

University of Nevada

Reno

✓  
STUDIES OF THE EFFECT OF DATA TRANSFORMATION  
ON THE KRIGING ESTIMATION

A thesis submitted in partial fulfillment of the  
requirements for the degree of Master of  
Science in Geology

by

Ikjung Nam  
!!!

James R. Carr/Thesis Advisor

Copyright March 1991

ALL RIGHTS RESERVED

000-0000

MINES  
LIBRARY

Tijssis

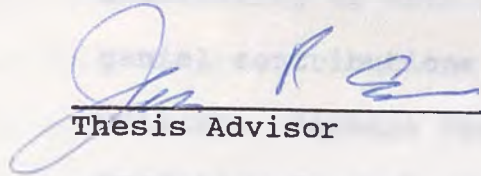
2783

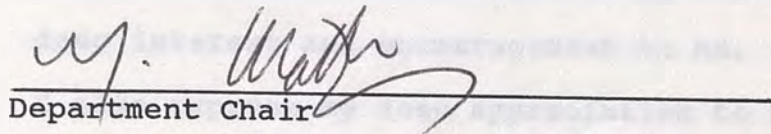
Copyright by Ikjung Nam 1991

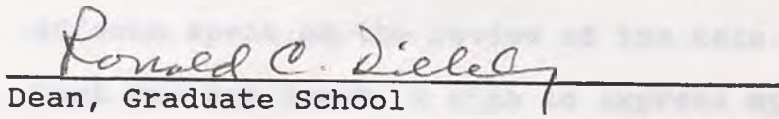
All Right Reserved

UNR LIBRARY

The thesis of Ikjung Nam is approved:

  
Thesis Advisor

  
Department Chair

  
Dean, Graduate School

University of Nevada

Reno

March 1991

## ACKNOWLEDGMENTS

I am most deeply indebted to Professor Dr. James R. Carr for his invaluable encouragement, kind guidance the undertaking of this thesis from time to time, and his genial contributions of time for helping my English writing. Perhaps the most important contribution Professor James R. Carr made to my thesis has been his deep interest and encouragement on me.

I also express my deep appreciation to Professor Dr. Robert D. Davis and Dr. Liang-chi Hsu for their time and effects spent on the review of the this paper.

Last but not least, I wish to express my deepest gratitude to God and my parents.

## ABSTRACT

The objective of this paper is to study the necessity and effect of data transformation to normality on kriging estimation and the applicability of using Hermite polynomials for data transfer function. For this purpose, two data sets are used and mean error, error variance, the result of cross-validation and scatter diagrams are compared for transformed data with untransformed data. The results show that the improvement in accuracy of kriging estimation by data transformation is more effective for highly skewed data. This means that data transformation is prerequisite step for highly skewed data on kriging estimation.

## Table of contents

CHAPTER	PAGE
ABSTRACT	
I. INTRODUCTION-----	1
II. METHODS-----	3
1. Data Sets	
2. Data Transformation	
3. Structural analysis	
4. Cross-validation of Kriging	
III. RESULTS AND COMPARISON-----	25
1. Raw Data and Transformed Data	
2. Variograms	
3. Cross-validation	
4. Untransformed Data and Retransformed Data	
5. Scatter Diagrams	
IV. DISCUSSION-----	48
V. CONCLUSION-----	50
VI. APPENDIX-----	52
VII. BIBLIOGRAPHY-----	53

## I. INTRODUCTION

The use of "kriging" in geostatistics has come to be synonymous with "optimal prediction" of values at given spatial locations from observations taken at known nearby locations. The various kriging techniques are all based on the linear unbiased estimator model  $Z^*(X_0) = \lambda_1 Z(X_1) + \lambda_2 Z(X_2) + \lambda_3 Z(X_3) + \dots + \lambda_n Z(X_n)$  where  $Z^*(X_0)$  is the estimator of true value  $Z(X_0)$  at location  $X_0$  and the  $\lambda_i$  are the weights allotted to each observation  $Z(X_i)$ . The weights are chosen such that the error associated with the estimator is less than error of any other linear sum, such that the weights  $\lambda_i$  minimize the estimation variance subject to the non-bias condition  $\sum \lambda_i = 1$ .

Kriging also assumes that the data are normally distributed because, with kriging,  $E[Z^*(X_0)] = m$ , where  $m$  is the mean. The best possible estimation of any unknown variable  $Z(X_0)$  is based on this theoretical premise, and the best estimator can be obtained from the normally distributed variables  $Z(X_i)$ . Linear estimation by kriging is not optimal for many geological data sets because many geological data are non-normal; in fact, one of the more common distributions is skewed with a long tail to the right. Based on this observation, the stationary random variables (geological data) need to be transformed for more accurate estimation into stationary centered gaussian random variables, that is, distributed normally, before kriging is performed. After this

process, the transformed variables are kriged or estimated and then estimated results are finally retransformed to original units. The widely used transfer function for geological data is the common logarithm whereby skewed distributions are more or less normalized by taking logarithms.

The objective of this paper is to study the necessity and effect of data transformation to normality for kriging by comparison between results of kriging of untransformed data and transformed data, drawn from two geological data sets. The comparison is based on mean error, error variance, the results of cross-validation, and scatter diagrams. A linear combination of Hermite polynomials, instead of common logarithms, is used for the transfer function in this study. The study is also aimed at investigating the applicability of using Hermite polynomials as transfer functions for geological data in the estimation by kriging.

## II. METHODS

The following steps are applied to two geological data sets and establish the goals of this study:

1. Raw data of the original sample values are transformed to another set of values that are normally distributed. The transformation function  $\phi$  is the expansion of Hermite polynomials and expresses the variable  $Z(\mathbf{X})$  in terms of centered gaussian variables (the standard normal distribution)  $Y(\mathbf{X})$ , i.e.,  $Y(\mathbf{X}) = \phi Z(\mathbf{X})$ , where  $Y$  has a mean of 0 and variance of 1.

2. The variogram of raw data and transformed data are computed for estimating spatial relationships, a basic requirement for any kriging.

3. Cross-validation by kriging is used as the basis for comparison of the accuracy of kriging estimation of raw data and transformed data because cross-validation is an appropriate tool for testing estimation methods.

4. The values of cross-validation of transformed data are inverted (retransformed) by  $Z(\mathbf{X}) = \phi^{-1} [Y(\mathbf{X})]$ , such that the comparison of the results of cross-validation of raw data and retransformed data is made in practice for the effect of data transformation on the kriging. The scatter diagrams are used for a visual test of accuracy. The main steps in this method are represented schematically in Figure 1 and are given in detail by the following section

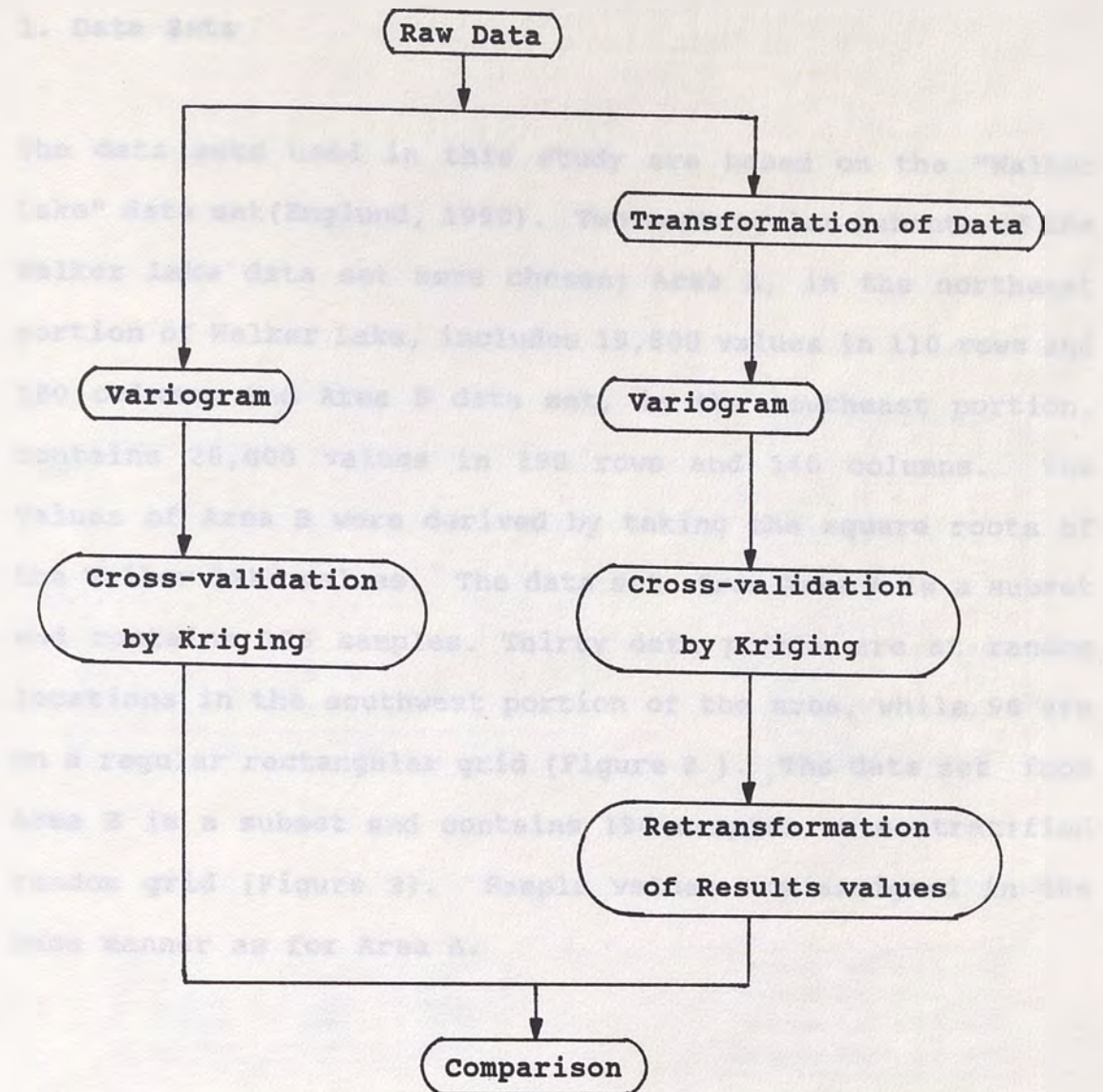
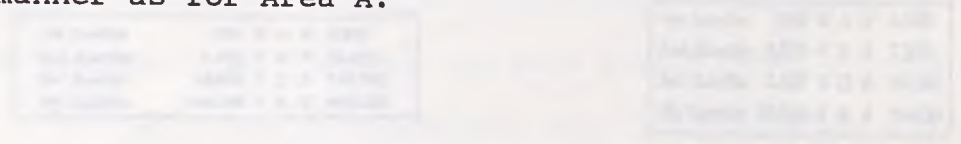


Figure 1. Steps of the method applied to two geological data sets.

