2.5

IMPROVING FORECAST VERIFICATION THROUGH NETWORK DESIGN

Eric Gilleland*

National Center for Atmospheric Research

1 INTRODUCTION

Methods are explored here for optimizing the network design for verification of forecasts of cloud ceiling height and visibility. Meteorological Aeronautical Report (METAR) data from surface stations are used for verification of these forecasts and the placement of these stations may affect the results of the verification. For instance, verification analyses in areas with densely located METAR stations may penalize poor forecasts–that cover several grid points in a region–several times for essentially the same mistake. It is also possible for forecasts to not be penalized enough in areas where METAR stations are only sparsely located. This second problem is not as serious because the forecasts are generally considered more important where METAR stations are abundant.

Although the motivation for this work is in forecast verification, actual verification techniques will not be discussed here. For more information on verification analyses please see Jolliffe (2003) or Wilks (1995). Here, the concern will be focused on spatial statistical methods. Theoretical development of spatial statistical methods can be found in Cressie (1993) or Stein (1999). A more basic knowledge can be found in Reich (2003) or Isaaks (1990).

Various spatial methods are explored for optimizing the network design. Section 3 illustrates these methods. Often, spatial sampling design is used for sampling points from a grid of locations and being able to rearrange the design locations (see, for example, Angulo (2003) or Müller (2003)). Here, however, points are not from a grid nor is it possible to move stations; it is desired only to thin an existing network. Standard spatial techniques are not appropriate for these data because even though they can be viewed as continuous, both are actually too discrete for successfull interpolation. Categorical kriging is more appropriate for these data. However, such techniques have not been used for this type of analysis and so using categorical kriging for finding an optimal network design, particularly for spatiotemporal data, is a matter for further research and is beyond the scope of this paper.

The covariance function used for categorical kriging, however, has a useful interpretation for finding an optimal network design. Therefore, a coverage design was employed using this covariance as a dissimilarity metric. Results showed that the network can be thinned for visibility data, but for ceiling height such thinning based on this analysis is not appropriate.

*Corresponding author address: Eric Gilleland, National Center for Atmospheric Research, Research Applications Program, Boulder, CO 80307-3000; email: ericg@ucar.edu

2 DATASETS

Data used here are hourly data collected from METAR stations from January 1 to January 30, 2003. For this analysis a subset of 48 stations in Northern California and parts of Nevada as shown in Figure 1 are used.

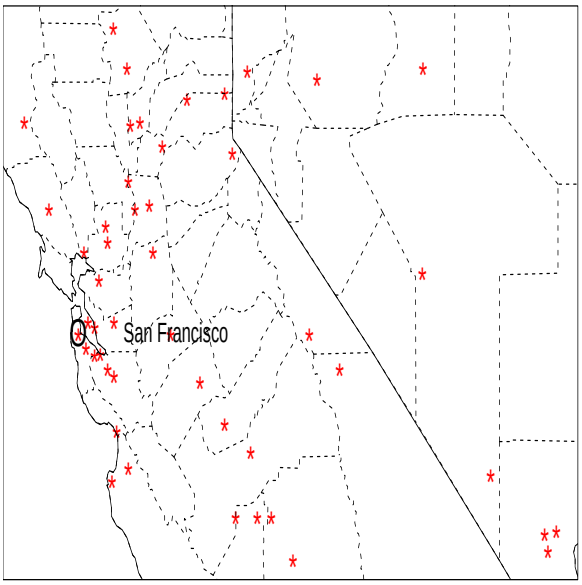

Figure 1: *METAR station locations used for exploring network design issues for verification of cloud ceiling height and visibility.*

For each of these 48 locations, there are 719 hourly time points yielding a total of $34,512$ observations. For cloud ceiling height there are a total of $3,836$ missing data points and for visibility there are $3,844$. Both of

Table 1: *Designations for Low Instrument Flight Rules (LIFR), Instrument Flight Rules (IFR) Marginal Visual Flight Rules (MVFR), and Visual Flight Rules (VFR).*

Flight rules	Cloud ceiling height	Visibility
LIFR	< 500 feet	< 1 mile
IFR	< 1000 feet	< 3 miles
MVFR	< 3000 feet	< 5 miles
VFR	> 3000 feet	> 5 miles

these datasets are theoretically continuous. That is, cloud ceiling height is the distance from the ground to the lowest layer of cloud and visibility is the distance one can see horizontally. However, the data are measured and recorded somewhat discretely as can be seen in Figures 2 and 3.

