ESTIMATION OF CLIMATIC VARIABLES USING STANDARD MODELING TECHNIQUES AND RANDOM FIELD THEORY

B. Bajusz, J. Schumaker,
L. Pochop and R. Burman

1984
Conference Proceedings WWRC-84-26

In

Conference Volume

Third Conference on Mountain Meteorology
American Meteorological Society

B. Bajusz
J. Schumaker
L. Pochop
R. Burman

Department of Agricultural Engineering
University of Wyoming
Laramie, Wyoming

Conference Volume: Third Conference on
Mountain Meteorology, 16-19 Oct. 1984,
Portland, Ore. Published by the American
Meteorological Society, Boston, Mass. USA

5.10

ESTIMATION OF CLIMATIC VARIABLES USING STANDARD MODELING TECHNIQUES AND RANDOM FIELD THEORY

Barbara Bajusz, Research Associate
Joan Schumaker, Research Assistant
Larry Pochop, Professor
Robert Burman, Professor

Agricultural Engineering Department
University of Wyoming
Laramie, Wyoming 82071

1. INTRODUCTION

The objective of this study was to develop a method for estimating the mean monthly temperature for each year from 1948 to 1978, and for the months of April through October, at various ungaged sites in the Upper Green River Basin of Wyoming. The data available for this thirty-one year period were collected from twelve National Weather Service stations located within the Basin.

This paper describes a sequence of modeling techniques designed to maximize the information extracted from the data, and to reduce the estimation problem to a series of smaller, more tractable problems.

2. PROBLEM DISCUSSION

Meteorological conditions over a specific area at a given point in time can be thought of as a two-dimensional random variable with spatial autocorrelations (i.e. a random field). Special statistical estimation and modeling techniques, such as kriging, are generally required for handling data from random fields. Whereas the observations from a random field (RF) are correlated by definition, the majority of common statistical techniques require independence of the observations. The lack of independence has both advantages and disadvantages. Information can be extracted from the correlation structure in the data as well as from the data values themselves, but at the cost of increasing the complexity of the estimation problem and the necessary sampling rate. Therefore kriging, the most commong technique, is only feasible when the sampling is very dense over the area of interest, or some very restrictive assumptions about the RF can be made.

The data for this study were collected from a sparse gaging network. Therefore kriging is possible only if (1) the expected value and the variance of the RF are constant over the field of interest, and (2) the correlation between the values at any two points within the field depends only on the separation between those points. The first assumption states that all the variability in the data is due to spatially correlated, chance deviations from an overall mean. This is not a realistic assumption in topographically heterogeneous areas. Part of the variability in the data is due to non-random factors such as differences in elevation. Thus, these are at least two sources of information in the data which must be taken into account. The first is the relationship of meteorological variables to topographic characteristics; the second is the spatial correlation between values at neighboring locations.

In the estimation process proposed below, the RF is written as the sum of two independent components, one for each of the sources of variability. The problem of estimating the value of the RF at an ungaged sites then reduces to a series of two simpler estimation problems.

3. MODEL FORMULATION

Let $Z_{jt}(\underline{x})$ denote the monthly mean temperature in year t, month j, and location $\underline{x} = (x_1,x_2)$, where x_1 and x_2 are the coordinates of point $\underline{x}$. The RF is the sum of two components,

$$Z_{jt}(\underline{x}) = \mu_j(\underline{x}) + R_{jt}(\underline{x})$$

where $\mu_j(\underline{x})$ is the expected value of the RF acting in the jth month at location $\underline{x}$, and $R_{jt}(\underline{x})$ is the deviation from that value observed in year t. The term $\mu_j(\underline{x})$ is that part of $Z_{jt}(\underline{x})$ which remains constant from year to year; its value is largely determined by the topographic characteristics of the Basin. The term $R_{jt}(\underline{x}) = Z_{jt}(\underline{x}) - \mu_j(\underline{x})$ represents the year-to-year fluctuations of the value of the RF around its mean at location $\underline{x}$. This fluctuation is that part of $Z_{jt}(\underline{x})$ which is purely random (i.e. its value does not depend on $\underline{x}$ or the value of $\mu_j(\underline{x})$), and exhibits spatial autocorrelation.

Two simplifying assumptions were made in this formulation. First, the mean of the RF for each month was assumed to be stationary from year to year for the time period of interest. Second, independent RFs were assumed to be operating in each month.

4. ESTIMATION OF $\mu_j(\underline{x})$

The term $\mu_j(\underline{x})$ denotes the expected value of the RF at location $\underline{x}$ in month j, and is assumed to be constant from one year to the next. Its value is largely determined by nonrandom factors such as topographical features and recurring meteorologial patterns. A small spatial correlation probably exists among the values. This correlation structure was ignored since the information it might contribute in the estimation process is negligible in comparison to the difficulties its presence introduces.

