NISS

Design and Analysis for Modeling and Predicting Spatial Contamination

Markus Abt, William J. Welch, and Jerome Sacks

Technical Report Number 82
March, 1998

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

Design and Analysis for Modeling and Predicting Spatial Contamination

Markus Abt[1], William J. Welch[1], and Jerome Sacks[2]

Sampling and prediction strategies relevant at the planning stage of the cleanup of environmental hazards are discussed. Sampling designs and models are compared using an extensive set of data on dioxin contamination at Piazza Road, Missouri. To meet the assumptions of the statistical model, such data are often transformed by taking logarithms. Predicted values may be required on the untransformed scale, however, and several predictors are also compared.

Fairly small designs turn out to be sufficient for model fitting and for predicting. For fitting, taking replicates ensures a positive measurement error variance and smooths the predictor. This is strongly advised for standard predictors. Alternatively, we propose a predictor linear in the untransformed data, with coefficients derived from a model fitted to the logarithms of the data. It performs well on the Piazza Road data, even with no replication.

KEY WORDS: Best linear unbiased prediction, dioxin contamination, Gaussian stochastic process, lognormal kriging, ordinary kriging, spatial statistics.

[1]Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada.

[2]National Institute of Statistical Sciences, Research Triangle Park, PO Box 14006, NC 27709-4006, U.S.A.

INTRODUCTION

Piazza Road is an Environmental Protection Agency (EPA) Superfund site located in Rosati, Missouri. In 1971 a waste oil mixture later found to be contaminated with dioxin was applied to a dirt road as a dust suppressant. For the subsequent cleanup, the site was divided into 150 rectangular exposure units (EUs) of approximately 5000 ft^2 each. The layout of some of these EUs was shown by Ryti, Neptune, and Groskinsky (1992, Figures 3–5). Prior to the cleanup, a pilot study with four EUs was carried out. The goal was to provide information on the large-scale and small-scale variability of the contaminant. A simulation study using these results showed that the size of the cost optimal cleanup unit was 14×14 ft^2 (Ryti, 1993). Based on this, the actual cleanup of the site was described by Ryti, Neptune, and Groskinsky (1992). Cleanup units were remediated, if necessary, in an EU to bring the EU average concentration below 1 ppb with a probability of at least 0.95.

As observations taken at nearby locations typically show higher correlation than observations taken further apart, our statistical models will have a very flexible class of correlation functions. Spatial correlations were not considered in the earlier works by Ryti (1993) or Ryti, Neptune, and Groskinsky (1992).

In this article we consider the problem of pointwise prediction of the dioxin concentration. This is relevant when very highly contaminated areas are to be identified. Furthermore, averaging a pointwise predictor can be a simple and computationally straightforward way to predict the average contamination over larger regions. We will return to this point.

Several diagnostics considered below suggest transforming the Piazza Road data by taking logarithms to satisfy the model assumptions. To obtain predictions on the original scale, however, this transformation needs to be reversed. We consider several methods.

We also investigate several issues in the design of the sampling scheme. Data are typically used in two ways. First, they appear in the predictor *implicitly* in the estimation of the covariance parameters in the statistical model, typically via maximum likelihood in this article. For model fitting, including nearby observations or even replicates could be

beneficial to estimate variability at different scales and is investigated. Secondly, data are used *explicitly*, as we focus on predictors that are (transformations of) linear combinations of the (transformed) data. The effect the number of observations used in the predictor has on prediction accuracy is also investigated.

For cost reasons, we might predict in an EU based on sparsely sampled data in that EU. Covariance parameters could be estimated, however, from extensive sampling in only a few representative EUs. We explore the impact on prediction accuracy from transferring the covariance-parameter estimates from one EU to another.

In the next section we discuss the Piazza Road data in greater detail. After demonstrating the need for taking logarithms of the dioxin concentrations, we consider various models and predictors of concentration on the original scale. The predictors and the design issues mentioned above are then investigated in a factorial experiment, which is analyzed by graphical methods. We also compare the results from variogram estimation of correlation properties and from some simpler approaches. Finally, we present some conclusions.

THE PIAZZA ROAD PILOT DATA

The Piazza Road pilot data are unusual because of the extensive sampling undertaken. Laboratory analysis was relatively inexpensive, enabling sampling at fine resolution. With such data we can select subsets and explore many alternative sampling strategies.

Figure 1 of Ryti (1993) shows the location of the 2×2 arrangement of the four EUs chosen for the pilot study. We use two of these EUs, named EU1 and EU2 here. In the data, dioxin concentrations at or below the detection limit of 0.3 ppb are reported as 0.3 ppb. In EU1 and in EU2 the proportion of such observations is less than 5%. Zirschky and others (1985) replaced the 0.3 ppb observations with 0.15 ppb. To be conservative from a risk perspective, we leave the 0.3 ppb observations unchanged when analyzing EU1 and EU2. EU3 and EU4 have a larger proportion of observations less than or equal to the detection limit (38% and 74%), and are therefore excluded from our analysis.

The four pilot EUs are 100 ft long and 50 ft wide. Two types of sample were taken in each. First, station markers were laid out on a regular 10 by 5 grid, spaced 10 ft apart. Then, 1 ft either side of each station marker, pairs of tablespoon samples were taken 1/2 in apart. This gives $10 \times 5 \times 2 \times 2 = 200$ observations in each EU. We will refer to these as the *grid* samples. Figure 1 shows the locations of the grid samples in EU1. Secondly, samples were obtained at 50 randomly selected locations in each pilot EU. These are also shown in Figure 1 for EU1. At each random location, nine tablespoon samples were taken within a 1 ft sampling frame on a 4 in square grid. These nine samples were mixed and three separate aliquots were selected for analysis. The resulting $50 \times 3 = 150$ observations per EU will be called *random* samples below. (Fig. 1)

Table 1 summarizes some characteristics of the data. EU1 shows a higher average concentration of the contaminant along with higher variability. (Table 1)

DATA TRANSFORMATIONS

The dependence of the variability in concentration on the mean suggested by Table 1 is typically not accounted for in simple linear statistical models or in those we will use below. Therefore, data transformations are applied to separate the standard deviation η functionally from the mean μ. In many applications, a dependence of the approximate form $\eta = \gamma\mu^{\delta}$ is observed. Taking logarithms on both sides, this relation turns into the linear equation $\log\eta = \log\gamma + \delta\log\mu$. Using the combined grid samples in EU1 and EU2, let s_i be the empirical standard deviation computed from the two observations in the i-th 1/2 in pair and let $\bar{y}_i$ be the corresponding average for $i = 1, \ldots, 200$. Figure 2 shows $\log s_i$ plotted against $\log\bar{y}_i$. Fitting the regression model $\log s_i = \log\gamma + \delta\log\bar{y}_i$ by least squares gives $\hat{\delta} = 1.013 \pm 0.073$. With $\delta = 1$, the functional dependence of the standard deviation on the mean can be removed by taking logarithms (Box and Draper, 1987, p. 284). Figure 2 also shows a plot of the standard deviation within each 1/2 in pair against the mean after taking logarithms. No functional relationship is apparent. An analysis based on the 100 triplets of random samples similarly indicates taking logarithms. (Fig. 2)

The skewness ζ and the kurtosis κ can provide further diagnostics of the need for a

transformation (Howarth and Earle, 1979). The models in the following section involve Gaussian stochastic processes. Thus, the standardized data should have ζ close to 0 and κ close to 3, the values for the standard normal distribution. Table 2 shows the estimated skewness and kurtosis from the grid data, for the untransformed concentrations and after taking logarithms. In both EU1 and EU2, the estimates show much better agreement with the theoretical values after transformation. (Table 2)

Further diagnostics supporting the need for taking logarithms are presented in the next section.

MODELS AND PREDICTORS

Let $Y(t)$ be the dioxin concentration at a location $t = (x, y)$ in a given EU. We will fit a model to the logarithm of concentration,

$$W(t) = \log Y(t) = \beta + Z(t) + \epsilon(t), \tag{1}$$

for each EU. Here, $Z(t)$ and $\epsilon(t)$ are assumed to be independent zero mean Gaussian stochastic processes. The stochastic process $Z(t)$, through its correlation structure described below, represents large scale variability due to systematic departures of the contaminant from the mean β, whereas $\epsilon(t)$ reflects uncorrelated measurement error and very local variation (nugget effect). The mean β is assumed to be constant within an EU. By replacing β with a more general regression in t, we could allow the mean of the process $W(t)$ to be a function of t as well. However, some preliminary experimentation with the data showed that this does not improve the fit of the models.