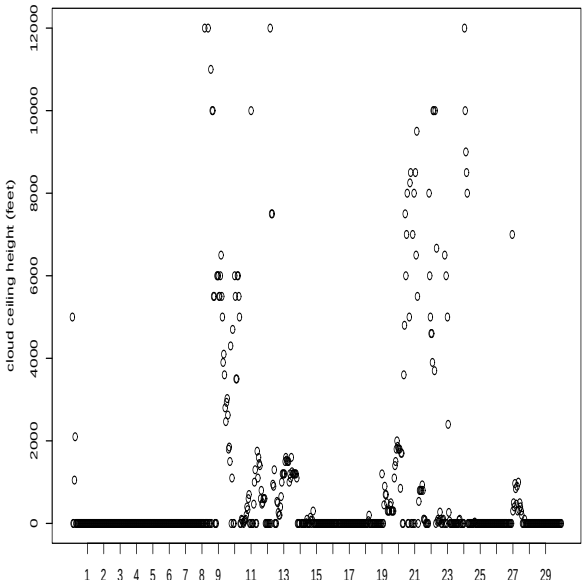

Figure 2: *Example scatter plot of ceiling height data for one station.*

Such discreteness in the data suggests the use of categories. Perhaps the most natural categories to use are the flight rules shown in Table 1.

3 STATISTICAL ANALYSES

For network design it is of interest to know whether information on a variable of interest from a particular group of sites is sufficient for a particular area or not. To learn this information it is generally necessary to perform some type of interpolation using a subset of the spatially located sites to determine if they predict well onto the entire set of sites or not. The standard statistical tools for this analysis involve finding the

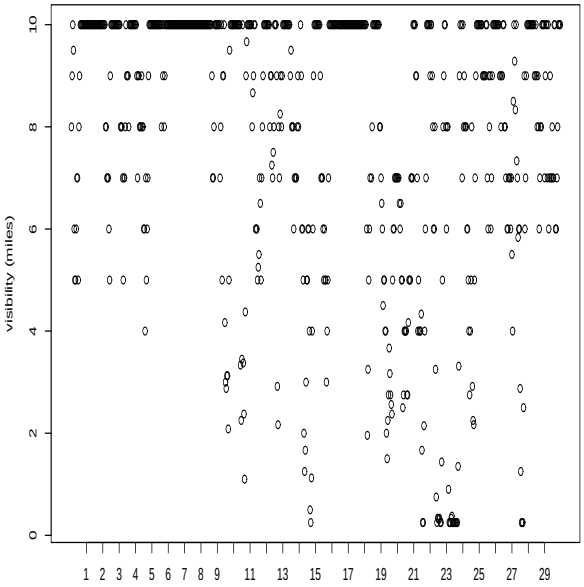

Figure 3: *Example scatter plot of visibility data for one station.*

best linear ubiased estimator using a covariance function that accounts for spatial correlations; a technique widely known as kriging. In this case, however, kriging is not appropriate because it requires the data to follow a Gaussian distribution and due to the discrete nature of these data they are certainly not Gaussian.