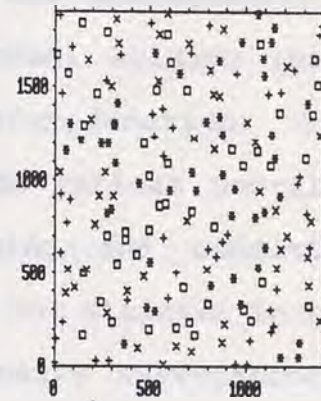
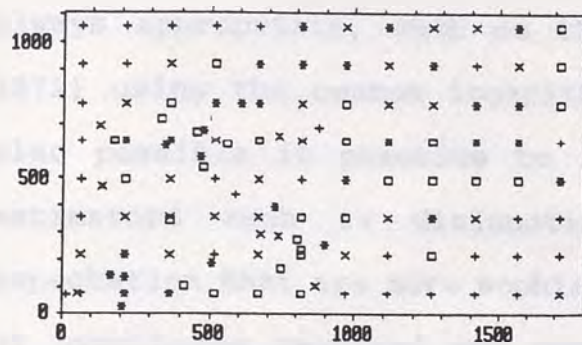
### 1. Data Sets

The data sets used in this study are based on the "Walker Lake" data set (Englund, 1990). Two rectangular subsets of the Walker Lake data set were chosen; Area A, in the northeast portion of Walker Lake, includes 19,800 values in 110 rows and 180 columns and Area B data set, in the southeast portion, contains 26,600 values in 190 rows and 140 columns. The values of Area B were derived by taking the square roots of the Walker Lake values. The data set from Area A is a subset and contains 126 samples. Thirty data points are at random locations in the southwest portion of the area, while 96 are on a regular rectangular grid (Figure 2). The data set from Area B is a subset and contains 190 samples on a stratified random grid (Figure 2). Sample values are assigned in the same manner as for Area A.



(A) Walker Lake A (B) Walker Lake B

Figure 2. Map of Walker Lake showing the locations of the data sets (Englund, 1990).



1st Quartile:	$.000 \leq + \leq 5.380$
2nd Quartile:	$5.380 < x \leq 33.600$
3rd Quartile:	$33.600 < \square \leq 146.090$
4th Quartile:	$146.090 < * \leq 992.200$

1st Quartile:	$.000 \leq + \leq 3.230$
2nd Quartile:	$3.230 < x \leq 7.330$
3rd Quartile:	$7.330 < \square \leq 16.290$
4th Quartile:	$16.290 < * \leq 54.830$

(A)

(B)

Figure 2. Map of locations of the data sets

(England, 1990).

A) Data set A    B) Data set B

## 2. Data Transformation

Among kriging estimators, there are methods based on data transformation to normality using transfer functions, in order to satisfy the assumption of linear kriging estimation and to get more accurate results of estimation. Some of these methods are naive and easy to apply but consequently not always appropriate, such as the lognormal kriging (David, 1972) using the common logarithm transfer function. It is also possible in practice to infer the various non-linear estimators such as disjunctive kriging and conditional expectation that are more sophisticated but stricter in terms of hypotheses required and computationally heavy (Matheron, 1976).

The method presented in this study is the simple linear kriging estimator of gaussian transformed variable  $Y(X)$  that is transformed from original raw data  $Z(X)$  by  $Y(X) = \phi [Z(X)]$ , where  $\phi$  is the transfer function. The transfer function used for data transformation is some order of expansion of Hermite polynomials.

In considering the Hermite polynomial transfer function, let two data sets be represented by a stationary variable of random function  $Z$ , with a specific observation represented as  $Z(X)$ . As noted by Matheron (1976, a), it is always possible to find a non-decreasing transformation  $\phi$  and another stationary R.F. (random function)  $Y$  that is normally distributed analog of

Z and also has a mean of 0 and a variance of 1. The variables in R.F. Z and Y are related through a function :

$$Y(X) = \phi [Z(X)]$$

in which  $\phi$  is written as a linear combination of Hermite polynomials (Journal and Huijbregts, 1978):

$$\phi (y) = \sum_{k=0}^{\infty} C_k H_k(y)$$

, where  $H_k(y)$  is a Hermite polynomial of order  $k$  and  $C_k$  are coefficients of this expansion calculated from data values.  $H_k(y)$  is given by Rodrigues' formula:

$$H_k(y) = (-1)^k e^{[y^2/2]} (d / dy )^k e^{-[y^2/2]}$$

and  $H_k(y)$  is easily calculated using the following recurrence relationship:

$$H_{k+1}(y) = yH_k(y) - kH_{k-1}(y)$$

, where  $H_0(y) = 1$ , and  $H_1(y) = y$ . With respect to the standard normal distribution, the Hermite polynomials are orthogonal, i.e.,

$$\int_{-\infty}^{\infty} [H_k(y)H_{k'}(y) e^{-[y^2/2]} / \sqrt{(2\pi)}] dy = 0 \quad \text{for } k \neq k'$$

$$\int_{-\infty}^{\infty} [H_k(y)H_k(y) e^{-[y^2/2]} / \sqrt{(2\pi)}] dy = k! \quad \text{for all } k$$

Orthogonality also implies that

$$C_k = 1/k! \int_{-\infty}^{\infty} [ \phi (y)H_k(y) \exp(-y^2/2) / \sqrt{(2\pi)}] dy$$

The coefficients  $C_k$ ,  $k=1, \dots$ , can be calculated through Hermite

integration:

$$C_k = \frac{1}{k!|2\pi|} \sum_{i=1}^I w_i e^{(y_i)^2/2} \phi(y_i) H_k(y_i)$$

where the  $y_i$ 's are the abscissas and the  $w_i$ 's are corresponding weight factors. Values of  $y_i$  and  $w_i$  are tabulated by Abramowitz and Stegun(1970) in Table 1. To calculate  $C_k$  using Hermite integration, it is necessary to determine the values of  $(y_i)$  for specific values of  $y_i$ . By definition, is the function which relates the data variables  $x_i$  to the corresponding standard normal variate  $y_i$ . A  $y_i$  corresponding to a given  $x_i$  is calculated from the cumulative distribution of  $x$ . This procedure can be shown graphically on Figure 3. Given the values of  $y_i$ , the corresponding values of  $Y_i = \phi(z_i)$  are following the inverse procedure; i.e., from  $y$  and the cumulative frequencies of standard normal distribution, a value  $p_i$  is calculated through tables or an approximation. This value of  $p_i$  is used to interpolate between the values of  $Z$  that bracket  $p_i$ , to give  $\phi(y_i)$ . From a table of the cumulative distribution, a value of  $y_i$  is found for each  $P_i$ . It also can be shown that  $C_0$  is the mean of the theoretical distribution of  $Z$  and the sum of  $C_k(k=1,2,\dots, k-1)$  is the variance of this distribution.

In practice, the expansion of the Hermite polynomial is carried out over  $k$  steps

$$Y(X) = \sum_{k=0}^{k-1} C_k H_k[Y(X)]$$

the  $k$  steps applied for this study are 5, 10, and 15, such that each value of data  $X_i$  is transformed to a value  $Y_i$  with coordinates  $C_0H_0(y)$ ,  $C_1H_1(y)$ , ..... ,  $C_{k-1}H_{k-1}(y)$  for each  $k$  step. Two data sets, having respectively 126 values and 190 values, are used in this study. Each variable of data sets is transformed by the above transfer function with 5, 10, and 15 coefficients into normal distribution.

Table 1. Abscissas and weight factors for Hermite integration.

$\int_{-\infty}^{\infty} e^{-x^2} f(x) dx = \sum_{i=1}^n w_i f(x_i)$ Abscissas = $\pm x_i$ (Zeros of Hermite Polynomials)				$\int_{-\infty}^{\infty} g(x) dx = \sum_{i=1}^n w_i e^{x_i^2} g(x_i)$ Weight Factors = $w_i$			
$\pm x_i$	$w_i$	$w_i e^{x_i^2}$	$n$	$\pm x_i$	$w_i$	$w_i e^{x_i^2}$	$n$
<b>n=2</b>				<b>n=10</b>			
0.70710 67811 86548	(-1)8.86226 92545 28	1.46114 11826 611		0.34290 13272 23705	(-1)6.10852 63373 53	0.68708 18539 513	
<b>n=3</b>				<b>n=12</b>			
0.09000 00000 00000	(0)1.18163 59006 04	1.18163 59006 037		0.94778 83912 40164	(-1)2.60492 31026 42	0.63962 12320 203	
1.22474 48713 91589	(-1)2.95408 97515 09	1.32393 11752 136		1.59768 26351 52605	(-2)5.16079 85615 88	0.66266 27732 669	
<b>n=4</b>				<b>n=16</b>			
0.52464 76232 75290	(-1)8.04914 09000 55	1.05996 44828 950		0.27348 10461 3815	(-1)5.07929 47901 66	0.54737 52050 378	
1.65068 01238 85785	(-2)8.13128 35447 25	1.24022 58176 958		0.82295 14491 4466	(-1)2.80647 45852 85	0.55244 19573 475	
<b>n=5</b>				<b>n=20</b>			
0.00000 00000 00000	(-1)9.45308 72048 29	0.94530 87204 829		0.73743 37285 454	(-1)2.86675 50536 28	0.49384 33852 721	
0.95857 24646 11819	(-1)3.93519 32315 22	0.98658 09967 514		1.23407 62153 953	(-1)1.09017 20602 00	0.49992 08713 363	
2.02018 28704 56086	(-2)1.99532 42059 05	1.18148 86255 360		1.73853 77121 166	(-2)2.48105 20887 44	0.50967 90271 175	
<b>n=6</b>				<b>n=20</b>			
0.43607 74119 27617	(-1)7.24629 59522 44	0.87640 13344 362		2.25497 40020 893	(-3)3.24377 33422 38	0.52408 01509 406	
1.39584 90740 13697	(-1)1.57067 32032 29	0.93558 05576 312		2.78880 60584 281	(-4)2.28338 63601 63	0.54485 17421 844	
2.35060 49736 74492	(-3)4.53000 99055 09	1.13690 83326 745		3.34785 45613 832	(-6)7.80255 64785 32	0.57526 24428 529	
<b>n=7</b>				<b>n=20</b>			
0.00000 00000 00000	(-1)8.10264 61755 68	0.81026 46175 568		3.94476 40401 156	(-7)1.08606 93701 69	0.62227 86961 914	
0.81628 78828 58965	(-1)4.25607 25261 01	0.82868 73032 836		4.60368 24495 507	(-10)4.39934 09922 73	0.70433 29611 748	
1.67355 16287 67471	(-2)5.45155 82819 13	0.89718 46002 252		5.38748 08900 112	(-13)2.22939 36455 34	0.89859 19614 532	
2.65196 13568 35233	(-4)9.71781 24509 95	1.10133 07296 103					
<b>n=8</b>				<b>n=20</b>			
0.38118 69902 07322	(-1)6.61147 01255 82	0.76454 41286 517					
1.15719 37124 46780	(-1)2.07802 32581 49	0.79289 00483 864					
1.98165 67566 95843	(-2)1.70779 83007 41	0.86675 26065 634					
2.93063 74202 57244	(-4)1.99604 07221 14	1.07193 01442 480					
<b>n=9</b>				<b>n=20</b>			
0.00000 00000 00000	(-1)7.20235 21560 61	0.72023 52156 061					
0.72355 10187 52838	(-1)4.32651 55900 26	0.73030 24527 451					
1.46855 32892 16668	(-2)8.84745 27394 38	0.76460 81250 946					
2.26458 05845 31843	(-3)4.94362 42755 37	0.84175 27014 787					
3.19099 32017 81528	(-5)3.96069 77263 26	1.04700 35809 767					

