MODELING PATTERNS OF

SMALL SCALE SPATIAL VARIATION IN SOIL

by

Fang Huang

A Project Report

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Master of Science

in

Applied Statistics

January 2006

APPROVED:

_____

Dr. Jayson D. Wilbur, Advisor

_____

Dr. William J. Martin, Associate Department Head

Abstract

The microbial communities found in soils are inherently heterogeneous and often exhibit spatial variations on a small scale. Becker et al. (2006) investigate this phenomenon and present statistical analyses to support their findings. In this project, alternative statistical methods and models are considered and employed in a re-analysis of the data from Becker. First, parametric nested random effects models are considered as an alternative to the nonparametric semivariogram models and kriging methods employed by Becker to analyze patterns of spatial variation. Second, multiple logistic regression models are employed to investigate factors influencing microbial community structure as an alternative to the simple logistic models used by Becker. Additionally, the microbial community profile data of Becker were unobservable at several points in the spatial grid. The Becker analysis assumes that the data are missing completely at random and as such have relatively little impact on inference. In this re-analysis, this assumption is investigated and it is shown that the pattern of missingness is correlated with both metabolic potential and spatial coordinates and thus provides useful information that was previously ignored by Becker. Multiple imputation methods are employed to incorporate the information present in the missing data pattern and results are compared with those of Becker.

Contents

List of Tables v

List of Figures vi

Chapter 1 Introduction 1

Chapter 2 Modeling Spatial Dependence 5

2.1 Methods 5

2.1.1 Hierarchical Nested Model 5

2.1.2 Semivariogram Models and Kriging method 11

2.2 Results 15

2.3 Summary and Future Work 21

Chapter 3 Microbial Community Analysis 23

3.1 Methods 23

3.1.1 Logistic Regression 23

3.1.2 Missing Data and Multiple Imputation 24

3.1.3 False Discover Rate 26

3.2 Results 27

3.3 Conclusion and Future work 33

Bibliography 35

Appendix: SAS Code 37

Acknowledgements

I would like to express my thanks to Professor Jayson D. Wilbur, whose guidance, advising and comments helped me to follow a right direction and to find a right approach during the whole course of this project.

In addition, I would also like to thank Josey Becker, Allan Konopka and Cindy Nakatsu from Purdue University and Tim Parkin from Iowa State University for their original experiment, soil sample data and analysis. Josey was available to answer questions about the data throughout this project and the original experiment was funded by a grant from the Department of Energy to Allan and Cindy. So, without them, I wouldn't have had the data for this project.

List of Tables

Table 2-1 ANOVA table for 3-way nested model for unbalanced data from Becker et al. 10

Table 2-2 3-way nested model fitting criteria (R^2) for log(Pb), log(Cr) and log(Potential) 15

Table 2-3 The correlations between two responses 15

Table 2-4 The parameters and model fit criteria for fitting model to variogram..... 16

Table 2-5 SSEs and $\sum |e|$ s from 3-way nested model and spherical semivariogram model..... 19

Table 3-1 SAS output of logistic regression test for missingmess using all covariates ... 28

Table 3-2 SAS output of logistic regression test for missingmess based on metabolic potential and coordinates x and y..... 28

Table 3-3 The overall hypothesis tests results for 68 bands. 31

Table 3-4 Significant tests for the effect of lead, chromium and metabolic potential among 9 selected bands. 32

List of Figures

Figure 1-1 The 3-level hierarchical structure of soil samples in the experiment of Becker et al. (2006) 2

Figure 2-1 The 3-way nested design for the soil samples from Becker et al. (only illustrates the structure of array 1) 6

Figure 2-2 A typical theoretical semivariogram 13

Figure 2-3 Kriging maps for log(Pb), log(Cr) and log(Activity). 18

Figure 2-4 Contour maps of the difference between the nested model and the kriging model for log(Pb), log(Cr) and log(Activity) 20

Figure 3-1 Histogram of the estimated probability of non-missing. 29

Figure 3-2 Estimated non-missing probabilities by coordinates (X and Y). 30

Chapter 1 Introduction

The structure of the microbial community in soils is inherently heterogeneous as a result of both adaptation to environmental gradients and the intrinsic biological processes among the microbial community (Franklin et al., 2003). Therefore, the variability and heterogeneity of the microbial community in soil samples often exhibit patterns of spatial variation on a small scale (Etterna et al., 2002).

In a recent study, Becker et al. (2006) have investigated the spatial relationship within and between metabolic potential and heavy metal contaminants, such as lead and chromium, at small scale at a chronically contaminated site. The purpose of the present study is to investigate alternative models and make improvements to the statistical analyses. First, a hierarchical model was used to replace the semivariogram models and kriging used by Becker for spatial analysis. Then the covariate information for all sample points with unobservable responses were included in the microbial community analysis using multiple imputation methods.

To better describe the hierarchical model used later, the schematics of the sampling plan were considered as a hierarchical structure (Zhu et al., 2004), as shown in Fig. 1-1. The soil samples of Becker et al. (2006) were collected from 5 different arrays. Only array 1, 2 and 3 are shown here. The distance between the centers of two arrays was 50 cm. For each array, there were seven hexagonal sub-arrays. The center of each sub-array was 15 cm away from its adjacent sub-arrays within one primary array. For each sub-array, there were seven hexagonal sub-sub-arrays. The center of each sub-sub-array was 5 cm away from its adjacent sub-sub-arrays within one sub-array. Three samples were

collected on each sub-sub-array area. In total there were 645 soil samples (i.e., 3 samples × 5 arrays × [1 center point + 6 sub-array ×7 sub-sub-arrays]). The concentrations of the metal contaminants lead and chromium, the metabolic potential, and a profile of microbial community structure were measured at each sample point..

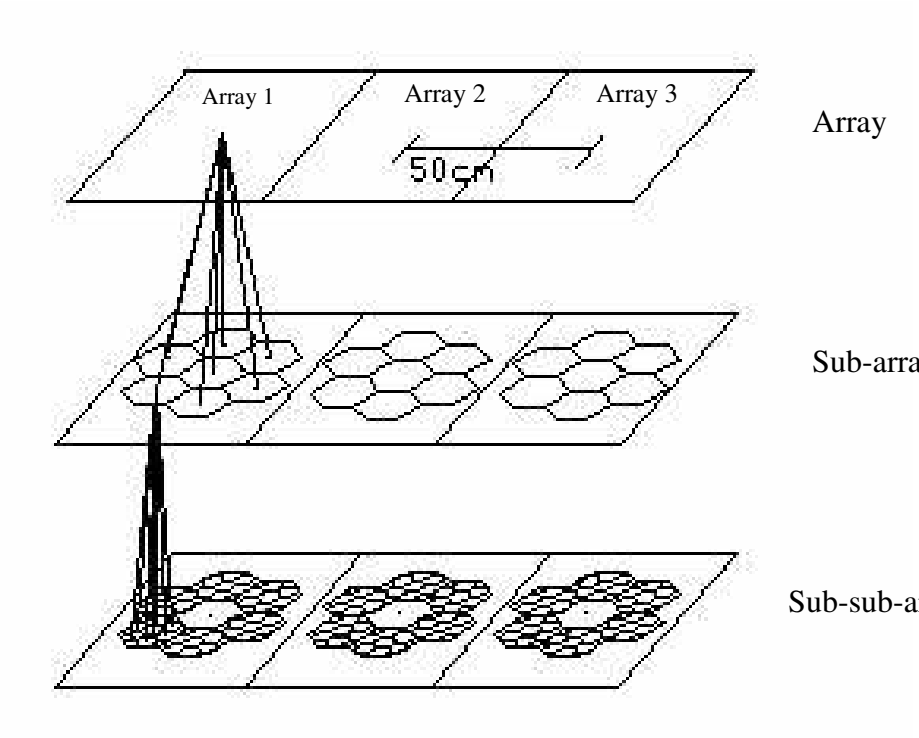

Figure 1-1 The 3-level hierarchical structure of soil samples in the experiment of Becker et al. (2006)

A strong spatial dependence both within and between lead, chromium and metabolic potential was found by Becker. The semivariogram models were used to describe the spatial dependence and kriging was used to estimate the response surface for the entire spatial grid. However, only qualitative comparisons and associations were made. The objective of this present work is to investigate whether a parametric statistical model based on the spatial dependent design might fit the data better and allow for more formal statistical inference, rather than only qualitative comparisons.

In Becker, the kriging maps showed that some areas with increased metal concentrations corresponded to the areas with decreased metabolic activity. But this pattern did not strictly hold true for the whole area. Also there were no qualitative comparisons because the direct relationship between the metabolic potential and the metal contaminations was not found in independent samples on the finest scale (< 1cm) (Becker et al., 2006). However, a multivariate hierarchical model for lead, chromium and metabolic potential may allow for statistical inference about relationships between lead, chromium and metabolic potential based on covariance components. Then the effect of the heavy metals on the microbial community and the spatial variation in chronically contaminated sites could be found.

In Chapter 2, the hierarchically nested random effects model used for the soil samples will be explained in detail, and the features of the model will be discussed. The model fit and test results for each of chromium concentration lead concentration and metabolic potential will be presented and comparisons between the results from the hierarchical spatial model and from the kriging model will be made.

In the Becker study, the microbial community profiles obtained by denaturing gradient gel electrophoresis (DGGE) were unobservable at several sample points. Primarily, samples with both high metal contents and high metabolic activity (137 out of 645) were obtained in the bacterial community experiment, and others samples were unobservable. When the logistic regression was employed to identify significant microbial populations with respect to each of lead, chromium and metabolic activity, the missing data which were unobservable had not been taken into account. If the data were missing completely at random, the inference from the observed data can be applied to

both observed and missing data (Little et al., 1987). Otherwise, if the missing data is not completely at random, the analysis based on only the observed data will lose all useful information from missing data. Especially, when the observed data is only a small fraction of the total number of data, the inference drawn from the observed values is questionable. Thus, it was of interest to find out the relationship between the missing data and the other covariates, such as metal contaminations metabolic potential and spatial information, and to include the information from the missing data into the analysis of microbial community structure.

In Chapter 3, the logistic regression was employed to investigate the relationship between the missingness and the concentration of the metal contaminants, the metabolic potential and the spatial information. A complete regression model combining the observed data as well as the covariate information from missing data was then fit using multiple imputations. The effects of lead, chromium and metabolic potential on the identification of significant bands (i.e., microbial populations) were then compared with the results from Becker et al.

Chapter 2 Modeling Spatial Dependence

2.1 Methods

2.1.1 Hierarchical Nested Model

A multi-scale tree-structured spatial model was applied on soil properties in previous work by Zhu et al. (2004). Similarly to the model from Zhu et al, a hierarchical model (3-way nested model) with random effects could be used to model the soil samples in Becker et al.

A linear 3-way nested model with random effects was used to analyze the spatial dependence of the metabolic potential, lead (Pb) and chromium (Cr) contaminants. The 3-way nested design is shown in Figure 2-1. Five arrays, seven sub-arrays for each array and seven sub-sub-arrays for each sub-array from sub-array1 to 6 were used in the studies. Yet the inferences are not to be confined to the particular arrays, sub-arrays and sub-sub-arrays selected in the study, but rather they are to pertain to all possible locations on each level of the hierarchical structure. Therefore, the effects of array, sub-array and sub-sub-array were considered as random factors. Each of the three sets of factor levels may be considered as the result of sampling a population about which inferences are to be drawn.

A 3-way nested model is

$$y_{ijkl} = \mu + \alpha_i + \beta_{j(i)} + \tau_{k(ij)} + \varepsilon_{l(ijk)};$$
$$i = 1,2,...,a,\ j = 1,2,...,b_i,\ k = 1,2,...,c_{ij}, \text{ and } l = 1,2,...,n_{ijk}.$$

The model represents the l^{th} observation from the k^{th} sub-sub-array within the j^{th} sub-array of array i. The random effects $(\alpha, \beta, \tau, \varepsilon)$ represent the effects of array, sub-array,

sub-sub-array and measurement error. They are assumed to have mutually independent normal distribution as followings:

$$\alpha_i \sim N(0,\sigma_\alpha{}^2), \quad i=1,2,\cdots,5;$$
$$\beta_{j(i)} \sim N(0,\sigma_\beta{}^2), \quad j(i)=0,1,\cdots,6;$$
$$\tau_{k(ij)} \sim N(0,\sigma_\tau{}^2), \quad k(ij)=\begin{cases}0,1,\cdots,6, & \text{if } j=1,2,\cdots,6;\\ 0, & \text{if } j=0.\end{cases}$$
$$\varepsilon_{l(ijk)} \sim N(0,\sigma^2), \quad l(ijk)=1,2,3.$$

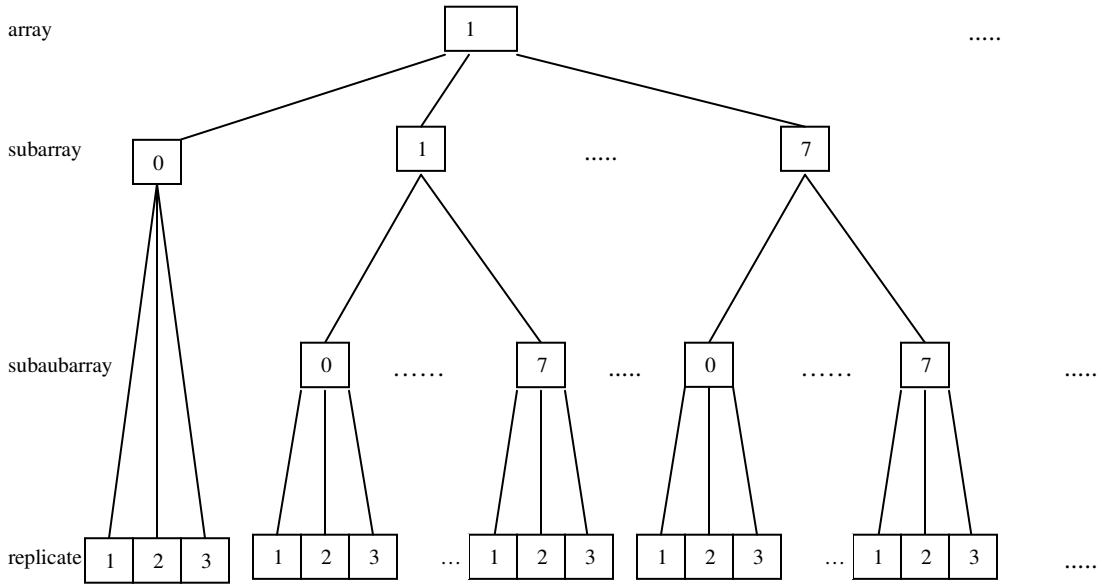

Figure 2-1 The 3-way nested design for the soil samples from Becker et al. (only illustrates the structure of array 1)

The variance of any observation is $(\sigma_\alpha^2+\sigma_\beta^2+\sigma_\tau^2+\sigma^2)$ where $\sigma_\alpha^2,\sigma_\beta^2,\sigma_\tau^2$ and σ^2 are referred to as variance components. Different responses are assumed to be independent except for the responses from the same array and/or from the same sub-array and/or from the same sub-sub-array. The spatial dependence of two responses on different locations was evaluated by the following covariance and correlation.

