Constructing spatial designs using regression subset selection

Doug Nychka* Qing Yang
J. Andrew Royle
National Institute of Statistical Sciences and
Department of Statistics, North Carolina State University

February 5, 1997

Abstract

An important problem in spatial statistics is determining where to make measurements. For example, in monitoring environmental pollutants one would like to know how to place a network of measuring instruments to make most efficient use of resources. A basic model for observations is a Gaussian random field with an isotropic covariance function. Given this model there are simple estimators (kriging, BLUE) for predicting the field at locations where measurements are not taken. However, even under this homogeneous model choosing design points to minimize the prediction variance is difficult numerically and is not feasible for large numbers of observation points .

This work presents two strategies for constructing designs that are readily computable and offer the promise of interactive use. Given a set of candidate points it is shown that there is an equivalence between choosing an optimal subset for predicting a linear functional of the field and a regression subset selection problem. Besides using the exhaustive branch and bound algorithm for subset selection (in Splus the "leaps" function), a new method, the lasso, is also considered. The lasso is a recent development by Robert Tibshirani and is based on a regularization/Bayesian principle. These designs are compared to space-filling designs over a range of isotropic correlation functions and the results suggest that both designs based on subset selection are as good or better than a space filling criterion. The methods are also applied to shrink an ozone monitoring network for the Chicago urban area.

1 Introduction

An important practical problem in monitoring environmental pollutants is the efficient location of monitoring instruments. The placement and the number of stations may be based on many factors some of which may depend on limited resources, federal and state regulations and local conditions. However, even restricting attention to the statistical issues related to network accuracy, the network design problem is difficult. In this paper we will suggest some new approaches for thinning an existing network of stations and also evaluate these ideas for developing new networks.

*Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203

Experimental design is a large and rich subdiscipline of statistics. For the purposes of environmental monitoring we will focus on designs that locate points over space. Also we will assume that the measurements of interest (or their transformed values) can be modeled as a Gaussian random field. The traditional approach for statistical design is to select a particular criterion that measures the merit of a particular design and then optimize this criterion over all possible choices of locations (Welch et al., 1992, Sacks et al., 1989). A technical problem with identifying optimal designs is the difficulty in computing the optimal configuration of locations. Lengthy computation times make straight forward approaches for design construction unsuitable for use in interactive programs.

In this work it will be assumed that the candidate design points are drawn from some discrete set. If one wanted to thin an existing network the candidate set would be the full set of network locations. For a new design, one could start with a large regular grid of points or some other space-filling design as the candidate set, $\mathcal{C}$. The possible designs would then consist of subsets of $\mathcal{C}$. Given this discretization of the designs space, we propose to construct designs using an equivalence between the optimal prediction variance and regression subset selection. In this setup each location in the candidate set is equated with a regression variable. Inclusion of this location in the design is equivalent to selecting this regression variable for a regression model. The advantage of this correspondence is that one can make use of efficient algorithms for subset selection. The idea of using branch and bound algorithms for finding optimal designs is not new (e.g. Welch, 1982) but the direct connection with using off the shelf regression software has not been given much attention.

This particular research was originally motivated by an admittedly *ad hoc* exercise to thin the Chicago area network of ozone monitoring stations and it will be helpful to introduce this problem by outlining this example. More details are given in Section 6. Over the summer of 1987, 89 daily ozone averages were measured at 20 locations in the Chicago urban area. The ozone field is not stationary over space but we can expect that observations separated a few days apart are nearly ncorrelated. Form the average of these 20 stations and 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 the 89 days and X_J is a matrix with the columns being the observed ozone values for a subset of J stations. An independent variable in this regression equation is identified with a particular station. Thus selecting a subset of variables is equivalent to identifying a subset of the monitoring network. Initially we used the `leaps` subset procedure in Splus 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 1 4. Despite the *ad hoc* nature of this procedure and an 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. Furthermore, what is the analogous procedure to generate a design when data is not available? An investigation of this strategy and extension to the case without data is the subject of this paper.

The next section gives the details of the spatial models and the relationship to the regression subset selection problem. Section 3 reviews the leaps and lasso subset selection methods and Section 4 describes an alternative way to construct designs using a purely geometric criterion. Section 5 evaluates these designs on a regular grid for several different design criteria and different covariances for the random field. In Section 6, these methods are applied to thin an existing ozone monitoring network for the Chicago urban area. This is followed by a discussion of the results and some future directions.

2 Random surfaces and subset selection

2.1 Random field model

The mathematical basis for this work depends on the assumption that the surface of interest is a realization of a spatial process. 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. Let $Z(\boldsymbol{x})$ be a Gaussian random field for $\boldsymbol{x} \in \Re^2$ 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 $1 \le j \le N$ be the N locations where the field is measured and $K_{i,j} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ denote the covariance matrix for this vector. This full set of locations will be referred to as the candidate set for the spatial design.

Now let $\lambda(Z)$ be a linear functional of the field and $\gamma_j = \mathrm{Cov}(\lambda(Z), Z(\boldsymbol{x}_j))$. In the examples in this paper λ is taken to be the average value of Z over the set of candidate locations but any other linear functional of the field can be handled. For Gaussian random variables the conditional expectation of $\lambda(Z)$ given $Z(\boldsymbol{x}_j)$ for $1 \le j \le N$ is a linear function of the observations and has the form $\hat{\lambda} = \boldsymbol{z}^T K^{-1}\gamma$ where $\boldsymbol{z}^T = \{Z(\boldsymbol{x}_1), ..., Z(\boldsymbol{x}_n)\}$. This "Kriging" estimate is unbiased and has minimum variance among all linear estimators. Moreover, it is straight forward to derive the formula for the mean squared error:

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

The key observation for the purposes of this work is that the mean squared error for $\hat{\lambda}$ can formally be expressed as the residual sum of squares from a linear regression. Recall the formula

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

Thus if the cross products satisfy the relationships $X^TX = K$, $X^TY = \gamma$ and $Y^TY = VAR(\lambda)$, then 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.

2.2 Properties of the Motivating Example

At this point one can identify a simple theoretical model for the *ad hoc* analysis applied to the Chicago ozone network described in the introduction. Just work backwards from the regression model form to identify the spatial covariance. If the ozone measurements and the network average are centered by their means and divided by the square root of the number of days then X^TX is the sample covariance matrix for this field. Similarly X^TY will be the sample covariance between each location and the network average (λ in this case) and Y^TY will be the sample variance for λ.

Under the assumption that the means are known the design selected by the leaps procedure will be the optimal one for predicting the network average. This follows from the correspondence in the previous section. The subset selected by minimizing the residual sums of squares in the regression problem is the same as the one that minimizes average prediction error.

2.3 A Pseudo regression problem for a given covariance

To capitalize on regression subset selection algorithms it is also of interest to understand how to create pseudo regression data from a given covariance function. Given K and λ, how can one construct X and Y so that the residual sums of squares will be equal to the mean squared error of the kriging estimate? First assume that the functional of interest is just a linear combination of the field at the candidate locations

$$\lambda(Z) = \sum_{k=1}^{N} w_k Z(\boldsymbol{x}_k)$$

thus $VAR(\lambda) = \boldsymbol{w}^T K \boldsymbol{w}$ and $\gamma = K\boldsymbol{w}$. One way to choose X and Y to achieve these relationships is $X = K^{1/2}$ and $Y = K^{-1/2}\gamma$. Clearly $Y^T Y = \gamma^T K^{-1}\gamma = VAR(\lambda)$ and we will argue that the second term must be equal to the correct quadratic form. Part of the reason for this exposition is to emphasize the distinction between the full set of N candidate locations and a subset of these of size J. Accordingly, let K_J denote the covariance matrix corresponding to a subset of J locations and let X_J denote the submatrix of X consisting of the relevant J columns. Then $X_J^T X_J = K_J$ and, keeping in mind that X_J is the submatrix of the full square root matrix of K (*not* the square root of the reduced matrix),

$$X_J^T Y = K_J^{1/2} K^{-1/2}\gamma = E\gamma$$

where E is a $N \times J$ matrix with $E_{ij} = 1$ if the i^{th} location in the subset is equal to the j^{th} location in the full candidate set and zero otherwise. Thus we have $(E\gamma)_i = COV(Z(\boldsymbol{x}_i), \lambda)$.

In the case when λ is not a linear combination of the field at the candidate locations the functional can be decomposed as $\lambda = \lambda - \lambda_0 + \lambda_0$ with $\lambda_0 = \sum_{k=1}^{N} w_k Z(\boldsymbol{x}_k)$ and $\boldsymbol{w} = K^{-1}\gamma$. With this choice

$$\lambda_0 = E\left\{\lambda | Z(\boldsymbol{x}_1), ..., Z(\boldsymbol{x}_N)\right\}$$

and thus $(\lambda - \lambda_0)$ must be uncorrelated with any linear combination of the field at the candidate points. Thus for any estimate based on a subset of the candidates

$$E(\lambda - \hat{\lambda}) = E(\lambda - \lambda_0)^2 + E(\lambda_0 - \hat{\lambda})^2$$

and so minimizing the mean squared error over points in the candidate set is equivalent to minimizing the second part. This second term can be interpreted as the mean squared error for estimating λ_0, a functional that is a linear combination of the field at the candidate points. Thus the general problem has been reduced to the special case previously examined.