For $t_1 = (x_1, y_1)$ and $t_2 = (x_2, y_2)$, the spatial correlation of the contaminant is modeled by the family of generalized exponential covariances given by

$$\mathrm{Cov}[Z(t_1), Z(t_2)] = \sigma^2 \exp(-\theta_x |x_1 - x_2|^{2-\alpha_x}) \exp(-\theta_y |y_1 - y_2|^{2-\alpha_y}), \tag{2}$$

with $\theta_x, \theta_y, \sigma^2 \in (0, \infty)$ and $\alpha_x, \alpha_y \in [0, 2)$. Values for α_x and α_y equal to zero lead to smooth surfaces for $Z(t)$, whereas values greater than zero lead to a more erratic structure. Large or small values of θ_x or θ_y represent weak or strong correlations of the

contaminant, respectively. This covariance family thus allows much flexibility in modeling spatial dependence. The errors $\epsilon(t)$ are assumed to have constant variance $\tau^2 \geq 0$, with no correlation between distinct observations, including replicates. Note that the covariance structure in (2) can alternatively be defined in terms of a variogram (e.g., Cressie, 1993, p. 67).

Let $W_D = [\log Y(t_1), \ldots, \log Y(t_n)]^T$ denote the vector of logarithms of dioxin concentrations obtained from taking measurements at the n sites in the design $D = \{t_1, \ldots, t_n\}$. The best linear unbiased predictor (BLUP) $\hat{W}(t)$ of the logarithm of dioxin concentration $W(t)$ at site t under model (1) is given by

$$\hat{W}(t) = c^T(t) W_D \tag{3}$$

with

$$c(t) = \Sigma^{-1}\sigma^2 r(t) + \Sigma^{-1}1_n(1_n^T\Sigma^{-1}1_n)^{-1}[1 - 1_n^T\Sigma^{-1}\sigma^2 r(t)].$$

Here, the $n \times n$ matrix Σ is given by $\sigma^2 R + \tau^2 I_n$, the $n \times n$ matrix R has element (i,j) given by $\mathrm{Corr}[Z(t_i), Z(t_j)]$ from (2), I_n is the $n \times n$ identity matrix, $r(t)$ is the $n \times 1$ vector of correlations with element i given by $\mathrm{Corr}[Z(t_i), Z(t)]$ from (2), and 1_n is the $n \times 1$ vector of 1's. The derivation of the BLUP and its standard error is described by Sacks, Schiller, and Welch (1989), for example. In the geostatistical literature (e.g., Cressie, 1993, p. 119–123), $\hat{W}(t)$ is called ordinary kriging.

Cross validation provides a diagnostic check of model (1). We again consider the grid data. For each EU, maximum likelihood estimates $\hat{\theta}_x$, $\hat{\theta}_y$, $\hat{\alpha}_x$, $\hat{\alpha}_y$, $\hat{\sigma}^2$, and $\hat{\tau}^2$ of the covariance parameters are obtained using the logarithms of the 200 observations. For cross-validation, each observation is then left out in turn and its predictor $\hat{W}(t)$ is computed from (3) using the remaining 199 observations. The first normal probability plot in Figure 3 shows the standardized residuals from cross validating in EU2 against (Fig. 3) quantiles of the standard normal. The points lie roughly on the straight line of unit slope also shown. Thus, there is good agreement with the normal distribution. Moreover, the approximate unit slope shows that the standard errors used to normalize the prediction errors are realistic. Thus, model (1) is consistent with the data.

For comparison, Figure 3 also shows the analogous normal probability plot when cross validating the model

$$Y(t) = \beta + Z(t) + \epsilon(t), \tag{4}$$

with the assumptions for Z and ϵ as in model (1). Discrepancies between the standardized residuals and the quantiles of the standard normal are apparent, providing further evidence of the need for transformation. Leaving out and predicting both observations in each 1/2 in pair leads to the same conclusions, as do analogous diagnostic plots for EU1.

A model of the logarithm of concentration seems to be more appropriate, yet we are typically interested in predicting the contamination on the original scale. We consider various strategies. They are characterized by the use or not of the logarithm transformation in the model and in the linear predictor.

A standard approach is to fit the model using the logarithms of the observed data, compute the BLUP $\hat{W}(t)$, which is a linear predictor in the logarithms of the data, and then exponentiate the BLUP. A multiplicative correction for bias is described in the Appendix. This is known as the ordinary lognormal kriging predictor in the geostatistical literature (Journel, 1980 or Rendu, 1979). Here, we refer to it as the log-log (or LL) predictor as the transformation is applied in model fitting and in the linear predictor. Among linear predictors, the LL predictor minimizes the mean squared error of prediction on the logarithm scale under the constraint of unbiasedness. Accordingly, Dowd (1982) suggested a predictor which is of the same form as the LL predictor, but minimizes the prediction error on the untransformed scale. We refer to it as the LL-D predictor. The optimization must be carried out numerically, a potential disadvantage in applications. Rivoirard (1990) gives a summary of these predictors. Some further details are in the Appendix.

As an alternative to the above predictors, we propose fitting the more appropriate model (1) for the logarithm of concentration but predicting the untransformed concentration $Y(t)$ using a predictor linear in the *untransformed* data. We therefore call it the log-untransformed (or LU) predictor. In the Appendix we derive the optimal coefficients in the predictor from the first and second moments of the lognormal distribution. Thus, under model (1) for the logarithm of concentration, this predictor of $Y(t)$ is the best (i.e.,

has minimum mean squared error) amongst predictors linear in the untransformed data.

For comparison we also consider using untransformed data for the fitted model and for prediction. This is called the UU predictor here. Its derivation (based on assumptions inappropriate for the Piazza Road data) and computation are analogous to those for $\hat{W}(t)$ under model (1).

For the LL predictor, the LL-D variant, and the LU predictor, maximum likelihood estimates of the covariance parameters are obtained from the logarithms of concentration. The same estimates serve for all three predictors. The UU predictor requires different maximum likelihood estimates from the untransformed data. Computations for model fitting are carried out using the software GaSP (Gaussian Stochastic Processes) developed by the second author. Unless mentioned otherwise all remaining computations, including optimization of the Dowd (1982) LL-D predictor, are performed with GAUSS (Aptech, 1996).

If average contamination over an area is of interest, a global estimator (also called block kriging) can be employed (Journel and Huijbregts, 1978, p. 320–324). Averaging pointwise predictors provides a numerically easy implementation (e.g., Istok and Cooper, 1988 or Weber and Englund, 1992, 1994). A best linear predictor results when using a best linear pointwise predictor such as the LU predictor. When the region of interest is of simple geometry, numerical integration of the predictor with a single variance estimate might be possible. Dowd's (1982) predictor can be modified to give a best log-linear predictor of the average that requires only one numerical minimization.

A FACTORIAL EXPERIMENT

The Piazza Road pilot study data provide a rich opportunity to investigate several issues that might arise in the planning stage of a cleanup of a contaminated site. We will assume below that primary interest is in pointwise prediction of the contaminant at unsampled locations.

There are two types of data: grid and random samples. The grid data will be used for model fitting and predicting, while the random data will be reserved for assessing the

accuracy of prediction. Predictions from the grid samples will be made at each of the 50 random locations within an EU, and the 50 averages across replicates from the random samples will be taken as true concentrations.

Before continuing we need to justify this use of the random samples. Recall that at each of the 50 random locations, nine tablespoon samples were taken within a 1 ft sampling frame on a 4 in square grid and composited. From the mixture, three replicates were analyzed. We assume that the true contamination is roughly constant over the relatively small sampling frame at a given location and that averaging the three observations provides a good approximation to the true concentration. To check the approximation, let s_i^2 be the sample variance of the three replicates at location i. If the three replicates are statistically independent, their mean has estimated variance $s_i^2/3$. The root of the average of these estimated variances across the 50 locations in an EU, i.e.,

$$\sqrt{\frac{1}{50}\sum_{i=1}^{50}\frac{s_i^2}{3}},$$

is 2.46 ppb in EU1 and .461 ppb in EU2. These values are fairly small relative to the root mean squared errors of prediction reported in the next section. Thus, the impact of not knowing the true dioxin concentration on the results of the factorial experiment is fairly small (and constant for all the methods compared).

Minimizing the root mean squared error (RMSE) of prediction over the 50 averages at the random locations will be the criterion for assessing the various sampling designs and predictors. It is computed from

$$\sqrt{\frac{1}{50}\sum_{i=1}^{50}(\hat{Y}_i - \bar{y}_i)^2},$$

where $\hat{Y}_i$ is the estimated dioxin concentration at location i and $\bar{y}_i$ is the average of the three replicates.

The factors investigated and their levels are listed in Table 3. The first three factors relate to the sampling design for fitting the covariance parameters in the models. All designs considered are subsets of the grid data. Recall that in each EU, 50 station markers were laid out at 10 ft spacing. Factor DF-10 is the number of these station markers used: (Table 3)

either all 50 or 25. Figure 4 illustrates for EU1, where a station is denoted by '1', '2', or a pair of '1' symbols, distinctions that will be explained shortly. The subset of 25 stations is arranged as a Latin hypercube (McKay, Conover, and Beckman, 1979), i.e., there are five in each column and five in rows 1 and 2, in rows 3 and 4, etc. The same patterns are used in EU2. Factor DF-2 is the number of stations with a pair of 2 ft samples. Both designs in Figure 4 have 15 such pairs, denoted by two '1' symbols close together. The 2 ft pairs are also arranged as a Latin hypercube. Factor DF-R refers to the number of 1/2 in (or replicate) pairs included. We call them replicate pairs as the coordinates within a pair are not distinguished in the data. Both designs in Figure 4 have 10 replicate pairs, denoted by the plotting symbol '2'. They are at two stations in each column and at one in each row. None of the designs considered has a replicate pair and a 2 ft pair at a single station. (Fig. 4)

Factors DF-10, DF-2, and DF-R relate to estimation of long-range, intermediate-range, and measurement error (or nugget effect) variability. By looking at all combinations of these factors we will be able to assess the importance of these components for prediction accuracy.

