashi, K. and Kurihara, K. (2014b). The optimal number of lags in variogram estimation in spatial data analysis. *Proceedings of COMPSTAT 2014 21st International Conference on Computational Statistics hosting the 5th IASC World Conference*, In: Gilli, M., Gonzalez-Rodriguez, G. and Nieto-Reyes, A.(Eds.), CD–ROM, 507–514.

Kitanidis, P. K. (1983). Statistical estimation of polynomial generalized covariance function and hydrologic application. *Water Resources Research*, **19**(4), 909–921.

Kitanidis, P. K. (1987). Parametric estimation of covariances of regionalized variables. *Water Resources Bulletin*, **23**(4), 557–567.

Lamorey, G. and Jacobson, E. (1995). Estimation of semivariogram parameters and evaluation of the effects of data sparsity, *Mathematical Geology*, **27**(3), 327–358.

Mardia, K. V. and Marshall, R. J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, **71**(1), 135–146.

Matheron, G. (1962). Traite de geostatistique appliquee, Tome **I**. *Memoires du Bureau de Recherches Geologigues et minieres*, **No. 14.** Editions Technip, Paris.

Matheron, G. (1963). Principles of geostatistics. *Economic geology*, **58**(8), 1246–1266.

Matheron, G. (1971). The theory of regionalized variables and its application. *Cahiers du Centre de Morphologie Mathematique*, **No. 5**. Fontainebleau, France.

Myers, D. E. (1989). Borden field data and multivariate geostatistics. In Hydraulic Engineering, M. A. Ports(Ed.), *American Society of Civil Engineering*, New York, 795–800.

Nirel, R., Mugglestone, M. A. and Barnett, V. (1998). Outlier–robust spectral estimation for spatial lattice processes. *Communications in Statistics: Theory and Methods*, **27**(12), 3095–3111.

Piazza, A., Menozzi, P. and Cavalli–Sforza, L. (1981). *The making and testing of geographic gene frequency maps*. International Biometric Society, **37**(4), 635–659.

SAS Institute Inc. (1999). *SAS/STAT User's Guide, Version 8.* SAS Institute, Cary, NC.

Tanaka, Y. (1994). Recent advance in sensitivity analysis in multivariate statistical methods. *Journal of the Japanese Society of Computational Statistics*, **7**(1), 1–25.

Tanimura, S. (2010). *Geospatial Data Analysis*. Kyoritsu Shuppan Co., Ltd. Tokyo.

Wackernagel, H. (1995). *Multivariate Geostatistics*. Springer. Germany.

Watson, G. S. (1972). Trend surface analysis and spatial correlation. *Geological Society of America, Special Paper*, **146**, 39–46.

Webster, R. (1985). Quantitative spatial analysis of soil in the field. In *Advances in Soil Science*, B. A. Stewart(Ed.), Springer–Verlag, New York, **3**(1), 1–70.

Yoo, S. M. and Um, I. (1999). On the estimation of semivariogram and spatial outliers with rainfall intensity data. *The Korean Journal of Applied Statistics*, **12**(1), 125–141.

http://www.jma.go.jp/jma/index.html

Acknowledgements

First of all, I express my deepest sense of gratitude to my respected supervisor, Professor Koji Kurihara of the Graduate School of Environmental and Life Science, Okayama University, for his constant help and guidance throughout my research work. His great knowledge and experience in research field acted as a torch to my research work.

I would like to extend my heartfelt thanks to Professor Wataru Sakamoto, Associate Professor Kaoru Fueda of the Graduate School of Environmental and Life Science and Professor Masaya Iizuka of the Admission Center, Okayama University, for their encouragement and assistance.

I would also like to extend my heartfelt thanks to Professor Seungbae Choi, Professor Kyukon Kim, Professor Manki Kang and Professor Changwan Kang of the Department of Data Information Science, Dongeui University and Professor Sungho Moon of the Department of Data Management, Busan University of Foreign Studies.

I would like to thank Dr. Myungjin Na, Dr. Fumio Ishioka and Dr. Kuniyoshi Hayashi for his most helpful input and advice. This research could not have been completed without their suggestions, contributions and hard work. They inspiration, keen interest and constant motivation throughout my research work helped me achieve a high level of success.

I am very thankful to my friends, seniors, juniors and colleagues for their encouragement and support.

Last but not the least, to my family; they are the reasons for what I am doing right now. No words can truly express my deepest gratitude and appreciation for the help and moral support given by each and every member of my family especially my parents.

Finally I would like to extend my heartfelt thanks to all of you.