3 Variable selection for linear regression models

The regression subset selection problem is, given a set of variables, to find the subset that minimizes some selection criterion. Perhaps the most common method is to consider all subsets of the variables of a given size and choose the combination that gives the minimum residual sums of squares. We also consider an alternative approach that depends on a nonlinear constraint on the parameter vector. To unify both of these methods consider the constrained minimization of the residual sum of squares:

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

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 Splus by the `leaps` function (Furnival and Wilson, 1974, Becker, Chambers and Wilks, 1988).

A recently proposed strategy for subset selection uses a different constraint set. The *lasso* procedure (Tibshirani, 1994) requires that the sum of the absolute values of the parameter vectors be less than a fixed value.

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

At first sight it is not clear why the solution to this constrained problem is related to subset selection. However, due to the properties of the absolute value function, the constrained solution will force some of the components of β to be identically zero. The reader is referred to Tibshirani's article for some examples to show why this makes sense geometrically. The number of zeroes 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 a lasso solution. In a straight regression setting one might use the parameter estimates from the lasso procedure for prediction. But in this work the lasso results are only used to identify subsets of the variables (locations). Once these locations are found the usual BLUE estimates are used to evaluate the efficiency of the design.

Both the leaps and lasso procedures run fairly quickly on a Sun Sparc II workstation. The leaps procedure attempts to compute the best subset for all subset sizes. For N variables this amounts to searching through the 2^N different possible subsets. Clearly a direct enumeration is not feasible even for moderate size N. The strategy is to eliminate most of the subsets using the fundamental inequality that the residual sum of squares (RSS) will never increase if additional variable are added to the model. For example, suppose that the RSS for variables 1 and 2 is less than the RSS for variables 3, 4 and 5, then any subset of size two taken from $\{3, 4, 5\}$ will be greater than the RSS associated with the subset $\{1, 2\}$. Therefore the RSS for these subsets of size two do not need to be computed. Another efficient feature in the leaps algorithm is that the inverse of $X_J^T X_J$ for subsets of variables is derived in an efficient manner from the inverse of that from the full regression problem. In this way the subset regressions can be rapidly computed. One disadvantage of the leaps method is that binary tree of possible subsets increases exponentially with the number of variables. If the residual sums of squares for many subsets are similar the tree can not be efficiently pruned. As part of the leaps implementation there is a tolerance of .0001 that will ignore differences below this bound. When this tolerance is enforced the leaps algorithm may give an approximate solution not the optimal one. Also in Splus the `leaps` function has an artificial upper limit of 31 variables based on coding the location of a node in a binary tree as bits in a single 32 bit signed integer. For this reason larger numbers of variables were not explored.

The computation of the lasso solution can be reduced to a standard problem in nonlinear programming: optimizing a quadratic cost function with inequality constraints on the parameters. Although the solution is iterative, the main computational burden in each step involves a weighted least squares regression. For this reason there is some hope that the lasso will be able to handle much larger variable selection problems than the leaps method. Subsets (i.e. designs) of different sizes were found by searching over a grid of values of the constraint parameter, t. More will be said about this grid search in Section 5.

The reader should keep in mind that the lasso results need not be the same as simply minimizing the residual sum off squares for a fixed subset. Although the lasso designs have useful properties we do not expect them to be optimal based on mean squared error for estimating the network average.

4 Space filling designs

Besides deriving designs based on a covariance function one can also use a geometric criterion to distribute design points. As in the case of using subset selection methods, it will be assumed that the designs are the solution to the discrete problem of selecting a subset of points from a larger set of candidates. Let $\mathcal{C}$ denote the set of N candidate points and $\mathcal{D}$ a subset constituting a design of size J. Let $p < 0$. A metric for the distance between a point 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 criterion can be thought of as measuring how well the design *covers* the point $\boldsymbol{x}$ and for $p < 0$ it is easy to show that $d_p(\boldsymbol{x}, \mathcal{D}) \rightarrow 0$ as $\boldsymbol{x}$ converges to a member of $\mathcal{D}$. Taking $q > 0$, an overall coverage criterion is an L_q average of coverages for each candidate point

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

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 $p \rightarrow -\infty$ and $q \rightarrow \infty$, $C_{p,q}$ converges to the criterion commonly associated with minimax, space filling sets of points and these designs are computed by the Gosset design package (Hardin and Sloane, 1994). (The terminology is confusing because $C_{-\infty,\infty}$ is actually maximum over minimums.) See also Johnson et al. (1990) for some theoretical connections between space filling designs and those based on prediction error for a spatial process. In general, coverage designs can be computed by the SAS procedure PROC OPTEX (Tobias, 1995). In this work coverage designs were found with $p = -5$ and $q = 1$ and the programming was done in Splus. Based on the examples of Tobias this choice was a compromise between designs that were close to the minimax solution but, because p is finite, were more stable to compute. Finding the optimum was accomplished by using random starting configurations and decreasing the coverage criterion by swapping a candidate point with a design point. Convergence usually occurred after 2 or 3 passes through the set of design points.

5 Comparing designs on regular grids

A numerical study was carried out to determine some differences in performance for the designs described in Sections 3 and 4. This study has a $2 \times 7 \times 3$ factorial layout where the factors are the type of regular lattice (5×5 square, 30 point hexagon) the range parameter of the covariance function (7 values), and the type of design (leaps, lasso, coverage). The regular lattices are plotted in Figure 1 where the minimum distance between points on the lattice is scaled to unit length. The lattices were not perfectly regular but were randomly perturbed by a small amount to break the four- or six-fold symmetry.

The fields were assumed to have zero mean with an isotropic exponential covariance function given by $k(\boldsymbol{x}, \boldsymbol{x}') = e^{-\|\boldsymbol{x}-\boldsymbol{x}'\|/\theta}$. Seven values of θ were used corresponding to nearest neighbor correlations ($e^{-1/\theta}$) of $(.01, .1, .3, .5, .7, .9, .99)$ for the lattice points.

The leaps and lasso procedures were used to select subsets based on predicting the average of the field at the lattice points. The space filling designs used are based on the coverage criterion with $p = -5$ and $q = 1$ where, of course, the candidate set is restricted to the lattice. This last type of design does not depend on the assumed covariance and will be the same for all values of θ.

Figures 2-6 are plots of some of these designs for different subset sizes and correlation scales. The most striking result is that it is not easy to generate lasso designs for all subsets sizes. The design as a function of t appears to be quantized, jumping between numbers of points that can provide a symmetrical pattern. For example, in the square lattice grid the favored designs are one point in the center, adding the corners to give five points, adding four more middle points to give a nine point symmetrical pattern. Oddly enough this behavior also depends on θ. Using a bisection search with respect to t it is possible to find designs of other sizes. However, it is not clear that these intermediate solutions are true solutions to the constrained problem or just the result of numerical roundoff errors. For this reason, and the difficulty in finding them, the lasso designs for these intermediate subset sizes were not considered.

5.1 Performance measures

The designs were evaluated according to three different measures of their performance that can be computed explicitly using the covariance function, and are mean squared errors for estimates of the field. Let $\hat{Z}(\boldsymbol{x}')$ denote the BLUE estimate of the field at an arbitrary location based on observations at the J design points. Note that $\hat{\lambda} = \sum_{j=1}^{N} \hat{Z}(\boldsymbol{x}_j)$, due to the linearity of the kriging estimator. Also let $\boldsymbol{A}$ denote the rectangle or hexagonal design region containing the set of candidate points. The 3 performance measures are:

1 Prediction variance of the average (PVA) for the design:

$$\mathrm{VAR}\left\{\frac{1}{n}\sum_{j=1}^{N} Z(\boldsymbol{x}_j) - \hat{\lambda}\right\}$$

2 Maximum prediction variance (MPV) for the design region.

$$\max_{\boldsymbol{x}\in\boldsymbol{A}} \mathrm{VAR}\{Z(\boldsymbol{x}) - \hat{Z}(\boldsymbol{x})\}$$

3 Average prediction variance (APV) for the design region.

$$\int_{\boldsymbol{x}\in\boldsymbol{A}} \mathrm{VAR}\left\{Z(\boldsymbol{x}) - \hat{Z}(\boldsymbol{x})\right\} d\boldsymbol{x}$$

Maxima and integrals over $\mathcal{A}$ were approximated by considering the values of the field on a fine grid of 40×40 points. Based on the continuity of the functions this level of discretization did not introduce any significant biases in the calculations of the performance measures.

5.2 Results

Figures 7-10 help to summarize the values of the performance measures across different correlation structures and designs. The horizontal axis in each case is the number of points in the design. The vertical axis is one of the performance measures or, in the case of the second and third plot for PVA, it is the ratio of the variance for the leaps designs to the other two. Lines connect the values for the leaps and space filling designs and symbols indicate the values for the lasso designs. Recall that lasso designs are only available for specific subset sizes and thus the values associated with different subsets are plotted as points instead of broken lines.

The leaps designs are optimal and so as expected they have the smallest variances for the predicted average. The efficiencies in the second and third plots of Figures 7 and 9 indicate that the coverage designs only incur a small increase, usually less than 10 %, in variance as compared to the leaps results. In contrast the lasso designs often have poor performance especially for the models very high correlations. The performance of the designs with respect to prediction variance is similar. The differences among the average prediction variance are most apparent for the smallest subset sizes and moderate size correlations. The maximum prediction variance is a more sensitive measure of the design configuration and provides a sharper distinction between the designs. Perhaps the most surprising feature is the nonmonotonicity of the APV as a function of subset size.