The next factor in Table 3, DP, is the number of samples in the design for predicting. These data appear explicitly in the predictors. The five levels considered are illustrated in Figure 5 for EU1 (with the same patterns in EU2). The five designs are generated by using all the grid data (level 200), then removing at random one of the replicates from each pair (level 100), then removing at random one of the 2 ft samples from each pair (level 50), then thinning the 50 stations (levels 24 or 15). These designs aim to spread the samples as uniformly as possible (within the limitations of the existing grid data) over the EU. There is particular interest in seeing whether sparse designs (levels 24 or 15) are adequate. (Fig. 5)

Factor P is the predictor used: The LL, LL-D, LU, and UU predictors outlined in the previous section are investigated.

Factor EUF is the EU used for fitting the model. Data from the fitting design in this EU will be used implicitly in maximum likelihood estimation of the covariance parameters.

Factor EUP is the EU for which predictions are made. Level EU1, for example, means

that data from the prediction design in EU1 are used explicitly to predict at the 50 random locations in EU1.

Combining the levels of the seven factors leads to a full factorial experiment with $2 \times 4 \times 3 \times 5 \times 4 \times 2 \times 2 = 1920$ statistical analyses. As an example, run 1295 reads

Run	DF-10	DF-2	DF-R	DP	P	EUF	EUP	RMSE
1295	25	15	10	24	LU	EU1	EU2	2.683

The fitting design has 25 stations, 15 pairs of 2 ft samples, and 10 pairs of replicates (it is the second in Figure 4). The prediction design has 24 observations (the fourth in Figure 5). As the LU predictor is used, estimates of the covariance parameters in the model (1) are required. Fitting is carried out using data from EU1, whereas the LU predictor is a linear combination of data from EU2. The resulting RMSE is for the 50 random locations in EU2.

Evaluating the RMSE for each of the 1920 analyses requires maximum likelihood estimation of the covariance parameters for 24 different fitting designs, in both EU1 and EU2, and for models with and without the logarithm transformation.

No numerical difficulties are encountered, even when the measurement error variance τ^2 is estimated to be zero or close to zero. As a referee pointed out, ill-conditioned matrices can arise for small values of $\hat{\alpha}_x$ and $\hat{\alpha}_y$ in (2), especially when $\hat{\theta}_x$ and $\hat{\theta}_y$ are small as well. For applications where numerical difficulties arise, restricting the number of locations used for fitting or for prediction to a smaller neighborhood around the location of interest would improve the conditioning.

RESULTS

The issues mentioned in the Introduction will be addressed mainly by graphical analysis of the results from the factorial experiment. All graphs show the RMSE of prediction averaged over the levels of all factors not appearing in the graph. EU1, with larger average dioxin concentration and variability, has much larger prediction errors than EU2. Thus, we present the results for the two prediction EUs separately.

An initial look at the RMSE values shows that the accuracies of the LL predictor and the LL-D variant are very similar. Over the 240 combinations of design for fitting, design for predicting, and EU for fitting, the maximum difference in the root empirical mean squared errors between these two predictors is only 0.502 ppb when predicting in EU1 and 0.006 ppb in EU2. In Figures 6 to 11, discussed below, the two predictors would have almost identical average RMSE. Thus, we show only the LL predictor.

We first consider the factors describing the design for fitting the covariance parameters. There is an important interaction between the number of 2 ft pairs (DF-2), the number of replicates (DF-R), and the predictor (P) that uses the estimated covariance parameters. Thus, we need to look at factors DF-2, DF-R, and P together.

Figure 6 shows the LL predictor's average RMSE as a function of DF-2 for each level of DF-R. It is seen that designs for fitting with no 2 ft pairs and no replicate pairs have much larger average RMSE, particularly when predicting in EU1. Performance is much improved with either five 2 ft pairs or five replicate pairs. In EU1, fitting designs with 10 replicate pairs show a small further improvement, but there is no further gain from more 2 ft pairs. In EU2, all fitting designs with some 2 ft pairs or replicates have similar prediction accuracy. (Fig. 6)

The analogous Figure 7 for the LU predictor demonstrates that it is much more stable with respect to factors DF-2 and DF-R. Even with no 2 ft samples and no replication there is little impact on average RMSE. There are minor gains from 10 replicate pairs in EU1; the number of 2 ft pairs has little effect in either EU. (Fig. 7)

If we ignore the highly discrepant fitting designs with neither 2 ft samples nor replication, comparison of Figures 6 and 7 indicates that the LU predictor performs marginally better than the LL predictor in EU1, and vice versa in EU2.

The UU predictor, which does not account for the necessary data transformations, performs poorly in both EUs. Figure 8 shows that, like the LL predictor, with neither 2 ft pairs nor replication a much higher average RMSE results. Comparison with Figures 6 and 7 indicates that for other levels of DF-2 and DF-R the UU predictor typically has the worst prediction accuracy. (Fig. 8)

The stability of the LU predictor even in the absence of 2 ft pairs or replicates is

explained by inspection of the covariance-parameter estimates for the various fitting designs. With no 2 ft pairs or replicates the estimated measurement error variance τ^2 may be zero. Consequently, all three predictors become interpolators following the measured prediction-design concentrations exactly rather than smoothing them. Inspection of some of the predicted surfaces, however, indicates that the LU predictor, even when it is an interpolator, has far less erratic behavior between the design points. This accounts for its robustness.

The poor performance of the LU and UU predictors with neither 2 ft pairs nor replication in the fitting design obscures the results for the remaining factors. As replication is the best guarantee of a nonzero estimated error variance, we henceforth drop level 0 for DF-R from the factorial experiment.

Figure 9 plots the average RMSE against the number of 10 ft station markers in the fitting design. It is seen that 25 are as good as 50, except for the UU predictor in EU1. Again, it is evident that the LU predictor marginally outperforms the LL predictor in EU1, and vice versa in EU2. The UU predictor is always dominated. (Fig. 9)

Cost considerations might rule out extensive sampling in every EU for model fitting. Ideally, we would like to obtain parameter estimates in one or a few EUs and carry them over to compute predictions from sparser data in other EUs. Thus, the data used implicitly and explicitly when constructing the predictor could come from different EUs. Figure 10 shows the average RMSE when predicting in EU1 is much the same when using the parameter estimates from EU1 or from EU2. When predicting in EU2, there is actually a reduction, probably just fortuitous, in average RMSE when the parameter estimates come from EU1, with the LU predictor again showing more stability. Overall, there is no evidence against transferability of the estimated parameters, even though the two EUs are fairly different in terms of dioxin average level and magnitude of variability. Note however, that this is only true as far as prediction accuracy is concerned and might not hold when estimation of the prediction standard deviation is of interest. (Fig. 10)

Figure 11 explores the role of the size of the design for predicting (factor DP) that provides data to be used explicitly in the predictor. (The five prediction designs are shown in Figure 5.) There is a small improvement in average RMSE with design size in EU1 (Fig. 11)

for the LL and LU predictors. In EU2, no systematic trends are apparent. For practical purposes, a small prediction design consisting of 15 observations is adequate here. In particular, including replicates or 2 ft pairs (DP at levels 100 or 200) in the design for predicting seems unnecessary for any of the predictors.

The results of the factorial experiment can be summarized as follows. For the fitting design, a much smaller pilot study would suffice for the model fitting stage of a similar problem. As few as 25 station markers at 10 ft spacing would be adequate. The LU predictor appears not to require replication or 2 ft pairs in the fitting design. For the LL predictor, some replication (or 2 ft pairs) is advised. Similarly, a small prediction design with about 15 samples suffices here. The LU predictor performs marginally better than the LL predictor in the high-variability EU1, whereas the LL predictor performs slightly better in EU2. In new applications, cross-validation could be performed to choose between these two predictors. The UU predictor, which clearly violates the model assumptions, is inferior here.

OTHER APPROACHES

The predictors considered in this article are based on fitting model (1) or model (4). The factorial experiment necessitated each model to be fit 48 times. Accordingly, we used a fairly automatic method, maximum likelihood estimation. Here we compare maximum likelihood with estimation of the covariance parameters via variograms. We also compare the predictors with simpler methods based on nearest neighbors.