3. Structural Analysis

Structural analysis is the name given to the procedure of characterizing the structure of the spatial distribution of variables considered. It is the first and indispensable step of any quantitative study. This step consists of defining below a set of variables and of characterizing the spatial structure of the variables studied. In most parts of quantitative theory the variables studied are assumed to be normally distributed. All the methods of analysis are based on this assumption and aim at a quantitative summary of all the available structural information.

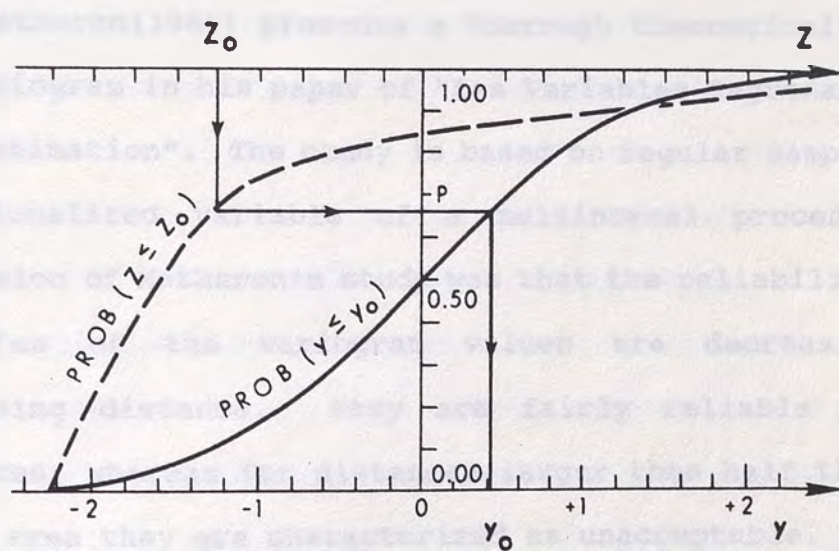


Figure 3. Graphical transformation for raw data to normal distribution

### 3. Structural Analysis

"Structural analysis" is the name given to the procedure of characterizing the structures of the spatial distribution of variables considered. It is the first and indispensable step of any geostatistical study. Thus every geostatistical study begins with the construction of a model designed to characterize the spatial structure of the regionalized variable studied. In most parts of geostatistical theory the variogram is assumed known or available through reliable estimates and acts as a quantified summary of all the available structural information.

Matheron(1965) presents a thorough theoretical study of the variogram in his paper of "Les Variables Regionalisees et Leur Estimation". The study is based on regular sampling from a regionalized variable of a multinormal process. One conclusion of Matheron's study was that the reliability of the estimates of the variogram values are decreasing with increasing distance. They are fairly reliable at short distances, whereas for distances larger than half the extent of the area they are characterized as unacceptable. The main reason for this is the decreasing number of observation pairs available for estimation(Omre, 1984). These conclusions show that the variogram quantifies the structural information for use in kriging estimation.

Variograms measure spatial variation in a regionalized

variable and define as the variance of the increment  $[Z(X) - Z(Y)]$ ;

$$2\gamma(X,Y) = \text{var}[Z(X) - Z(Y)]$$

where  $Z(X)$ ,  $Z(Y)$  are regionalized variables associated with location  $X$  and  $Y$ . Under the intrinsic hypothesis of geostatistics, that the increments  $[Z(X) - Z(Y)]$  associated with a small distance  $|h|$  are weakly stationary or stationary, this variance reduces to

$$2\gamma(h) = \text{var}[Z(X) - Z(X+h)]$$

where  $|h|$  is the vector separating the points  $X$  and  $X + |h| (=Y)$  restricted to the intrinsic case. Under this assumption, the first moment of the increment and its expected value, is constant or at most only slowly varying with spatial position  $X$ ; and the second moment is also invariant with spatial location.

For the variable  $Z$  at different locations, it becomes necessary to index the locations as  $X_i$ , where  $i=1,2,3,\dots,n$ , corresponding to  $n$  observations of data values. The estimator of the variogram from  $n$  data values  $\{Z(X_i), i=1,2,3,\dots,n\}$  is the arithmetic mean of the squared differences  $[Z(X_i) - Z(X_i + h)]$ :

$$2\gamma(h) = \sum_{i=1}^{n(h)} \frac{[Z(X_i) - Z(X_i + h)]^2}{n(h)}$$

where  $n(h)$  is the number of data pairs used for calculation in  $h$  distance. The factor "2" in front of the is there for

mathematical convenience. The term  $\gamma(h)$  was defined by Matheron as the semivariogram because it is one-half of the spatial variance:

$$\gamma(h) = \frac{1}{2} \sum_{i=1}^{n(h)} \frac{[Z(X_i) - Z(X_i + h)]^2}{n(h)}$$

But many subsequent authors have referred to  $\gamma$  as variogram, and David(1977) advocates that "for the sake of simplicity" this common usage should be adopted.

The value of  $\gamma(h)$  is calculated from regionalized variables of such a data set if there are  $n$  possible pairs of observations which are separated by distances of exactly  $h$ . It is possible to compute  $\gamma$  for other spacings: for example, by taking alternate observations,  $h_1 = 2h$  and  $\gamma(h_1)$  is computed in exactly the same way using the  $n$  possible pairs of observations which are separated by a distance of exactly  $2h$ . A succession of estimated values  $\gamma(h)$  is obtained using different spacings of  $h$ ,  $2h$ ,  $3h$ , ..... and these are plotted on the graph, in which the distance between the pairs of samples is plotted along the horizontal axis and the value of the variogram along the vertical as is shown in Figure 4.

The variogram as computed from the data will not tend to be perfect curves, but rather lumpy, noisy and less regular. But this lumpy graph can be likened to one or other of a small number of ideal curves. The ideal curves are defined as simple mathematical functions which relate  $\gamma$  to  $h$ . It is usually

These values are plotted against the distance between the two points. The resulting curve is the variogram. The variogram is a measure of the spatial correlation of the data. It is a function of the distance between the two points. The variogram is a measure of the spatial correlation of the data. It is a function of the distance between the two points.

The variogram is a measure of the spatial correlation of the data. It is a function of the distance between the two points. The variogram is a measure of the spatial correlation of the data. It is a function of the distance between the two points.

$$\gamma(h) = \frac{1}{2} \left( \frac{z(x) - z(x+h)}{h} \right)^2$$

The variogram is a measure of the spatial correlation of the data. It is a function of the distance between the two points. The variogram is a measure of the spatial correlation of the data. It is a function of the distance between the two points.

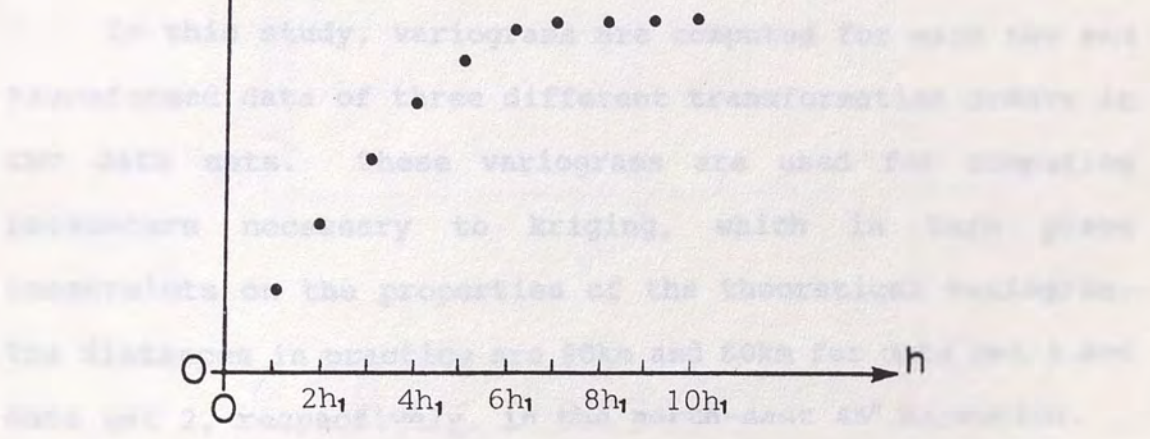


Figure 4. Usual method of plotting a variogram.

these ideal curves which are used in the subsequent kriging methods. Because the kriging methods need the relationship between point samples established by the variogram to estimate local values from surrounding point samples, the variogram must be modeled by one of these curves.