Focusing on the leaps designs when the correlation is .7, the MPV increases by about 20 % as the subset size increases from 7 to 8 and then decreases by 33 % for a subset size of 9. This unusual behavior can be understood by looking at the prediction variance surfaces (Figure 11). The MPV is sensitive to the corners in the design region and this effect can be explained by the leaps design favoring interior coverage verses pinning down the corners. The emphasis on the interior leaves a corner open for the subset of size 7 and hence a large prediction variance at this location. A similar phenomenon is associated with the peak in the space filling designs at a subset size of 7.

For the hexagonal lattice the lasso designs do not do as well as leaps or coverage designs especially with respect to PVA. The very large inefficiency at a correlation of .99 and a subset size of 9 corresponds to an unusual, asymmetric lasso design that does not spread out points well. In contrast to the square lattice MPV there are fewer instances of nonmonotonicity for the MPV criterion. For example the unusual behavior detailed in Figure 11 above for the leaps designs does not seem to occur for the hexagonal region.

5.3 Random designs

It is useful to calibrate the performance of these designs relative to locations that are chosen at random. In this way one can judge the benefit of a deliberate design over a purely random choice. For subsets of size 5 and 10, 100 random designs were generated by choosing a random sample of points from the square lattice. The distributions of the PVA and the MPV variance are summarized by boxplots in Figure 12. In addition, the values corresponding to the leaps designs are indicated by open circles. Intermediate correlation scales yield the greatest diversity among the random designs' performances. In the range of .3 to .7, we see that the median for random design is at least 50% higher than the same value for the leaps design.

6 Chicago area ozone monitoring network

As a companion to studying regular arrays of points the design ideas were applied to the collection of 20 stations used to monitor ambient ozone in the Chicago area (see figure 13). The interest is to identify subsets of stations that can accurately reproduce summaries for the full complement of stations. The basic idea of this example is to use the 1987 data to construct designs and then validate these designs based on the next three years of data.

Typically ozone is recorded on an hourly basis and is aggregated into a daily summary. For this work we consider average ozone in the period 8am-5pm (8Hr Average). The summary was calculated for each station over the period from June 3 to August 21 along with the network average.

6.1 Subset designs

Based on the discussion from section 2, the implicit spatial covariance used for the leaps and lasso designs is the empirical covariances among the stations and the covariance between stations and the network average. The leaps method selects subsets of locations that are visually reasonable (Figure 14). One consistent feature is that sites in the southwest section (station numbers 6, 10, and 16) do not seem to contribute to estimates of the network average. Analogous to the regular lattice results, the lasso procedure will only generate designs of particular sizes and the gaps in Figure 15 indicate design sizes that are not readily available. Unlike the leaps designs the lasso tends to cluster points in one area near Lake Michigan. Also the lasso designs are nested subsets.

As to be expected the space filling designs produce uniform distributions of points that include the southwest corner avoided by the other designs.

6.2 Design Performance

Because a long record of ozone measurements is available the easiest way to evaluate subsets is by validating with ozone measurements made at other years. Recall that the designs were constructed using a covariance estimated from the ozone data from 6/87-8/87. These data were also used to derive the prediction equations. For validation we used the linear prediction equations found for each design and applied these to ozone measurements made in the following three summers. As a benchmark the reader should note that the unconditional standard deviation for the full network over the three year validation period is 16.1 PPB. Figure 15 summarizes the mean squared errors between predictions and observed values for these years. Because of the optimal properties of the leaps procedure the prediction variance (or RSS) is smaller than the other designs over the training sample. When the designs are compared over subsequent years their performance is similar.

A common feature in the subset sizes is the rapid increase in the design efficiency as the number of sites increases. For example, a design containing 5 locations appears to do well in estimating the average of the full network. In quantitative terms a 5 station network has a prediction root mean squared error that is only 2.5 PPB.

7 Discussion

This work has demonstrated the feasibility of selecting spatial designs based on algorithms for subset selection in regression. For the regular lattices the leaps and coverage designs have similar performance. Although the leaps designs are optimal for estimating the average of the field, it is also encouraging that they are comparable to the coverage designs with respect to the mean prediction variance. Because the coverage designs tend to distribute points evenly one would expect them to do well with respect to MPV. Given the difficulty in finding the optimal design with respect to MPV the coverage design performance is a practical benchmark.

The lasso designs do not do as well as the leaps and coverage designs for regular lattices. This poor performance is especially apparent in estimating the network average. However, the lasso designs still do better than most random chosen designs and their performance improves with larger subset sizes. Also, given that the lasso computational algorithm is a quadratic programming problem with inequality constraints it should scale linearly with sample size. For this reason it remains a useful designs tool. It may be possible to improve the lasso by using a weighted average as the target linear functional. The weights could be small for candidate points near the boundaries and tend to push the lasso designs towards the interior of the design region.

The results for the Chicago ozone network are most important in suggesting the efficiency of a much smaller network for estimating the network average. Unfortunately this application does not help to distinguish among the different design strategies because, for the three years used for validation, performance is similar. For an irregular set of candidate points with a nonstationary covariance, the variance of the predicted average may be less sensitive to the design. Mean and maximum prediction variance may help to distinguish among the design strategies. Future work on this problem is to identify a nonstationary covariance function for this spatial process and evaluate the other measures of design performance.

The careful reader will note that theoretical basis in Section 2 rests on the assumption that the field has a zero mean. In practice one would like to include a linear term depending on spatial location in modeling the mean or "trend" of the spatial field. So for example the average field would be estimated by taking the average over the estimated linear component and adding this to the average for the remaining part attributed to the random field. Although one can include these terms as keep variables in the selection algorithm the RSS does not account for the additional variability in estimating the mean or trend terms.

An important aspect of this problem is the robustness of these designs to a range of covariances. Typically one does not know the covariance function precisely and it is important to have designs that are not overly sensitive to the exact form of the covariance. One intriguing question is whether a design problem considering a family of covariance functions has a corresponding regression subset selection problem.

References

Becker, R.A., Chambers, J.M. and Wilks, A.R. (1988). *The New S Language*, Wadsworth & Brooks/Cole, Pacific Grove, California.

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

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

Hardin, R.H. and Sloane, N.J.A. (1994). *Operating manual for Gosset: A general purpose program for constructing experimental designs*. AT&T Bell Laboratories, Murray Hill, NJ.

Sacks, J., Welch, W.J., Mitchell, T.J. and Wynn, H.P. (1989). Design and analysis of computer experiments. *Statistical Science* 4, 409-435.

Welch, W.J., Buck, R.J., Sacks, J., Wynn, H.P., Mitchell, T.J. and Morris, M.D. (1992). Screening, Predicting and Computer Experiments. *Technometrics* 34, 15-25.

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

Tibshirani, R., 1994, Regression shrinkage and selection via the lasso, Technical Report, Department of Statistics, University of Toronto, Jan. 1994

Tobias, R. (1995). *SAS QC Software. Volume 1: Usage and Reference.* SAS Institute, Inc.

Welch, W.J. (1982). "Branch-and-bound Search for Experimental Designs Based on D Optimality and Other Criteria," *Technometrics* **24**, 41-48.

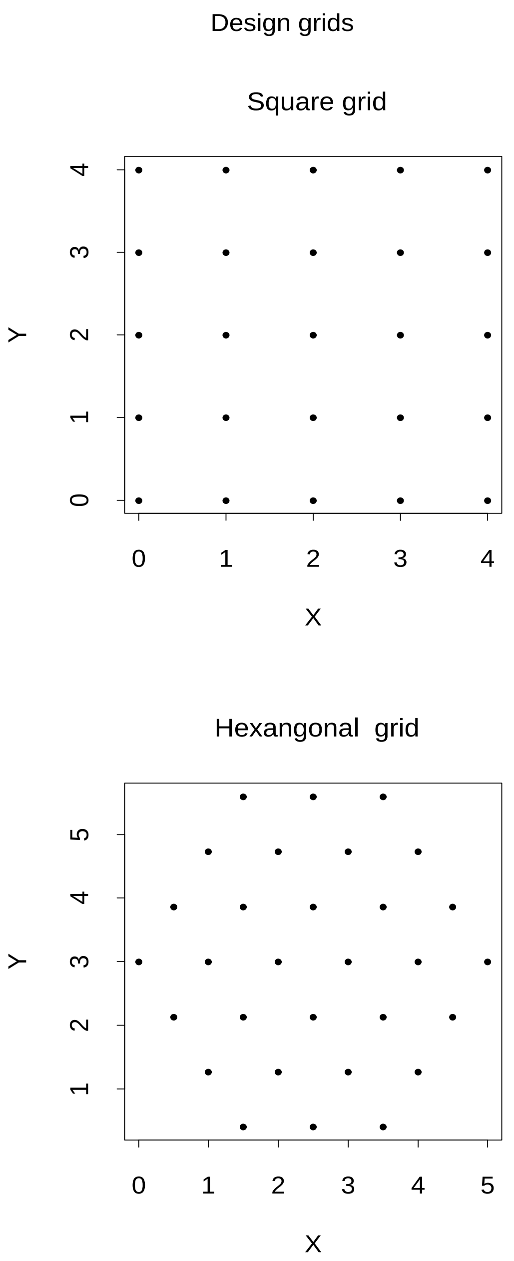

Figure 1: Regular lattices used to evaluate methods for choosing designs. The two regular candidate sets consist of a 25 point square lattice and a 30 point hexagonal lattice. The spacing between nearest neighbor point is taken to be of unit distance.

