DOE/EY/22874-56

STATISTICAL CONSIDERATIONS IN NETWORK DESIGN

PAUL SWITZER
Department of Statistics
Stanford University

WORKING PAPER #10

AUGUST 1979

Working papers represent work which is either not intended for publication or not ready for publication. They are circulated in "rough" form because of their topical interest, for comment and criticism by qualified persons associated with the SIMS Project. For these reasons, working papers are not intended for general circulation.

This research was supported by Grant from SIMS.

Prepared For
THE U.S. DEPARTMENT OF ENERGY
UNDER CONTRACT NO. EY-76-S-02-2874

DISCLAIMER

Statistical Considerations in Network Design

by

Paul Switzer

Abstract. Following some general remarks on the importance and unimportance of optimization in spatial network design we take up, in modest detail, how one might exploit spatial autocorrelations and covariate information. We point out that spatial autocorrelations, themselves, require care in their estimation and then proceed with two illustrations to show how probabilistic error calculations are made for mapping problems using network station data. One illustration uses a quantitative mapping variable and the other uses a qualitative mapping variable.

Remarks

There must be many different hydrological network design problems about which statisticians might have something to say. I will confine my remarks to problems in which the spatial or geographic aspects are the main feature. Such problems typically involve a set of fixed stations, called monitoring stations, which record one or more measurements more or less continuously over time.

Consider the problem of estimation of a space average for a fixed time period, e.g., number of acre-feet of water falling as precipitation on a given drainage basin during a given day or a given year. We assume that the data collection system consists of fixed-site monitoring stations measuring precipitation. As well, we may have covariate data being collected at the monitoring network such as air temperature and also covariate data not related to the monitoring network such as a topographic map, barometric pressure maps, etc. A feature of this problem which distinguishes it from some others we will consider is the

fact that we are estimating a spatial aggregate--as opposed to problems in which we attempt to interpolate or contour the precipitation field.

The network design considerations for such problems depend on whether a network is being designed de nouveau or whether it is an existing network which will be expanded or contracted. The number of monitoring stations needed to meet whatever statistical estimation criteria will depend mainly on factors intrinsic to the variability of precipitation in the given basin and much less on the statistician's cunning in using the data. Of course, if relevant covariate data, such as topography, is ignored by the statistician, then the presumed required network density may be appreciably overstated.

A statistical estimation criterion such as expected root-mean-squared-error (rmse) is a calculation derived from a probability model. Statisticians who do such calculations believe that if true root-mean-squared-errors could be measured then they would on average be equal to their calculated values. Since space averages for rainfall can never be measured the justification of the statistician's faith in probability-based calculations may be drawn from use of the same probability models for spatial interpolation where actual mean-squared-errors could be compared with calculations.

The reasonable assessment of the magnitude of errors of estimation is the statistician's main job. The probability methods he uses serve in lieu of actual measurement of errors where these are never available or are available only at a later time. Probability-error calculations can be grossly inaccurate if the probability model is poorly chosen; for example, this may happen if spatial autocorrelation of data is ignored.

A statistical criterion such as expected rmse may be used to compare different methods of estimating a spatial average from the same available data and even perhaps to suggest an "optimum" estimation method. For example, if we use only the monitoring station data for the time period of interest, how good is an area-weighted estimate versus some other data-weighting scheme? In the context of any probability model, the area-weighted rmse is always suboptimal; however, its rmse is almost always within an epsilon of the optimum. In any event, finding the optimum weighting is a highly model-dependent procedure and it is usually a wasteful exercise. The important exercise is the calculation of the rmse, for whatever data-weighting scheme one has, as I have previsouly emphasized.

The rmse criterion might also be used to resolve certain network design questions, for example finding the optimum configuration of a given number n of monitoring stations. In typical situations the rmse is insensitive to the configuration, provided one avoids cluster-

ing of stations and is somewhat careful to allocate more stations to that part of the basin whose rainfall is thought to be less predictable. To try to optimize the configuration is another one of those wasteful exercises because of the insensitivity of the rmse, the mathematical difficulty of the problem, and the strong model dependence.