One commonly used model for this ideal curve is the spherical model that is used in this study. Figure 5 shows an example of a spherical model of a variogram that consists of two separation functions with a discontinuity:

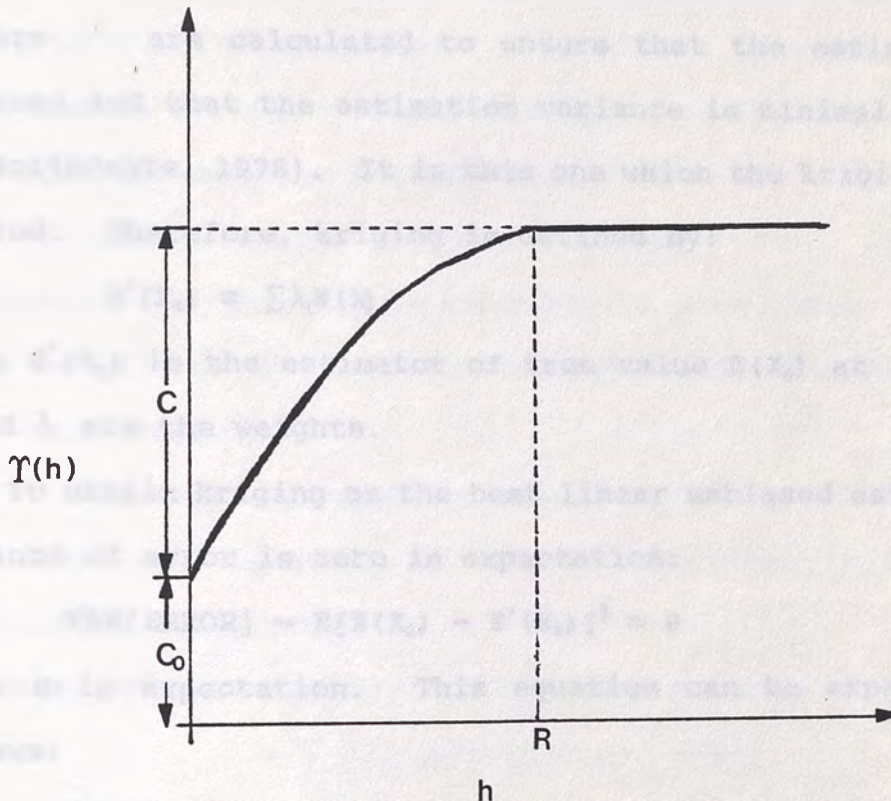
$$\begin{aligned} \gamma(h) &= C \left[ \frac{3h}{2R} - \frac{h^3}{2R^3} \right] + C_0 && \text{when } h \leq R \\ \gamma(h) &= C + C_0 && \text{when } h > R \end{aligned}$$

The separation  $R$  is called the "range",  $C + C_0$  is called the "sill" and  $C_0$  is known as the "nugget effect" that represents the discontinuity of the variogram at the origin and is due to measurement errors and some other factors.

In this study, variograms are computed for each raw and transformed data of three different transformation orders in two data sets. These variograms are used for computing parameters necessary to kriging, which in turn place constraints on the properties of the theoretical variogram. The distances in practice are 90km and 80km for data set 1 and data set 2, respectively, in the north-east  $45^\circ$  direction.

4. Cross-validation being applied.

The kriging technique estimates values at given spatial locations from observations made at other locations and a weighted averaging scheme. In which the weights are chosen such that the error associated with the prediction is less than that for any other set. The weights depend upon the location of the data and upon the variogram model. The weights are calculated to ensure that the estimator is unbiased and that the estimation variance is minimal (see Journé and Holsinger, 1975). It is this one which the kriging variance is used.



$C$ : Sill - Nugget,  $C_0$ : Nugget,  $R$ : Range

Figure 5. The spherical model of a variogram.

#### 4. Cross-validation using Kriging

The kriging technique estimates values at given spatial locations from observations made at other locations and a weighted moving average technique in which the weights are chosen such that the error associated with the "predictor" is less than that for any other sum. The weights depend upon the location of the data and upon the variogram model. The  $n$  weights are calculated to ensure that the estimator is unbiased and that the estimation variance is minimal (Journel and Huijbregts, 1978). It is this one which the kriging seeks to find. Therefore, kriging is defined by:

$$Z^*(X_0) = \sum \lambda_i Z(X_i)$$

where  $Z^*(X_0)$  is the estimator of true value  $Z(X_0)$  at location  $X_0$  and  $\lambda_i$  are the weights.

To obtain kriging as the best linear unbiased estimator, variance of error is zero in expectation:

$$\text{VAR}[\text{ERROR}] = E[Z(X_0) - Z^*(X_0)]^2 = 0$$

where  $E$  is expectation. This equation can be expanded as follows:

$$\begin{aligned} \text{VAR}[\text{ERROR}] &= E[Z(X_0) - Z^*(X_0)]^2 \\ &= E[Z(X_0)^2 - 2Z(X_0) \sum_{i=1}^n \lambda_i Z(X_i) + (\sum_{i=1}^n \lambda_i Z(X_i))^2] \\ &= E[Z(X_0)^2] - 2E[Z(X_0) \sum_{i=1}^n \lambda_i Z(X_i)] \\ &\quad + E[\sum_{i=1}^n \lambda_i Z(X_i) [\sum_{i=1}^n \lambda_i Z(X_i)]] \\ &= E[Z(X_0)^2] - 2E[Z(X_0) \sum_{i=1}^n \lambda_i Z(X_i)] \end{aligned}$$

$$+ E \left[ \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j Z(X_i) Z(X_j) \right] \dots \dots \dots (1)$$

To get spatial correlation from data values, the concept of spatial covariance is used:

$$COV[Z(X_i) Z(X_j)] = E[(Z(X_i) - m_i) (Z(X_j) - m_j)] \dots \dots \dots (2)$$

where COV is a covariance between Z(X<sub>i</sub>) and Z(X<sub>j</sub>) that are spatial data values at locations X<sub>i</sub> and X<sub>j</sub>; m<sub>i</sub> and m<sub>j</sub> are mean values of X and Y variables.

The equation (2) can be rewritten by;

$COV[Z(X_i) Z(X_j)] = E [Z(X_i) Z(X_j)] - m^2$  because all geostatistical analyses assume the intrinsic hypothesis: i.e. the data is second-order stationary. Thus the means of X<sub>i</sub> and X<sub>j</sub> variables are spatially constant in a stationary region;

$$m_i = m_j = m \text{ (global mean)}$$

From this, the equation (1) can be rewritten in terms of covariances as follows:

$$\begin{aligned} VAR[ERROR] &= E[Z(X_0)] - 2 \left[ \sum_{i=1}^n \lambda_i COV(Z(X_0) Z(X_i)) + m^2 \right] \\ &+ \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j COV[Z(X_i) Z(X_j)] + m^2 \\ &= VAR[Z_0] - 2 \sum_{i=1}^n \lambda_i COV(Z(X_0) Z(X_i)) \\ &+ \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j COV[Z(X_i) Z(X_j)] \dots \dots \dots (3) \end{aligned}$$

using the equation of  $VAR[Z_0] = E[Z(X_0)]^2 - m^2$  By the

definition of kriging, the estimate is unbiased and the estimation variance of error is minimized. This criterion is satisfied by using the Lagrangian mathematics techniques that

are used for constrained optimization problems. Therefore, the equation (3) can be rewritten using a Lagrange parameter:

$$\begin{aligned} \text{VAR}[\text{ERROR}] &= \text{VAR}[Z_0] - 2 \sum_{i=1}^n \lambda_i \text{COV}[Z(x_0) Z(x_i)] \\ &+ \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \text{COV}[Z(x_i) Z(x_j)] - 2\mu \left[ \sum_{i=1}^n \lambda_i - 1 \right] \end{aligned} \quad \dots\dots\dots (4)$$

where  $\mu$  is the Lagrangian multiplier. The estimation variance is to be minimized subject to the unbiased condition of  $\sum \lambda_i = 1$ , so that the optimal weights for kriging are obtained from the standard Lagrangian techniques in that the partial derivatives in equation (4) with respect to  $\lambda_i$  and  $\mu$  become zero;

$$\begin{aligned} \frac{\partial \text{VAR}[\text{ERROR}]}{\partial \lambda_i} &= -2 \text{COV}[Z(x_0) Z(x_i)] \\ &+ 2 \sum_{j=1}^n \lambda_j \text{COV}[Z(x_i) Z(x_j)] - 2\mu = 0 \\ \therefore \sum_{j=1}^n \lambda_j \text{COV}[Z(x_i) Z(x_j)] - \mu &= \text{COV}[Z(x_0) Z(x_i)] \end{aligned} \quad \dots\dots\dots (5)$$

$$\begin{aligned} \frac{\partial \text{VAR}[\text{ERROR}]}{\partial \mu} &= -2 \sum_{i=1}^n \lambda_i + 2 = 0 \\ \therefore \sum_{i=1}^n \lambda_i &= 1 \end{aligned} \quad \dots\dots\dots (6)$$

This set of equations can be expressed in matrix form;

$$\begin{bmatrix} \text{COV}[Z(x_0)Z(x_0)] & \dots & \text{COV}[Z(x_0)Z(x_n)] & 1 \\ \vdots & & \vdots & \vdots \\ \vdots & & \vdots & \vdots \\ \vdots & & \vdots & \vdots \end{bmatrix} \cdot \begin{pmatrix} \lambda_1 \\ \vdots \\ \vdots \\ \vdots \end{pmatrix} = \begin{pmatrix} \text{COV}[Z(x_0)Z(x_1)] \\ \vdots \\ \vdots \\ \vdots \end{pmatrix}$$

$$\begin{bmatrix} \vdots & & \vdots & \vdots \\ \text{COV}[Z(X_n)Z(X_1)] & \dots & \text{COV}[Z(X_n)Z(X_n)] & 1 \\ 1 & \dots & 1 & 0 \end{bmatrix} \begin{pmatrix} \lambda_n \\ -\mu \end{pmatrix} = \begin{pmatrix} \vdots \\ \text{COV}[Z(X)Z(X_n)] \\ 1 \end{pmatrix}$$

Let the first matrix be  $[K]$ , second be  $[\lambda]$ , and the third be  $[M2]$ :

$$[K] [\lambda] = [M2]$$

To find the weight values of each sample in the matrix  $[\lambda]$ , to obtain invert the matrix  $[K]$  and multiply through by  $[K]^{-1}$

$$[\lambda] = [K]^{-1}[M2]$$

The values of covariances between any two variables in matrix  $[K]$  is calculated by using the variogram:

$$\begin{aligned} \gamma(h) &= E\{[Z(X+h) - Z(X)]^2\} \\ &= \text{COV}(0) - \text{COV}(h) \end{aligned}$$

$$\text{COV}(h) = \text{COV}(0) - \gamma(h)$$

\*  $\text{COV}(h)$  is covariance of any two values of distance  $h$ .