Observations on the same array have the following correlation.

$$\begin{aligned} Cov(Y_{ijkl}, Y_{ij'k'l'}) &= E[(\alpha_i + \beta_{j(i)} + \gamma_{k(ij)} + \varepsilon_{l(ijk)})(\alpha_i + \beta_{j'(i)} + \gamma_{k'(ij')} + \varepsilon_{l'(ij'k')})] \\ &= E(\alpha_i \alpha_i) + E(\beta_{j(i)} \beta_{j'(i)}) + E(\gamma_{k(ij)} \gamma_{k'(ij')}) + E(\varepsilon_{l(ijk)} \varepsilon_{l'(ij'k')}) \\ &= {\sigma_\alpha}^2 \end{aligned}$$

$$Cor(Y_{ijkl}, Y_{ij'k'l'}) = \frac{{\sigma_\alpha}^2}{{\sigma_\alpha}^2 + {\sigma_\beta}^2 + {\sigma_\gamma}^2 + \sigma^2}$$

Observations on the same array and same sub-array have the following correlation.

$$\begin{aligned} Cov(Y_{ijkl}, Y_{ijk'l'}) &= E[(\alpha_i + \beta_{j(i)} + \gamma_{k(ij)} + \varepsilon_{l(ijk)})(\alpha_i + \beta_{j(i)} + \gamma_{k'(ij)} + \varepsilon_{l'(ijk')})] \\ &= E(\alpha_i \alpha_i) + E(\beta_{j(i)} \beta_{j(i)}) + E(\gamma_{k(ij)} \gamma_{k'(ij)}) + E(\varepsilon_{l(ijk)} \varepsilon_{l'(ijk')}) \\ &= {\sigma_\alpha}^2 + {\sigma_\beta}^2 \end{aligned}$$

$$Cor(Y_{ijkl}, Y_{ijk'l'}) = \frac{{\sigma_\alpha}^2 + {\sigma_\beta}^2}{{\sigma_\alpha}^2 + {\sigma_\beta}^2 + {\sigma_\gamma}^2 + \sigma^2}$$

Observations on the same array, same sub-array and sub-sub-array have the following correlation.

$$\begin{aligned} Cov(Y_{ijkl}, Y_{ijkl'}) &= E[(\alpha_i + \beta_{j(i)} + \gamma_{k(ij)} + \varepsilon_{l(ijk)})(\alpha_i + \beta_{j(i)} + \gamma_{k(ij)} + \varepsilon_{l'(ijk)})] \\ &= E(\alpha_i \alpha_i) + E(\beta_{j(i)} \beta_{j(i)}) + E(\gamma_{k(ij)} \gamma_{k(ij)}) + E(\varepsilon_{l(ijk)} \varepsilon_{l'(ijk)}) \\ &= {\sigma_\alpha}^2 + {\sigma_\beta}^2 + {\sigma_\gamma}^2 \end{aligned}$$

$$Cor(Y_{ijkl}, Y_{ijkl'}) = \frac{{\sigma_\alpha}^2 + {\sigma_\beta}^2 + {\sigma_\gamma}^2}{{\sigma_\alpha}^2 + {\sigma_\beta}^2 + {\sigma_\gamma}^2 + \sigma^2}$$

Variance components can be estimated based on the analysis of variance method, which estimates expected mean square by the corresponding observed mean squares and solving for the variance components. The detailed formulas are given by D. M. Mahamunulu (1963) and shown in book (Searle et al., 1992, Appendix F.3). Those

formulas were applied to the soil sample data from Becker et al. (2006), which led to the following results.

$$b_{\cdot} = \sum_i b_i = 5\times 7 = 35;$$
$$c_{i\cdot} = \sum_j c_{ij} = 6\times 7 + 1 = 43;$$
$$c_{\cdot j} = \sum_i c_{ij} = \begin{cases} c_7 = 1\times 5 = 5; \\ c_1 = c_2 = c_3 = c_4 = c_5 = c_6 = 7\times 5 = 35; \end{cases}$$
$$N = \sum_i \sum_j \sum_k n_{ijk} = 645.$$

where b_i is the total number of sub-array in array i; c_{ij} is the total number of sub-sub-array in sub-array j under array i; n_{ijk} is the number of replicates in sub-sub array k within sub-array j under array i.

$$k_1 = \sum_i n_{i\cdot\cdot}^2 / N = (7\times 3\times 6 + 3)^2 \times 5 / 645 = 129;$$
$$k_2 = \sum_i \sum_j n_{ij\cdot}^2 / N = [((7\times 3)^2 \times 6 + 3^2)\times 5] / 645 = 20.58;$$
$$k_3 = \sum_i \sum_j \sum_k n_{ijk}^2 / N = 3^2 \times 43 \times 5 / 645 = 3;$$

$$k_4 = \sum_i \sum_j n_{ij\cdot}^2 / n_{i\cdot\cdot} = [(3\times 7)^2 \times 6 + 3^2] / 129 = 102.907;$$
$$k_5 = \sum_i \sum_j \sum_k n_{ijk}^2 / n_{i\cdot\cdot} = (3^2 \times 7 \times 6 + 3^2)\times 5 / 129 = 15;$$
$$k_6 = \sum_i \sum_j \sum_k n_{ijk}^2 / n_{ij\cdot} = 3^2 \times 7 \times 6 \times 5 / (3\times 7) + 3^2 \times 1 \times 5 / 3 = 105.$$

$$v_1 = N - k_1 = 645 - 129 = 516; \qquad v_2 = k_4 - k_2 = 102.907 - 20.58 = 82.327;$$
$$v_3 = k_5 - k_3 = 15 - 3 = 12; \qquad v_4 = a - 1 = 5 - 1 = 4;$$
$$v_5 = N - k_4 = 645 - 102.907 = 542.093; \qquad v_6 = k_6 - k_5 = 105 - 15 = 90;$$
$$v_7 = b_{\cdot} - a = 35 - 5 = 30; \qquad v_8 = N - k_6 = 645 - 105 = 540;$$
$$v_9 = c_{\cdot\cdot} - b_{\cdot} = 43\times 5 - 35 = 180; \qquad v_{10} = N - c_{\cdot\cdot} = 645 - 43\times 5 = 430.$$

The followings sum of squares can be calculated from data.

$$T_0 = \sum_i \sum_j \sum_k \sum_l y_{ijkl}^2, \quad T_A = \sum_i y_{i\cdots}^2 / n_{i\cdot\cdot}, \quad T_{AB} = \sum_i \sum_j y_{ij\cdot\cdot}^2 / n_{ij\cdot},$$
$$T_{ABC} = \sum_i \sum_j \sum_k y_{ijk\cdot}^2 / n_{ijk}, \text{ and } \quad T_\mu = y_{\cdots\cdot}^2 / N.$$

Then by equating the observed mean squares to their expected mean square values

$$\hat{E}(MS\alpha) = (T_A - T_\mu)/\nu_4 = \frac{\nu_1}{\nu_4}\hat{\sigma}_\alpha^2 + \frac{\nu_2}{\nu_4}\hat{\sigma}_\beta^2 + \frac{\nu_3}{\nu_4}\hat{\sigma}_\tau^2 + \hat{\sigma}^2 = 129\hat{\sigma}_\alpha^2 + 20.58\hat{\sigma}_\beta^2 + 3\hat{\sigma}_\tau^2 + \hat{\sigma}^2;$$

$$\hat{E}(MS\beta) = (T_{AB}\text{-}T_A)/\nu_7 = \frac{\nu_5}{\nu_7}\hat{\sigma}_\beta^2 + \frac{\nu_6}{\nu_7}\hat{\sigma}_\tau^2 + \hat{\sigma}^2 = 18.07\hat{\sigma}_\beta^2 + 3\hat{\sigma}_\tau^2 + \hat{\sigma}^2;$$

$$\hat{E}(MS\tau) = (T_{ABC}\text{-}T_{AB})/\nu_9 = \frac{\nu_8}{\nu_9}\hat{\sigma}_\tau^2 + \hat{\sigma}^2 = 3\hat{\sigma}_\tau^2 + \hat{\sigma}^2;$$

$$\hat{E}(MSE) = (T_0 - T_{ABC})/\nu_{10}. = \hat{\sigma}^2.$$

And the variance components were estimated as

$$\hat{\sigma}^2 = (T_0 - T_{ABC})/\nu_{10}, \quad \hat{\sigma}_\tau^2 = (T_{ABC} - T_{AB} - \nu_9\hat{\sigma}^2)/\nu_8,$$
$$\hat{\sigma}_\beta^2 = (T_{AB} - T_A - \nu_7\hat{\sigma}^2 - \nu_6\hat{\sigma}_\tau^2)/\nu_5,$$
$$\hat{\sigma}_\alpha^2 = (T_A - T_\mu - \nu_4\hat{\sigma}^2 - \nu_3\hat{\sigma}_\tau^2 - \nu_2\hat{\sigma}_\beta^2)/\nu_1.$$

For the above model with random factors and unequal treatment sample sizes, the analysis of variance (ANOVA) table for the above 3-way nested model is shown in Table 2-1. The sums of squares are calculated in the same way as in the fixed factors case (Neter et al., 1990). However, the test statistics are based on the expected mean squares for the random effects. Thus for the hypotheses $H_0 : \sigma_\alpha^2 = 0, H_0 : \sigma_\beta^2 = 0, H_0 : \sigma_\tau^2 = 0.$ the test statistics are $F_\alpha = \frac{MS\alpha}{MS\beta}$, $F_\beta = \frac{MS\beta}{MS\gamma}$ and $F_\gamma = \frac{MS\gamma}{MS\varepsilon}$.

Maximum likelihood method was used in the analysis of variance components. Rao and Heckler (1997) suggest that the maximum likelihood (ML) had a lower MSE when estimating variances in unbalanced design while restricted maximum likelihood (REML) had the less bias, and. also takes the fix effects into account. Since our case is unbalanced random design with random effects, the ML method was chosen.

Table 2-1 ANOVA table for 3-way nested model for unbalanced data from Becker et al.

(v_4, v_7, v_9, v_{10} and N can be found on previous page)

Source of Variation	Sum of Squares	Degrees of Freedom
Array (α)	$T_A - T_\mu$ $T_A = \sum_i y_{i...}^2 / n_{i..}$, $T_\mu = y_{....}^2 / N.$	v_4
Sub-array(within array) (β)	$T_{AB} - T_A$ $T_A = \sum_i y_{i...}^2 / n_{i..}$, $T_{AB} = \sum_i \sum_j y_{ij..}^2 / n_{ij.}$.	v_7
Sub-sub-array (within sub–array) (τ)	$T_{ABC} - T_{AB}$ $T_{AB} = \sum_i \sum_j y_{ij..}^2 / n_{ij.}$, $T_{ABC} = \sum_i \sum_j \sum_k y_{ijk.}^2 / n_{ijk}$.	v_9
Error (ε)	$T_0 - T_{ABC}$ $T_0 = \sum_i \sum_j \sum_k \sum_l y_{ijkl}^2$, $T_{ABC} = \sum_i \sum_j \sum_k y_{ijk.}^2 / n_{ijk}$,.	v_{10}
Total	$T_0 - T_\mu$ $T_0 = \sum_i \sum_j \sum_k \sum_l y_{ijkl}^2$, $T_\mu = y_{....}^2 / N.$	$v_4 + v_7 + v_9 + v_{10}$

It is noted that there are some drawbacks of this linear 3-way nested random effects model. It could be seen in Figure 1-1 that there are some areas in arrays which do not belong to any sub-array. And there are some areas in sub-arrays which do not belong to any sub-sub-array. Also the model addressed that two points from one array will always be more correlated than two points from different arrays. For example, according to above model, the points on the right side of array 1 are not assumed to be more correlated with points on the left side of array 2, but this is unlikely if strong patterns of spatial dependence are present.

2.1.2 Semivariogram Models and Kriging method

In order to compare the hierarchical model with the kriging model, SAS software (SAS Institute, Cary NC) was used to reproduce the analysis in the previous study by Becker et al. in GS+ software (Gamma Design Software, LLC, Plainwell, Michigan).

Using kriging, spatial dependence in soil properties can be described with a predictable spatial pattern. It is usually specified in the form of covariance or *semivariogram*, which qualifies the strength of association between neighbors as a function of pairwise distance. It is assumed that the spatial correlation does not depend on the locations of a pair of observations, but just on the distance between two observations.

The semivariogram is a half of mean squared difference of a pair of observation. It is similar in form to the typical variance estimator $Var(x) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$. A semivariogram $\gamma(h)$ can be calculated by $\gamma(h) = \frac{1}{2N(h)}\sum_{i=1}^{n}[Z_i - Z_{i+h})]^2$. *N(h)* represents the number of the observation pairs separated by distance *h*. Z_i and Z_{i+h} are the values of the observation at location *i* and *(i+h)* respectively. The relationship between the semivariogram and the covariance function is given by

$$\gamma(h) = \frac{1}{2}Var(Z_i - Z_j) = \frac{1}{2}\{Var(Z_i) + Var(Z_j) - 2Cov(Z_i, Z_j)\}.$$

If assuming that the mean and variance are constant over the region (2nd order stationary), such as, $Var(Z_i) = Var(Z_j) = \sigma^2$ and $Cov(Z_i, Z_j) = Cov(h)$, then $\gamma(h) = \sigma^2 - Cov(h) = \sigma^2[1 - Corr(h)]$.

The sample semivariogram is derived from the data and the pairwise distance *h*. Based on the sample semivariogram, the best fitted theoretical semivariogram model is obtained. The predictions on the unobserved locations are calculated from theoretical semivariogram model.

When two measurements taken at the same location are different, there exists *nugget effect*. *Nugget* (c_1) indicates the micro-scale variation or measurement error. It is the intercept of the semivariogram. *Sill* ($c_0 + c_1$) is the value of the semivariogram when two observations are far enough apart to be considered nearly uncorrelated. *Range* (a_0) is the distance beyond which the two observations are almost uncorrelated. It is the value of pairwise distance where the semivariogram reach the sill value. The *spatial dependence* $\left(\frac{c_0}{c_0 + c_1}\right)$ is defined by the proportion of the difference between sill and nugget in sill (Ettema et al., 2002). A typical semivariogram are shown in Figure 2-2 to illustrate the parameters in spherical semivariogram model and the meanings of spatial dependence.