Another design problem is the choice of the number n of monitoring stations needed to meet an rmse requirement for the estimation of a spatial average. The required n will depend on the availability and usefulness of covariate data as well as on the size, variability, and autocorrelation properties of the precipitation field itself. So this design problem cannot be solved without first having data--a common statistical conundrum. The usual resolution is to implement a network in stages. The first-stage network is used to estimate the gross statistical properties needed to determine how many, if any, additional stations are needed to meet the rmse requirement. If a second-stage network is implemented, the additional data provided can be used for more refined statistical model building.

The deletion of a station from a network is a simpler design problem inasmuch as the configuration is already fixed and we need only calculate which deleted station will result in the smallest increase in rmse. The addition of a station to a network is not generally proposed in order to reduce the rmse of an estimated spatial average; rather, it is proposed in order to have specific information about a particular site and the siting problem is therefore not completely statistical.

Of course, estimation of spatial averages is not likely to be the only task of the data network. In particular one may wish to estimate other functionals of the spatial frequency function (over a given time interval) such as its median or its quartiles: one-half of the basin area has a rainfall exceeding the spatial median, and quartiles are defined similarly. For an evenly-spaced monitoring network, the usual sample median and quartiles of the station data are reasonable estimates. However, assessing the magnitude of possible estimation errors does require some sophistication because of autocorrelation introduced through spatial continuity.

We now turn to the problem of interpolating a map of the spatial precipitation field using the station data. At each location the map will be in error and the magnitude of this error will, on average, be larger at locations far from a station and at locations of lesser spatial continuity. As an overall statistical criterion of error, we may take the root-mean-squared interpolation error averaged over the basin ($rmse_A$) or else the maximum rmse over the basin ($rmse_M$), for example. The statistician's usual approach to the estimation of these error criteria for any given or putative network will again involve modelling a spatial autocovariance function.

The number of stations needed to satisfy the error criterion should depend slightly, if at all, on their geometric configuration provided reasonable good sense prevails. That is, one puts more stations in areas of higher variability and avoids clustering of stations. The criterion $rmse_M$ will be more sensitive to the network configuration than $rmse_A$.

We suppose that interpolation is done by forming weighted averages of nearby station measurements. Provided stations are not clustered, simple weighting schemes based on distances to stations will be nearly optimal. An estimated "optimal" weighting scheme may be calculated from the estimated autocovariance function. Since the autocovariance function is needed anyway for rmse calculations, it may as well be used also for optimizing the data weighting scheme. If there is some trend or other obvious structure in the mean precipitation function, this should become part of the spatial model: it will affect the method of estimation of autocovariances and it will raise questions of bias avoidance for the map interpolation problem. Fortunately, interpolation bias may be more or less eliminated by imposing appropriate constraints on the choice of data-weighting schemes. This point will be taken up further in the illustrations which follow. For the map interpolation problem we are in the fortunate position that the estimated interpolation errors calculated from the statistician's probability models can be checked against reality to some extent. This is achieved by interpolating the station values themselves as though they were unknown, an exercise worth carrying out routinely as a check on the probability calculations.

It is a mathematical feature of interpolated maps that they are smoother than reality even when "optimally" constructed. One aspect of this extra smoothness is the suppression of precipitation peaks and valleys; for example, it is unusual to interpolate values which exceed the maximum station precipitation. One should be careful, therefore, when using interpolated maps as spatial simulations and, in particular, the spatial frequency function of an interpolated map will be a biased estimator of the true spatial frequency function. Also, the autocovariance function of an interpolated map will be considerably flatter than the autocovariance function estimated from station data only.

A data network may be used to interpolate maps of a qualitative (as opposed to numerical) variable, for example the presence-absence pattern of rainfall on a given day or something like a geologic map based on rock type or soil type. Here a reasonable statistical estimation criterion might be the proportion of the map area colored incorrectly. This criterion is equivalent to a spatially averaged mean-squared-error based on a zero-one error measure. The estimation of such error criteria, in the context of models of probabilistic parti-

tions of the basin, requires fitting spatial autocovariance functions for dichotomous variables, which in turn may be used for deciding upon an appropriate network density.