As with standard linear regression, however, it is possible to generalize the technique of kriging to non-Gaussian data. In particular, it is possible to use what Reich (2003) refers to as categorical kriging. First, denote the probability that a point $\mathbf{x}_s$ belongs to category i by

$$p_i(\mathbf{x}_s) = \text{prob}\{\mathbf{x}_s \in \text{category } i\},\ i = 1, \ldots, N$$

where N is the number of categories. It is desired to find the estimate of $p_i(\mathbf{x}_s)$, call it $\hat{p}_i(\mathbf{x}_s)$, for each category and then choose the category with the highest estimated "*probability*" as the value at location $\mathbf{x}_s$ subject to the constraint

$$\sum_{i=1}^{N} \hat{p}_i(\mathbf{x}_s) = 1 \tag{1}$$

for all locations, $\mathbf{x}_s$.

The general form of the estimate, $\hat{p}_i(\mathbf{x}_s)$, is similar to the usual kriging estimate. Namely,

$$\hat{p}_i(\mathbf{x}_s) = \sum_{\alpha} \lambda_{\alpha i} I(\mathbf{x}_\alpha \in i) \tag{2}$$

where $I(\mathbf{x}_\alpha \in i)$ is 1 if the value at location $\mathbf{x}_\alpha$ falls in category i and zero otherwise. To verify the constraint

(1) it is sufficient to assume that the weights, λ, are constant for all categories (i.e. $\lambda_{\alpha 1} = \lambda_{\alpha 2} = \ldots = \lambda_{\alpha N} = \lambda_\alpha$) and to impose the unbiasedness condition $\sum \lambda_\alpha = 1$ (see Journel (1983)).

Using a geostatistical approach for finding the new estimator

$$\hat{p}_i(\mathbf{x}_s) = \sum_\alpha \lambda_\alpha I(\mathbf{x}_\alpha \in i) \tag{3}$$

requires the use of a single covariance for all categories, which can be defined as the probability of two points separated by a vector, h, belonging to the same category (see Soares (1992)). Specifically, define

$$C(h) = E\{\sum_i (I(\mathbf{x}_s \in i) \cdot I(\mathbf{x}_s + h \in i))\} \tag{4}$$

It is desired to find the estimator that minimizes the expected sum of the sqaured differences between estimated and real values. That is, it is desired to minimize the quantity

$$\sum_i E\{I(\mathbf{x} \in i) - \hat{p}_i(\mathbf{x})\}^2 \tag{5}$$

The minimization of (5) does not imply that the error for each category is minimized. However, it is desired to find a consistent estimator for the entire set (see Soares (1992)).

Minimizing (5) subject to the constraint that $\sum_\alpha \lambda_\alpha = 1$ leads to the classical kriging system

$$\sum_\alpha \lambda_\alpha C(\mathbf{x}_\alpha, \mathbf{x}_\beta) + \mu = C(\mathbf{x}_\beta, \mathbf{x}) \tag{6}$$

where $C(\mathbf{x}, \mathbf{y}) = C(h)$ and h is the distance vector between the points $\mathbf{x}$ and $\mathbf{y}$. The resulting weights ensure that the constraint equation (1) is satisfied and that (5) is minimized. See Soares (1992) for more details on this type of kriging.

As stated earlier, the general practice is to obtain the estimates, $\hat{p}_i(\mathbf{x})$, for each location, $\mathbf{x}$, and use as the predicted category for a given location that category whose associated estimate is the highest. Assuming a multinomial distribution, an estimate of the standard error of prediction associated with the probability of classifying a location in a given category is given by

$$\sigma(\hat{p}_i(\mathbf{x})) = \sqrt{\hat{p}_i(\mathbf{x}_s)(1 - \hat{p}_i(\mathbf{x}_s))/(n-1)} \tag{7}$$

where n is the number of nearest neighbors used in the kriging (see Reich (2003)).

The standard errors given in (7) have some drawbacks as far as deciding on an optimal network design. Particularly, it is possible to select a particular category, say i, for a site, $\mathbf{x}_s$, with $\hat{p}_i(\mathbf{x}_s) = 1$ even if the true category is not i. Despite this, the standard error of prediction is zero; implying that the prediction cannot be wrong even though it actually is wrong.

