United States
Environmental Protection
Agency

Office of Research and
Development
Washington DC 20460

EPA/600/R-00/034
August 2000

Statistical Estimation and Visualization of Ground-Water Contamination Data

EPA/600/R-00/034
August 2000

Statistical Estimation and Visualization of Ground-Water Contamination Data

Rachel K. Boeckenhauer
Dennis D. Cox
Katherine B. Ensor
Philip B. Bedient
Anthony W. Holder
Rice University
Houston, Texas 77005-1892

Cooperative Agreement No. CR-821906

Project Officer
Mary E. Gonsoulin
Subsurface Protection and Remediation Division
National Risk Management Research Laboratory
Ada, Oklahoma 74820

National Risk Management Research Laboratory
Office of Research and Development
U.S. Environmental Protection Agency
Cincinnati, OH 45268

Notice

The U.S. Environmental Protection Agency through its Office of Research and Development partially funded and collaborated in the research described here under Cooperative Agreement No. CR-821906 to Rice University. It has been subjected to the Agency's peer and administrative review and has been approved for publication as an EPA document. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

All research projects making conclusions or recommendations based on environmentally related measurements and funded by the Environmental Protection Agency are required to participate in the Agency Quality Assurance Program. This project was conducted under an approved Quality Assurance Project Plan. The procedures specified in this plan were used without exception. Information on the plan and documentation of the quality assurance activities and results are available from the Principal Investigator.

Foreword

The U.S. Environmental Protection Agency is charged by Congress with protecting the Nation’s land, air, and water resources. Under a mandate of national environmental laws, the Agency strives to formulate and implement actions leading to a compatible balance between human activities and the ability of natural systems to support and nurture life. To meet these mandates, EPA’s research program is providing data and technical support for solving environmental problems today and building a science knowledge base necessary to manage our ecological resources wisely, understand how pollutants affect our health, and prevent or reduce environmental risks in the future.

The National Risk Management Research Laboratory is the Agency’s center for investigation of technological and management approaches for reducing risks from threats to human health and the environment. The focus of the Laboratory’s research program is on methods for the prevention and control of pollution to air, land, water, and subsurface resources; protection of water quality in public water systems; remediation of contaminated sites and ground water; and prevention and control of indoor air pollution. The goal of this research effort is to catalyze the development and implementation of innovative, cost-effective environmental technologies; develop scientific and engineering information needed by EPA to support regulatory and policy decisions; and provide technical support and information transfer to ensure effective implementation of environmental regulations and strategies.

This work presents methods of visualizing and animating statistical estimates of ground water and/or soil contamination over a region from observations of the contaminant for that region. The primary statistical methods used to produce the regional estimates are nonparametric regression and geostatistical modeling (kriging). Nonparametric regression can be used as a more "rough and ready" method to produce surface estimates with little outside intervention, whereas geostatistical modeling produces prediction errors. Finally, a method is proposed for estimating the total amount of contaminant present in a region. This report is published and made available by EPA's Office of Research and Development to assist the user community.

Clinton W. Hall, Director
Subsurface Protection and Remediation Division
National Risk Management Research Laboratory

Abstract

This work presents methods of visualizing and animating statistical estimates of ground water and/or soil contamination over a region from observations of the contaminant for that region. The primary statistical methods used to produce the regional estimates are nonparametric regression and geostatistical modeling (kriging). Nonparametric regression can be used as a more "rough and ready" method to produce surface estimates with little outside intervention, whereas geostatistical modeling produces prediction errors.

Animation of changes in the estimated level of contaminant or chemical as observations are removed illustrate the effect of each individual measurement on the overall estimate and the error or variance of this estimate. Such methods are applied to the Eglin Air Force Base (AFB) Florida site. The benefit of animating surface estimates in data which is taken over time is clearly seen by an example from a site near Phoenix, AZ, where aberrations in the data for one or several years were readily apparent by viewing a smoothed animation.

Finally, a method is proposed for estimating the total amount of contaminant present in a region. The proposed method models the data as a realization of a lognormal stochastic process and then capitalizes on conditional simulation to generate realizations of the modeled process from which the distribution of the total contaminant (or integral of the process) is estimated.

This report was submitted in fulfillment of cooperative agreement No. CR-821906 by Rice University under the sponsorship of the United States Environmental Protection Agency. This report covers a period from 10/01/93 to 03/31/97, and work was completed as of 11/03/98.

Contents

List of Figures

Tables

Acknowledgments

This project was initiated by Dr. M. E. Gonsoulin and Dr. C. Enfield, U. S. Environmental Protection Agency. We thank Dr. S. Schmelling of the U.S. Environmental Protection Agency, Dr. R. Michler of the University of North Texas, and Dr. D. McLaughlin of the Department of Civil and Environmental Engineering Massachusetts Institute of Technology, for their reviews and valuable comments. Also, we thank Ms. J. Elliott of the U. S. Environmental Protection Agency and Ms. M. Williams of Orbiting Astronomical Observatory Corp. for providing graphical support and developing the final format of the report.

Chapter 1
Introduction

Understanding the plume of any contaminant is a multi-faceted problem. Often data is limited and modeling exercises are tedious. To aid the environmental researcher, we present several techniques for capitalizing on the measurements of the level of contaminant in soil samples from a region. Our techniques rely on visual displays of statistical estimates of the contaminant plume. We explore two- and three-dimensional estimation and visualization techniques and ways to examine related contaminants. Furthermore, we propose a method for quantification of the total amount of contaminant within a region. Both of these methods are investigated in the context of a site specific example, but the tools generalize to other similar problems.

As a simple method for displaying surface estimates from field data, exploratory visualization of the Eglin AFB in Section 4.1 was performed using nonparametric regression to produce the surface estimates. This methodology requires estimation of only one parameter, the bandwidth, as opposed to the several parameters required by the more complicated art of variogram fitting used in geostatistical modeling. Surface estimates were obtained for both two- and three-dimensional data for the Eglin AFB site, using the program *Geomview* on a Silicon Graphics machine to facilitate display and animation. *Geomview* allows the viewer to rotate images in real time, which aids greatly in examining the surface (i.e., looking for peaks and valleys, etc.). For data in two dimensions, the third dimension (i.e., the *z* direction) may be used to plot the surface estimate as a perspective plot. Color contours on the surface can be set to levels of interest to highlight areas where the contamination is above a fixed level. This tool can be useful in cases where environmental regulations require contamination to be below some specific level, to identify regions of high contaminant concentration, and to follow the movement of a contaminant over time. In an analogous fashion, for three-dimensional visualization, shell contours are plotted at certain levels to illustrate regions of higher concentrations.

Estimates of prediction errors in both the two-dimensional and three-dimensional setting provide an understanding of the differing levels of uncertainty of the estimate of the level of the contaminant over the region. In the case of Eglin AFB ground water, we also use visual tools in conjunction with cross-validation to ascertain the effect each of the data values has on the estimate of the level of contaminant for the region. An animation of estimates produced excluding individual data points, alternated with the overall surface estimate, lends insight to the question of where to obtain new samples. This sort of display also helps us to determine the level of error in our estimates of the contaminant plume.

A second site providing a different type of complexity was examined from a statistical estimation and visualization perspective. This second site, in Arizona, yields observations of several contaminants collected over a period of several years. Animations of estimates capitalizing on the temporal component clearly illustrate major trends and aberrations in the data, which can then be investigated more closely. Also, there were several different contaminant substances measured at the Arizona site, such as TCE, DCE, and SO_4^{2-} . The behavior of these contaminants is expected to be interrelated. We present suggestions on how to best visualize simultaneously two or more related substances in order to highlight possible relationships among the series. The Arizona site also includes a common problem in that the region where measurements were taken increases over time. This problem is addressed here, along with some possible solutions.

Another primary focus of this research is to answer the question of how much total contamination is present at a site and how to best estimate this quantity given soil core samples from the site. It is important when producing such estimates to also understand the level of uncertainty in the estimate, in other words to obtain the standard error of the estimate. A byproduct of carefully implemented geostatistical methods such as *kriging* is standard errors for the estimated mean level over the region. Estimation of total contaminant involves estimation of the integral of the modeled process over a region. We pursue estimation of the total contaminant for Eglin AFB. The level of ground water BTEX was modeled as a realization of a lognormal stochastic process, and estimates of the distribution of the integral were produced by Monte-Carlo simulation of the process conditional on the observed data.

The data used in this research, from Eglin AFB and a site near Phoenix, AZ, are presented in Chapter 2. Chapter 3 contains discussion of the statistical methods used, including estimation by nonparametric regression and kriging. Chapter 4 contains results for Eglin AFB, including both exploratory data analysis and estimation of total contaminant. Chapter 6 contains discussion of the visualization and animation for the Arizona site, including discussions of temporal data and visualization of two related substances. Finally, Chapter 7 contains conclusions and suggestions for future research.

Chapter 2
Sites and Data

2.1 Eglin AFB

In Chapters 4 and 5, we consider data from Eglin AFB, an example of a shallow aquifer in sandy soil. A leak of 30,000 - 40,000 gallons of JP-4 jet fuel was detected at Eglin AFB in Florida by Air Force personnel in 1984 (Boeckenhauer, *et.al.*,1995). The contamination measured here is from BTEX, including benzene, toluene, ethylbenzene, and *m*-, *o*-, and *p*-xylene, which are typically contained in petroleum fuels and are hazardous substances regulated by the U.S. Environmental Protection Agency (Sweed, et al.,1996).

Two data sets are available for this site, namely: (1) Ground-water BTEX concentrations in two dimensions, measured in *μg/L* and (2) Soil BTEX concentrations in three dimensions, measured in *mg/kg*. Also, for the exploratory analysis, we used the soil data which are approximately 7.0 and 7.6 *ft* below the water table as two different two-dimensional data sets. The 22 ground-water data points were collected by researchers from Rice University in March 1993 using a cone penetrometer. These data range from 0.001 to over 9*mg/L*. A plot of the ground-water data is shown in Figure 2.1.

Anaerobic soil cores were collected in March and July, 1993, and March, 1994. The soil data set contains 336 points at 20 different locations, with values ranging from 0 to approximately 750*mg/kg*. A plot of the three-dimensional soil data points is shown in Figure 2.2. The actual vertical range of the region is 21.6 *ft*, whereas the longitude encompasses 230.3 *ft* and latitude 286.4 *ft* The *X*s connected by the dotted line indicate the location of the source of contaminant. The larger blocks denote observations of measured concentration exceeding 25*mg/kg*, whereas the smaller blocks depict observations with measurements between 0 and 25*mg/kg*. Also, the location of the water table is shown with stripes. It is observed from this figure that very few of the data points actually have values greater than 25*mg/kg* (only 16 of the 336), and *all* of these lie below the water table.

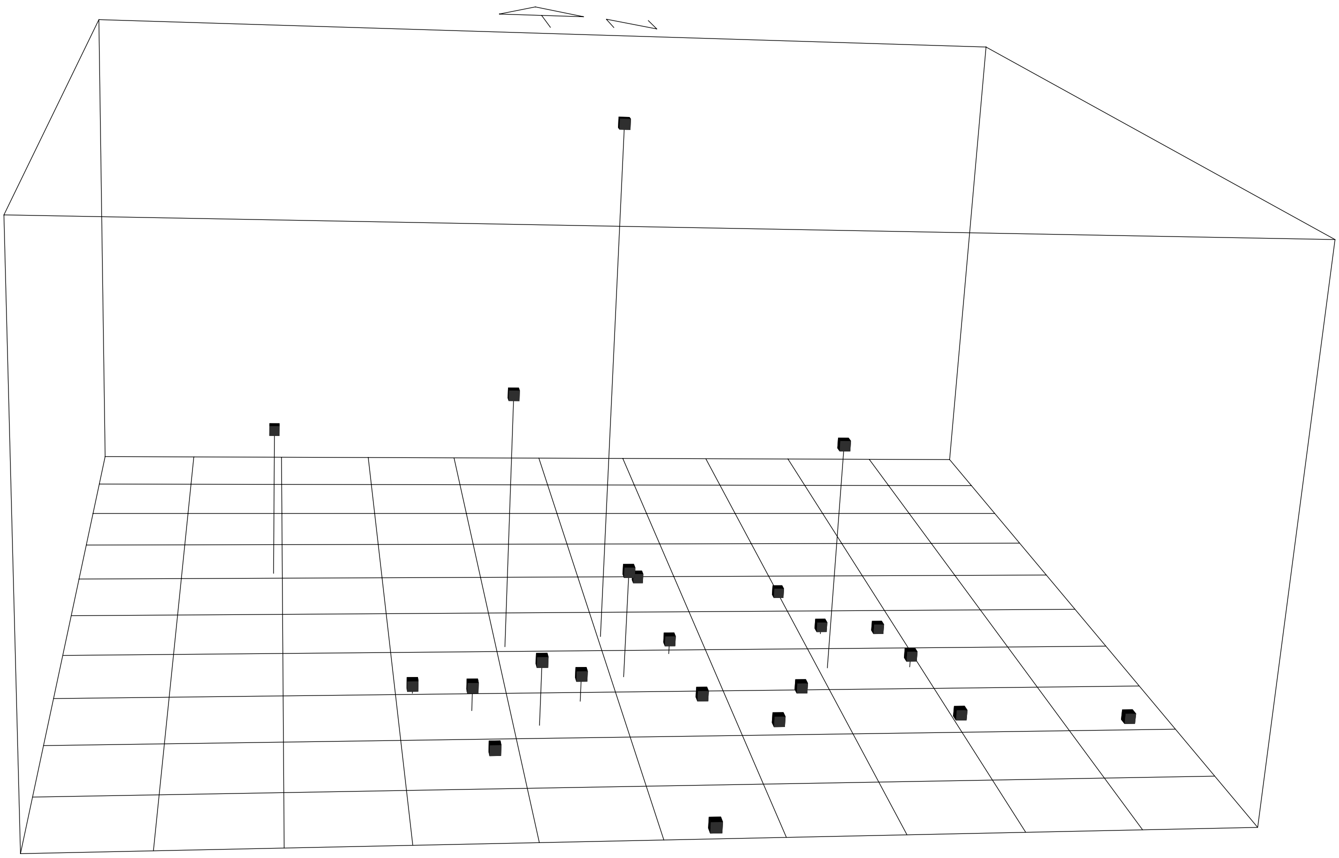

Figure 2.1 Eglin AFB ground-water data points. Coordinates: depth-axis: 4400 to 5100; width-axis: 4900 to 5600 and vertical-axis: 0 mg/kg to 9.1 mg/kg.

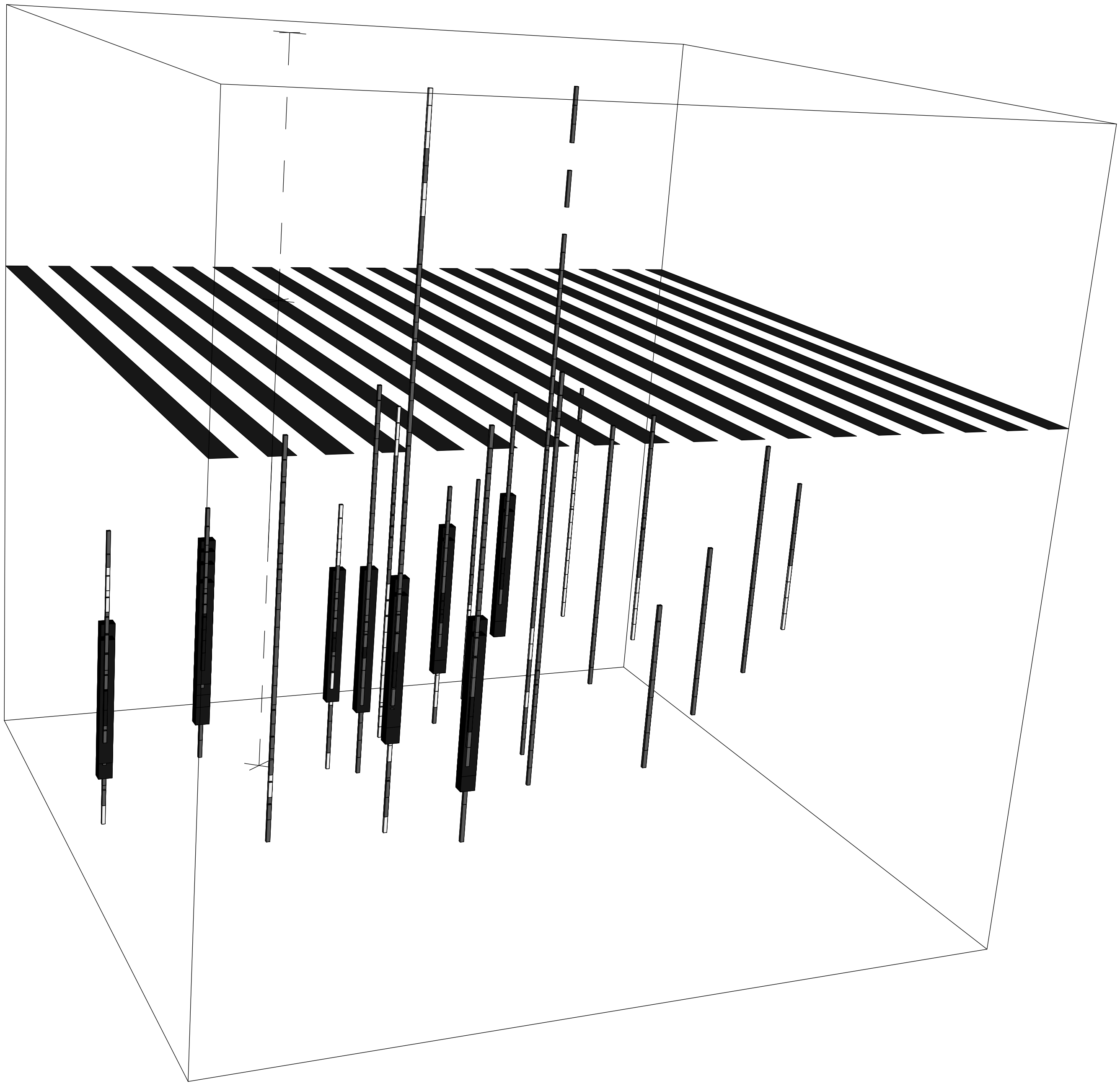

Figure 2.2 Eglin AFB soil data points.

2.2 Arizona

In Chapter 6, we consider data from a contaminated site near Phoenix, AZ. Contaminants measured include trichloroethylene (TCE), dichloroethylene (DCE), and sulfate (SO_4^{2-}), all measured in *μg/L*. Other measured contaminants contained somewhat sparse data and were not used at this time. The DCE "measurements" are actually sums of measured values of 1,1-DCE and 1,2-DCE, so the measurement locations here are only used if measurements of *both* of these are available. These data were gathered from 1985 to 1993.

Chapter 3
Statistical Methods

3.1 Nonparametric Regression

For the exploratory analysis of Eglin AFB in Section 4.1, we used *nonparametric regression* to produce a surface estimate of the plume. The model used for the contaminant plume is $u(x) = f(x) + \varepsilon(x)$ where:

$\mathbf{x}$ = a point in the region of interest,

$u(\mathbf{x})$ = the observed level of contaminant at x,

$f(\mathbf{x})$ = the true level of contaminant at x, and

$\varepsilon(\mathbf{x})$ = random noise in the measurement process.

The model assumes:

- The observation locations $\mathbf{x}_1,\ldots,\mathbf{x}_n$ are randomly chosen.
- The unknown function f(•) is twice continuously differentiable. Note that f(•) is not a random process.
- The random noise $\varepsilon(\mathbf{x}_1),\ldots,\varepsilon(\mathbf{x}_n)$ is independent, but not necessarily identically distributed.

An estimate of f($\mathbf{x}$) over the region of interest can be obtained via nonparametric regression methods (Scott, 1992) and is given by:

$$\hat{f}(x) = \sum_{i=1}^{n} u_i w_h(\boldsymbol{x},\boldsymbol{x}_i) \tag{3.1}$$

where the weights are defined as:

$$w_h(\boldsymbol{x},\boldsymbol{x}_i) = \frac{K_h(x - x_i)}{\sum_{j=1}^{n} K_h(x - x_j)} \tag{3.2}$$

and $u_i \equiv u(x_i)$. The function $K_h(\bullet)$ is referred to as the scaled *kernel* function with *bandwidth*, or smoothing parameter, *h*. Note that the bandwidth determines the smoothness of the surface estimate; a larger bandwidth yields a smoother estimate. For some standardized (or unscaled) kernel function $K(\bullet)$, we define $K_h(\bullet)$ as $K_h(t) = K(t/h)/h$, so as *h increases*, the value of $t/h = (x - x_i)/h$ *decreases*.

3.2 Kriging and Variograms

Another method which we will use to estimate contaminant levels over the region is the geostatistical spatial prediction method known as *kriging*. A complete and thorough exposition of geostatistical methods is given in (Cressie, 1993). A very brief overview is provided here for purposes of introduction and definition of notation. The general idea of kriging is to first use the observed levels of a contaminant to produce a model of the spatial covariance structure of the process. This spatial covariance model is then used to obtain the "optimal" predictor $p(\boldsymbol{Z}; s_0)$ of $Z(s_0)$, the value of the random process at s_0. This predictor is $E[Z(s_0)|\mathbf{Z}]$, which is precisely the same as $p^O(\boldsymbol{Z}; s_0)$ in (5.2) in the case where $g(\bullet)$ is simply $g(Z) = Z$.

Assume that the data $\mathbf{z} = (z(s_1), \ldots, z(s_n))$ are a sample from a realization of the stochastic process $\{Z(s) : s \in A\}$. In order to do inference from the data, we need to make some assumptions. A common practice is to assume *second-order stationarity*. That is, assume that

$$E[Z(s)] = \mu \qquad \forall s \in A \tag{3.3}$$

or that $F_s(z) \equiv Pr(Z(s) \leq z)$ does not depend on $\mathbf{s}$ and

$$\boldsymbol{cov}\left(Z(s_i), Z(s_j)\right) = C(s_i - s_j) \qquad \forall s_i, s_j \in A, \tag{3.4}$$

i.e., $cov(Z(s_i), Z(s_j))$ depends only on the vector $s_i - s_j$. Furthermore, if

$$\boldsymbol{cov}\left(Z(s_i), Z(s_j)\right) = C(|s_i - s_j|) \qquad \forall s_i, s_j \in A, \tag{3.5}$$

i.e., $cov(Z(s_i),Z(s_j))$ depends only on the *distance* $|s_i - s_j|$, then the process is said to be *isotropic*.

If second-order stationarity holds, a convenient way to model the covariance structure of the process is through use of a *variogram* function

$$var\left(Z(s_i)-Z(s_j)\right)=2\gamma\left(s_i-s_j\right) \qquad\qquad \forall s_i,s_j \in A, \tag{3.6}$$

The function $\gamma(s_i - s_j)$ is referred to as the *semivariogram.* Note that it is easy to show that $\gamma(h) = \gamma(-h)$ and $\gamma(\mathbf{0}) = 0$. If we have:

$$\lim_{h\to 0} \gamma(h)=c_0>0 \tag{3.7}$$

then c_0 is what is known as a *nugget effect*. This may be due either to some microscale variation or to measurement error. The term nugget effect comes from spatial prediction's origins in mining, and refers to a variation caused by small *nuggets* of ore. Regardless, in all real data, there is some measurement error and so we would be remiss to model our process without a nugget effect. On the other end, as $|\mathbf{h}|\rightarrow\infty$, the semivariogram converges to the process variance. This follows easily by noting that we assume that the covariance between two values of the process diminishes to zero as the distance between them increases.

In the case that the process is isotropic; i.e., the spatial covariance between values of the process depends *only* on the distance between the observations, there are a number of standard variogram models available (Cressie, 1993). However, if the process is *anisotropic,* it is sometimes possible to transform the locations so that an isotropic variogram model remains appropriate. Specifically, such a transformation is possible in cases of *geometric anisotropy;* i.e., where rotating and scaling the locations produces an isotropic process. For example, it is typically the case with ground-water data that the correlation is higher for points a distance ***h*** apart if they lie in the direction of ground-water flow rather than perpendicular to it. In this case, the variogram is of the form:

$$2\gamma(\mathbf{h}) = 2\gamma_0\left(|\mathbf{B}\,\mathbf{h}|\right) \qquad\qquad \mathbf{h} \in A \subset \Re^d \tag{3.8}$$

where ***B*** is a d xd matrix and γ_0 is an isotropic variogram. We will be using this type of transformation on the Eglin AFB ground-water contamination data where $d = 2$. In the case $d = 2$, the matrix B is given by:

$$\mathbf{B}=\begin{pmatrix} \cos^2(\theta)+r*\sin^2(\theta) & (1-r)*\sin(\theta)*\cos(\theta) \\ (1-r)*\sin(\theta)*\cos(\theta) & \sin^2(\theta)+r*\cos^2(\theta) \end{pmatrix} \tag{3.9}$$

meaning that θ is the angle clockwise from North at which the scale is multiplied by r. The other axis is then the one which is perpendicular to this, and the scale in this direction is not altered. For example, then, if we were working with ground-water data where the flow was along the northwest-southeast direction, we might use $\theta\cong 135$ deg and some $r > 1$.

Assuming the modeled spatial covariance structure of the random process, we can now obtain optimal predictions. If our optimization criteria is minimization of the squared-error loss then the optimal predictor of the random process at any point $\mathbf{s}_0$ is given by the expectation of the random process conditional on the observed values of the process; in other words, the best predictor of $Z(s_0)$ is $p_0(Z,s_0) = E(Z(s_0)|Z)$, where Z denotes the vector of data as in (Boeckenhauer, 1996). In the case that $Z(\bullet)$ is a Gaussian process, this predictor is linear. Here we will be using the form of spatial prediction known as *ordinary kriging* (Cressie, 1993) which requires the two assumptions:

1. There is a constant mean, i.e.

$$Z(\mathbf{s})=\mu+\delta(\mathbf{s}) \tag{3.10}$$

for $\mathbf{s}\in A$ and $\mu\in\Re$ unknown.

