Optimal Network Design for Spatial Prediction, Covariance Parameter Estimation, and Empirical Prediction

Dale L. Zimmerman[1]

[1]Dale L. Zimmerman is Professor, Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA 52242 (E-mail: dzimmer@stat.uiowa.edu; Phone: 319-335-0818; Fax: 319-335-3017).

Abstract

Inferences for spatial data are affected substantially by the spatial configuration of the network of sites where measurements are taken. In this article, criteria for network design that emphasize the utility of the network for prediction (kriging) of unobserved responses assuming known spatial covariance parameters are contrasted with criteria that emphasize the estimation of the covariance parameters themselves. It is shown, via a series of related examples, that these two main design objectives are largely antithetical and thus lead to quite different "optimal" designs. Furthermore, a hybrid design criterion that accounts for the effect that the sampling variation of estimated spatial covariance parameters has on prediction is described and illustrated. Situations in which the hybrid optimal design resembles designs that are optimal with respect to each of the other two criteria are identified. An application to the optimal augmentation of an acid deposition monitoring network in the eastern U.S. is presented.

1 Introduction

Inferences for spatial data are affected substantially by the spatial configuration of the network of sites where measurements are taken. Because kriging is usually the ultimate objective of a geostatistical analysis, most authors who have considered network design for geostatistical data have emphasized the utility of designs for prediction. Invariably, these design considerations are made under an assumption that the second-order dependence (as characterized by either the covariance function or the semivariogram) is known, except possibly up to a scale parameter. Examples of this work include Bras and Rodriguez-Iturbe (1976), McBratney, Webster, and Burgess (1981), Yfantis, Flatman, and Behar (1987), Barnes (1989), and Cressie, Gotway, and Grondona (1990). The design criteria considered by these authors are generally either the average kriging variance (prediction error variance) or the maximum kriging variance over the region of interest, or some minor modification thereof.

Inasmuch as the second-order dependence of geostatistical data must in practice be estimated, however, a total emphasis on prediction under an assumption of known dependence parameters, with no regard for the utility of the design for the estimation of those parameters, may be ill-advised. Motivated by this concern, several authors have considered design criteria for good estimation of dependence parameters. Russo (1984) considered a design to be optimal if it minimized the dispersion of lags (i.e. distances between data locations) within the lag classes used for semivariogram estimation, while Warrick and Myers (1987) proposed a criterion that measures how well the lag distribution corresponding to a design conforms to a prespecified distribution. Zimmerman and Homer (1991), Müller and Zimmerman (1995, 1999), and Zhu and Stein (2004) focused instead on criteria more directly related to estimation quality, such as minimizing the determinant of the asymptotic covariance matrix of covariance parameter estimates.

This article has two purposes. The first is to demonstrate that the two design objectives of efficient prediction assuming known dependence on the one hand, and efficient estimation of dependence parameters on the other, are largely antithetical and thus often lead to very different optimal designs. The second purpose is to introduce a hybrid design criterion that emphasizes prediction but accounts for the additional prediction uncertainty incurred due to the estimation of covariance parameters, and show how designs that are optimal with respect to this criterion compare to designs that are optimal with respect to the other two types of criteria. Unlike the design objective of mean parameter estimation, none of the three design objectives considered herein has a well-developed optimal design theory; consequently, we will use a series of examples, some small and some large, to investigate these issues empirically. An application to an acid deposition network in the eastern U.S. will also show how designs that are optimal with respect to the three types of criteria compare to each other.

2 Design for Kriging with Known Dependence Parameters

Consider the following standard data-model framework for geostatistical data. We suppose that a continuous, spatially-varying quantity, Z, is to be observed at a predetermined number, n, of points $\mathbf{s}_1, \ldots, \mathbf{s}_n$ in a region of interest D. Let $Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)$ represent the observations taken at these points. These observations are modelled statistically as a spatially incomplete sample of one realization of a random field or stochastic process $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$. Assume further that the random field's mean is a linear function of unknown parameters, i.e. $E[Z(\mathbf{s})] = \sum_{i=0}^{p} \beta_i f_i(\mathbf{s})$ where $f_0(\mathbf{s}) \equiv 1$ and the remaining f_i's are known functions of observed covariates (for example, the spatial coordinates of $\mathbf{s}$). Also define the covariance function by $C(\mathbf{s}, \mathbf{u}) = \text{cov}[Z(\mathbf{s}), Z(\mathbf{u})]$. In this section we assume that the covariance function is known.

Under the model just described, the best (in the sense of minimizing the prediction error variance) linear unbiased predictor (BLUP) of $Z(\mathbf{s}_0)$, the realized but unobserved value of Z

at an arbitrary point $\mathbf{s}_0 \in D$, is the so-called universal kriging predictor, which is given by

$$\hat{Z}(\mathbf{s}_0) = [\mathbf{c} + \mathbf{X}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}(\mathbf{x} - \mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{c})]'\mathbf{\Sigma}^{-1}\mathbf{z}. \tag{1}$$

Here, $\mathbf{c}$ is the vector whose ith element is $C(\mathbf{s}_i, \mathbf{s}_0)$; $\mathbf{X}$ is the matrix whose ijth element is $f_{j-1}(\mathbf{s}_i)$ and which is assumed to have full column rank; $\mathbf{\Sigma}$ is the matrix whose ijth element is $C(\mathbf{s}_i, \mathbf{s}_j)$; $\mathbf{x}$ is the vector whose jth element is $f_{j-1}(\mathbf{s}_0)$; and $\mathbf{z}$ is the data vector with ith element $Z(\mathbf{s}_i)$. The minimized prediction error variance associated with $\hat{Z}(\mathbf{s}_0)$, also called the kriging variance, is given by

$$\sigma_K^2(\mathbf{s}_0) = \mathrm{var}(\hat{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0)) = C(\mathbf{s}_0, \mathbf{s}_0) - \mathbf{c}'\mathbf{\Sigma}^{-1}\mathbf{c} + (\mathbf{x} - \mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{c})'(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}(\mathbf{x} - \mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{c}). \tag{2}$$

Now let $\mathcal{S}$ be the set of all possible points where observations may be taken. We assume throughout that $\mathcal{S}$ is either D, a subregion of D, or a finite set of points in D. Suppose we wish to choose a subset of n points from $\mathcal{S}$ that are optimal, in some sense, for the purpose of kriging. As noted previously, the two most commonly used design criteria for this purpose are the average kriging variance

$$\frac{1}{|\mathcal{S}|}\int_D \sigma_K^2(\mathbf{s})\,d\mathbf{s} \quad \text{or} \quad \frac{1}{|\mathcal{S}|}\sum_{i \in \mathcal{S}} \sigma_K^2(\mathbf{s}_i)$$

(depending on whether $\mathcal{S}$ is a region or a finite point set) and the maximum kriging variance, $\max_{\mathbf{s} \in \mathcal{S}} \sigma_K^2(\mathbf{s})$. Here $|\mathcal{S}|$ is either the volume of $\mathcal{S}$ (if $\mathcal{S}$ is D or a subregion of it) or the number of points in $\mathcal{S}$. An n-point design is optimal with respect to one of these criteria if it minimizes the criterion over all possible n-point designs, each point taken from $\mathcal{S}$.

The top panel of Figure 1 displays optimal designs for the first in a series of toy examples. Throughout the entire series, $\mathcal{S}$ is a 5×5 square grid of points with unit spacing and $n = 4$. In this first example, observations are assumed to arise from a stationary process on $\mathcal{S}$ with unknown constant mean and an isotropic exponential covariance function given by $C(\mathbf{s}, \mathbf{u}) = \sigma^2 \exp(-\|\mathbf{s} - \mathbf{u}\|/\theta)$. For greater interpretability we put $\rho = \exp(-1/\theta)$ and reparameterize the covariance function as $C(\mathbf{s}, \mathbf{u}) = \sigma^2 \rho^{\|\mathbf{s}-\mathbf{u}\|}$; then ρ represents the correlation between adjacent points in $\mathcal{S}$. For each value of $\rho = 0.01, 0.02, \ldots, 0.99$, we obtained the 4-point design that minimizes

$$K(\boldsymbol{\theta}) = \max_{\mathbf{s} \in \mathcal{S}} \sigma_K^2(\mathbf{s}; \boldsymbol{\theta}) \tag{3}$$

by complete enumeration of all possible designs. There are $\begin{pmatrix} 25 \\ 4 \end{pmatrix}$=12650 four point arrangements in $\mathcal{S}$, but due to the symmetry of $\mathcal{S}$ with respect to rotations of 90, 180, and 270 degrees and with respect to reflections, many of these are equivalent to each other. In fact, there are 2012 equivalence classes, or distinct designs, and in our figures we shall always display just one member of a class.

