Design of Air-Quality Monitoring Networks

Douglas Nychka
North Carolina State University

Nancy Saltzman
National Institute of Statistical Sciences

1 Introduction

Where should ozone be measured? It is well accepted that high levels of ozone are not only damaging to human health but also reduce crop yield and damage vegetation.[1] However, the continuous measurement of ozone at a location is relatively expensive and so the number and locations of instruments need to be chosen judiciously. This question, although deceptively simple, raises a host of fundamental issues. Most importantly, how can we infer ozone levels at places where measurements are not made? What does it mean to measure ozone well and how many monitoring instruments are really necessary?

1.1 Environmental Issues

Although we will find that these questions, and some answers, can be formulated in terms of statistical and mathematical concepts, this work is rooted in the practical problem of adequately monitoring the environment with limited resources. Figure 1 indicates the locations of the ozone monitoring network for a portion of the Midwest. This chapter is about tackling some specific design problems for this region of the United States. These include *augmenting* a network in size for better spatial representation or *reducing* a network in a way to maintain the best coverage of the region. One intriguing result that emerges from these design case studies is that under certain measures of performance, the monitoring network for ozone is not significantly compromised when the number of stations is reduced by half. The technical contributions from this work are the construction of efficient arrays of stations and the ability to draw objective statistical conclusions about performance. There is currently interest in monitoring other atmospheric pollutants. Reducing the network for a well-studied pollutant, such as ozone, would free up resources to track other constituents.

Given the importance of ozone for air quality, it may seem strange that the design and assessment of regional networks, such as the area in Figure 1, is not already a standard practice. Two reasons for this are the regulatory ozone standards and the difficulty of modeling the ozone field. The location of many

[1] See Chapter 1 for a more detailed discussion of the problems with ambient ozone pollution.

stations has been dictated by the regulatory requirements of detecting peaks of very high ozone in urban areas. For example, stations might be placed near known sources of high ozone or high precursor emissions (such as a refinery). To account for the transport of the pollutant, some stations may be sited down-wind of known sources or urban cores. Thus, the care in placement has not been driven by interest in measuring, say, a spatial average, but to have sensitivity for extreme ozone conditions that may occur over a short period of time and at some specific location.

Determining the spatial distribution of ozone is a much more difficult task and is in contrast to the succinct regulatory statistics derived from the network maxima. Ozone has a nonstationary covariance, with the variance and the correlations depending on where it is measured. Typical methods developed in geostatistics do not work well for nonstationary fields such as ozone. Thus, statistical techniques have not been readily available to determine the uncertainty (e.g., standard errors) associated with predictions made from the monitoring locations. But before outlining how these statistical problems might be overcome, it is appropriate to explain how the monitoring data might be used in addition to measuring the regulatory standard.

1.2 Why Find Spatial Predictions for Ozone?

Besides meeting the regulatory requirements, two other applications of the monitoring data will be mentioned: determining exposure of the population and validating numerical (physical) models. First, a spatial average of ozone concentrations could measure the general exposure of individuals to ozone in a region. This measure is in contrast to the regulatory statistics that register extreme events, possibly restricted to a small part of an urban region. Although a spatial average is a reasonable measure, the exact spatial weighting used to form the average is open to debate. For example, a simple (geographic) spatial average of the ozone concentrations over an urban area may ignore the fact that residential areas have lower ozone levels but have more people. This suggests that the spatial weights might be related to population density instead of area and, so, residential areas would be given greater weight. However, places where people work (during the afternoon hours of peak ozone) might be located in core urban areas with high levels of the pollutant. Examples like these suggest that no one particular weighting scheme over space will always be appropriate or accepted. Moreover, the weighting may change over time as activity patterns and the production and transport of ozone in a region change.

Also, an urban area that has poor air quality may be required by law to take steps to decrease the amount of pollution. But how can policymakers decide what measures will be effective? The salient feature of this problem for a statistician is that it *cannot* be answered directly by the analysis of contemporaneous observational data. A natural tool in the decision-making process is to use a physical model that will provide predictions of pollutant concentrations if pollution-generating activities are changed from current practices. In the case of ozone, one needs a physical model describing the creation and transport of ozone as a function of the chemical emissions and relevant weather patterns. Creating a large numerical model for predicting a pollutant, such as the Regional Oxidant Model (ROM) for ozone (Pierce *et al.* 1994; Alapaty *et al.* 1995), is a substantial undertaking and represents a significant scientific contribution. However, in order to use this model for prediction, it is crucial to collect observational data that can validate the numerical model and provide guidance on mechanisms that the modelers have overlooked. Such a comparison is not straightforward and must account for the agreement or disagreement of the model output and observed measurements at different levels of spatial resolution. For example, the model and observed data may be similar if averaged over a 100-square mile block but may have some differences when compared at specific locations. Similar to measures of exposure, we see that model validation also requires a flexibility in the precise way the estimated ozone field is to be used.

1.3 Designs and Data Analysis

The use of monitoring data for measuring exposure or validating numerical models makes it clear that one must be able to predict ozone at spatial locations other than those where measurements are taken. These are complicated problems; the estimated ozone field will be used in many ways, either by being evaluated at particular points or integrated over different regions. Thus, there is no single statistic that will provide all the information needed to solve either problem. A simple way to formalize this flexibility is to demand that ozone be predicted well on the *average* in the design region, or for *all* points in the design region. Those familiar with spatial design (or the design of experiments) will recognize these as criteria based on the prediction variance and suggesting A-optimal (minimizing mean squared error) or G-optimal (minimizing maximum mean squared error) designs (e.g., Johnson *et al.* 1990).

A surprising feature of these studies is that finding the best design according to a specific criterion plays a small role in the project. This is in contrast to a more formal textbook treatment of designs. Why this difference? The pure statistical design problem *assumes* an underlying statistical model for the measurements and so this abstract problem rarely meets data and does not address multiple objectives from the measurements. For the network design problem, the model for ozone must be estimated and so part of this chapter is concerned with the modeling of the ozone field from data. The computation of optimal designs for large numbers of points can be onerous, not only discouraging interactive use but also requiring a large investment in software development. Practical experience (and some theory) suggests that simpler, geometric criteria yield designs that work nearly as well as optimal ones. Also, because the spatial estimates will serve several purposes, applying extensive effort to achieve the optimal design based on a specific criterion may be misguided. The emphasis in the studies discussed below will be on designs that can be readily computed and that fill out the design region. Although the resulting designs will be evaluated according to standard design criteria, we do not expect them to be optimal.

The linchpin for evaluating network designs is to formulate a statistical model for the distribution of ozone over space. We will use some standard spatial statistical methods, but there are some twists required for air-quality measurements such as ozone. In particular, nonstationary spatial covariance functions will be fit to the ozone field. Air-quality data has a time component that can be turned to good benefit in fitting models. This is very different than the usual situation in geostatistics where one has only one "sample" and must estimate the covariance function with only a single observation at each location. Because the covariance model for the ozone field is central to all the design applications, this is presented in a section separate from the case studies. Usually, numerical models such as ROM are validated against observed data. However, for reasons that will be explained below, we will also use the ROM output to enhance the observational record and fit models to it. Turning the tables and using model output for estimating spatial covariance functions is new and not without some controversy.

1.4 Chapter Outline

Because modeling is central to all the design problems discussed herein, the description of the ozone data and the models are abstracted from the case studies of interest and follow in the next two sections. Section 3 ends with a short discussion and an example of design evaluation.

The design problems tackled in this chapter appear in Sections 4, 5, and 6, and are ordered in increasing complexity. They reflect the actual research path on this topic. Also, to give the reader a feeling for the interplay between substantive environmental questions and the development of new statistics, the statistical methodology will be introduced in context as it is needed. The first project is small in scale and investigates *reducing* the size of the 20-station network for the Chicago urban area. By adapting algorithms from regression subset selection, it is possible to search through the large number of possible subsets (2^{20}) to find groups with good properties. The next design problem is to *augment* a region in rural Illinois for better measurement of rural ozone levels. Here, we introduce a tool for constructing designs based on filling space. These designs are not "optimal," but they can be computed rapidly and can adapt to the practical constraints of the geographic regions. Indeed, use of the space-filling designs became the key step in completing this project. The third design problem considers the impact of augmenting or reducing the network for the midwest region depicted in Figure 1. The last section of the chapter draws some general conclusions and suggests areas for future work.

The reader is referred to the *web companion* for specific data sets and software that are related to the case studies in this chapter.

2 Data

In the case studies two forms of ozone data were used in modeling: observational data from the NAMS/SLAMS network, and output from a run of the Regional Oxidant Model (ROM) for the period 6/1/87 – 8/30/87.

2.1 Hourly Ozone and Related Daily Summaries

There are approximately 500 stations in the combined NAMS/SLAMS[2] network that measure ozone in the United States. This network varies in size over time and tends to concentrate stations in urban areas. Figure 1 locates some of these stations for the midwest region of the United States.