\*  $\text{COV}(0)$  is covariance between any location itself, such that it is same to the sill value of variogram.

\*  $\gamma(h)$  is variogram value of two different variables of distance  $h$ .

From the above relation between covariance and variogram, the values of matrix  $[K]$  are computed from which the inverse matrix  $[K]^{-1}$  is obtained. It is also possible to obtain the

matrix [M2] by the values obtained from variogram parameters. Using the inverse matrix [K]<sup>-1</sup> and matrix [M2], it is possible to get weight values of [λ]. Therefore, the kriging estimation Z\*(X<sub>0</sub>) is computed:

$$Z^*(X_0) = \sum_{i=1}^n \lambda_i Z(X_i)$$

The effect of data transformation on the kriging estimator is assessed by cross-validation because it is an appropriate tool for testing estimation methods. The practice of cross-validation in this paper follows the procedure of comparing estimated values with observed values of raw data and transformed data in two data sets. The procedure is as follows: For each value X<sub>i</sub> (i=1,2,3,.....,n) in two data sets, compute a kriged value at the same location from the nearest neighboring data, without that sample value; compare the estimated values Z<sub>e</sub>(X<sub>i</sub>) with observed values Z(X<sub>i</sub>) for each data values. Clark(1986) reviews the history of cross-validation and its usefulness in geostatistics. She notes that this type of comparison is used initially to compare methods of estimation and to justify the use of kriging as an estimation method(Hohn, 1988). In this study, kriging is arbitrarily restricted to using the 10 nearest neighboring samples for estimation. Moreover, the radius of the window used for kriging is arbitrarily chosen as one-half the range of the variogram.

Finally, the results from cross-validation by kriging of the transformed data are retransformed by  $Z(X) = \phi^{-1}Y(X)$  into

original data values in order to compare results of cross-validation by kriging for raw data and transformed data.

4. Raw Data and Transformed Data

The global statistics of these two data sets are presented in Table 2 and histograms are shown in Figure 4. Data set 1, containing 120 samples, shows a highly skewed distribution and values range between 0.00 and 100.000. The histogram of data set 2 containing 120 samples illustrates that the distribution is less skewed than that of data set 1. The data range is between 0.00 and 34.71.

After data transformation, the two data sets represent distributions close to gaussian (Figure 5). The first group of data set 2 has a distribution more similar to that of the histogram of data set 1. There is no important difference for histograms between different transformation orders such as  $k=4$ ,  $k=10$ , and  $k=11$ . The global statistics for transformed data groups in Figure 5 are presented in Table 3. The transformation using Bernoulli polynomials results in a mean of 2 and variance of 1. However, the mean and variances calculated in this study are not exactly 2 and 1 because some tails exist in gaussian distribution derived from transformation. Nevertheless, inverse transforming the transformed data using the Bernoulli polynomials in this case produces statistical moments such as those shown in Table 4. These values of transformed data do not significantly

### III. RESULTS AND COMPARISON

#### 1. Raw data and Transformed data

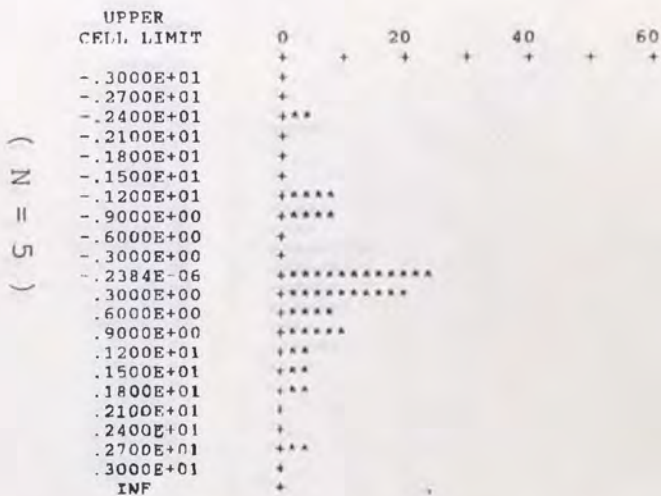
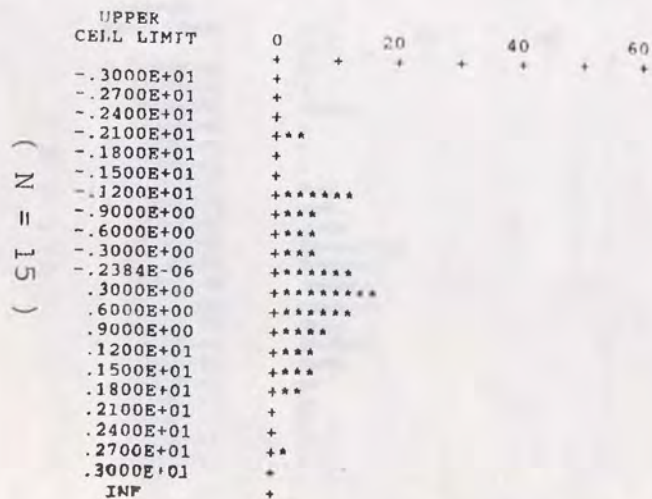
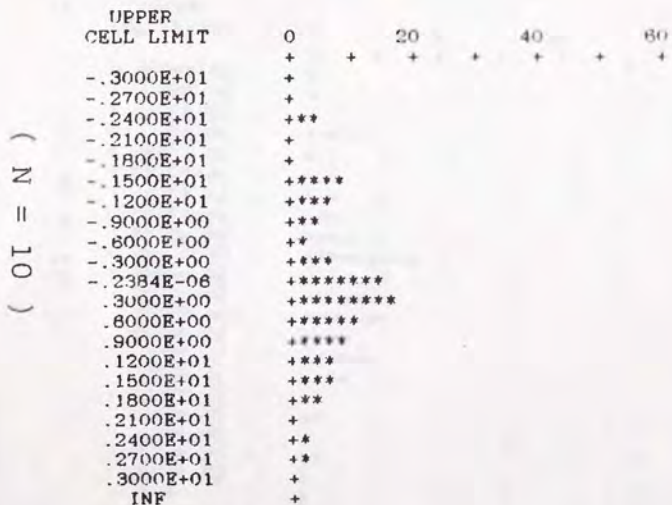
The global statistics of these two data sets are summarized in Table 2 and histograms are shown in Figure 6. Data set A, containing 126 samples, shows a highly skewed distribution and values range between 0.00 and 992.200. The histogram of data set B containing 190 samples illustrates that the distribution is less skewed than that of data set A. The data range is between 0.00 and 54.83.

After data transformation, the two data sets represent distributions close to gaussian shape (Fig.7); the histogram of data set B has a distribution more similar to normal than the histogram of data set A. There is no important difference for histograms between different transformation orders such as  $k=5$ ,  $k=10$ , and  $k=15$ . The global statistics for transformed data plotted in figure 7 are summarized in Table 3. The transformation using Hermite polynomials yields data that have a mean of 0 and variance of 1. However, the means and variances calculated in this study are not exactly 0 and 1 because some tails exist in gaussian distribution derived from transformation. Nevertheless, inverse transforming the transformed data using the Hermite model results in data whose statistical moments match closely those of the raw data (Table 4). These values of transformed data do not significantly

Table 2. The global statistics of two data sets

	Data set A	Data set B
<b>N. of samples</b>	126	190
<b>Mean</b>	125.121	10.581
<b>Variance</b>	43277.970	103.753
<b>Std. Dev.</b>	208.034	10.186
<b>Skewness</b>	2.448	1.384
<b>Median</b>	33.905	7.375

Figure 6. Histograms of the two data sets.



(A). Data set A

Figure 7. Histograms of transformed data.

Fig. 7. Continued.

(B) . Data set B

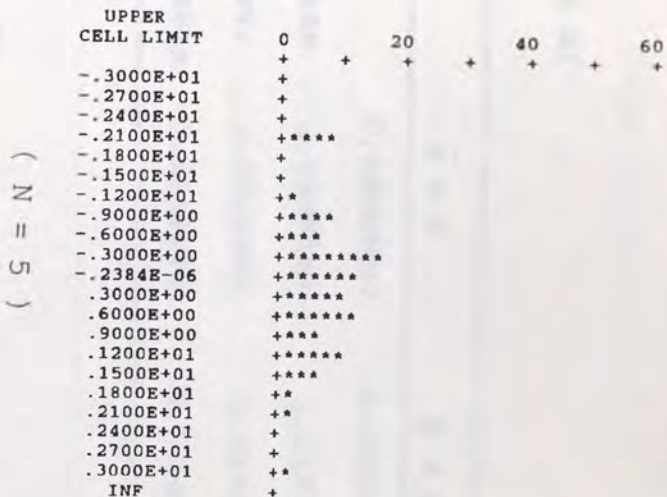
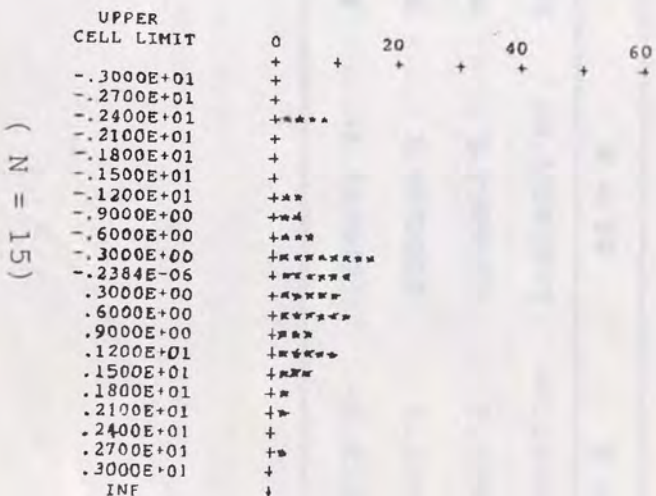


Table 3. The global statistics for transformed data.