We call a design that minimizes $K(\boldsymbol{\theta})$ for given $\boldsymbol{\theta}$ a "locally K-optimal" design (where the K is for kriging). The values of ρ considered here result in three distinct K-optimal designs. In all three of these the four design points tend to be relatively far apart, expanding outward slightly as the spatial correlation increases. This type of design is intuitively reasonable, as it prevents any prediction site from being too far from any design point. We note that if the mean was known, the K-optimal design would minimize the maximum distance between design points and prediction sites. The four-point design that achieves this would have an "interior square" arrangement, with each design point being along a diagonal of the 5×5 grid, equidistant from the center point and a corner point of $\mathcal{S}$. On the other hand, if all of our interest was on estimating an unknown mean as precisely as possible, the optimal design would be the "four-corner" design, i.e. the design with points at the four extreme corners of $\mathcal{S}$, because this design renders the observations at design points as nearly uncorrelated as possible. The locally K-optimal designs represent compromises between the interior square and four-corner designs.

The bottom panel of Figure 1 displays locally K-optimal designs for a situation exactly the same as that just described, except that the process is taken to have a planar mean function, i.e. $E[Z(\mathbf{s})] = \beta_0 + \beta_1 x + \beta_2 y$ (with $\beta_0, \beta_1, \beta_2$ unknown). Because the model matrix,

$\mathbf{X}$, is different in this case, it is not surprising that the locally K-optimal designs are different, at least for some values of ρ. (Note also that a few of the 2012 distinct designs, for example those with all four points in the same row or column, must be excluded from consideration when the mean is planar because they render the mean function parameters non-estimable.) It is interesting that the four-corner design is optimal when the spatial correlation is weak to moderate, but when the spatial correlation is sufficiently strong the optimal design contracts slightly inward to coincide with the locally K-optimal design for the constant mean case. The optimality of the four-corner design when the spatial correlation is weak to moderate is reasonable in light of the well-known fact that the four-corner design is D-optimal for estimating a planar mean, i.e. it minimizes the determinant of the covariance matrix of estimated planar mean parameters, when there is no spatial correlation.

Adding a nugget to the covariance function can also affect which designs are locally K-optimal, though it appears that the effect is slight. When the mean was constant and a nugget equal to either 25% or 50% of the overall sill was added, the three designs displayed in the top panel of Figure 1 remained locally K-optimal, but the range for ρ over which the design was optimal shifted very slightly (e.g. the lower endpoint for the third design moved from 0.24 to 0.22 as the nugget increased from 0% to 50% of the sill). In the case of a planar mean and the same two nuggets, the four-corner design was the K-optimal design uniformly over all values of ρ.

The limitations of studying a four-point design in a 5×5 design space are obvious, hence we also considered a series of larger examples in which we took $\mathcal{S}$ to be a 100×100 grid and $n = 50$. We considered the same mean functions (constant and planar) and nuggets (0, 25, and 50 per cent of the sill) as in the toy examples; however, since complete enumeration of all designs is not feasible in this case we limited ρ to three values (0.2, 0.5, and 0.8) and obtained designs by a simulated annealing algorithm, starting from a completely random initial design. (For details on the use of simulated annealing in a spatial design context, see van Groenigen and Stein (1998) or Lark (2002).) Here ρ represents the correlation, in the no-nugget case, between sites in $\mathcal{S}$ 20 grid spacings apart, which makes the correlation scale for the overall extent of $\mathcal{S}$ comparable to what it was for the toy examples. For each combination of mean function, nugget, and ρ the algorithm was repeated three times, from a different initial design each time, to verify reproducibility of the essential features of the design. The results were consistent across these repetitions, so for brevity we just display a design obtained from an abitrarily chosen one.

Figure 2 displays the locally K-optimal designs obtained for the larger example. Only the no-nugget case is included in Figure 2, as the effect on the designs of adding a nugget was imperceptible. The top panel, corresponding to the constant mean case, shows the locally K-optimal designs to be "regular," in the sense that the points are rather uniformly dispersed over the study region, with no two points close to each other. This is true regardless of the strength of spatial correlation, and is consistent with the toy example and with previously published findings (e.g. McBratney, Webster, and Burgess, 1981; Yfantis, Flatman, and Behar, 1987). In the planar mean case shown in the bottom panel, however, the locally K-optimal designs have a preponderance of points near the periphery of the design space; indeed the points tend to pile up in two opposite corners. This too is consistent with the toy example, and essentially the same interpretation made for the toy example is plausible here, that is, that prediction in the planar mean model is enhanced substantially by improved estimation of the slope parameters, which will occur for designs with more high-leverage (peripheral) points.

We also examined designs that minimize the average (rather than the maximum) kriging variance over the grid points, both for the toy example and the larger example. While these designs were not necessarily identical to the locally K-optimal designs shown in Figures 1 and 2, qualitatively they were very similar.

Our overall conclusion is that the strength of the spatial correlation and the size of the nugget have relatively little effect on the K-optimal design, but that the choice of mean function has a tremendous effect. When the mean is constant, the best designs are very regular; when the mean is planar the best designs have most of their points along the periphery of the design space.

3 Design for Estimation of Covariance Parameters

Now let us suppose that the covariance function is known only up to a parameter vector $\boldsymbol{\theta}$, and that we wish to design a network that is optimal, in some sense, for estimating $\boldsymbol{\theta}$. Of course, such a design may depend on the particular estimator, $\hat{\boldsymbol{\theta}}$, and on the particular measure of that estimator's quality that are to be used. One design approach, which was proposed by Müller and Zimmerman (1999), is to maximize the determinant of the information matrix associated with a nonlinear generalized least squares estimator of $\boldsymbol{\theta}$. A related approach, which was used by Zhu and Stein (2004) and is also used here, is to assume that the underlying process is Gaussian and maximize the determinant of the information matrix associated with either the maximum likelihood (ML) estimator or residual maximum likelihood (REML) estimator of $\boldsymbol{\theta}$. The rationale for these criteria is that the inverse of the information matrix is, under regularity conditions, the asymptotic covariance matrix of the corresponding estimator of $\boldsymbol{\theta}$; thus for sufficiently large samples it should provide a reasonable approximation to the determinant of the intractable mean squared error (MSE) matrix

$$\mathbf{M}(\boldsymbol{\theta}) = E[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'].$$

Let $\mathbf{I}_{ML}(\boldsymbol{\theta})$ and $\mathbf{I}_{REML}(\boldsymbol{\theta})$ represent the information matrices associated with the ML and REML estimators of $\boldsymbol{\theta}$, respectively. The ijth element of $\mathbf{I}_{ML}(\boldsymbol{\theta})$ is given by

$$\frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_i}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_j}),$$

and the ijth element of $\mathbf{I}_{REML}(\boldsymbol{\theta})$ is given by

$$\frac{1}{2}\mathrm{tr}(\mathbf{P}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_i}\mathbf{P}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_j}),$$

where $\mathbf{P} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}$. For brevity we shall limit our presentation to designs that are optimal with respect to ML estimation; results corresponding to REML estimation are quite similar except when the size of the design is small and the number of mean parameters is large. Note that maximizing $|\mathbf{I}_{ML}(\boldsymbol{\theta})|$ is equivalent to minimizing $|\mathbf{B}_{ML}(\boldsymbol{\theta})|$, where $\mathbf{B}_{ML}(\boldsymbol{\theta}) = \mathbf{I}_{ML}^{-1}(\boldsymbol{\theta})$, the large-sample approximation to $\mathbf{M}_{ML}(\boldsymbol{\theta})$. We shall use $|\mathbf{B}_{ML}(\boldsymbol{\theta})|$ as a design criterion, calling a design that minimizes it for a particular $\boldsymbol{\theta}$ a locally CP-optimal design (where CP refers to covariance parameter). For economy of expression we shall cease using the subscript ML and will simply denote $\mathbf{B}_{ML(\boldsymbol{\theta})}$ by $\mathbf{B}(\boldsymbol{\theta})$.