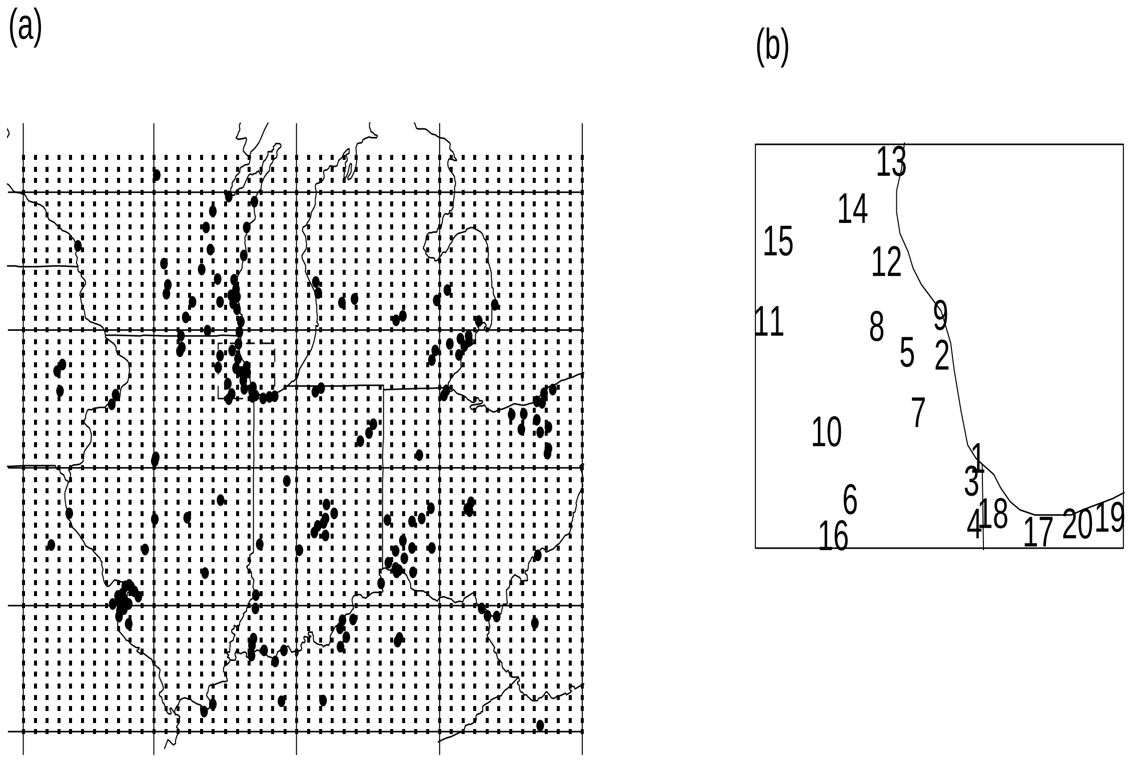

FIGURE 1. (a) locations of the NAMS/SLAMS ozone monitoring stations for 1987 midwestern United States, the ROM grid cell centers, and the 4 × 4 blocks used to investigate nonstationarity of the ozone field (see Figure 8). The second figure (b) is an enlargement of the dashed box from (a), highlighting the the Chicago urban area.

[2]The National Air Monitoring System and the State and Local Monitoring Systems. These data are available in the EPA AIRS database. We would especially like to acknowledge the help of William Cox from EPA in making these data available.

To be comparable with the ROM output, the most extensive set of observational data studied were measurements made over the summer of 1987 and consist of approximately 160 stations in the Great Lakes region. In addition, data from the Chicago urban area over the period 1987–1991 were used. Ozone measurements were in units of parts per billion (ppb) and were usually aggregated into hourly measurements.

There are several common ways of converting the 24-hour recording in a day to a daily summary statistic. The main daily summary used in this work is the eight hour average over the period from hours 9 through 17.[3] One advantage of the 8-hour average is that by averaging over time, it can adjust for some of the local transport of ozone. Also, for spatial modeling and prediction, it is helpful for the field to be approximately normally distributed. A statistic based on an average will be more likely to have a distribution that is close to normal. Finally the 8-hour average has relevance given that recent air-quality standards have been phrased with respect to this statistic. Some other useful daily summaries are the maximum and the two truncated sums SUM06 and SUM08. SUM06 (SUM08) is defined as the sum over all measurements that exceed 60 (80) ppb over the day. Note that because of the hard threshold in their definitions, SUM06 and SUM08 can be identically zero for days when ozone measurements are below the threshold.

2.2 Handling Missing Data

Ozone records for a station often contain missing observations and any careful analysis of these data must deal with this problem. For the modeling used in this work, we have filled in missing hourly values using the median polish techniques of Davis *et al.* described in Chapter 2. However, if a substantial number of hourly measurements are missing, the day's record is coded as missing and skipped. In contrast to a time-series analysis where complete time records are very convenient, it is not necessary to fill in the missing day's measurements to derive spatial covariance information. For example, under the assumption that the ozone field is stationary in time, one can estimate the covariance between two locations by the sample covariance based on the available data for these two locations. Although this "pairwise" computation of covariances uses as many data values as possible, it has some problems and more will be said about this in the next section.

2.3 Model Output

It may seem strange to use simulated data (ROM output) to model the covariances of the ozone field. The difficulty with building models for network

[3]The more conventional summary is the *maximum* eight hour average found over a 24-hour period. Since ozone levels peak in early afternoon, however, the maximum 8-hour average will usually be over the window from 9–17 hours.

design is that observational data are sparse just in the areas where information is needed. This is the primary reason for considering modifications to the network. To overcome gaps in the observational data, ozone concentrations from the output of ROM are used as surrogate data to model the spatial covariances.

At its most basic form, ROM is a boundary layer photochemical grid model with a resolution that is useful for modeling large (1000 km) domains (Lamb 1983; Young *et al.* 1989). The model attempts to simulate ozone under different emission inputs and different meteorology, and aids in studying the effect of hypothetical emission patterns on the production of ozone. The model run studied here is from ROM (version 2.2) where weather inputs have been set to follow the "ozone season" for 1987, 6/3/1987 – 8/30/1987 and where emission patterns follow the estimated 1990 inventory. The model output is reported on a grid with a spacing of $1/4°$ latitude by $1/6°$ longitude (18.5 km $\times$ 18.5 km) and hourly values for ozone concentrations are available for each grid cell. Model output was validated against observational data (Davis *et al.* 1998) and the covariance structure was compared over a larger region. The positive results of this comparison were encouraging enough to use the ROM output for modeling.

3 Spatial Models

The key to any method of spatial interpolation or extrapolation is posing a model for the unknown surface. In the context of pollutants, it is useful to assume that the concentrations, or a suitable transformation of them, are a realization of a Gaussian random surface. The reader is referred to Cressie (1991) for background on spatial statistics related to random fields and a review of the literature in this area. The main goal of this section is to review some aspects of spatial statistics and describe some nonstandard details for ozone.

3.1 Random Fields

Formally, let $Z(\boldsymbol{x})$ denote the measured amount of pollutant (or transformed pollutant) at location $\boldsymbol{x} \in \Re^2$. Assume $Z(\boldsymbol{x})$ is normally distributed with $EZ(\boldsymbol{x}) = 0$ and $\mathrm{Cov}(Z(\boldsymbol{x}), Z(\boldsymbol{x}')) = k(\boldsymbol{x}, \boldsymbol{x}')$. Also, let $\boldsymbol{x}_j$ for $1 \leq j \leq N$ be the N locations where the field is observed; $Z_j = Z(\boldsymbol{x}_j)$ and $K_{i,j} = k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \mathrm{Cov}(\boldsymbol{Z}_i, \boldsymbol{Z}_j)$ denote the covariance matrix for this observation vector.

Under the assumption of isotropy,[4] a common covariance function used for a spatial prediction is the exponential function

$$k(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2 e^{-||\boldsymbol{x}-\boldsymbol{x}'||/\theta},$$

[4]Here, isotropy is taken to mean that the covariance function, k, only depends on $\boldsymbol{x}$ and $\boldsymbol{x}'$ through the distance that they are separated.

where $||\cdot||$ is a measure of distance between two locations. In this chapter, the flat Euclidean distance, $||\boldsymbol{u}|| = \sqrt{u_1^2 + u_2^2}$, is often used for computational convenience, but when possible, great circle distance is used to adjust for curvature of the earth. Note that for any pair of locations separated by a fixed distance, the correlation will be the *same*. In addition, the marginal variances of the fields are assumed to be constant: $\mathrm{Var}(Z(\boldsymbol{x})) = \sigma^2$. For ozone fields, and many other pollutant fields, this isotropic model is not adequate because the marginal variance, σ^2, is not constant with respect to locations. Indeed, the correlations often depend on the locations as well. Figure 2 indicates the nonstationary nature of the observed ozone correlations. Here, we see that the correlations in the first plot tend to decrease more rapidly with the distance of separation than in the second plot.