Estimation of Spatial Variability

All network design problems require a probabilistic assessment of error which in turn requires specification of a spatial autocovariance function (SAF) or a spatial variogram (SV). [In principle we may have different SAF's operating during different time periods. It is convenient, but not always essential, that the SAF be made time-invariant.] If we denote the space-time precipitation field by $Z(x;t)$ then we model it as a random process with mean function $m(x;t)$ and "residual" field $\varepsilon(x;t) = Z(x;t)-m(x;t)$. Its SAF is given by

$$C(x',x'') = E[\varepsilon(x';t)\cdot\varepsilon(x'';t)]$$

where x',x'' are two points in the basin and the expectation operator E is taken with respect to the generating random process. Suppose the n stations in the network are positioned at $x_1,x_2,\ldots,x_n$. It is not straightforward to estimate the spatial autocovariance even between station pairs, viz. $C(x_i,x_j)$, contrary to what is sometimes supposed. For example the "usual" empirical covariance calculation between the two time series of a station-pair does not estimate the spatial C, but instead something complicated involving the mean function and the temporal autocorrelations and temporal crosscorrelations as well as the spatial C. On the other hand, one may first try to estimate the mean function and then calculate the empirical covariance between the two time series of empirical residuals. This is better but still not right.

Rather than estimate the SAF we may estimate a closely related function called the spatial variogram (SV). It is defined as

$$\gamma(x',x'') = \frac{1}{2}E[\varepsilon(x';t)-\varepsilon(x'';t)]^2 .$$

A reasonable method to estimate the SV will now be described; it is adapted from a method used by Pierre Delfiner and others. Suppose the mean function at time t is modeled to be a linear function of the values of time t of a set of K measured covariates $\{v_j\}$, i.e.,

$$m(x;t) = \sum_{j=1}^{K} b_{jt}\, v_j(x;t) + b_{0t} . \qquad (1)$$

The measured covariates may be elevation above sea level, geographic coordinates, air temperature, etc., and the coefficients $b_{0t},b_{1t},\ldots,b_{Kt}$ may vary with time. Suppose the station rainfall measurements

$Z(x_1;t)$,..., $Z(x_n;t)$ are used to obtain an ordinary least-squares estimate of the mean function m, i.e., we do a linear regression of rainfall Z on the covariates $\{v_j\}$ separately at each fixed time t. We obtain calculated residuals D_{it} at each station i of the network at each time period t. Now it can be shown that $E[D_{it}^2]$ is a quantity which does not involve the unknown b's and which can be computed directly from the spatial variogram γ. Hence the D_{it}^2 may be used to estimate the SV without becoming involved in temporal autocorrelations. This method shows that even for a class of time-varying mean functions we may estimate simple combinations of the purely spatial variogram using historical time series. The next step is to fit values of the SV parameters.

To illustrate we use a simple class of spatial variograms depending on two parameters c_1, c_2 which will be estimated from station data records:

$$2\gamma(x',x'') = E\{\varepsilon(x';t)-\varepsilon(x'';t)\}^2 = [c_1 + \| a'-a''\| c_2] \cdot \| x'-x''\| \qquad (2)$$

where $\| x'-x''\|$ is the spatial distance between position x' and x'' and $\| a'-a''\|$ is the difference in altitude covariate. Hence residuals which are further apart spatially are more likely to be different, and residuals corresponding to large altitude disparities are more likely to be different. Suppose further we take the mean function to depend linearly on the altitude covariate, i.e.,

$$m(x;t) = b_{0t} + b_{1t}\, a(x) , \qquad (3)$$

and that we have a four-station network as shown in Figure 1, with station-to-station distances as indicated. The rainfall at each station for each of five time periods is shown in Table 1. This table also shows the altitude of each station. Then the ordinary least-squares regression of rainfall on altitude, at each time t, produces calculated residuals at stations x_2 and x_3 which are linearly independent and proportional, respectively, to