To explore the kinds of designs that are optimal with respect to this criterion, we begin again with the series of toy examples introduced in the previous section. We consider only the case of a constant mean function, since $\mathbf{B}(\boldsymbol{\theta})$ is invariant to the choice of mean function. (As noted previously, estimability considerations require that a few designs be excluded when the mean is planar, but this has minimal effect on which designs are optimal.) Also, $\mathbf{B}(\boldsymbol{\theta})$ is invariant to translations of the design; hence, the aforementioned collection of equivalence classes of designs with respect to rotations and reflections can be reduced to an even smaller collection of equivalence classes, and again we will display only one member of a given class.

Figure 3 shows the seven 4-point locally CP-optimal designs obtained by taking $\rho = 0.01, 0.02, \ldots, 0.99$, for a model with no nugget. In all of these designs, the points are clustered to some degree. When ρ is sufficiently small the design is a compact cluster, but as ρ increases the cluster tends to elongate into a linear "strand" and may disperse somewhat. This is in marked contrast to the characteristics of locally K-optimal designs. It would appear that the presence of some small lags in the design is very useful for precise estimation of covariance parameters but not for prediction (with known covariance parameters). Indeed, these two design objectives are largely antithetical; for instance, the K-optimal design when

$\rho = 0.5$ and the mean is constant (the design in the rightmost panel in the top row of Figure 1) is the fourteenth worst design (out of 2012) for minimizing $|\mathbf{B}(\boldsymbol{\theta})|$, and the CP-optimal design under the same circumstances (middle panel in top row of Figure 3) is the fifteenth worst for minimizing the K-optimality criterion. This is established more definitively by examining the rank correlations between $K(\boldsymbol{\theta})$ and $|\mathbf{B}(\boldsymbol{\theta})|$ over the entire population of distinct four-point designs (see Table 1). These correlations, denoted by $r_{K,CP}$, are negative except when the spatial correlation is very strong.

Locally CP-optimal designs for a model with a nugget equal to zero are displayed in Figure 4. Note that this model is not equivalent to the no-nugget model when using $|\mathbf{B}(\boldsymbol{\theta})|$ as a criterion, since $\mathbf{I}(\boldsymbol{\theta})$ for this model is of larger dimensions (3×3 versus 2×2) and some of its terms involve partial derivatives of the covariance matrix with respect to the nugget. Figure 5 displays locally CP-optimal designs for a model with nugget equal to 50% of the sill. We see that the locally CP-optimal designs corresponding to the zero-nugget model are quite similar to their counterparts for the no-nugget model. The 50% nugget case is not greatly different, but there is some indication that the clusters of points are less inclined to be aligned linearly.

It may be argued that the appropriateness of using a criterion based on asymptotic theory in the context of four-point designs is highly questionable. Note, however, that the usefulness of $|\mathbf{B}(\boldsymbol{\theta})|$ as a design criterion relies not on its quality as an approximation to $|\mathbf{M}(\boldsymbol{\theta})|$ but on the monotonicity of its relationship with $|\mathbf{M}(\boldsymbol{\theta})|$. If this relationship is monotonic, then the two criteria will rank designs in the same order. In order to investigate the extent to which this relationship is monotonic, we obtained an estimate, $\hat{\mathbf{M}}(\boldsymbol{\theta})$, of the ML estimator's finite-sample MSE matrix for a given $\boldsymbol{\theta}$, and used its determinant as a criterion in place of $|\mathbf{B}(\boldsymbol{\theta})|$. It is not computationally feasible to do this for all 99 values of ρ and all 2012 distinct four-point designs, hence we considered only three values of ρ (0.2, 0.5, and 0.8) and only 102 four-point designs for each ρ. These designs were obtained by first ranking all 2012 designs according to their value of the K-optimality criterion for that ρ, and then selecting the designs in positions 1, 21, 41, ..., 2001, 2012 in the list. The estimation of the MSE matrix was accomplished by simulating 1000 realizations of a Gaussian process over the design space, and obtaining, from the subset of simulated data corresponding to each design, the ML estimator of $\boldsymbol{\theta}$, say $\hat{\boldsymbol{\theta}}_{ij}$ (where i indexes the simulations and j indexes the designs). From these estimators we calculated

$$\hat{\mathbf{M}}_j(\boldsymbol{\theta}) = \frac{1}{1000}\sum_{i=1}^{1000}(\hat{\boldsymbol{\theta}}_{ij} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_{ij} - \boldsymbol{\theta})',$$

which estimates $\mathbf{M}_j(\boldsymbol{\theta})$, the MSE matrix corresponding to the jth design. The rank correlations between $|\hat{\mathbf{M}}_j(\boldsymbol{\theta})|$ so obtained and the jth design's value of $|\mathbf{B}(\boldsymbol{\theta})|$ ranged from 0.37 when $\rho = 0.2$ to 0.50 when $\rho = 0.8$. More to the point, the particular design that minimized $|\mathbf{B}(\boldsymbol{\theta})|$ also generated the smallest value of $|\hat{\mathbf{M}}_j(\boldsymbol{\theta})|$ when $\rho = 0.5$, and the second smallest value of $|\hat{\mathbf{M}}_j(\boldsymbol{\theta})|$ when $\rho = 0.2$ and $\rho = 0.8$. Based on this, and in light of the sampling variation of $\hat{\mathbf{M}}_j(\boldsymbol{\theta})$, it seems reasonable to conclude that the general features of designs that have small values of the determinant of the ML estimator's exact mean squared error matrix will coincide with those that have small values of $|\mathbf{B}(\boldsymbol{\theta})|$.

We also investigated the use of $|\mathbf{B}(\boldsymbol{\theta})|$ as a design criterion for the larger example introduced in the previous section. For this example we deem the sample size (which is 50) to be sufficiently large for the ranking of designs with respect to $|\mathbf{B}(\boldsymbol{\theta})|$ to be a very good approximation to the ranking with respect to $|\mathbf{M}(\boldsymbol{\theta})|$. Figure 6 displays locally CP-optimal designs corresponding to constant and planar mean functions and to the same values of ρ and nugget used previously for the larger example. Looking first at the case of a constant mean and no nugget, we see that the three locally optimal designs all have a substantial proportion of points regularly spaced along the boundary, and the remainder are in a small number of relatively large clusters of contiguous points, the clusters themselves located rather haphazardly in the design space. Some of these clusters are linear strands or have linear segments, a phenomenon we also observed in the toy example. The interior of the design space is rather empty, save for the occasional cluster. This pattern of points has the effect of creating many small lags and many large lags. The case of a planar mean and no nugget is not greatly different, though there appear to be a few less points within clusters and a few more points regularly spaced within the interior. The designs are significantly different, however, when the nugget is 50% of the sill. In both the constant and planar mean cases with a 50% nugget, the locally optimal design consists of clusters of a small number (2-5) of points. As ρ increases, so does the average number of points per cluster and the dispersion of those points within clusters. Furthermore, when the mean is constant, the vast majority of clusters are located near the design space's periphery and are regularly spaced along it, whereas when the mean is planar most clusters are regularly spaced within the interior (although there is a cluster in each corner).

The overall conclusion we make from these results is that for purposes of good estimation of covariance parameters, a design should have a much greater number of small lags than will occur in a K-optimal design or, for that matter, in a completely random arrangement of points. The design should have many large lags as well. Such a distribution of lags is achieved by a pattern of regularly spaced clusters, most of which lie along the periphery of the design space. Another conclusion is that the mean function and nugget have a much greater effect than the strength of spatial correlation on CP-optimal design.

4 Design for Kriging with Estimated Covariance Parameters

Neither of the design approaches discussed in the previous two sections fully address the prediction problem of most practical interest, which is to predict the unobserved value of $Z(\mathbf{s}_0)$ when the spatial dependence parameters are unknown. The standard predictor in this situation, known as the E-BLUP (empirical BLUP), is given by an expression identical to

(1) but with $\mathbf{\Sigma}$ evaluated at $\hat{\boldsymbol{\theta}}$, an estimate of $\boldsymbol{\theta}$, rather than at the hitherto-assumed-known $\boldsymbol{\theta}$. That is, if we let $p_1(\mathbf{z}, \mathbf{s}_0; \boldsymbol{\theta})$ be the known-covariance-parameter BLUP at $\mathbf{s}_0$ given by (1), then the E-BLUP at $\mathbf{s}_0$ is given by $p_2(\mathbf{z}, \mathbf{s}_0) = p_1(\mathbf{z}, \mathbf{s}_0; \hat{\boldsymbol{\theta}})$.