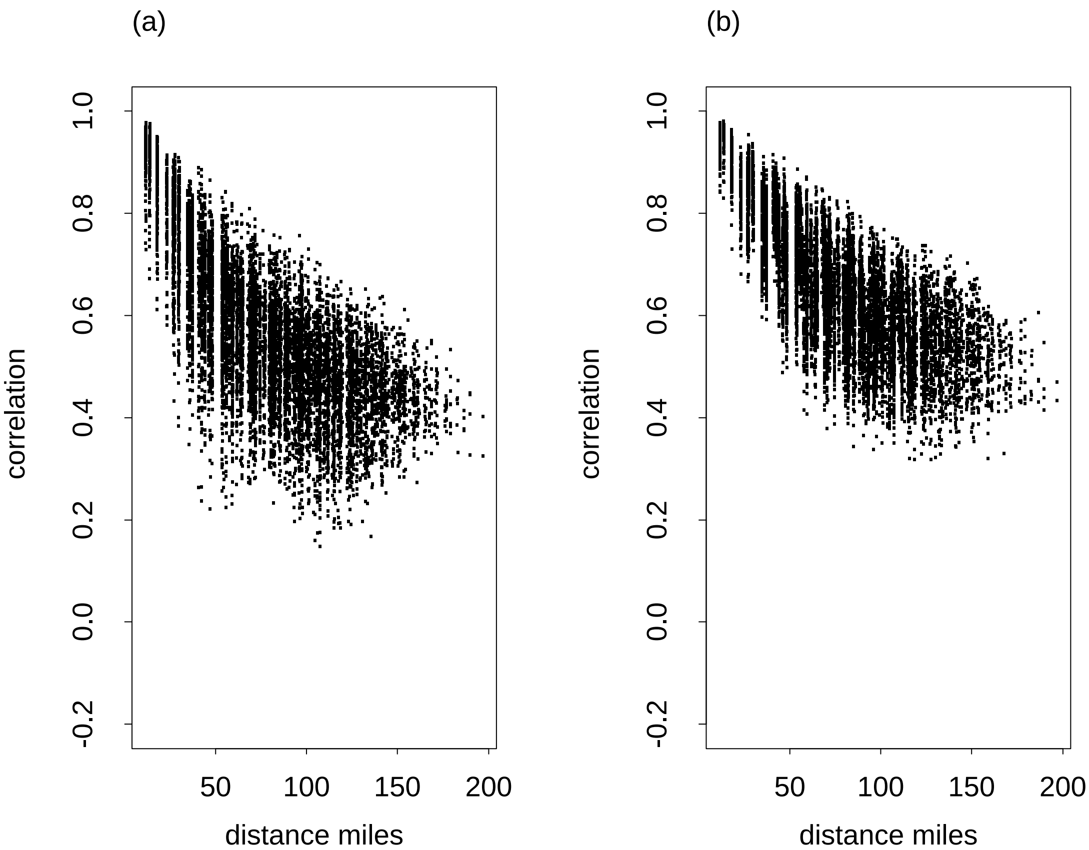

FIGURE 2. Correlograms of ROM output for two blocks for the Great Lakes region. The two plots are correlograms for ROM output in the second and third blocks from the second row from the 4×4 arrangement in Figure 1a. Note that these blocks cover the Chicago area, including southern Lake Michigan and the region covering the common borders of Indiana, Michigan, and Ohio, respectively. Pairwise correlations are calculated for pairs of ROM cells based on the 8-hour average over the summer 1987 model run. For each pair of ROM cells in a block, the correlation is plotted against the distance between cells.

A useful extension allows for different marginal variances

$$k(\boldsymbol{x}, \boldsymbol{x}') = \sigma(\boldsymbol{x})\sigma(\boldsymbol{x}')e^{-||\boldsymbol{x}-\boldsymbol{x}'||/\theta}$$

but still retains constant (isotropic) correlations. A further generalization is to consider a nonstationary covariance that has a parametric part, such as the exponential model, and a nonparametric covariance, represented as a series of eigenvalues and eigenfunctions:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \sigma(\boldsymbol{x})\sigma(\boldsymbol{x}')\left(\rho e^{-||\boldsymbol{x}-\boldsymbol{x}'||/\theta} + \sum_{\nu=1}^{M} \lambda_\nu \psi_\nu(\boldsymbol{x})\psi_\nu(\boldsymbol{x}')\right),$$

where $0 \leq \rho \leq 1$ and $\{\lambda_\nu > 0\}$. Here, the eigenexpansion is computed from the difference between the sample correlation matrix of the data and the isotropic part of the model. This form is a hybrid between the parametric models typically used in spatial statistics and empirical orthogonal function (EOF) expansions used in the atmospheric sciences (Nychka *et al.* 1998). This nonstationary model has a simple interpretation in terms of the observed field. Let

$$Z(\boldsymbol{x}) = \sigma(\boldsymbol{x})\left(\rho S(\boldsymbol{x}) + \sum_{\nu=1}^{M} a_\nu \lambda_\nu^{1/2} \psi_\nu(\boldsymbol{x})\right),$$

where $S(\boldsymbol{x})$ is an isotropic process with a marginal variance of one and mean of zero, and $\{a_\nu\}$ are independent $N(0,1)$ random variables. Thus, the field is a sum of an isotropic process and a linear combination of M additional functions that have random coefficients. The fraction of variance contributed by the parametric part is ρ. A full development of this approach and some background on other nonstationary models can be found in Nychka *et al.* (1998).

In the standard geostatistics framework where only a single realization of a field is available, these models would be difficult to estimate. This is due to the variability associated with the variogram. Because we are concerned with nonstationary covariances, the individual differences of the variograms cannot be aggregated into distance classes. So, with just a single realization, only one squared difference is available as an empirical estimate of the covariance between two locations. In this work, the data from 69 days are used to estimate the covariance where we assume that the covariance structure is consistent over time. Thus, there are 69 observations each the sample covariance between two points; thus, fitting nonstationary models to the empirical covariances is much more stable.

Based on the data from the summer period in 1987, the two covariance models listed above were fit to the observation data *and* the ROM output. The marginal variances were estimated from the ROM or observational data and interpolated to a grid. For the first model, the range parameter was taken to be the median from fitting exponential models to the 16 blocks of data indicated in Figure 1a. For the second covariance model, the parameter θ and the eigenexpansion were estimated from the middle 69 days of the summer. The remaining 10 days at either end were used for cross-validation in order to determine ρ. The results were $\rho = .5$, $\theta = 140$ (miles) based on the ROM output and $\rho = .25$ and $\theta = 120$ (miles) based on the observational data. For both kinds of data, the number of eigenfunctions in the expansion (M) was fixed (subjectively) at 5 although this parameter might be also estimated by cross-validation. Based on the ROM output field for ozone, the isotropic component is fairly long ranged but only accounts for approximately one-half of the variability in field.

3.2 Spatial Estimates

The spatial prediction problem is quite simple: Determine the pollutant levels at points where they are not observed. More formally, the goal is to estimate $Z(\boldsymbol{x})$ given Z_j for $1 \leq j \leq N$. Under the assumption of normality, the best linear unbiased estimate of $Z(\boldsymbol{x})$ is given by

$$\hat{Z}(\boldsymbol{x}) = \gamma^T K^{-1} \boldsymbol{Z},$$

where $\gamma_j = \mathrm{Cov}(Z(\boldsymbol{x}), Z_j)$ and this estimate has prediction variance

$$E(Z(\boldsymbol{x}) - \hat{Z}(\boldsymbol{x}))^2 = \mathrm{Var}(Z(\boldsymbol{x})) - \gamma^T K^{-1} \gamma. \tag{4.1}$$

Although this optimality should not be taken too seriously, these estimates have been widely successful for fitting and interpolating spatial data. It should be noted that the simple estimate outlined above can be improved by adding in a low-order polynomial term to account for a general trend, and for this more general estimator, the form for the prediction variance is similar. One problem, however, is that the prediction variance formula is sensitive to misspecification of the covariance. Because in practice the covariance is unknown, this may introduce biases in the reported prediction variances.

Prediction variance or its square root, prediction standard error (PSE), is a useful criterion to measure how well a design covers a region. Indeed, if all the modeling assumptions are met, the PSE is proportional to the width of the confidence interval for the spatial predictions. Thus, a good spatial design will tend to make the prediction variance small at all points in a region of interest.

To interpret the design procedures based on subset selection, it is also helpful to define estimates for a linear combination of the field at a discrete set of locations. For example, let $\lambda(Z) = (1/n) \sum_{k=1}^{n} Z(\boldsymbol{u}_k)$ be the average value of the pollutant over a set of points. Let $\gamma_j = \mathrm{Cov}(\lambda(Z), Z(\boldsymbol{x}_j))$ and $\hat{\lambda} = \gamma^T K^{-1} \boldsymbol{z}$, where $\boldsymbol{z}^T = \{Z(\boldsymbol{x}_1), ..., Z(\boldsymbol{x}_n)\}$. Similar to the point predictions, this Kriging estimate is unbiased and has minimum variance among all linear estimators. The corresponding formula for the mean squared error has the same form as above:

$$E(\lambda - \hat{\lambda})^2 = \mathrm{Var}(\lambda) - \gamma^T K^{-1} \gamma. \tag{4.2}$$

3.3 Design Evaluation

Based on the discussion of the uses of the predicted ozone surface, it is reasonable to focus on selecting designs based on the prediction error either for individual points in the design region or for estimating a spatial average. To understand the graphical summaries appearing in the case studies, it is helpful to give an example using estimated prediction variance surfaces for two spatial designs.