2. The predictor is linear in the observations, i.e.

$$p(Z;s_0)=\sum_{i=1}^{n}\lambda_i Z(\mathbf{t}_i) \tag{3.11}$$

where the observations are at locations $\mathbf{t_1}, ..., \mathbf{t_n}$ and $\sum_{i=1}^{n}\lambda_i=1$. (Recall that this predictor is an estimate of $E[Z(s_0)|Z(t_1),...,Z(t_n)]$.)

Requiring $\sum_{i=1}^{n}\lambda_i=1$ guarantees uniform unbiasedness, i.e. $E(p(Z;s_0)) = \mu = E(Z(s_0))$. For further information on optimal prediction using the *kriging equations*, see (Cressie, 1993).

3.3 Estimating an Integral via Sample-Mean Monte Carlo

The question of the total amount of contaminant within a given region is equivalent to estimating the integral over the region of the estimated spatial process for this contaminant. Monte Carlo methods provide an estimate of the integral (see also (Rubinstein,1981) and (Hammersley and Handscomb, 1964)) by viewing the integral as an expectation and simulating the sample mean as an estimate of this expectation. For example, suppose we wish to estimate the integral over a region ***A*** of some function $q(s)$:

$$\Phi = \int_A q(\mathbf{s})\, ds \tag{3.12}$$

The basis of this method is to represent the integral Φ as the expected value of a random variable. For example, suppose that **S** is a random variable which has density $f_S(s)$ on ***A***. We may then rewrite the integral Φ in (3.12) as:

$$\Phi = \int_A q(\mathbf{s})\, ds = \int_A \frac{q(\mathbf{s})}{f_S(\mathbf{s})} f_S(s) ds = E\left[\frac{q(\mathbf{S})}{f_S(\mathbf{S})}\right] \tag{3.13}$$

provided that $f_S(\mathbf{s}) > 0$ when $q(s) \neq 0$.

In particular, suppose **S** is uniform on ***A***. That is, **S** has density

$$f_S(\mathbf{s}) = \frac{1}{|\boldsymbol{A}|} I_A(\mathbf{s}) \tag{3.14}$$

where

$$\mathrm{I}_A(\mathbf{s}) = \begin{cases} 1 & \text{if } \mathbf{s}\ \varepsilon\, \mathrm{A} \\ 0 & \text{otherwise} \end{cases} \tag{3.15}$$

is the indicator function on ***A*** and |***A***| is the norm of ***A*** (e.g., the area or volume of ***A*** if ***A*** is two- or three-dimensional, respectively). We may then simplify the integral in (3.13)

$$\Phi = \int_A \mathrm{q}(\mathrm{s})\, \mathrm{ds} = \mathrm{E}\left[\frac{\mathrm{q}(\mathbf{s})}{\mathrm{f}_S(\mathbf{s})}\right] = \mathrm{E}\left[\frac{\mathrm{q}(\mathbf{s})}{1/|\mathrm{A}|}\right] = |\mathrm{A}|\mathrm{E}\left[\mathrm{q}(\mathbf{S})\right] \tag{3.16}$$

To use this method to estimate Φ, then, we will

1. Generate a "large" (say *p*) number of locations, $s_1,...,s_p$ uniformly over the region ***A***.

2. Evaluate q(s) at each of these locations, yielding $q(s_1),...,q(s_p)$.

3. Compute the sample mean of these evaluations to yield an estimate of $E[q(\mathbf{S})]$, i.e. $\frac{1}{p}\sum_{j=1}^{p} q(s_j)$.

4. Use the sample mean in (3.) to estimate the integral Φ, or

$$\hat{\Phi} = \frac{|A|}{p}\sum_{j=1}^{p} q(\mathbf{s_j}) \tag{3.17}$$

The error inherent in this method relates to the randomness of the sampled sites and the number of sites which we sample. Specifically, the variance of the integral estimate for the function *q* is given by:

$$\begin{aligned} Var(\Phi) &= Var\left(\frac{|A|}{p}\sum_{j=1}^{p} q(\mathbf{s}_j)\right) \\ &= \frac{|A|^2}{p^2} Var\left(\sum_{j=1}^{p} q(\mathbf{s}_j)\right) \end{aligned} \tag{3.18}$$

This variance may be estimated by:

$$\begin{aligned} \hat{V}ar(\hat{\Phi}) &= \frac{|A|^2}{p^2} ps^2 \\ &= \frac{|A|^2 s^2}{p} \end{aligned} \tag{3.19}$$

where s^2 is the sample variance of $q(s_1),...,q(s_p)$.

In order to estimate the total level of contaminant say for the Eglin AFB, we model the spatial process of interest. Using our model as the truth, we can generate a *p*-dimensional Multivariate Normal random vector with the appropriate mean structure and spatial covariance structure as given by our estimated model (see, for example, Johnson, 1987; Stewart, 1973). This simulated random vector is then used in the above algorithm to ascertain the total amount of contaminant present, and the standard error of this estimate.

Chapter 4
Eglin AFB: Visualization and Exploratory Analysis

Our exploratory analysis of the observations of BTEX from Eglin AFB provides an understanding of the location and shape of the contaminant plume. For the two-dimensional data, perspective plots of the surface estimate with color contours visually display both the level of contamination and the rate of change over the region. The color contours can be set to specific concentrations of interest, such as regulatory levels. In the three-dimensional case, nested contour visualization is used to provide immediate characterization of the plume. Again, the contours could be keyed to concentrations of interest. Such exploratory analyses are greatly enhanced by on-line manipulations of the visual tools provided. For example, it is possible to rotate the surfaces to search for high levels of contamination which may be visible only from certain vantage points. Furthermore, color facilitates identification of trouble spots. However, even the gray-scale static versions of the plots presented here are useful for providing visual understanding of the plume.

The surface estimates for the exploratory visualization were produced using *nonparametric regression*, as discussed in Section 3.1. In order to obtain an accurate surface estimate of the plume, it is necessary to choose appropriate bandwidths to produce the nonparametric regression estimate. (Recall that the larger the bandwidth, the smoother the final surface estimate will be.) Appendix A contains the details of bandwidth selection, via cross-validation, for both the two-dimensional and three-dimensional Eglin AFB data.

Properties of spatial estimates and/or nonparametric regression estimates rely on asymptotic theory which, due to the small number of observations available, is certainly not the situation with the data at hand. Therefore, we explore the robustness of our estimate by examining the change in the estimate as sample points are removed from the estimation process. The uses of this examination are two-fold: (1) a better understanding of the the magnitude of the error of our estimates is obtained, and (2) areas where additional observations are needed are highlighted. In other words, intuitively, we would take additional observations in the region where the estimated level of contamination changed the most as data was removed. If removing an observation has little effect on the estimate resulting in a small error, additional data would not be needed in that region. This result was visualized in two ways:

1. by viewing an animation of the estimated surface alternated with surfaces estimated by the removal of one of the sample points, and
2. by visualizations of the absolute differences (errors) between the surface estimated without point *i* and the measured value at point *i*.

The animation in case (1) is very useful when viewed on the SGI, but does not appear here. Case (2) appears in Section 4.2.

By viewing a smoothed version of the absolute error of our plume estimate, both in the two-dimensional and three-dimensional cases, it is clear where additional observations are needed. One suggestion for future sampling sites would be to sample in the region where both the estimated level of contamination and the error associated with the estimate are high. Of course, any measure of *error* or the amount of information contained in the data could be displayed in a similar fashion.

4.1 Visualization of Estimated Plumes

The program *Geomview* for an SGI was used to display plume estimates of the BTEX concentration. Geomview was written at the NSF Geometry Center, University of Minnesota, and is available through anonymous *ftp* from *ftp.geom.umn.edu.* The program *ashreg*, a modification of *ashn* (Scott, 1992), was used to produce plume estimates from the three-dimensional data. A biweight kernel was used in all cases.

4.1.1 Two-Dimensional Data

Figure 4.2 contains a plot of the estimated plume for the Eglin AFB ground-water data. Figure 4.1 shows the legends for these contour levels and those for the later figures, signifying estimated concentrations of:

$$\begin{array}{lrclr}
(a) & 5000 & \leq & \hat{u} & < & 10000 \\
(b) & 2000 & \leq & \hat{u} & < & 5000 \\
(c) & 1000 & \leq & \hat{u} & < & 2000 \\
(d) & 500 & \leq & \hat{u} & < & 1000 \\
(e) & 100 & \leq & \hat{u} & < & 500 \\
(f) & 10 & \leq & \hat{u} & < & 100 \\
(g) & 0 & < & \hat{u} & < & 10 \\
(h) & & & \hat{u} & = & 0 \\
(i) & & & \multicolumn{3}{l}{No\ data}
\end{array} \tag{4.1}$$

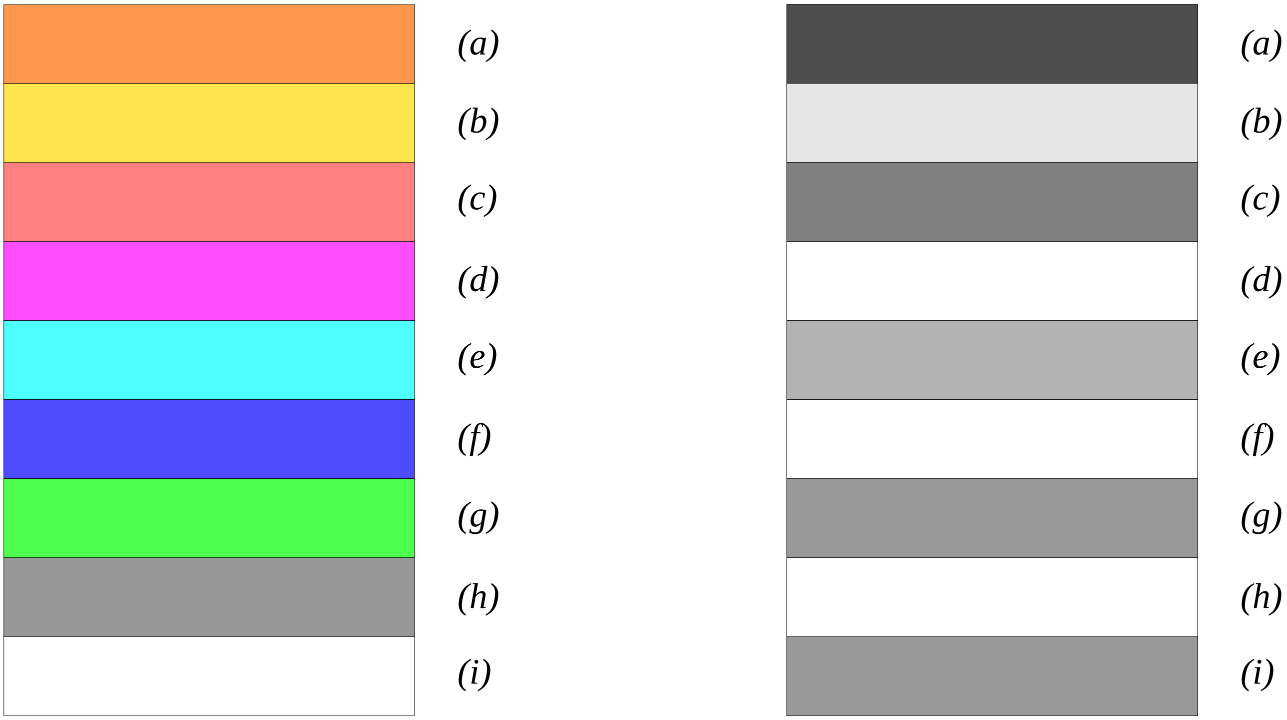

Figure 4.1 Legends for two-dimensional perspective plot contours.

where all of the concentrations are in *μg/L*. The level labeled *No data* includes areas where there is no estimate as we are too far from any of the measured points for the given bandwidth. (Note that, for the black and white figures, some of the lower levels are shown in the same shade, but we are primarily interested in areas where there are *high* concentrations, and these are distinctly different.) The same legend applies to the plot of the estimated plume for the Eglin AFB soil data at a depth of approximately 7.0 *ft* given in Figure 4.3, except that here the concentrations are:

$$\begin{array}{lrcccr}
(c) & 50 & \leq & \hat{u} & < & 100 \\
(d) & 10 & \leq & \hat{u} & < & 50 \\
(e) & 5 & \leq & \hat{u} & < & 10 \\
(f) & 1 & \leq & \hat{u} & < & 5 \\
(g) & 0 & < & \hat{u} & < & 1 \\
(h) & & & \hat{u} & = & 0 \\
(i) & & & \multicolumn{3}{l}{\textit{No data}}
\end{array}$$

with all measurements in *mg/kg*. (Note that at this particular depth, the estimated concentration does not exceed 50*mg/kg*.)

For both of the two-dimensional plots, the actual data locations are marked on the grid at the top of the plot. The large point which is connected to the perspective plot by a vertical line is the mode of the estimate. An arrow pointing north indicates the orientation of the plot.

4.1.2 Three-Dimensional Data

Figure 4.4 contains a plot of the estimated plume for the Eglin AFB soil data. Two different levels of contamination are represented, with the lower (outer) shell being sliced so that we can see the higher (inner) one. The outer and inner shells represent concentration levels of approximately 0.73 and 7.33*mg/kg*, respectively. Also, the location of the water table and the source are clearly marked.

In an on-line version of the 3-D visualization of the plume, nested contours can be displayed using the transparency feature of the SGI and Geomview. This allows the use of solid color nested contours, where the inner shells can be seen through the outer ones. The use of transparency and the ability to rotate the graph greatly enhance the informative 7.0 *ft* below the Water Table value of the plot to the user.

4.2 Error Visualization

We simultaneously display a perspective plot of the two-dimensional ground-water BTEX plume estimate and the absolute errors of this estimate in Figure 4.5. Note that the heights of these two plots are on the same scale. The contours for the errors, using the legend in Figure 4.1, are the same as given in 4.1 except for levels (a) and (b):

$$(a) \quad 3000 \quad \leq \quad \hat{u} \quad < \quad 5000$$

$$(b) \quad 2000 \quad \leq \quad \hat{u} \quad < \quad 3000$$

with errors in absolute *µg/L*. In Figure 4.6, we show a perspective plot of the ground-water data where the contours are determined by the smoothed absolute errors. By combining the plume estimate and the absolute errors into a single plot, we more readily identify regions of high concentration and regions with a large amount of uncertainty in the estimated value.

To visualize the errors for the three-dimensional soil data, we simply took the absolute errors as calculated for cross-validation in A.2 (i.e. $|u_1— \hat{u}|$, . . . , $|u_n —\hat{u}_n||$) and plotted a smoothed contour shells (see Figure 4.7). Here the outer and inner shells represent contaminant concentrations of approximately 0.48 and 4.76*mg/kg*, respectively.

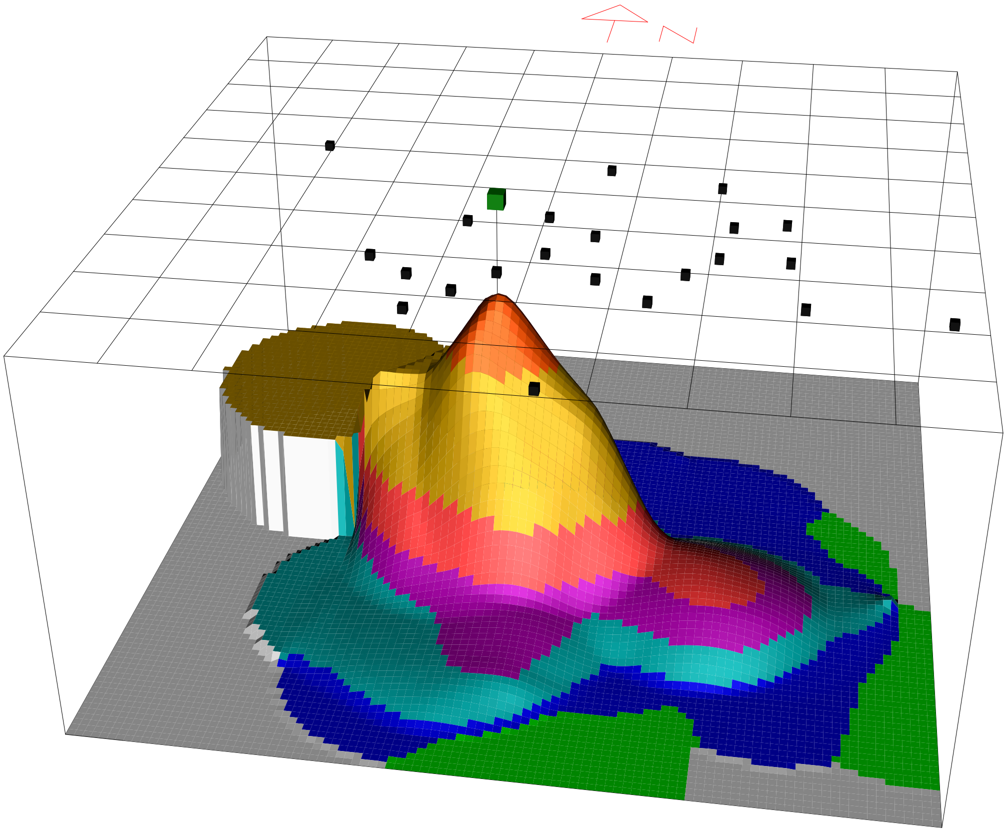

Figure 4.2 Estimate of contaminant plume for Eglin AFB ground-water data.

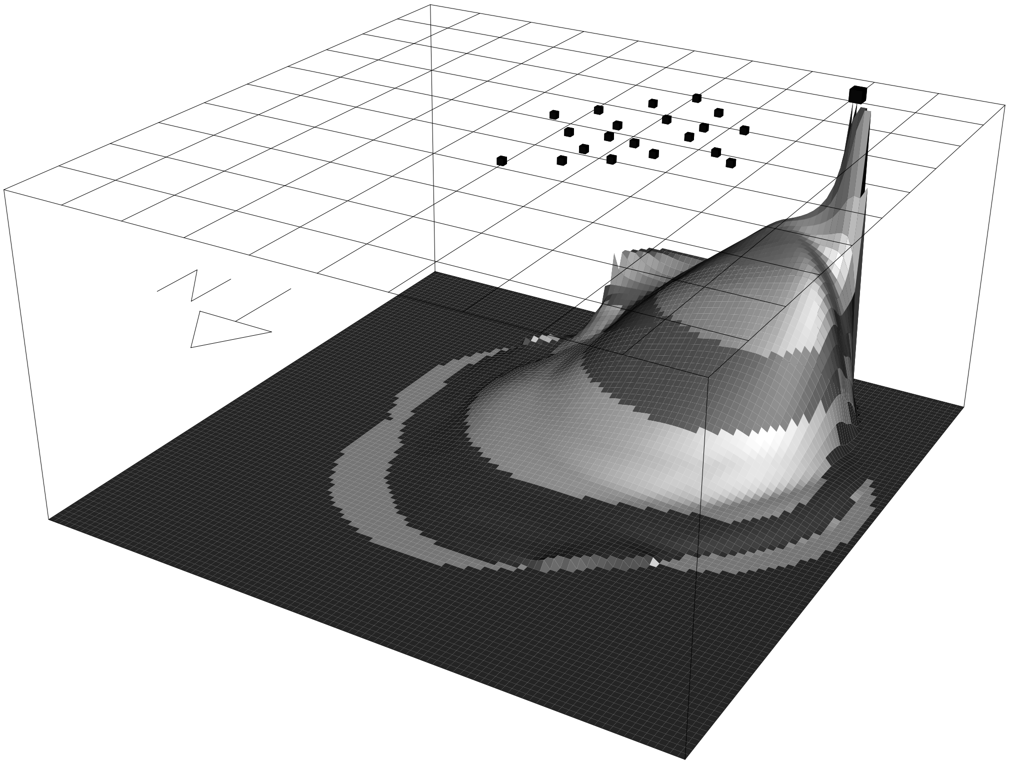

Figure 4.3 Estimate of soil contamination 7.0 *ft* below the water table.

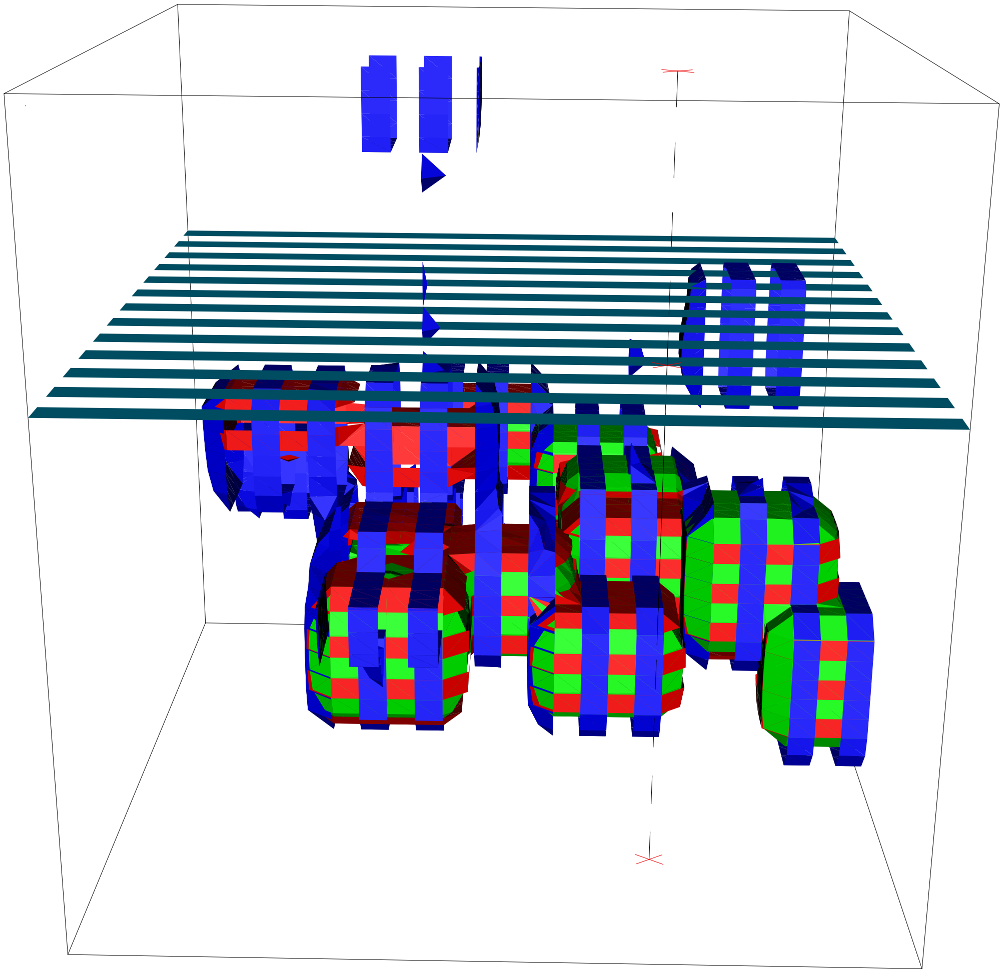

Figure 4.4 Estimate of soil contamination (BTEX) at Eglin AFB.

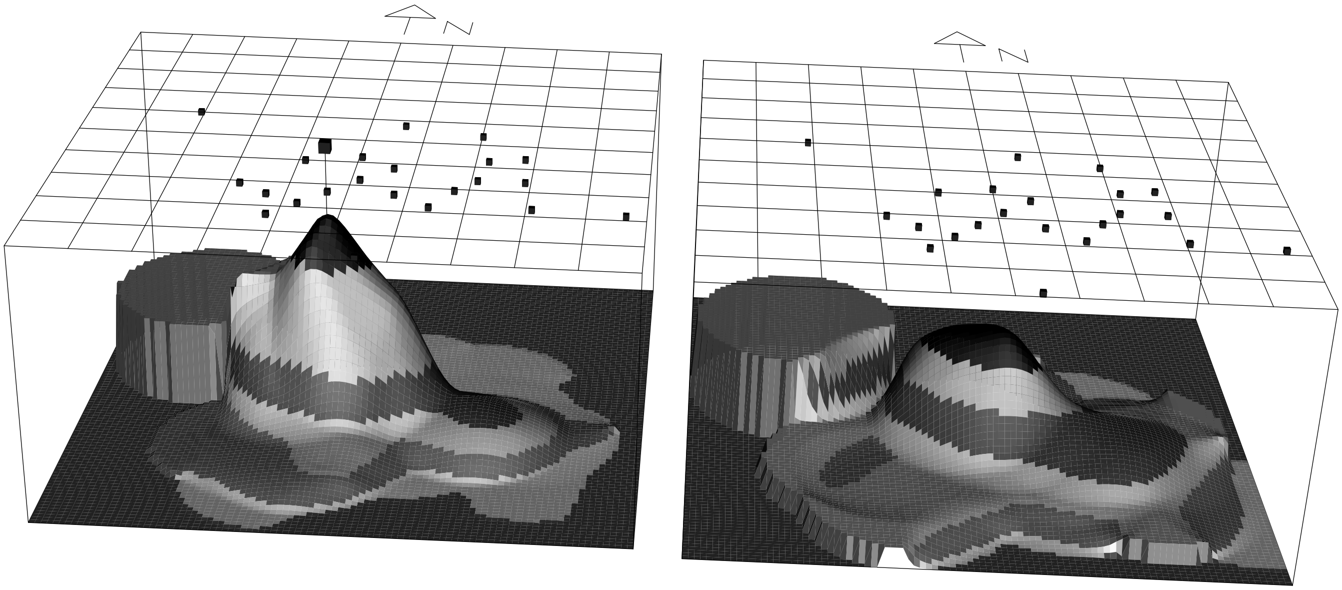

Figure 4.5: Estimates of ground-water contamination and absolute errors.