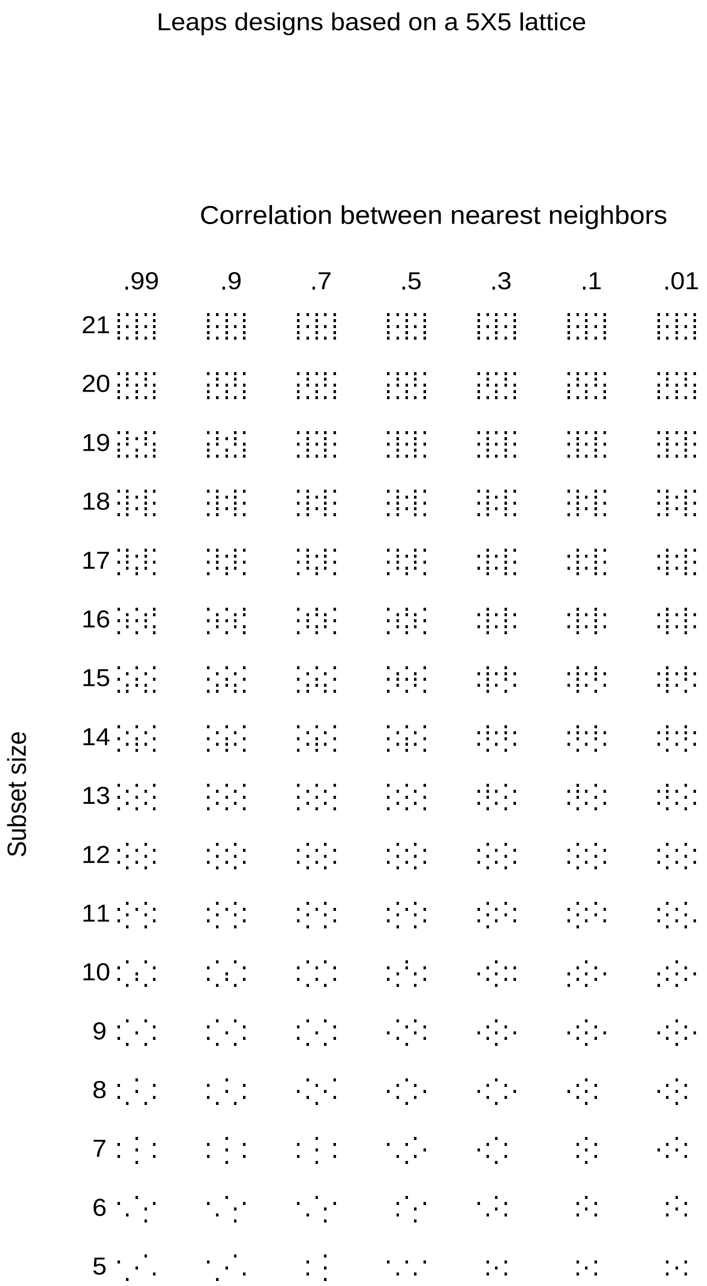

Figure 2: Leaps Designs based on a square lattice and exponential covariance. Vertical axis is the number of points in the design and the columns are for different scale parameter values of an isotropic exponential covariance function. In this model $\theta = -1/log(\rho)$ where ρ is the correlation between locations separated by unit distance.