Variogram estimation of the covariance parameters has been used in environmental applications by, for example, Cooper and Istok (1988) and Zirschky and others (1985). Details of the method can be found in the textbooks by Cressie (1993, p. 69–83, 90–104) and Journel and Huijbregts (1978, p. 192–194). The latter authors recommended that at least 30 distinct pairs of observations be available at each lag to obtain a reliable estimate for the variogram. The smaller fitting designs amongst the 24 in the factorial experiment do not satisfy this condition. Therefore, we estimate the variograms for EU1 and for EU2 using all 200 samples available. As the UU predictor performed poorly we consider only

model (1) for the logarithm of concentration. Figure 12 shows the estimated variograms in the x- and in the y-direction for EU2, for example. The spatial statistics package available under S-Plus (Mathsoft, 1996) was adequate for these computations. (Fig. 12)

Various models can be fit to the empirical variograms by nonlinear least squares. In addition to the linear, spherical, and rational quadratic variogram models (Cressie, 1993, p. 61), we also fit the generalized exponential model. In notation analogous to Cressie's it is given by

$$\gamma(h;\theta) = \begin{cases} 0, & \text{for } h = 0, \\ c_0 + c_g\{1 - \exp[-(h/a_g)^{2-\alpha_g}]\}, & \text{for } h \neq 0, \end{cases}$$

where h is a distance in a given coordinate, and $\theta = (c_0, c_g, a_g, \alpha_g)$ for $c_0, c_g, a_g > 0$ and $\alpha_g \in [0, 1]$. The exponential model (Cressie, 1993, p. 61) and the Gaussian model (Cressie, 1993, p. 89) are included as special cases by putting $\alpha_g = 1$ or $\alpha_g = 0$, respectively. Choosing the model with smallest residual sums of squares points to the generalized exponential model in the x-direction ($c_0 = 0.249, c_g = 0.352, a_g = 8.305, \alpha_g = 0.313$) and the exponential model in the y-direction ($c_0 = 0.017, c_g = 0.476, a_g = 9.956, \alpha_g = 1$). These fitted models are also shown in Figure 12.

Using the fitted variograms, we can again construct the LL and LU predictors. To distinguish them from the earlier predictors from models fitted by maximum likelihood, we call them the LL_γ and the LU_γ predictors. The RMSE for these predictors is reported in Table 4 for the five prediction designs in Figure 5. For comparison, Table 4 also gives results for the LL and LU predictors with maximum likelihood estimation of the covariance parameters. For these predictors the RMSE is averaged over the levels of the factors DF-10, DF-2, DF-R (excluding level 0), and EU-F, i.e., fitting designs with 30–75 samples. In contrast, recall that variogram estimation is based on all 200 samples in an EU. Nonetheless, Table 4 shows that better prediction accuracy in EU1 follows from maximum likelihood estimation of the covariance parameters. In EU2 the two methods of estimation give similar results. (Table 4)

A numerically much simpler way to predict the dioxin concentration at one of the 50 random locations, say (x_0, y_0), is to average all observations in the prediction design that are within a distance d of (x_0, y_0). We refer to this nearest-neighbors approach as N-d.

Table 4 gives the RMSE of prediction, again using the five prediction designs in Figure 5, with d taking the values 10, 15, 20, 25, and ∞ ft. Allowing d to be infinity means that, for every one of the 50 random locations, the predicted value is just the average of all the prediction-design samples in the exposure unit. In EU1, the N-d predictor typically performs worse than the LU predictor based on maximum likelihood, the best of the kriging predictors considered. Smaller distances d tend to give smaller RMSE, presumably because the local behavior of the contaminant is better captured. For four out of five of the prediction designs, even in the best value of d is chosen, the N-d predictor has larger RMSE than the LU predictor. Therefore, modeling the spatial correlation seems advantageous. A disadvantage of the N-d approach is that for small distances there may be no neighboring observations to average (NA in Table 4). In the lower variability EU2, there is little difference between the best N-d predictor and the LL predictor or its LL_γ counterpart.

CONCLUSIONS

We have used the Piazza Road pilot data to provide guidance on sampling and analysis strategies for point prediction of contamination. For similar sites the main recommendations are as follows.

For fitting the covariance parameters, we found that 25 sampling stations at 10 ft spacing are as good as 50. The main advantage of replicate (1/2 in) pairs seems to be to ensure a positive estimate for the measurement error variance, and hence smooth the predicted surface. Five such pairs are adequate. With some replication, 2 ft pairs do not appear to be necessary.

For prediction in a given EU, 15 sampling stations at 10 ft spacing seem adequate. Increasing the number of sampling stations or including 2 ft or replicate pairs provides little advantage for prediction. A fairly small sampling design could be augmented by further sampling in regions of high prediction uncertainty in a second stage, if necessary. Such sequential strategies will be reported elsewhere.

There is little impact on prediction accuracy when transferring covariance parameter

estimates from one EU to another. This allows the relatively extensive sampling needed for model fitting to be confined to a limited area.

In many environmental applications, a data transformation, often logarithmic, is required to justify the distributional assumptions of the statistical model. Predictions are usually required on the original scale, however. The LL predictor was found to be slightly superior in the low concentration EU, whereas the LU predictor performed better in the higher concentration EU. The LU predictor is based on a model for the logarithm of concentration, but uses linear combinations of the untransformed data. Thus, exponentiating to return to the original scale is unnecessary. Avoiding the potentially unstable exponentiation may be the reason why the LU predictor performs well even when no replicates are included in the fitting design. We would tentatively recommend the LU predictor if prediction on the untransformed scale is required. To check this choice, cross validation could be used to compare the LU predictor with, say, the LL predictor.

Although the Piazza Road pilot data are extensive, they did not allow exploration of other reasonable sampling designs. For example, none of the designs for fitting or for predicting had, say, 5 ft spacing.

Estimating the correlation parameters via maximum likelihood performed well even with only 25–30 samples. In contrast, variogram estimation requires more data, at least if standard practice is followed, yet performed worse. Furthermore, maximum likelihood estimation is more automatic for less experienced practitioners.

The generality of all these findings should be explored by investigating other sites and contaminants.

Further work is in progress. Here we have investigated actual prediction accuracy. In practice, one often needs to estimate accuracy by computing a standard error for each prediction. The impact of design and analysis strategies on the reliability of standard errors is of interest. We have also chosen EUs with minimal data at the detection limit. Other EUs in the pilot data have a considerable proportion of data censored by the detection limit, and new methods of model fitting and predicting are required. Ultimately, cleanup decisions are based on average exposures, and the impact of design and analysis on estimating averages should be investigated.

ACKNOWLEDGMENTS

Financial support from the US EPA under cooperative agreement EPA CR 819638-01-0 to the National Institute of Statistical Sciences is gratefully acknowledged. This work has not been subjected to EPA peer review, however. It may not necessarily reflect the views of the EPA, and no official endorsement should be inferred. Welch's research was also funded by NSERC of Canada. We would also like to thank Evan Englund, Dave Higdon, Max Morris, and Don Ylvisaker for numerous discussions, and an anonymous referee for helful comments.

APPENDIX

We describe a linear and two nonlinear predictors for the dioxin contamination $Y(t)$ at site t under model (1). The discussion assumes the covariance parameters to be known. To compute the predictors, these parameters are replaced by estimates. Notation is as in the section entitled Models and Predictors.

The LL predictor is obtained by exponentiating $\hat{W}(t) = c^T(t)W_D$ and applying a multiplicative bias correction, $K(t)$. This leads to

$$\hat{Y}_{LL}(t) = \exp[K(t) + c^T(t)W_D],$$

where $K(t) = [\sigma^2 + \tau^2 - c^T(t)\Sigma c(t)]/2$. This is the ordinary lognormal kriging predictor suggested by Cressie (1993, p. 135–136), Journel (1980), and Rendu (1979). Among predictors of the form $\hat{Y}(t) = \exp[a(t)+b^T(t)W_D]$, the LL predictor minimizes $\mathrm{E}[\log \hat{Y}(t) - \log Y(t)]^2$ under constraints on $a(t)$ and $b(t)$ to ensure unbiasedness. Retaining the form of the predictor, we can also minimize $\mathrm{E}[\hat{Y}(t) - Y(t)]^2$ with respect to $a(t)$ and $b(t)$, as suggested by Dowd (1982). This minimization must be carried out numerically.

As an alternative to the above predictors which are nonlinear in the data, we propose a linear predictor for $Y(t)$ of the form $\hat{Y}_{LU}(t) = d^T(t)Y_D$. The optimal coefficients $d(t)$ are based on the log-transformed model (1). From properties of the lognormal distribution, $Y(t)$ has expectation $\lambda = \exp[\beta + (\sigma^2 + \tau^2)/2]$. Thus, for unbiasedness, we need the

constraint $d^T(t)\mathbf{1}_n = 1$. Similarly, the property

$$\mathrm{Cov}[Y(t_1), Y(t_2)] = \lambda^2\{\exp[\mathrm{Cov}(Z(t_1), Z(t_2))] - 1\}$$

leads to a mean squared error of prediction. Some algebra applying the constraint $d^T(t)\mathbf{1}_n = 1$ several times is required. We find that

$$\mathrm{MSE}[\hat{Y}_{LU}(t)] = \lambda^2 \left[d^T(t)\Gamma d(t) - 2d^T(t)q(t) + \exp(\sigma^2 + \tau^2)\right], \tag{5}$$

where $q(t) = \exp[\sigma^2 r(t)]$ and $\Gamma = \exp(\Sigma)$, and exponentiation of the vector and matrix is element-wise. Introducing a Lagrange multiplier for the constraint $d^T(t)\mathbf{1}_n = 1$, the constrained minimization of (5) with respect to $d(t)$ gives

$$d(t) = \Gamma^{-1}q(t) + \Gamma^{-1}\mathbf{1}_n(\mathbf{1}_n^T\Gamma^{-1}\mathbf{1}_n)^{-1}[1 - \mathbf{1}_n^T\Gamma^{-1}q(t)],$$

which has a form similar to (3).