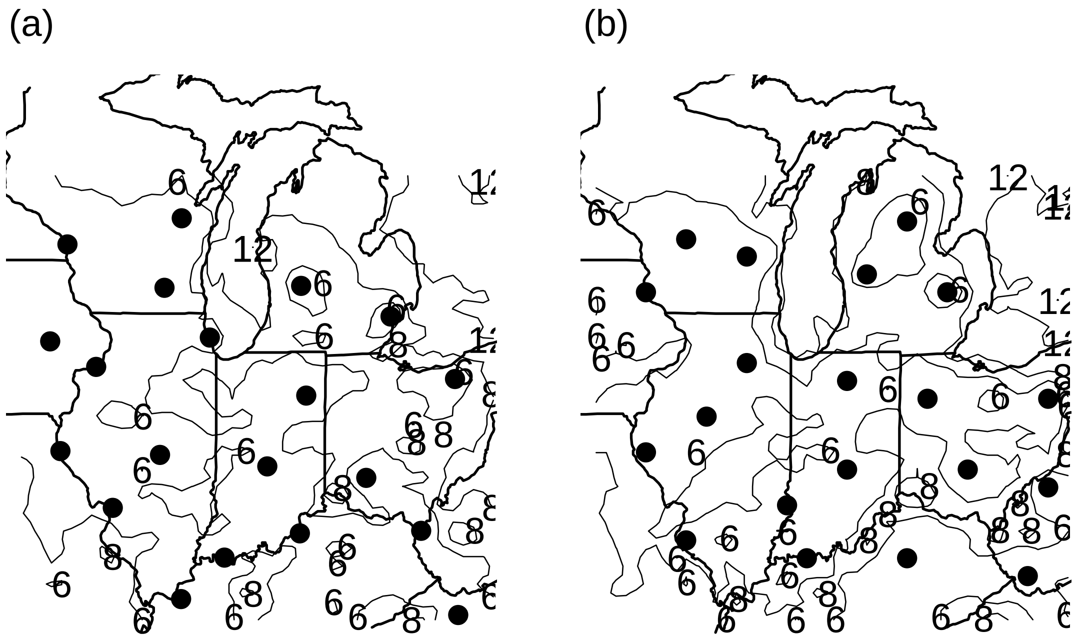

FIGURE 3. A comparison of the prediction errors for two networks consisting of 20 stations. Part (a) is based on selecting a subset of 20 locations from the existing network of 168 (see Figure 1). Part (b) is based on a network of 20 stations selected from a large uniform grid of points. In both cases, the designs were generated by minimizing the coverage criterion $C_{p,q}(\mathcal{D})$, as defined in Section 5.1, with $p = -5$ and $q = 5$.

The first illustration gives contours of prediction standard errors for a 20-point network constrained to be a subset from existing stations, and the second network is generated from a larger candidate set. As might be expected, areas of high prediction variance occur at the edges of the study region, with the constrained network tending to higher PSE.

Besides reporting summary statistics such as the average and maximum prediction variance over the design region, area/variance curves can be used to summarize the fraction of area in the design region with a prediction variance less than a particular value. Examples of these graphical summaries are given in Figure 4 and the difference between the two 20-point designs is shown in this

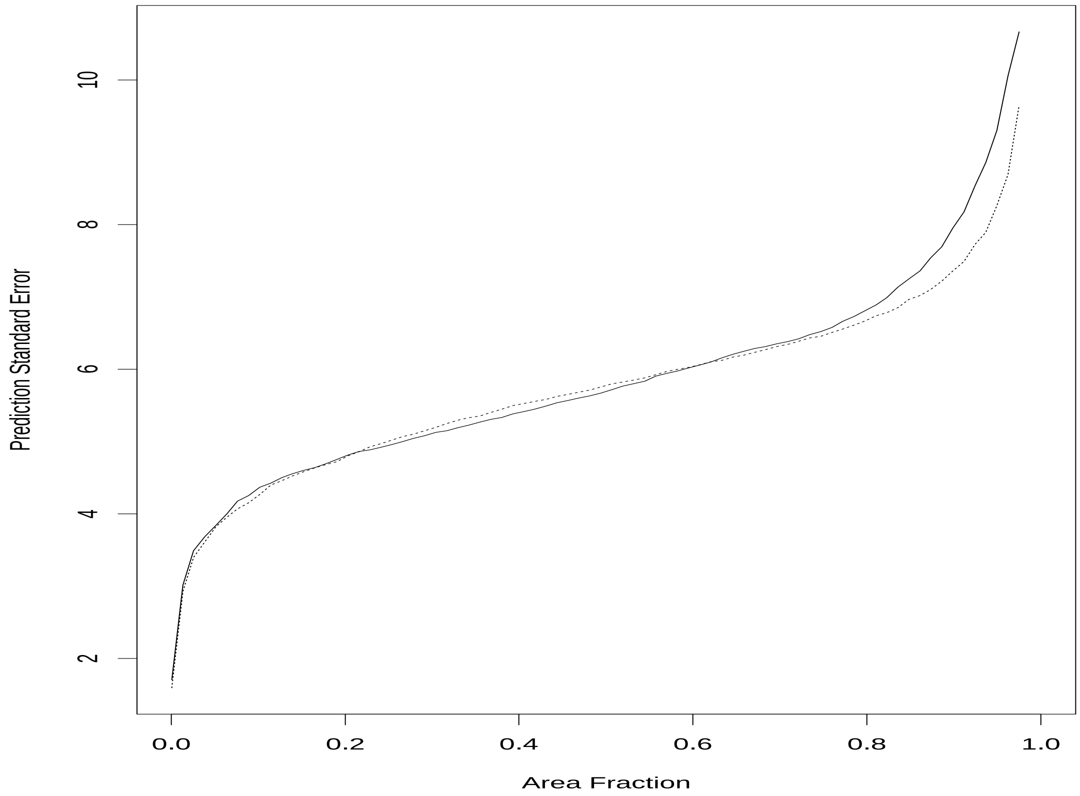

FIGURE 4. The prediction error as a function of the fraction of area of the design region. The solid line is performance of the 20-point constrained network from Figure 3a and the dashed line is performance of the 20-point network from Figure 3b.

Figure. In terms of area, the two networks exhibit similar performance over approximately 80% of the design region. However, for the remaining fraction of area, the PSEs diverge, with the maximum PSEs being substantially different: 10.7 ppb for the constrained network and 9.7 ppb for the unconstrained one.

Another evaluation criterion is based on the concept of thinning a network. Any subset of the network can be used to predict the average for the full (unthinned) network. The variance of this estimate is one measure of how well a particular subset can represent the full network and, conversely, of the efficiency of the original network.

4 Thinning a Small Urban Network

The first design problem studies how the small network for the Chicago urban area might be thinned in an efficient manner. In this study, we will concentrate on the accuracy of estimates of a spatial average; this has a direct connection with exposure measures or other spatial summaries. The basic outline of this work came about by an interest in doing something simple, avoiding the involved modeling of the spatial covariances, but exploiting the time replicates of the data. Also, we wanted to consider several different daily summaries of hourly ozone and their impact on the design choice.

4.1 Preliminary Results

The basic idea for thinning this small set of station information came from an admittedly *ad hoc* exercise using regression on the observed data. For each day, form the average of the daily summaries from the 20 stations and take the 20 individual station values as the "explanatory" variables that can be used to predict the full average. Now, consider the problem of finding a subset of J stations that gives the best linear prediction of the network average. Phrased in this way, the problem is just a regression model

$$Y = \alpha + X_J \beta_J + e,$$

where Y is the network average for each day. X_J is a matrix with the columns being the observed ozone values for a subset of J stations and the rows being the daily summaries for the stations on a given day. An independent variable in this regression equation is identified with a particular station; therefore, selecting a subset of variables is equivalent to identifying a subset of the monitoring network. Initially, we used the *leaps* subset procedure in S-PLUS to find the subsets of different sizes that minimized the residual sums of squares. Some of the resulting subsets of stations are plotted in the first column in Figure 6. Despite the apparent abuse of regression methodology, the results seemed promising. Of course, the fundamental question is: What kind of designs are produced and whether they are robust to several measures of design performance? In order to go any further, it was important to figure out what this procedure was actually doing in selecting a design.

4.2 Designs from Subset Selection

The *ad hoc* selection strategy described above can be justified by making the connection between regression subset selection and minimizing the prediction variance of the estimated average. The basic algorithms for doing this are well known (e.g., Cressie 1991, Sec. 5.9), but the convenience of using off-the-shelf regression software has not been emphasized.

Let λ be the average over the candidate points (or full network) and let $\hat{\lambda}$ be the estimate of this average based on a subset. A good design will estimate λ with a small variance. An important connection is that the mean squared error for $\hat{\lambda}$ can be expressed formally as the residual sum of squares (RSS) from a linear regression. Moreover, regression variables are identified with particular locations. Recall the formula

$$(Y - \hat{Y})^T (Y - \hat{Y}) = Y^T (I - X(X^T X)^{-1} X^T) Y = Y^T Y - Y^T X (X^T X)^{-1} X^T Y.$$

Thus, if the cross products satisfy the relationships $X^TX = K$, $X^TY = \gamma$, and $Y^TY = \mathrm{Var}(\lambda)$, the residual sum of squares from the regression of Y on X will be equal to the mean squared error of the Kriging estimate for λ, Equation (4.2). We can now interpret the regression exercise introduced at the beginning of this section. If the data are centered about their means, then $(1/n)K$ and $(1/n)\gamma$ are just the sample covariances based on the observed data. The subset selection finds the best subset of stations based on these choices for the covariances.

