Bulletin of the International Statistical Institute, **51**, 1, 219-222.

SPATIAL–TEMPORAL MODELING OF METEOROLOGICAL FIELDS WITH APPLICATION TO CLIMATE CHANGE

Mark S. Handcock

Department of Statistics
The Pennsylvania State University
311 Joab L. Thomas Building
University Park, PA 16802-2111, USA
email: handcock@stat.psu.edu

1. INTRODUCTION

There is an extensive literature on statistical spatial–temporal modeling of meteorological and environmental fields. See, for example, Guttorp & Sampson (1994), Mardia & Goodall (1994), Brown, Le & Zidek (1994), Niu & Tiao (1995), Jones & Hulme (1996), Smith (1996) and Luo, Wahba, & Johnson, (1997). These papers focus on developing realistic models for possibly heterogeneous spatial–temporal dependence structures. In practice the characteristics of the underlying stochastic process are seldom completely known and are usually modeled from environmental data. The perspective taken here is that the model specification and statistical inference should not be regarded as two separate procedures when the objective is prediction. In particular it is essential that the choice of the class of models take into account the various sources of uncertainty that will effect the predictions based on that class. This will not only improve the quality of the predictions but also improve the quality of the assessment of the uncertainty of the prediction. The sources of uncertainty in prediction are sampling variability, model uncertainty and model misspecification. Sampling variability is the variation in the prediction given that the specified model is correct. The model uncertainty is the uncertainty about the correct model within the specified class of models given the available data. The model misspecification is the degree to which the specified class of models does not represent the spatial–temporal variation of the phenomena. The current available statistical approaches typically only take into account the first source when they assess the overall uncertainty. In general complicated (e.g., semi–parametric) model classes have smaller model misspecification while simple (e.g., small parametric) model classes have smaller model uncertainty.

Here we develop a spatial-temporal model for the spring temperature over a region in the northern United States covering eastern Montana through the Dakotas ($90^\circ - 107^\circ$ in longitude) and northern Nebraska up to the Canadian border ($41^\circ - 49^\circ$ in latitude). The empirical work of Lettenmaier, Wood and Wallis (1994) suggests that the temperatures over the spring period might exhibit a temporal pattern not found in the winter months. In addition the relatively stable and simple topography of the region help to ensure homogeneity and the minimization of localized effects.

A companion study reported in Handcock & Wallis (1994) considered the winter months and this region because GCM predictions of climatic change $(4^\circ F - 10^\circ F)$ induced by increased greenhouse gases are expected to be at maximum for high latitudes during the winter months (Mitchell (1989), IPCC (1990)). However there was no indication that the areal mean temperature for this time of the year in this region has changed over the last half century. There the posterior predictive distributions were used as a basis for calibrating temperature shifts by the historical record. In particular, the objective was to understand how soon gradual increases in temperature over this region would be discernible from the year-to-year variation. A similar analysis could be undertaken for the spring season considered here.

2. METHODOLOGY

Suppose $Z(x)$ is a real-valued stationary Gaussian random field on R with mean $E\{Z(x)\} = f'(x)\beta$, where $f(x) = \big\{f_1(x), \ldots, f_q(x)\big\}'$ is a known vector-valued function and β is a vector of unknown regression coefficients. Furthermore, the covariance function is represented by $\operatorname{cov}\big\{Z(x), Z(y)\big\} = \alpha K_\theta(x, y)$ for $x, y \in R$ where $\alpha > 0$ is a scale parameter, $\theta \in \Theta$ is a $p \times 1$ vector of structural parameters and Θ is an open set in $\mathbb{R}^p$. In the general case, we observe $\{Z(x_1), \ldots, Z(x_n)\}' = Z$ and will focus on the prediction of $Z(x_0)$. In our application $x_1, \ldots, x_n$ are the spatial locations of the stations in the network. We will focus on the prediction of $Z(x_0)$, where x_0 is a new location in the region of interest. The Kriging predictor is the best linear unbiased predictor of the form $\widehat{Z}_\theta(x_0) = \lambda'(\theta)Z$; that is, the unbiased linear combination of the observations that minimizes the variance of the prediction error. In the application $x = (u_1,\ u_2)$ and we can take $f_1(x) = u_1$ and $f_2(x) = u_2$, the latitude and longitude of locations within the geographic region, respectively. A third component of the mean will be added in Section 3. The covariance function represents the covariance between the temperature at the locations $x = (x_1,\ x_2)$ and $y = (y_1,\ y_2)$.