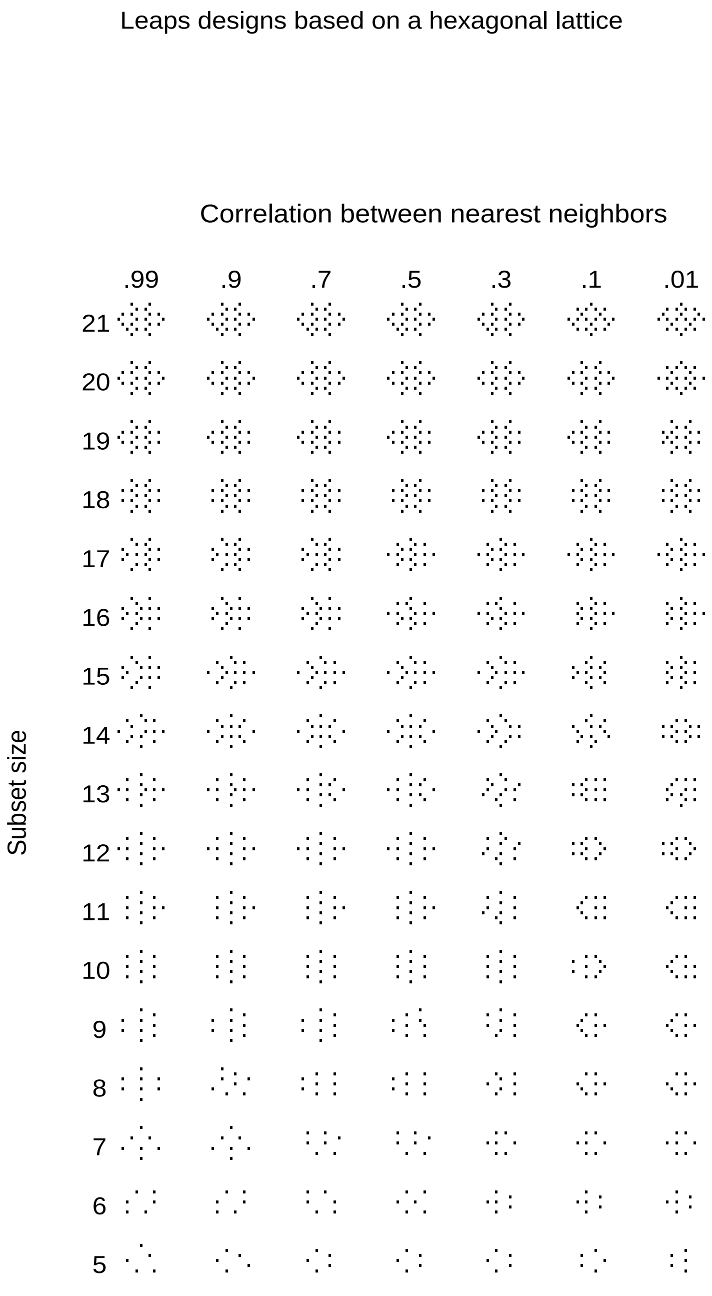

Figure 3: Leaps Designs based on a hexagonal lattice

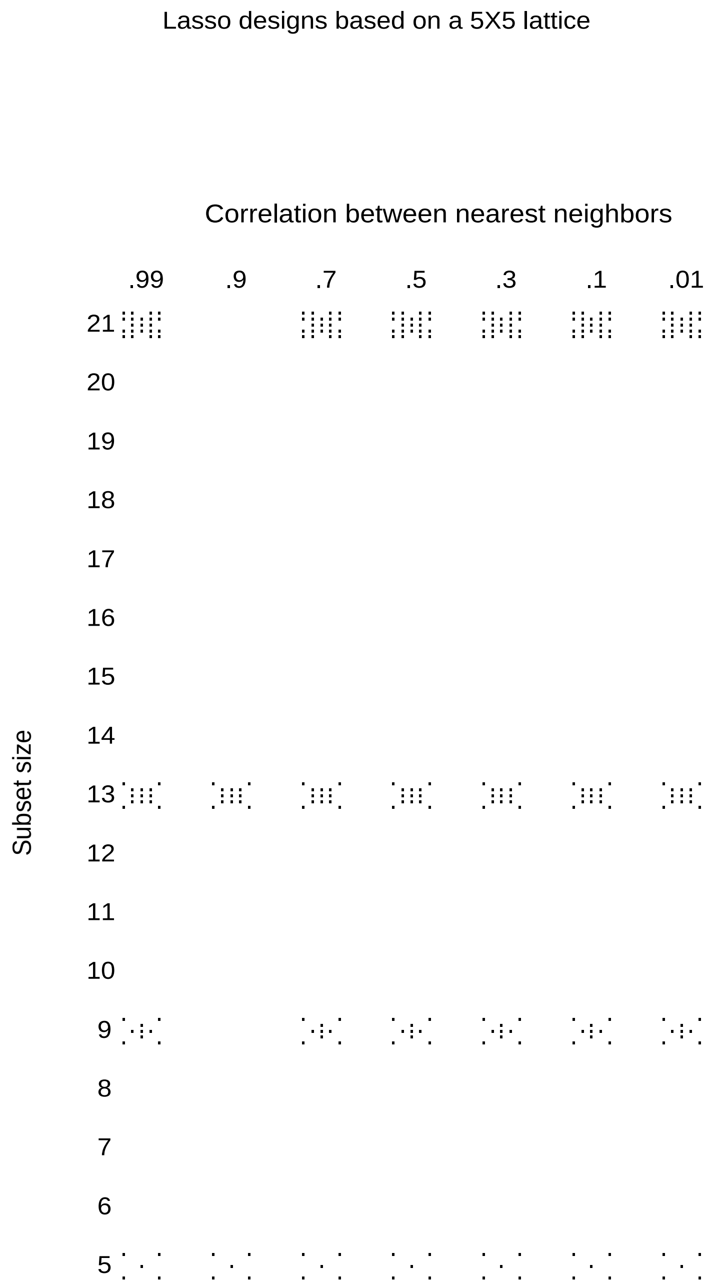

Figure 4: Lasso Designs based on a square lattice

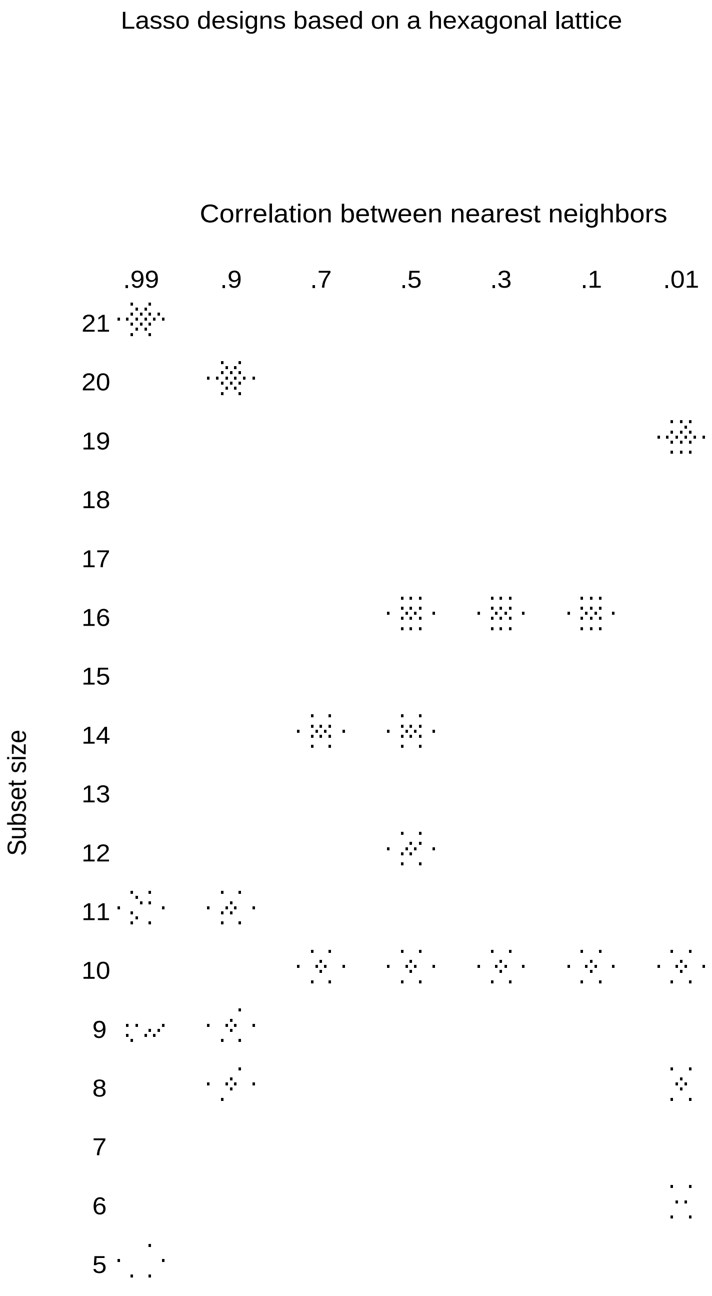

Figure 5: Lasso Designs based on a hexagonal lattice

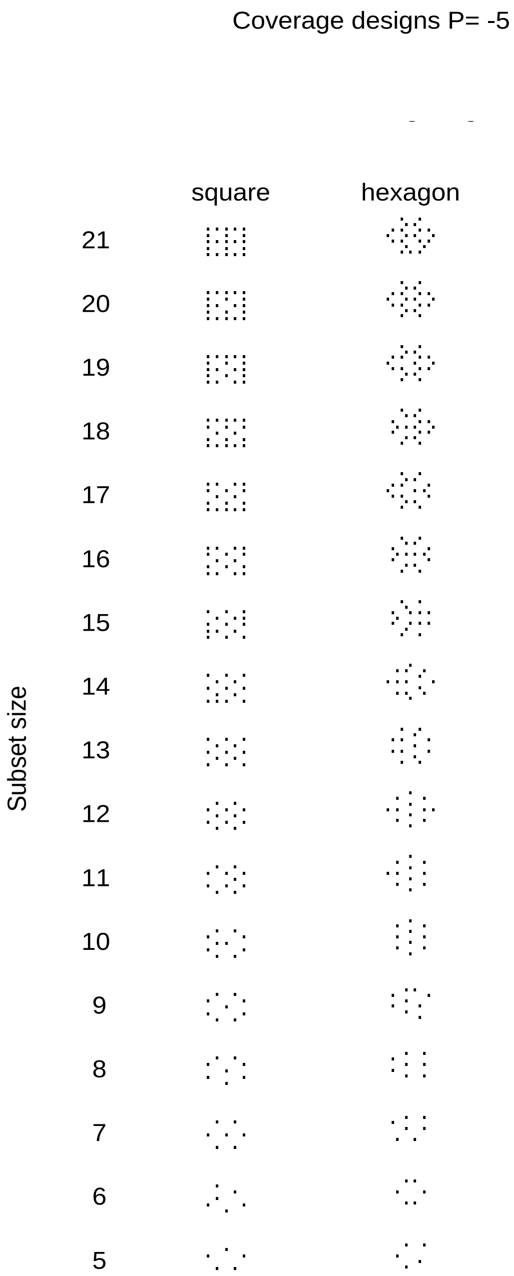

Figure 6: Coverage design with $p = 5$ and $q = 1$ for the square and hexagonal lattices.

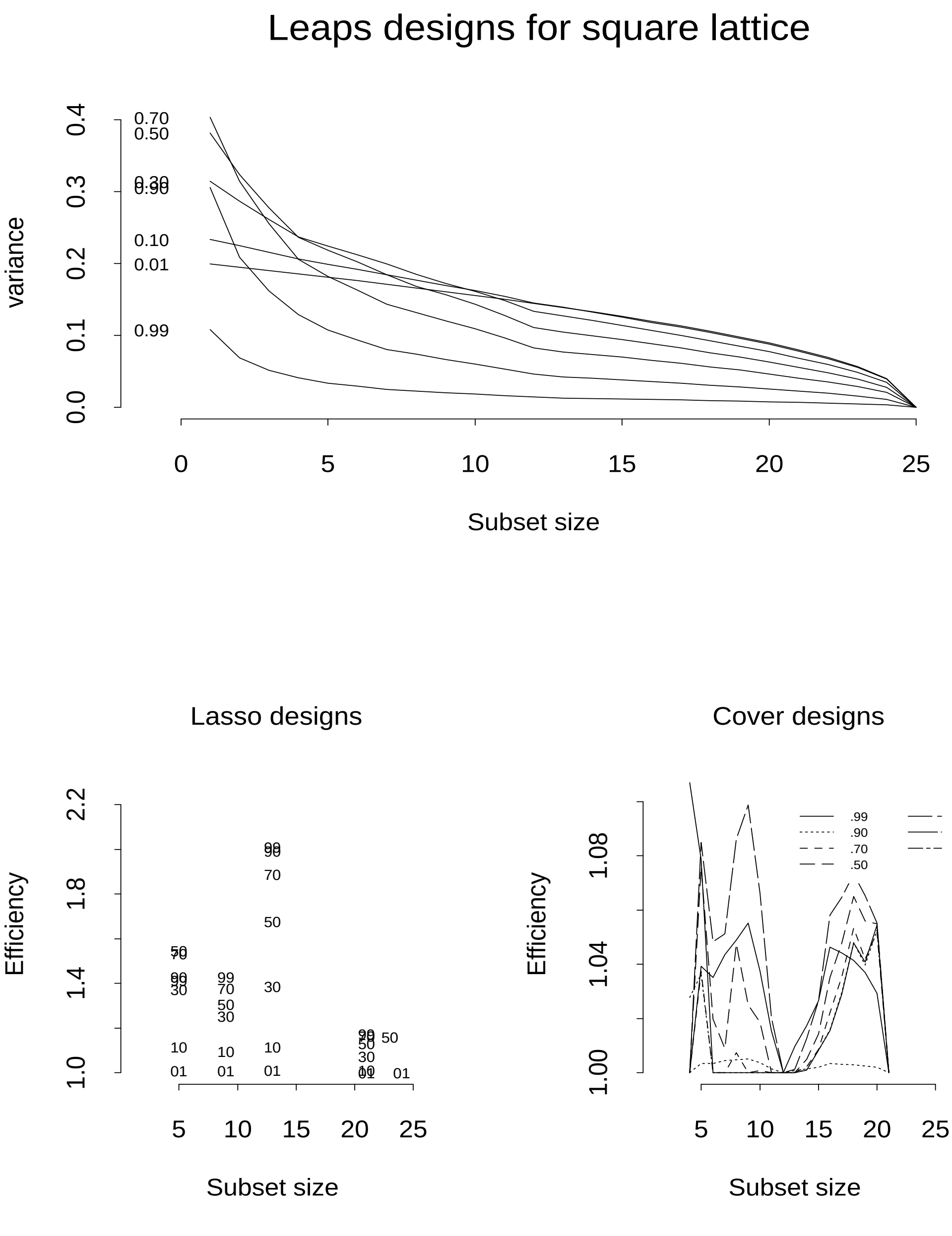

Figure 7: Comparison of the variance of the predicted average for the square candidate set. The first plot gives the optimal prediction variance for different size leaps designs and different parameters of the exponential covariance function. The nearest neighbor correlations for each choice are given next to the vertical axis. The second plot are the efficiencies for the lasso and coverage designs relative to the leaps designs. Coverage efficiencies are joined by lines and the lasso efficiencies are indicated by numerals for the different covariance models.