A general form for the regression subset selection problem is to find the solution to

$$\min_{\beta \in \mathcal{M}} (Y - X\beta)^T (Y - X\beta)$$

for some set $\mathcal{M}$. For usual subset selection, if J is the number of nonzero parameter values, then $\beta \in \mathcal{M}$ if $N - J$ values of the parameter vector are zero. We will refer to this estimate of the subset as the *leaps* procedure since it is computed by a leaps and bounds algorithm and is implemented in S-PLUS by the *leaps* function (Furnival and Wilson 1974; Becker *et al.* 1988). If X and Y are constructed as prescribed above, then minimizing the RSS is the same as minimizing the variance of $\hat{\lambda}$. Moreover, the best regression "subset" corresponds to the subset of candidate points for the spatial design. Given this correspondence, the best regression subset is then the optimal subset of locations with respect to predicting λ.

An alternative strategy for subset selection uses a different constraint set. The *lasso* procedure (Tibshirani 1996) requires that the sum of the absolute values of the parameter vectors be less than a fixed value, i.e.,

$$\mathcal{M} = \left\{ \beta : \sum_{j=1}^{N} |\beta_j| < t \right\}.$$

At first sight, it is not clear why the solution to this constrained problem is related to subset selection. Due to the properties of the absolute value function, however, the constrained solution will force some of the components of β to be identically zero. The number of zeros in the solution can be decreased by increasing t and one identifies the best subset with those variables that have nonzero parameter values. Of course, if t is made large enough, the full least squares solution will satisfy the constraint and thus will also be the *lasso* solution. It is helpful to scale the X variables so that the least squares solution has a constraint of one. Thus, t has a range of $[0, 1]$. Figure 5 illustrates the operation of the *lasso* when it is applied to pick subset designs for the Chicago urban network.

We have found that the *lasso* tends to favor points in the center of the design region and along the edges. The designs based on four- and five- point cluster stations in the center. For this particular example, the sequence of *lasso* subsets are also nested. This is a useful property if one wanted to identify a

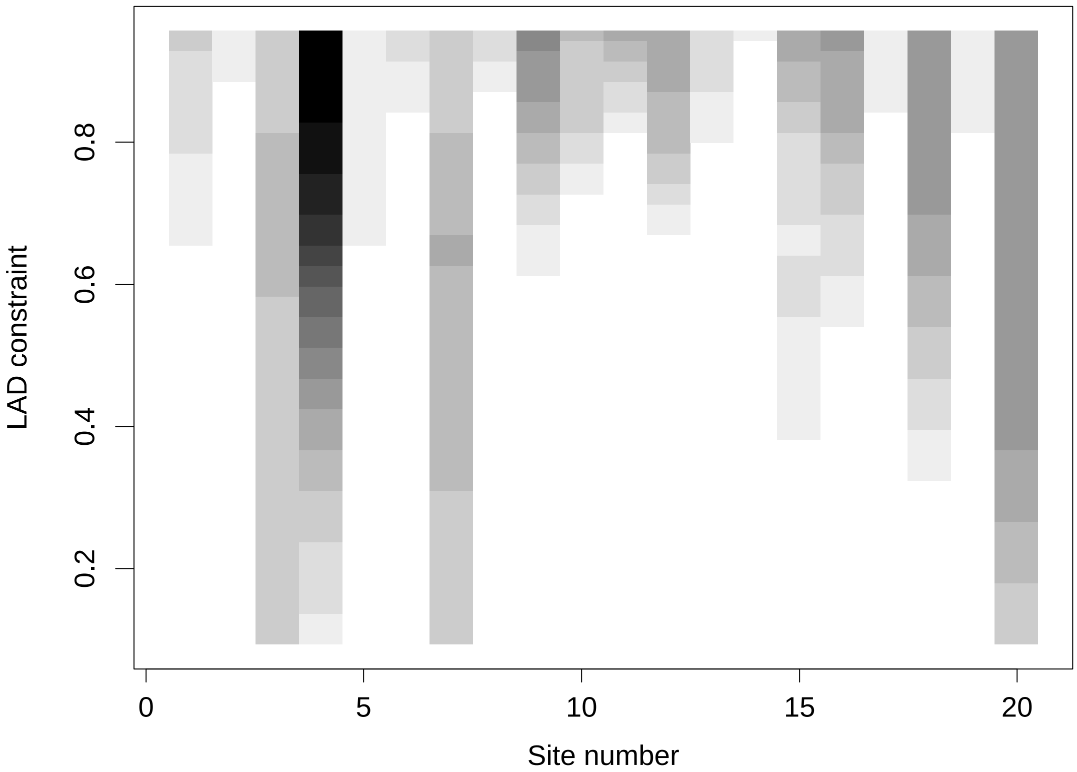

FIGURE 5. Estimated parameters for predicting the full network average using the *lasso* subset selection method. The least absolute deviations (LAD) constraint is the value of the *lasso* equation parameter t as a fraction of its value obtained with the ordinary least square (OLS) parameters. The absolute values of the parameter sizes β_j are represented as a gray-scale value and the site numbers along the horizontal axis can be matched with those in Figure 1b. Stations are eliminated from the solution by decreasing the LAD constraint from one, the OLS solution for all stations.

design consisting of a core of several stations that could then be expanded to include more locations.

Both the *leaps* and *lasso* procedures run fairly quickly on UNIX workstation and so the subset regressions can be computed rapidly. The *leaps* procedure is implemented as part of the interactive spatial design package DI (Nychka *et al.* 1996a); see Appendix B.

4.3 Results

For the period 1987–1991 there were 20 stations in operation in the Chicago urban area (see Figure 1b). The observations recorded for the summer of 1987 were used to estimate a sample covariance for ozone at the network locations and generate designs that were subsets of the full 20-station network. The remaining three years were used to validate the designs and calculate the prediction error. Note that without further modeling, the covariance is only available at the network locations and so the designs were compared based on how well they could estimate the full network average. Figure 6 reports some of the designs found by the *leaps* procedure and the *lasso* for the eight hour average daily summary. The third column of designs in this figure are based on a geometric criterion that is described in the next case study. The designs generated from the summer 1987 data were then validated using the subsequent three years; the results for the eight hour average are summarized in Figure 7. The most important feature is the rapid increase in the design efficiency as the number of sites increases. For example, a design containing five locations appears to do well in estimating the average of the full network.

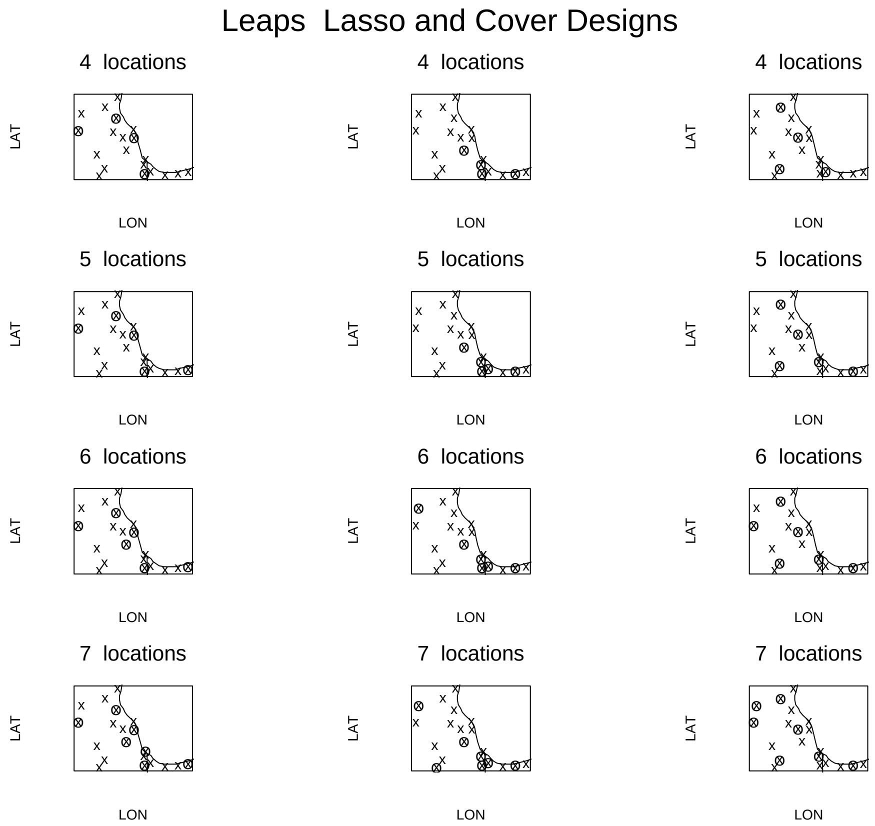

FIGURE 6. Subset designs for the Chicago urban area. The first and second columns are *leaps* and *lasso* designs, respectively, based on the daily 8-hour average over the period 6/3/87 to 8/30/87. In the third column are coverage designs.