The sampling distribution of the E-BLUP is only partially understood. Remarkably, it is known to be unbiased under very weak conditions; however, an explicit expression for the variance of the E-BLUP's prediction error, i.e.

$$m_2(\mathbf{s}_0, \boldsymbol{\theta}) = \mathrm{var}(p_2(\mathbf{z}, \mathbf{s}_0) - Z(\mathbf{s}_0)),$$

has not been derived. In practice $m_2(\mathbf{s}_0, \boldsymbol{\theta})$ is usually estimated by substituting $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ in (2), the expression for the variance of the BLUP's prediction error. Harville and Jeske (1992) and Zimmerman and Cressie (1992) considered the performance of this method for estimating $m_2(\boldsymbol{\theta})$, and also proposed the following approximation:

$$m_2(\mathbf{s}_0, \boldsymbol{\theta}) \doteq \sigma_K^2(\mathbf{s}_0; \boldsymbol{\theta}) + \mathrm{tr}[\mathbf{A}(\mathbf{s}_0, \boldsymbol{\theta})\mathbf{H}(\boldsymbol{\theta})] \tag{4}$$

where $\mathbf{A}(\mathbf{s}_0, \boldsymbol{\theta}) = \mathrm{var}[\mathbf{d}(\mathbf{z}, \mathbf{s}_0; \boldsymbol{\theta})]$, $\mathbf{d}(\mathbf{z}, \mathbf{s}_0; \boldsymbol{\theta}) = \partial p_1(\mathbf{z}, \mathbf{s}_0; \boldsymbol{\theta})/\partial \boldsymbol{\theta}$, $\mathbf{H}(\boldsymbol{\theta})$ is a matrix that either equals or approximates the MSE matrix $\mathbf{M}(\boldsymbol{\theta})$, and $\mathrm{tr}(\cdot)$ denotes the trace of a square matrix. If $\boldsymbol{\theta}$ is estimated by maximum likelihood, then a natural choice for $\mathbf{H}(\boldsymbol{\theta})$ is $\mathbf{B}(\boldsymbol{\theta})$. Zimmerman and Cressie (1992) and later Abt (1999) found this approximation to be much closer than $\sigma_K^2(\mathbf{s}_0, \boldsymbol{\theta})$ to $m_2(\mathbf{s}_0, \boldsymbol{\theta})$ in many, though not all, situations they considered.

Accordingly, here we investigate optimal design for kriging with estimated covariance parameters using as a criterion the maximum value of the E-BLUP's asymptotic approximate prediction error variance over all sites in $\mathcal{S}$,

$$E(\boldsymbol{\theta}) = \max_{\mathbf{s} \in \mathcal{S}} \{\sigma_K^2(\mathbf{s}; \boldsymbol{\theta}) + \mathrm{tr}[\mathbf{A}(\mathbf{s}, \boldsymbol{\theta})\mathbf{B}(\boldsymbol{\theta})]\}. \tag{5}$$

We call a design that minimizes this criterion for a given $\boldsymbol{\theta}$ a locally EK-optimal design (where EK stands for empirical kriging). Observe that criterion (5) combines a measure of quality of design with respect to prediction with known covariance parameters, namely $\sigma_K^2(\mathbf{s}; \boldsymbol{\theta})$, with a measure of quality with respect to covariance parameter estimation, tr$[\mathbf{A}(\mathbf{s},\boldsymbol{\theta})\mathbf{B}(\boldsymbol{\theta})]$. We may therefore expect that EK-optimal designs may lie somewhere between the extremes of K-optimal and CP-optimal designs.

Figures 7, 8, and 9 display locally EK-optimal designs for the toy example, in the case of a constant mean and a model with no nugget, 0% nugget, and 50% nugget respectively. In all three cases of nugget, the optimal design is highly clustered when ρ is small and thus is identical or bears a resemblance to the corresponding CP-optimal design. In the no-nugget and 0% nugget cases, however, the design points become dispersed as ρ increases and and the EK-optimal design begins to resemble or coincide with the corresponding K-optimal design. In the case of a 50% nugget, the attenuation of clustering is much less pronounced, so that the EK-optimal design resembles or coincides with the CP-optimal design for all values of ρ we considered. Further corroboration of this behavior is obtained by examining the rank correlations between $K(\boldsymbol{\theta})$ and $EK(\boldsymbol{\theta})$, denoted by $r_{K,EK}$, and the rank correlations between $|\mathbf{B}(\boldsymbol{\theta})|$ and $EK(\boldsymbol{\theta})$, denoted by $r_{CP,EK}$, across the population of distinct four-point designs (see Table 1). When the mean is constant and there is no nugget, $r_{K,EK}$ moves from very close to -1 when the spatial correlation is weak, to very close to $+1$ when the spatial correlation is strong. In the case of a 50% nugget, $r_{K,EK}$ is negative for all values of ρ considered, though this negative correlation does weaken as ρ increases. Note that the behavior of $r_{CP,EK}$ is non-monotonic, but that it is close to 1 over all values of ρ in the case of a 50% nugget.

The behavior of the locally EK-optimal designs in the case of a planar mean is generally quite similar to that for the constant mean case, so we will not elaborate on it.

The quality of approximation (5) in this toy example is again open to question, indeed perhaps even more so than the approximation of the ML estimator's finite-sample covariance matrix by $\mathbf{B}(\boldsymbol{\theta})$, due to the additional level of approximation introduced in (3). Consequently, we performed another simulation study, which was similar to the previous one but went a step further. The extra step was to use the ML estimate of $\boldsymbol{\theta}$ computed from each simulation to obtain the E-BLUP at all 25 grid sites, then square the difference of these E-BLUPs and the simulated data at corresponding sites, and finally average these over the 1000 simulations. This average squared difference at each site is an estimate of the E-BLUP's finite-sample prediction error variance at that site, and its maximum over $\mathcal{S}$ can be used instead of (5) as a design criterion. We computed rank correlations, over the 102 designs, between (5) and its estimated finite-sample counterpart. These correlations ranged from 0.63 when $\rho = 0.2$, to 0.97 when $\rho = 0.8$. Based on this, and in light of the sampling variation in the estimates of the E-BLUPs' finite-sample prediction error variances, it seems reasonable to conclude that, even in a case with a sample as small as in our toy example, the general features of designs that have small values of $EK(\boldsymbol{\theta})$ will coincide with those that minimize the maximum of the E-BLUP's finite-sample prediction error variances.

Finally we consider EK-optimal designs in the context of the larger example. Figure 10 shows locally EK-optimal designs corresponding to the same combinations of ρ, mean function, and nugget used previously for this example. We see that in all cases the EK-optimal design is similar to its K-optimal counterpart (Figure 2) on a global scale, i.e. in terms of overall spatial coverage and extent, but that it has a few small clusters in it, thereby borrowing the small-scale features of its CP-optimal counterpart (Figure 6). Thus, we conclude that a modest number of very small lags in the design, while of no use for prediction with known covariance parameters, is beneficial for empirical prediction.

5 Application: Optimal Augmentation of an Acid Deposition Monitoring Network

As a final example, and one which is perhaps more representative of the types of design problems actually encountered in practice, we consider the optimal augmentation of one site to an existing acid deposition network. We base our analysis on data collected in 1983 at the 19 cities of the Utility Acid Precipitation Study Program (UAPSP) network in the eastern U.S. (see Figure 11). Several analytes were measured in the study, but we restrict attention to total annual hydrogen ion deposition. The data (in units of μ mole H^+/ cm^2) and their corresponding latitude-longitude coordinates are given by Cressie, Gotway, and Grondona (1990), who also considered a one-site augmentation problem for this network, but one that assumed known covariance parameters and used a different prediction-based criterion than we use here. Zimmerman and Homer (1991) likewise used these data to obtain one-site augmentations of the network that were optimal with respect to estimation of certain functions of the covariance parameters.