REFERENCES

Aptech, 1996, GAUSS Mathematical and Statistical System, System and Graphics Manual: Version 3.2.33: Aptech Systems, Maple Valley, 259 p.

Box, G.E.P. and Draper, N.R., 1987, Empirical Model-Building and Response Surfaces: Wiley, New York, 669 p.

Cooper, R.M. and Istok, J.D., 1988, Geostatistics Applied to Groundwater Contamination. II: Application: Jour. Environ. Eng., v. 114, n. 2, p. 287–299.

Cressie, N.A.C., 1993, Statistics for Spatial Data: Wiley, New York, 900 p.

Dowd, P.A., 1982, Lognormal Kriging—The General Case: Math. Geol., v. 14, n. 5, p. 475–499.

Howarth, R.J. and Earle, S.A.M., 1979, Application of a Generalized Power Transformation to Geochemical Data: Math. Geol., v. 11, n. 1, p. 45–62.

Istok, J.D. and Cooper, R.M., 1988, Geostatistics Applied to Groundwater Pollution. III: Global Estimates: Jour. Environ. Eng., v. 114, n. 4, p. 915–928.

Journel, A.G., 1980, The Lognormal Approach to Predicting Local Distributions of Selective Mining Unit Grades: Math. Geol., v. 12, n. 4, p. 285–303.

Journel, A.G. and Huijbregts, C.J., 1978, Mining Geostatistics: Academic Press, New York, 600 p.

McKay, M.D., Conover, W.J., and Beckman, R.J., 1979, A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output From a Computer Code: Technometrics, v. 21, n. 2, p. 239–245.

Mathsoft, 1996, S+SPATIALSTATS User's Manual: Version 1.0: Mathsoft, Seattle, 228 p.

Rendu, J.M., 1979, Normal and Lognormal Estimation: Math. Geol., v. 11, n. 4, p. 407–422.

Rivoirard, J., 1990, A Review of Lognormal Estimators for In Situ Reserves: Math. Geol., v. 22, n. 2, p. 213–221.

Ryti, R.T., Neptune, D., and Groskinsky, B., 1992, Superfund Soil Cleanup: Environ. Testing and Analysis, v. 1, n. 1, p. 26-31, 67.

Ryti, R.T., 1993, Superfund Soil Cleanup: Developing the Piazza Road Remedial Design: Jour. of the Air and Waste Management Assoc., v. 43, n. 2, p. 197–202.

Sacks, J., Schiller, S.B., and Welch, W.J., 1989, Designs for Computer Experiments: Technometrics, v. 31, n. 1, p. 41–47.

Weber, D. and Englund, E., 1992, Evaluation and Comparison of Spatial Interpolators: Math. Geol., v. 24, n. 4, p. 381–391.

Weber, D.D. and Englund, E.J., 1994, Evaluation and Comparison of Spatial Interpolators II: Math. Geol., v. 26, n. 5, p. 589–603.

Zirschky, J., Keary, G.P., Gilbert, R.O., and Middlebrooks, E.J., 1985, Spatial Estimation of Hazardous Waste Site Data: Jour. Environ. Eng., v. 111, n. 6, p. 777–789.

	Mean	Median	SD	IQR
EU1	13.130	9.790	14.456	16.215
EU2	2.779	2.315	2.036	2.015

Table 1: Mean, Median, Standard Deviation (SD), and Interquartile Range (IQR) of Dioxin Concentration (ppb)

	Skewness		Kurtosis	
Transformation	EU1	EU2	EU1	EU2
None	3.18	2.08	21.60	10.08
Log	-0.73	-0.47	2.54	3.33

Table 2: Skewness and Kurtosis of the Grid Concentrations

Factor	Description	Levels
DF-10	Number of 10 ft station markers in the design for fitting	25, 50
DF-2	Number of 2 ft pairs in the design for fitting	0, 5, 10, 15
DF-R	Number of replicate pairs in the design for fitting	0, 5, 10
DP	Number of observations in the design for predicting	15, 24, 50, 100, 200
P	Predictor	LL, LL-D, LU, UU
EUF	EU for fitting	EU1, EU2
EUP	EU for predicting	EU1, EU2

Table 3: Factors Investigated in the Factorial Experiment

	Number of observations for predicting in EU1					Number of observations for predicting in EU2				
Predictor	200	100	50	24	15	200	100	50	24	15
LL_γ	17.976	17.312	17.577	17.366	17.291	2.564	2.524	2.392	2.654	2.670
LU_γ	16.876	16.556	17.266	16.443	16.573	2.603	2.557	2.481	2.707	2.641
LL	16.318	15.935	16.709	16.741	16.971	2.544	2.481	2.363	2.597	2.617
LU	15.463	15.179	16.213	15.843	16.416	2.688	2.598	2.531	2.665	2.606
N-10	17.942	17.758	16.635	15.247	NA†	2.589	2.580	2.739	NA	NA
N-15	17.907	17.044	17.967	16.082	NA	2.565	2.527	2.417	2.639	2.620
N-20	19.502	18.813	19.971	19.485	19.817	2.726	2.688	2.597	2.828	2.782
N-25	20.177	19.618	20.486	19.198	18.954	2.917	2.886	2.817	3.051	3.008
N-∞	22.131	22.079	22.340	21.929	21.863	2.988	2.964	2.884	2.934	2.867

† Not Available

Table 4: Root Mean Squared Error (ppb) for Various Predictors

Figure captions:

Figure 1. Locations of grid samples and random samples in EU1. One pair of observations was taken 1/2 in apart at each of 100 grid locations; three replicated samples were obtained at each of 50 random locations.

Figure 2. Left: Logarithm of standard deviation versus logarithm of mean for untransformed 1/2 in pairs. Right: Standard deviation versus mean for logarithm-transformed 1/2 in pairs.

Figure 3. Normal probability plots of standardized cross-validation residuals in EU2. Left: model (1). Right: model (4).

Figure 4. Two designs with 50 and 25 grid stations used for model fitting. In both, there are 15 locations with 2 ft pairs and 10 locations with replicate pairs (plotting symbol '2').

Figure 5. Five designs for fitting in EU1 with 200, 100, 50, 24, and 15 samples, respectively ('2' denotes a replicate pair).

Figure 6. Average RMSE of prediction for LL predictor versus number of 2-foot pairs in design for fitting (DF-2), by number of replicates (DF-R).

Figure 7. Average RMSE of prediction for LU predictor versus number of 2-foot pairs in design for fitting (DF-2), by number of replicates (DF-R).

Figure 8. Average RMSE of prediction for UU predictor versus number of 2-foot pairs in design for fitting (DF-2), by number of replicates (DF-R).

Figure 9. Average RMSE of prediction versus number of 10 ft station markers in design for fitting (DF-10), by predictor (P).

Figure 10. Average RMSE of prediction versus EU used for fitting (EUF), by predictor (P).

Figure 11. Average RMSE of prediction versus number of observations in design for predicting (DP), by predictor (P).

Figure 12. Estimated variograms from grid-sample logarithms of dioxin concentration in EU2. Left: x-direction. Right: y-direction. Solid lines show the variogram models fit by nonlinear least squares.