$$D_{2t} = Z(x_1;t) - 3Z(x_2;t) + Z(x_3;t) + Z(x_4;t)$$

$$D_{3t} = Z(x_1;t) + Z(x_2;t) - 3Z(x_3;t) + Z(x_4;t)$$

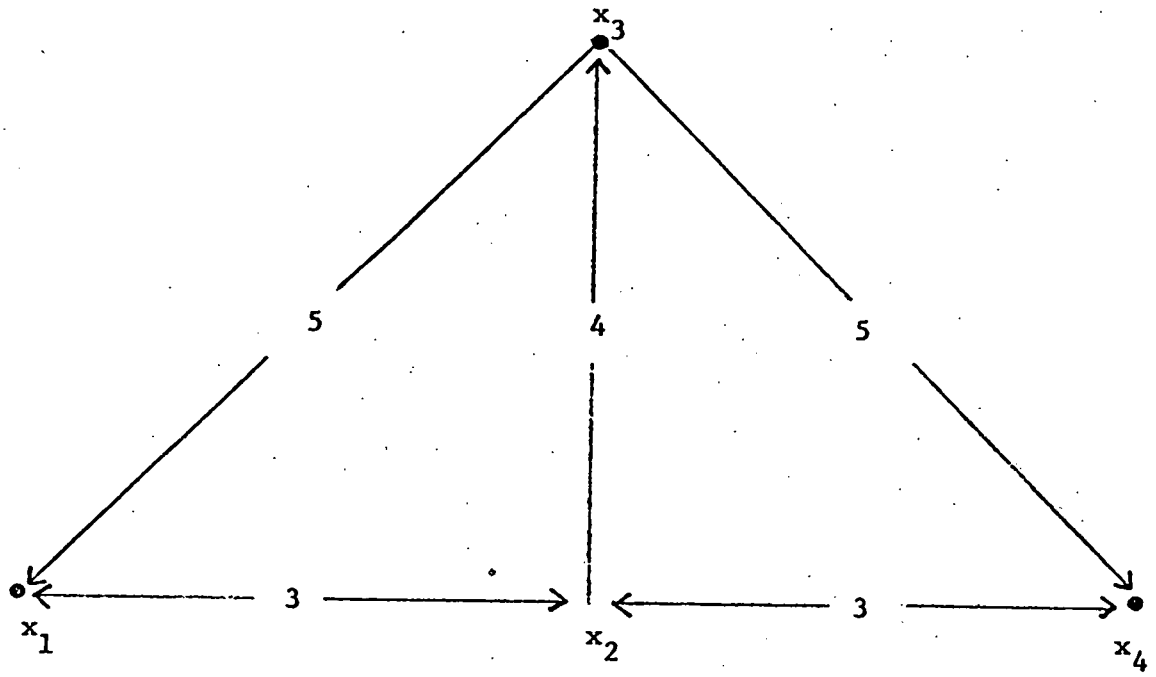

Figure 1.

Station No.	Elevation	Rainfall during five different time periods				
1	-1	5	7	4	4	10
2	0	4	10	4	4	12
3	0	6	9	4	3	13
4	+1	8	12	6	5	16

Table 1.

With the variogram (2), it may be shown that D^2_{2t} estimates $14c_1-4c_2$ for any time period t, whereas D^2_{3t} estimates $30c_1-12c_2$ for any t. Therefore, solving for the parameters, we get

$$\hat{c}_{1t} = (3D^2_{2t} - D^2_{3t})/12$$

$$\hat{c}_{2t} = (15D^2_{2t} - 7D^2_{3t})/24 .$$

Taking the median of each of these parameter estimates over time, using the data of Table 1, provides the final values

$$\hat{c}_1 = 0.67$$

$$\hat{c}_2 = 1.33$$

In this example the covariate, altitude, does not change with time. This is convenient but not essential; we might have used an air tempera- covariate and used the difference $\| a'_t - a''_t \|$ in the variogram model (2). Similarly we might have used a multivariate time-varying covariate. The fitted SV now allows us to calculate autocovariances of rainfall for any pair of points in the basin and thereby we may assess probabilistically the magnitudes of errors (mse) in various estimation problems using a spatial data network. The use of the SAF for this purpose is demon- strated in the next section.

Illustrations of Error Calculations

As before, $Z(x;t)$ is used to denote the precipitation variable at position x in the basin at time t. The data network consists of stations at positions $x_1, x_2, \ldots, x_n$ which record Z, possibly with error, at all times t. For purposes of estimation error calculations, Z is modeled probabilistically as a space-time random process with mean function $m(x;t)$. We consider examples of probabilistic error calcula- tions for mapping (i.e., spatial interpolation) problems based on sta- tion network data.