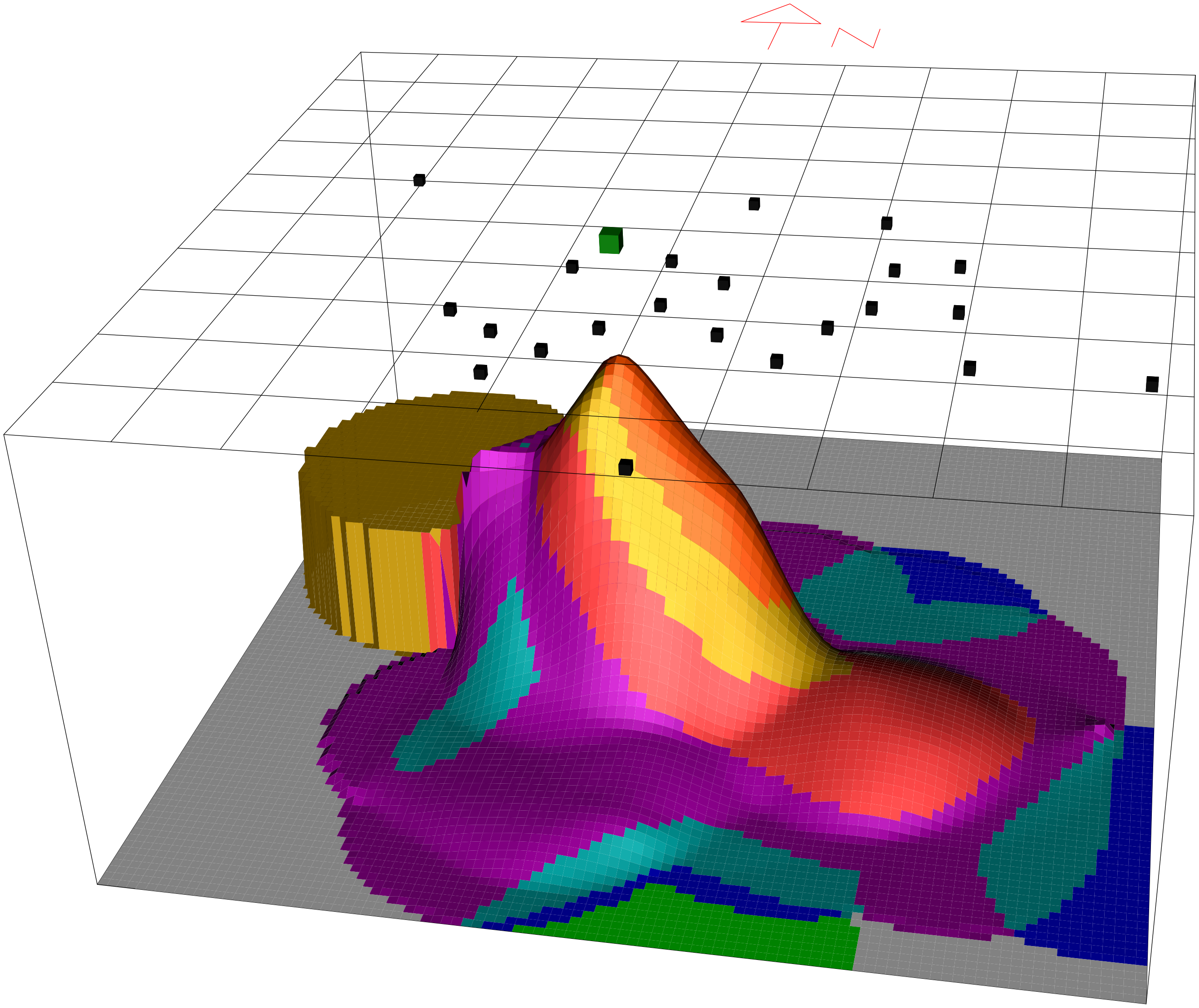

Figure 4.6 Estimate of ground-water contamination: contours represent magnitude of error.

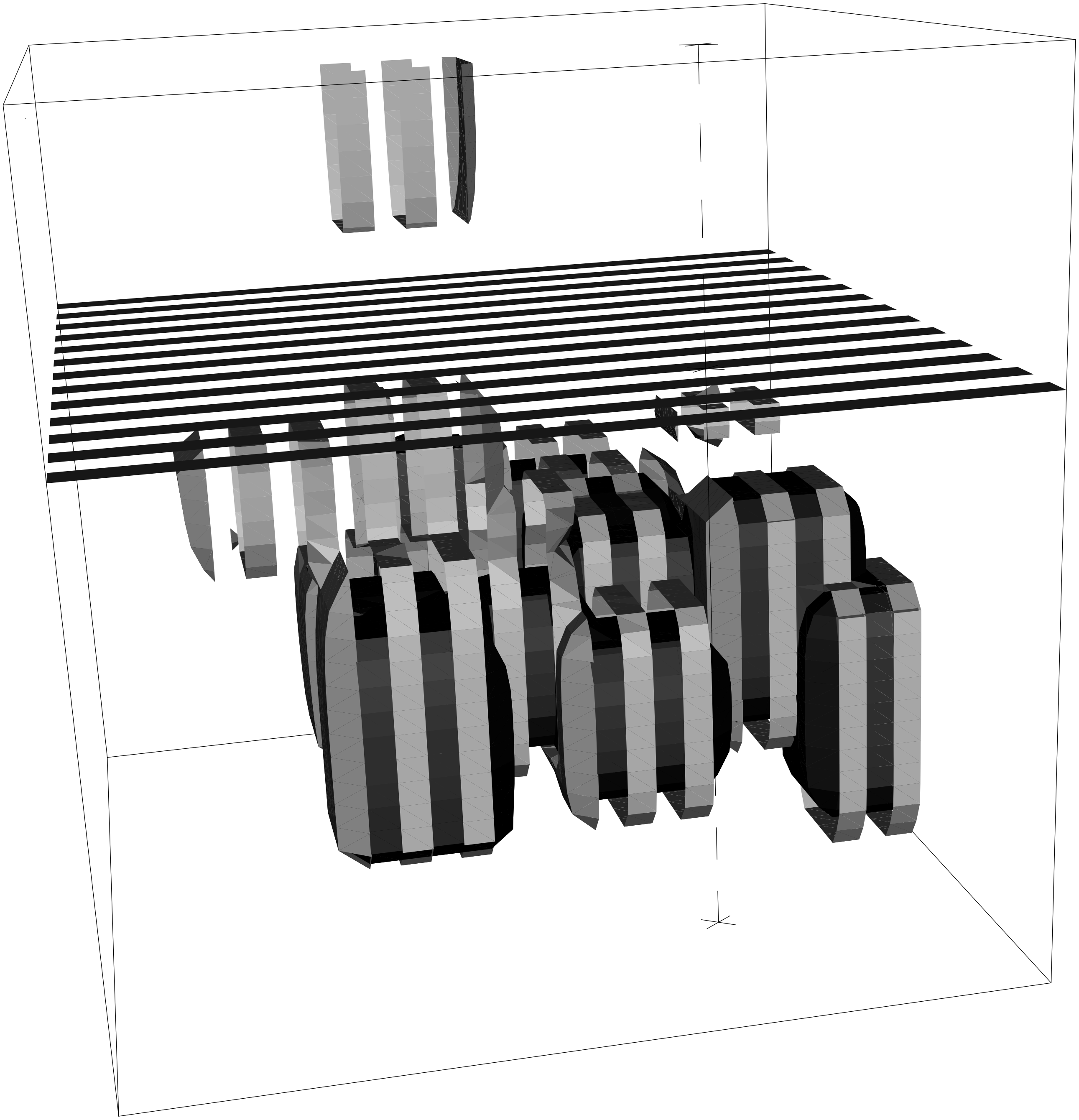

Figure 4.7 Smoothed absolute error estimate for Eglin AFB soil data.

Chapter 5
Estimation of Non-Linear Functionals of Random Processes for Environmental Problems

In analyzing environmental sampling data, it is often of interest to estimate some function of the data. For example, one may be interested in the maximum concentration attained within some region, the location of this maximum, the region for which the concentration exceeds some set value, or the total amount of contaminant present in a region. For example, government regulations on levels of ambient ozone typically involve exceedence of some threshhold deemed unsafe for human beings at *any* location within a region (Cox, et al.,1995). In this case, one would wish to estimate the maximum concentration attained in the region of interest. In the case where one is attempting to clean up ground water or soil contamination, it is of interest to know the total amount of contaminant present in a region. The estimation of total contaminant involves estimating an integral over a region and is what will be addressed in this section.

In this chapter, we will model the Eglin AFB ground-water BTEX observations as a realization of a stochastic process using the methods described in §3.2. The goal, then, will be to estimate the distribution of the integral of the process over some set region. Now in the case where this process is Gaussian, estimation of this distribution is a solved problem. However, it is commonly the case that environmental data are *lognormal* or well approximated by a lognormal distribution. Such is the case for the Eglin AFB BTEX observations.

The approach taken here is to estimate the integral of the lognormal process by Monte Carlo simulation of the process conditional on the data. The "conditional simulation" referred to here is similar to that discussed by Englund and Heravi (Englund, et al., 1995) in reference to Deutsch and Journel (Deutsch, et al., 1993). That is, one fits a semivariogram model (Cressie, 1993) to the data, then simulates from this model at some locations of interest. The actual measurements are honored at measured locations, possibly with some error variation. Englund and Heravi discuss the simulation as being along a regular grid; here we will be discussing simulation at random locations.

5.1 Description of the Problem

As previously stated, estimation of the total amount of contaminant present in some region is equivalent to estimation of the stochastic integral $Y = \int_A X(\mathbf{s})ds$ where $\langle X(\mathbf{s})\rangle$ is some random process, based on some (possibly noisy) observations $x(s_1),\ldots,x(s_n)$ where $x(s_i) = X(s_i) + \varepsilon_i$ and $\varepsilon_i \overset{i.i.d.}{\sim} N(0,\sigma_\varepsilon^2)$. The definition of a stochastic integral is given in (Boeckenhauer, 1996). In particular, we are interested in a point estimate of Y and an interval estimate (i.e., prediction interval). For the point estimate of a general function of the process g(X), assume initially squared error loss:

$$L(g(X),p(\mathrm{x};g(X))) = (g(X) - p(\mathrm{x};g(X)))^2 \tag{5.1}$$

where x is the vector of data and $g(\bullet)$ is some function of ⟨X⟩. Then the optimal predictor (i.e., the predictor $p(\bullet)$ which minimizes $E[L(\bullet)|\mathrm{x}]$) is:

$$p^O(\mathrm{x};g(X)) = E[g(X)|\mathrm{x}] \tag{5.2}$$

(Cressie, 1993). In particular, then, assuming squared error loss, the optimal predictor of the integral Y is

$$p^O(\mathrm{x};g(X)) = E[Y|\mathrm{x}] = E\left[\int_A X(s)\,ds|\mathrm{x}\right] \tag{5.3}$$

Now if Z(s) = X(s) is a Gaussian process, then it is shown in (Boeckenhauer, 1996) that the conditional distribution of *Y* given ***Z*** is normal, and formulas for the mean and variance are given. While these formulas for the conditional mean and variance involve the nontrivial task of calculating integrals of the mean and covariance functions, it is nonetheless in theory a solved problem.

However, suppose that instead ⟨X(s)⟩ is a lognormal process. That is, $X(\mathrm{s}) = e^{\boldsymbol{Z}}(\mathrm{s})\ \forall \mathbf{s} \in A$, where ⟨Z(s)⟩ is a Gaussian process. (Note, here we will actually be using base 10 logs, so we will instead define $X(\mathrm{s}) = 10^{Z(\mathrm{s})}$, but this does not affect the argument.) Now we can estimate $E[f_A\ Z(\mathrm{s})ds\ |\mathrm{z}]$ as above for the logged process. However, since exp(•) is a convex function, we have by Jensen's Inequality (Lehmann,1983),

$$\begin{aligned} E\left[\int_A X(\mathbf{s})\,ds|\mathbf{x}\right] &= E\left[\int_A \exp(Z(\mathbf{s}))ds|\mathbf{x}\right] \\ &= \int_A \left(\left[\exp(Z(\mathbf{s}))|\mathbf{x}\right]\right)ds \\ &= \int_A \left(\exp\left(E\left[Z(\mathbf{s})|\mathbf{x}\right]\right)\right)ds \end{aligned} \tag{5.4}$$

provided *var*(Z(s)| x) > 0, so we cannot simply use the conditional mean of the ⟨Z⟩ process to obtain the conditional mean of the ⟨X⟩ process. Furthermore, the distribution of the integral in this case will definitely *not* be normal, so some other method must be used to produce prediction intervals.

5.2 Using Monte Carlo to Estimate the Distribution of a Stochastic Integral

The integral which we estimate here is not merely the integral of a function, but rather the integral of a random process over a region, and thus is itself a random variable. We will still be able to use the method discussed in Section 3.3, but instead of evaluating the function at each location, we will simulate the value of the process at each location by generating from the model. We will actually simulate *several* realizations of the process, which will then give us an estimate of the *conditional distribution* of the integral, given the data.

The first step is to use our transformed observations (base 10 log) to produce a model for the process using the geostatistical methods in Section 3.2 and (Cressie,1993). Secondly, we will generate uniform locations over the region of interest, $\boldsymbol{A}$, as discussed in Section 3.3. Variation is introduced here, as discussed in Section 3.3, because we are using observations of the process at only certain locations in $\boldsymbol{A}$ to estimate the integral over the entire region $\boldsymbol{A}$. We must then calculate the covariance matrix and mean vector from the geostatistical model for these particular locations.

To simulate a realization of the random process, we first generate a multivariate normal process based on the mean and covariance obtained from our geostatistical modeling, then we exponentiate the simulated observations. From this realization, we may find the average as discussed in Section 3.3 to produce an estimate of the integral of *that realization*. That is, if we generated p locations $\mathbf{s_{1,1}}$ **,...,**$\mathbf{s_{1,p}}$ uniformly in $\boldsymbol{A}$, then simulated values $z_1(\mathbf{s_{1,1}}$), ...,$z_1(\mathbf{s_{1,p}}$) at these locations from the Geostatistical model, we are estimating the integral

$$\Phi_1 = \int_A 10^{z_1(\mathbf{S})} ds \tag{5.5}$$

where $10^{z1(\mathbf{s})}$, $\mathbf{s} \in \boldsymbol{A}$ denotes this realization. We then estimate the integral of the realization in the manner discussed in §3.3:

$$\hat{\Phi}_1 = \frac{|A|}{p} \sum_{j=1}^{p} 10^{z_1(\mathbf{s_{1,j}})} \tag{5.6}$$

The variance of the integral estimate for the realization, conditional on the data and the realization, may be estimated by

$$\begin{aligned} Var\left(\hat{\Phi}_1\right) &= Var\left(\frac{|A|}{p} \sum_{j=1}^{p} 10^{z_1(\mathbf{s_{1,j}})}\right) \\ &= \frac{|A|^2 s^2}{p} \end{aligned} \tag{5.7}$$

as in (3.19). To estimate the variance of the integral of the process, notice that

$$\begin{aligned} Var\left[\hat{\Phi}|\mathbf{x}\right] &= Var\left[E\left[\hat{\Phi}|X(\cdot),x\right]|x\right] + E\left[Var\left[\hat{\Phi}|X(\cdot),x\right]|x\right] \\ &= Var\left[\Phi|x\right] + E\left[Var\left[\hat{\Phi}|X(\cdot)\right]|x\right] \end{aligned} \tag{5.8}$$

(Recall that $\mathbf{x}$ is the vector of data.) An unbiased estimate of $E\left[Var\left[\hat{\Phi}|X(\bullet)\right]|data\right]$ is obtained from the sample mean of the variances of the $\hat{\Phi}_1$, i = 1,..., m. Also, $Var\left[\hat{\Phi}|\mathbf{x}\right]$ may be estimated by the sample variance of the integral estimates from the realizations. Sampling from a distribution which is not uniform may decrease the variance of the estimate in some cases, however, it is the large variation between the integral estimates from different realizations which should be a cause of concern. This latter error can only be reduced by either (1) improving the model, or (2) gathering additional data.

The above steps are repeated to obtain a large number of realizations thereby reducing error introduced by Monte Carlo estimation. For each simulation i, i = 1,..., m, we:

1. Generate p locations $\mathbf{s}_{i,1}$,...,$\mathbf{s_{i,p}}$ uniformly in $\boldsymbol{A}$.
2. For these locations, calculate the covariance matrix and mean vector, conditional on the data, using the geostatistical model.

3. Generate multivariate normal data $z_1(\mathbf{s_{i,1}}),\dots,z_1(\mathbf{s_{i,p}})$ using this mean vector and covariance matrix, and exponentiate them (using the same base as for the logs we took of the data to produce the model).

4. Produce an estimate of the integral

$$\Phi_i = \int_A 10^{z_i(\mathbf{s})} ds \tag{5.9}$$

using the sample-mean Monte Carlo method, i.e.

$$\hat{\Phi}_i = \frac{|A|}{p} \sum_{j=1}^{p} 10^{z_i(\mathbf{s_{i,j}})} \tag{5.10}$$

We may then get an estimate of the mean of the integral of the process

$$E\left[\Phi|\mathrm{x}\right] = E\left[\int_A 10^{Z(\mathrm{s})} ds|\mathrm{x}\right] \tag{5.11}$$

where x again denotes the vector of data, by, for example, taking the mean

$$\hat{E}\left[\Phi|\mathrm{x}\right] = \frac{1}{m} \sum_{i=1}^{m} \hat{\Phi}_i \tag{5.12}$$

More importantly, we are able to estimate the conditional distribution of the integral and get prediction intervals using the quantiles of the $\hat{\Phi}_i$. We may also estimate the variance of the integral

$$\begin{aligned} Var\left[\Phi|\boldsymbol{x}\right] &= Var\left[\hat{\Phi}|\boldsymbol{x}\right] - E\left[Var\left[\hat{\Phi}|X(\bullet),x\right]|\boldsymbol{x}\right] \\ &= \hat{Var}\,\hat{\Phi}_1,\dots,\hat{\Phi}_m - \frac{1}{m}\sum_{i=1}^{m} \hat{Var}\left(\hat{\Phi}_i\right) \end{aligned} \tag{5.13}$$

as discussed above.

5.3 Application

5.3.1 Discussion of Data

A QQ-plot of the log of the ground-water BTEX data from Eglin AFB, Figure 5.1, indicates that lognormality is a reasonable assumption in this instance. For further reference, we also plot the log of the data at the sampled spatial locations (Figure 5.2).

We now discuss the covariance modeling and integral estimation for the Eglin AFB site using the ground-water data introduced in Section 2.1. Recall that these data appear to be lognormal. In Section 5.3.1, we will estimate a variogram from the log 10 data to model the covariance structure, as discussed in Section 3.2. Section 5.3.1 contains details regarding integral estimation for this data, and thus estimation of the total amount of contaminant.

Variogram Estimation

Before performing spatial estimation on any data, we must first model the covariance structure of the data. The empirical semivariogram for the logged Eglin AFB ground-water data is shown in Figure 5.3. While this semivariogram does, in general, seem to increase with distance as expected, it nonetheless is rather undesirable. In particular, there is high variability, which (a) will make it difficult to estimate well with a vaiogram model and (b) will cause any estimates made from such a model to have large error variance.

It should be noted that the above variogram was done assuming that the data was *isotropic*, that is, that the covariance of the process at two locations depends only on the distance between these locations and not the direction. In fact, this does not really appear to be the case here, as Figure 5.2 seems to indicate that the plume extends along the northwest-southeast direction, due to the fact that the "large" observations seem to fall about this line. So it is possible here that we can get a better variogram, and thus a better fit, by first transforming the coordinate axes through a rotation and then scaling one of the axes. This allows us to take (this particular type of) anisotropic data and model it using a standard one-dimensional, isotropic variogram model. Using the Splus spatial module function *anisotropy.plot,* we are able to try various rotations and scalings of the coordinate axes in an attempt to produce a more appropriate variogram. Examples of these empirical variograms, along with loess smooth lines, are shown in Figures 5.4 and 5.5. Note that a number of these are obviously bad choices, as they do not even have γ increase with distance. We are looking for something with the points "well-clustered" about an *increasing* line. We concluded the best

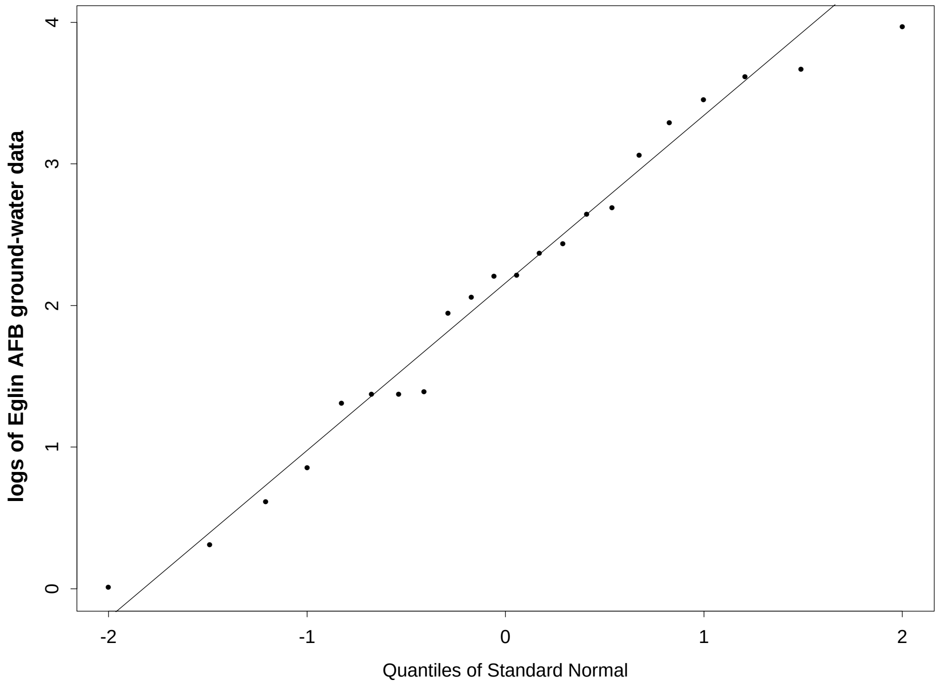

Figure 5.1 QQ-plot of logs of Eglin AFB ground-water data.

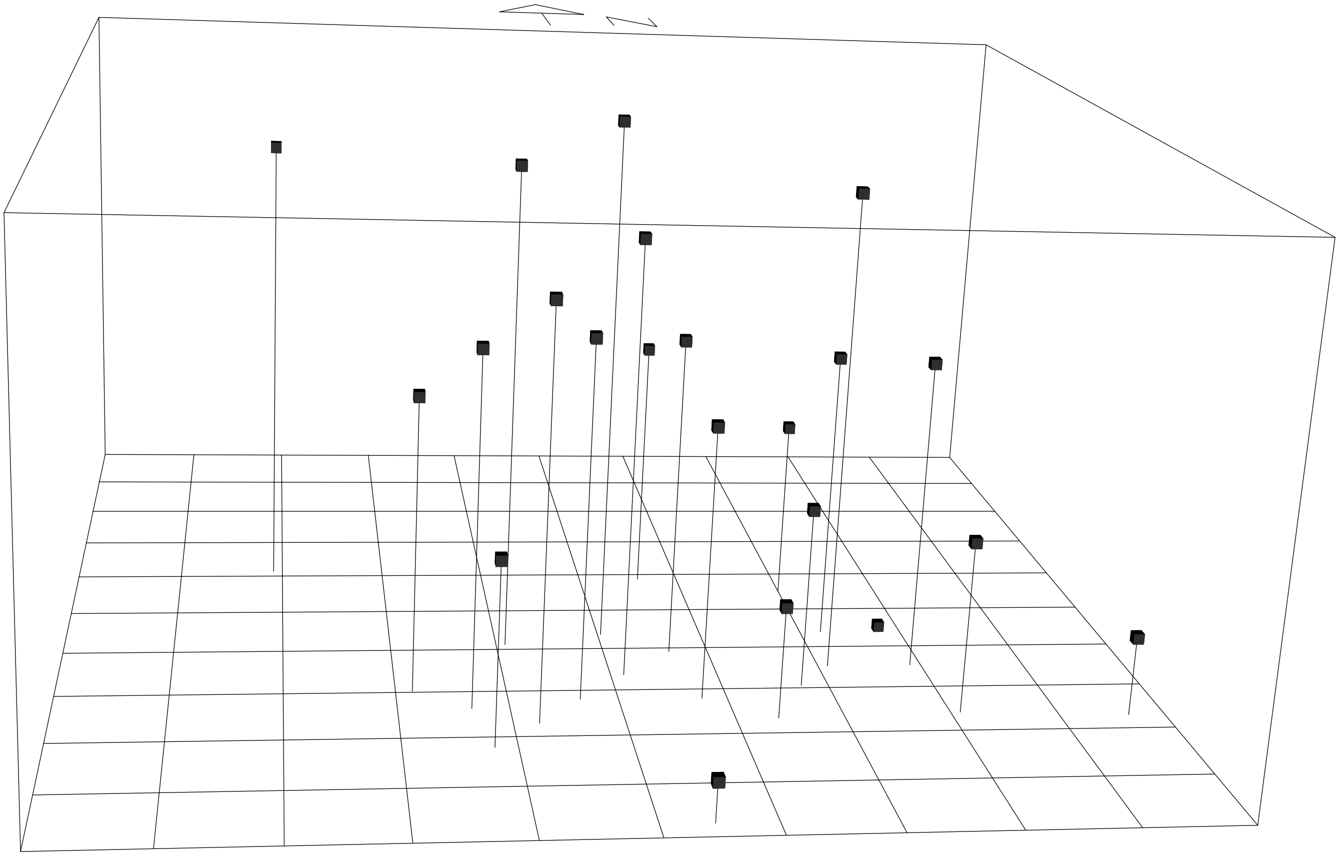

Figure 5.2 Logs of Eglin AFB ground-water data points.