Additional stations produce some decrease in the variance, but this improvement is marginal relative to the large decrease up to five stations. In quantitative terms, a five-station network has a prediction root mean squared error of 2.5 ppb. As a benchmark, the reader should note that the unconditional standard deviation for the full network average over the three year validation period is much higher: 16.1 ppb. Using the other daily summaries

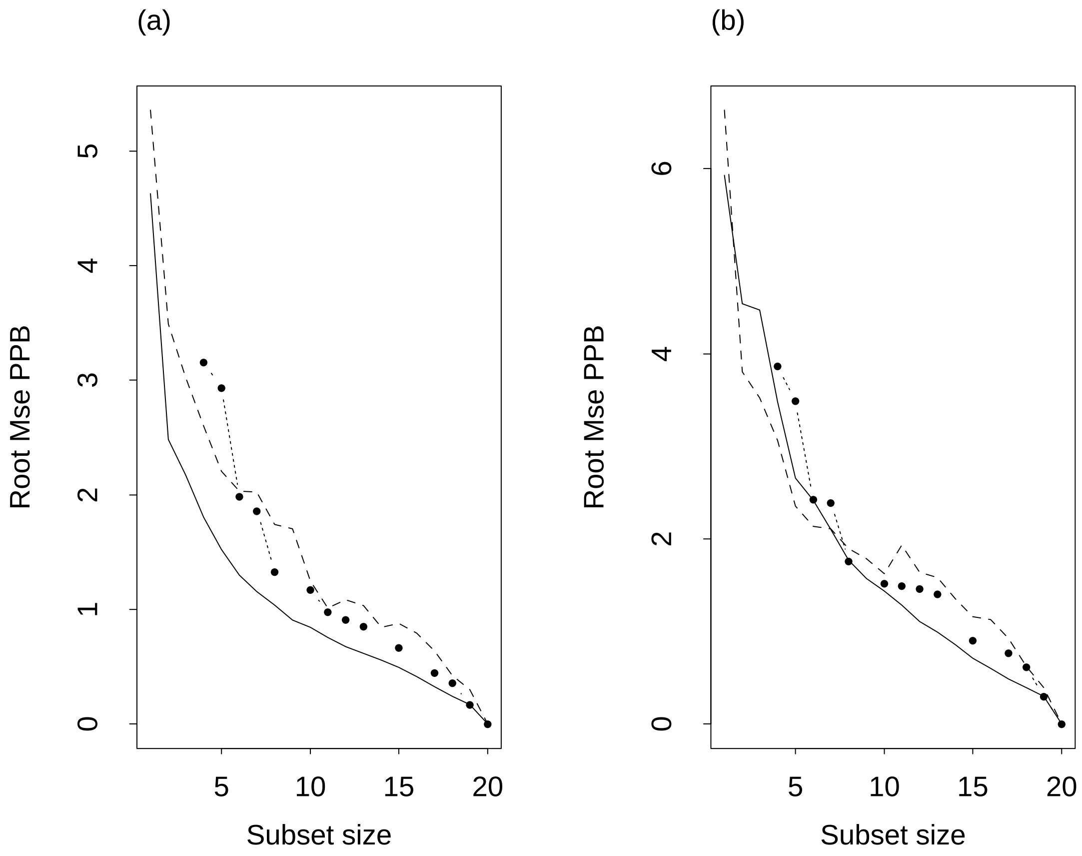

FIGURE 7. Root mean squared error for predicting the full Chicago area network average based on several different subset figures/designs. Plot (a) is the root mean squared error between the predicted average and the full network average observed for each day from 6/3/87 to 8/30/87. The data over this period have also been used to select the figures/designs. Plot (b) is the same comparison, but for the three subsequent summers, 1988, 1989, and 1990: *leaps* designs (solid), *lasso* (points), and coverage (dashed line).

produced different subsets, but similar patterns emerge when the designs were validated.

5 Adding Rural Stations to Northern Illinois

One issue often raised in discussions with EPA scientists was the adequacy of the monitoring network for measuring ozone in rural areas. Few stations are sited in rural areas and a natural question is how much improvement might be expected by adding more locations. A section of northern Illinois was used for testing the methodology of estimating nonstationary covariance functions and generating designs (Nychka *et al.* 1996b). This region coincides with an 18×6 rectangle of ROM grid cells and is part of the validation region in Davis *et al.* (1998) and was chosen because there was already some familiarity with the ozone field in this area. For this work, only the 8-hour average daily summary was used. Although the SUM06 and SUM08 statistics are also relevant, an average was chosen for convenience because of its distributional properties.

The designs for this problem and the larger region in the next section were created using a geometric (space-filling) criterion. This is very different than the usual spatial designs constructed using the covariance function. This choice was made largely on computational grounds. Recall that the prediction variance from Section 3 depends on the inverse of the covariance matrix of the observed points and, so, the objective function to find an optimal design will also include this term. In this case, the number of observations is the total number of stations (163) plus the extra stations (5–10) allocated to the rural study region. Given that the matrix inversion must be done for every evaluation of the objective function, one would expect the minimization of the average prediction error to be computationally intensive. This task is compounded by the fact that typically there are many local minima in the design criterion and, therefore, a conservative strategy is to repeat the optimization with a suite of starting designs.

Note that even though *construction* of the designs are not based on a covariance model, the *evaluation* of the designs are based on their performance for spatial prediction. So if a space-filling design exhibits acceptable properties for, say, average prediction variance, the corresponding A-optimal design must do as well or better. In this way, the space-filling designs in this study set upper bounds on what is possible.

5.1 Space-Filling Designs

A family of design criteria, independent of the assumed covariance function, may be based on geometric measures of how well a given design covers the design region. We will see that these designs are relatively easy to generate and can build in the natural geographic constraints common with environmental problems.

The basic simplification is to reduce the design region to a large, but finite, set of *candidate* points, $\mathcal{C}$. Let $\mathcal{D} \subset \mathcal{C}$ denote the set of N design points. A metric for the distance between any point $\boldsymbol{x}$ and a particular design is

$$d_p(\boldsymbol{x}, \mathcal{D}) = \left(\sum_{\boldsymbol{u} \in \mathcal{D}} ||\boldsymbol{x} - \boldsymbol{u}||^p \right)^{(1/p)} .$$

This metric measures how well the design *covers* the point $\boldsymbol{x}$. For $p < 0$, it is easy to show that $d_p(\boldsymbol{x}, \mathcal{D}) \to 0$ as $\boldsymbol{x}$ converges to a member of $\mathcal{D}$. This relation makes sense because one would expect the design points to cover themselves perfectly. For $q > 0$, an overall coverage criterion is an L_q average of coverage points in the design region

$$C_{p,q}(\mathcal{D}) = \left(\sum_{\boldsymbol{u} \in C} d_p(\boldsymbol{x}, \mathcal{D})^q \right)^{(1/q)} .$$

The coverage design for a given size is the subset that minimizes $C_{p,q}(\mathcal{D})$ for all $\mathcal{D} \subset \mathcal{C}$. In the limit as $p \to -\infty$ and $q \to \infty$, $C_{p,q}$ converges to a criterion used to define *minimax* space-filling designs. Explicitly, the minimax design minimizes the maximum of nearest-neighbor distances among points in the candidate set to those in the design. It helps to interpret this criterion by analogy to the problem of locating convenience stores in a city. The stores should be sited to be close to the customers and the *minimax* design solution locates stores so that the maximum distance that any customer has to travel to his closest store is minimized. Johnson *et al.* (1990) give some theoretical connections between space-filling designs and those based on prediction error for a spatial process. They show that as the correlations tend toward independence, the minimax designs and those based on minimizing the maximum prediction error are identical.

Coverage designs are generated using a simple "swapping" algorithm. One successively swaps each design point for a candidate point and determines whether the criterion is reduced. If so, the design point is replaced. This is continued until one cannot make any productive swaps. By keeping careful track of how the coverage criterion changes when two points are swapped, one can greatly reduce the number of computations. This is possible because many of the pairwise distances between design points and candidates do not change when a single swap is made. The resulting algorithm is simple and although not gradient based, it has the advantage of being readily implemented in a higher-level language such as S-PLUS. This makes it possible to consider fairly general forms of coverage metrics, because the metric need only be written as a high level function in S-PLUS. Also, no structure is assumed for the candidate set, so complicated and practical constraints can be built into the design. For example, in the Great Lakes region, it is important to keep the monitoring sites on land! This constraint is difficult to parameterize, however, due to irregular lake shorelines. The swapping strategy allows one to simply exclude any point over water from the candidate set. It should be noted that the swapping algorithm will always converge but is not guaranteed to produce a global optimum or even give the same answer from different initial designs. Thus, some care should be used in computing designs using this technique.