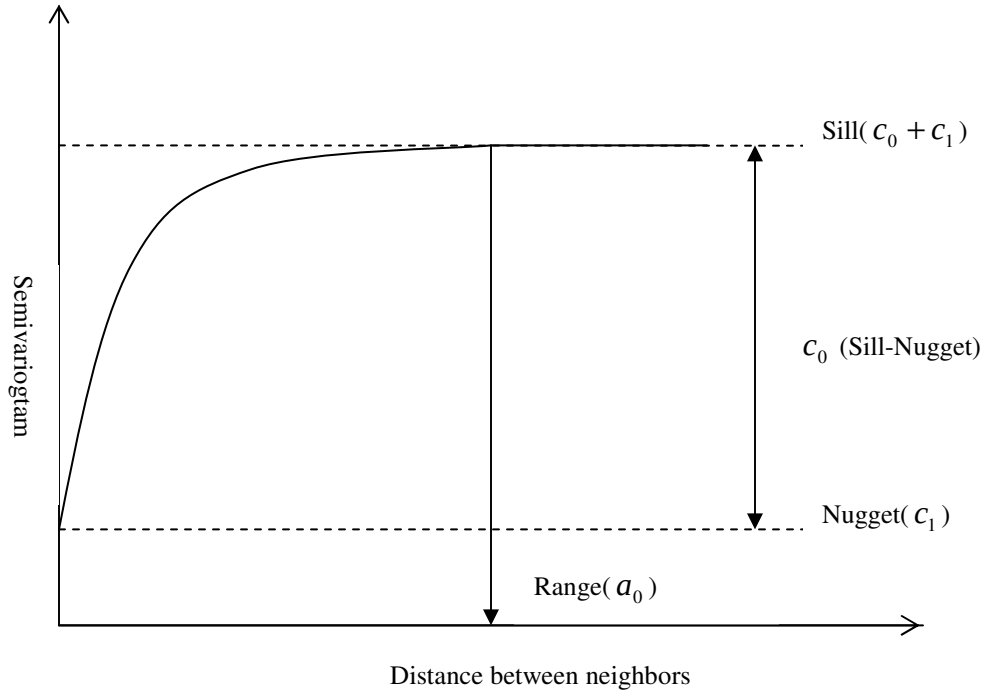

Figure 2-2 A typical theoretical semivariogram

There are several theoretical semivariograms which model the possible underlying spatial correlation, such as spherical semivariogram, Gaussian semivariogram, exponential semivariogram, power semivariogram and nested model (SAS Institute Inc., 2004). A spherical semivariogram model is given by

$$\gamma_z(h) = \begin{cases} c_1 + c_0\left[\frac{3}{2}\frac{h}{a_0} - \frac{1}{2}\left(\frac{h}{a_0}\right)^3\right], & \text{for } h \le a_0 \\ c_1 + c_0, & \text{for } h > a_0 \end{cases}.$$

A Gaussian semivariogram model is given by

$$\gamma_z(h) = c_1 + c_0\left[1 - \exp\left(-\frac{h^2}{a_0^2}\right)\right].$$

An exponential semivariogram model is given by

$$\gamma_z(h) = c_1 + c_0\left[1 - \exp\left(-\frac{h}{a_0}\right)\right].$$

And a power semivariogram model is given by

$$\gamma_z(h) = c_1 + c_0 h^{a_0}.$$

A so-called nested semivariogram model is a combination of any of above models.

In spatial analysis, if the correlation between the observations at two points depends not only on the distance, but also on the orientation of the two points, the model is called anisotropic. Otherwise, it is called isotropic. It is possible that in some directions the correlations are more than other directions.

Kriging is an interpolation method, which makes predictions to the unobserved values of the random variable Z. SAS procedures VARIOGRAM and KRIG2D were used to plot variogram and plot the predicted data surface. Since the NLIN procedure in SAS can find the least squares or weighted least squares estimates of the parameters of a nonlinear model, it was used to find the least square estimators of the parameters in the chosen theoretical semivariogram model. The R-squares of the model fitting were calculated as variance explained by model over the total variance.

To compare the hierarchical model and kriging model, the residuals from the 3-way nested model and the spherical semivariogram model for the observed points were calculated. Only array 1-3 were used in the comparison, because those three arrays were located on one plate and contiguous to each other. Further comparison was conducted on the predicted values on kriging grids from two models. The predicted values from kriging model were straightforward using SAS software. However, for the nested model, the

transformations on x-y coordinates to hexagonal coordinates were used. Arrays were assigned only according to the x coordinate.

2.2 Results

The analysis results from 3-way nested model for each of metabolic potential, lead and chromium concentration showed that the metabolic potential and metal contamination levels varied between arrays, sub-arrays and sub-sub-arrays. The R^2 shown in Table 2-2 indicate that the 3-way nested models fit well.

Table 2-2 3-way nested model fitting criteria (R^2) for log(Pb), log(Cr) and log(Potential)

	R^2
Log(Pb)	0.96
Log(Cr)	0.93
Log(Metabolic Potential)	0.94

The intra-class correlations induced by the hierarchical model (See Table 2-3) show that the two responses from the same array and the same sub-array are more correlated and the two responses from the same array sub-array and sub-sub-array are the most correlated. The correlations increase a lot when two responses are from same sub-array under the same array. The correlations increase slightly when two responses are from same sub-sub-array under the same array and sub-array.

Table 2-3 The correlations between two responses

Correlations	Y=Log(Pb)	Y=Log(Cr)	Y=log(Metabolic Potential)
$Cor(Y_{ijkl}, Y_{ij'k'l'})$	0.508	0.469	0.343
$Cor(Y_{ijkl}, Y_{ij'k'l'})$	0.778	0.748	0.616
$Cor(Y_{ijkl}, Y_{ij'k'l'})$	0.940	0.900	0.922

Variograms and kriging maps were reproduced in SAS and were compared with that from Becker et al. For simplicity, only isotropic semivariograms was considered here. After the different theoretical semivariogram models were compared to the sample semivariograms, the spherical models were found to be the closest models. The parameters of our fitted spherical models and R-squares are shown in Table 2-4. The parameters for metal contents are only slightly different from results by Becker et al. But the parameters for activity are much more different. Although the fittings of our kriging model from SAS are not as good as that from Becker et al. (R-squares in our test are slightly lower), our model showed that the samples had higher spatial dependence.

Table 2-4 The parameters and model fit criteria for fitting model to variogram.

	Model	Methods	Nugget (C1)	Sill-Nugget (C0)	Range (a0)	R^2	Spatial dependence C0/(C0-C1)
Log(Pb)	spherical	variogram	0.1846	3.1625	25.0277	0.4614	0.9448
Log(Cr)	spherical	variogram	0.1847	3.1625	25.0278	0.46	0.9448
Log(Activity)	spherical	variogram	0.1847	3.1625	25.0278	0.465	0.9448

The reproduced kriging maps are shown in Figure 2-3, which matches the kriging maps from Becker well. The kriging maps show that in some areas the increasing metal concentrations correspond to the decreasing net activity, while this relationship does not hold true everywhere in the plots. Some areas even show the contrary relationship. For example, on the right part of the plots (34<x<67), from the upright corner ((x, y) = (67, 21)) to ((x, y) = (34, -10.5)) the kriging maps show the increasing pattern for lead and

chromium levels and decreasing pattern of net activity. But on the lower left corner area, from ((x, y) = (-67, -10.5)) to ((x, y) = (-51, -21)) the level of lead, chromium and net activity have the same increasing fashion.

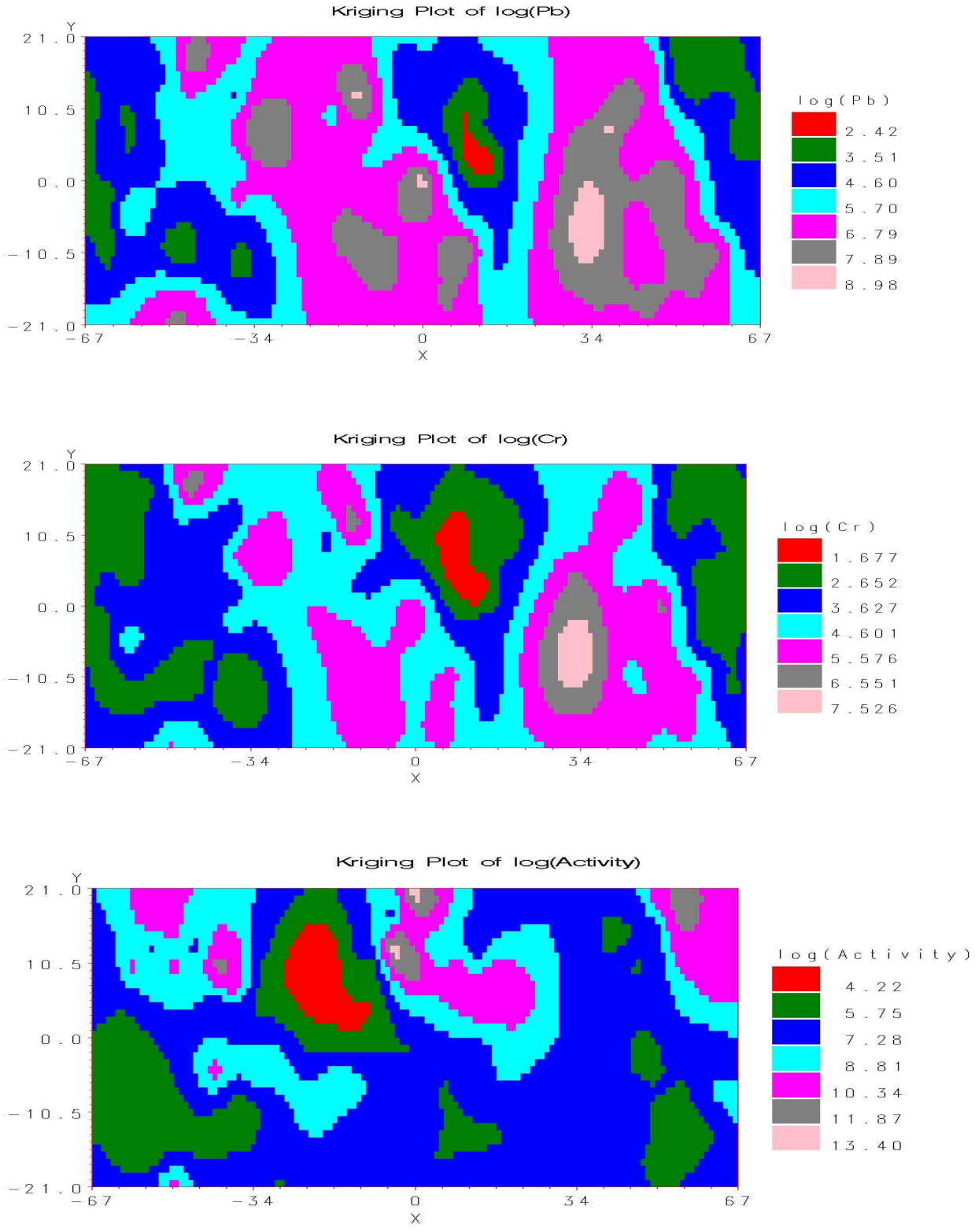

Figure 2-3 Kriging maps for log(Pb), log(Cr) and log(Activity).

To compare the model fit of the 3-way nested model (hierarchical model) and the kriging model, the sum of squared errors (SSE) and the sum of absolute errors ($\sum|e|$) on the observation points were calculated and listed in Table 2-5. Because the 3-way nested model (hierarchical model) had smaller SSE and smaller $\sum|e|$, it is clear that the 3-way nested model (hierarchical model) fitted better than kriging model.

Table 2-5 SSE and $\sum|e|$ s from 3-way nested model and spherical semivariogram model.

		Hierarchical	kriging
Log(Pb)	SSE	69.7710	78.1993
	$\sum\|e\|$	108.6877	121.0323
Log(Cr)	SSE	88.4821	94.3125
	$\sum\|e\|$	115.1850	122.8490
Log(Activity)	SSE	139.1124	160.1952
	$\sum\|e\|$	134.9042	160.3095

The contour plots of the differences between the nested model and the spherical semivariogram model on kriging grids were shown in Figure 2-4. Those contours of difference have the nearly opposite pattern to the kriging plots in Figure 2.3. It means that when the predictions from the kriging model are high, the predictions from the nested model are less than the predictions from the kriging model. When predictions from the kriging models are low, the predictions from nested model are greater than the predictions from the kriging model. Therefore, the nested model smoothes better.

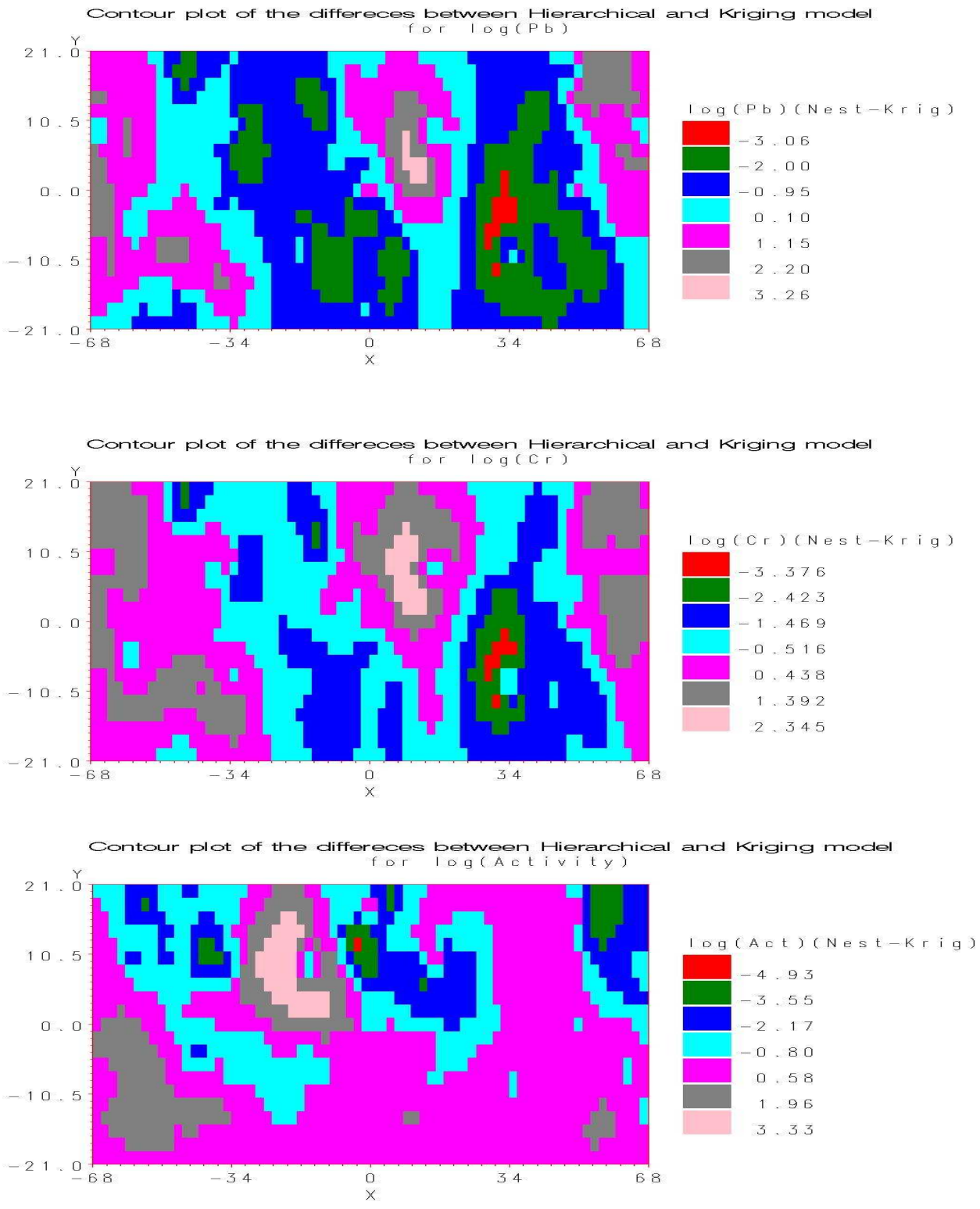

Figure 2-4 Contour maps of the difference between the nested model and the kriging model for log(Pb), log(Cr) and log(Activity)

2.3 Summary and Future Work

The above analysis shows that the improved 3-way nested model represents the spatial dependence of the metabolic potential and the metal contents in sampling area. By comparing it to the previous spherical model and kriging method, 3-way nested model fits the data better.