Regression analysis, a standard statistical technique, was used to quantify the relationship between the data values and selected independent, or predictor, variables. When the correlation among the data values is in fact, significant, regression analysis will still produce unbiased estimates of $\mu_j(\underline{x})$, but the variance estimates will tend to underestimate the true variances.

The data used in the regression analyses are the average monthly mean temperatures for each location and month, and are denoted by

$$\bar{Z}_j(\underline{x}) = \sum_{t=1}^{T(\underline{x})} Z_{jt}(\underline{x})/T(\underline{x}),$$

where T(x) is the number of years of record for site $\underline{x}$. The factors selected for use as independent variables in the regression were those characteristics identified by the literature as having the potential to influence temperature, and which were easily obtainable for all sites in the Basin.

A stepwise regression was used to select from the independent variables those factors which best described the changes in $\bar{Z}_j(\underline{x})$ from one station to the next. Weighted least squares (WLS) regression was then performed to get *minimum variance* estimates of the regression coefficients.

The monthly R^2 values for the WLS regression indicated that 88 to 98 percent of the variability in the $\bar{Z}_j(\underline{x})$ values was accounted for by the regression function relating $\bar{Z}_j(\underline{x})$ to nonrandom factors. The remaining two to twelve percent of the variability is due to lack of fit of the regression model, plus a pure error (i.e. random) component which may exhibit spatial correlation.

An examination of the residual plots indicated that the model was underfit, and that at least second-order terms were needed for some of the predictor variables. An underfit regression model will not give unbiased estimates of $\mu_j(\underline{x})$. If the estimates of $\mu_j(\underline{x})$ are biased, the final estimates of $Z_{jt}(\underline{x})$ will also be biased. Additional terms were not, however, added to the model. With data from only twelve stations, the addition of higher-order terms could create a prediction bias, and thus reduce the models predictive capabilities at ungaged sites. In addition, the residual pattern may be due in part to spatial correlation among the error terms. The effect of lack of fit and pure error are not, however, separable without true replication. Thus their combined effects were implicitly modeled by spatially interpolating the regression residuals using the isohyetal method. The isohyetal method does not use the structural information in the data, and the properties of its estimates are unknown. This algorithm was used, however, because it does not require any underlying assumptions about the data.

The final estimate of $\mu_j(\underline{x})$ denoted by $\hat{\mu}_j(\underline{x})$, is then the value predicted by the regression function for location $\underline{x}$ plus the estimated regression residual as read from the isohyets at that location. This estimator returns the observed average, $\bar{Z}_j(\underline{x})$, at all gaged sites, and thus is unbiased at these locations. The properties of this estimator for ungaged sites are unknown.

Details of this estimation step are given by Schumaker et al. (1984).

5. ESTIMATION OF $R_{jt}(\underline{x})$

The term $R_{jt}(\underline{x}) = Z_{jt}(\underline{x}) - \mu_j(\underline{x})$ denotes the deviation in year t from the expected value of the RF at site $\underline{x}$ in month j. This term carries the random, spatially correlated component of $Z_{jt}(\underline{x})$. $R_{jt}(\underline{x})$ is thus a RF. Two simplifying assumptions about $R_{jt}(\underline{x})$ were made. As with $Z_{jt}(\underline{x})$, independent RFs were assumed to be operating in each month. Second, each of the RFs were assumed to be stationary both in their mean and in their covariance structure from year to year.

The expected value of $R_{jt}(\underline{x})$ is by definition equal to zero for all locations in the Basin. The variance of $R_{jt}(\underline{x})$ is equal to the variance of $Z_{jt}(\underline{x})$ for a given month and site, and therefore was estimated at each of the twelve weather stations with the sample variance of the observed $Z_{jt}(\underline{x})$ values. A homogeneity of variances test was conducted to test the hypothesis that the variance of $R_{jt}(\underline{x})$ is constant across the Basin in any given month. This hypothesis could not be rejected at the 0.10 significance level in any month. The RF for $R_{jt}(\underline{x})$ thus meets the first assumption necessary for kriging. The second assumption required for kriging, that of covariance stationarity, was examined through variogram construction and analysis. Since the true values of $\mu_j(\underline{x})$ are unknown even for gaged sites, the values of $R_{jt}(\underline{x})$ were estimated by replacing $\mu_j(\underline{x})$ with its best linear unbiased estimator, $\bar{Z}_j(\underline{x})$.