If we construct an estimate of rainfall for an unobserved position x_0 at time t, using a linear weighting of the network data at time t, then the mean-squared-error of the estimator is

$$E[Z(x_0;t) - \Sigma w_i\ Z(x_i;t)]^2 \equiv \text{mse}(\underset{\sim}{w}) ,$$

where $\underset{\sim}{w} = (w_1, w_2, \ldots, w_n)$ are the station weights. These weights should depend on the position of x_0 relative to the stations; for example, if x_0 is very close to x_i then we may expect w_i to be nearly 1 and all other weights to be nearly zero. In general, it is reasonable to assign weight zero to all stations except a handful in the vicinity of x_0, a procedure common to contouring programs.

A little manipulation shows that

$$\text{mse}(w) = \Sigma\Sigma w_i w_j [\gamma(x_i,x_0) + \gamma(x_j,x_0) - \gamma(x_i,x_j)] + B^2 , \qquad (4)$$

where

$$B = m(x_0;t) - \Sigma w_i\ m(x_i;t) .$$

For any arbitrary weighting scheme, the mse calculation depends on the mean function through its B term. However, when we postulate a simple enough mean function then we may choose weights $\underset{\sim}{w}$ so that B is guaranteed to vanish, and only the variogram γ is needed for the mse estimate. Such estimates are called "unbiased". For example, using the mean function (1) we have an unbiased estimate of $Z(x_0;t)$ at any time period t by choosing station weights to satisfy

$$\Sigma w_i = 1$$

as well as

$$\Sigma w_i\ v_j(x;t) = v_j(x_0;t) \qquad (5)$$

for each

$$j = 1, 2, \ldots, K .$$

It is easily conceivable that unbiased interpolation will have larger average estimation errors than biased interpolation, but having an mse depending only on the spatial variogram is a distinct advantage. It seems that the main disadvantage of unbiased methods is that the restriction (5) depends on the position of x_0. And one must not forget that the unbiasedness property is relative to the postulated form of the mean function. To illustrate using the station network of Figure 1 and mean function (3), if we wish to interpolate rainfall at position x_0 having altitude $a(x_0) = 0$, then the unbiased weighting requirement is $w_1 = w_4$ (as well as $\Sigma w_i = 1$). Now suppose the position of x_0 is such that it forms a square with the three stations x_1, x_3, x_4 of Figure 1. Among unbiased interpolations, there is one which minimizes the mse (4) with the variogram (2) estimated from the data of Table 1. The minimization may be accomplished by elementary methods giving so-called optimum weights $w_1 = w_4 = -0.4$, $w_2 = 2.0$, and $w_3 = -0.2$, with minimum $\text{rmse}(w_0) = 1.46$. For comparison, equal weighting of all four stations gives rmse = 1.89; putting all weight at the station x_2 gives rmse = 1.63.

Once we have some idea of the rainfall spatial variogram we may begin to answer simple design questions. Suppose a square-grid network is planned so that the maximum interpolation rmse does not exceed a specified σ_M. For interpolation at some x_0 we will use only the four station values forming the grid square in which x_0 is found and suppose that we make no use of covariate information. Then the maximum rmse for optimum unbiased interpolation occurs at the center of any grid square. If the variogram is given by $\gamma(x',x'') = c\|x'-x''\|$, then with grid spacing Δ the maximum rmse for interpolation is approximately $0.75\sqrt{\Delta c}$. Therefore, the required $\Delta = 1.33\ \sigma_M^2/c$, where the parameter c of the spatial variogram has presumably been estimated somehow. For the problem of estimating spatial averages the rmse calculations are more complicated and involve integration of the variogram.

The mapping of spatially varying qualitative variables, such as presence-absence patterns, presents somewhat different statistical problems. The network of n stations may be used to partition the basin into station-polygons $R_1, R_2, \ldots, R_n$ based on a nearest-point rule. The estimated map at time t may then be constructed by shading each of these station-polygons according to the station observation at time t. Such nearest-point rules produce qualitative maps with improbably jagged boundaries, but the jaggedness may be a virtue since it clearly reflects the density of the data network.