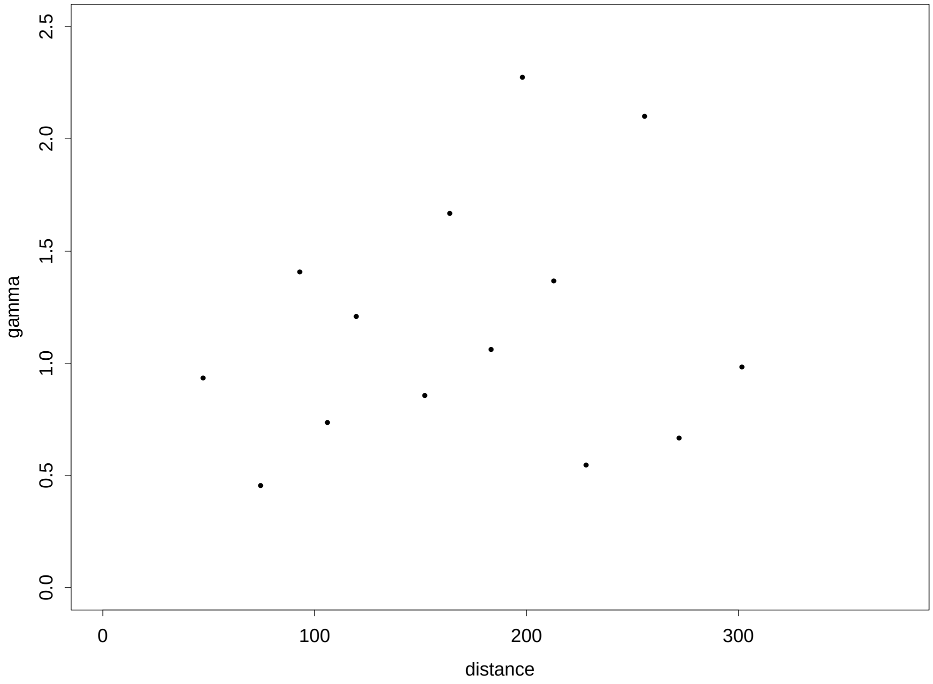

Figure 5.3 Empirical semivariogram of logged Eglin AFB ground-water data.

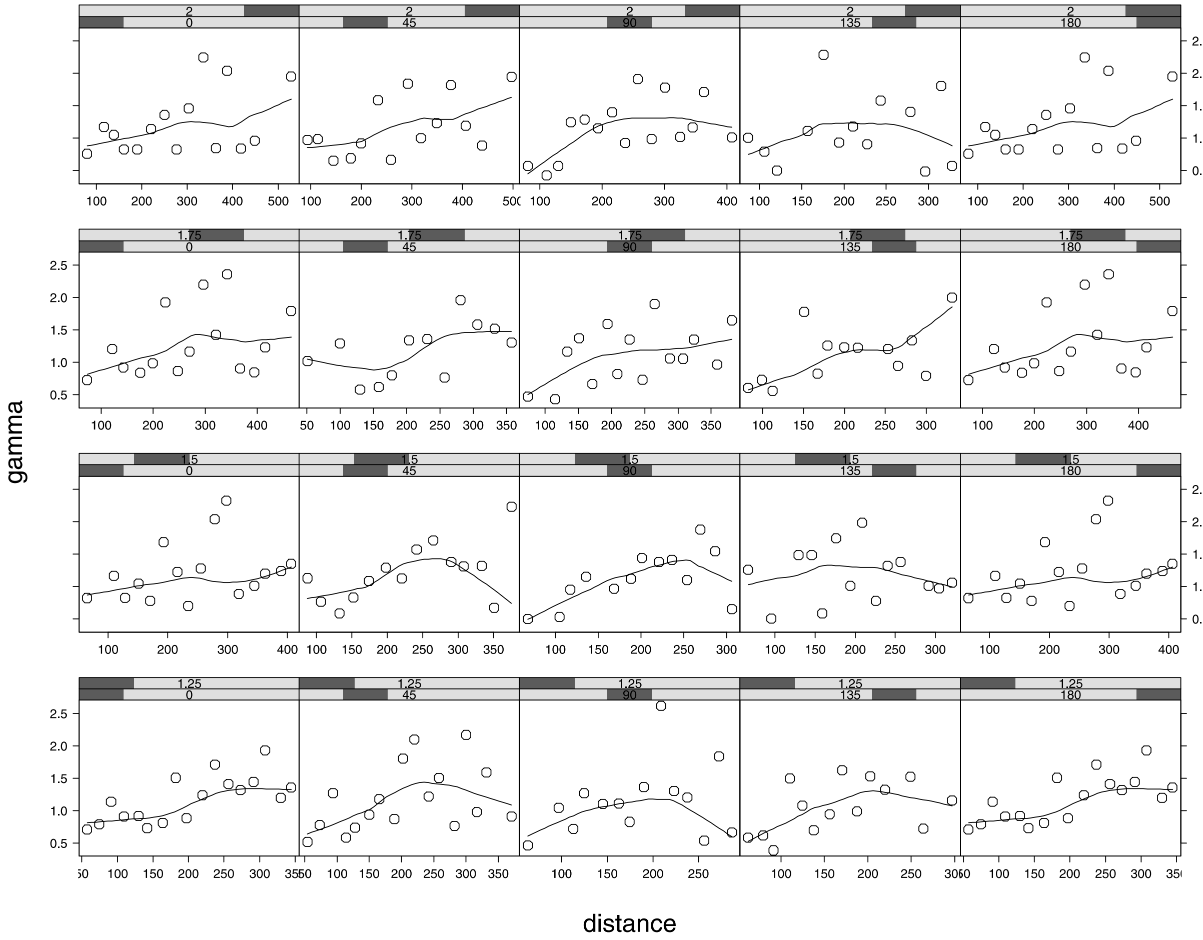

Figure 5.4 Empirical semivariograms with rotation angles from 0° on the left to 180° on the right and ratios from 1:25 at the bottom to 2 at the top.

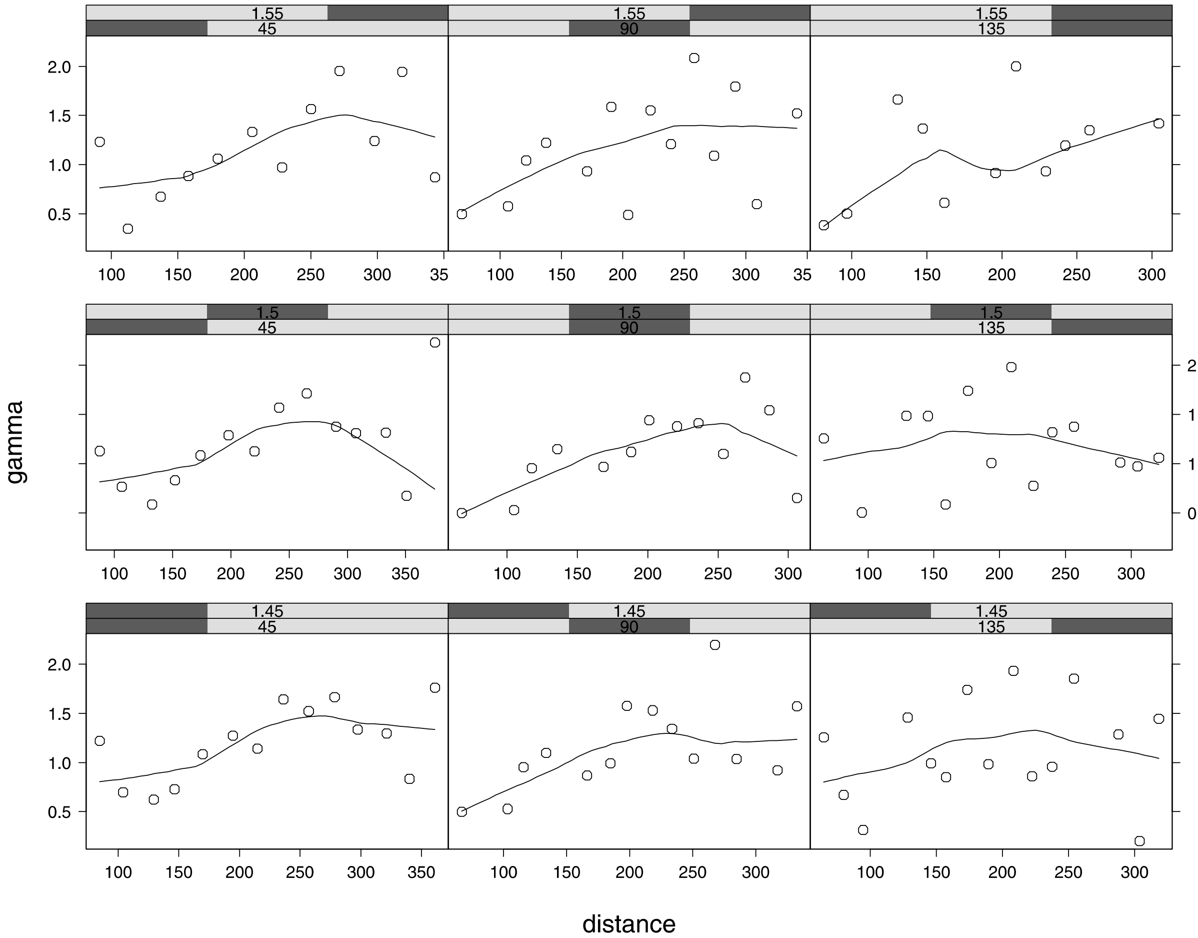

Figure 5.5 Empirical semivariograms with rotation angles from 45° on the left to 135° on the right and ratios from 1:45 at the bottom to 1:55 at the top.

option was a rotation angle of 45° and an axis ratio of 1.45. That is, for the i^{th} data point at location ($x[i]$, $y[i]$), we multiply the location by the matrix **B** where

$$\begin{aligned}\mathbf{B} &= \begin{pmatrix} \cos^2(45) + 1.45{*}\sin^2(45) & (1-1.45){*}\sin(45){*}\cos(45) \\ (1-1.45){*}\sin(45){*}\cos(45) & \sin^2(45) + 1.45{*}\cos^2(45) \end{pmatrix} \\ &= \begin{pmatrix} 1.225 & -0.225 \\ -0.225 & 1.225 \end{pmatrix} \end{aligned} \tag{5.14}$$

as in (3.9). This empirical semivariogram, along with two different semivariogram models, is shown in Figure 5.6. The spherical semivariogram, with the formula

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ c_0 + c_s\left(1.5{*}\left(h/a_s\right) - .5{*}\left(h/a_s\right)^3\right) & 0 < h \le a_s \\ c_0 + c_s & h \ge a_s \end{cases}$$

(Cressie, 1993) was produced with nugget $c_0 = 0$, (partial) sill $c_s = 1.51$, and range $a_s = 325$. This model appears as the dashed line in Figure 5.6. The rational quadratic semivariogram has formula

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ c_0 + \dfrac{c_r * h^2}{1 + h^2/a_r} & h > 0 \end{cases} \tag{5.15}$$

(Cressie, 1993) and was calculated with nugget $c_0 = 0.19$, $c_r = 1/14580$, and $a_r = 24750$. This variogram model appears as the solid line in the figure. In both cases, h refers to the *distance* (in the transformed space) between the data points. The two variogram models appear very similar for most distances in Figure 5.6, but are different in the tails. While either one would likely work reasonably well, we chose to use the rational quadratic model as it produces a plausible variogram which has a positive nugget effect (and a nugget effect of zero would assume that there was no measurement error, which is certainly an unreasonable assumption). The kriged surface estimate produced using this variogram model is shown in Figure 5.7 and the standard error surface is depicted in Figure 5.8. The kriging here was done using the spatial module of S-PLUS (MathSoft, 1995), which estimated the constant mean to be $\mu = 1.55344$ on the log_{10} scale.

Details of the Integral Estimation

Programs to estimate the integral for the Eglin AFB ground-water data were written in the C programming language and appear in (Boeckenhauer, 1996). The method is essentially that described in Section 5.2. That is, for each of `M` realizations of the process, we simulate values at `N` locations within the region, where `N` and `M` are contained in the file `constants.h`. The main program, contained in `runsim.c`, first uses the data locations to produce Σ_{11} , the covariance matrix for these locations. As discussed above, the locations are first corrected for geometric anisotropy by multiplying each pair of locations $\mathbf{t}_i = (x,y)$ by the matrix B in (5.14). Then the distance (in transformed space) between each pair of points is calculated, followed by the covariance between the locations using the rational quadratic model selected above. Specifically, notice that since we are assuming the data are of the form

$$z(\mathbf{t}_i) = Z(\mathbf{t}_i) + \varepsilon_i \qquad i = 1,\ldots,22 \tag{5.16}$$

where ε_i and ε_j are independent for $i \neq j$ and $Z(\mathbf{t}_j)$ is independent of ε_i for every i and j, we have

$$\begin{aligned} Cov\left(z(\mathbf{t}_i), z(\mathbf{t}_j)\right) &= Cov\left(Z(\mathbf{t}_i) + \varepsilon_i, Z(\mathbf{t}_j)\varepsilon_j\right) \\ &= Cov\left(Z(\mathbf{t}_i), Z(\mathbf{t}_j)\right) + Cov\left(Z(\mathbf{t}_i), \varepsilon_j\right) + Cov\left(\varepsilon_i, Z(\mathbf{t}_j)\right) + Cov\left(\varepsilon_i, \varepsilon_j\right) \\ &= Cov\left(Z(\mathbf{t}_i), Z(\mathbf{t}_j)\right) + Cov\left(\varepsilon_i, \varepsilon_j\right) \\ &= \begin{cases} Cov\left(Z(\mathbf{t}_i), Z(\mathbf{t}_j)\right) & \text{if } i \neq j \\ Var\left(Z(\mathbf{t}_i) + Var(\varepsilon_i)\right) & \text{if } i = j \end{cases} \end{aligned} \tag{5.17}$$

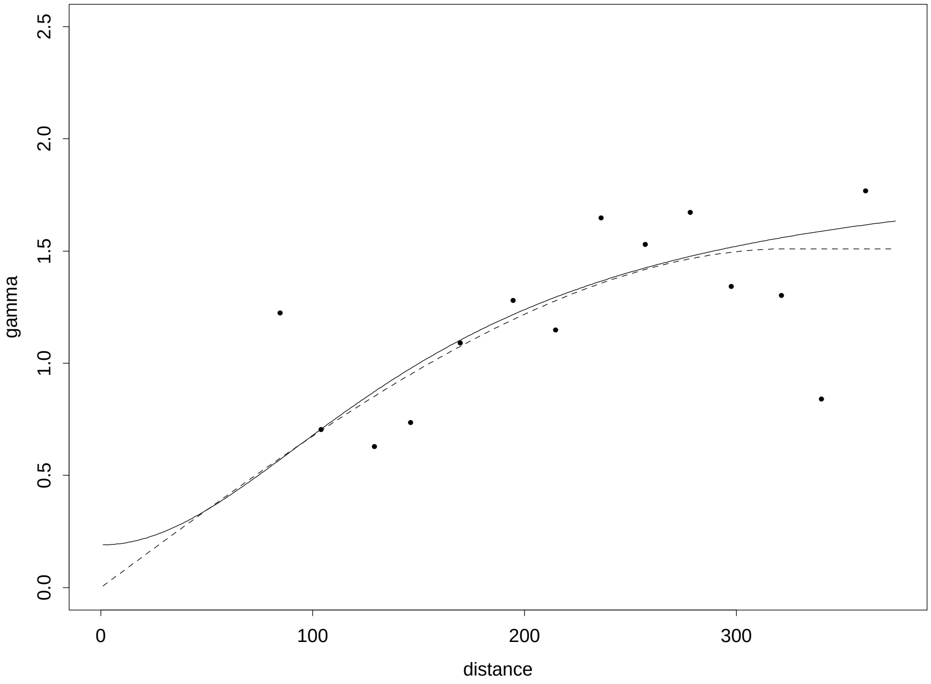

Figure 5.6 Semivariogram of logged Eglin AFB ground-water data with transformed locations.

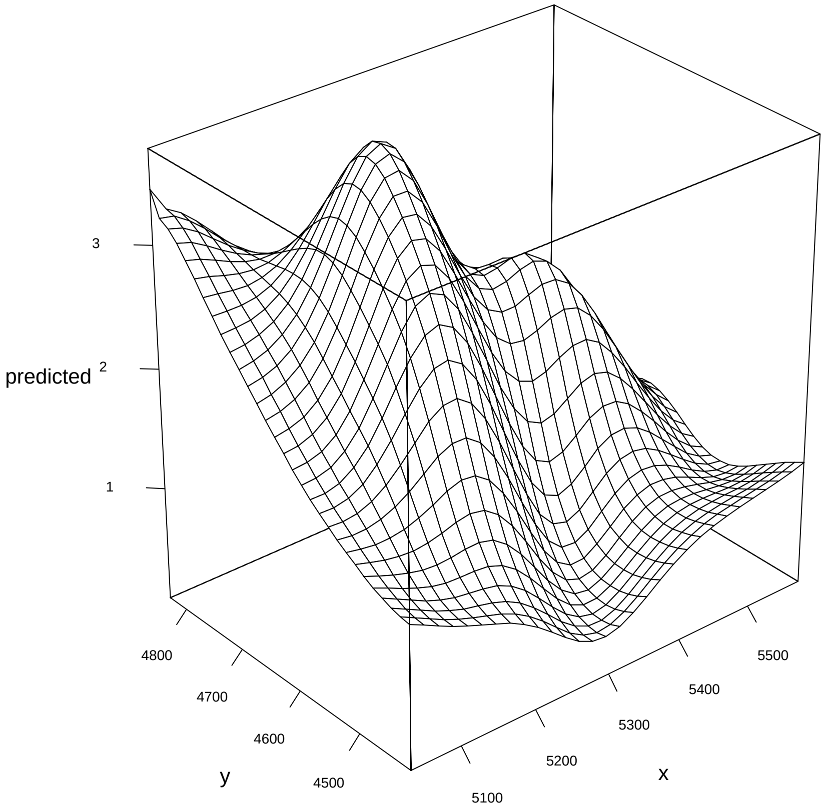

Figure 5.7 Kriged surface estimate of logged Eglin AFB ground-water data using rational quadratic variogram details of the integral estimation.

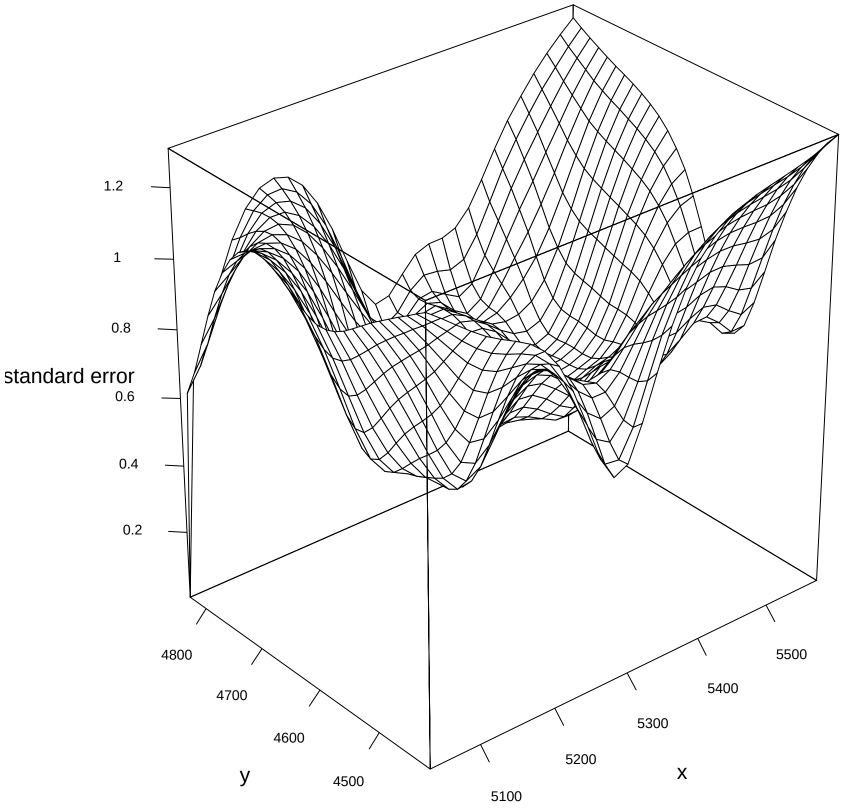

Figure 5.8 Standard errors for kriged surface estimate of logged Eglin AFB ground-water data using a rational quadratic variogram.

Recall that we are using the rational quadratic model. So letting $h = \left|\mathbf{t}_i^{'} - \mathbf{t}_j^{'}\right|$ (i.e., *h* is the distance between locations i and j after transforming locations), we have:

$$Cov\left(Z(\mathbf{t}_i), Z(\mathbf{t}_j)\right) = C(h) = c_r a_r - \frac{c_r * h^2}{1 + h^2/a_r} \tag{5.18}$$

and

$$Var(\varepsilon_i) = c_0 \tag{5.19}$$

(i.e., the nugget effect). Thus the *i, jth* element of Σ_{11} is given by:

$$\begin{aligned}\left(\Sigma_{11}\right)_{i,j} &= Cov\left(z(\mathbf{t}_i), z(\mathbf{t}_j)\right) \\ &= \begin{cases} c_r a_r - \dfrac{c_r h^2}{1 + h^2/a_r} & = \dfrac{c_r a_r}{1 + h^2/a_r} \quad \text{for } i \neq j \\ c_0 + c_r a_r - \dfrac{c_r * h^2}{1 + h^2/a_r} & = c_0 + c_r a \quad \text{for } i = j \end{cases}\end{aligned} \tag{5.20}$$

Recall that c_0 = 0.19, c_r = 1/14580, and a_r = 24750 were the parameters used for our model in the previous section.

The function *sim*, contained in `sim.c`, is then called `M` times, each time producing a sample from a realization of the process of size `N`. For each simulation, we first generate `N` locations uniformly over the region using the function *loc* in `loc.c`, first generating all of the *x* values, then all of the *y*s. Then the function *cov*, contained in `cov.c` is used to calculate the conditional mean vector and covariance matrix for these generated locations given the data. Using standard multivariate normal theory (Mardia, et al., 1979), we first calculate the unconditional matrix for the generated locations using the rational quadratic model. Letting s_1 ,..., s_N be the simulated sampling locations, this yields the *NxN* matrix $\sum_{22}$ where the *i, j*th element is given by

$$\begin{aligned}\left(\sum_{22}\right)_{i,j} &= Cov\left(z(\mathbf{s}_i), z(\mathbf{s}_j)\right) \\ &= Cov\left(Z(\mathbf{s}_i), Z(\mathbf{s}_j)\right) \\ &= C\left(\left|\left(\mathbf{s}_i - \mathbf{s}_j\right)\mathbf{B}\right|\right) \\ &= c_r a_r - \frac{c_r h^2}{1 + h^2/a_r} \\ &= \frac{c_r a^2}{1 + h^2/a_r}\end{aligned} \tag{5.21}$$

where *h* is the distance between the locations after transformation; i.e., $h = \left|\mathbf{s}_i\mathbf{B} - \mathbf{s}_j\mathbf{B}\right| = \left|\left(\mathbf{s}_i - \mathbf{s}_j\right)\mathbf{B}\right|$. Similarly, we calculate the *N*x*n* cross-covariance matrix $\sum_{21}$ with the *i, j*th element given by:

$$\begin{aligned}\left(\sum_{21}\right)_{i,j} &= Cov\left(z(\mathbf{s}_i), z(\mathbf{t}_j)\right) \\ &= Cov\left(Z(\mathbf{s}_i), Z(\mathbf{t}_j) + \varepsilon_j\right) \\ &= Cov\left(Z(\mathbf{s}_i), Z(\mathbf{t}_j)\right) + Cov\left(Z(\mathbf{s}_i), \varepsilon_j\right) \\ &= C\left(\left|\left(\mathbf{s}_i - \mathbf{t}_j\right)\mathbf{B}\right|\right) \\ &= c_r a_r - \frac{c_r h^2}{1 + h^2/a_r} \\ &= \frac{c_r a^2}{1 + h^2/a_r}\end{aligned} \tag{5.22}$$

with *h* in this case being $h = \left|\mathbf{s}_i\mathbf{B} - \mathbf{t}_j\mathbf{B}\right| = \left|\left(\mathbf{s}_i - \mathbf{t}_j\right)\mathbf{B}\right|$. We then can find the conditional covariance matrix

$$\sum = \sum_{22} - \sum_{21} \sum_{11}^{-1} \sum_{21}^{T} \tag{5.23}$$

and the conditional mean vector

$$\mu = \mu\mathbf{1}_N + \sum_{21} \sum_{11}^{-1} \left(\boldsymbol{Z} - \mu\mathbf{1}_n\right) \tag{5.24}$$

where $\mathbf{1}_N$ ($\mathbf{1}_n$) is a vector of 1s of length *N* (*n*), and μ is the constant mean as estimated by the kriging (1.55344) and $\boldsymbol{Z}$ is the vector of log_{10} observations. Calculation of $\sum_{11}^{-1} \sum_{21}^{\boldsymbol{T}}$ was accomplished with the use of the LAPACK function dspsv (Anderson, et al., 1992).