The above methods provide a way to employ information from surrounding sites to predict the category for another site. Nevertheless, it is not clear how to use this information to determine which design is "best". These methods have previously been used more to determine, for example, soil composition. In that setting zones of uncertainty can be established to give the researcher an idea of what soil type is most likely to be present in a given area. In this setting, one would need to find the zones of uncertainty for each time point; in this case 719 of them making it difficult to use as a measure of "best" design. The covariance function obtained using these methods may be useful as a measure of similarity (or dissimilarity) in a coverage design (see Nychka (1998) or Johnson (1990)).

Specifically, for a given set of candidate points, C, denote the set of n design points as D where $D \subset C$, then an overall average criterion is an L_q average of cover points in the design region. Namely,

$$(\sum_{\mathbf{x}' \in C} (\sum_{\mathbf{x} \in D} d(\mathbf{x}, \mathbf{x}')^p)^{q/p})^{1/q} \tag{8}$$

where $p < 0$, $q > 0$ are parameters and $d(\mathbf{x}, \mathbf{x}')$ is a distance metric or in the present case a dissimilarity metric. That is, if the covariance function (4) is thought of as a correlation matrix (all values are between zero and one) then a dissimilarity metric would be $d(\mathbf{x}, \mathbf{x}') = 1 - \rho(\mathbf{x}, \mathbf{x}')$, where $\rho(\mathbf{x}, \mathbf{x}') = C(h)$ from (4) and h is the distance vector between $\mathbf{x}$ and $\mathbf{x}'$. Large negative values of p tend to yield designs that are more spread out and as $q \longrightarrow \infty$ and $p \longrightarrow -\infty$ the result gives a classic minimax design.

Criterion (8) is minimized over several space-filling designs of a given size to obtain a "coverage design" from among the class of space-filling designs. It is possible to fix points in the design so that they cannot be swapped out. Generally, the initial design is chosen at random and choice of starting design may affect the outcome. Criterion (8) is guaranteed to converge (Nychka (1998) or Johnson (1990)).

Note that this method gives a subset of a predetermined size, n, that is "best" based on the dissimilarity metric, which in this case is the probability of two points separated by a distance vector, h, belonging to the same class; it does not determine the "best" size, n, of the network. To attempt to find the "best" size, one possibility is to use the coverage design algorithm to find several designs of varying sizes and then use generalized cross-validation (GCV) with categorical kriging to help decide on the design size. This method is still somewhat crude because the larger the design size, the smaller the GCV will be, but it can give some

idea of how each design size performs.

4 RESULTS

The analyses described in the above section allow for an anisotropic covariance function; meaning that the covariance may depend on direction as well as distance. In the following analyses all covariances are assumed to be isotropic; although it may be of interest to look at the case of anisotropy. Additionally, the covariances (probability that two stations separated by a distance vector h are in the same class) are calculated using the temporal component of the data. So that (4) becomes

$$C(h) =$$

$$E_t\{E\{\sum_i (I(\mathbf{x}_s \in i; t) \cdot I(\mathbf{x}_s + h \in i; t))\}\} \qquad (9)$$

where E_t is the expectation over time.

Figures 4 and 5 show the scatter plots for the empirical covariance functions (9) for cloud ceiling height and visibility respectively. In each case, the plots appear to be quite scattered and suggest that it may be inappropriate to thin these networks. Nevertheless, in the case of visibility, a mixture of exponential functions is fit to these correlations in order to apply the coverage design technique discussed in section 3.