## Data set A)

	k = 5	k = 10	k = 15
Mean	0.09884547	0.01396579	0.05847715
Variance	1.1269230	1.2751510	1.1149950
Std.Dev.	1.1061566	1.129960	1.0559330
Skewness	-0.2349486	-0.1416749	0.08206316

## Data set B)

	k = 5	k = 10	k = 15
Mean	-0.01990485	-0.02595637	-0.04997877
Variance	1.1538960	1.2060450	1.3184270
Std. Dev.	1.1074196	1.0982010	1.1482280
Skewness	-0.1621768	-0.2138797	-0.4214138

Table 4. Comparisons of mean and variance between raw data and Hermite model.

Data set A)

	Mean	Diff(%)	Variance	Diff(%)
From Raw data	125.12080		43277.970	
From Hermite k = 5	108.5896	-13.2	31550.96	-27.1
Model k = 10	108.5896	-13.2	32045.18	-26.0
k = 15	108.5896	-13.2	32051.18	-25.9

Data set B)

	Mean	Diff(%)	Variance	Diff(%)
From Raw data	10.348		93.844	
From Hermite k = 5	10.197	-1.5	90.158	-3.9
Model k = 10	10.383	-1.9	96.269	-7.2
k = 15	10.383	-1.9	96.291	-7.2

differ between transformation orders, except for  $k=5$  for data set B. The values of coefficients are listed in the Table 5.

One problem encountered in transforming data is that the highest value 54.830, when transformation order  $k=5$  is used for data set B, has not transformed properly; the program prints out a warning message. This phenomenon can happen when data having low end values or upper end values of data distribution are transformed to gaussian shape (Hohn, 1988). In this case, those values have to be deleted from data set and then transformation is carried out. Therefore, in this study, one largest value of data set B is trimmed for data transformation.

## 2. Variograms

As described before, the variogram model and parameters of the variogram are important for the optimal estimation by kriging because the ideal variogram model and parameters chosen from the values of observed experimental variograms are used in computing kriging weights. The experimental variograms were calculated from raw data and transformed data resulting from the three different transformation orders. The graphical results from these variograms are shown on Figure 8 and quantitative results are summarized on Table 6.

Table 5. The coefficients for Hermite polynomials transfer function.

Coefficients	Data set A	Data set B
$C_0$	108.58960	10.382980
$C_1$	143.79940	9.2289490
$C_2$	70.12870	2.2812980
$C_3$	9.053464	-0.1968090
$C_4$	-4.764523	-0.11519360
$C_5$	-1.913547	0.0085020410
$C_6$	-0.03550424	-0.0015762770
$C_7$	0.09378555	-0.0042804920
$C_8$	0.005598922	-0.00072181840
$C_9$	-0.004788297	0.00018211850
$C_{10}$	-0.0001804899	0.000073894600
$C_{11}$	0.0003401696	0.0000063284330
$C_{12}$	0.00003940154	-0.0000006700042
$C_{13}$	-0.000007947422	-0.00000002294676
$C_{14}$	-0.0000005750427	-0.000000022701710

## Data set A)

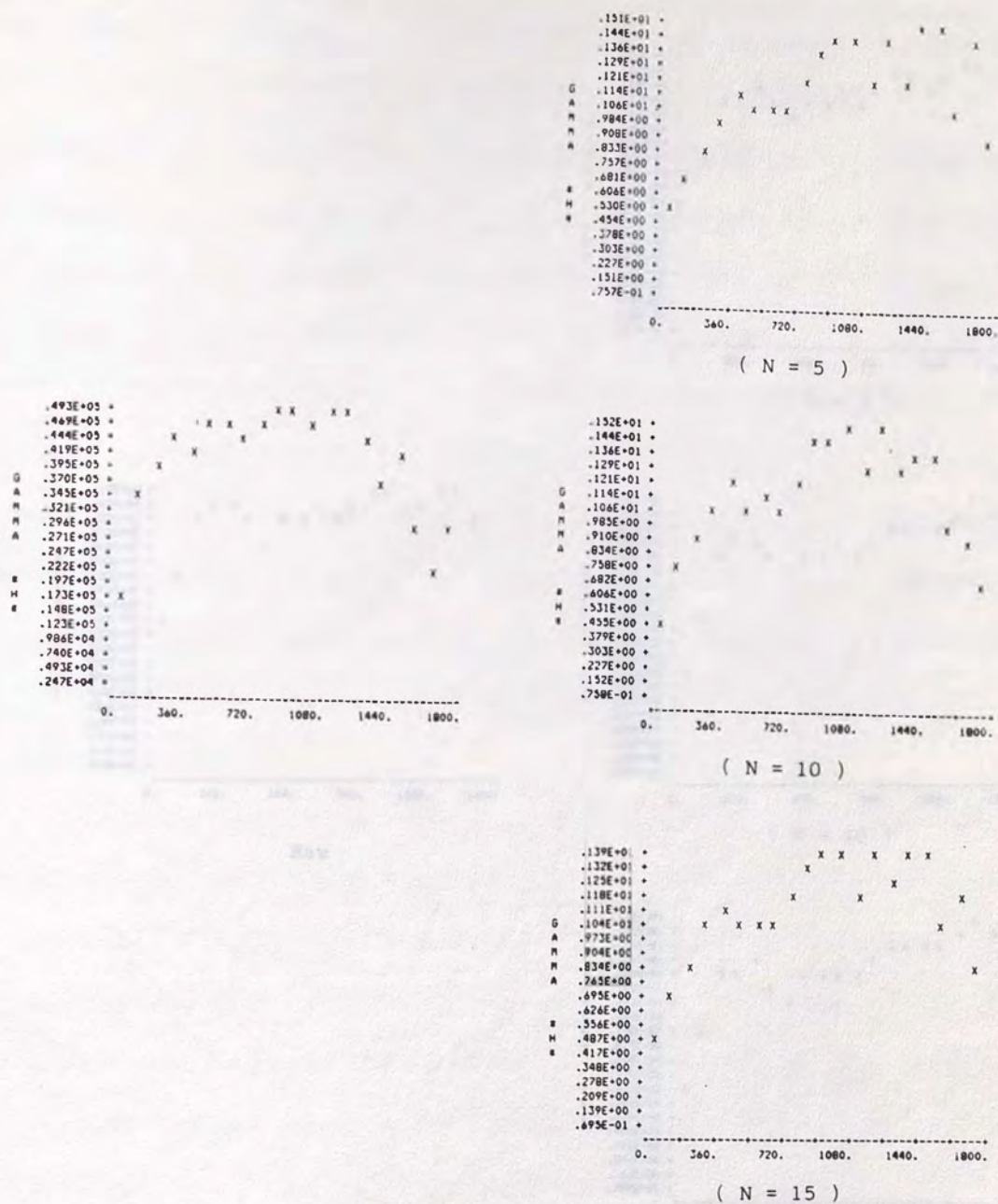


Fig. 8. Figure 8. Experimental and modelled variograms

## Data set B)

```

-1.13E+01 *
-1.27E+01 *
-1.20E+01 *
-1.13E+01 *
-1.07E+01 *
-1.00E+01 *
-9.9E+00 *
-8.27E+00 *
-8.01E+00 *
-7.34E+00 *
-6.67E+00 *
-6.00E+00 *
-5.34E+00 *
-4.67E+00 *
-4.00E+00 *
-3.34E+00 *
-2.67E+00 *
-2.00E+00 *
-1.33E+00 *
-6.67E-01 *
0.
-----
0. 320. 640. 960. 1280. 1600.

```

( N = 5 )

```

-1.14E+03 *
-1.09E+03 *
-1.03E+03 *
-9.72E+02 *
-9.14E+02 *
-8.57E+02 *
-8.00E+02 *
-7.43E+02 *
-6.86E+02 *
-6.29E+02 *
-5.72E+02 *
-5.14E+02 *
-4.57E+02 *
-4.00E+02 *
-3.43E+02 *
-2.86E+02 *
-2.29E+02 *
-1.71E+02 *
-1.14E+02 *
-5.72E+01 *
0.
-----
0. 320. 640. 960. 1280. 1600.

```

Raw

```

-1.44E+01 *
-1.37E+01 *
-1.30E+01 *
-1.23E+01 *
-1.15E+01 *
-1.08E+01 *
-1.01E+01 *
-9.59E+00 *
-9.07E+00 *
-8.55E+00 *
-8.03E+00 *
-7.51E+00 *
-6.99E+00 *
-6.47E+00 *
-5.95E+00 *
-5.43E+00 *
-4.91E+00 *
-4.39E+00 *
-3.87E+00 *
-3.35E+00 *
-2.83E+00 *
-2.31E+00 *
-1.79E+00 *
-1.27E+00 *
-7.5E-01 *
0.
-----
0. 320. 640. 960. 1280. 1600.

```

( N = 10 )

```

-1.40E+01 *
-1.32E+01 *
-1.48E+01 *
-1.36E+01 *
-1.28E+01 *
-1.20E+01 *
-1.12E+01 *
-1.04E+01 *
-9.99E+00 *
-8.79E+00 *
-7.19E+00 *
-6.39E+00 *
-5.59E+00 *
-4.79E+00 *
-3.99E+00 *
-3.19E+00 *
-2.39E+00 *
-1.59E+00 *
-7.99E-01 *
0.
-----
0. 320. 640. 960. 1280. 1600.

```

( N = 15 )

Fig.8. Continued.

Table 6. Variogram parameters

Data set	Nugget	Sill	Range (km)	Model
Raw	9860.00	43300.00	680	Spherical
k=5	0.34	1.13	756	Spherical
A k=10	0.33	1.28	720	Spherical
k=15	0.28	1.10	760	Spherical
Raw	40.00	104.00	860	Spherical
k=5	0.33	1.15	880	Spherical
B k=10	0.26	1.21	900	Spherical
k=15	0.38	1.32	900	Spherical

procedure.