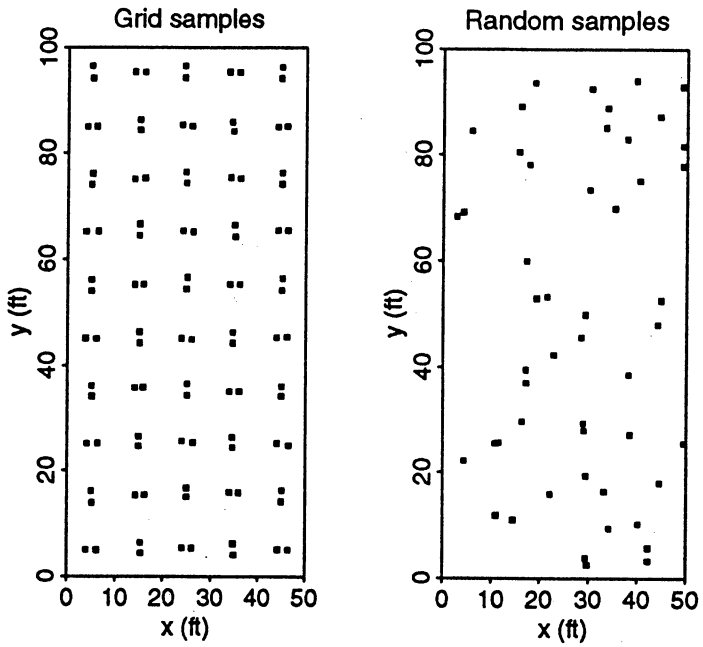


Figure 1, MS# 96-78, Abt, Welch, and Sacks

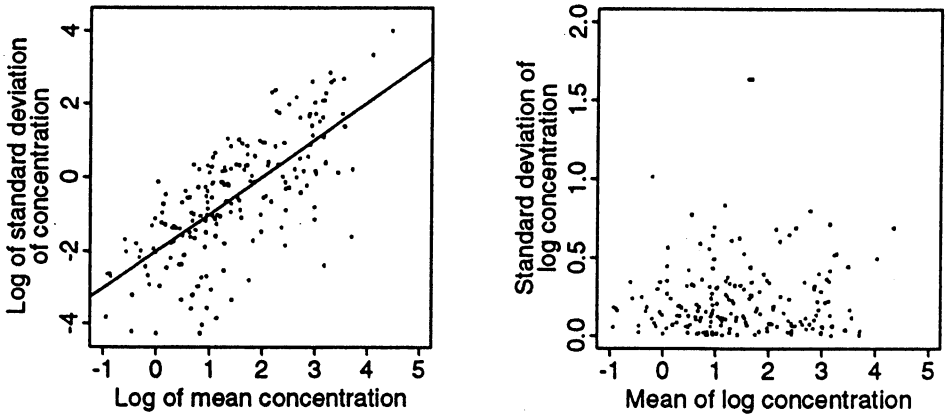


Figure 2, MS# 96-78, Abt, Welch, and Sacks

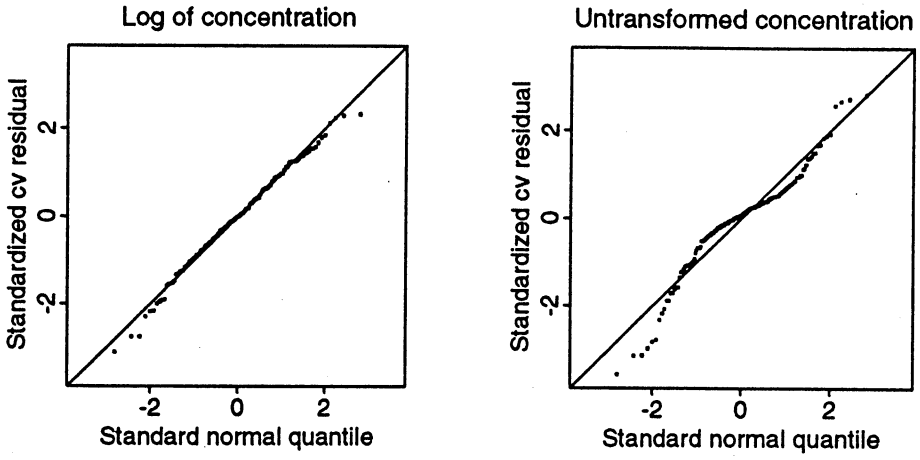


Figure 3, MS# 96-78, Abt, Welch, and Sacks

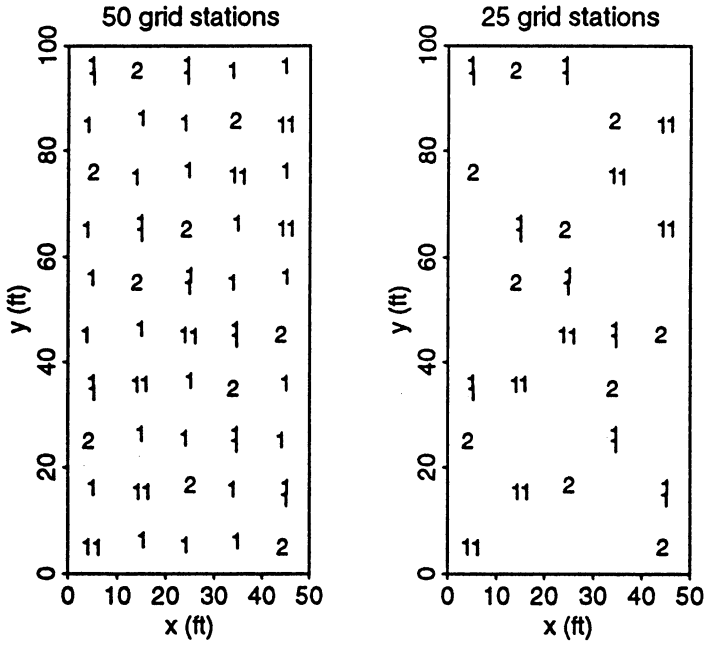


Figure 4, MS# 96-78, Abt, Welch, and Sacks