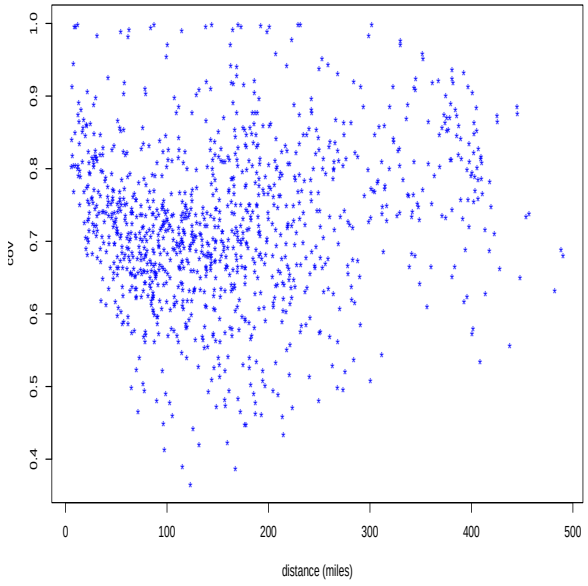

Figure 4: *Scatter plot of distance and probability of two classes separated by given distance being in the same class (covariance) for cloud ceiling height data.*

Using the covariance fit shown in Figure 5 for the visibility data, a coverage design is found for 20 design

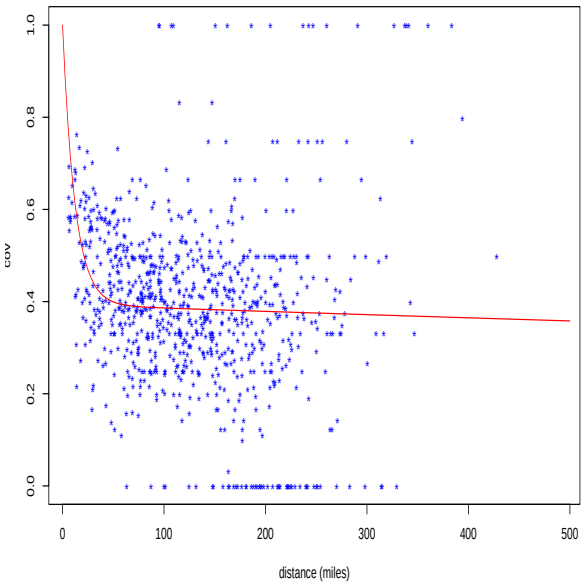

Figure 5: *Scatter plot of distance and probability of two classes separated by given distance being in the same class (covariance) for visibility data. Red line indicates the fitted mixture of exponential functions to these data. Specifically,* $C(h) = 0.61\exp(-h/12.15) + (1-0.61)\exp(-h/5325.39)$.

points out of the total 48 METAR stations. Values of $p = -20$ and $q = 20$ are used in order to get a reasonably spread out design that is close to a minimax design. The resulting design is shown in Figure 6. The R software package *fields* (Nychka (2003)) is used to find the coverage design; specifically the function *cover.design*.

Bootstrapping is used to obtain some kind of inference about the parameters for the covariance shown in Figure 5. Values for the mixing parameter are very tight with a standard deviation of only 0.011 and the values range from about 0.57 to 0.62 indicating that the short range exponential is more important than the long range one. For the short range parameter the values are also very tight with a standard deviation of about 1.1 miles and a range of about 9 miles to about 15 miles. The long range parameter varies widely, but its minimum value is roughly 1200 miles indicating that this term is nearly zero across all simulations. These results suggest that a single exponential is adequate. However, the mixture has the nice property of capturing the short range correlation much better than a single exponential.

The map from Figure 1 suggests a smaller region should be studied independently. Specifically, the nine stations on the southern edge of the San Francisco Bay should be considered separately. Figures 7 and 8 show the empirical covariances (9), but restricted to