For the next stage, to obtain the effects of the heavy metal contents on metabolic potential, the overall hierarchical model based on all variables (lead, chromium and metabolic potential) should be analyzed. From the covariance components, the relationships between lead, chromium and metabolic potential can be found.

The overall hierarchical model will be

$$\begin{pmatrix} y_{1\,ijkl} \\ y_{2\,ijkl} \\ y_{3\,ijkl} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + \begin{pmatrix} \alpha_i^{(1)} \\ \alpha_i^{(2)} \\ \alpha_i^{(3)} \end{pmatrix} + \begin{pmatrix} \beta_{j(i)}^{(1)} \\ \beta_{j(i)}^{(2)} \\ \beta_{j(i)}^{(3)} \end{pmatrix} + \begin{pmatrix} \gamma_{k(ij)}^{(1)} \\ \gamma_{k(ij)}^{(2)} \\ \gamma_{k(ij)}^{(3)} \end{pmatrix} + \begin{pmatrix} \varepsilon_{l(ijk)}^{(1)} \\ \varepsilon_{l(ijk)}^{(2)} \\ \varepsilon_{l(ijk)}^{(3)} \end{pmatrix} .$$

where $\underset{\sim}{y}$ is the observed vector of lead, chromium and metabolic potential on each observation point. Vector $\underset{\sim}{\alpha}$ is the effect of array. Vector $\underset{\sim}{\beta}$ is the effect of sub-array. Vector $\underset{\sim}{\gamma}$ is the effect of sub-sub-array. Vector $\underset{\sim}{\varepsilon}$ is the effect of measurement error. The superscripts (1)-(3) are corresponding to variable lead, chromium and metabolic potential. The assumptions are $\underset{\sim i}{\alpha} \sim N(0,\Sigma_\alpha)$, $\underset{\sim j(i)}{\beta} \sim N(0,\Sigma_\beta)$, $\underset{\sim k(ij)}{\gamma} \sim N(0,\Sigma_\gamma)$ and $\underset{\sim l(ijk)}{\varepsilon} \sim N(0,\Sigma_\varepsilon)$. The variance and covariance matrices $\Sigma_\alpha, \Sigma_\beta$ and Σ_γ are assumed to have the forms below

$$\Sigma_\alpha = \begin{pmatrix} \sigma^2_{\alpha_{(1)}} & \tau_{\alpha_{(12)}} & \tau_{\alpha_{(13)}} \\ \tau_{\alpha_{(21)}} & \sigma^2_{\alpha_{(2)}} & \tau_{\alpha_{(23)}} \\ \tau_{\alpha_{(31)}} & \tau_{\alpha_{(32)}} & \sigma^2_{\alpha_{(3)}} \end{pmatrix} \otimes I_a , \qquad \Sigma_\beta = \begin{pmatrix} \sigma^2_{\beta_{(1)}} & \tau_{\beta_{(12)}} & \tau_{\beta_{(13)}} \\ \tau_{\beta_{(21)}} & \sigma^2_{\beta_{(2)}} & \tau_{\beta_{(23)}} \\ \tau_{\beta_{(31)}} & \tau_{\beta_{(32)}} & \sigma^2_{\beta_{(3)}} \end{pmatrix} \otimes I_{b_i} \quad \text{and}$$

$$\Sigma_\gamma = \begin{pmatrix} \sigma^2_{\gamma_{(1)}} & \tau_{\lambda_{(12)}} & \tau_{\gamma_{(13)}} \\ \tau_{\gamma_{(21)}} & \sigma^2_{\gamma_{(2)}} & \tau_{\gamma_{(23)}} \\ \tau_{\gamma_{(31)}} & \tau_{\gamma_{(32)}} & \sigma^2_{\gamma_{(3)}} \end{pmatrix} \otimes I_{c_{ij}} ,$$

where a, b_i and c_{ij} were defined as in 3-way nested model in page 6. The parameters $\tau_{\alpha_{(12)}}$, $\tau_{\alpha_{(13)}}$, $\tau_{\alpha_{(23)}}$, $\tau_{\beta_{(12)}}$, $\tau_{\beta_{(13)}}$, $\tau_{\beta_{(23)}}$, $\tau_{\gamma_{(12)}}$, $\tau_{\gamma_{(13)}}$ and $\tau_{\gamma_{(23)}}$ are the covariance components which we are interested in.

The current version of SAS (v. 9.1) is not able to perform analysis of multivariate nested models with random effects. Therefore, within the limited time of this project, the overall model fit and analyses of covariance components were not investigated using other software. Further study could continuously analyze the multivariate model on random effects and the covariance components.

Chapter 3 Microbial Community Analysis

3.1 Methods

3.1.1 Logistic Regression

Logistic regression model was used to determine whether there was a relationship between missingness of the presence of a band and the metal contents, the metabolic potential and the spatial coordinates. If let π_i represent the probability that sample i was not missing and 1-π_i the probability that sample i was missing, the logistic regression model can be written as follows.

$$\log it(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \underset{\sim}{\beta} \times \underset{\sim}{\mathrm{X}}$$

where $\underset{\sim}{X} = (1, \mathrm{Pb_i}, \mathrm{Cr_i}, \mathrm{Potential_i}, \mathrm{x_i}, \mathrm{y_i}, \mathrm{z_i})'$

$$\underset{\sim}{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6).$$

x_i, y_i and z_i are the three-dimensional spatial coordinates of the i^{th} sample.

To find the significant effects of lead, chromium, metabolic potential and spatial coordinates x and y on missingness, the hypotheses tests on each contribution from $\underset{\sim}{X}$ were carried out, such as, $H_0 : \beta_j = 0, j = 1,...,6$. It was based on the property of asymptotic distribution of the likelihood ratio test: $-2\log\lambda(X) \to \chi_1^2$ in distribution, where $\lambda(X)$ is the likelihood ratio test statistic and χ_1^2 is the chi-square distribution with 1 degree of freedom. If L_1 represents the maximized log-likelihood for the model with $\beta_j \neq 0$ and L_0 represents the maximized log-likelihood for the model with $\beta_j = 0$, the p value are calculated by 1- Chi(-2 (L_0 - L_1),1) . Chi(-2 (L_0 - L_1), 1) represents the

cumulative probability of Chi-square distribution with 1 degree of freedom at - 2 (L_0 - L_1). Then the elements of $\underset{\sim}{X}$ without significant effects were eliminated from the original logistic model. After the final model was selected, the least square estimator of parameters are $\underset{\sim}{\hat{\beta}}$ and the estimated probability with sample i not missing is

$$\hat{\pi}_i = \frac{\exp(\underset{\sim}{\hat{\beta}} \times \underset{\sim}{X})}{1 + \exp(\underset{\sim}{\hat{\beta}} \times \underset{\sim}{X})}.$$

3.1.2 Missing Data and Multiple Imputation

Since there were missing data during the experiment of obtaining DGGE gel, the missing data pattern need to be identified at first. In general, there are three general types of missing-data mechanisms (Little et al., 1987): missing completely at random (MCAR), missing at random (MAR) and neither missing at random nor observed at random. For MCAR, the probability of missingness is independent of response variable Y and covariates X. For MAR, the probability of missingness only depends on covariates X but not on response variable Y. For the third case, the probability of missingness depends on response variable Y and possible covariates X as well. In the previous microbial community analysis by Becker et al., the missing data were assumed to be MCAR. However, the missingness was found not independent of covariate X using logistic regression model as shown in section 3.2. It means that the assumption MCAR was not true for the data. In our study, the missing data is assumed to be MAR

To include the missing data in the analysis of the presence of a significant band, multiple imputations for missing data methods were used. For each missing value, the multiple imputation procedure generates a set of values and those values are combined

with the original data to be a set of complete data. Then they were used for analysis and the results are combined for inference (Yuan, 2000).

In our case, the presence of a band was represented by a binary variable (y), which was assumed to have a Bernoulli distribution with a probability ϕ. The probability of presence (ϕ) depended on covariates X, such as lead, chromium and metabolic potential. The relationship between them can be described by logistic regression as follow.

$$\log it(\phi_i) = \log\left(\frac{\phi_i}{1-\phi_i}\right) = \underset{\sim}{\beta} \times \underset{\sim}{X}$$

where $\underset{\sim}{X} = (1, \mathrm{Pb_i}, \mathrm{Cr_i}, \text{Potential})'$.

$\underset{\sim}{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$.

As mentioned before, some of data y were unobservable (missing). For each missing data y_i, the probability of its missingness (ϕ_i) was modeled by logistic regression on metal contents, metabolic potential and spatial coordinates as described in section 3.1.1. Note that none of $\underset{\sim}{X}$ were missing.

To estimate the parameter $\underset{\sim}{\beta}$ and the significant effects by lead, chromium and potential, the following procedures based on the logistic regression imputation method were carried out automatically by the MI proc of SAS. First, a logistic regression model for the presence of a band was fit based only on the observed data, and the estimated parameter $\underset{\sim}{\hat{\beta}}$ was obtained. Then based on this fitted model with $\underset{\sim}{\hat{\beta}}$, probability ϕ was calculated for each missing data. For each ϕ, multiple draws (y_j) from Bernoulli (ϕ) were combined with the observed data as multiple complete data sets. Finally, the logistic

regression for presence of a band was reapplied on each complete data set, and the estimations of parameter were combined to give results (Allison, 2005).

After the final logistic regression model was obtained, the significant effects of lead, chromium and metabolic potential were analyzed using hypothesis tests. Both overall logistic regression on all covariates (lead, chromium and metabolic potential) and the logistic regression on each individual covariate were fit for 68 bands data. In order to compare the results with that from Becker et al., the logistic regression on each of lead, chromium and net activity was also tested. The adjusted p values for controlling false discovery rate in multiple tests were used as described in details in the next section. The results of significant effects from overall logistic regression tests and contributions from each individual covariate were obtained. The adjusted p values and the significance on bands derived from individual covariate test were compared with the results from Becker et al.

3.1.3 False Discovery Rate

In testing the effect of lead, chromium and metabolic potential on the presence of bands in the microbial community profile, the overall test was based on multiple inferences of 68 multiple logistic regression tests. For multiple inferences, family-wise error rate control and false discovery rate (FDR) control are two methods commonly used to reduce the increased false positive (significance) rate. To obtain as many as possible discoveries (bands on which the covariates have significant effects) in our case, family-wised error rate control was not needed and controlling FDR is used (Benjamini et al., 1995).

FDR is the expected proportion of the false rejected null hypothesizes among all rejections of null hypothesizes (Benjamini et al., 1995). If the numbers of false rejections (rejection of null hypothesis when null hypothesis is true) is V and the number of correct rejections (the rejection of null hypothesis when alternative hypothesis is true) is S, the FDR (Q_e) is defined by

$$Q_e = \begin{cases} E\left[\frac{V}{V+S}\right], & \text{if } V+S \neq 0; \\ 0 & \text{if } V+S = 0. \end{cases}$$

To control the FDR in a multiple-test procedure, the p-values ($p_1, p_2, ..., p_m$) of m null hypothesizes $H_1, H_2, ..., H_m$ should be ordered as $p_{(1)}, p_{(2)}, ..., p_{(m)}$. For FDR at α, all $H_{(i)}, i = 1, 2, ..., k$ should be rejected, for k is the largest i which satisfies $P_{(i)} \leq \frac{i}{m}\alpha$ (Benjamini et al., 1995). The adjusted p-values (s) controlling the FDR used in SAS software is just an alternative way of the above calculation. The FDR adjusted *p*-values (s) are defined in step-up fashion (SAS Institute Inc., 1999):

$s_m = p_m$
$s_{(m-1)} = \min(s_m, [m/(m-1)]p_{(m-1)})$
$s_{(m-2)} = \min(s_{(m-1)}, [m/(m-2)]p_{(m-2)})$
...

3.2 Results

The logistic regression model for missingness was fitted for all considered covariates (lead, chromium, metabolic potential and coordinates) firstly. The test results from SAS outputs were listed in Table 3-1. At significant level $\alpha = 0.05$, p values of log(Pb),

log(Cr) and coordinates z are great than 0.05. Therefore, the effects of lead, chromium and coordinates z on the missingness were not significant. After those three variables were deleted, the new model had the test results from SAS outputs shown in Table 3-2. It showed that the missingness was significantly related with the metabolic potential and the spatial coordinates x and y.