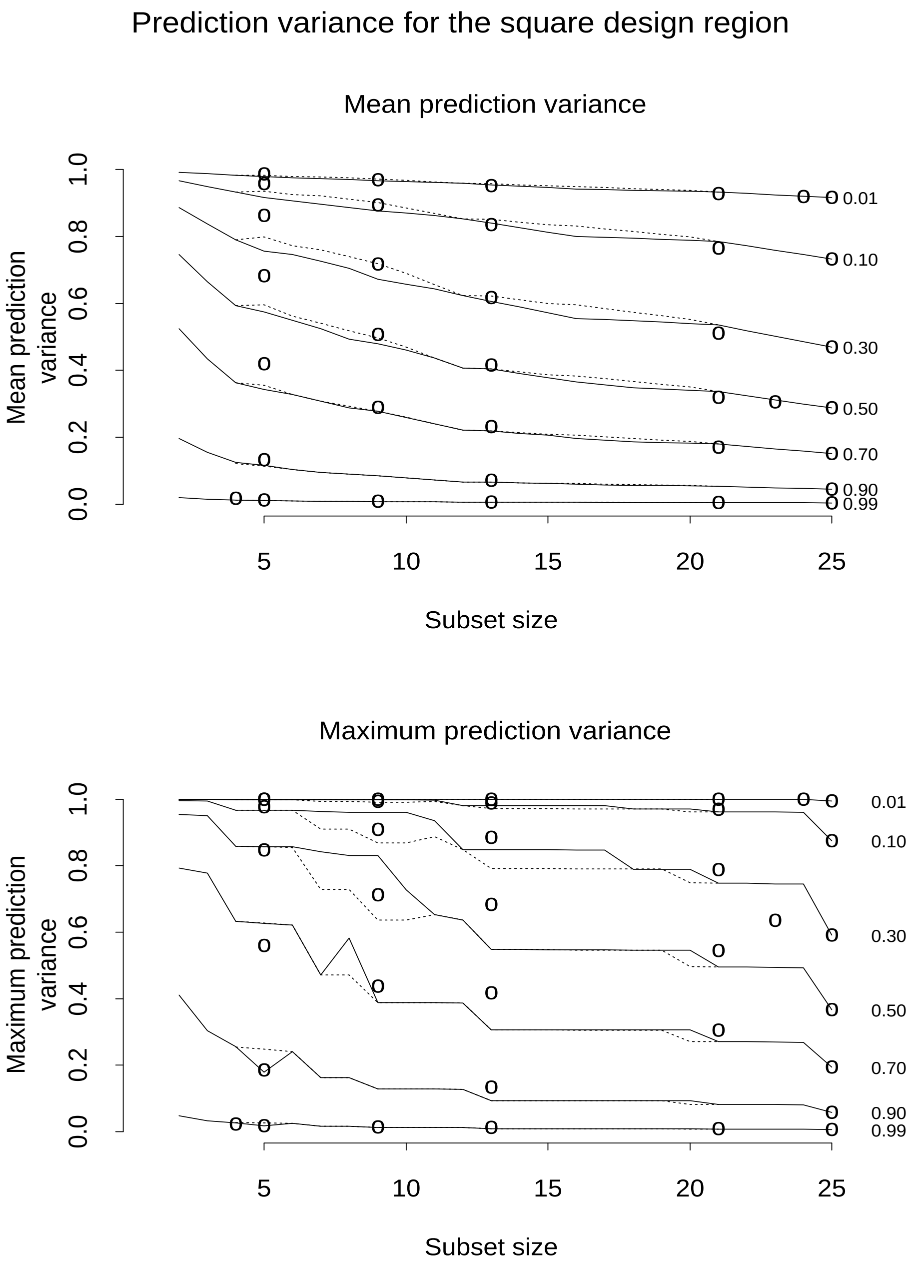

Figure 8: Comparison of the prediction variance of designs for a square lattice. The two plots indicate the performance of the leaps (solid line), lasso (points) and coverage (dashed line) for the average prediction variance and the maximum prediction variance over the different exponential covariance models. Both summary statistics were calculated from a 40×40 grid covering the design region.

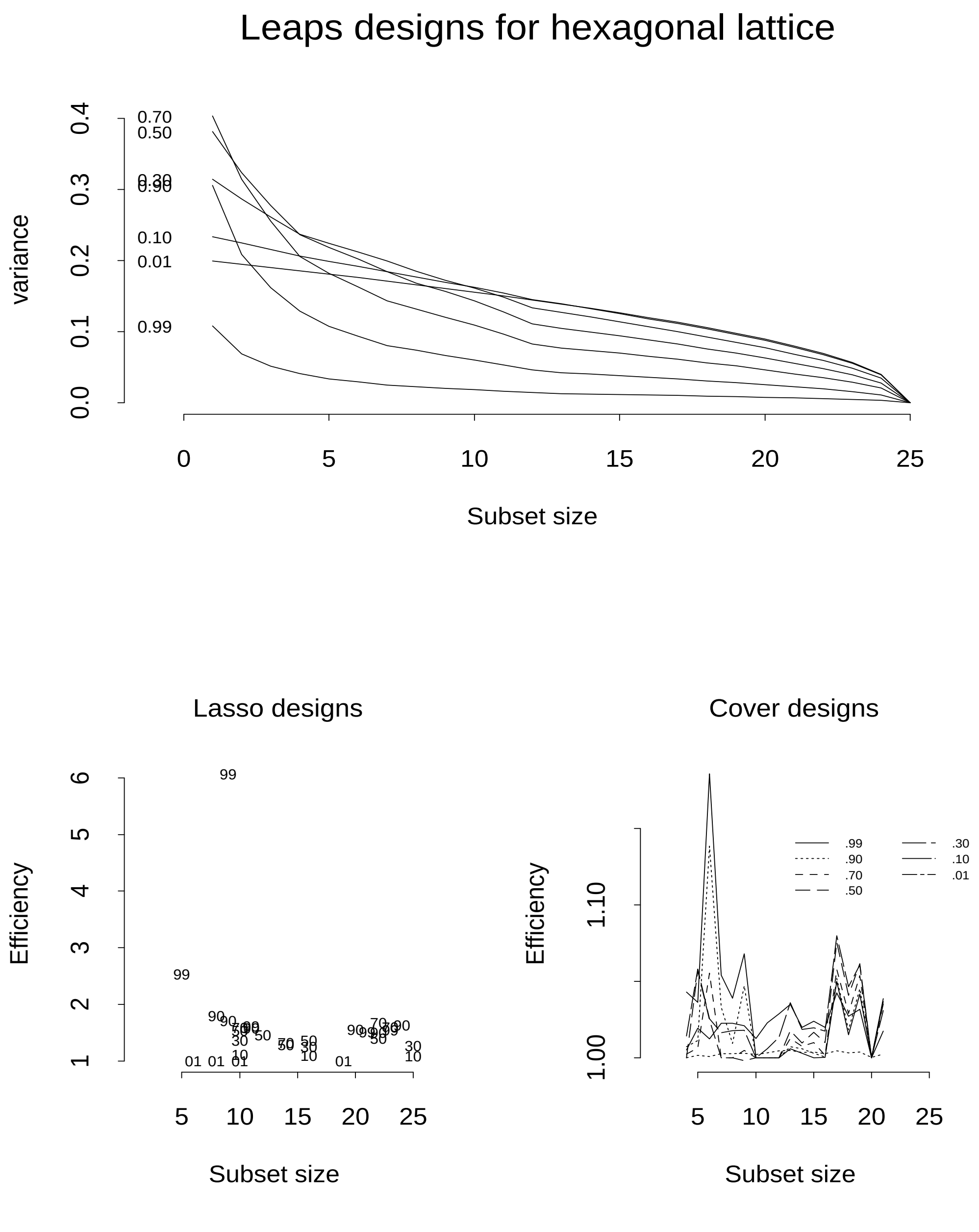

Figure 9: Comparsion of variance of the predicted average for the hexagonal lattice.