Directional variograms were constructed for each month and year of interest, and then pooled over years based on the assumed stationarity of the RF.

In theory, the variogram for a covariance-stationary RF should rise monotonically from zero, and plateau at the value of the variance of the RF. Let the separation between sites at which the plateau is reached be denoted by h*, where the distance between two locations $\underline{x}_a$ and $\underline{x}_b$ is defined by $h_{ab} = \|\underline{x}_a - \underline{x}_b\|^{1/2}$. The values of the RF at two locations are correlated if and only if their distance of separation is less than h*. The value h* is often referred to as the zone of influence. The information in the variogram required for kriging is the value of h*, the height of the plateau, and the shape of the curve as it rises from the origin.

Twenty-eight directional variograms were constructed, four for each of the seven months of interest. None of these variograms indicated a departure from the assumption of covariance stationarity. In most cases the zone of influence was less than the minimum separation between the weather stations. This means that neither the value of h*, nor the correlation structure within the zone of influence, can be estimated from this data. In this situation, the best linear unbiased estimate of the value of $R_{jt}(\underline{x})$ at an ungaged site is the sample mean of the observed values, denoted by $\bar{R}_{jt}$. The sample variance of $\bar{R}_{jt}$ gives an upper bound to the variance of the estimated value.

The most striking feature in all of the variograms is that the level of the plateau ranges from one-tenth to one-half of its theoretical value as estimated by the sample variance of the $Z_{jt}(\underline{x})$ values. This indicates that a major portion of the total variability in the $R_{jt}(\underline{x})$ value is unaccounted for by within-year variability. The implication is that the RF is not stationary from year to year as was originally assumed.

If the nonstationarity is only a shift in the location or mean of the RF between years, the above estimation procedures for $R_{jt}(\underline{x})$ are still valid. Kriging only requires that the mean be constant across the field of interest rather than equal to zero. The variogram construction as given above is also invariant to shifts in the mean across years.

If the correlation structure also changes from year to year, kriging may still be used, but the variograms cannot be pooled across years. Variogram construction and analysis must be performed for each year of interest using the $R_{jt}(\underline{x})$ values observed in that year.

The simplifying assumption of stationarity in the mean and covariance structure of the RFs should be relaxed only as far as needed. In this study there were insufficient stations reporting in any one year to produce reliable directional variograms on a yearly basis. Thus, it was not possible to determine whether the nonstationarity in the data extended to the covariance structure of the RF. The original stationarity assumptions were therefore replaced with the following: for each of the RFs it is assumed that the mean of the RF is itself a random variable having expectation equal to zero, and the covariance structure is stationary across years.

6. RESULTS AND DISCUSSION

The final estimator for the value of $Z_{jt}(\underline{x})$ at an ungaged site is

$$\hat{Z}_{jt}(\underline{x}) = \hat{\mu}_j(\underline{x}) + \hat{R}_{jt}(\underline{x}).$$

In the absence of any information for the month and year of interest, $\hat{R}_{jt}(\underline{x})$ is replaced with its expected value, and the estimate reduces to $\hat{\mu}_j(\underline{x})$, the estimate of the mean of the RF. The term $\hat{R}_{jt}(\underline{x})$ is used to update or correct $\hat{\mu}_j(\underline{x})$ when information for month j and year t are available.

The estimator $\hat{Z}_{jt}(\underline{x})$ is unbiased if $\hat{\mu}_j(\underline{x})$ is unbiased. In this study the data were too sparse to obtain an estimate for $\mu_j(\underline{x})$ which was known to be unbiased for all locations within the Basin.

When appropriate statistical tools are used at each stage of the modeling process, the variance of the estimator $\hat{Z}_{jt}(\underline{x})$ is just the sum of the variances of the component estimates.

It is possible to partition the total variability in the data in a manner analagous to analysis of variance, if complete historical records for a meteorological variable are available at several gaging stations over a fixed period of years. This information can be used to determine which component of the problem accounts for most of the variability in the data and thus should receive the greatest modeling effort.

It is important to have a continuous data record for each station over the entire time period when the year-to-year component of variability is significant. If the data for one or more stations cover only a portion of the time period, the effects of time and site differences will be confounded together in the $\bar{Z}_j(\underline{x})$ values. This will introduce a bias in $\hat{\mu}_j(\underline{x})$ which is a function of these values.

7. REFERENCES

Schumaker, J., B. Bajusz, L. Pochop, and J. Bornelli, 1984. Predicting temperature and precipitation at ungaged sites in the Upper Green River Basin. Proceedings of the Third Conferences on Mountain Meteorology, Portland, OR.