Table 3-1 SAS output of logistic regression test for missingness using all covariates

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.9102	0.1755	118.4341	<.0001
Total_Pb	1	0.000010	0.000092	0.0128	0.9100
Total_Cr	1	-0.00004	0.000329	0.0184	0.8920
Potential	1	0.1071	0.0157	46.2861	<.0001
X	1	-0.00903	0.00340	7.0518	0.0079
Y	1	0.0616	0.0117	27.6984	<.0001
Z	1	0.000910	0.00194	0.2197	0.6392

Table 3-2 SAS output of logistic regression test for missingness based on metabolic potential and coordinates x and y

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.7514	0.1606	118.8560	<.0001
X	1	-0.0118	0.00336	12.3452	0.0004
Y	1	0.0453	0.0141	10.3822	0.0013
Potential	1	0.0639	0.0271	5.5710	0.0183

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	379.696	346.626
SC	383.654	362.460
-2 Log L	377.696	338.626

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	39.0696	3	<.0001
Score	40.4463	3	<.0001
Wald	34.2098	3	<.0001

From the above estimated logistic regression model, the probability of non-missing $\hat{\pi}_i$ was estimated for each band data. The histogram of $\hat{\pi}_i$ was plotted in Figure 3-1. The plot indicates that the band data were observed without missing most likely with probability around 0.06-0.24.

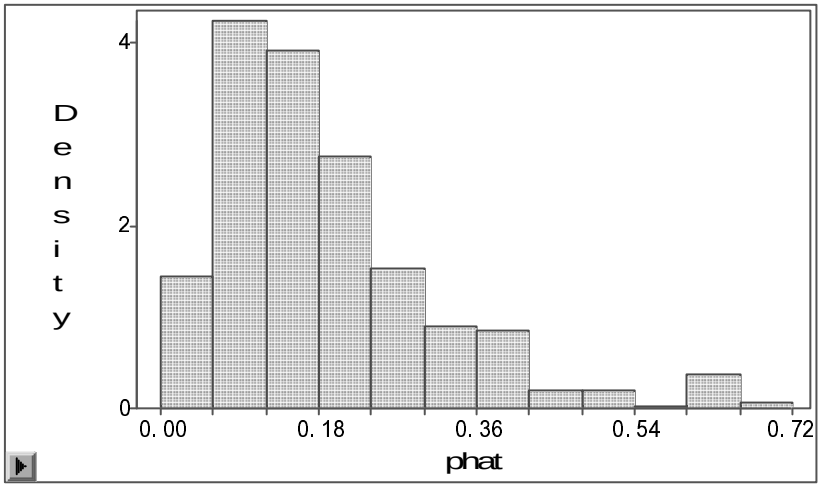

Figure 3-1 Histogram of the estimated probability of non-missing.

The plot of estimated non-missing probability by coordinates x and y (shown in Figure 3-2) gave an intuitive understanding of how missing data occurred in array 1-3. Three similar patterns from the left to right in the whole area of the plot correspond to the array 1, 2 and 3 in soil sample area. The plot indicates that the samples observed more likely appear around array 1 since very small probabilities of non-missing dominate the most areas in array 2 and 3 (in green and red color).

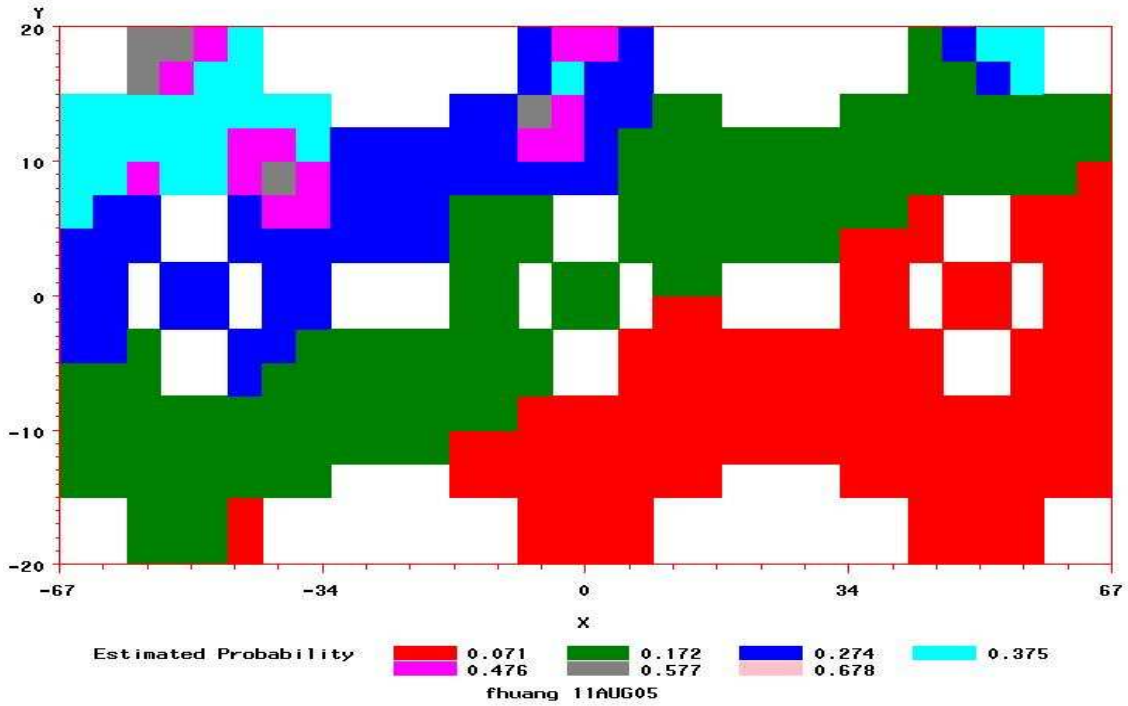

Figure 3-2 Estimated non-missing probabilities by coordinates (X and Y).

With the control of FDR, the overall logistic regression model for 68 significant bands data based on 1000 imputations for missing data showed that lead, chromium and metabolic potential had significant effects on band #3, #21, #41, #49, #52, #56, #22, #12 and #14 (overall test results shown in Table 3-2). Among those bands, band #3 was significantly affected by lead and metabolic potential, band #12 was significantly affected by all three variables (lead, Chromium and metabolic potential) and other bands were only significantly affected by metabolic potential (shown in Table 3-3).

Comparing the test results from the logistic regression models for each of lead, chromium and metabolic potential using 500 imputations on missing data with the results from Becker et al., the adjusted p-value by controlling FDR with imputations were larger than those from Becker et al. That means strong controls for false rejections in the tests using imputation because that the generated data from imputation for missing data

increased the information for analysis. In the analysis of logistic regression for 68 bands on chromium with 500 imputations, no significant effect was found on any band. But from the analysis of Becker et al. (2006), chromium had significant effect on band #22.

Table 3-3 The overall hypothesis tests results for 68 bands.

raw_p is the p-value of independent hypothesis test for each band. fdr_p is the adjusted p-values controlling FDR. The rows with green shading color indicate the bands which the overall effects of lead, chromium and metabolic potential are significant.

band	raw_p	fdr_p
3	0.0001	0.00136
21	0.0001	0.00136
41	0.0001	0.00136
49	0.0001	0.00136
52	0.0001	0.00136
56	0.0013	0.01473
22	0.0033	0.03206
12	0.005	0.03778
14	0.005	0.03778
16	0.0091	0.06182
2	0.01	0.06182
61	0.0139	0.07877
62	0.0315	0.15834
7	0.0326	0.15834
58	0.0408	0.18496
35	0.0448	0.1904
67	0.0496	0.1984
57	0.0716	0.24869
65	0.0738	0.24869
32	0.0749	0.24869
44	0.0768	0.24869
53	0.1004	0.31033
10	0.1235	0.35048
9	0.1237	0.35048
20	0.1706	0.4454
5	0.176	0.4454
39	0.1795	0.4454
4	0.1834	0.4454
31	0.2007	0.47061
15	0.2159	0.48937
27	0.2238	0.48939
29	0.2303	0.48939
51	0.2448	0.4968
28	0.2484	0.4968
24	0.3032	0.58907
30	0.3301	0.62352

46	0.3438	0.63185
54	0.3602	0.64457
18	0.4009	0.69901
40	0.421	0.7157
42	0.4577	0.75911
47	0.4793	0.77601
11	0.5302	0.81467
45	0.5365	0.81467
43	0.544	0.81467
36	0.5511	0.81467
50	0.5904	0.8542
37	0.6242	0.88067
1	0.6346	0.88067
55	0.6522	0.88073
63	0.6728	0.88073
64	0.6861	0.88073
13	0.6889	0.88073
60	0.6994	0.88073
38	0.7368	0.91095
8	0.7774	0.93091
6	0.7894	0.93091
17	0.8009	0.93091
68	0.8077	0.93091
33	0.8461	0.9555
23	0.8823	0.9555
59	0.8882	0.9555
48	0.9115	0.9555
34	0.9163	0.9555
25	0.9367	0.9555
26	0.9478	0.9555
66	0.9535	0.9555
19	0.9555	0.9555

Table 3-4 Significant tests for the effect of lead, chromium and metabolic potential among 9 selected bands.

The rows with green shading color indicate the metabolic potential have significant effects on the bands.
The rows with yellow shading color indicate the lead have significant effects on the bands
The rows with pink shading color indicate the chromium have significant effects on the bands.

Obs	Parm	Probt	band
1	Total_Pb	0.0014	3
2	Total_Cr	0.6427	3
3	Potential	0.0001	3
4	Total_Pb	0.9552	21
5	Total_Cr	0.589	21
6	Potential	<.0001	21
7	Total_Pb	0.5168	41
8	Total_Cr	0.4086	41

9	Potential	<.0001	41
10	Total_Pb	0.3199	49
11	Total_Cr	0.2783	49
12	Potential	<.0001	49
13	Total_Pb	0.9119	52
14	Total_Cr	0.984	52
15	Potential	<.0001	52
16	Total_Pb	0.0424	56
17	Total_Cr	0.0465	56
18	Potential	0.0006	56
19	Total_Pb	0.2252	22
20	Total_Cr	0.1168	22
21	Potential	0.0009	22
22	Total_Pb	0.0367	12
23	Total_Cr	0.0426	12
24	Potential	0.0265	12
25	Total_Pb	0.251	14
26	Total_Cr	0.2043	14
27	Potential	0.0016	14

3.3 Conclusion and Future work

The missing data is related to the metabolic potential and the spatial coordinates. The logistic regression model based on observed and unobservable DGGE fingerprint data indicated that the metabolic potential and the metal contents had significant effects on band #3, #21, #41, #49, #52, #56, #22, #12 and #14. Compared to the significant effects from analysis by Becker et al. where the missing data were ignored, the significant bands affected by chromium are different. This showed that the missingness has an effect on the microbial community analysis. Therefore, the missingness was not completely at random and should be taken in account.

The results of the effect of heavy metal contents on the presence of the bands in the logistic regression analysis might also be of interest to biologist. It was showed that the lead had a significant effect on band #3 and both lead and chromium had significant effects on band #12. Further experiments and analysis on those two bands would provide

more information on the microbial community changes caused by heavy metal contaminants.

Bibliography

Allison, P. D. Imputation of categorical variables with PROC MI. SAS Institute Inc., *Proceedings of the Thirtieth Annual SAS Users Group International Conference.* Cary, NC: SAS Institute Inc., 2005.

Becker, J. M., Parkin, T., Nakatsu, C. H., Wilbur, J. D. and Konopka, A. Bacterial activity, community structure, and cm-scale spatial heterogeneity in contaminated soil. *Microbial Ecology,* 2006. (in press)

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological), Vol.57, No.1 (1995), 289-300*

Etterna, C. H. and Wardle, D. A. Spatial soil ecology. *Trends in Ecology & Evolution, Vol.17 No.4 April 2002.*

Franklin, R. B. and Mills, A. L. Multi-scale variation in spatial heterogeneity for microbial community structure in an eastern Virginia agricultural field. *FEMS Microbiological Ecology 44(2003) 335-346.*

Little, R. J. A. and Rubin, D. B. Statistical analysis with missing data. *Wiley Series in Probability and Mathematical Statistics, 1987.*

Mahamunulu, D. M. Sampling variances of the estimates of variance components in the unbalanced 3-way nested classification. *The Annals of Mathematics Statistics, Vol.34, No. 2 (Jun., 1963), 521-527.*

Neter, J., Kutner, M. H., Nachtsheim, C. J. and Wasserman, W. Applied Statistical Models (Fourth Edition). *IRWIN Series in Statistics, 1990.*

Rao, P.S.R.S. and Heckler, C. E. The three-fold random effects model. *Journal of Statistical Planning and Inference*, *64 (1997) 341-352.*

SAS Institute Inc. SAS 8 Document. The MULTTEST Procedure. *SAS Institute, INC, Cary, NC. 1999.*

SAS Institute Inc. Documentation for SAS 9.1. Chapter 37 the KRIGE2D Procedure. *SAS Institute, INC, Cary, NC. 2004*.

Searle, S. R. Casella, G. and McCulloch, C. E. Variance components. *Wiley Series in Probability and Mathematical Statistics*, *1992.*

Yuan, Y. C. Multiple Imputation for Missing Data: Concepts and new development. SAS Institute Inc., *Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference.* Cary, NC: SAS Institute Inc., 2000.