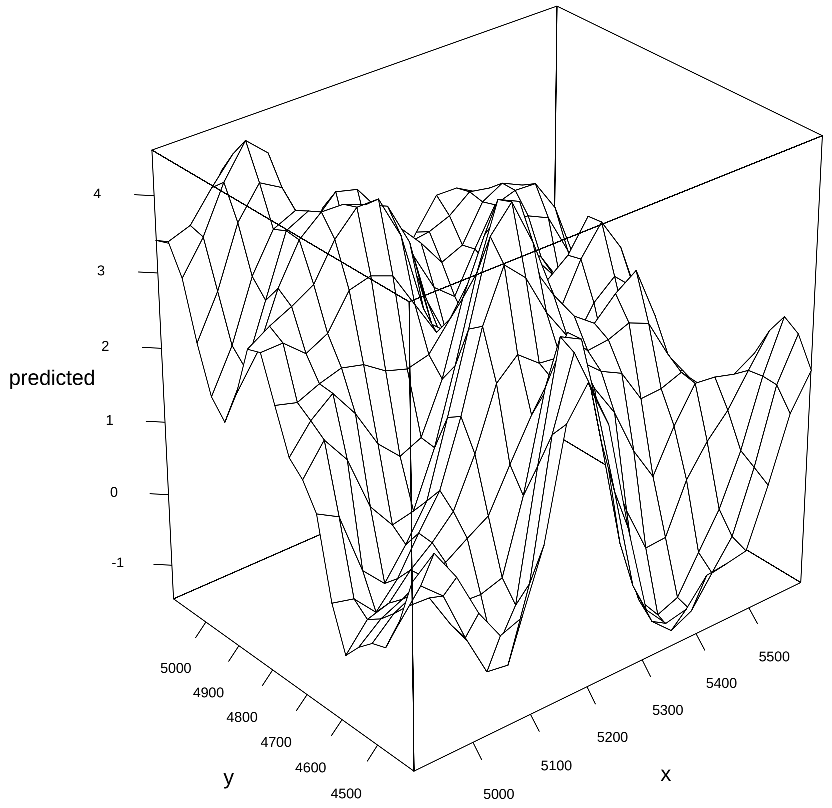

Figure 5.9 Surface of a grid simulation.

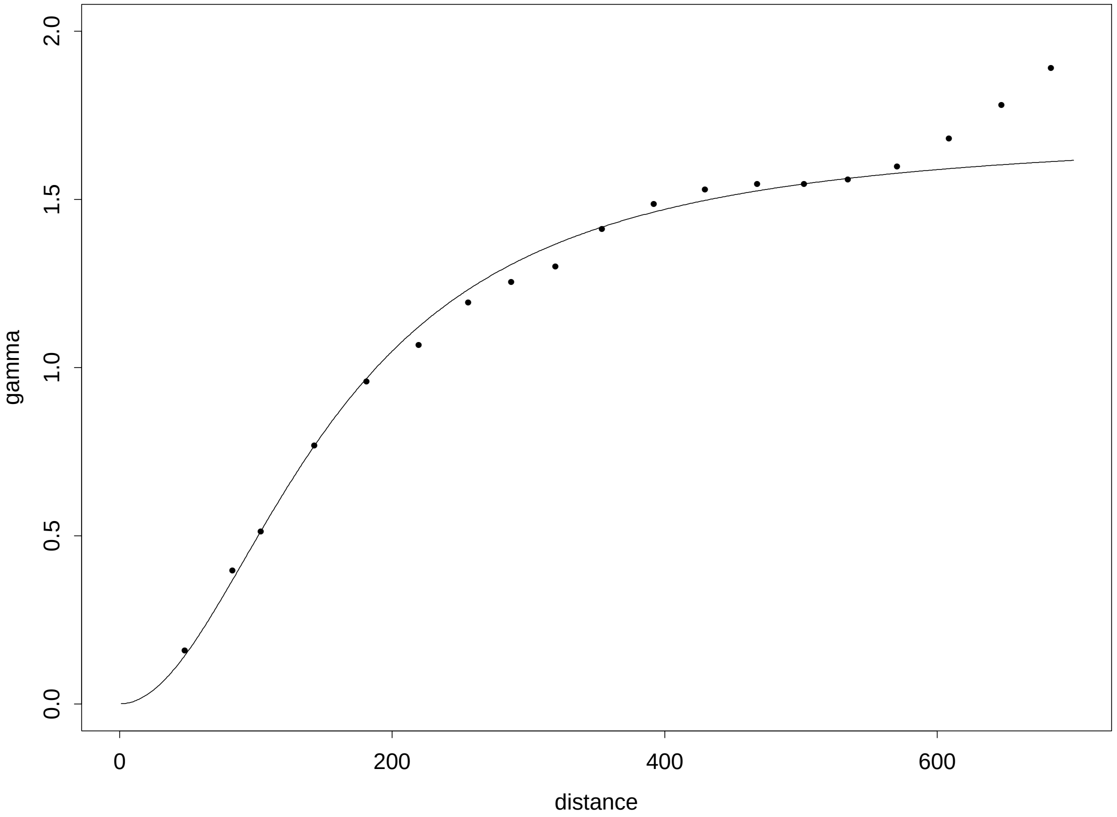

Figure 5.10 Semivariogram calculated from a grid simulation.

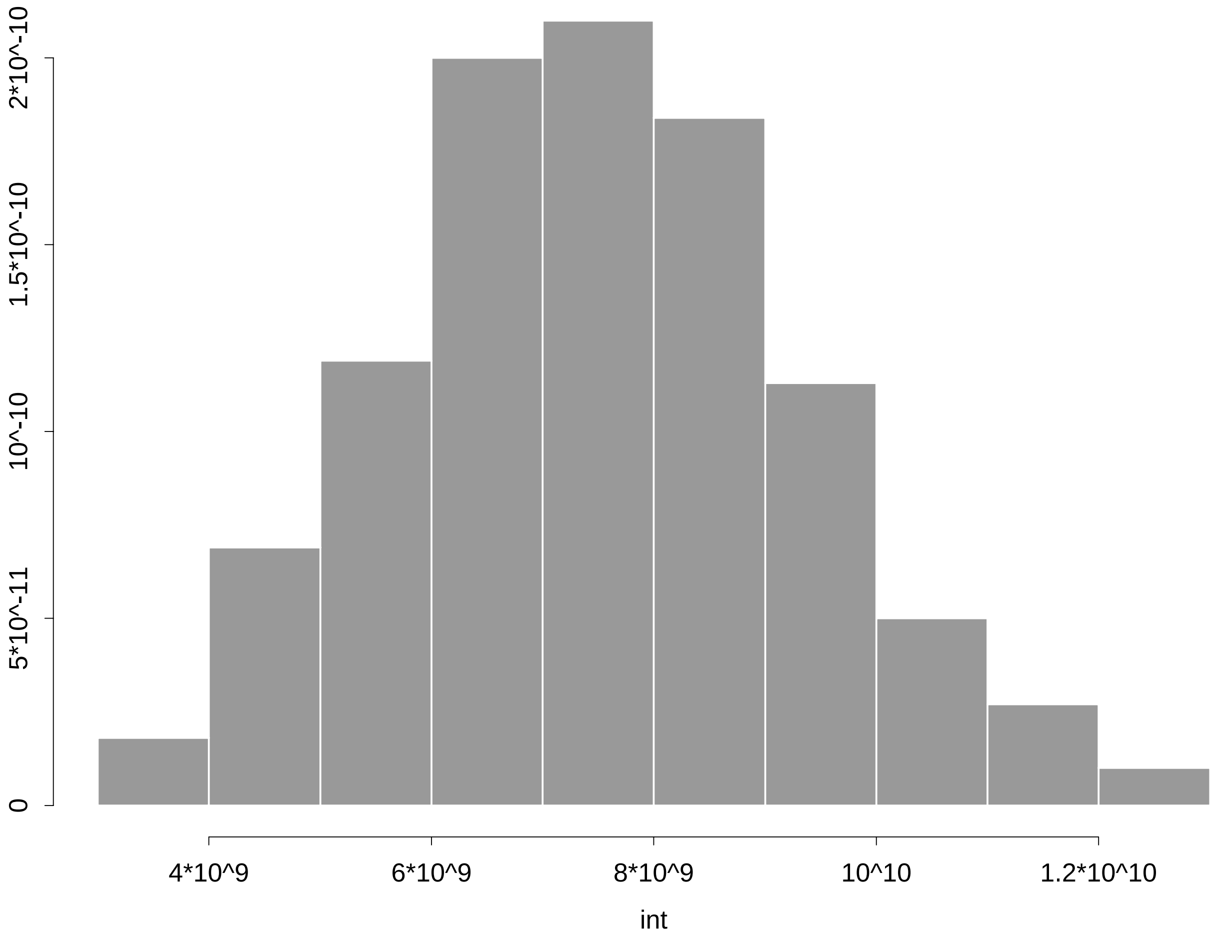

Figure 5.11 Histogram of integral estimates from 1000 samples of size 500: one realization.

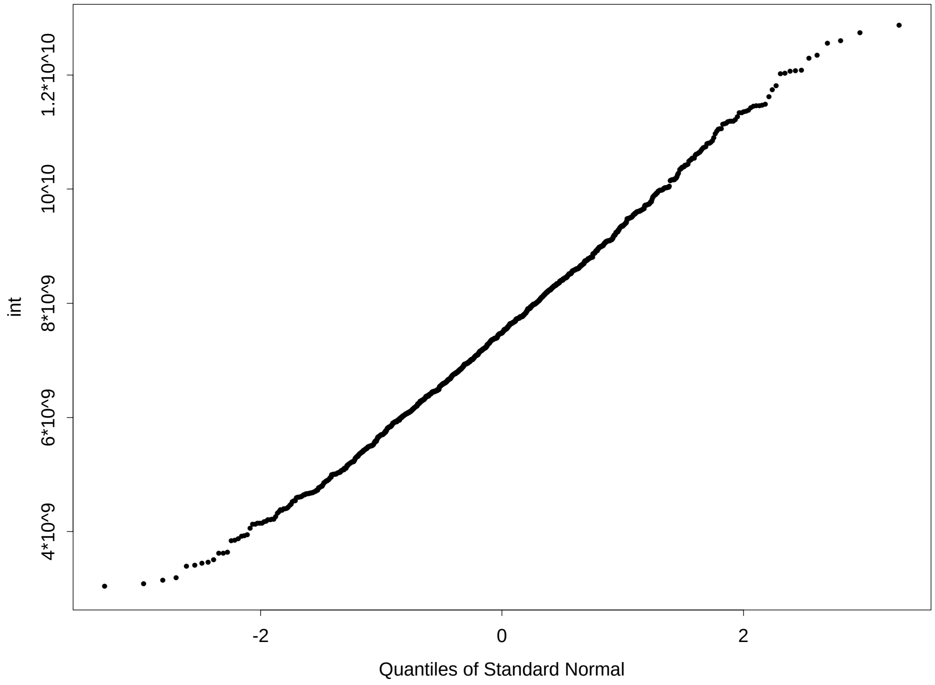

Figure 5.12 QQ-plot of integral estimates from 1000 samples of size 500: one realization.

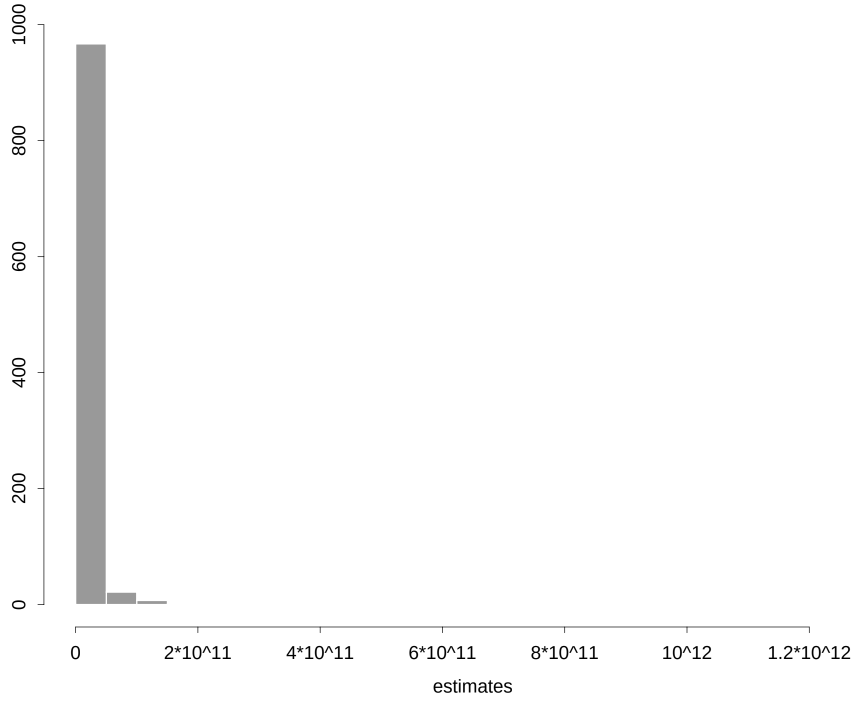

Figure 5.13 Histogram of integral estimates from 1000 samples of size 500: different realizations.

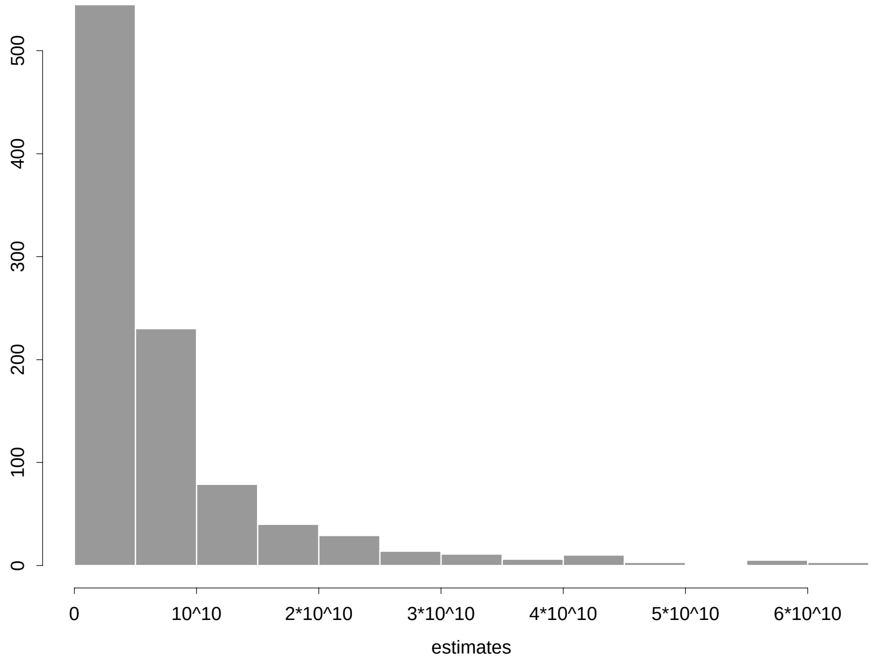

Figure 5.14 Histogram of lower 97:5% integral estimates from 1000 samples of size 500: different realizations.

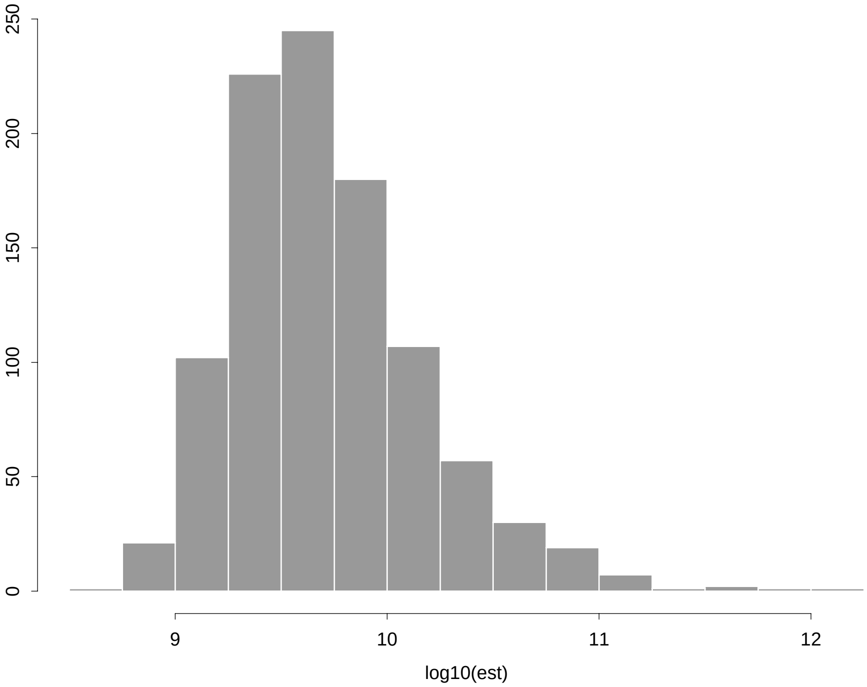

Figure 5.15 Histogram of logged integral estimates from 1000 samples of size 500: different realizations.

The function *sim* then calls the function *multnorm*, located in `multnorm.c`, which uses $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ to generate multivariate normal random variables from $N_{\boldsymbol{N}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We first generate an N-vector $\mathbf{Y}$ of standard normal random variates, i.e. $\mathbf{Y} \sim N_p(0_p, I_p)$ where $\mathbf{0}_p$ is a p-vector of zeroes and $\mathbf{I}_p$ is the pxp identity matrix. This is done using the function *gauss* (Reilly, 1995) in `boxmul.c`. The lower triangular Cholesky decomposition matrix $\mathbf{L}$ is calculated by using the LAPACK function *dpptrf* (Anderson, et al.,1992). (Recall that $\mathbf{L}$ is the unique lower triangular matrix such that $\Sigma = \mathbf{L}\mathbf{L}^{\boldsymbol{T}}$.). Finally, we calculate $\mathbf{X} = \mathbf{L}\mathbf{Y} + \boldsymbol{\mu}$. Since *dpptrf* calculates $\mathbf{L}$ in packed format (Anderson, et al., 1992), it was necessary to write code to do the matrix multiplication. This was done using as few operations as possible.

Now predictions of the process at the simulated locations are obtained by letting $Z(\mathbf{s_j}) = 10^{X}(\mathbf{s_j}) = 10^{\mathbf{X_j}}$. The integral of the realization is then estimated by

$$\hat{\Phi} = \frac{k^2|A|}{N}\sum_{j=1}^{N} Z(\mathbf{s_j}) \qquad (5.25)$$

as discussed in §3.3. Note that in this case A is rectangular, so $|A|$ is easy to calculate. The constant $k = 3.048$ refers to the number of *dm/ft*, since the (x,y) locations are in *ft* and the concentration predictions $\boldsymbol{Z}(\mathbf{s}_j)$ are in *µg/L*. Since this region is only 2-dimensional, the integral estimate here is in *µg/dm*, and would need to be multiplied by a depth in *dm* to give an estimate of the total amount of contaminant. Also, the variance of the integral estimate is calculated in the manner discussed in §3.3. However, instead of calculating $\mathrm{v\hat{a}r}(\hat{\Phi}_i) = (k^4|A|^2/N^2)\,\mathrm{v\hat{a}r}\left(\sum_{j=1}^{N} 10^{z_i(\mathbf{s_{i,j}})}\right)$, it is calculated as

$$\mathrm{v\hat{a}r}(\hat{\Phi}_i) = (1/N^2)\,\mathrm{v\hat{a}r}\left(\sum_{j=1}^{N} W_{i,j}\right) = (1/N)\mathbf{s}^2_{W_i} \qquad (5.26)$$

where $W_{i,j} = k^2|A|10^{z_i(\mathbf{s_{i,j}})}$ and $\mathbf{s}^2_W$ is the sample variance of $W_{i,1}$, ... ,$W_{i,N}$. This amounts to the same thing but is easier to program, as we can then use the integral estimate in calculating the sample variance.

This entire procedure is performed `M` times and the resulting integral estimates from the realizations are then used to estimate the conditional distribution of the integral of the process, given the data. A point estimate of the integral may be found by, for example, taking the mean of the integral estimates for the realizations. Approximate 100(1— α)% prediction intervals may be found by using the α/2 and 1— α/2 quantiles of the simulated integral estimates.

5.3.2 Results

As a quick assessment that our computer code is simulating the desired process, we generate, from the assumed model, a single realization over a 20x20 grid which encompasses the region, where the x-values range from 4900 to 5600 and the y-values range from 4400 to 5100. A perspective plot of this single realization is given in Figure 5.9. The realization behaves as one might expect given the conditional mean shown in Figure 5.7. Also, the empirical variogram of this realization compares well to the rational quadratic variogram model used to generate the data (see Figure 5.10).

Recall that the estimate of the integral for a realization, as discussed in §3.3, is given by

$$\hat{\Phi} = \frac{|A|}{N} \sum_{j=1}^{N} q(\mathbf{s_j}) \tag{5.27}$$

where the $q(\mathbf{s_j})$, $j = 1,..., N$ are the values of realization at locations $s_1,..., s_N$. Noting that this is in fact a sample mean of the $|A|q(s_j), j = 1,..., N$, we have by the central limit theorem that, for a single realization, the estimate of the integral should have a normal distribution as $N \to \infty$. We would like to check the integral estimates produced from samples of a *single realization* to see that this is in fact the case.

Ideally, we would like to be able to take a large number of samples from a single realization of size P, where the samples are sufficiently sized for the asymptotics to "kick in" (say $N = 500$ or 1000). However, since we are simulating the realization, we must simulate all of the values of the realization which we wish to sample together initially. If there are P such values of the realization, this means that we must not only calculate a PxP covariance matrix, but must then calculate the Cholesky decomposition of this matrix. As P gets large, near singularities (i.e., singularities within machine precision) in the conditional covariance matrix Σ cause the Cholesky decomposition routine to fail. The largest value of P for which the code could be successfully run with any consistency was 2000, which does not allow for a great number of independent samples of size 500, to say the least.

In lieu of independent samples we rely on dependent samples; i.e., samples which share some of the same realization values. A single realization of size 2000 was simulated and subsamples of size 500 were obtained from the 2000. The subsampling was performed with replacement. So each sample is then a sample from this realization taken at 500 independent uniform locations over the region. From each of these samples, an integral estimate was calculated using the integral estimate (5.25). Figure 5.11 shows a histogram of the integral estimates taken from 1000 such samples. The histogram appears to be approximately normal. Figure 5.12 contains a qq-plot of the 1,000 integral estimates. The qq-plot is very close to a straight line, although a bit off in the tails. Again, this indicates that these integral estimates are approximately normally distributed. Also, the integral estimate produced by the entire realization was approximately 7.548×10^9, which is also very close to the center of the histogram and the median of the integral estimates from the samples. Thus, the distribution of integral estimates for samples from a single realization appears to be fairly normal, as it should be.

Finally, to actually estimate the distribution of the integral, we generated samples of 500 from each of 1000 different realizations. Figure 5.13 shows a histogram of the 1000 different integral estimates. This histogram is obviously *very* skewed, to the extent that we cannot see any of the detail in the lower part of the histogram, where most of the values reside. Figure 5.14 shows a histogram of the lower 97.5% of these estimates and Figure 5.15 shows a histogram of the $\log_{10}$ estimates. These allow us to see more detail in the lower part of the histogram. Note that the units for the estimates are *μg/dm*, and they would have to be multiplied by a measurement of depth in *dm* to provide an estimate of total contaminant for a *three*-dimensional region. Tables 5.1 and 5.2 contain summary statistics about the integral estimates for the realizations. We can see the effect of the skewness in these summary statistics. For example, the value of the sample mean is more than *twice* the value of the sample median. Furthermore, the values of the α% trimmed means *rapidly* approach the value of the median as α increases.

Table 5.1 Statistics of Integral Estimates for 1000 Realizations

minimum	0.025 quantile	median	0.975 quantile	maximum
5.604×10^8	1.034×10^9	4.527×10^9	6.354×10^{10}	1.252×10^{12}

Table 5.2 Means of Integral Estimates for 1000 Realizations

mean	5% trimmed mean	10% trimmed mean	20% trimmed mean
1.204×10^{10}	7.776×10^9	6.849×10^9	5.946×10^9

Recall that we said in Section 5.1 that the conditional mean (i.e., the mean calculated above) was an optimal estimate of the integral under squared error loss. Now either the median or any of the means could be used as a point estimate of the integral of the process. In particular, in the case of absolute error loss

$$L(g(\boldsymbol{X}), p(\mathbf{x}; g(\boldsymbol{X}))) = \left| g(\boldsymbol{X}) - p(\mathbf{x}; g(\boldsymbol{X})) \right| \tag{5.28}$$

the median is actually optimal. However, due to the high skewness and the large difference between the median and the mean, we would do well to exercise caution in our choice. More investigation is necessary to determine which of these is more representative of the truth, or if some other statistic would be better, or if what is "better" depends on the particular application.

It is important to reiterate here that even though the distribution of integral estimates from samples from *one* realization is quite normal, the distribution of the integral of the process is not even close to normal. However, this is not surprising as these are two entirely different distributions. The first is the distribution of estimates for a realization with variation coming only from the "sampling error", i.e. the error induced by estimating a value of the integral for a region with only a finite number of points. The second involves this sampling error, along with the actual variation of the process.