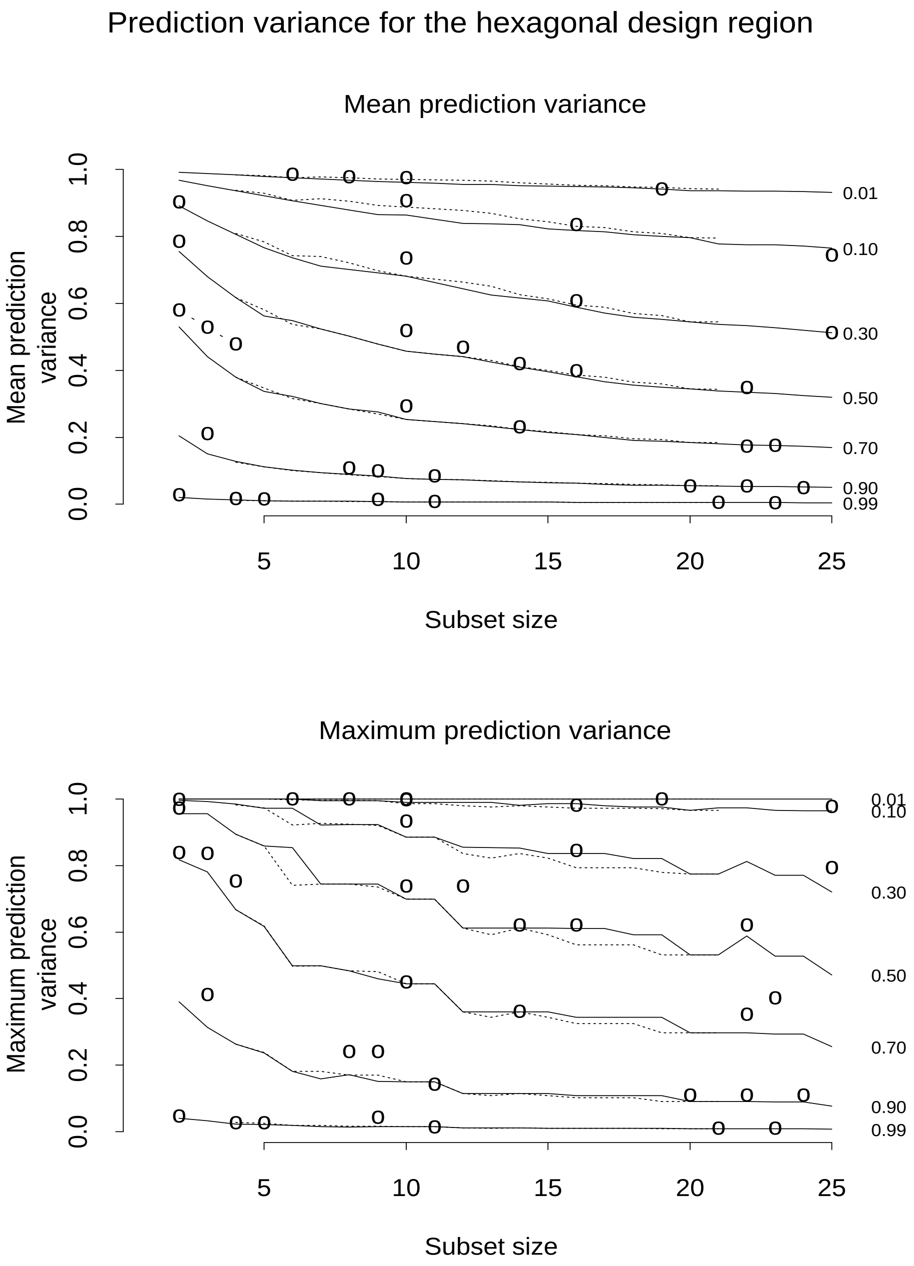

Figure 10: Comparsion of prediction variance for the hexagonal lattice.

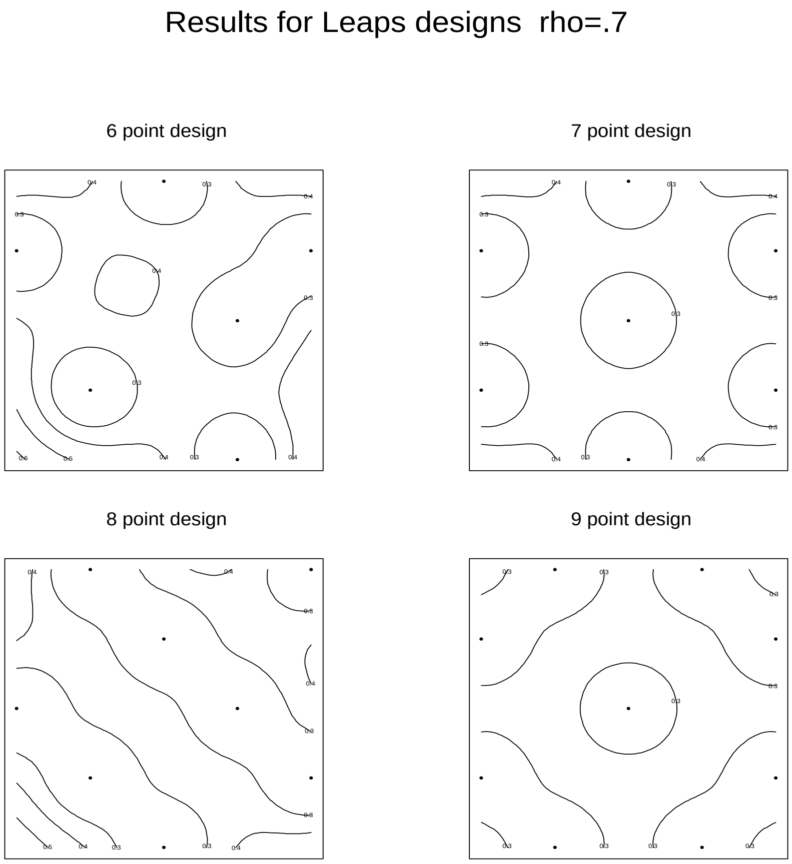

Figure 11: The prediction variance surfaces of leaps designs for the square lattice. Contour plots indicate the prediction variance for design sizes of 6,7,8 and 9 points and an exponential covariance function with correlation .7 .

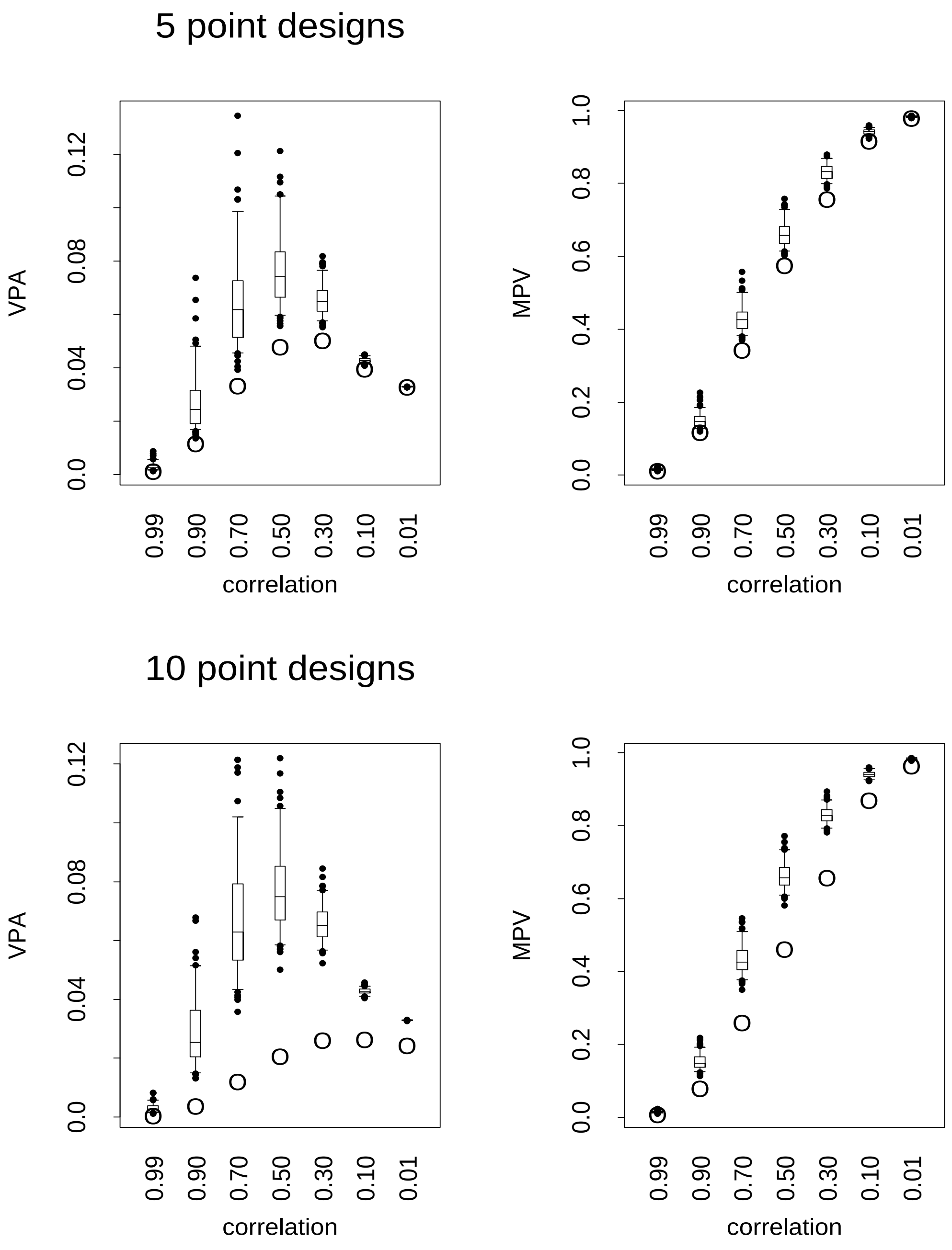

Figure 12: Comparison of leaps designs to designs based on random sampling. The boxplots are all based on 100 designs that were generated by taking random samples of the candidate set. (The whisker length of the boxplots are determined by the 5 and 95 percentiles of the data.) The two top plots consider designs of size 5 and the second are of size 10. The labels along the horizontal axis indicate the exponential covariance and the plotted circles are the values associated with the leaps designs.