In practical situations in which an existing network may be augmented, it is often the case, due to political or logistical considerations, that the candidates for addition are rather few. Accordingly, we suppose that the site to be added to the network comes from the set of eleven candidates displayed together with the existing sites in Figure 11. Incidentally, this is the same set of candidates used in both references cited in the previous paragraph. Furthermore, based on analyses presented in these previous works, we adopted a Gaussian random field model for these data, with quadratic mean function

$$E[Z(\mathbf{s})] = \beta_0 + \beta_1 x + \beta_2 y + \beta_3 x^2 + \beta_4 xy + \beta_5 y^2$$

and isotropic exponential covariance function $C(r) = \sigma^2 \exp(-r/\theta)$. Here, r represents great-circle distance in units of 100 miles. ML estimates (and their estimated standard errors) of the quadratic mean parameters were as follows: $\hat{\beta}_0 = 5.24$ (0.40), $\hat{\beta}_1 = 0.203$ (0.071), $\hat{\beta}_2 = 0.056$ (0.13), $\hat{\beta}_3 = -0.047$ (0.016), $\hat{\beta}_4 = 0.057$ (0.024), $\hat{\beta}_5 = -0.057$ (0.031). ML estimates of the covariance parameters were $\hat{\sigma}^2_{ML} = 0.901$ and $\hat{\theta}_{ML} = 0.517$.

Assuming that the fitted model is the correct model, we can determine which of the 11 candidate sites for augmentation is optimal with respect to each of the three design criteria we have featured in this article. The K-optimality and EK-optimality criteria require the definition of a set $\mathcal{S}$ of prediction sites on which to evaluate the kriging variance. We took this set to be a regular grid of 1980 prediction sites with 25-mile spacing, overlaid upon the eastern U.S.

We found that augmenting the existing network by Charlottesville, VA yielded the K-optimal and EK-optimal designs, while augmentation by Charleston, WV resulted in the CP-optimal design (see Figure 11 for the locations of these particular sites). In light of our findings in section 3, the CP-optimality of the design with Charleston is not surprising, as augmentation by Charleston adds the largest number of small lags to the network. That Charlottesville would yield the K-optimal design is less obvious, but part of the explanation may be that it lies near the periphery (as do many other candidate sites, however) and at a latitude where the fitted quadratic mean surface has large curvature. The coincidence of the K-optimal and EK-optimal designs here is probably attributable to the relatively weak spatial correlation among the residuals from the quadratic mean surface, which renders the ranking of E-BLUP prediction error variances relatively insensitive to spatial correlation estimates of varying quality and thus more greatly influenced by the kriging variances.

We repeated the entire analysis but this time adopting a constant, rather than quadratic, mean function. In this case the optimal design with respect to all three criteria was Minneapolis, MN, which is the northwesternmost candidate site. Thus, as we observed with our previous examples, the choice of mean function can have a large effect on the optimal design.

6 Conclusions

We have considered optimal spatial network design for three design objectives: efficient prediction assuming that covariance parameters are known, efficient estimation of those covariance parameters, and efficient prediction when the covariance parameters are unknown. The first two objectives were found to be largely antithetical; for example, when the mean is constant and there is no nugget, designs optimal with respect to the first objective tend to consist of a regular pattern dispersed across the design space, whereas designs optimal with respect to the latter tend to include a few large clusters. Our hybrid criterion for the third design objective attempts to balance the opposing tensions of the first two. Situations in which the hybrid optimal design closely resembles designs that are optimal with respect to either of the first two criteria were identified. These situations were rather intricate to describe, as they depend quite substantially on the strength of the spatial correlation, the mean function used, and the existence and size of the nugget. Nevertheless, it is hoped that the toy examples and larger examples presented herein have laid out some general principles of good spatial design for various objectives that can be put to use by practitioners in designing or augmenting sampling networks.

REFERENCES

Abt, M. (1999). Estimating the prediction mean squared error in Gaussian stochastic processes with exponential correlation structure. *Scandinavian Journal of Statistics*, **26**, 563-578.

Barnes, R.J. (1989). Sample design for geologic site characterization. In *Geostatistics*, Vol. 2, M. Armstrong, ed. Kluwer, Dordrecht, pp. 809-822.

Bras, R.L. and Rodriguez-Iturbe, I. (1976). Network design for the estimation of areal mean of rainfall events. *Water Resources Research*, **12**, 1185-1195.

Cressie, N., Gotway, C.A., and Grondona, M.O. (1990). Spatial prediction from networks. *Chemometrics and Intelligent Laboratory Systems*, **7**, 251-271.

Lark, R. (2002). Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood. *Geoderma*, **105**, 49-80.

McBratney, A.B., Webster, R., and Burgess, T.M. (1981). The design of optimal sampling schemes for local estimation and mapping of regionalized variables - I. *Computers and Geosciences*, **7**, 331-334.

Müller, W.G., and Zimmerman, D.L. (1995). An algorithm for sampling optimization for semivariogram estimation. Pp. 173-178 in *MODA4 - Advances in Model-Oriented Data Analysis*, C.P. Kitsos, and W.G. Müller, eds., Physica, Heidelberg.

Müller, W.G. and Zimmerman, D.L. (1999). Optimal designs for variogram estimation. *Environmetrics*, **10**, 23-37.

Russo, D. (1984). Design of an optimal sampling metwork for estimating the variogram. *Soil Science Society of America Journal*, **48**, 708-716.

van Groenigen, J.W., and Stein, A. (1998). Constrained optimization of spatial sampling using continuous simulated annealing. *Journal of Environmental Quality*, **43**, 684-691.

Warrick, A. W., and Myers, D. E. (1987). Optimization of sampling locations for variogram calculations. *Water Resources Research*, **23**, 496-500.

Yfantis, E.A., Flatman, G.T., and Behar, J.V. (1987). Efficiency of kriging estimation for square, triangular, and hexagonal grids. *Mathematical Geology*, **19**, 183-205.

Zhu, Z. and Stein, M.L. (2004). Spatial sampling design for parameter estimation of the covariance function. Submitted for publication.

Zimmerman, D.L. and Cressie, N. (1992). Mean squared prediction error in the spatial linear model with estimated covariance parameters. *Annals of the Institute for Statistical Mathematics*, **44**, 27-43.

Zimmerman, D.L. and Homer, K.E. (1991). A network design criterion for estimating selected attributes of the semivariogram. *Environmetrics*, **2**, 425-441.

Table 1: Rank correlations between the K-optimality, CP-optimality, and EK-optimality criteria over the population of 2012 distinct designs in the toy example.

	Constant mean, no nugget			Constant mean, 50% nugget		
ρ	$r_{K,CP}$	$r_{K,EK}$	$r_{CP,EK}$	$r_{K,CP}$	$r_{K,EK}$	$r_{CP,EK}$
0.1	-0.97	-0.95	0.97	-0.92	-0.86	0.98
0.2	-0.93	-0.89	0.96	-0.87	-0.79	0.97
0.3	-0.88	-0.75	0.88	-0.80	-0.69	0.97
0.4	-0.81	-0.07	0.20	-0.72	-0.57	0.94
0.5	-0.74	0.73	-0.52	-0.65	-0.42	0.87
0.6	-0.64	0.94	-0.55	-0.50	-0.38	0.91
0.7	-0.27	0.98	-0.17	-0.30	-0.37	0.97
0.8	0.21	0.99	0.25	-0.16	-0.34	0.93
0.9	0.29	1.00	0.30	0.03	-0.28	0.83

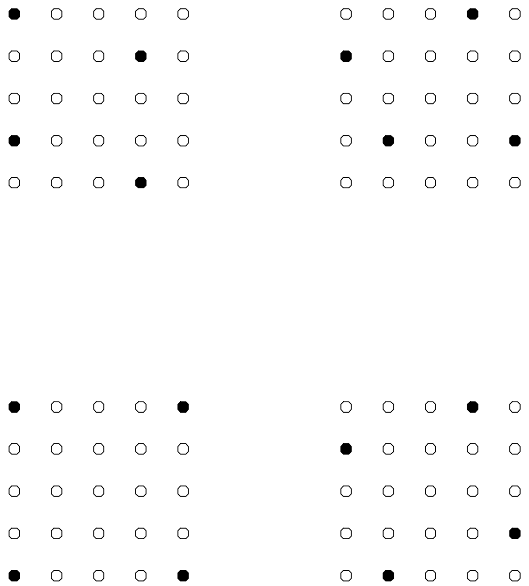

Figure 1: Locally K-optimal designs for the toy example. All designs pertain to the no-nugget model. Top row: constant mean case, with left panel corresponding to $\rho = 0.01, 0.02, \ldots, 0.21$, middle panel corresponding to $\rho = 0.22, 0.23$, and right panel corresponding to $\rho = 0.24, \ldots, 0.99$. Bottom row: planar mean case, with left panel corresponding to $\rho = 0.01, 0.02, \ldots, 0.64$ and remaining panel corresponding to $\rho = 0.65, \ldots, 0.99$.