An important feature of the distance metric, $d_p(\cdot,\cdot)$ is that at least in a qualitative sense, it is similar to a prediction variance over the design region. Figure 8c is a contour plot of $d_p(\boldsymbol{x}, \mathcal{D})$ with $p = -1$ for a 20-point subset of the NAMS/SLAMS stations. This surface is zero at the design points and has local maxima at points that are in gaps of the network or at the boundaries. The general pattern follows the prediction variance surface based on an isotropic model (Figure 8a). Keeping in mind that the coverage criterion involves some form of averaging over the coverage surface, it is reasonable to expect that averages of the prediction variance should correspond to coverage. Specifically, with $q = 1$, one recovers an average value of the surface, and, of course, when

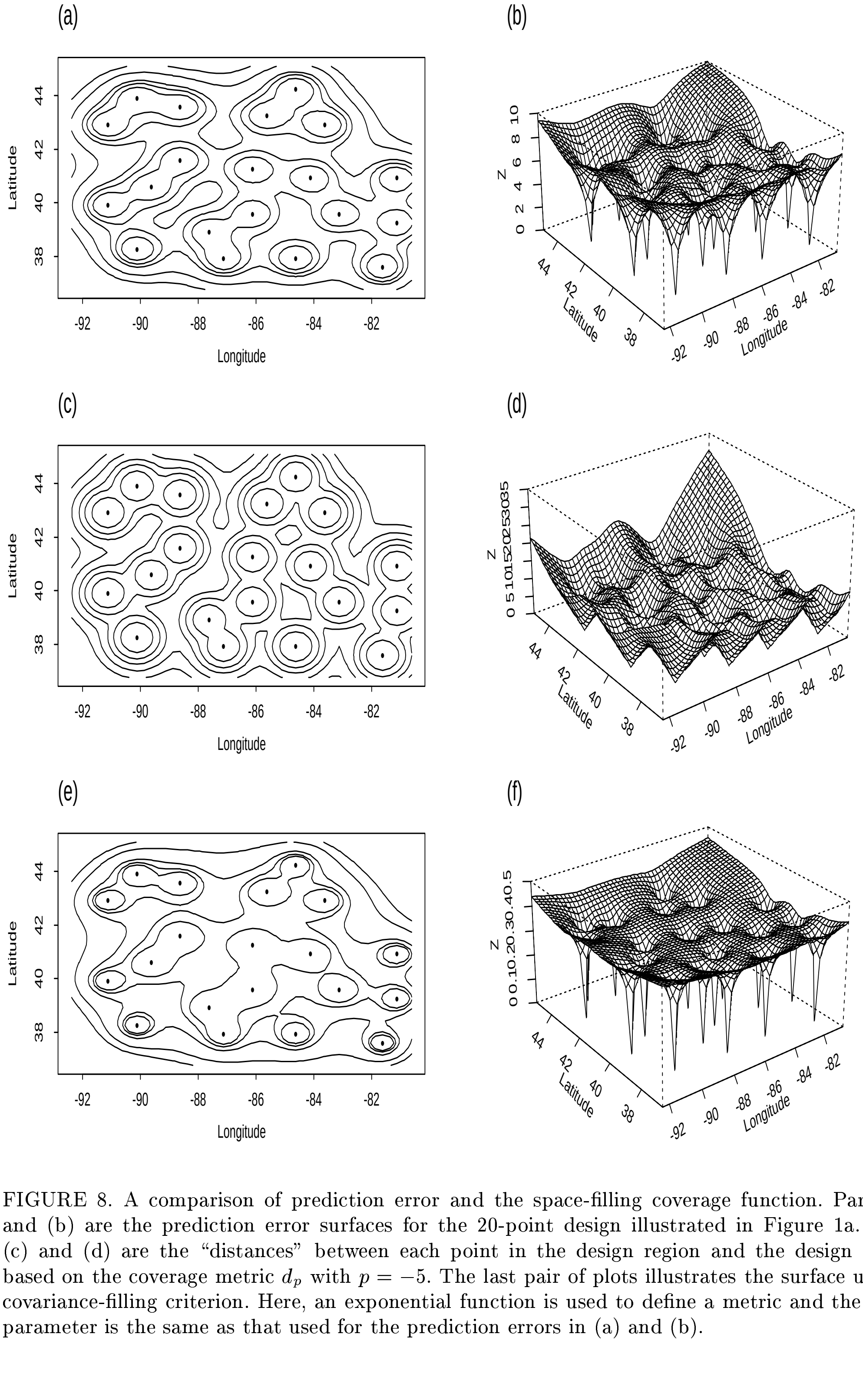

FIGURE 8. A comparison of prediction error and the space-filling coverage function. Parts (a) and (b) are the prediction error surfaces for the 20-point design illustrated in Figure 1a. Parts (c) and (d) are the "distances" between each point in the design region and the design points based on the coverage metric d_p with $p = -5$. The last pair of plots illustrates the surface using a covariance-filling criterion. Here, an exponential function is used to define a metric and the range parameter is the same as that used for the prediction errors in (a) and (b).

$q = \infty$, the coverage criterion is the maximum value of the surface.

One may also modify the coverage metric to be closer to a prediction variance and to reflect the correlation scales associated with a particular spatial process. Quite simply, we take $p = -1$ and replace the Euclidean norm in d_p with

$$k(\boldsymbol{x}, \boldsymbol{x}) - k(\boldsymbol{x}, \boldsymbol{u}) = \sigma(\boldsymbol{x})^2(1 - C(\boldsymbol{x}, \boldsymbol{u})),$$

where $C(\cdot, \cdot)$ is the correlation of the field at two locations. Specifically in the case of the isotropic exponential covariance, a *covariance-filling* criterion with $q = 1$ is

$$\left(\frac{1}{\sigma^2}\right) \sum_{u \in \mathcal{C}} \left[\sum_{x \in \mathcal{D}} \frac{1}{1 - e^{-||\boldsymbol{u} - \boldsymbol{x}||/\theta}} \right]^{-1}.$$

Figures 8e and 8f are contour and perspective plots, respective of the surface using this modification with a range parameter that is appropriate for ozone ($\theta = 161.3$). At least qualitatively, we see that this metric gives better agreement with the actual prediction variance surface.

5.2 Results for Rural Illinois

Without incorporating stations that border the region, designs will favor the edges of the region to the detriment of adding interior points. To adjust for this effect, the existing monitoring stations were added to the design as fixed points, and new points were selected from within the design region. The candidate set was taken to be an arbitrary grid of 432 (36×12) points within the test region. Figure 9 summarizes some of the results of filling in the rural area with additional monitoring sites. In this example, 5 and 10 points were added using the coverage criterion ($p = -5$ and $q = 5$). Because of the interest in filling in this rectangular subregion, the prediction error was only evaluated within this area. Overall, the addition of five points decreased the median prediction standard deviation by 25%, from 3.9 ppb to 3.0 ppb. The addition of 10 points provided a greater decrease, to 2.8 ppb.

6 Modifying Regional Networks

The last design study quantified the results of thinning, augmenting, and creating a new ozone network for the Great Lakes region in Figure 1a. The practical considerations here involved use of the current locations or additions to increase flexibility by changing the sites. The differences between these two choices is illustrated in Figure 2 for two 20-point networks. Here, the disadvantage of using the current monitors is that voids such as the one in northern Michigan, where there are no stations, cannot be filled in. In this study, stations outside this region were not included in the spatial prediction or used

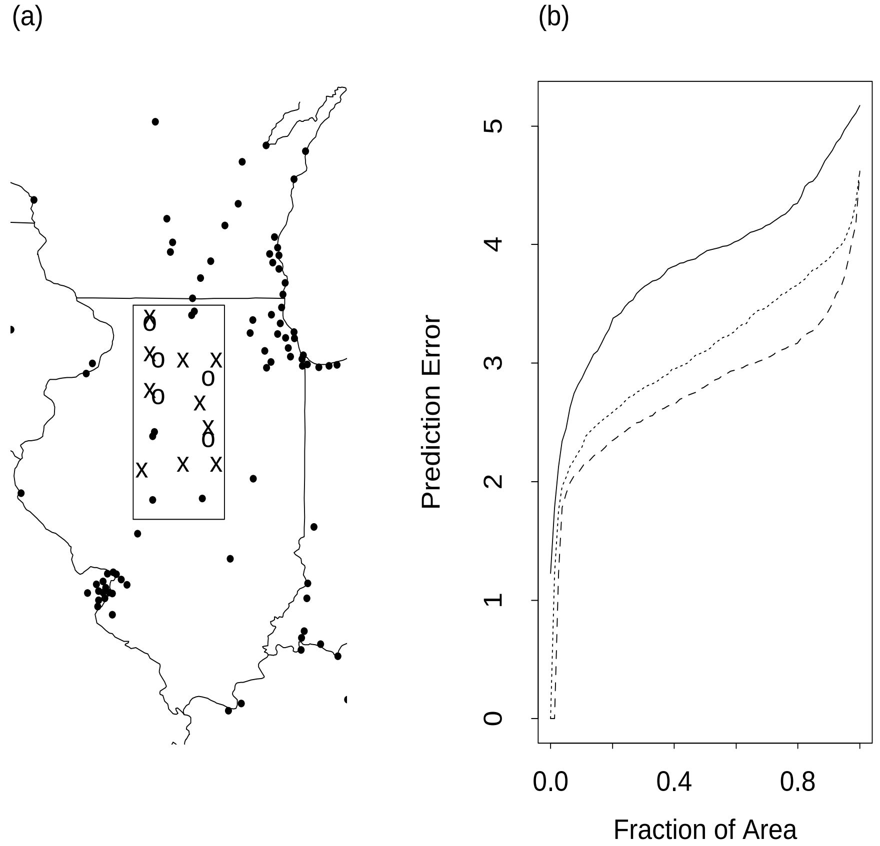

FIGURE 9. Coverage designs adding 5 (o) or 10 (x) additional locations within a rectangular area in northern Illinois. Part (b) summarizes the prediction standard errors for the original network of 168 stations (solid), adding 5 stations (short dashes), or adding 10 stations (long dashes).

as fixed points in the design. Thus the resulting designs could be sensitive to boundary effects. The hope was that designs with a larger number of points might reduce the edge effects caused by ignoring monitoring sites outside this region.