In this study, hypothesis testing is applied to determine whether the results of cross-validation with transformed data are significantly different from results using raw data. The parametric test for hypothesis testing is applied to test differences between the results of cross-validation for raw data and transformed data because the errors are normally distributed. The  $F$  test is applied in this study and the results of this test are given in Table 3. The results of a statistic for raw data and transformed data are given in Table 3. From these values, no evidence suggests that the results computed by cross-validation using raw data and transformed data are

### 3. Cross-validation by kriging

Cross-validation was used to evaluate the accuracy of the kriging and the results of cross-validation using raw data and transformed data are compared. For each data location, the true value and estimated value are compared and their difference calculated. The results of this cross-validation are given in Table 7.

Davis(1987) urged that cross-validation is not a hypothesis testing method. For this limitation on cross-validation, hypothesis testing for cross-validation results was introduced by Carr and Robert(1989) to accept or reject the hypothesis of superior estimation. They stated that cross-validation can contributed to a hypothesis testing procedure.

In this study, hypothesis testing is applied to determine whether the results of cross-validation using transformed data are significantly different from results using raw data. The parametric test for hypothesis testing is applied to test differences between the results of cross-validation for raw data and transformed data because the errors are normally distributed. The Z test(Appendix A) is applied and the results of this test are given in Table 8. The results of Z statistics for two data sets approach almost zero. From these values, no evidence suggests that the results computed by cross-validation using raw data and transformed data are

Table 7. The results of cross-validation

Data set		M.E.	E.V.	A.K.V.
A	Raw	0.365593	54211.230	9308.562
	Trn. (k=5)	-0.005222	0.77952780	0.565036
	Trn. (k=10)	-0.002703	0.89316510	0.597901
	Trn. (k=15)	-0.002480	0.81987950	0.501349
B	Raw	-0.223447	88.862620	46.548810
	Trn. (k=5)	-0.0170376	0.9463240	0.524627
	Trn. (k=10)	-0.011305	0.9507783	0.421823
	Trn. (k=15)	-0.020829	1.0614860	0.594817

M.E.: Mean Error

E.V.: Error Variance

A.K.V.: Average of kriging Variance

Table 8. The results of Z test.

Data set	Trn. order	value of Z test	Table(5%)
A	k = 5	0.02	2.33
	k = 10	0.02	2.33
	k = 15	0.02	2.33
B	k = 5	0.24	2.33
	k = 10	0.24	2.33
	k = 15	0.23	2.33

different.

#### 4. Untransformed data and Retransformed data

The results from cross-validation of transformed data are inverted(retransformed) to original units. Then, the kriging results of raw data can be compared with that of transformed data. An alternative analysis to parametric hypothesis test is performed on the results from cross-validation. The analysis is based on counting the number of locations where transformed data yield closer estimated value to true value than untransformed data, and also comparing the magnitude of reducing estimation error between where estimated values of transformed data are more accurate than those of raw data and where estimated values of raw data are more accurate than those of transformed data(Table 9).

Ratios are calculated by dividing the number of data locations at which the estimation method using transformed data is superior by the total number of data locations. From Table 9, all ratios of data set A and B are greater than 0.5 and the data set A has larger ratio values than data set B. The ratio greater than 0.5 means that the number of estimated values of transformed data, which are estimated more accurately than raw data, is larger than the number of estimated values of raw data, which are estimated more

Table 9. The results of superiority test between raw data and transformed data

Data set	trn. order	ratio	A.R.E. (Trn.)	A.R.E. (Raw)
A	k = 5	0.54	76.44	29.66
	k = 10	0.64	74.82	44.00
	k = 15	0.64	74.36	43.63
B	k = 5	0.57	2.62	1.81
	k = 10	0.58	2.53	1.19
	k = 15	0.57	2.43	2.19

NOTE:

Ratio is the number of location at which the results from cross-validation using retransformed data are closer to true values than these using raw data; this number is divided by total data number.

A.R.E.(Trn.) is average value of reducing error difference, that is mean of all difference values between estimation errors of retransformed data and raw data at location where estimated value using transformed data is closer to true value than estimated value using raw data.

A.R.E. (Raw) is average of reducing error difference at location where estimated value using raw data

accurately than is closer to true value than estimated value of data have a greater accuracy using transformed data.

\* Equation;

$$\text{A.R.E. (Trn)} = \frac{1}{N} \sum_{i=1}^N (|ER_i - T_i| - |ET_i - T_i|)$$

$$\text{A.R.E. (Raw)} = \frac{1}{M} \sum_{i=1}^M (|ET_i - T_i| - |ER_i - T_i|)$$

**N:** The total number of locations where estimated value using transformed data is closer to true value than estimated value using raw data.

**M:** The total number of locations where estimated value using raw data is closer to true value than estimated value using transformed data.

**ER<sub>i</sub>:** Estimated value using raw data.

**ET<sub>i</sub>:** Estimated value using transformed data.

**T<sub>i</sub> :** True value.

accurately than transformed data. This means that transformed data have a greater number of estimated values which are more accurate than for raw data, and demonstrates that the data transformation improves the accuracy of kriging estimation. The effect of this improvement comes out more satisfactorily in data set A than data set B because data set A is more highly skewed. The other value from Table 9, A.R.E., supports evidence of the decision that data transformation improves the accuracy of kriging estimation even though the ratios of data set B and ratio of data set A with transformed order  $N=5$  are not much greater than 0.5 which indicates equal performance between raw data and transformed data. Because the magnitude of reducing error difference at locations where the estimation of transformed data are more accurate is greater than the magnitude of reducing error difference at location where the estimation of raw data are more accurate, then using transformed data provides improvement. The big difference of magnitude between these two values in data set A means the estimated values are changed to become more close to the true value when transformed data are used in kriging estimation. In data set B, the results of transformation is not improved as much as in data set A. There is no significant difference of results between different transformation orders in each data set, except ratio values of  $k=5$  order in data set 1.

## 5. Scatter diagram

A visual comparison between raw data and retransformed data is obtained using scatter diagrams (Fig. 9). The scatter diagrams show estimated values versus true values. Theoretically, a plot of estimated values versus true values yields a straight line with a slope of 1 when estimation is perfect. However, this plot shows scatter around the 1:1 line in a practical case. The scatter diagrams of data sets A and B give some correlation that is used for comparing the results of kriging using raw data with the results of kriging using transformed data.

The comparisons of scatter diagrams of raw data with transformed data in data set A indicates that the estimated values by kriging using transformed data are closer to original true values and less dispersed than these using raw data; specifically around low values of original data or below the mean value. Transformed data with  $k=15$  and  $k=10$  orders represent better results than with  $k=5$ . However, the estimated values are not improved and changed by data transformation near the above mean values. Moreover, there is a tendency for underestimation of some high values.

The scatter diagram for the raw data of set B does not show as good a result in comparison to data set A and not much difference between the results of transformed data and raw data. However, there is also a certain degree of

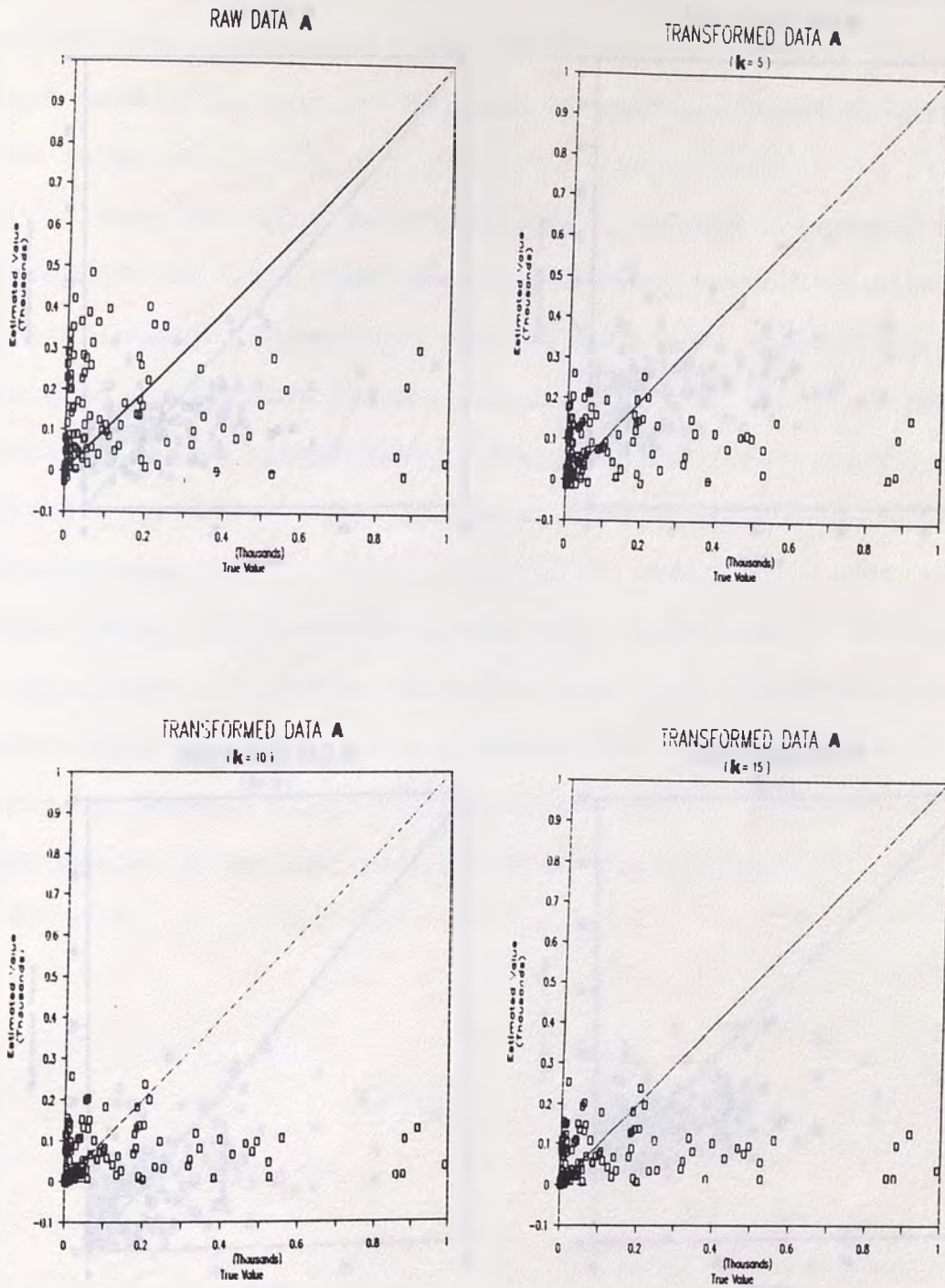


Figure 9. Scatter diagrams.