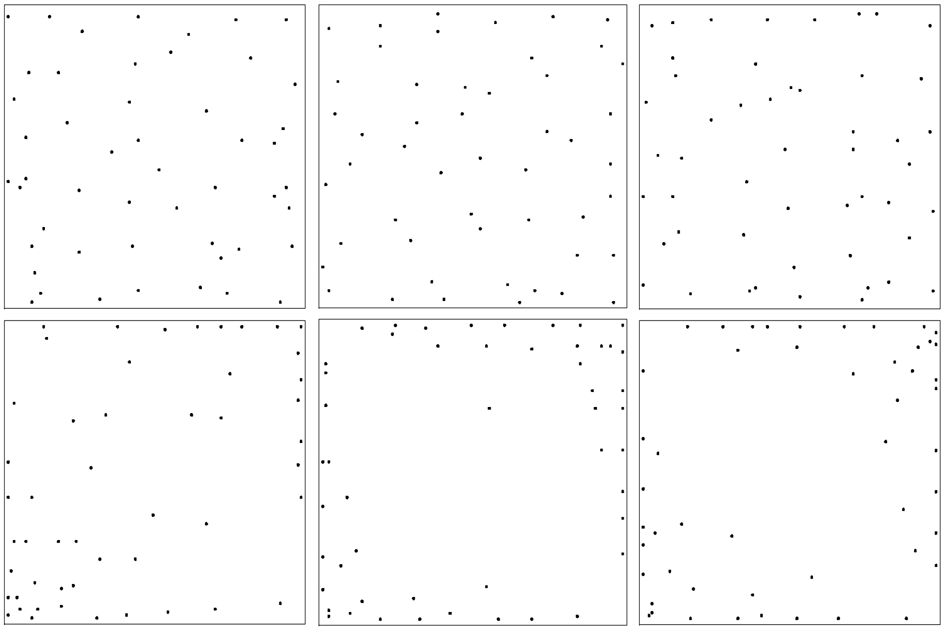

Figure 2: Locally K-optimal designs for the larger example. Rows of panels, from top to bottom, correspond to constant and planar means; columns, from left to right, correspond to $\rho = 0.2$, 0.5, 0.8.

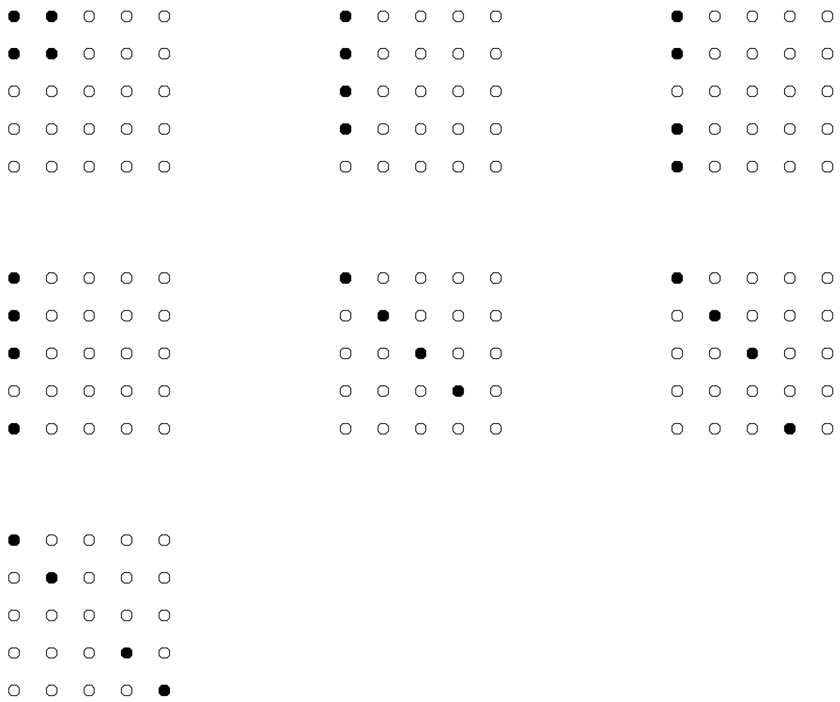

Figure 3: Locally CP-optimal designs for the toy example, in the case of a constant mean and no nugget. Moving from left to right within rows, the panels correspond to the following ranges for ρ:

$\rho = 0.01, \ldots, 0.33$

$\rho = 0.34, \ldots, 0.61$

$\rho = 0.62, 0.63$

$\rho = 0.64, 0.65$

$\rho = 0.66, \ldots, 0.70$

$\rho = 0.71$

$\rho = 0.72, \ldots, 0.99$

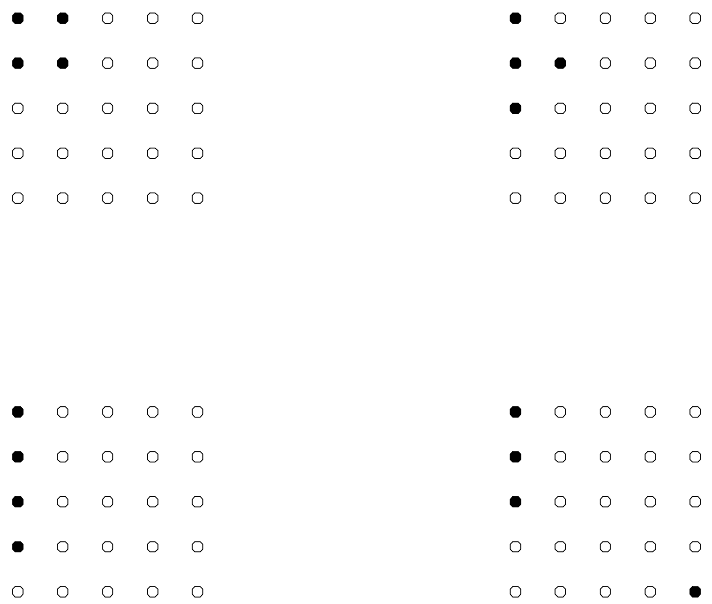

Figure 4: Locally CP-optimal designs for the toy example, in the case of a constant mean and 0% nugget. Moving from left to right within rows, the panels correspond to the following ranges for ρ:

$\rho = 0.01, \ldots, 0.20$

$\rho = 0.21, \ldots, 0.29$

$\rho = 0.30, \ldots, 0.86$

$\rho = 0.87, \ldots, 0.99$

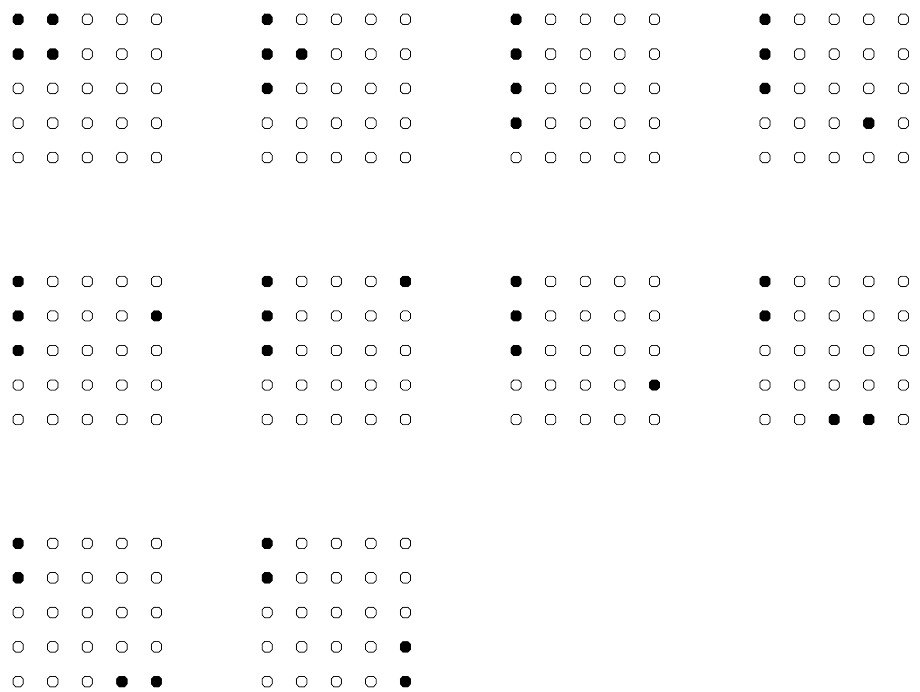

Figure 5: Locally CP-optimal designs for the toy example, in the case of a constant mean and 50% nugget. Moving from left to right within rows, the panels correspond to the following ranges for ρ:

$\rho = 0.01, \ldots, 0.18$

$\rho = 0.19, \ldots, 0.30$

$\rho = 0.31, \ldots, 0.70$

$\rho = 0.71, 0.72$

$\rho = 0.73, \ldots, 0.75$

$\rho = 0.76$

$\rho = 0.77, \ldots, 0.79$

$\rho = 0.80$

$\rho = 0.81, \ldots, 0.88$

$\rho = 0.89, \ldots, 0.99$

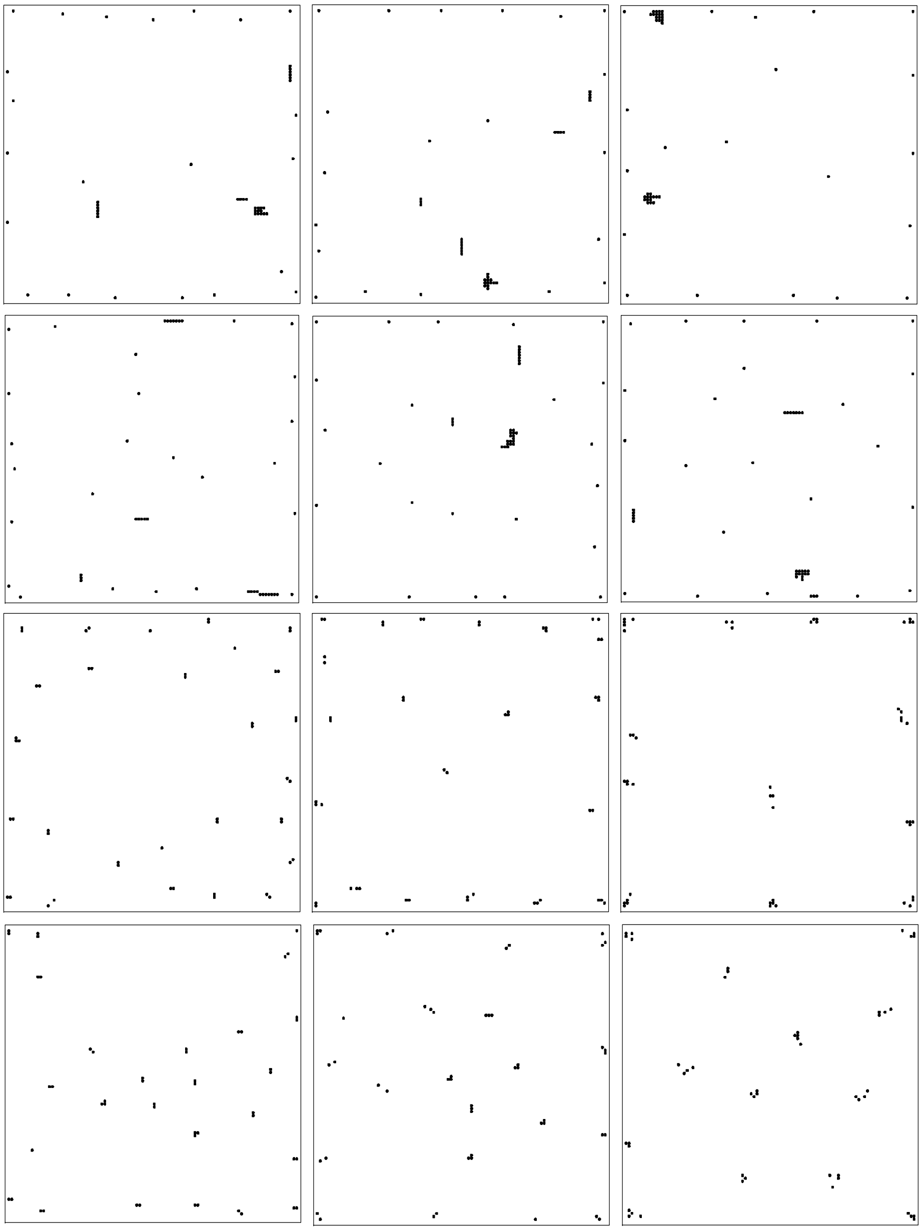

Figure 6: Locally CP-optimal designs for the larger example. Rows of panels, from top to bottom, correspond to constant mean and no nugget, planar mean and no nugget, constant mean and 50% nugget, planar mean and 50% nugget. Columns, from left to right, correspond to $\rho = 0.2,\ 0.5,\ 0.8$.

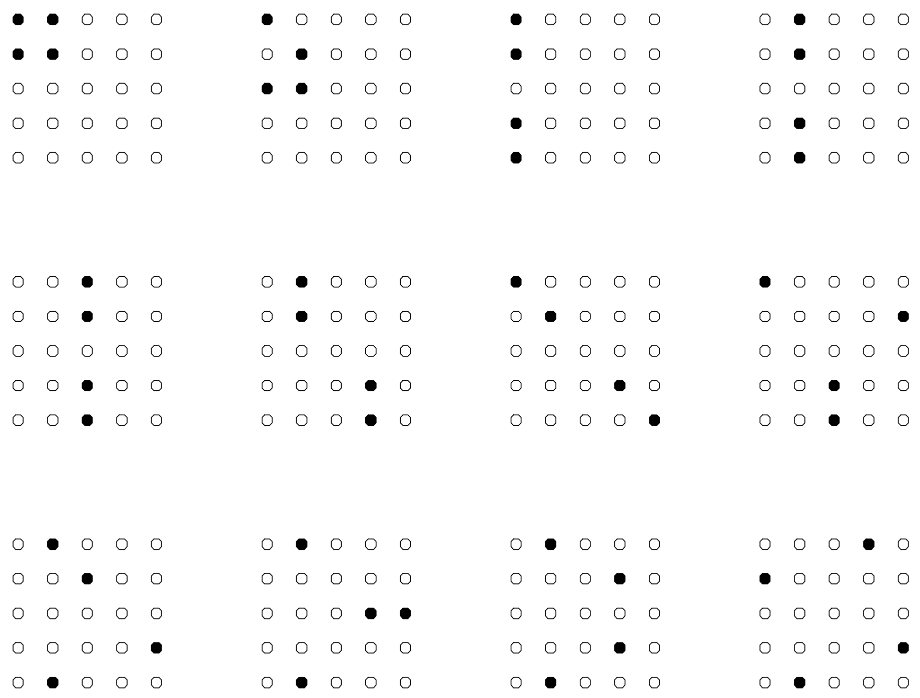

Figure 7: Locally EK-optimal designs for the toy example, in the case of a constant mean and no nugget. Moving from left to right within rows, the panels correspond to the following ranges for ρ:

$\rho = 0.01, \ldots, 0.17$

$\rho = 0.18, \ldots, 0.21$

$\rho = 0.22$

$\rho = 0.23, 0.24$

$\rho = 0.25, 0.26$

$\rho = 0.27, \ldots, 0.31$

$\rho = 0.32, \ldots, 0.41$

$\rho = 0.42$

$\rho = 0.43$

$\rho = 0.44, \ldots, 0.47$

$\rho = 0.48, 0.49$

$\rho = 0.50, \ldots, 0.99$

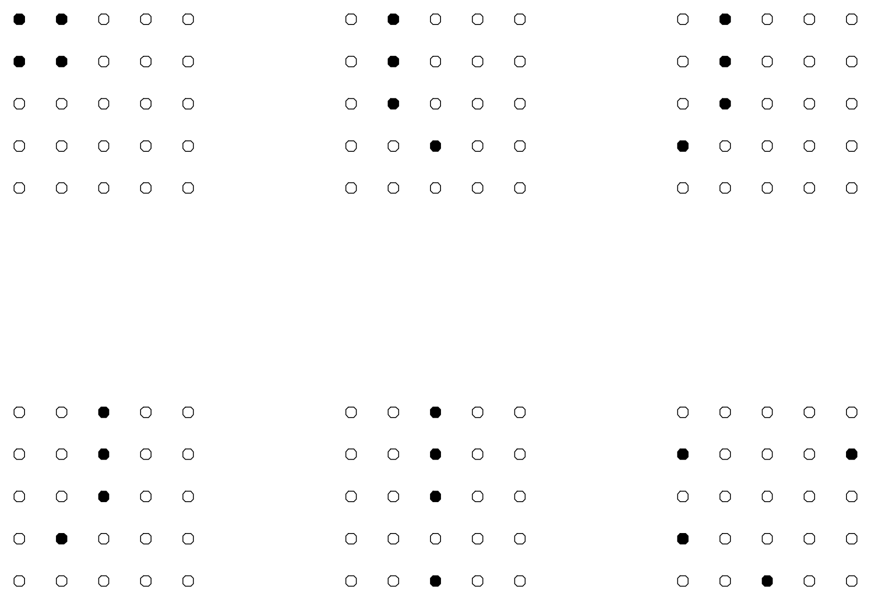

Figure 8: Locally EK-optimal designs for the toy example, in the case a constant mean and 0% nugget. Moving from left to right within rows, the panels correspond to the following ranges for ρ:

$\rho = 0.01, \ldots, 0.33$

$\rho = 0.34$

$\rho = 0.35, 0.36$

$\rho = 0.37, \ldots, 0.40$

$\rho = 0.41, \ldots, 0.73$

$\rho = 0.74, \ldots, 0.99$

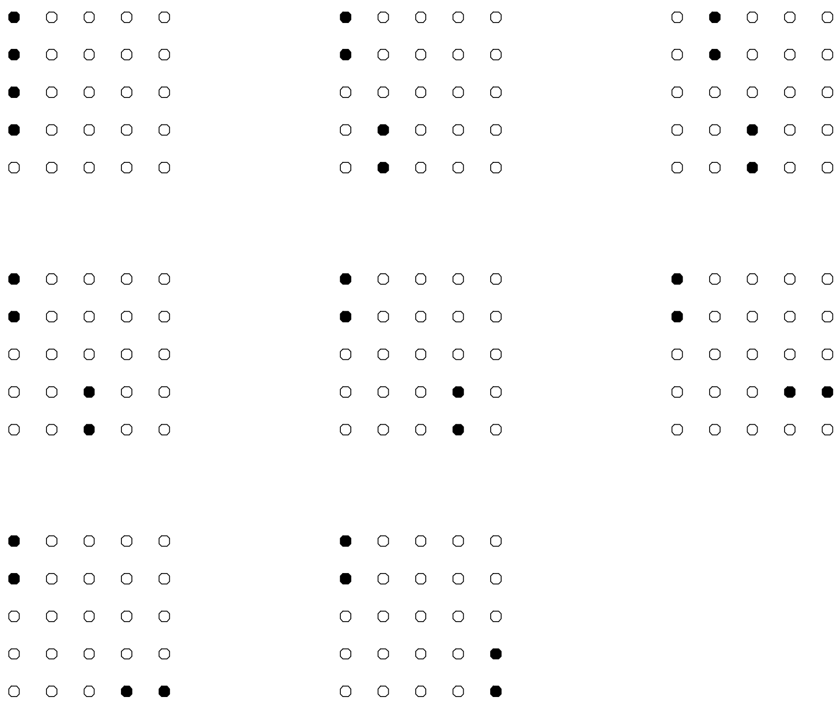

Figure 9: Locally EK-optimal designs for the toy example, in the case of a constant mean and 50% nugget. Moving from left to right within rows, the panels correspond to the following ranges for ρ:

$\rho = 0.01, \ldots, 0.56$

$\rho = 0.57, \ldots, 0.61$, 0.67

$\rho = 0.62, \ldots, 0.66$

$\rho = 0.68, \ldots, 0.74$

$\rho = 0.75, \ldots, 0.77$

$\rho = 0.78$

$\rho = 0.79, \ldots, 0.85$

$\rho = 0.86, \ldots, 0.99$

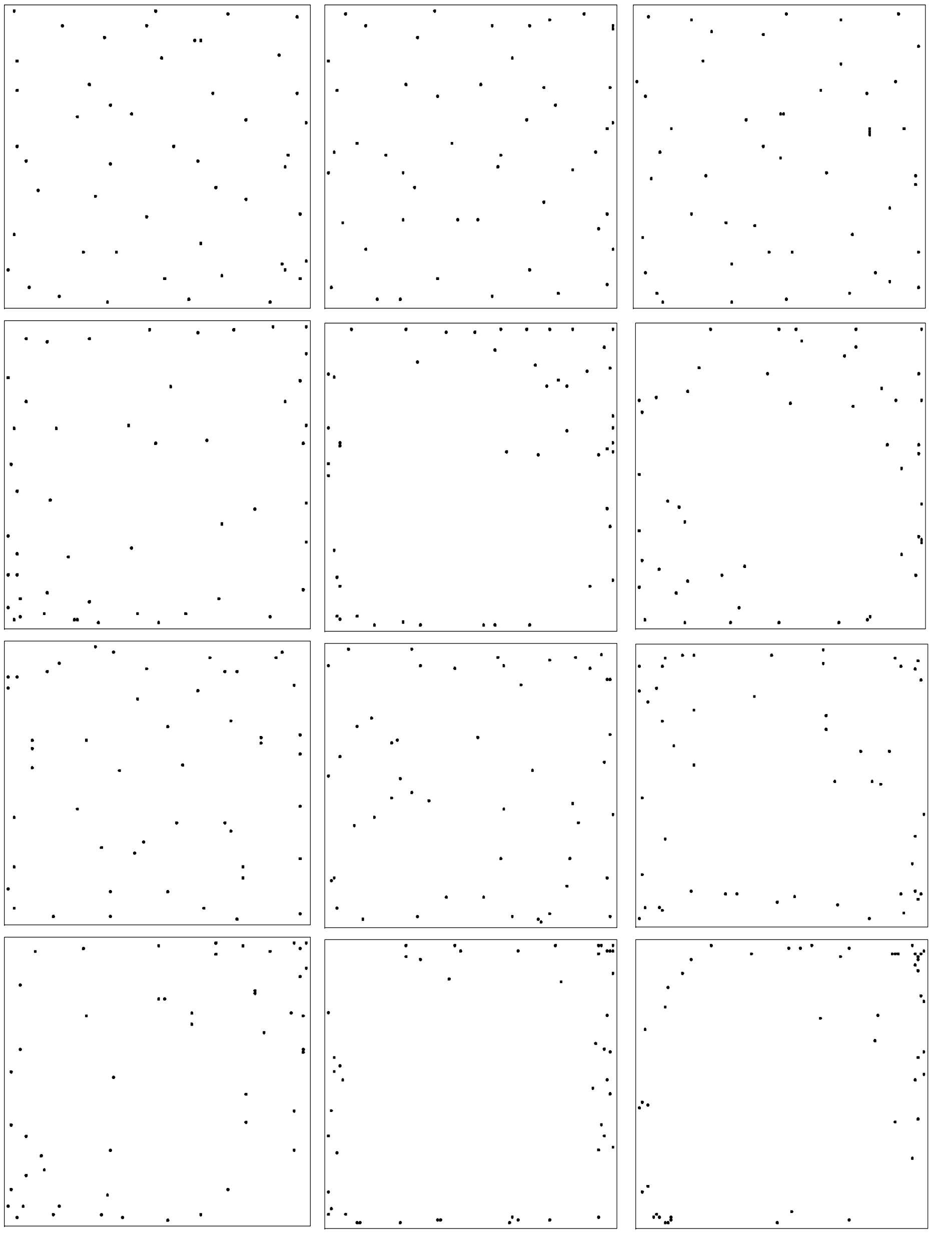

Figure 10: Locally EK-optimal designs for the larger example. Rows of panels, from top to bottom, correspond to constant mean and no nugget, planar mean and no nugget, constant mean and 50% nugget, planar mean and 50% nugget. Columns, from left to right, correspond to $\rho = 0.2,\ 0.5,\ 0.8$.

Figure 11: Map of the Utility Acid Precipitation STudy Program network in 1983 and candidate sites for augmenting the network. Network sites are represented by a black circle; candidate sites are represented by open circles, an open triangle (the K-optimal and EK-optimal site), or an open square (the CP-optimal site).