Zhu, J., Morgan, C. L.S., Norman, J. M., Yue, W. and Lowery, B. Combined mapping of soil properties using a multi-scale tree-structured spatial model. *Elsevier Geoderma, 118 (2004), 821-334.*

Appendix: SAS Code

1. SAS code for 3-way nested model for each of lead, chromium and metabolic potential.

Calculate SSN and $\sum |e|$ for each model.

```
/* using nested model to analysis Pb,Cr and Metabolic Potential*/
/* calculate SSE and sum of absolute errors                    */

*imput data;
PROC IMPORT OUT= WORK.PbCrNet
            DATAFILE= "R:\project\DATA.xls"
            DBMS=EXCEL REPLACE;
     SHEET="docdata$";
     GETNAMES=YES;
     MIXED=NO;
     SCANTEXT=YES;
     USEDATE=YES;
     SCANTIME=YES;
RUN;
*****************************************************************;
* take log;
data logPbCrNet;
  set PbCrNet;
  l_Pb=log(Total_Pb);
  l_Cr=log(Total_Cr);
  l_Metabolic_Potential=log(Metabolic_Potential+1);
run;
*****************************************************************;
* macro to calculate SSE and Sum of the ABS(error);
%macro sse;
Proc iml;
  use resids;
        read all var {Z} into x;
        SSE=t(x)*x; * sum of square of errors;
        print SSE;
run;
data resids;
  set resids;
  ABSZ=abs(Z); * absolute errors;
  run;

Proc iml;
  use resids;
        read all var {ABSZ} into x;
        S_ABSE=sum(x); * Sum of absolute errors;
        print S_ABSE;
run;
quit;
%mend;
*****************************************************************;

/* Fit a nested model to log(lead) data;*/
proc glm data=logPbCrNet;
        class Array Sub Subsub Rep;
        Model l_Pb=Array Sub(Array) Subsub(Sub Array);
        random Array Sub(Array) Subsub(Sub Array)/test;
        output out=resids predicted=PRED residual=Z;
run;

%sse; *calculate SSE and Sum of the ABS(error);

Proc VARCOMP Method=ML; * Estimating variance components;
        Class Array Sub Subsub Rep;
        Model l_Pb=Array Sub(Array) Subsub(Sub Array);
```

```
run;
* check model assumption;
Proc gplot data=resids;
       plot  Z*PRED=Array Z*PRED=Sub Z*PRED=Subsub;
       plot  Z*Array Z*Sub Z*Subsub; * variability between different
arrays,subarrays,subsubarrays;
run;
quit;
* check normality of residuals;
proc univariate data=resids normal ;
       qqplot Z;
run;

/* Fit a nested model to log(Cr);*/

proc glm data=logPbCrNet;
       class Array Sub Subsub Rep;
       Model l_Cr=Array Sub(Array) Subsub(Sub Array);
       random Array Sub(Array) Subsub(Sub Array)/test;
       output out=resids predicted=PRED residual=Z;
run;
Proc VARCOMP Method=ML; * Estimating variance components;
       Class Array Sub Subsub Rep;
       Model l_Cr=Array Sub(Array) Subsub(Sub Array);
run;
Proc gplot data=resids;
       plot  Z*PRED=Array Z*PRED=Sub Z*PRED=Subsub;
       plot  Z*Array Z*Sub Z*Subsub; * variability between different
arrays,subarrays,subsubarrays;
run;
quit;
* check normality of residuals;
proc univariate data=resids noprint ;
       qqplot Z;
run;

%sse; * calculate SSE and Sum of the ABS(error);

/* Fit a nested model to log(Metabolic potential);*/

proc glm data=logPbCrNet;
       class Array Sub Subsub Rep;
       Model l_Metabolic_Potential=Array Sub(Array) Subsub(Sub Array);
       random Array Sub(Array) Subsub(Sub Array)/test;
       output out=resids predicted=PRED residual=Z;
run;
Proc VARCOMP Method=ML; * Estimating variance components;
       Class Array Sub Subsub Rep;
       Model l_Metabolic_Potential=Array Sub(Array) Subsub(Sub Array);
run;
Proc gplot data=resids;
       plot  Z*PRED=Array Z*PRED=Sub Z*PRED=Subsub;
       plot  Z*Array Z*Sub Z*Subsub; * variability between different
arrays,subarrays,subsubarrays;
run;
quit;
* check normality of residuals;
proc capability data=resids noprint;
       qqplot Z;
run;
%sse; * calculate SSE and Sum of the ABS(error);
```

2. SAS code for estimating theoretical semivariogram model and plotting kriging map.

```
* SAS Code for variogram and kriging;
*imput data;
PROC IMPORT OUT= WORK.PbCrNet
            DATAFILE= "R:\project\DATA.xls"
```

```
            DBMS=EXCEL REPLACE;
     SHEET="docdata$";
     GETNAMES=YES;
     MIXED=NO;
     SCANTEXT=YES;
     USEDATE=YES;
     SCANTIME=YES;
RUN;
*************************************************************************;
* take log;

data krigedata;
  set PbCrNet
  l_Pb=log(Total_Pb);
  l_Cr=log(Total_Cr);
  l_Activity=log(Activity+620);
  if (array gt 3) then delete;
run;

*****************************************************************;
* Using variogram model and Kriging on log lead data;
*****************************************************************;

* 3D plot-surface plot;
proc g3d data=krigedata;
      title 'Surface Plot';
      scatter X*Y=l_Pb / xticknum=5 yticknum=5
         grid zmin=0 zmax=15;
      label X = 'X'
            Y = 'Y'
            l_Pb = 'lead'
      ;
   run;
* using variogram to estimate the number of the lags;
proc variogram data=krigedata outdistance=outd;
      compute novariogram;
      coordinates xc=X yc=Y;
      var l_Pb;
   run;

   title 'OUTDISTANCE= Data Set Showing Distance Intervals';
   proc print data=outd;
   run;

   data outd; set outd;
      mdpt=round((lb+ub)/2,.1);
      label mdpt = 'Midpoint of Interval';
   run;

   axis1 minor=none;
   axis2 minor=none label=(angle=90 rotate=0);
   title 'Distribution of Pairwise Distances';
   proc gchart data=outd;
      vbar mdpt / type=sum sumvar=count discrete frame
                  cframe=ligr gaxis=axis1 raxis=axis2 nolegend;
   run;


* distribution of the pairwised distances;
proc variogram data=krigedata outdistance=outd;
      compute novariogram nhclasses=40;
      coordinates xc=X yc=Y;
      var l_Pb;
   run;
  title 'OUTDISTANCE= Data Set Showing Distance Intervals';
   proc print data=outd;
   run;

   data outd; set outd;
      mdpt=round((lb+ub)/2,.1);
      label mdpt = 'Midpoint of Interval';
```

```
   run;

   axis1 minor=none;
   axis2 minor=none label=(angle=90 rotate=0);
   title 'Distribution of Pairwise Distances';
   proc gchart data=outd;
      vbar mdpt / type=sum sumvar=count discrete frame
                  cframe=ligr gaxis=axis1 raxis=axis2 nolegend;
   run;


proc variogram data=krigedata outv=outv;
      compute lagd=1.75 maxlag=38 robust;
      coordinates xc=X yc=Y;
      var l_Pb;
   run;

   title 'OUTVAR= Data Set Showing Sample Variogram Results';
   proc print data=outv label;
      var lag count distance variog rvario;
   run;

   data outv2; set outv;
      vari=variog; type = 'regular'; output;
      vari=rvario; type = 'robust'; output;
   run;
Proc Print data=outv2;
run;
   title 'Standard and Robust Semivariogram for logPb Data';
   proc gplot data=outv2;
      plot vari*distance=type / frame cframe=ligr vaxis=axis2
                                haxis=axis1;
      symbol1 i=join l=1 c=blue   /* v=star   */;
      symbol2 i=join l=1 c=yellow /* v=square */;
      axis1 minor=none
            label=(c=black 'Lag Distance') /* offset=(3,3) */;
      axis2 order=(0 to 5 by 1) minor=none
            label=(angle=90 rotate=0 c=black 'Variogram')
            /* offset=(3,3) */;
   run;
   quit;

   proc print data=outv2;
   run;

*****************;
* Spherical model;
*****************;
* optimize spherical model;
 proc nlin data=outv2 method=Gauss hougaard;

      parms c0=2 to 5 by 0.5
                     c1=0 to 1 by 0.1
            a0=10 to 50 by 1;
      if distance gt a0 then
         model  variog =c0+c1; * variog and rvario are different;
          else
         model  variog = c1+c0*((3/2)*(distance /a0)-
0.5*(distance*distance*distance)/(a0*a0*a0));
          output out=variomod pred=gvhat;
run;


* fit optimized spherical model;
data outv3; set outv;
      c0=3.1625; a0=25.0277; c1=0.1846;
      if  distance gt a0 then vari=c0+c1;
      else vari = c1+c0*((3/2)*(distance /a0)-
0.5*(distance*distance*distance)/(a0*a0*a0));
      type = 'Spherical'; output;
      vari = variog; type = 'regular'; output;
```

```
      vari = rvario; type = 'robust'; output;
   run;

   title 'Theoretical and Sample Semivariogram for l_Pb';
   proc gplot data=outv3;
      plot vari*distance=type / frame cframe=ligr vaxis=axis2
                                haxis=axis1;
      symbol1 i=join l=1 c=blue    /* v=star    */;
      symbol2 i=join l=1 c=yellow  /* v=square  */;
      symbol3 i=join l=1 c=cyan    /* v=diamond */;
      axis1 minor=none
            label=(c=black 'Lag Distance') /* offset=(3,3) */;
      axis2 order=(0 to 5 by 1) minor=none
            label=(angle=90 rotate=0 c=black 'Variogram')
            /* offset=(3,3) */;
   run;
   quit;

* using proc krig2D to predict the unobserved data;
proc krige2d data=krigedata outest=est;
      pred var=l_Pb r=60;
      model nugget=0.1846 scale=3.1625 range=25.0277 form=SPHERICAL;
      coord xc=X yc=Y;
      grid x=-67 to 67 by 1 y=-21 to 21 by 1;
   run;


   proc g3d data=est;
      title 'Surface Plot of Kriged l_Pb';
      plot gxc*gyc=estimate/rotate=30;
 *     scatter gxc*gyc=estimate / grid;
      label gyc      = 'Y'
            gxc      = 'X'
            estimate = 'l_Pb'
            ;
   run;
goptions htitle=2 htext=2;

footnote1 ;
axis1 label = ("X");
axis2 label = ("Y");

 legend1 position=(right middle)
        label=(position=top 'log(Pb)')
        value=(height=2)
               SHAPE=BAR(6,4)
        across=1;
proc gcontour data=est;
   title  'Kriging Plot of log(Pb)';
   plot gyc*gxc=estimate / pattern
                        autolabel=(check=none)
                                        haxis=axis1
                        vaxis=axis2
                        legend=legend1;
run;
quit;
* plot the standard errors;
proc g3d data=est;
      title 'Surface Plot of Standard Errors of Kriging Estimates';
      scatter gxc*gyc=stderr / grid;
      label gyc    = 'Y'
            gxc    = 'X'
            stderr = 'Std Error'
            ;
   run;
*****************************************************************;
* Using semivariogram model and Kriging on log Chromium data;
*****************************************************************;

* 3D plot-surface plot;
proc g3d data=krigedata;
```

```
      title 'Surface Plot';
      scatter X*Y=l_Cr / xticknum=5 yticknum=5
         grid zmin=0 zmax=15;
      label X  = 'X'
            Y = 'Y'
            l_Cr = 'Chromate'
      ;
   run;
* using variogram to estimate the number of the lags;
proc variogram data=krigedata outdistance=outd;
      compute novariogram;
      coordinates xc=X yc=Y;
      var l_Cr;
   run;

   title 'OUTDISTANCE= Data Set Showing Distance Intervals';
   proc print data=outd;
   run;

   data outd; set outd;
      mdpt=round((lb+ub)/2,.1);
      label mdpt = 'Midpoint of Interval';
   run;

   axis1 minor=none;
   axis2 minor=none label=(angle=90 rotate=0);
   title 'Distribution of Pairwise Distances';
   proc gchart data=outd;
      vbar mdpt / type=sum sumvar=count discrete frame
                  cframe=ligr gaxis=axis1 raxis=axis2 nolegend;
   run;


* distribution of the pairwised distances;
proc variogram data=krigedata outdistance=outd;
      compute novariogram nhclasses=40;
      coordinates xc=X yc=Y;
      var l_Cr;
   run;
  title 'OUTDISTANCE= Data Set Showing Distance Intervals';
   proc print data=outd;
   run;

   data outd; set outd;
      mdpt=round((lb+ub)/2,.1);
      label mdpt = 'Midpoint of Interval';
   run;

   axis1 minor=none;
   axis2 minor=none label=(angle=90 rotate=0);
   title 'Distribution of Pairwise Distances';
   proc gchart data=outd;
      vbar mdpt / type=sum sumvar=count discrete frame
                  cframe=ligr gaxis=axis1 raxis=axis2 nolegend;
   run;


proc variogram data=krigedata outv=outv;
      compute lagd=1.75 maxlag=38 robust;
      coordinates xc=X yc=YZ;
      var l_Cr;
   run;

   title 'OUTVAR= Data Set Showing Sample Variogram Results';
   proc print data=outv label;
      var lag count distance variog rvario;
   run;

   data outv2; set outv;
      vari=variog; type = 'regular'; output;
      vari=rvario; type = 'robust'; output;
```

```
   run;

   title 'Standard and Robust Semivariogram for log(Cr)';
   proc gplot data=outv2;
      plot vari*distance=type / frame cframe=ligr vaxis=axis2
                                haxis=axis1;
      symbol1 i=join l=1 c=blue   /* v=star   */;
      symbol2 i=join l=1 c=yellow /* v=square */;
      axis1 minor=none
            label=(c=black 'Lag Distance') /* offset=(3,3) */;
      axis2 order=(0 to 5 by 1) minor=none
            label=(angle=90 rotate=0 c=black 'Variogram')
            /* offset=(3,3) */;
   run;
   quit;
******************;
* Spherical Model;
******************;
* optimize spherical model;
 proc nlin data=outv2 method=Gauss hougaard;

      parms c0=1 to 4 by 0.5
                     c1=0 to 1 by 0.1
            a0=10 to 50 by 2;
      if distance gt a0 then
         model  variog=c0+c1; * using variog and rvario are different;
          else
         model  variog = c1+c0*((3/2)*(distance /a0)-
0.5*(distance*distance*distance)/(a0*a0*a0));
          output out=variomod pred=gvhat;
run;

* fit optimized spherical model;
data outv3; set outv;
      c0=3.1625; a0=25.0278; c1=0.1847;
      if  distance gt a0 then vari=c0+c1;
      else vari = c1+c0*((3/2)*(distance /a0)-
0.5*(distance*distance*distance)/(a0*a0*a0));
      type = 'Spherical'; output;
      vari = variog; type = 'regular'; output;
      vari = rvario; type = 'robust'; output;
   run;

   title 'Theoretical and Sample Semivariogram for l_Cr';
   proc gplot data=outv3;
      plot vari*distance=type / frame cframe=ligr vaxis=axis2
                                haxis=axis1;
      symbol1 i=join l=1 c=blue    /* v=star    */;
      symbol2 i=join l=1 c=yellow  /* v=square  */;
      symbol3 i=join l=1 c=cyan    /* v=diamond */;
      axis1 minor=none
            label=(c=black 'Lag Distance') /* offset=(3,3) */;
      axis2 order=(0 to 5 by 1) minor=none
            label=(angle=90 rotate=0 c=black 'Variogram')
            /* offset=(3,3) */;
   run;
   quit;
* using proc krig2D to predict the unobserved data;
proc krige2d data=krigedata outest=est;
      pred var=l_Cr r=60;
      model nugget=0.1847 scale=3.1625 range=25.0278 form=SPHERICAL;
      coord xc=X yc=Y;
      grid x=-67 to 67 by 1 y=-21 to 21 by 1;
   run;

   proc g3d data=est;
      title 'Surface Plot of Kriged l_Cr';
      plot gxc*gyc=estimate/rotate=30;
               *scatter gxc*gyc=estimate / grid;
      label gyc      = 'Y'
            gxc      = 'X'
```

```
            estimate = 'l_Cr'
            ;
   run;
goptions htitle=2 htext=2;

footnote1 ;
axis1 label = ("X");
axis2 label = ("Y");

 legend1 position=(right middle)
        label=(position=top 'log(Cr)')
        value=(height=2)
               SHAPE=BAR(6,4)
        across=1;
 proc gcontour data=est;
   title  'Kriging Plot of log(Cr)';
   plot gyc*gxc=estimate / pattern
                        autolabel=(check=none)
                                        haxis=axis1
                        vaxis=axis2
                        legend=legend1;
run;
quit;
* plot the standard errors;
   proc g3d data=est;
      title 'Surface Plot of Standard Errors of Kriging Estimates';
      scatter gxc*gyc=stderr / grid;
      label gyc    = 'Y'
            gxc    = 'X'
            stderr = 'Std Error'
            ;
   run;

*****************************************************************;
* Using semivariogram model Kriging on log Activity data;
*****************************************************************;

* 3D plot-surface plot;
proc g3d data=krigedata;
      title 'Surface Plot';
      scatter X*Y=l_Activity / xticknum=5 yticknum=5
         grid zmin=0 zmax=15;
      label X  = 'X'
            YZ = 'Y'
            l_Activity = 'ln(Activity)'
      ;
   run;
* using variogram to estimate the number of the lags;
proc variogram data=krigedata outdistance=outd;
      compute novariogram;
      coordinates xc=X yc=Y;
      var l_Activity;
   run;

   title 'OUTDISTANCE= Data Set Showing Distance Intervals';
   proc print data=outd;
   run;

   data outd; set outd;
      mdpt=round((lb+ub)/2,.1);
      label mdpt = 'Midpoint of Interval';
   run;

   axis1 minor=none;
   axis2 minor=none label=(angle=90 rotate=0);
   title 'Distribution of Pairwise Distances';
   proc gchart data=outd;
      vbar mdpt / type=sum sumvar=count discrete frame
                  cframe=ligr gaxis=axis1 raxis=axis2 nolegend;
   run;
```

```
* distribution of the pairwised distances;
proc variogram data=krigedata outdistance=outd;
      compute novariogram nhclasses=40;
      coordinates xc=X yc=Y;
      var l_Activity;
   run;
  title 'OUTDISTANCE= Data Set Showing Distance Intervals';
   proc print data=outd;
   run;

   data outd; set outd;
      mdpt=round((lb+ub)/2,.1);
      label mdpt = 'Midpoint of Interval';
   run;

   axis1 minor=none;
   axis2 minor=none label=(angle=90 rotate=0);
   title 'Distribution of Pairwise Distances';
   proc gchart data=outd;
      vbar mdpt / type=sum sumvar=count discrete frame
                  cframe=ligr gaxis=axis1 raxis=axis2 nolegend;
   run;


proc variogram data=krigedata outv=outv;
      compute lagd=1.75 maxlag=38 robust;
      coordinates xc=X yc=Y;
      var l_Activity;
   run;

   title 'OUTVAR= Data Set Showing Sample Variogram Results';
   proc print data=outv label;
      var lag count distance variog rvario;
   run;

   data outv2; set outv;
      vari=variog; type = 'regular'; output;
      vari=rvario; type = 'robust'; output;
   run;

   title 'Standard and Robust Semivariogram for l_Activity';
   proc gplot data=outv2;
      plot vari*distance=type / frame cframe=ligr vaxis=axis2
                                haxis=axis1;
      symbol1 i=join l=1 c=blue   /* v=star   */;
      symbol2 i=join l=1 c=yellow /* v=square */;
      axis1 minor=none
            label=(c=black 'Lag Distance') /* offset=(3,3) */;
      axis2 order=(0 to 5 by 1) minor=none
            label=(angle=90 rotate=0 c=black 'Variogram')
            /* offset=(3,3) */;
   run;
   quit;

******************;
* spherical Model;
******************;
* optimize spherical model;
 proc nlin data=outv2 method=Gauss hougaard;
      parms c0=1 to 5 by 0.5
                    c1=0 to 1 by 0.1
            a0=10 to 40 by 2;
      if distance gt a0 then
         model  variog=c0+c1; * using variog and rvario are different;
          else
         model  variog = c1+c0*((3/2)*(distance /a0)-
0.5*(distance*distance*distance)/(a0*a0*a0));
          output out=variomod pred=gvhat;
run;
```

```
* fit optimized spherical model;
data outv3; set outv;
      c0=3.1625; a0=25.0278; c1=0.1847;
      if  distance gt a0 then vari=c0+c1;
      else vari = c1+c0*((3/2)*(distance /a0)-
0.5*(distance*distance*distance)/(a0*a0*a0));
      type = 'Spherical'; output;
      vari = variog; type = 'regular'; output;
      vari = rvario; type = 'robust'; output;
   run;

   title 'Theoretical and Sample Semivariogram for l_Activity';
   proc gplot data=outv3;
      plot vari*distance=type / frame cframe=ligr vaxis=axis2
                                haxis=axis1;
      symbol1 i=join l=1 c=blue    /* v=star    */;
      symbol2 i=join l=1 c=yellow  /* v=square  */;
      symbol3 i=join l=1 c=cyan    /* v=diamond */;
      axis1 minor=none
            label=(c=black 'Lag Distance') /* offset=(3,3) */;
      axis2 order=(0 to 6 by 1) minor=none
            label=(angle=90 rotate=0 c=black 'Variogram')
            /* offset=(3,3) */;
   run;
   quit;

* using proc krig2D to predict the unobserved data;
proc krige2d data=krigedata outest=est;
      pred var=l_Activity r=60;
      model nugget=0.1847 scale=3.1625 range=25 form=SPHERICAL;
      coord xc=X yc=Y;
      grid x=-67 to 67 by 1 y=-21 to 21 by 1;
   run;

   proc g3d data=est;
      title 'Surface Plot of Kriged l_Activity';
      plot gyc*gxc=estimate/rotate=30;
*scatter gxc*gyc=estimate / grid;
      label gyc      = 'Y'
            gxc      = 'X'
            estimate = 'l_Activity'
            ;
   run;
 goptions htitle=2 htext=2;

footnote1 ;
axis1 label = ("X");
axis2 label = ("Y");

 legend1 position=(right middle)
        label=(position=top 'log(Activity)')
        value=(height=2)
               SHAPE=BAR(6,4)
        across=1;
 proc gcontour data=est;
   title  'Kriging Plot of log(Activity)';
   plot gyc*gxc=estimate / pattern
                        autolabel=(check=none)
                                        haxis=axis1
                        vaxis=axis2
                        legend=legend1;
run;
quit;
* plot the standard errors;
   proc g3d data=est;
      title 'Surface Plot of Standard Errors of Kriging Estimates';
      scatter gxc*gyc=stderr / grid;
      label gyc    = 'Y'
            gxc    = 'X'
            stderr = 'Std Error'
            ;
```

```
   run;
```