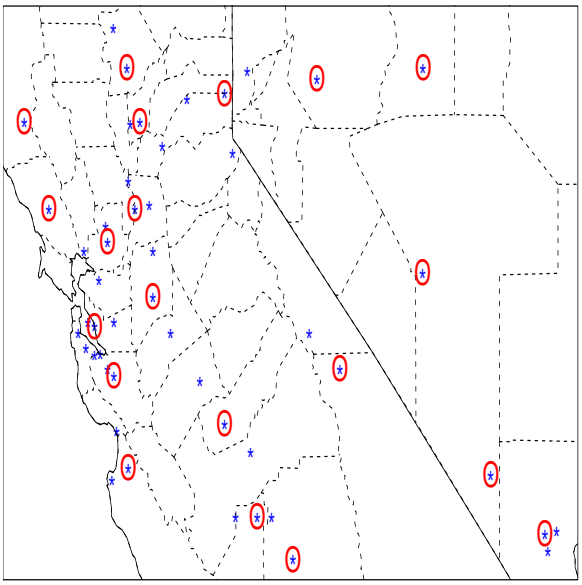

Figure 6: *Coverage design results (circled) for visibility data using a design of size 20 out of 48 candidate stations using dissimilarity metric* $1 - \rho(h)$ *where* $\rho(h) = 0.61\exp(-h/12.15) + (1 - 0.61)\exp(-h/5325.39)$.

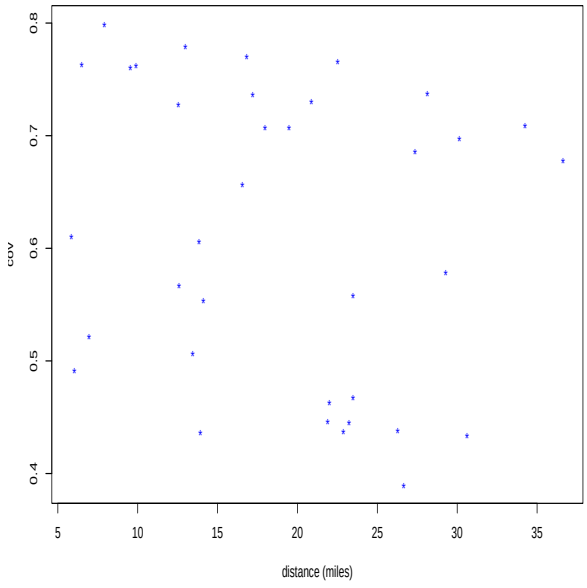

Figure 7: *Scatter plot of distance and probability of two classes separated by given distance being in the same class (covariance) for subset of cloud ceiling height data.*

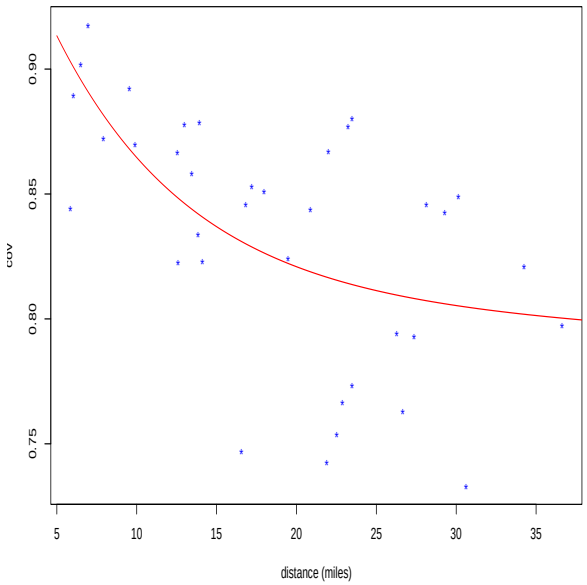

Figure 8: *Scatter plot of distance and probability of two classes separated by given distance being in the same class (covariance) for subset of visibility data. Red line indicates the fitted mixture of exponential functions to these data. Specifically,* $C(h) = 0.19\exp(-h/8.47) + (1 - 0.19)\exp(-h/2603.90)$.

this small region for each set of data. In the case of cloud ceiling height, the results do not change as the probabilities are still quite scattered from less than 50% to nearly 1 for distances less than 5 miles. On the other hand, the empirical covariance for the visibility data looks very promising with high probabilities for all distances, but still decreasing with distance. In fact, the probabilities do not drop below 0.8 until about 15 miles. There appears to be much more structure here. The resulting design for this subset is shown in Figure 9.