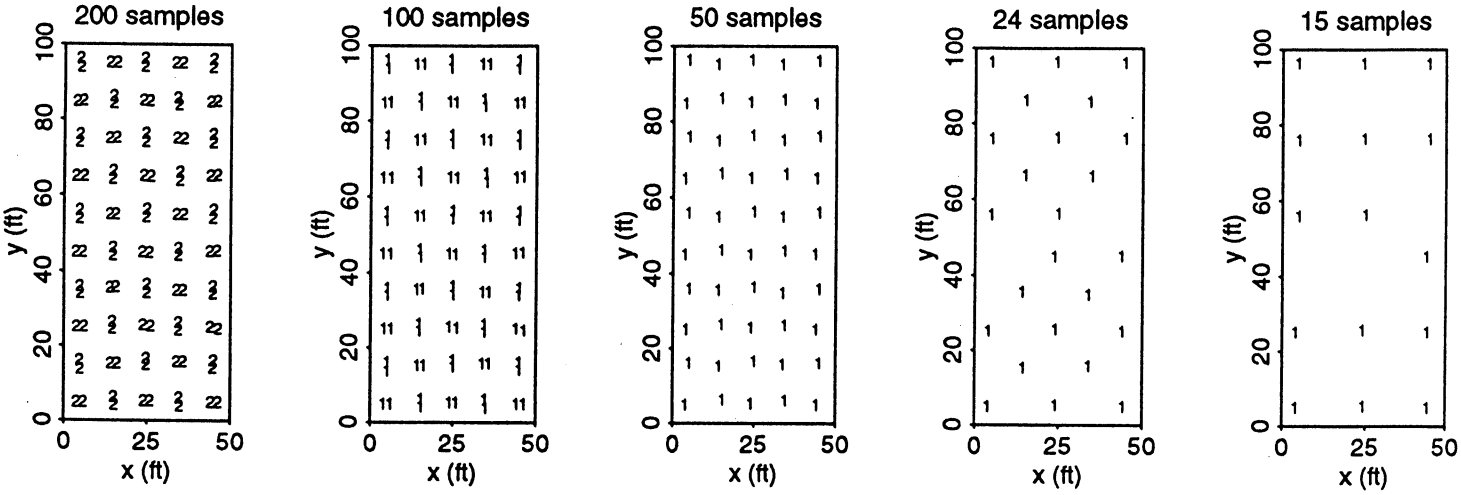


Figure 5, MS# 96-78, Abt, Welch, and Sacks

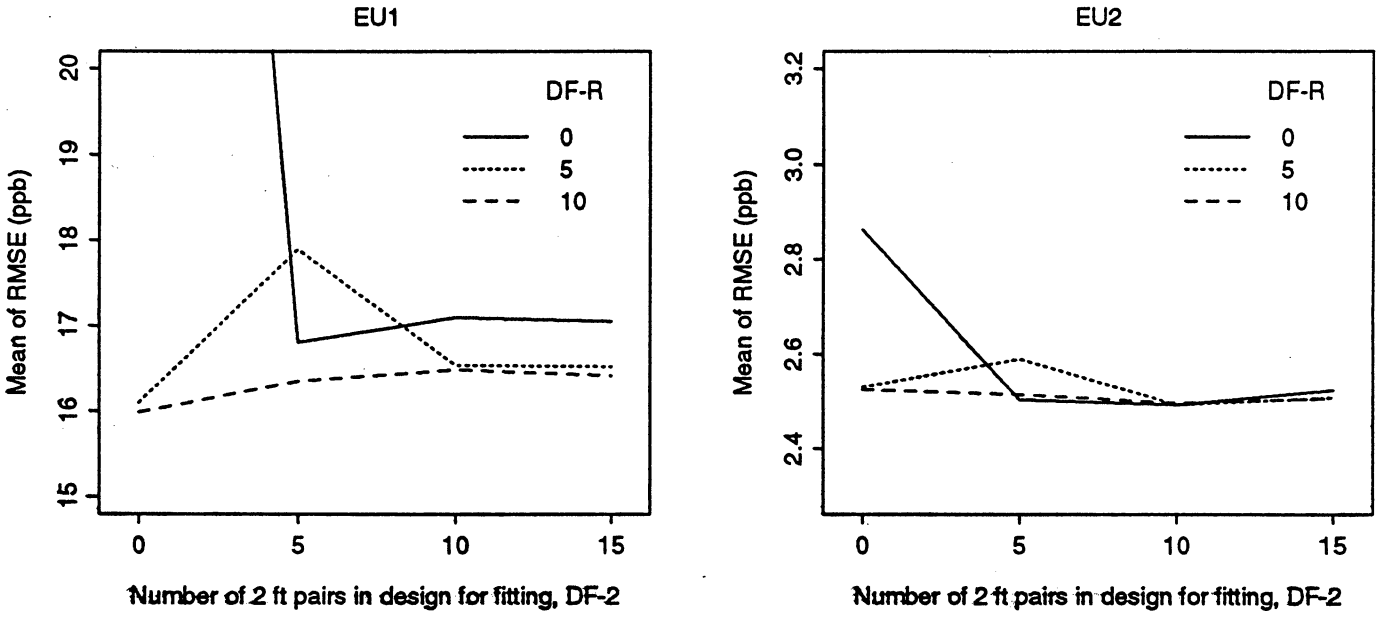


Figure 6, MS# 96-78, Abt, Welch, and Sacks

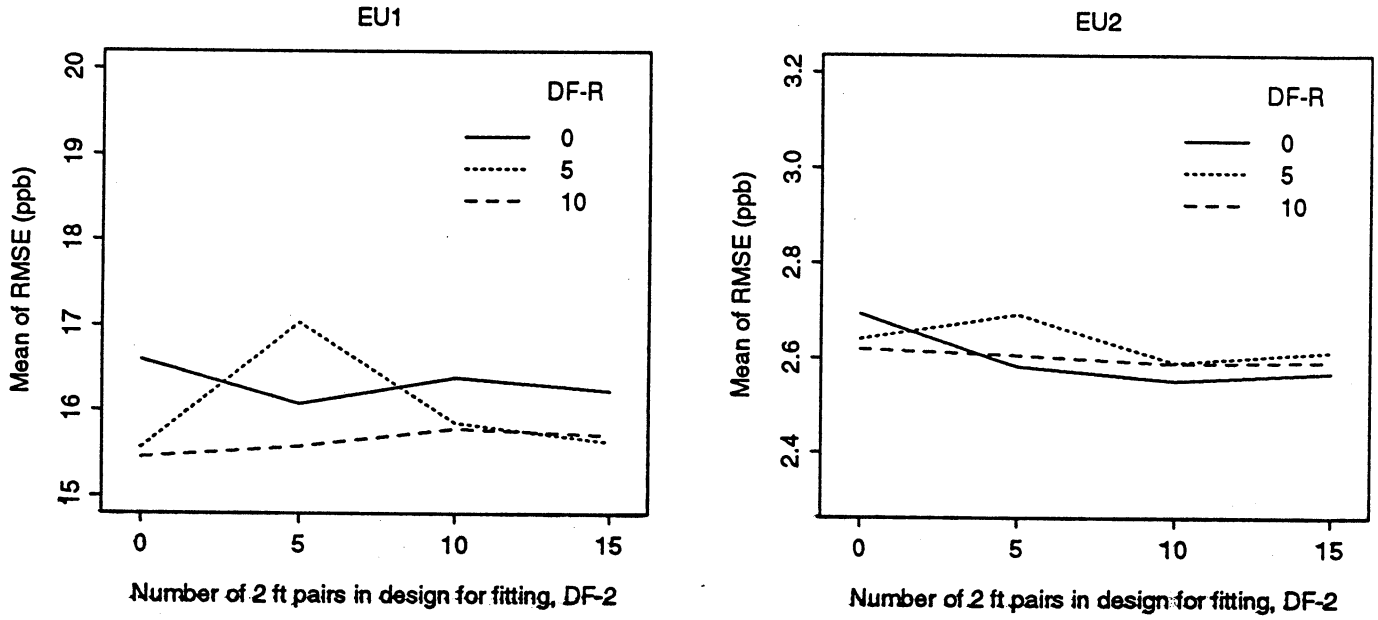


Figure 7, MS# 96-78, Abt, Welch, and Sacks

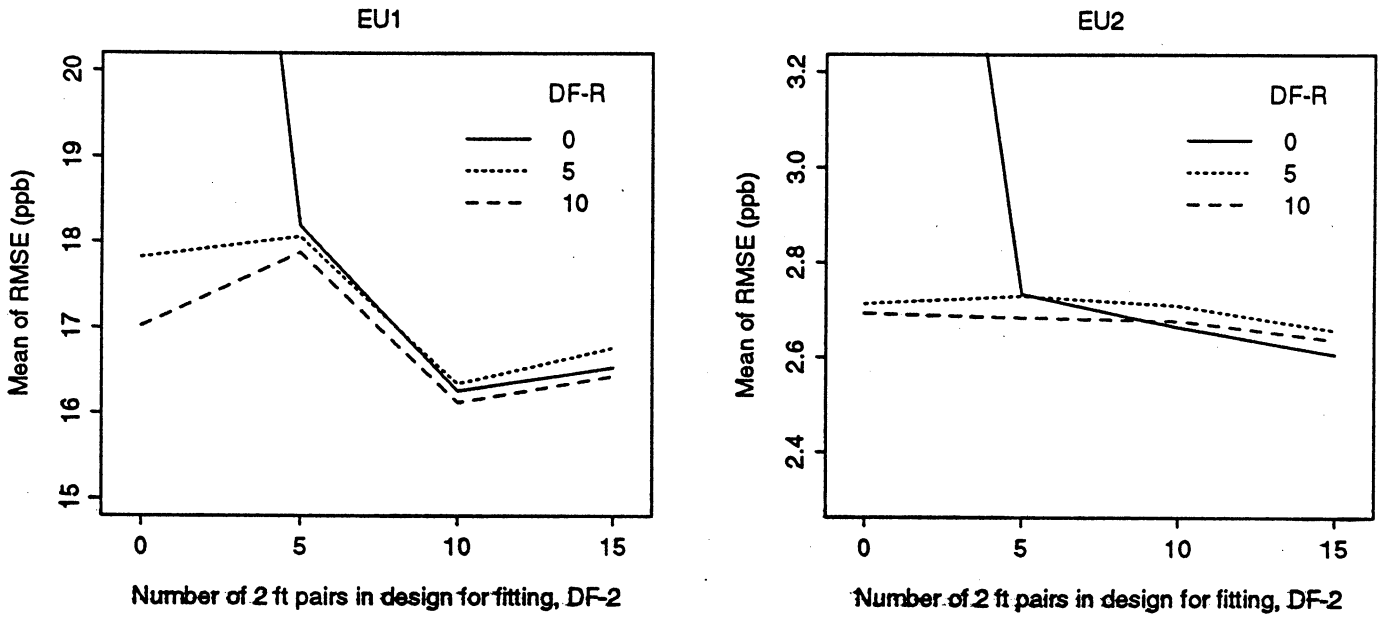


Figure 8, MS# 96-78, Abt, Welch, and Sacks

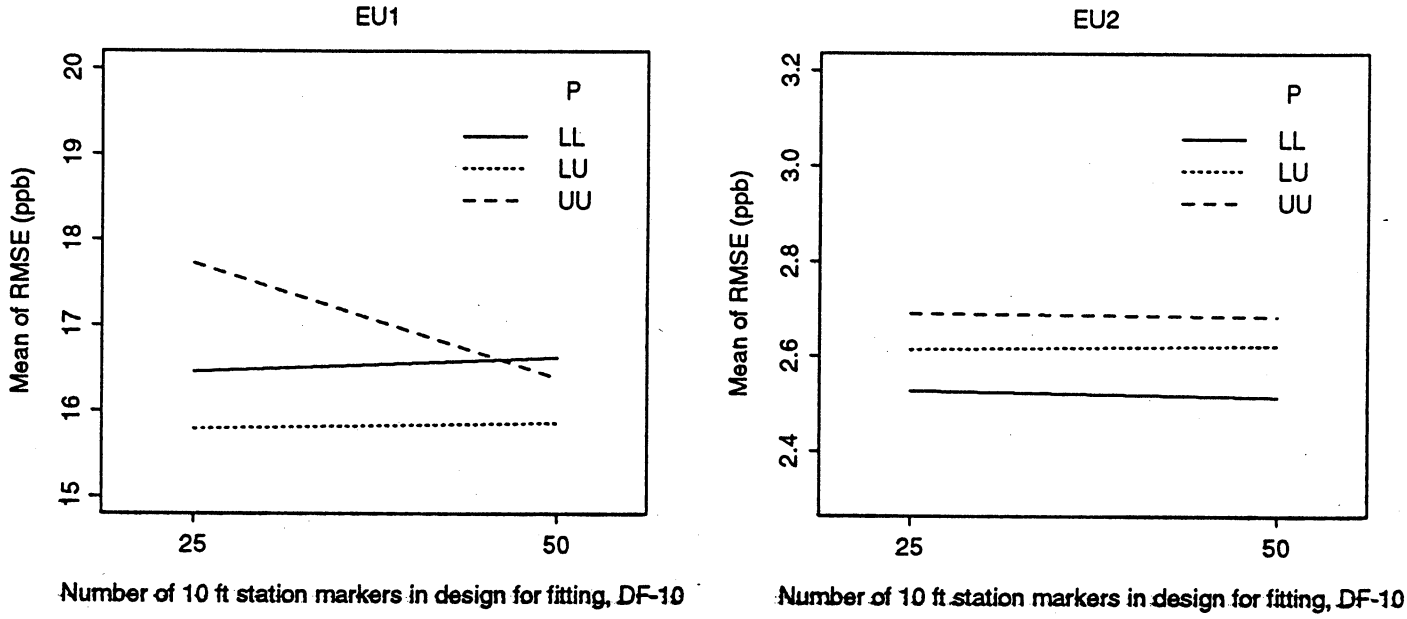