In evaluating the design, it was important to investigate the sensitivity of the results to the covariance function. This requirement is one reason why two different covariance models were fit to observational data and the ROM output.

6.1 Results for the Larger Midwest Network

The current network was increased in size (augmented) in the following way. The existing stations were set as fixed members of the design and the coverage criterion was minimized by adding a specific number of additional locations drawn from a candidate set. As a result, new network locations were placed at the edges of the region rather than filling in voids in the interior. Due to this placement, augmentation had little effect on the network performance except

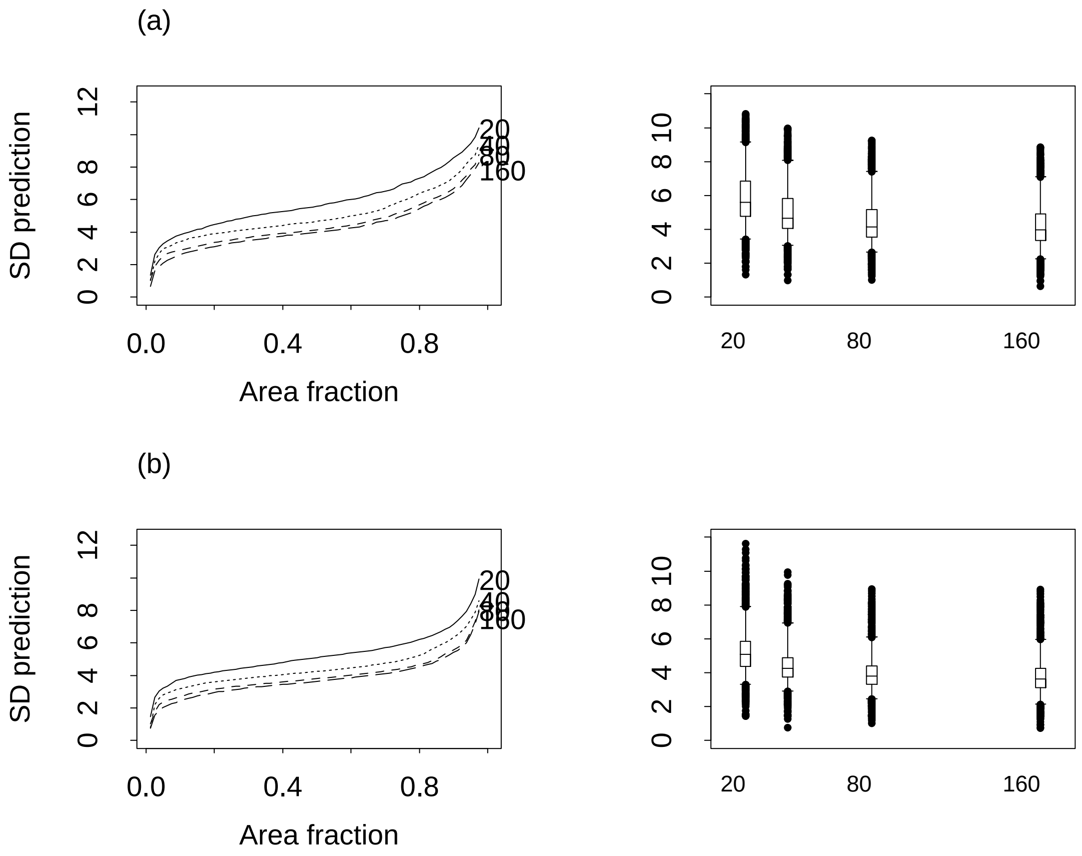

FIGURE 10. A comparison of coverage design ($p = -5$ and $q = 5$) performances for the Great Lakes region based on different network sizes and covariance models. Plot (a) evaluates the prediction error based on a nonstationary covariance function from the observed station data over the summer of 1987. The network sizes are 20, 40, 80, and 160 stations, and the curves are monotonically ordered in network size. Plot (b) is the same summary, except that the covariance function has been estimated from the ROM output. In the second column are boxplots of the distribution of the standard deviations.

along boundaries. One important exception was the placement of stations in northern Michigan, where the results suggested the need for additional sites.

Smaller or thinned networks were constructed by either considering subsets of the existing stations (constrained) or drawing from a larger uniform grid as the candidate set (unconstrained). For small numbers of stations, the performance of the constrained and unconstrained networks was similar. The differences in candidate sets became more apparent as the number of design points approached the total number of existing stations.

The variance/area plots (Figure 10) for different sized networks drawn from the existing NAMS/SLAMS sites were typical of the efficiency of a smaller monitoring network. For example, a subnetwork of 80 stations was comparable to the full network of 168 stations. The size of the prediction variances changed between the covariance models estimated from observational as opposed to

ROM output. This is not surprising given the different values for the range in the stationary component. However, the qualitative features relating the prediction variance to network size appeared to be fairly insensitive to different models for the covariance.

7 Scientific Contributions and Discussion

This research on spatial designs has shown that it is feasible to reduce or augment monitoring networks in an objective manner. Designs that are based solely on geometric criteria often do an excellent job of locating stations. In the absence of covariance information, this promises to be a useful design tool.

By validating against independent data in the case of the Chicago urban network, or by using an estimated covariance based on ROM output, it is possible to assess the prediction performance of the designs. One important physical feature of ozone is the fairly high correlations among concentrations at different locations. This can be quantified by the large range parameter (θ) estimated for the exponential portion of the covariance function. This property results in designed subnetworks having similar predictive performance to the full network. In particular, reducing the number of stations in the Great Lakes region to half the number of stations only inflates the median prediction standard error by about 10%. Practically speaking, a reduction in the number of monitors would have limited effect on the accuracy of the spatial extrapolations since the "error bars" would only be increased by this amount when half the network is used for prediction.

7.1 Future Directions

Trend detection is an important component in pollutant monitoring (Helsel and Hirsch 1988; Esterby *et al.* 1992; Styer 1994), and has two connections to network design. First, a network that yields more accurate and less biased estimates of regional summaries will reduce the standard error of a trend estimate. Second, a more extensive trend analysis can involve determining the differences in trends of the pollutant level at different locations. In this latter case, one could consider estimating a trend surface for an entire region. This spatial problem is different in character from just spatial prediction of the pollutant. Optimal solutions may require a different network because the spatial covariances for pollutant levels may be different from the the covariances for spatial trends. An example of thinning a wet deposition monitoring network with regard to trend estimates has been studied by Oehlert (1996).

This work clearly demonstrates the value of numerical models for understanding the properties of observational networks. In this role, it is important to validate these numerical models, not only in terms of mean levels but also in terms of covariance structure. This second-order validation for ROM is clearly needed to justify use of an estimated covariance function base on ROM output.

Also, the proposed designs attempt to spread points uniformly over the design region. Although this may be the best strategy for making the prediction variances small, the resulting designs will not contain much information about short- or medium-range covariance structure. Thus, the network may be efficient for prediction but have little power to check a covariance model. In particular, it may not be a good design with respect to validating second-order properties (such as covariances) of a numerical model. An important area of future work is to investigate designs that can also be used to estimate spatial covariance structure at different scales.

References

Alapaty, K., Olerud, D.T. and Hanna, A.F. (1995). Sensitivity of regional oxidant model predictions to prognostic and diagnostic meteorological fields. *Journal of Applied Meteorology* **34**, 1787–1801.

Becker, R.A., Chambers, J.M. and Wilks, A.R. (1988). *The New S Language: A Programming Environment for Data Analysis and Graphics.* Wadsworth & Brooks/Cole, Pacific Grove, CA.

Cressie, N.A.C. (1991). *Statistics for Spatial Data.* Wiley, New York.

Davis, J., Bailey, B., Nychka, D. and Vorburger, L. (1998). A comparison of the regional oxidant model with observational ozone measurements. Technical Report. National Institute of Statistical Sciences, Research Triangle Park, NC.

Esterby, S.R., El-Shaarawi, A.H. and Block, H.O. (1992). Detection of water quality changes along a river system. *Environmental Monitoring and Assessment* **23**, 219–242.

Furnival, G.M. and Wilson, R.W., Jr. (1974). Regression by leaps and bounds. *Technometrics* **16**, 499–511.

Helsel, D.R. and Hirsch, R.M. (1988). Discussion of "Applicability of the t-test for detecting trends in water quality variables." *Water Resources Bulletin* **24**, 201–204.

Johnson, M.E., Moore, L.M., and Ylvisaker, D. (1990). Minimax and maximin distance designs. *Journal of Statistical Planning and Inference* **26**, 131–148.

Lamb, R.G. (1983). A Regional Scale (100 km) Model of photochemical air pollution: Part 1 Theoretical foundation. Technical Report EPA-6000/3-83-035, Environmental Protection Agency, Washington, DC.