Bootstrapping results for the covariance parameters shown in Figure 9 for the Bay Area subset are similar to the results found for the entire 48 METAR stations, but the long range parameter appears to be more important here. The mixing parameter ranges from 0.33 to 0.52 with a standard deviation of about 0.04 suggesting that the two exponential components are more equally necessary than before. The short range parameter ranged from about 2 miles to about 6 miles with a standard deviation of about 0.85 miles and the long range parameter ranged from about 60 miles to about 1400 miles and was more varied than either of the other two parameters with a standard deviation of about 200 miles.

Given the unique structure of this particular subset of data, it is perhaps a good idea to fix this smaller design when doing the analysis on the entire network. Also, it may be a good idea to fix certain locations

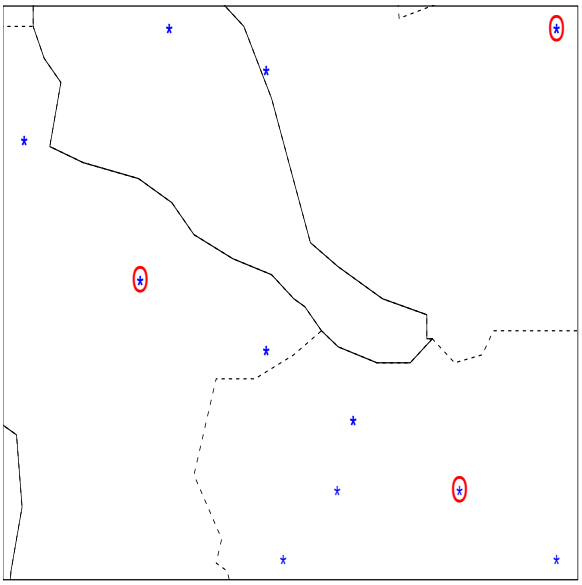

Figure 9: *Coverage design results (circled) for visibility data using a design of size 3 out of 9 candidate stations using dissimilarity metric* $1 - \rho(h)$ *where* $\rho(h) = 0.19\exp(-h/8.47) + (1 - 0.19)\exp(-h/2603.90)$.

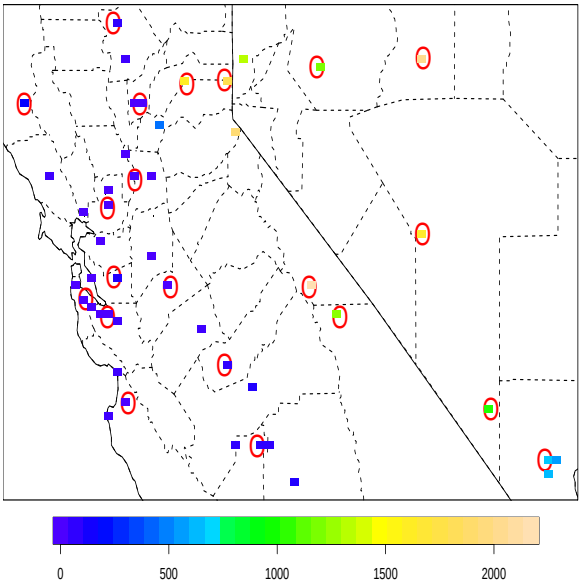

Figure 10: *Coverage design results (circled) for a design of size 20 (fixing some stations due to elevation differences and from finding a "best" design for a subset of stations near the San Francisco bay) out of 48 candidate stations. Elevations are shown in color.*

based on elevation differences. Figure 10 shows the results of performing the same coverage design analysis as in Figure 6, but with the three design points from Figure 9 fixed in the design along with five points where sharp changes in elevation occur. The coverage design algorithm was run several times with different randomly chosen starting designs to ensure that the choice of starting design would not affect the result. The generalized cross-validation value from categorical kriging obtained for this design is about 0.69, which is about 0.25 higher than the "best" design using 25 stations and nearly 0.3 lower than using only 15 stations.