3. Comparing predicted difference between hierarchical model and krige model

```
* SAS Code for comparing predicted difference between hierarchical model and krige model;

data array123;
 input Sample$ replicate$     Array Sub Subsub Rep X Y Z Activity TotalPb TotalCr
Potential;
 cards;
…
;
run;

* take log;
data array123;
  set array123;
  l_Pb=log(TotalPb);
  l_Cr=log(TotalCr);
  l_Activity=log(Activity+620);
run;

************************************************************;
* Transform the xy coordinates to array-sub-subsub coordinates;
************************************************************;

***********Assign Array # *******;
%macro array;
data tran;
 set krigest;
 if gxc < -25 then array=1;
 else if gxc > 25 then array=3;
 else array=2;
 run;

 proc print data=tran;
 run;
%mend;

***********Assign Sub # *********;
* Algorithm:
* 1. transform the original cardinal coordinates in krig model to the cardinal
coordinates using the center of each
*    hexagon as the origin.
* 2. transform the new cardinal coordinates (x,y) to the hexagonal coordinates.
* 3. repeat 1 & 2 for each array;


%macro sub;

%do l = 1 %to 3; * from array 1 to array 3;
  data tran&l;
  set tran;
  if array ne &l then delete;
  s0=0;s1=1;s2=2;s3=3;s4=4;s5=5;s6=6;* # of the subarray;
  dx0=0;dx1=(-1);dx2=(-1);dx3=0;dx4=1;dx5=1;dx6=0;* the unit offsets from the center of
subarray to (0,0)coordinates on x;
  dy0=0;dy1=(-1);dy2=1;dy3=2;dy4=1;dy5=(-1);dy6=(-2);* the unit offsets from the center
of subarray to (0,0)coordinates on y;

  x1=12; y1=6.93; * the magnitude on the unit offsets;
  %do m = 0 %to 6;
      if ABS(gxc + 50 * (2 - &l) + x1 * dx&m - (gyc + y1 * dy&m) / 1.732) <= 8
           and ABS(gxc + 50 * (2 - &l) + x1 * dx&m + (gyc + y1 * dy&m) / 1.732) <= 8
            and ABS((gyc + y1 * dy&m) / 0.866) <= 8 then sub = s&m;
      *else sub=7;* it did not work. why?;
  %end;

run;
%end;
```

```
data transub;
 set tran1 tran2 tran3;
run;
%mend;

******** assign subsub # ***********;
* split the data to two parts(in/out of array boundary);
%macro split;
Data outsub;
       set transub;
       if sub ne . then delete;
       subsub = .;
       keep VARNAME GXC GYC ESTIMATE  STDERR array sub subsub;
run;
proc print data=outsub;
run;
%mend;

%macro subsub;
%do l = 1 %to 3;
  data transub&l;
         set transub;
         if array ne &l then delete;
         x1=12; y1=6.93; * the magnitude on the unit offsets on subarray level;
         x2=4; y2=2.309; * the magnitude on the unit offsets on subsubarray level;
         %do n = 0 %to 6;
             data transub&l&n; * split data into 3*6 groups by array and subarray;
                    set transub&l;
                    if sub ne &n then delete;
                    %do m = 0 %to 6;
                         if ABS(gxc+50*(2-&l)+x1*dx&m+x2*dx&m -
(gyc+y1*dy&m+y2*dy&m)/1.732)<=2.666
                                and ABS(gxc+50*(2-&l)+x1*dx&m+x2*dx&m
+(gyc+y1*dy&m+y2*dy&m)/1.732)<=2.666
                                 and ABS((gyc+y1*dy&m+y2*dy&m)/0.866)<=2.666 then
subsub=s&m;
                        %end;
                run;


         %end;
         run;

%end;
data transs1;
       set transub10 transub11 transub12 transub13 transub14 transub15 transub16;
run;
data transs2;
              set transub20 transub21 transub22 transub23 transub24 transub25 transub26;
run;
data transs3;
              set transub30 transub31 transub32 transub33 transub34 transub35 transub36;
run;
data transubsub;
 set transs1 transs2 transs3;
run;
%mend;

%macro combinetran;
data transubsub;
  set transubsub;
  keep VARNAME GXC GYC ESTIMATE  STDERR array sub subsub;
run;

proc print data=transubsub;
run;

* combine two parts (in/out of array boundary) together;
data transubsub_T;
  set transubsub outsub;
run;
```

```
proc print data=transubsub_T;
run;

proc sort data=transubsub_T;
  by array;
  run;
%mend;

*************************************************************;
* predict the log(Pb/Cr/Activity) for transformed data on grids;
*************************************************************;

*********************************************************************;
* calcultate the difference between the predictions in kriging model and those in nested
model;
*********************************************************************;
%macro diff;

* delete observed data;
data nest_pred;
   set resids;
   if gxc=. then delete;
   run;

* sort data of predictions from both models in order to merge them together;
proc sort data=nest_pred;
 by gxc gyc;
 run;
proc sort data=krigest;
 by gxc gyc;
 run;
proc print data=nest_pred;
run;
proc print data=krigest;
run;

* calculate prediction from hierarchical - prediction from kriging on each grid;
data preddif;
   merge nest_pred krigest;
   nest_krig = pred - estimate;
   run;
proc print data=preddif;
run;
title'Scatter plot of the differeces between Hierarchical and Kriging model';
proc g3d data=preddif;
  scatter gyc*gxc=nest_krig;
  run;

%mend;

******************;
*  For Log(Pb)   *;
******************;
* using proc krig2D to predict the unobserved data;
proc krige2d data=array123 outest=krigest;
      pred var=l_Pb r=60;
      model nugget=0.1846 scale=3.1625 range=25.0277 form=SPHERICAL;
      coord xc=X yc=Y;
      grid x=-68 to 68 by 2 y=-21 to 21 by 2;
   run;
proc print data=krigest;
run;
*************************************************************;
* Transform the xy coordinates to array-sub-subsub coordinates;

* Assign Array # ;
%array;
* Assign Sub # ;
%sub;
proc print data=transub;
```

```
run;
* split the data to two parts(in/out of array boundary);
%split;
* Assign Subsub # ;
%subsub;
proc print data=transubsub;
run;
* combine two parts (in/out of array boundary) together;
%combinetran;

************************************************************;
* predict the log(Pb) for transformed data on grids;

*seperate the data by Pb , Cr and Potential;
data obs_Pb;
   set array123;
   *if array gt 3 then delete;
   gxc= . ; gyc= . ;
   keep array sub subsub gxc gyc rep l_Pb;
run;
Proc print data=obs_Pb; run;

* calculate mean from observed values;
* it will be used as the prediction on points without sub or subsub value;
Proc means data=obs_Pb;
  var l_Pb;
  output out=pbout mean=m_logPb;
  run;

proc print data=pbout ;run;

* prepare the identical structure for transformed data;

Data new;
  set transubsub_T;
  l_Pb = . ; * set the value of response which we want to predict as missing;
  rep = 1 ;  * glm do not predict new response without assign rep(any of 1,2,3 has same
results ;
  keep gxc gyc array sub subsub rep l_Pb;
run;
proc print data=new; run;

* put observed data and transformed data together;
data all;
       set obs_Pb new;
run;
proc print data=all; run;

* Fit a nested model to log data;
proc glm data=all;
       class Array Sub Subsub Rep;
       Model l_Pb=Array Sub(Array) Subsub(Sub Array);
       random Array Sub(Array) Subsub(Sub Array)/TEST;
       output out=resids predicted=PRED residual=Z;
run;
proc print data=resids;
run;
data resids;
       set resids;
       if sub=. then pred=5.768;  * set predicted values on other points to estimated
average(5.768);
       if subsub=. then pred=5.768;
run;

************************************************************************************
*****;
* calcultate the difference between the predictions in kriging model and those in nested
model;

%diff;
goptions htitle=2 htext=2;
```

```
footnote1 ;
axis1 label = ("X");
axis2 label = ("Y");

 legend1 position=(right middle)
        label=(position=top 'log(Pb)(Nest-Krig)')
        value=(height=2)
               SHAPE=BAR(6,4)
        across=1;

title1'Contour plot of the differeces between Hierarchical and Kriging model';
title2' for log(Pb)';
proc gcontour data=preddif;
  plot gyc*gxc=nest_krig/pattern
                                        autolabel=(check=none)
                                        haxis=axis1
                      vaxis=axis2
                      legend=legend1;;

  run;
  quit;
******************;
*  For Log(Cr)   *;
******************;

* using proc krig2D to predict the unobserved data;
proc krige2d data=array123 outest=krigest;
      pred var=l_Cr r=60;
      model nugget=0.1846 scale=3.1625 range=25.0277 form=SPHERICAL;
      coord xc=X yc=Y;
      grid x=-68 to 68 by 2 y=-21 to 21 by 2;
   run;
proc print data=krigest;
run;
************************************************************;
* Transform the xy coordinates to array-sub-subsub coordinates;

* Assign Array # ;
%array;
* Assign Sub # ;
%sub;
proc print data=transub;
run;
* split the data to two parts(in/out of array boundary);
%split;
* Assign Subsub # ;
%subsub;
proc print data=transubsub;
run;
* combine two parts (in/out of array boundary) together;
%combinetran;

************************************************************;
* predict the log(Cr) for transformed data on grids;

*seperate the data by Pb , Cr and Potential;
data obs_Cr;
   set array123;
   *if array gt 3 then delete;
   gxc= . ; gyc= . ;
   keep array sub subsub gxc gyc rep l_Cr;
run;
Proc print data=obs_Cr; run;

* calculate mean from observed values;
* it will be used as the prediction on points without sub or subsub value;
Proc means data=obs_Cr;
  var l_Cr;
  output out=Crout mean=m_logCr;
  run;
```

```
proc print data=Crout ;run;

* prepare the identical structure for transformed data;

Data new;
  set transubsub_T;
  l_Cr = . ; * set the value of response which we want to predict as missing;
  rep = 1 ;  * glm do not predict new response without assign rep(any of 1,2,3 has same
results ;
  keep gxc gyc array sub subsub rep l_Cr;
run;
proc print data=new; run;

* put observed data and transformed data together;
data all;
       set obs_Cr new;
run;
proc print data=all; run;

* Fit a nested model to log data;
proc glm data=all;
       class Array Sub Subsub Rep;
       Model l_Cr=Array Sub(Array) Subsub(Sub Array);
       random Array Sub(Array) Subsub(Sub Array)/TEST;
       output out=resids predicted=PRED residual=Z;
run;
proc print data=resids;
run;
data resids;
       set resids;
       if sub=. then pred=4.0794;  * set predicted values on other points to estimated
average(5.768);
       if subsub=. then pred=4.0794;
run;

*******************************************************************************************
*****;
* calcultate the difference between the predictions in kriging model and those in nested
model;

%diff;
goptions htitle=2 htext=2;

footnote1 ;
axis1 label = ("X");
axis2 label = ("Y");

 legend1 position=(right middle)
        label=(position=top 'log(Cr)(Nest-Krig)')
        value=(height=2)
               SHAPE=BAR(6,4)
        across=1;

title1'Contour plot of the differeces between Hierarchical and Kriging model';
title2' for log(Cr)';
proc gcontour data=preddif;
  plot gyc*gxc=nest_krig/pattern
                                        autolabel=(check=none)
                                        haxis=axis1
                      vaxis=axis2
                      legend=legend1;;

  run;
  quit;

***********************;
*  For Log(Activity)   *;
***********************;

* using proc krig2D to predict the unobserved data;
```

```
proc krige2d data=array123 outest=krigest;
      pred var=l_Activity r=60;
      model nugget=0.1846 scale=3.1625 range=25.0277 form=SPHERICAL;
      coord xc=X yc=Y;
      grid x=-68 to 68 by 2 y=-21 to 21 by 2;
   run;
proc print data=krigest;
run;
************************************************************;
* Transform the xy coordinates to array-sub-subsub coordinates;

* Assign Array # ;
%array;
* Assign Sub # ;
%sub;
proc print data=transub;
run;
* split the data to two parts(in/out of array boundary);
%split;
* Assign Subsub # ;
%subsub;
proc print data=transubsub;
run;
* combine two parts (in/out of array boundary) together;
%combinetran;

************************************************************;
* predict the log(Activity) for transformed data on grids;

*seperate the data by Pb , Cr and Potential;
data obs_Activity;
   set array123;
   *if array gt 3 then delete;
   gxc= . ; gyc= . ;
   keep array sub subsub gxc gyc rep l_Activity;
run;
Proc print data=obs_Activity; run;

* calculate mean from observed values;
* it will be used as the prediction on points without sub or subsub value;
Proc means data=obs_Activity;
  var l_Activity;
  output out=Actout mean=m_logact;
  run;

proc print data=Actout ;run;

* prepare the identical structure for transformed data;

Data new;
  set transubsub_T;
  l_Activity = . ; * set the value of response which we want to predict as missing;
  rep = 1 ;  * glm do not predict new response without assign rep(any of 1,2,3 has same
results ;
  keep gxc gyc array sub subsub rep l_Activity;
run;
proc print data=new; run;

* put observed data and transformed data together;
data all;
       set obs_Activity new;
run;
proc print data=all; run;

* Fit a nested model to log data;
proc glm data=all;
       class Array Sub Subsub Rep;
       Model l_Activity=Array Sub(Array) Subsub(Sub Array);
       random Array Sub(Array) Subsub(Sub Array)/TEST;
       output out=resids predicted=PRED residual=Z;
run;
```

```
proc print data=resids;
run;
data resids;
       set resids;
       if sub=. then pred=7.66825;  * set predicted values on other points to estimated
average(5.768);
       if subsub=. then pred=7.66825;
run;

*******************************************************************************************
*****;
* calcultate the difference between the predictions in kriging model and those in nested
model;

%diff;
goptions htitle=2 htext=2;

footnote1 ;
axis1 label = ("X");
axis2 label = ("Y");

 legend1 position=(right middle)
        label=(position=top 'log(Act)(Nest-Krig)')
        value=(height=2)
               SHAPE=BAR(6,4)
        across=1;

title1'Contour plot of the differeces between Hierarchical and Kriging model';
title2' for log(Activity)';
proc gcontour data=preddif;
  plot gyc*gxc=nest_krig/pattern
                                        autolabel=(check=none)
                                        haxis=axis1
                      vaxis=axis2
                      legend=legend1;;

  run;
  quit;
```