As the distribution of the integral is not normal, we obviously cannot use typical normal prediction intervals. We may, however, use quantiles of the distribution to get estimated prediction intervals for the integral of the process. For example, for a 95% prediction interval, we may use the values in Table 5.1 to get an interval of (1.034 × 10 ^{9} ; 6.354 × 10^{10}). Note, however, that this interval still involves the aforementioned sampling error.

The variances for the integral estimates for the realizations, $\hat{\boldsymbol{\Phi}}_i$, were estimated as discussed in Section 5.2. The estimated standard deviations ranged from 5.645 × 10^7 to 4.602 × 10^{11}. However, for the lower 97.5% of the integral estimates, the largest estimated standard deviation was 1.708 × 10^{10} , nearly 30 times smaller than the overall maximum. This would lead us to suspect that the highest integral estimates for realizations come about due to one or two very high simulated values in that realization, thus increasing the variance of the sample from the realization tremendously. Further investigation is necessary to determine if this is in fact the case.

The variance of the integral is estimated here as discussed in Section 5.2. That is

$$Var[\boldsymbol{\Phi}|\mathbf{x}] = Var\left[\hat{\boldsymbol{\Phi}}|\mathbf{x}\right] - E\left[Var\left[\hat{\boldsymbol{\Phi}}|\boldsymbol{X}(\cdot)\right]|\mathbf{x}\right] \tag{5.29}$$

where $Var\left[\hat{\boldsymbol{\Phi}}|\mathbf{x}\right]$ may be estimated by the sample variance of the integral estimates from the realizations and an unbiased estimate of $E\left[Var\left[\hat{\boldsymbol{\Phi}}|\boldsymbol{X}(\cdot)\right]|\mathbf{data}\right]$ is obtained by taking the sample mean of the variances of the $\hat{\boldsymbol{\Phi}}_i$ calculated from the 1000 realizations as discussed above. In this example, this yields

$$Var[\boldsymbol{\Phi}|\mathbf{x}] = 2.423 \times 10^{21} - 2.796 \times 10^{20} \cong 2.143 \times 10^{21} \tag{5.30}$$

So the estimated standard deviation of the integral is $\sqrt{Var[\boldsymbol{\Phi}|\mathbf{x}]} = 4.629 \times 10^{24}$. It is of concern here that this value is actually *larger* than the estimated mean of the integral of 1.204 × 10^{10} . This is caused by the large variation between the integral estimates for the different realizations, and the high skewness of the distribution.

Chapter 6

Phoenix, AZ: Visualization with a Time Component

In this chapter, we use methods of kriging and related spatial estimation (Cressie,1993) to study various concentration plumes at a site near Phoenix, Arizona. Since we have data from several years for this site, we are able to incorporate time into our analysis. In all cases, the plots were intended to represent a yearly average, so *all* of the data from each year was used. In cases where there was more than one measurement for a particular location in a year, the average of these values was used. In Section 6.1, we do an exploratory visualization of contaminant levels, along with accuracy assessments in the form of prediction standard errors. In Section 6.2, we present portions of animations of contaminants together, including a method for animation of two possibly related contaminants.

6.1 Exploratory Visualization

Figures 6.1, 6.3, 6.5, and 6.7 show "prediction" surfaces for TCE and DCE concentrations for two selected years. The predictions were done using the log of the data, so these surfaces are on a log scale. Figures 6.2, 6.4, 6.6, and 6.8 show the corresponding standard error surfaces. The x and y coordinates are the same for all of the plots in this section and the next, with the z coordinates varying slightly. The legends for the z coordinates are given in Figures 6.9 and 6.10 for the prediction surfaces and error surfaces, respectively.

Figures 6.9 and 6.10 indicate the color codes for different levels of concentrations and standard errors. (Note that the legend for the prediction surfaces in Figure 6.9 refers to the level of the original data, rather than the log of the data.) Predicted values tend to be most accurate in the neighborhood of wells where the data were taken, which explains the downward spikes in the standard error surfaces. (That is, each small downward spike represents an observation well location.)

One sees, when comparing TCE concentration maps (Figures 6.1 and 6.5 for the years 1991 and 1992, respectively) with the corresponding DCE concentrations (Figures 6.3 and 6.7), that both substances are highest in the northeast portion of the region, and the TCE plume drops off more quickly than the DCE plume. This suggests scavenging of TCE to create DCE over time as the plume is transported from northeast to southwest. Thus, a map of the ratio of concentration of DCE to TCE should show a large increase moving down the plume, but the sum of the two concentrations might be relatively constant, assuming the region is relatively "closed." Of course, conclusions drawn from such displays would have to be tempered by the accuracy of the estimated quantities.

6.2 Animation

6.2.1 Trichloroethylene and Dichloroethylene (TCE and DCE)

Since it is suspected that the TCE and DCE plumes are somehow interrelated, it is desirable to plot the two together in such a way as to make the relationship (if it exists) more visible. With this in mind, animation of the TCE and DCE data is accomplished by combining the estimated concentration plots for each on the same plot. This allows the two substances to be easily animated together in time, which in turn should facilitate our attempts to see relationships between the levels of the two substances through space and time. The animation has been performed using the measured data from the years 1985 to 1993, as discussed in Section 2.2.

Spatial estimation, via kriging, was then performed on the *logs* (base 10) of the *yearly* data. (See Appendix B for a brief discussion of the spatial modeling and an explanation of the problems there encountered.) For the animation, interpolations were done in between the years (5 time slices between each pair of years), allowing a smoother progression which aids greatly in seeing general trends in the data.

Examples of the plots used in the animation are shown in Figures 6.11, 6.12, 6.13, and 6.14, representing the years 1986, 1988, 1991, and 1993, respectively. The TCE surface is the upper one in all of the plots and the DCE the lower. The size of the region is approximately 18000 *ft* in the E-W direction and 8300 *ft* in the N-S direction. All plots are shown on the $\log_{10}$ scale; e.g., the purple regions which are labeled as 1 to 2 on the scale are regions where the estimated contaminant level is between 10 and 100 $\mu g/L$.

These plots all confirm our assumption that the values of the contaminants are generally highest in the Northeast region, where the source is located. These plots help to point out some general features in the data, as well as introducing some new questions. In particular, Figure 6.13 would seem to indicate that there is a ridge of high values of contaminant along a line from East-Northeast to West-Southwest, along with some low values next to this region of high values. Investigation into the site properties yielded the information that the ground-water flow is generally from E-NE to W-SW, so it would appear that the contaminant levels are highest along a direct line from the source in the direction of ground-water flow, perhaps indicating very slow dispersion of the contaminants through other means. The 1986 plot, Figure 6.11, is representative of the earlier years in that data were only taken in an area relatively close to the source. In later years, gradually points were added farther to the west of the source. (There were none added to the east, presumably since it was unexpected that the contaminant would spread greatly in a direction counter to the direction of ground-water flow.) By 1993, both contaminants appear to have dissipated greatly, and many of the mild variations seen are likely due to measurement error. Figure 6.12, from 1988, shows lower values of

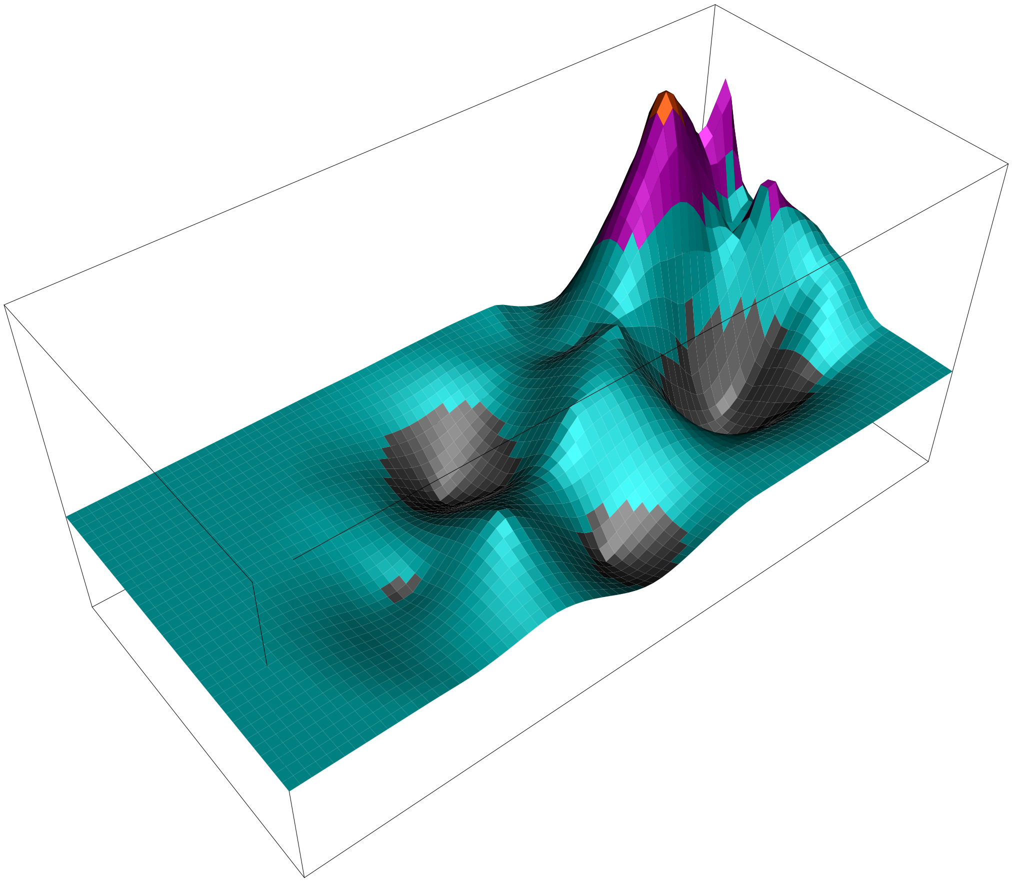

Figure 6.1 TCE prediction surface for 1991. Orientation: depth-axis: 892,600 to 896,800; width-axis: 478,000 to 484,000; vertical-axis: 0 to 2.25.

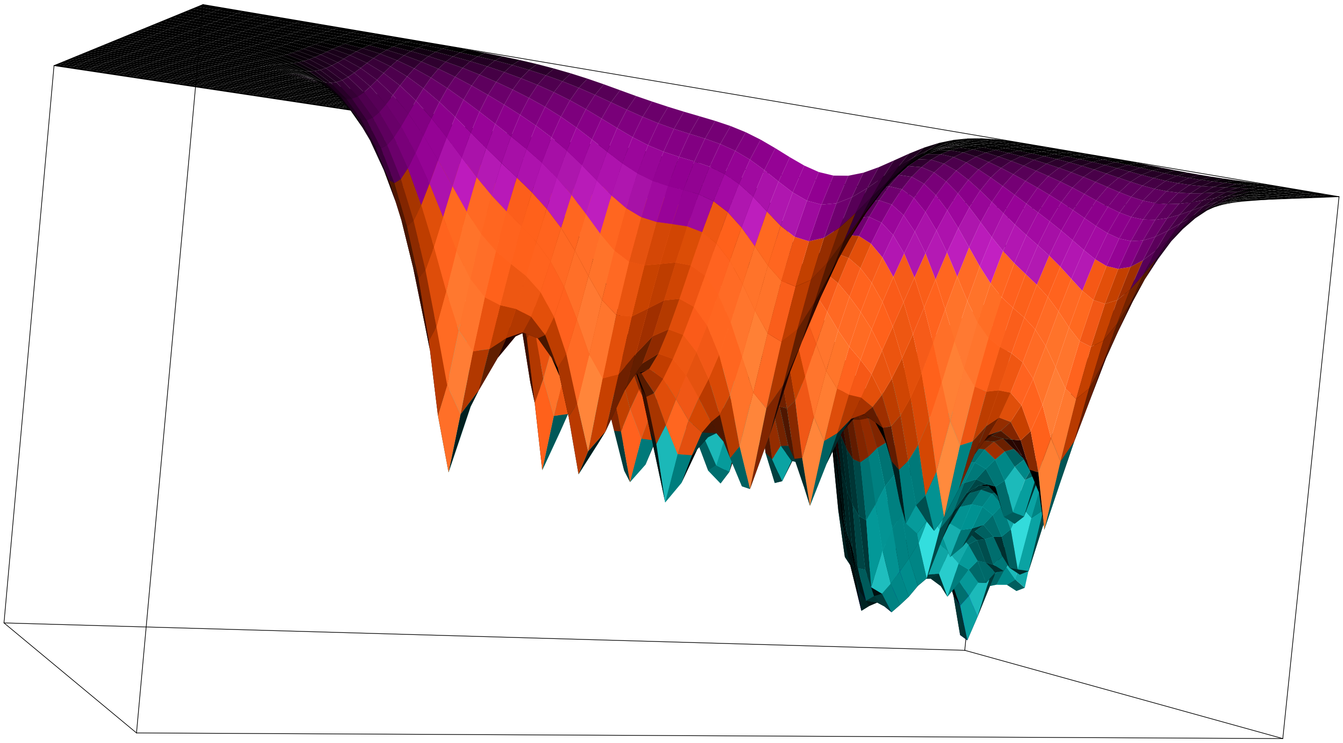

Figure 6.2 TCE standard errors of prediction for 1991.

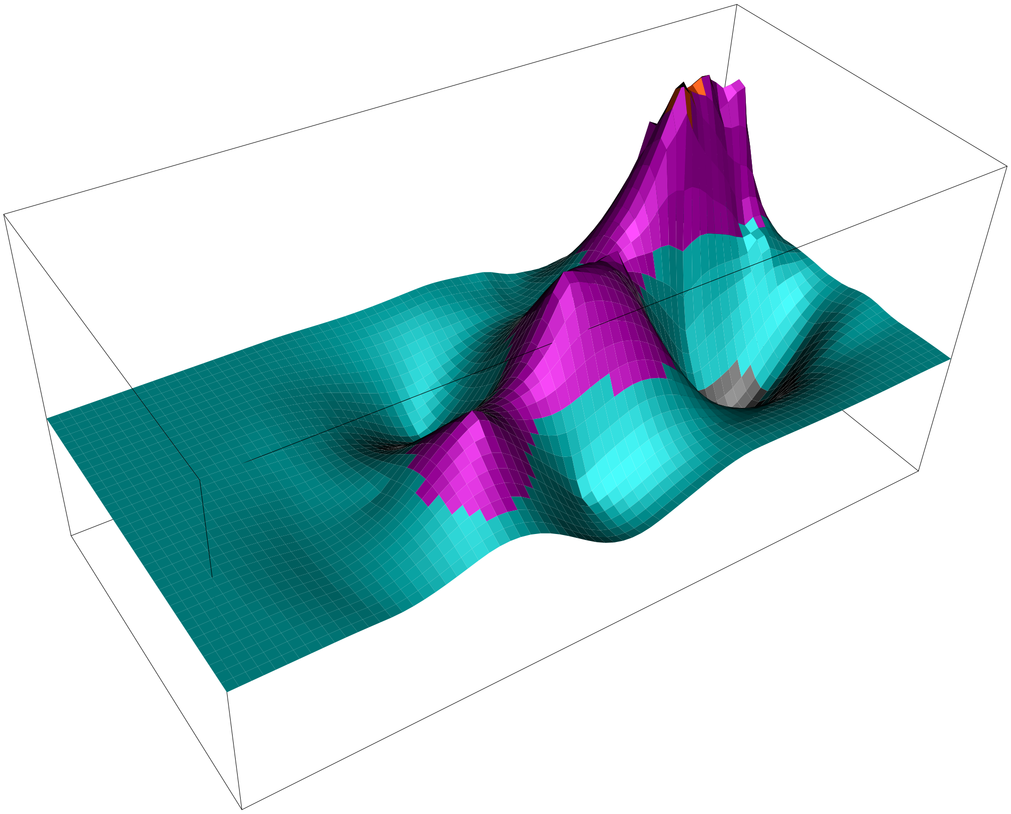

Figure 6.3 DCE prediction surface for 1991.

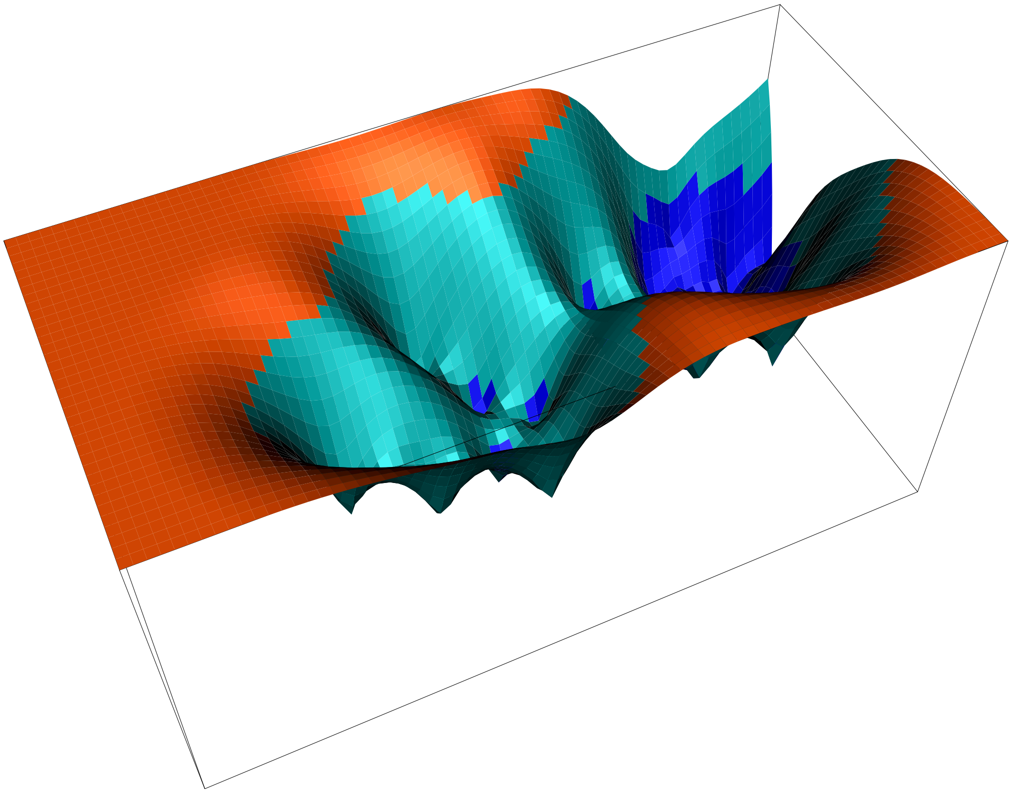

Figure 6.4 DCE standard errors of prediction for 1991.

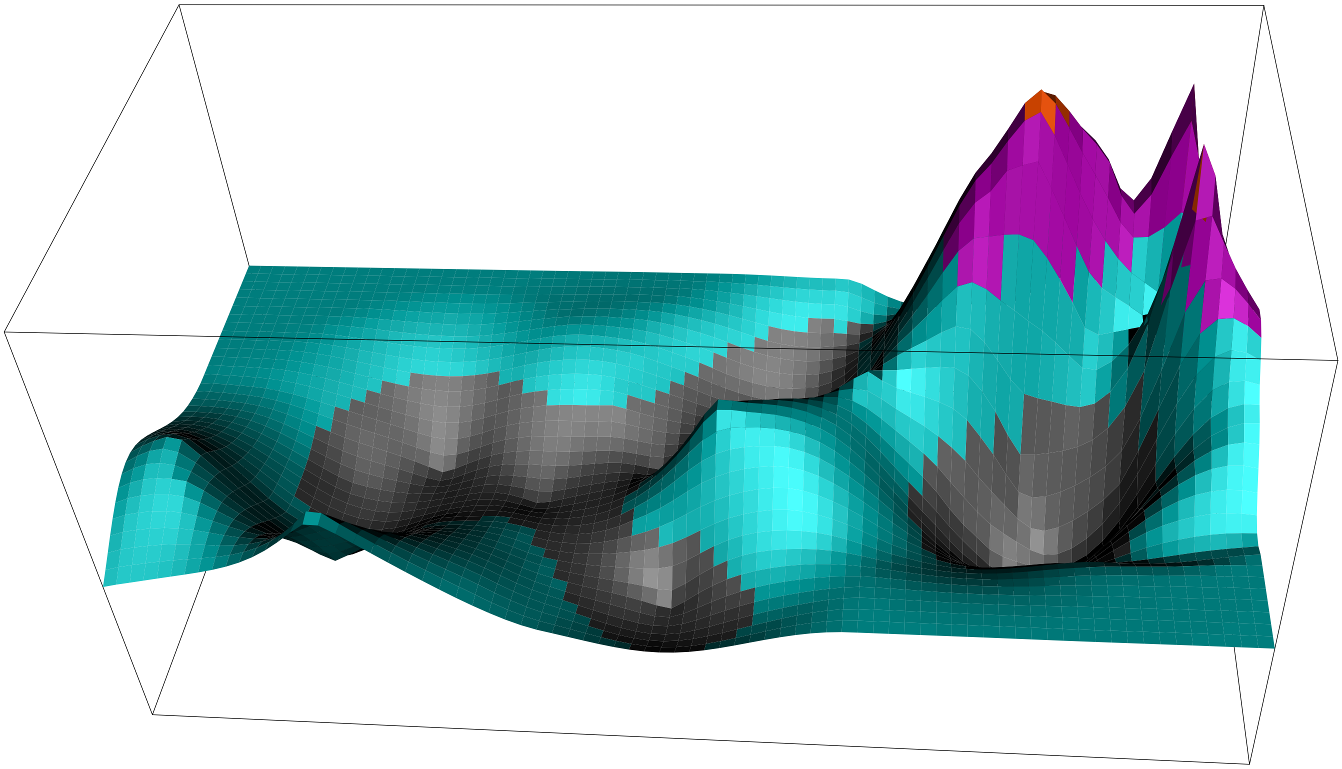

Figure 6.5 TCE prediction surface for 1992.

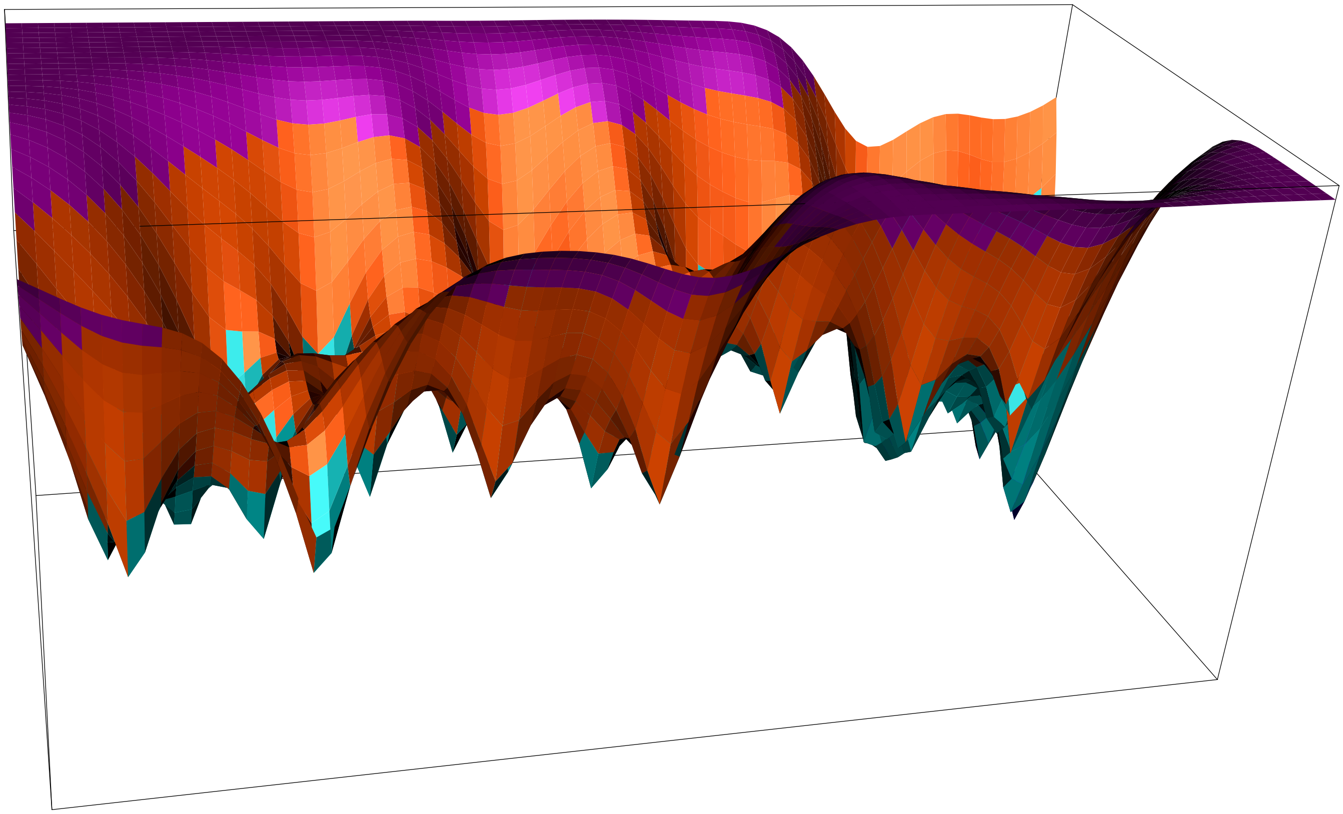

Figure 6.6 TCE standard errors of prediction for 1992.