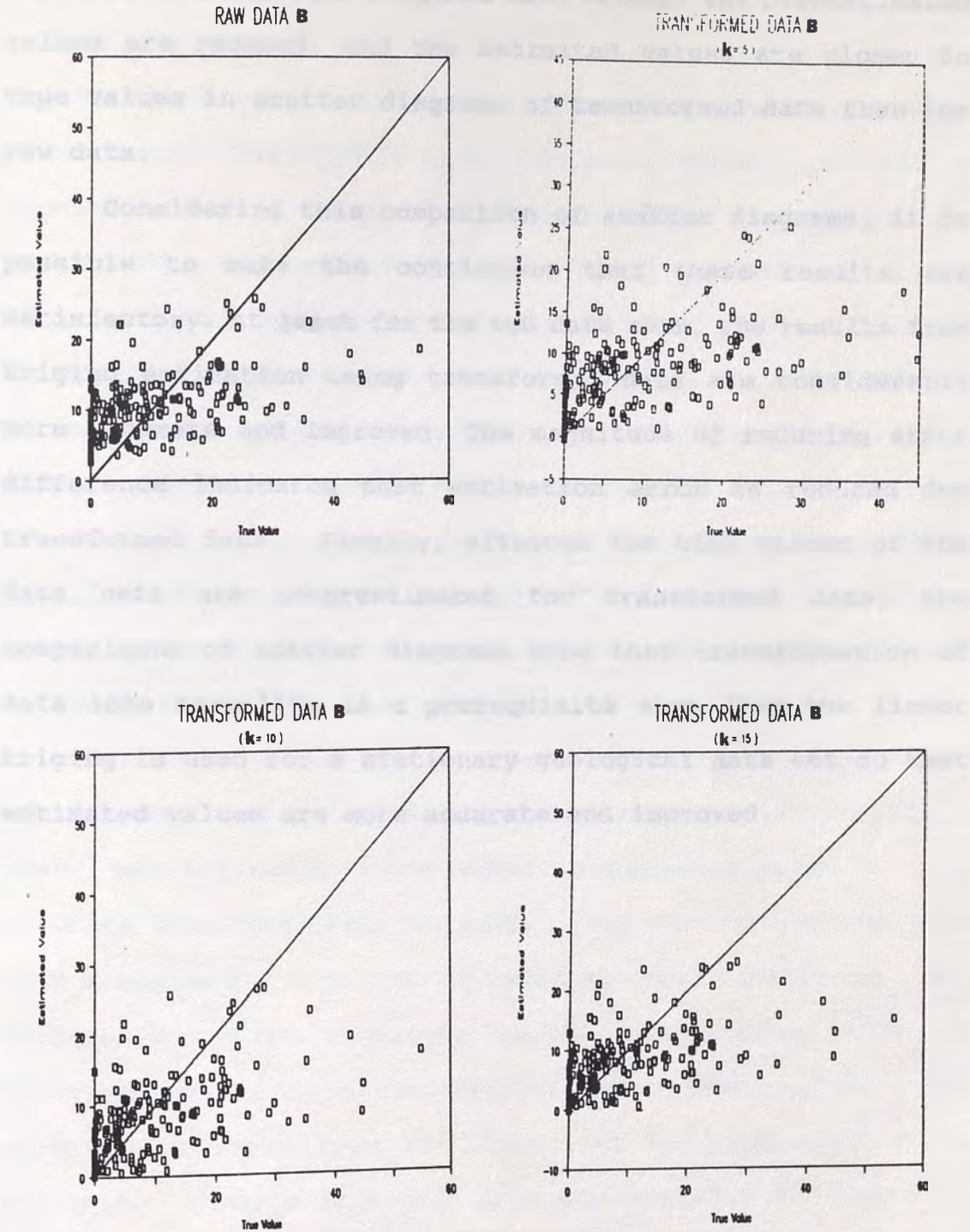


Figure 9. Continued.

improvement around low original data values; the overestimated values are reduced, and the estimated values are closer to true values in scatter diagrams of transformed data than for raw data.

Considering this comparison of scatter diagrams, it is possible to make the conclusion that these results are satisfactory. At least for the two data sets, the results from kriging estimation using transformed data are considerably more accurate and improved. The magnitude of reducing error difference indicates that estimation error is reduced for transformed data. Finally, although the high values of the data sets are underestimated for transformed data, the comparisons of scatter diagrams show that transformation of data into normality is a prerequisite step when the linear kriging is used for a stationary geological data set so that estimated values are more accurate and improved.

#### IV. DISCUSSION

Hypothesis testing was applied to cross-validation results even though David(1977) says cross-validation is not a hypothesis-test method. This test shows no evidence to suggest that the results of kriging using transformed and raw data are different. However, application of hypothesis tests to the results from cross-validation may not be valid because an important assumption for hypothesis testing requires random sampling(Carr and Roberts, 1989). They said that the errors resulting from cross-validation do not represent random sampling because the errors are not independent;a level of spatial interdependence is present. This interdependence may influence the hypothesis testing.

For this reason, one alternative analysis to hypothesis testing was used in this study; the counting of locations where the estimated value using transformed data is more accurate than that using raw data. This alternative analysis also examines the magnitude of reducing error difference. The difference of the magnitude values of reducing error at locations, where estimation results from transformed data are superior to these from raw data, and at locations, where estimation results from raw data are superior to these of transformed data, supports the conclusion that the estimation using transformed data is more accurate even though the ratio values for number of locations, at which the results from

transformed data are superior to those of raw data, are not much greater than 0.5.

There is a general tendency for overestimated values to be reduced, hence closer to true values in transformed data, especially for thresholds less than the mean values. But, the estimated values using transformed data are more underestimated than those using raw data in thresholds at high end values of the data distributions.

## V. CONCLUSION

The objective of this paper is to study the necessity and effect of data transformation to normality on kriging estimation. Also, the study is aimed at investigating the applicability of using Hermite polynomials for data transfer function. For this purpose, two data sets were used and the results from kriging estimation using transformed data and raw data were compared. All comparisons were fundamentally based on the results from cross-validation that cannot be used as a hypothesis testing method. Hypothesis testing of the results from cross-validation show no evidence to suggest that estimated values of transformed data and raw data are different. However, this conclusion by hypothesis testing is not valid in this study because spatial independence of errors used for hypothesis testing of cross-validation results is not a correct assumption for this test.

Another analysis was used to compare results: the counting of locations where estimated values of transformed data are closer to true value than for estimation using raw data and comparing the magnitude of reducing error difference between two locations, where the estimated values of transformed data are more accurate and the estimated values of raw data are more accurate. From this comparison, the results show that the accuracy of kriging estimation using transformed data, which are distributed normally, is improved so that the

estimated values are more accurate than those for raw data because the ratios are greater than 0.5 and the magnitude of reducing error difference is increased for transformed data.

The scatter diagrams also show that significant improvements in the accuracy of estimation are obtained by data transformation in thresholds less than mean values. However, there is no improvement of accuracy around high values. Moreover, some estimated values of high-end data are more underestimated in transformed data. Even though a tendency to underestimate high values exists, the estimation using transformed data results in superior accuracy in comparison to that using raw data. The differences between improvements for data set A and data set B are caused by the shape of their raw data distribution. Data set B has a less skewed distribution compared to data set A. This means that the improvement in accuracy of kriging estimation by data transformation is more effective for highly skewed data than for less skewed data.

## VI. APPENDIX A: HYPOTHESIS TEST

A parametric hypothesis test, Z test, is used in this study for comparing mean error of estimation using raw data with that using transformed data. The Z statistic is calculated as follows:

$$Z = (\text{Mean}_A - \text{Mean}_B) / K$$

where,  $\text{Mean}_A$  = Mean of sample

$\text{Mean}_B$  = Mean of sample B

$$K = (SP^2/N_A + SP^2/N_B)^{0.5}$$

$$SP^2 = [(N_A - 1)(\text{VAR A}) + (N_B - 1)(\text{VAR B})] / (N_A + N_B - 2)$$

$N_A$  = Number of values in sample A

$N_B$  = Number of values in sample B

In this study,  $\text{Mean}_A$  is the value for mean of raw data estimation errors and  $\text{Mean}_B$  is the value for mean estimation error for transformed data. The results for Z statistics are compared with table values to judge the hypothesis that  $\text{mean}_A = \text{mean}_B$ . If the absolute value of the Z statistic exceeds table values, this hypothesis is rejected.

**VII. BIBLIOGRAPHY**

- Abramowitz, M. and Stegun, D.A.**, 1970, Handbook of Mathematical Functions with Formulas, Graphs and Tables: National Bureau of Standards Applied Mathematics Series, 55, U.S. Government Printing Office, Washington, D.C., p.924.
- Carr, J.R. and Roberts, K.P.**, 1989, Application of Universal Kriging for Estimation of Earthquake Ground Motion: Math. Geol., Vol.21, No.2, p.255-265.
- Clark, I.**, 1986, The Art of Cross-Validation in Geostatistical Application, Proceedings of 19th APCOM Symposium, p.211-220.
- David, M.**, 1972, Tool for Planning Variance and Conditional Simulations, Proceedings of 11th APCOM Symposium: University of Arizona, p.D10-D23.
- David, M.**, 1977, Geostatistical Ore Reserve Estimation: Elsevier Scientific Publishing Co., New York, 364p.
- Davis, B.M.**, 1987, Uses and Abuses of Cross-validation in Geostatistics: Math. Geol., No.3, p.241-248.
- Englund, J.E.**, 1990, A variance of Geostatisticians: Math. Geol., Vol.22, No.4, p.417-455.
- Hohn, M.E.**, 1988, Geostatistics and Petroleum Geology: Van Nostrand Reinhold, New York, 183p.
- Journel, A. and Huijbregts, C.**, 1978, Mining Geostatistics: Academic Press, London, 600p.

**Matheron, G., 1965, Les Variables Regionalisees et Leur Estimation: Masson, Paris, 305p.**

**Matheron, G., 1976, Forecasting Block Grade Distribution: The Transfer Function: Advanced Geostatistics in the mining Industry, p.221-251.**

**Omre, 1984, Alternative Variogram Estimators in Geostatistics. PH.D. thesis, Stanford University.**