4 Logistic regressions for missingness on other variables

```
* find the relationship between missingness and other variables;

*imput data;
PROC IMPORT OUT= WORK.alldata
            DATAFILE= "R:\project\all Data.xls"
            DBMS=EXCEL REPLACE;
     SHEET="All Values$";
     GETNAMES=YES;
     MIXED=NO;
     SCANTEXT=YES;
     USEDATE=YES;
     SCANTIME=YES;
RUN;
proc print data=alldata;
run;

* Fit a logistic regression to censoring variable;
proc logistic data=alldata;
 model Censoring=Total_Pb Total_Cr Potential X Y;
 run;

* delete Total_Pb and Total_Cr because they don't have significant effect;
proc logistic data=alldata;
 model Censoring= X Y Potential;
output out=temp pred=phat;
where array le 3;
run;
```

```
* plot the missing data pattern with X and Y coordinates;
proc gcontour data=temp;
       plot Y*X=phat/pattern;
run;
proc freq data=alldata;
       table censoring*array/nopercent nocol norow;
run;

data prob;
 set temp;
 keep sample phat;
 run;

proc print data=prob;
run;
```

5 multiple imputations for missing data using logistic regression methods (overall test and individual tests on lead, chromium and activity)

```
/*code for multiple imputation estimate the parameters for 68 bande binary vaiable with
missing values*/

*imput data;
PROC IMPORT OUT= WORK.alldata
            DATAFILE= "R:\project\all Data.xls"
            DBMS=EXCEL REPLACE;
     SHEET="All Values$";
     GETNAMES=YES;
     MIXED=NO;
     SCANTEXT=YES;
     USEDATE=YES;
     SCANTIME=YES;
RUN;
*******************************************************************************;
* calculate overall model p-value for H0: beta1=beta2=beta3=0 without testing beta0;
*******************************************************************************;
%macro param;
data params;
run;
%do l = 1 %to 68;
     proc mi data=alldata out=outlog&l noprint nimpute=5;
       class x&l;
       var Total_Pb Total_Cr Potential x&l;
       * logistic regression imputation method;
       monotone logistic(x&l=Total_Pb Total_Cr Potential);
     run;

     proc logistic data=outlog&l outest=c&l covout noprint;
       model x&l=Total_Pb Total_Cr Potential;
       by _Imputation_;
     run;

     proc mianalyze data=c&l mult;
        ods output ParameterEstimates=parms&l;
        modeleffects Total_Pb Total_Cr Potential;
     run;
     data params;
       set params parms&l;
     run;
%end;
%mend;
%param;

************************************************;
* use FDR controlling to adjust multiple test P values;
```

```
************************************************;
*imput data;
PROC IMPORT OUT= WORK.overallp
            DATAFILE= "R:\project\overallp.xls"
            DBMS=EXCEL REPLACE;
     SHEET="sheet1$";
     GETNAMES=YES;
     MIXED=NO;
     SCANTEXT=YES;
     USEDATE=YES;
     SCANTIME=YES;
RUN;
proc print data=overallp;
run;
data overallp;
   set overallp;
   raw_p=p;
run;
proc sort data=overallp;
       by raw_p;
run;
* calculate the adjusted p values for controlling FDR;
proc multtest pdata=overallp fdr out=overalltest;
run;
proc print data=overalltest;
run;
data params;
run;

%macro param(l);

     proc mi data=alldata out=outlog&l noprint nimpute=1000;
       class x&l;
       var Total_Pb Total_Cr Potential x&l;
       monotone logistic(x&l=Total_Pb Total_Cr Potential);
     run;

     proc logistic data=outlog&l outest=c&l covout noprint;
       model x&l=Total_Pb Total_Cr Potential;
       by _Imputation_;
     run;

     proc mianalyze data=c&l mult;
         ods output ParameterEstimates=parms&l;
         modeleffects Total_Pb Total_Cr Potential;
     run;
     data params;
               set params parms&l;
     run;
%mend;

* multiple imputation test for each of the significant bands
* which the overall adjusted p values less than 0.05;
* (band #3,21,41,49.52.56.22.12.14)
%param(3);
%param(21);
%param(41);
%param(49);
%param(52);
%param(56);
%param(22);
%param(12);
%param(14);

* add the band number to estimators;
data params;
       set params;
       do i=1 to 9;
           if (_n_-1) gt ((i-1)*3) and (_n_-1)le (i*3) then id=i;
       end;
       if _n_=1 then delete;
```

```
           if (id eq 1) then band=3;
        if (id eq 2) then band=21;
               if (id eq 3) then band=41;
        if (id eq 4) then band=49;
               if (id eq 5) then band=52;
        if (id eq 6) then band=56;
               if (id eq 7) then band=22;
        if (id eq 8) then band=12;
               if (id eq 9) then band=14;
       keep parm probt band;
run;
proc print data=params;
run;


****************************************;
* tests for each of lead, Cr and activity;
****************************************;
* for lead;
%macro estimatelead;
data plead;
run;
%do l = 1 %to 68;
    proc mi data=alldata out=outlog&l noprint nimpute=5;
       class x&l;
       var Total_Pb x&l;
       monotone logistic(x&l=Total_Pb );
    run;

    proc logistic data=outlog&l outest=c&l covout noprint;
       model x&l=Total_Pb ;
       by _Imputation_;
    run;

    proc mianalyze data=c&l mult;
         ods output ParameterEstimates=plead&l;
         modeleffects intercept Total_Pb ;
    run;
    data plead;
               set plead plead&l;
    run;
%end;
%mend;
%estimatelead;

* add the band number to estimators;
data plead;
       set plead;
       do i=1 to 68;
          if (_n_-1) gt ((i-1)*2) and (_n_-1)le (i*2) then band=i;
       end;
       if _n_=1 then delete;
run;

proc print data=plead;
run;

data plead;
   set plead;
   if parm eq 'intercept' then delete;
   keep band probt estimate;
run;
proc print data=plead;
run;
data plead;
   set plead;
   raw_p=probt;
run;
proc sort data=plead;
       by raw_p;
run;
```

```
proc multtest pdata=plead fdr out=lead;
run;
proc print data=lead;
run;

* for Cr;
%macro estimateCr;
data pCr;
run;
%do l = 1 %to 68;
    proc mi data=alldata out=outlog&l noprint nimpute=5;
       class x&l;
       var Total_Cr x&l;
       monotone logistic(x&l=Total_Cr );
    run;

    proc logistic data=outlog&l outest=c&l covout noprint;
       model x&l=Total_Cr ;
       by _Imputation_;
    run;

    proc mianalyze data=c&l mult;
         ods output ParameterEstimates=pCr&l;
         modeleffects intercept Total_Cr ;
    run;
    data pCr;
       set pCr pCr&l;
    run;
%end;
%mend;
%estimateCr;
* add the band number to estimators;
data pCr;
       set pCr;
       do i=1 to 68;
           if (_n_-1) gt ((i-1)*2) and (_n_-1)le (i*2) then band=i;
       end;
       if _n_=1 then delete;
run;

proc print data=pCr;
run;

data pCr;
   set pCr;
   if parm eq 'intercept' then delete;
   keep band probt estimate;
run;
proc print data=pCr;
run;

data pCr;
   set pCr;
   raw_p=probt;
run;

proc sort data=pCr;
       by raw_p;
run;

proc multtest pdata=pCr fdr out=Cr;
run;
proc print data=Cr;
run;

* for Activity;
%macro estimateact;
data pact;
run;
%do l = 1 %to 68;
```

```
    proc mi data=alldata out=outlog&l noprint nimpute=5;
       class x&l;
       var Net_Activity x&l;
       monotone logistic(x&l = Net_Activity );
    run;

    proc logistic data=outlog&l outest=c&l covout noprint;
       model x&l=Net_Activity ;
       by _Imputation_;
    run;

    proc mianalyze data=c&l mult;
         ods output ParameterEstimates=pact&l;
         modeleffects intercept Net_Activity ;
     run;
     data pact;
       set pact pact&l;
     run;
%end;
%mend;
%estimateact;
* add the band number to estimators;
data pact;
       set pact;
       do i=1 to 68;
           if (_n_-1) gt ((i-1)*2) and (_n_-1)le (i*2) then band=i;
       end;
       if _n_=1 then delete;
run;

proc print data=pact;
run;

data pact;
   set pact;
   if parm eq 'intercept' then delete;
   keep band probt estimate;
run;
proc print data=pact;
run;

data pact;
   set pact;
   raw_p=probt;
run;

proc sort data=pact;
       by raw_p;
run;

proc multtest pdata=pact fdr out=activity;
run;

Proc print data=activity;
run;
```