5 DISCUSSION

Although the coverage design finds the best design for a particular design size, it does not find the "best" design size. Here, a "best" size design is found by running the coverage design analysis for several design sizes and then a cross validation analysis using the categorical kriging described in section 3 is used to obtain a heuristic idea of a good design size. As with all cross validation analyses, the value will increase with fewer stations and it is not clear how to decide how large of a value is too large. Inspection of plots of visibility values for several time points suggests that the final design (Figure 10) appears to be quite reasonable.

Results look quite good for visibility data, but not for cloud ceiling height where the empirical covariances suggested that all of the stations are necessary for verification. This may be due to isolated thunderstorms that may cover one station, but not other stations only a few miles away. In this case, network thinning is indeed not appropriate for forecast verification. However, the sensitivity of many ceilometers makes it possible for a single station to read a low ceiling height if, for example, something other than a cloud such as birds or aircraft were to fly over at the right time. In this case, network thinning may be reasonable. It may be possible to determine which is the case by using other data sources, such as satellite data, to determine if there are isolated thunderstorms or not. However, while the presence of clouds over stations may be detected by a satellite, the actual ceiling height could still differ substantially from one station to another.

ACKNOWLEDGEMENTS

This research is in response to requirements and funding by the Federal Aviation Administration (FAA). The views expressed are those of the authors and do not necessarily represent the official policy or position of the FAA. NCAR is sponsored by the National Science Foundation.

REFERENCES

Angulo, J.M., Ruiz-Medina, M.D., Alonso, F.J., Bueso, M.C., July 2003: Generalized approaches to spatial sampling design, *The ISI International Conference on Environmental Statistics and Health: Conference Proceedings*, 11-19.

Cressie, Noel A.C., 1993: Statistics for Spatial Data (Revised Edition), *Wiley Interscience*, 605 Third Avenue, New York, NY 10158-0012.

Isaaks, Edward H. and Srivastava, R. Mohan (contributor), 1990: An introduction to applied geostatistics, *Oxford University Press*, New York, NY 10010.

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

Jolliffe, Ian T. and Stephenson, David B., 2003: Forecast verification: a practitioner's guide in atmospheric science, *John Wiley and Sons Inc.*, 111 River Street, Hoboken, NJ 07030.

Journel, A.G., 1983: Non-parametric estimation of spatial distribution, *Mathematical Geology*, 15 (3), 445-468.

Müller, W.G., July 2003: Spatial design methods in case of correlated observations: a comparison, *The ISI International Conference on Environmental Statistics and Health: Conference Proceedings*, 21-32.

Nychka, Douglas and Saltzman, Nancy, 1998: Design of Air Quality Monitoring Networks, *Lecture Notes in Statistics: Case Studies in Environmental Statistics*, *Springer*, 175 Fifth Avenue, New York, NY 10010.

Nychka, D., Bailey, B., Ellner, S., Haaland, P., O'Connell, M., Hardy, S., Baik, J., Meiring, W., Royle, J.A., Fuentes, M., Hoar, T., Tebaldi, C. and Gilleland, E., 2003: fields: a collection of programs based in R/S for curve and function fitting with an emphasis on spatial data, *http://www.cgd.ucar.edu/stats/Software/Fields/*.

Reich, Robin M. and Davis, Richard, 2003: Quantitative spatial analysis, *Course Notes for NR/ST 523*, Colorado State University, Fort Collins, Colorado 80523

Soares, Amilcar, 1992: Geostatistical estimation of multi-phase structures, *Mathematical Geology*, 24, 149-160.

Stein, Michael L., 1999: Interpolation of spatial data: some theory for kriging, *Springer-Verlag*, 175 Fifth Ave., New York, N.Y. 10010.

Wilks, Daniel S., 1995: Statistical Methods in the Atmospheric Sciences, *Academic Press*, 525 B Street, Suite 1900, San Diego, CA 92101-4495.