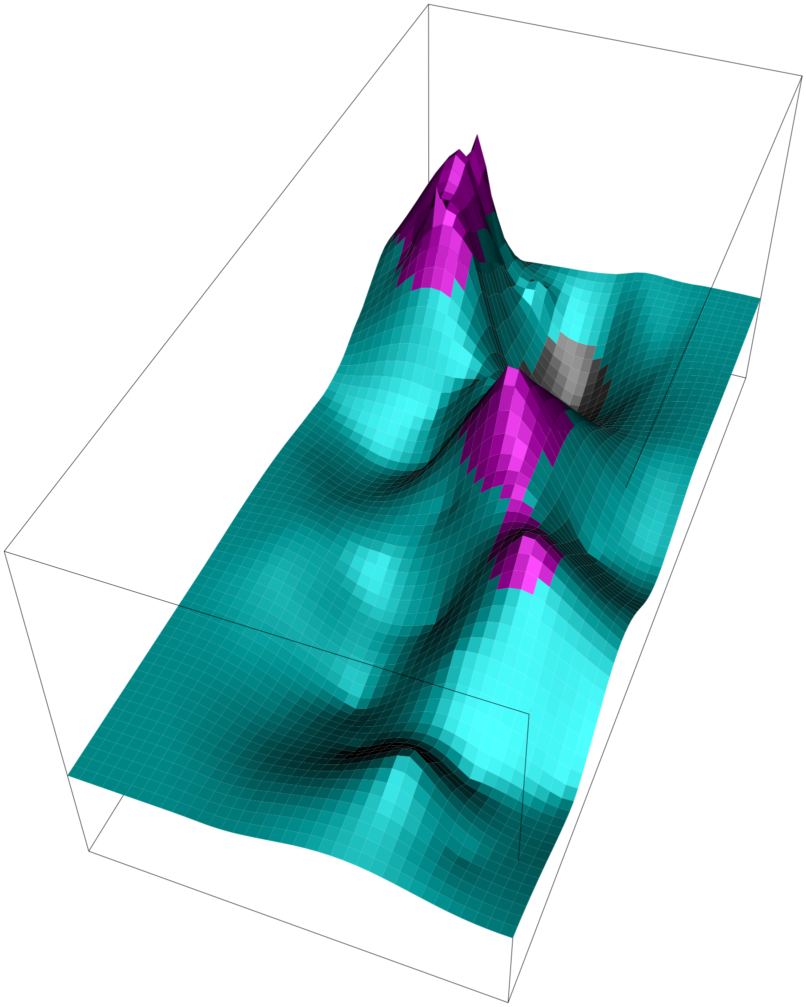

Figure 6.7 DCE prediction surface for 1992.

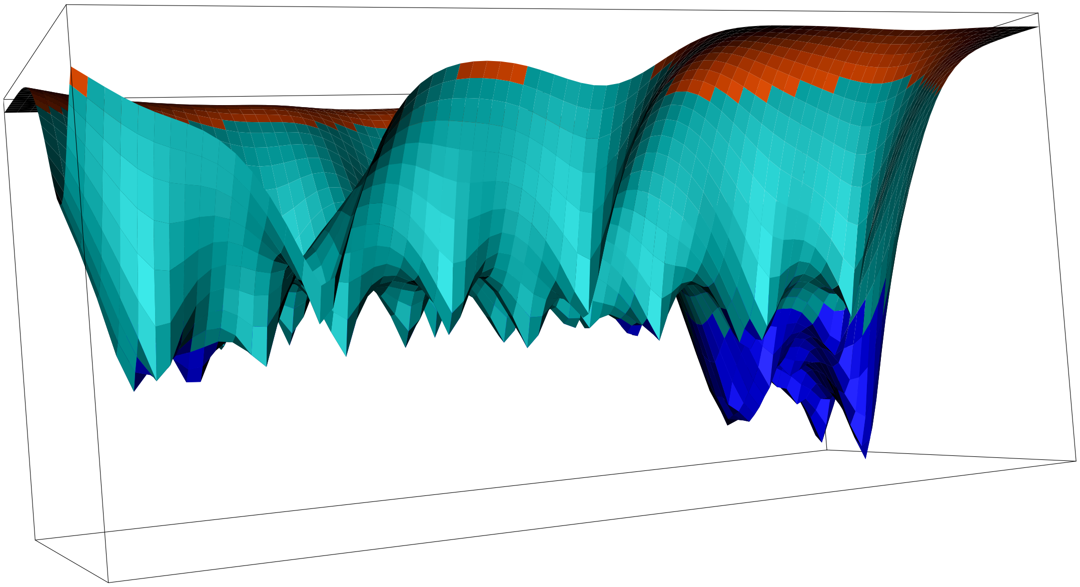

Figure 6.8 DCE standard errors of prediction for 1992.

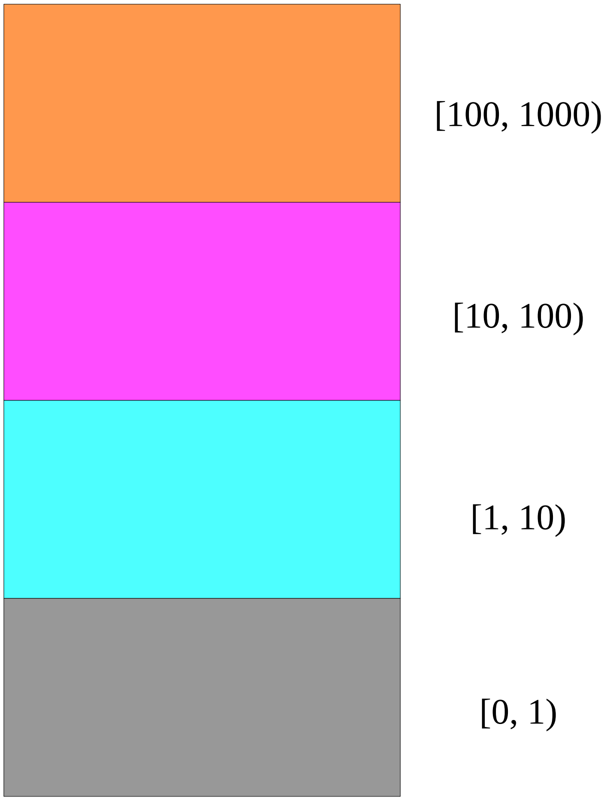

Figure 6.9 Legend for prediction surfaces (measurements in *mg/L*).

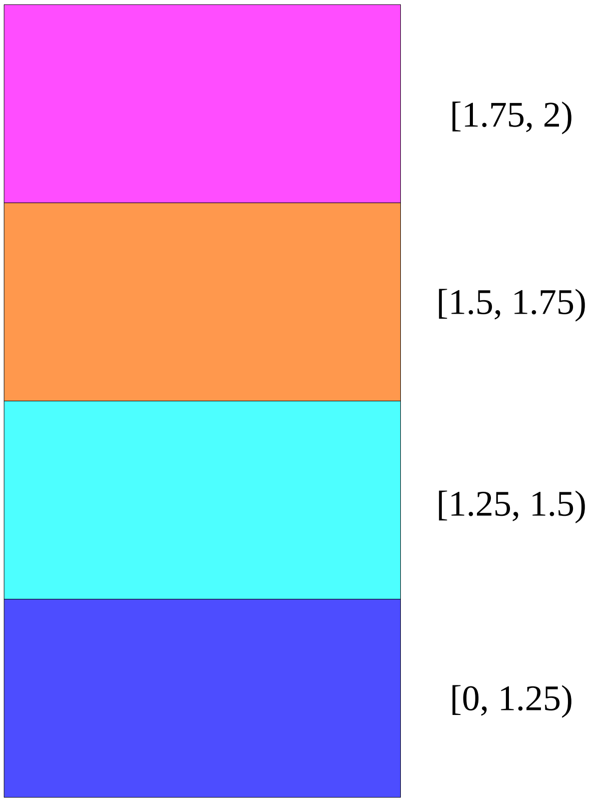

Figure 6.10 Legend for standard errors.

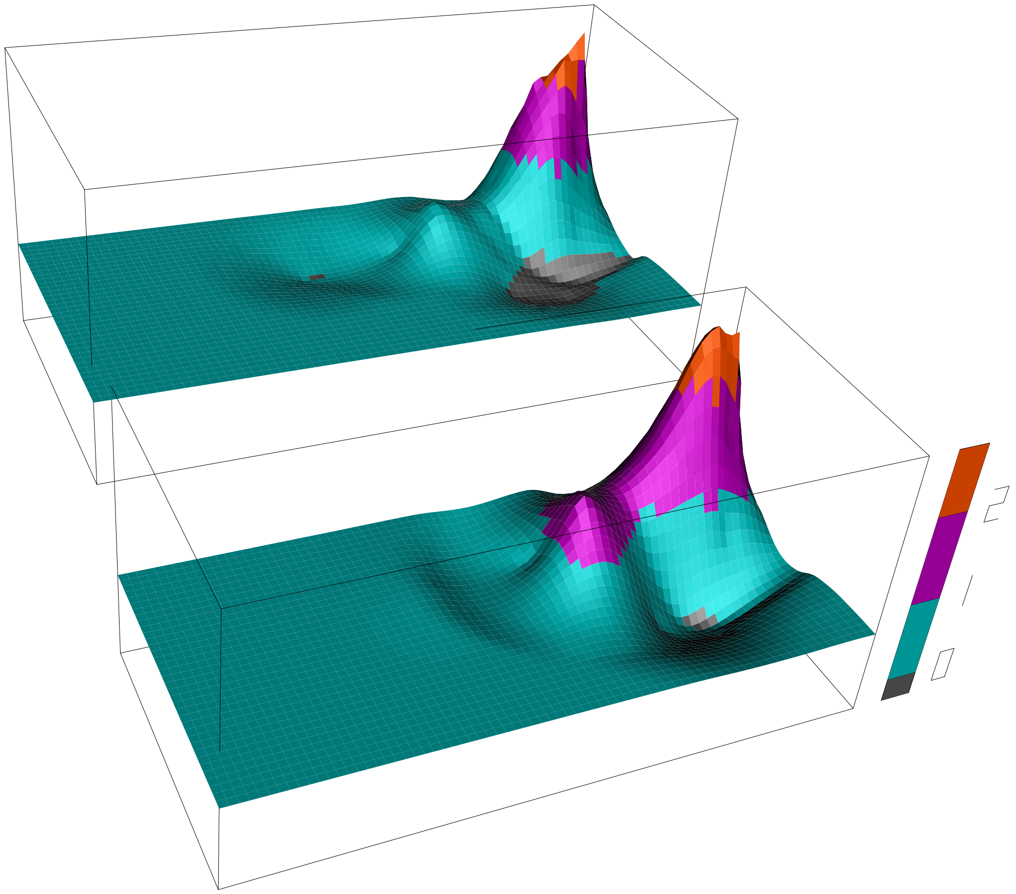

Figure 6.11 TCE and DCE surfaces for 1986.

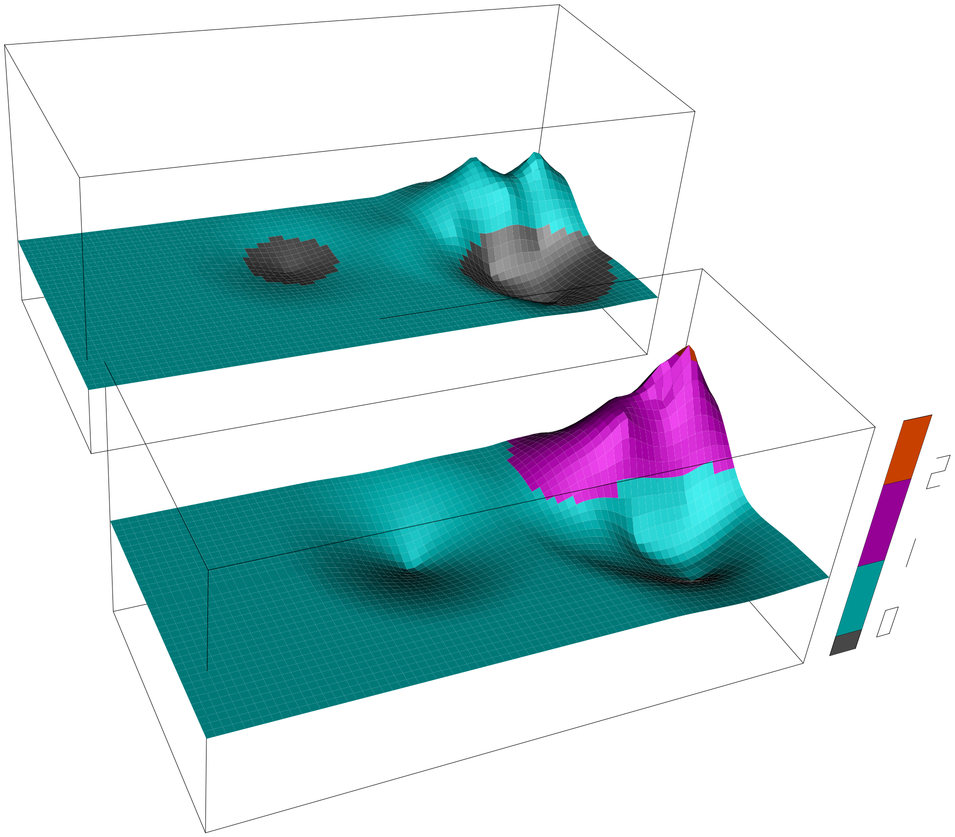

Figure 6.12 TCE and DCE surfaces for 1988.

this occurred. It has been supposed that perhaps an excessive amount of rainfall in this year would have diluted the concentrations of contaminant in the ground water, but it should be noted that this is definitely only suspicion and further investigation is necessary. In any event, however, animation of the spatial estimates through time allowed us to easily pick up on this seeming aberration.

6.2.2 Sulfate Ions

Measurements of sulfate ions (SO_4^{2-}) are of interest due to a supposed connection between levels of SO_4^{2-} , TCE, and DCE. That is, anaerobic bacteria which consume SO_4^{2-} also consume TCE, converting it into DCE. It is believed that regions of low sulfate, dubbed “sulfate holes,” indicate the presence of such bacteria. If this is true, we also would expect to see TCE drop in these regions. There are much fewer sulfate data than there are TCE and DCE data, to the extent that there are insufficient data in several of the years to do a reasonable surface estimate. For this reason, SO_4^{2-} surfaces were produced only for the years 1985 through 1989, inclusive, and 1991. As there was some concern that the large flat sections of the surfaces in the portions of the region with no data might be somewhat misleading, a new visualization technique was tried with the SO_4^{2-} data. For each year, the surface was only plotted in the area where there *were* data, with a bounding box to indicate the region and keep all years on the same scale. The success of this approach is perhaps mixed. It does indeed make it very clear in what regions we do not have any good estimates due to lack of data. However, it also makes it nearly impossible to interpolate between years to produce a smooth animation, and in fact increases the “jumpy” effect seen when viewing the surfaces in chronological order. (This is not such a problem with the sulfate data, as the time gap makes them not entirely suitable for animation, anyway.) Again, the plots are all on a $\log_{10}$ scale with the same color contours as for TCE and DCE.

Figures 6.15, 6.16, and 6.17 contain examples of these SO_4^{2-} surfaces. These surfaces have several interesting features. In all three, we see a small peak in sulfate levels near the source of TCE and DCE contamination. Also, in all three, we see one or more “sulfate holes” near to this peak. This indicates that perhaps TCE is being converted to DCE near to the source, which would lead to TCE values dropping off more rapidly than DCE as distance from the source increases. Efforts to see if such a relationship exists will be discussed further in Section 6.3. Figure 6.16 reveals another interesting and somewhat odd feature of

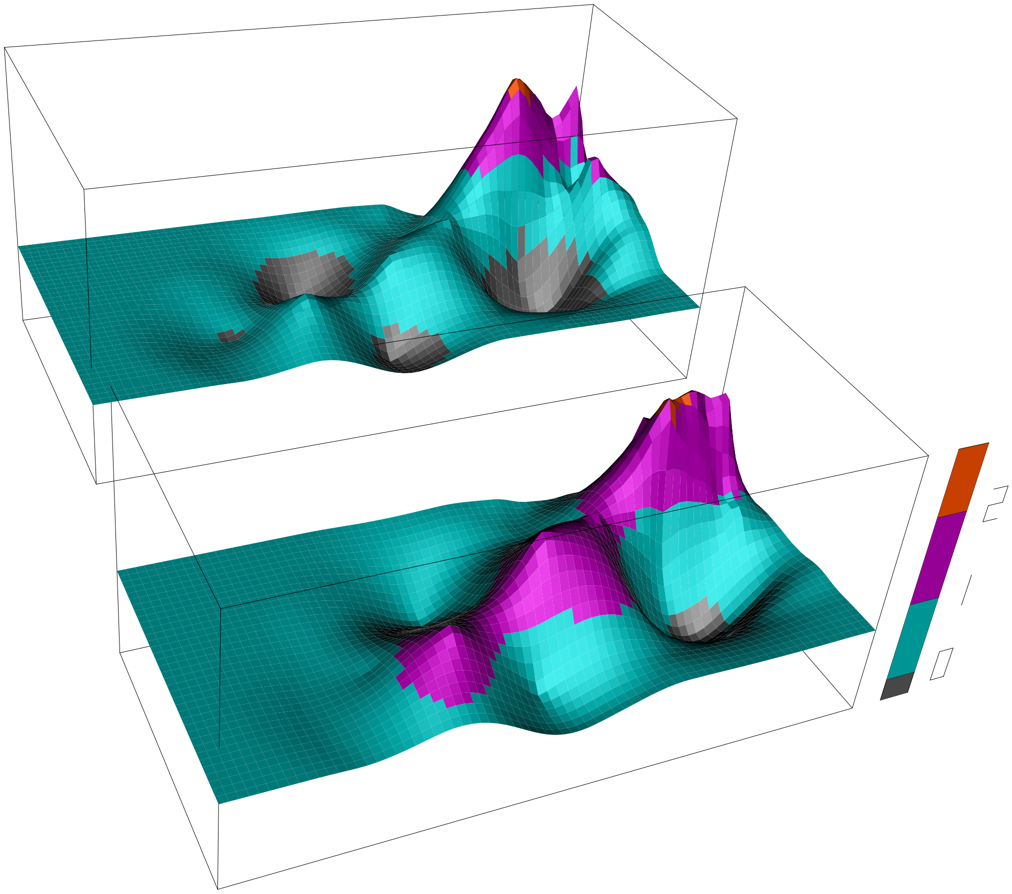

Figure 6.13 TCE and DCE surfaces for 1991.

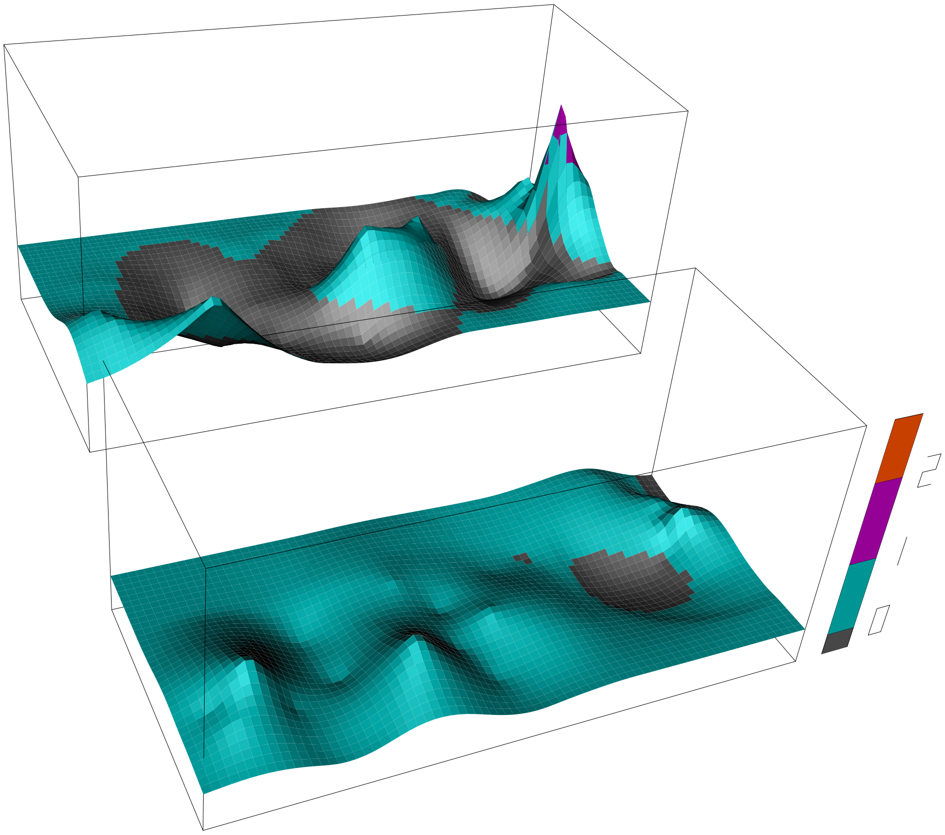

Figure 6.14 TCE and DCE surfaces for 1993.

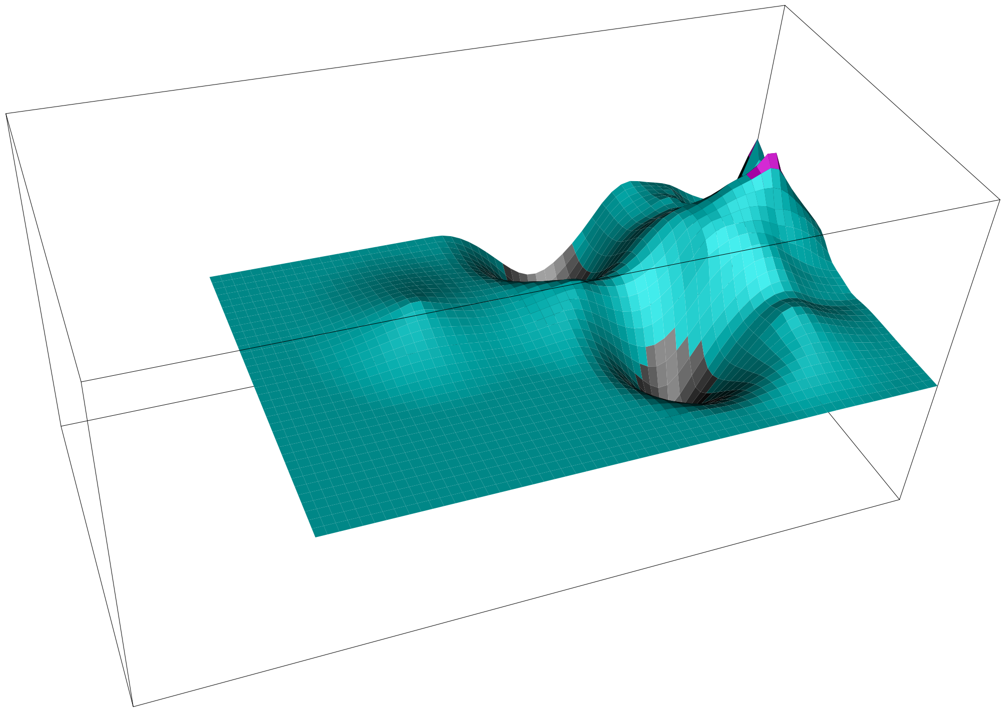

Figure 6.15 SO_4^{2-} surface for 1986.

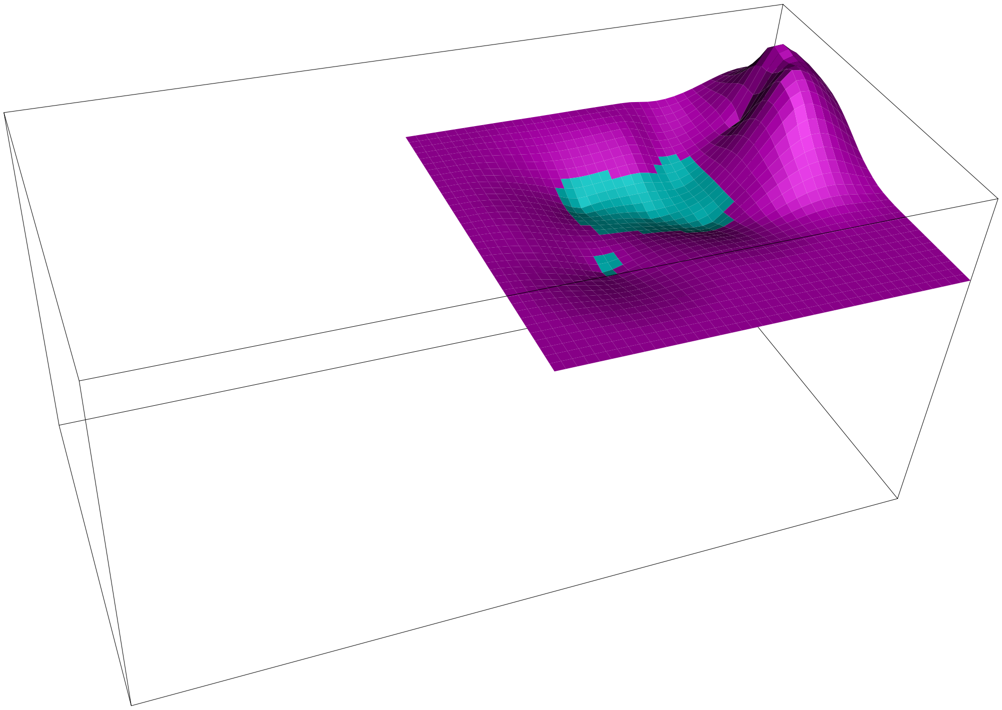

Figure 6.16 SO_4^{2-} surface for 1988.