Figure 9, MS# 96-78, Abt, Welch, and Sacks

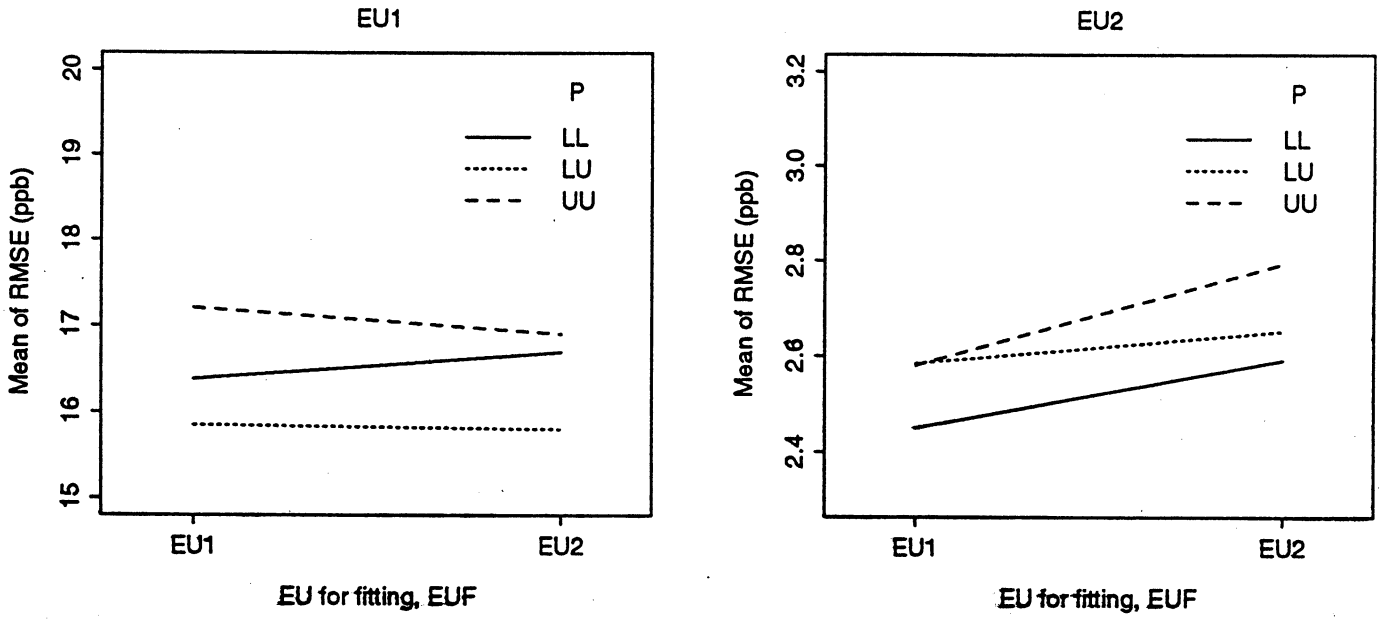


Figure 10, MS# 96-78, Abt, Welch, and Sacks

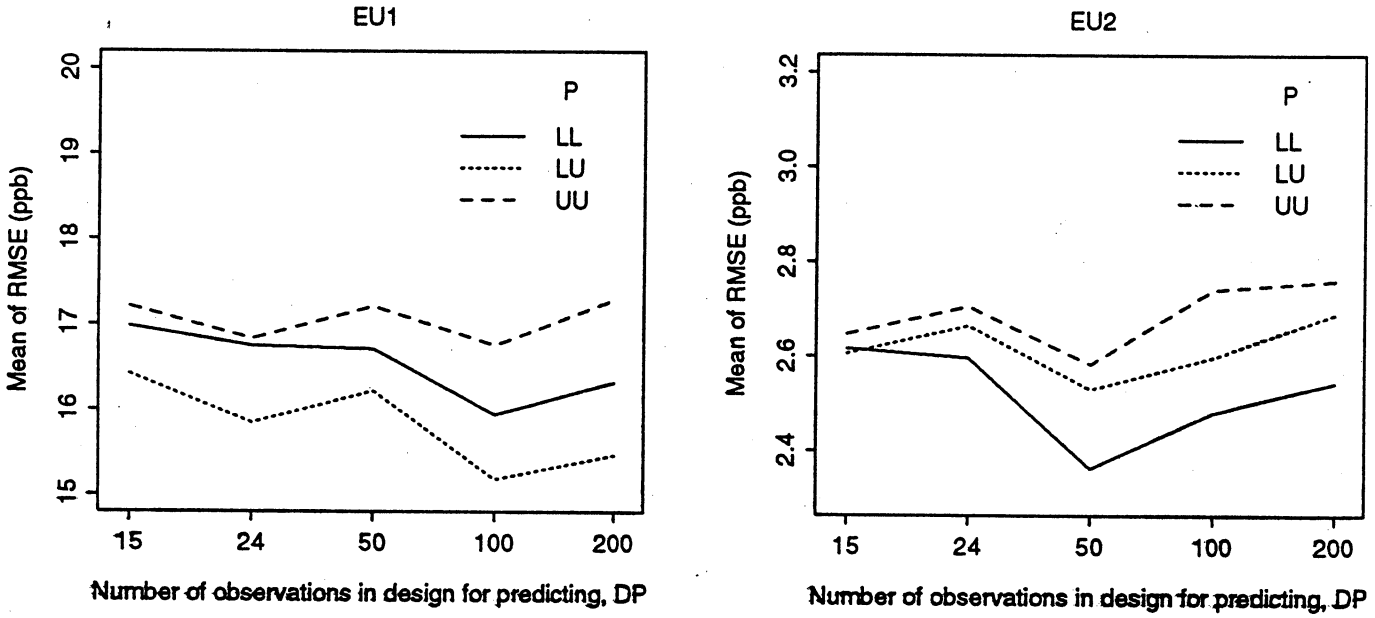
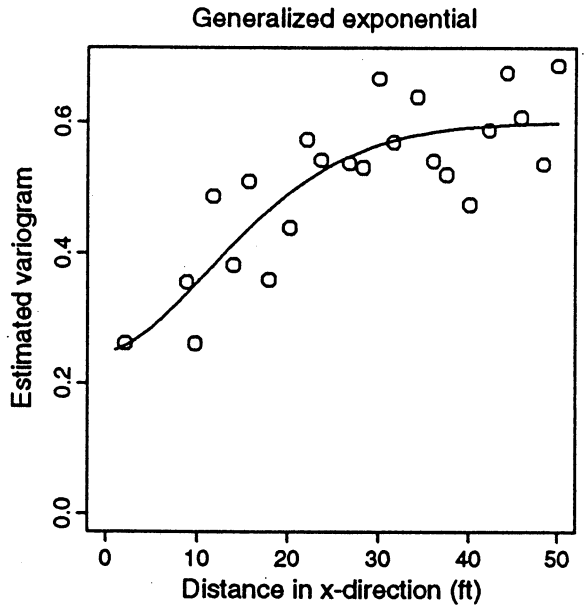


Figure 11, MS# 96-78, Abt, Welch, and Sacks



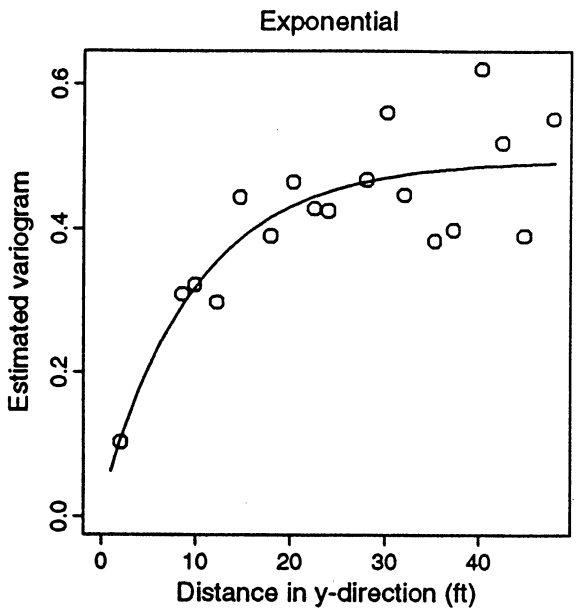


Figure 12, MS# 96-78, Abt, Welch, and Sacks