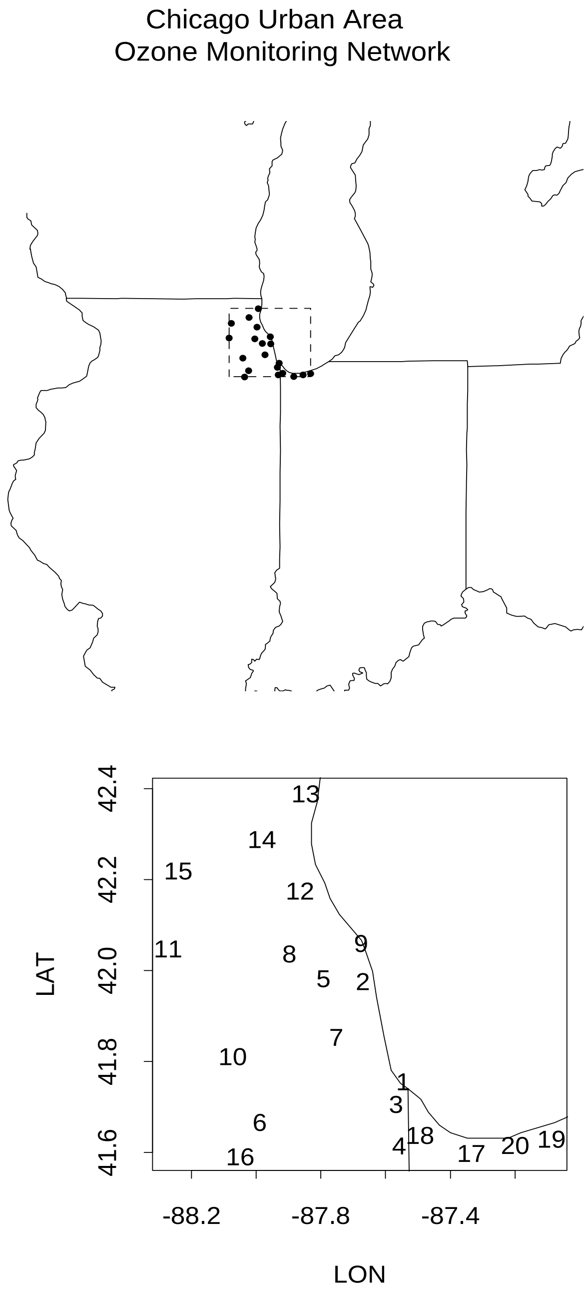

Figure 13: Locations of 20 ozone monitoring stations in the Chicago urban area.

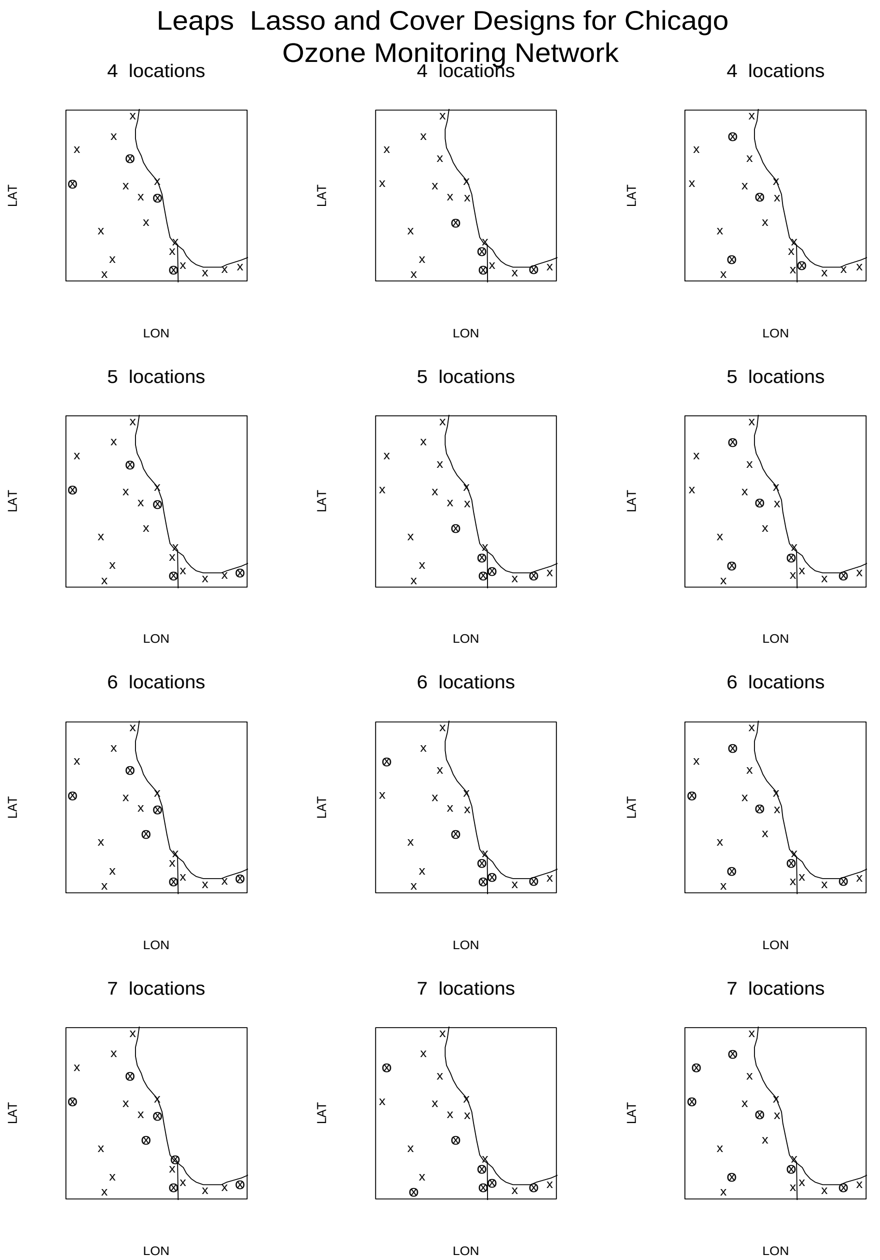

Figure 14: Leaps, Lasso and Coverage designs for the Chicago urban area based on observed ozone. The leaps and lasso designs were based on the daily 8 hour average ozone measured over the period 6/3/87- 8/30/87.

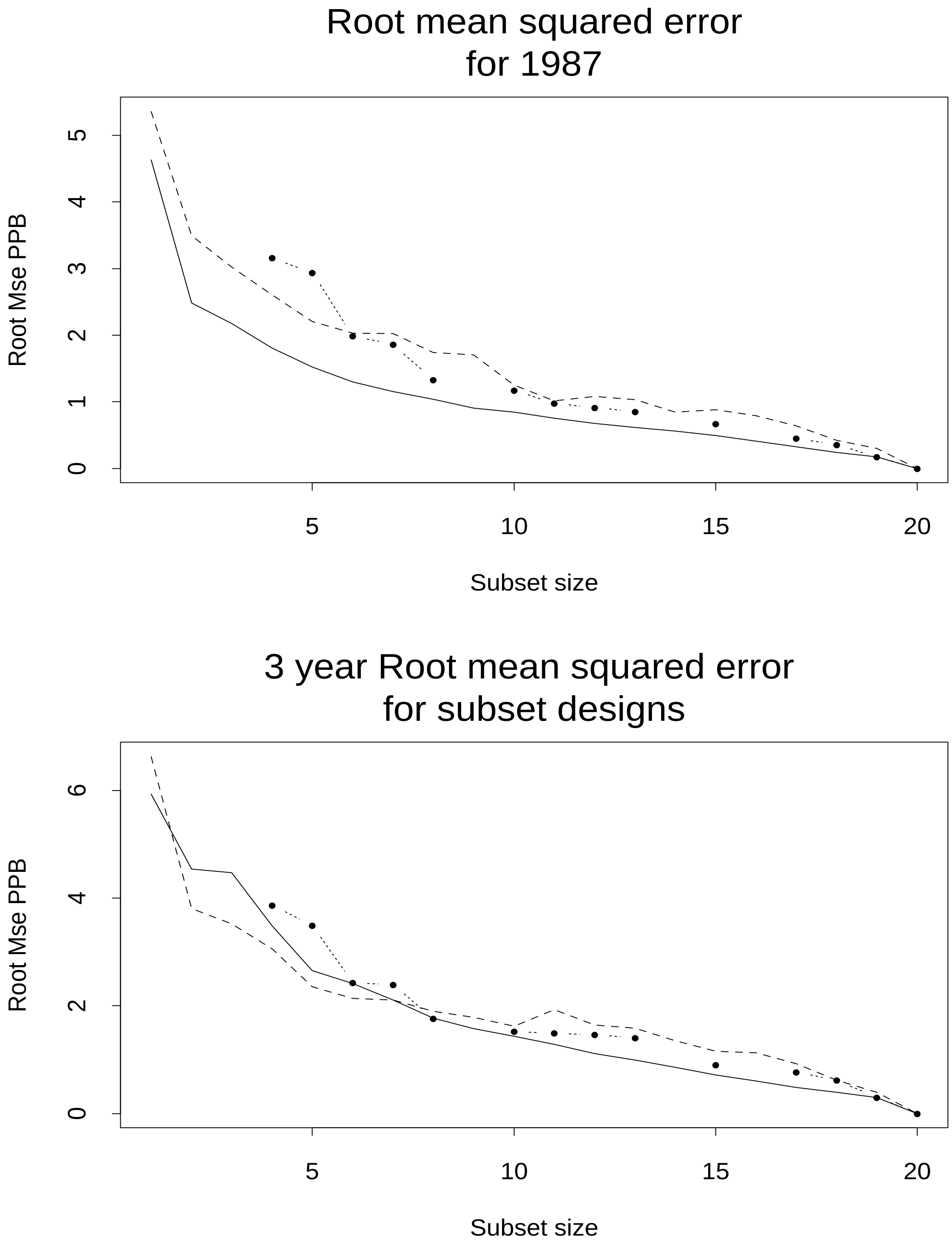

Figure 15: Root mean squared error for the network average for the different designs. The first plot is the root mean squared error between the predicted average and the full network average observed for each day over the 6/3/87-8/30/87. The data over this period has also been used to select the designs. The second plot is the same comparison but for the three subsequent summers 1988, 1989 and 1990. Leaps designs (solid), lasso (points) and coverage (dashed line).