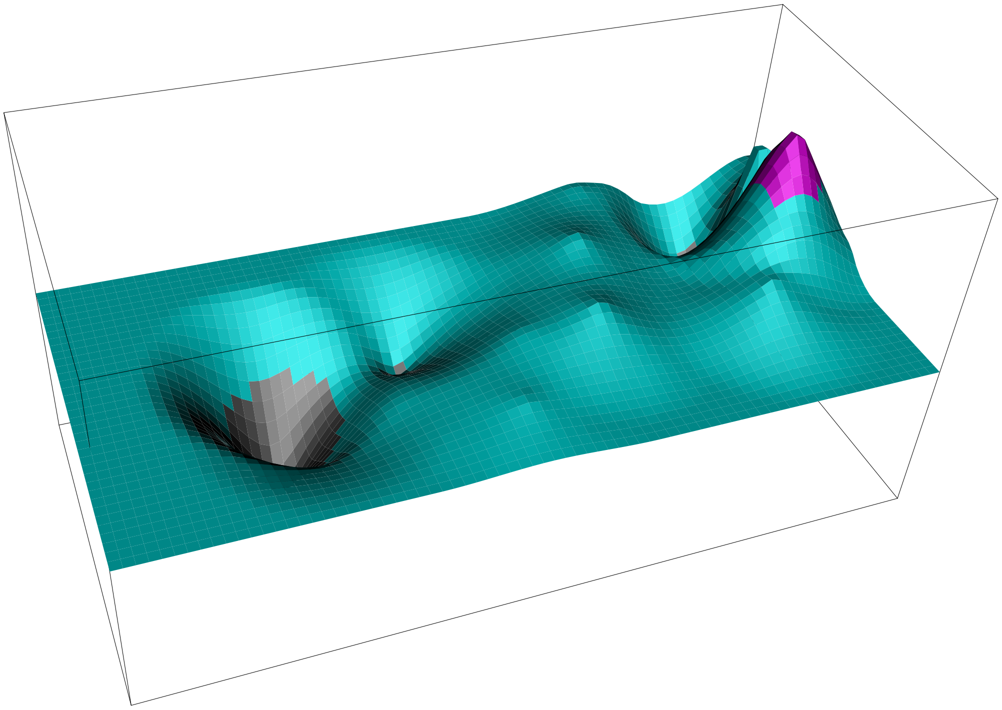

Figure 6.17 SO_4^{2-}surface for 1991.

the sulfate data. The overall sulfate levels for the years 1987, 1988, and 1989 over the region are significantly higher than for the other years studied. We have not yet been able to identify a reason for this, but it certainly seems to warrant further investigation.

6.3 Further Analytical Efforts

As mentioned previously, the region for which there were data available increased as time progressed. Specifically, in the earlier years (e.g., 1985), there were data points only relatively close to the source. Later, more data points were added to the south and especially to the west. This can create some problems with analysis, so it was decided following a conversation with Dr. Joe Hughes that we should try to do local analyses using some smaller region around the source. It is hoped that this will help us to better understand the chemical and transport processes at work. We expect that

- TCE should decrease in regions of high sulfate due to the likely presence of sulfate- consuming bacteria in these regions, and that
- TCE and SO_4^{2-} should decrease quicker with distance from the source than DCE.

Two different types of local analyses were attempted. The first involves simply using only a small rectangular region around the source; e.g., the Northeast corner of the total region. The other involves using a relatively narrow region from the source and extending in the direction of flow. For this second method, we then want to do spatial estimates of contaminant (i.e., $\log_{10}$ of TCE, DCE, and SO_4^{2-}) vs. the following:

1. distance from the source,
2. time since Jan. 1, 1985 and distance from the source, and
3. time since Jan. 1, 1985 and distance from the source at time 0.

As the region chosen in this case is narrow, the distance from the source is approximately equal to the distance from the source along the line of flow. This method seems reasonable because the primary method of transport of contaminant in this system is ground-water flow, with dispersion being decidedly less. So if the speed of ground-water flow is ***s*** *ft/day*, then the contaminant present a distance *st* from the source in the direction of flow on day *t* should have been at the source on day 0. This is precisely the motivation for the third case listed above, for which we calculate $d' = d - st$, the estimated distance from the source at time 0.

Unfortunately, there is one problem with this latter method which has not yet been resolved: it is not entirely clear where the source of contaminant *is*. The general vicinity is certainly known, but to pinpoint an “exact” location allowing us to place a narrow strip about the source has proved to be a difficult problem. It is possible that the source is in fact a large area and would not be well approximated by a point source. From recent investigations, what appears likely is that there are at least 2 point sources approximately 500 - 1000 *ft* apart. (We have hypothesized this after noticing that in the general vicinity of the source, there are two clusters of measurement locations which both contain very high levels of TCE. As it is common to place large numbers of wells near a known source, the presence of two sources known to previous investigators seems a logical hypothesis.) This presents some unforeseen problems and is what has led us to consider the first method mentioned; i.e., doing more standard estimation on a small region about the source.

Chapter 7

Summary and Conclusions

Site characterization and estimation of contaminant plumes is a complex problem which requires the compilation by the environmental researcher of many sources of information. Observations on the contaminant level over the region are expensive and sometimes difficult to obtain. In this research effort, we suggested several methods of examining such valuable data to further the researcher's understanding of the environmental problem under study.

Based on observational data, we explored analytical methods for estimating the level and extent of the contaminant plume. Nonparametric regression methods proved useful for quick summaries of the contaminant plume, whereas the more difficult to implement geostatistical methods were required for quantitative measures of the contaminant plume, such as the total amount of contaminant present.

In addition to exploring the analytical issues associated with estimation of the contaminant plume and functionals of this plume, we investigated how best to display this information through visualization methods. Two and three-dimensional perspective plots with color contours proved useful in our investigation. To associate the error in the estimated plume with the estimated level of the plume, we suggest associating the height of the perspective plot with the estimated level of contaminant and the color contours with the estimated amount of observed error in the estimate.

Our investigations also found that animation of the estimated level of contaminants or estimated errors was a useful exploratory tool. For the Eglin data, surface estimates produced with all but one point are animated alternately with the surface estimate using all the data points. This allows us to readily see the effect each data point has on the surface estimate. For points whose absence produces a large change in the surface estimate, it may be desirable to take additional samples near this point to help stabilize the estimate in this area; estimates of the prediction errors at these points are also useful for this reason. Animations through time, with smoothing, were used for the Arizona data allowing quick identification of atypical behavior in time. Also for the Arizona site, we investigated methods of simultaneously animating two related substances. Simultaneous animations of TCE and DCE helped identify the relationship between these two substances. Furthermore, we examined the issue of a growing region or plume. Both of these issue are the focus of further research.

We proposed in Section 5 a method for estimating the integral of a random process in the case where the process is lognormal by modeling the process through geostatistical methods and simulating the process conditional on the data. This is useful for estimating the total amount of contaminant present in a region. When implemented on the Eglin ground water observations for BTEX, this method produced reasonable point estimates but large confidence intervals. Large confidence intervals are to be expected from such a small number of observations, however we are hopeful that further research into improved statistical methods can yield tighter confidence intervals for small sample sizes.

Future Research

This work has surfaced a number of topics which would be appropriate for future research. It is still an open question how to best view two possibly related substances to see how they are related. For the TCE and DCE data from Arizona, it was thought that examining the ratio of the two substances might be useful, but this did not seem to reveal very much. An examination of functions of two such substances so they may be viewed as a single surface seems like a promising idea, however.

For the sulfate data from Arizona, we attempted to deal with the issue of an increasing design region over time, due to more information, and interest, on the part of those taking observations. Attempts to plot various portions of the estimated surface, dependent on the region, seem to be of dubious value. The idea of estimating a small region near the source, or perhaps estimating along the line of ground-water flow from the source, as discussed in Section 6.3, is a promising idea. The lack of a well-defined source in this instance made such an examination difficult, but such a method could certainly still be examined for this and other data.

The method for estimating the integral discussed in Section 5 is promising, but reductions of variance and better prediction intervals are areas which need to be addressed. Specifically, if one looks at the values obtained for the integral estimates and their standard deviations, one notices that in fact the standard deviations are *extremely* high. In particular, recall that if we use the mean of the integral estimates for the realizations as a point estimate for the integral of the process , we get an estimate of total contaminant of 1.204×10^{10} . However, the estimated variance of this value is then the mean of the variances for the realizations, giving a variance of 2.796×10^{20} , and a standard deviation of 1.672×10^{10} . That is, the estimated standard deviation of total concentration is actually *higher* than the estimated concentration. It is possible that by sampling locations using some distribution other than a uniform, i.e. *importance sampling* (Rubinstein, 1981), may yield integral estimates with a smaller variance. For example, we may wish to sample from a smaller, possibly non-rectangular region where the data are more dense. Or we may wish to sample with higher probability along the direction of geometric anisotropy than in the perpendicular direction. At any rate, it is desirable to investigate ways of reducing the variance of the integral estimates, and this is a topic of current research for some of the authors of this report.

For the prediction intervals, we used what are referred to as *equal tail* intervals. That is, the interval is a two-tailed interval with equal probability in either of the tails. This is not always the best type of interval to use, and particularly may not be in the case of such an asymmetric distribution. It would be worthwhile to investigate other types of intervals, particularly those known as *highest posterior density* (HPD) regions (Casella and Berger, 1990). In this case, the 1- α interval is chosen so as to be as short as possible. Specifically, if the posterior density is denoted by π, the 1- α HPD region is given by $\{x: \pi(x) \geq c\}$ where *c* is such that $1 - \alpha = \int_{\{x: \pi(x) \geq c\}} \pi(x)\, dx$. Also, the prediction intervals discussed here are actually *too large*, as they contain additional variation due to Monte Carlo sampling error. To get better prediction intervals, this factor needs to be corrected for.

Finally, it was mentioned at the beginning of Section 5 that there are actually *several* nonlinear functionals of random processes which are of interest to estimate. The integral of the process is the only one of these which we have investigated in detail to this point. Other functions which are of interest are:

- the maximum concentration attained within a region,
- the location where this maximum concentration occurs, and
- the region for which the concentration exceeds some set value.

These other three are quantities which are of interest for various types of environmental contamination and these warrant further investigation. Further, in ozone modeling, it is common to use a square root transform rather than a log transform as in Carroll et al.,1997. Thus it is also of interest to estimate the total contaminant in the case where the process is transformed by a square root rather than a log.

Appendix A
Cross-Validation

In all cases, it is assumed that bandwidths in the *x* and *y* directions should be the same, i.e. $\underline{h} = (h_1, h_2)$ where h_1 is the bandwidth in the longitude and latitude directions and h_2 is the bandwidth in the vertical direction. We will use $\underline{h}$ to denote either h_1 or (h_1, h_2) depending on whether the estimate is in 2-D or 3-D, respectively. The bandwidths were chosen by minimizing over $\underline{h}$

$$SSE(h) = \sum_{i=1}^{n}\left(u_i - \hat{u}_{i,h}\right)^2$$

where *n* is the number of data points, u_i is the i^{th} observation, and $\hat{u}_{i,h}$ is the nonparametric regression estimate, based on bandwidth $\underline{h}$, of the value at the i^{th} data point obtained when this point is removed. (Note: $\hat{u}_{i,h} = \hat{f}\left(\underline{x_i}\right)$ from 4.1.)

A.1 Two-dimensional Data

For the ground-water data, a single cross-validation was performed on bandwidths varying from 100 to 250*ft* in increments of ten. As seen in Figure A.1, the bandwidth selected here is 120*ft.* See 4.1 for a plot of the estimated plume.

Similarly, a cross-validation was performed for two-dimensional soil data from Eglin at approximate depths of 7.0 and 7.6*ft* below the water table. In both of these cases, the minimization of *SSE*(*h*) was performed by a grid search from 100 to 250*ft* in increments of ten. The bandwidth chosen for the depth of 7.0*ft* is 170*ft*, while 230*ft* was chosen for the depth of 7.6*ft.* See 4.1 for a plot of the estimated plume at a depth of 7.0*ft.*

In addition to an estimate of the contaminant plume, we also visualize a smoothed estimate of the absolute error of this estimate as each data point is removed. A cross-validation of the absolute errors for the ground-water data resulted in a bandwidth of 130*ft*, which is close to the 120*ft* found for the plume estimate. The plots for visualizing the error are given in Section 4.1.

A.2 Three-dimensional Data

For the three-dimensional soil data, we needed to perform cross-validation in both the *x* and *y* directions and the *z* direction. (Here the *x* direction is longitude, the *y* direction is latitude, and the *z* direction is distance above or below the water table.) First we performed cross-validation in the *x* and *y* directions, allowing the bandwidth in the *z* direction to vary along with the bandwidth in the *x* and *y* directions. This yielded an *x* and *y* bandwidth of 24*ft.* Then we examined the *SSE*(*h*) using *x* and *y* bandwidths varying from 20 to 40*ft* (which includes the minimum of 24*ft*), and six different *z* bandwidths ranging from approximately 1.5 to approximately 3. This caused us to choose an *xy* bandwidth of approximately 24*ft* and a *z* bandwidth of approximately 1.8*ft.*

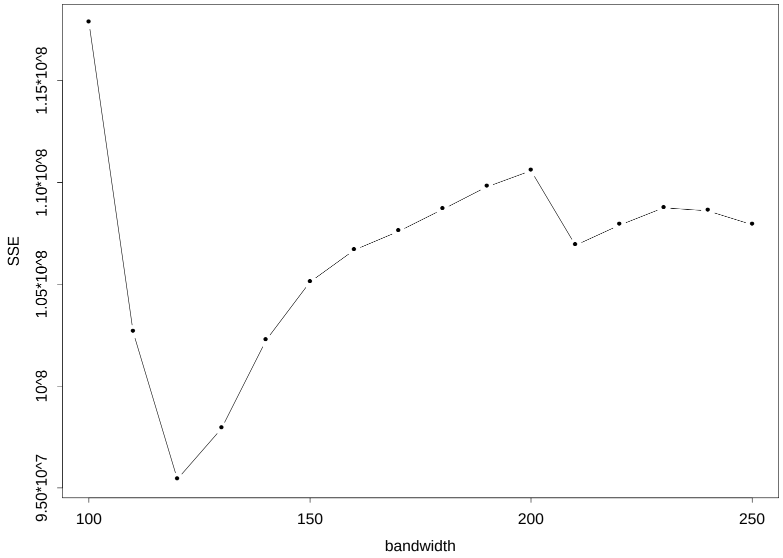

Figure A.1 Bandwidth selection for ground-water data.

Appendix B

Discussion of Spatial Estimation for Arizona

There were a number of difficulties encountered when doing the spatial estimates of TCE, DCE, and SO_4^{2-} for the Arizona site. The biggest problem was finding suitable variogram models. It was decided that it would be most reasonable to find only one variogram model for each substance, which would be used for all years of data. Many of the classical empirical variograms with standard default binwidths, etc., produced totally unreasonable variograms (e.g., variograms which were flat or indicated stronger correlation between points which were a long ways apart than for points which were close together). For example, consider the classical empirical variogram for 1990 TCE, shown in Figure B.1. This plot is highly variable and shows a general *decreasing* trend, whereas a variogram should be generally *increasing*. Cressie's robust variogram estimator did not usually solve these problems. It was conjectured that the high variability could largely be due to low numbers of pairs of data points at many of the higher distances. To combat this, we binned the data point pairs into groups with equal *numbers of pairs*, rather than equal width bins as in the classical estimator. We tried taking both means and medians within these groups, analogous to the classical and robust estimators. A plot of the mean case for the 1990 TCE data is shown in Figure B.2. Apart from the last bin, where the data point pairs used are so far apart in distance as to be suspicious anyway, this empirical variogram estimator looks much better than the classical one. Specifically, it is *much* less variable and has a decidedly increasing trend. Using this type of empirical variogram estimator, then, spherical variogram models were fit first for each year (and each substance). Then, from these, a variogram model for each substance was chosen which would be "best" in most years, and hopefully reasonable in all. This is intended to allow us to get satisfactory variogram estimates even for those years which do not contain much data and to provide a more unified approach.

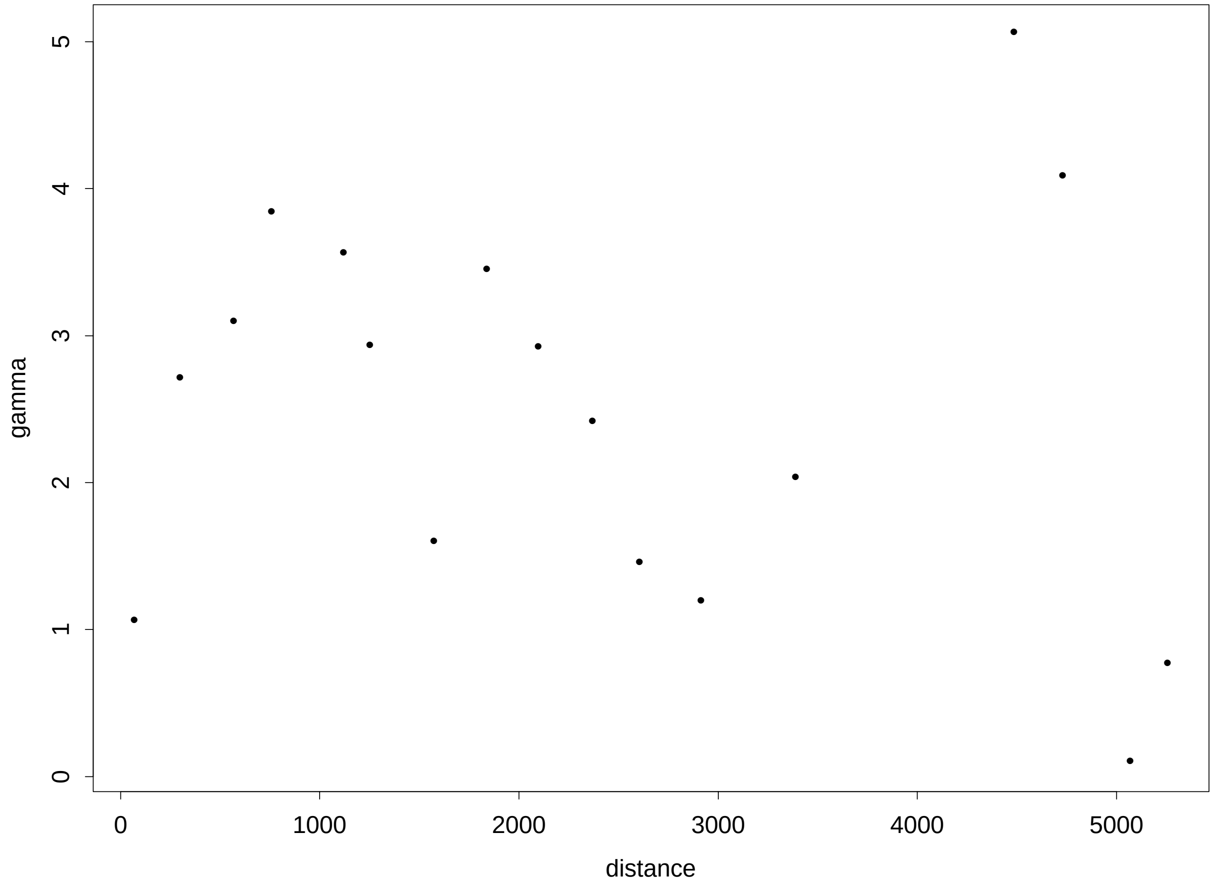

Figure B.1 Classical empirical variogram for 1990 TCE.

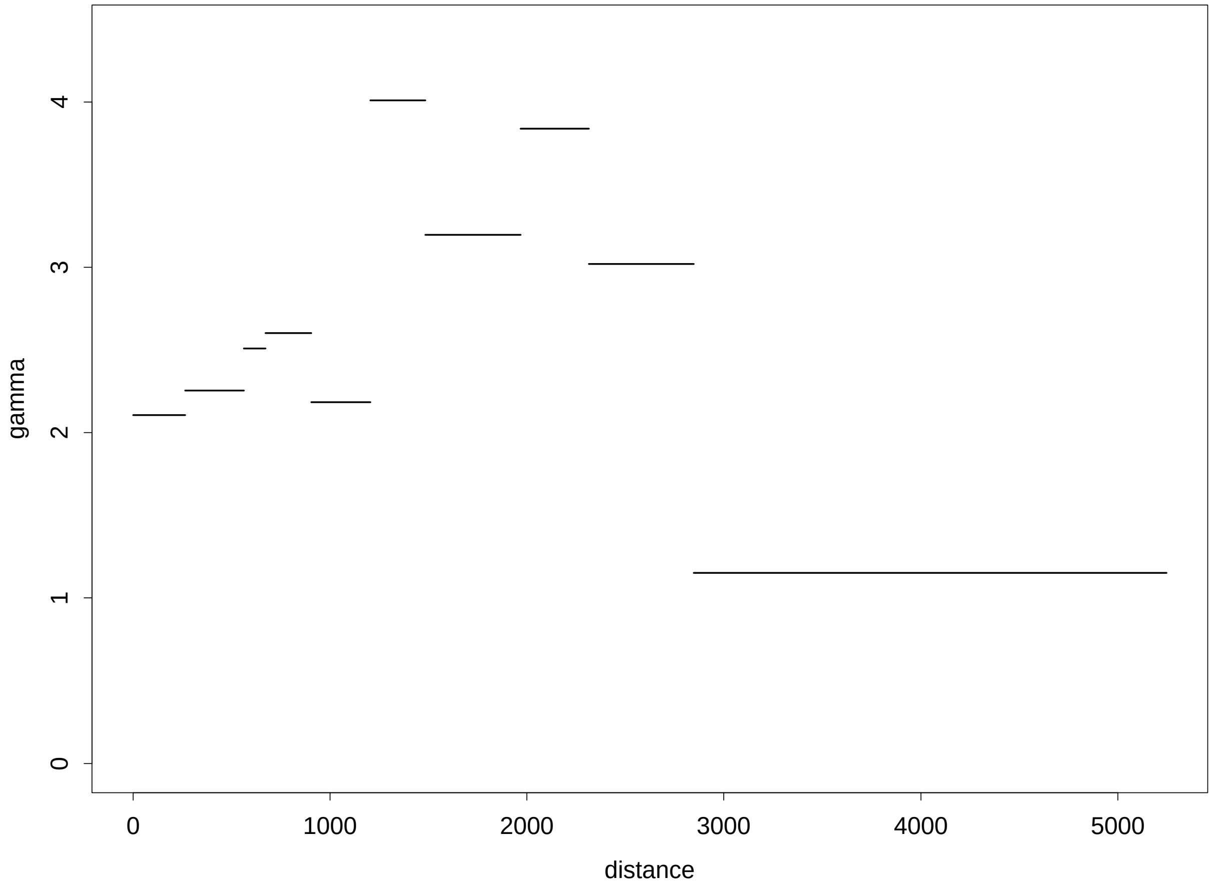

Figure B.2 Mean γ values for 1990 TCE.

Bibliography

E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen. LAPACK Users' Guide. SIAM, Philadelphia, 1992.

R.K. Boeckenhauer. Estimating nonlinear functionals of a random field. PhD. dissertation. Department of Science, Rice University, 2000.

R.K. Boeckenhauer, K.B. Ensor, D.W. Scott, and P.B. Bedient. Visualization of ground water contamination data. Proceedings of the 27th Interface Between Statistics and Computing Science, Editors Michael M. Meyer and James L. Rosenberger. Interface Foundation of North America, Fairfax, VA, pp 76-84, 1996.

R.J. Carroll, R. Chen, E.I. George, T.H. Li, H.J. Newton, H. Schmiediche, and N. Wang. Ozone exposure and population density in Harris County, Texas. Journal of the American Statistical Association, 92:392-415, 1997.

G. Casella and R.L. Berger. Statistical Inference. Duxbury Press, Belmont, CA, 1990.

D.D. Cox, L.H. Cox, and K.B. Ensor. Spatial Sampling for the Environment. Environmental and Ecological Statistics, Vol. 4, pp 219-233, 1996.

N.A.C. Cressie. Statistics for Spatial Data. John Wiley & Sons, New York, revised edition, 1993.

C.V. Deutsch and A.G. Journel. GSLIB: Geostatistical Software Library and User's Guide. Oxford University Press, 1993.

E.J. Englund and N. Heravi. Power curves for spatial sampling. Draft: 4/95.

J.M. Hammersley and D.C. Handscomb. Monte Carlo Methods. Methuen & Co. Ltd., London, 1964.

M.E. Johnson. Multivariate Statistical Simulation. John Wiley & Sons, New York, 1987.

E.L. Lehmann. Theory of Point Estimation. John Wiley & Sons, New York, 1983.

K.V. Mardia, J.T. Kent, and J.M. Bibby. Multivariate Analysis. Academic Press, Inc., San Diego, 1979.

MathSoft Inc. S-plus, 1988, 1995. Version 3.3 Release 1 for Sun SPARC, SunOS 4.1.x.

P.L. Reilly. Pmmlcg. IsoQuantic Technologies, LLC, 1994, 1995.

R.Y. Rubinstein. Simulation and the Monte Carlo Method. John Wiley & Sons, New York, 1981.

D.W. Scott. Multivariate Density Estimation. John Wiley & Sons, New York, 1992.

G.W. Stewart. Introduction to Matrix Computations. Academic Press, San Diego, 1973.

H.G. Sweed, P.B. Bedient, and S.R. Hutchins. Surface application system for in situ ground-water bioremediation: Site characterization and modeling. GroundWater, 34(2):211-222, March 1996.