Some portions of the estimated map (for a fixed time t) will be incorrectly colored or labeled. The total area of that portion of the map which is incorrectly labeled may serve as an error criterion L_t. If we utilize a probability model as a generator of the true map pattern

then we may calculate expected values of this error criterion as a function of network design and of the probability model parameters. We have

$$E\{L_t\} = \sum_{i=1}^{n} \int_{x \in R_i} p_t(x,x_i)\, d\mu(x) \qquad (6)$$

where

$$p_t(x,x_i) = \text{Prob}\{Z(x;t) \neq Z(x_i;t)\}$$

To illustrate, suppose we have a square-grid data network and a probability model for which

$$p_t(x',x'') = b_t \| x'-x'' \| \qquad (7)$$

is valid for small values of the inter-point distance $\| x'-x'' \|$. The p function is used in the calculation of the error criterion only for inter-point distances less than 0.7Δ where Δ is the network grid spacing, so we do not need to specify the behavior of p at larger distances. By performing the integration in (6) we get

$$E\{L_t\} = 0.383(\Delta b_t)$$

assuming distances are scaled so that the total map area is 1. We see that mapping errors are proportional to the station grid spacing and hence inversely proportional to the square of the number of stations, within the context of the probability model (7).

We may use the station data to estimate the p function of (7) directly for the distance $\| x'-x'' \| = \Delta$, the grid spacing, by finding the proportion of adjacent station-pairs with different observed Z-values. For the estimated three-color map of Figure 2 we find $\hat{p}_t(x',x'') \doteq 0.13$ for $\| x'-x'' \| = \Delta$ and our estimate of the model parameter is $b_t \doteq 0.13/\Delta$. Therefore, the estimated error in the Figure 2 map is $E\{L_t\} = 0.383(.13) = 5\%$ of the total map area. This map was based on a network of 625 stations. If we would be content with

a 20% error on three-color maps of similar complexity, then we could do with a network of about 625/16 = 40 stations.

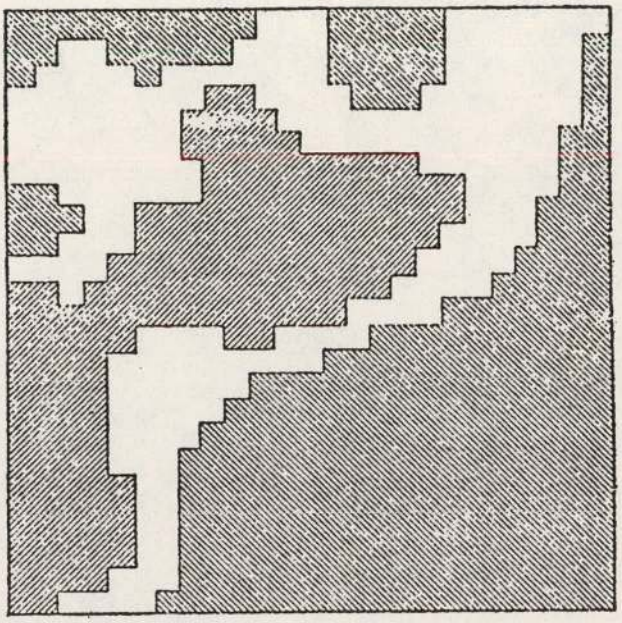

Figure 2

The crucial b_t parameter reflects the average patchiness of the underlying pattern at time t and it seems reasonable that more stations are needed for the accurate mapping of patchier phenomena. However, patchiness may vary from one time period to another so a given network will sometimes produce more accurate and less accurate maps. It is therefore reasonable that our error estimates should vary from one time period to another and that b_t should be estimated anew at each time period.

References

Delfiner, P., Linear estimation of nonstationary spatial phenomena, Advanced Geostatistics in the Mining Industry, 49-68. Reidel, 1976.

Matern, B., Spatial Variation, Almaenna Foerlaget, Stockholm, 1960.

Mathéron, G., The Theory of Regionalized Variables and its Applications, Centre de Morphologie Mathematique, ENSMP, Paris, 1971.

Switzer, P., Estimation of the accuracy of qualitative maps, Display and Analysis of Spatial Data, Wiley, 1975.