As β is a location parameter we expect that our prior opinions about β bear no relationship to those about α and *a priori* might expect α and β to be independent, leading to the use of the simple prior: $\mathrm{pr}(\alpha,\ \beta, \theta) \propto \mathrm{pr}(\theta)/\alpha$. The marginal posterior distribution of θ can be shown to be

$$\mathrm{pr}(\theta \mid Z) \propto \mathrm{pr}(\theta) \cdot |K_\theta|^{-1/2} |F' K_\theta^{-1} F|^{-1/2} \widehat{\alpha}(\theta)^{-(n-q)/2}$$

where $F = \{f_j(x_i)\}_{n \times q}, K_\theta = \{K_\theta(x_i, x_j)\}_{n \times n}$, and $\widehat{\alpha}(\theta)$ is the GLS of α given θ. The Bayesian predictive distribution for $Z(x_0)$ is

$$\mathrm{pr}(Z(x_0) \mid Z) \propto \int_\Theta \mathrm{pr}(Z(x_0) \mid \theta, Z) \cdot \mathrm{pr}(\theta \mid Z) d\theta$$

where the predictive distribution of $Z(x_0)$ conditional on θ and Z is a shifted t distribution on $n - q$ degrees of freedom with location $\widehat{Z_\theta}(x_0)$ and scale $\widehat{\alpha}(\theta) V_\theta$, the usual prediction error variance for $Z(x_0) - \widehat{Z_\theta}(x_0)$. This formulation does not restrict the covariance model to be isotropic or homogeneous, although the such a model is used in the application below. Depending on the influence of θ on the spread and location of $\mathrm{pr}(Z(x_0) \mid \theta, Z)$, the Bayesian predictive distribution might be wider or narrower than the predictive distribution derived by "plugging in" point estimates of α, β and $\theta.$. The location of the plug-in predictive distribution may also be quite different from the Bayesian predictive distribution. Typically the Bayesian predictive distribution will have no simple analytic form and must be determined numerically.

3. MODELING SPRING TEMPERATURE FIELDS

Initially we consider a spatial model appropriate for a meteorological field over a single time period. The field discussed here is the average spring temperature. The daily average temperature at a location is defined to be the mean of the daily maximum and the daily minimum at that location. The average spring temperature is defined to be the average daily average temperature over the months March, April and May.

The basis of the data is a network of 1219 stations (the HCN network) for the contiguous United States developed by the U. S. Carbon Dioxide Information Analysis Center "with the objective of compiling a data-set suitable for the detection of climatic change" (Karl, Williams, Quinlan & Boden (1990)). The actual data values are from Wallis, Lettenmaier and Wood (1991) which applied a enhanced method of adjusting for missing days (See Handcock & Wallis (1994)).

The components of the parametric mean function, $f_i(\cdot)$, should clearly include the latitude, longitude and elevation of each station. Other possibilities are polynomials in latitude, longitude, elevation, and the distance to the closest urban area or transformations of them. The ultimate choices for components for the mean function were latitude, longitude and elevation, as additional components did not have an appreciable effect on the likelihood ratios or on the likelihood function itself.

The parametric family of covariance functions used in this analysis is the Matérn class discussed in Matérn (1986) and Handcock & Stein (1993). In the form used here it is spatially isotropic and homogeneous: $K_\theta(x, y) \equiv K_\theta(|x - y|)$ is usually expressed as a function of a single scalar variable: $(x/\theta_1')^{\theta_2} \mathcal{K}_{\theta_2}(x/\theta_1')/(2^{\theta_2 - 1}\Gamma(\theta_2))$ where $\theta_1' = \theta_1/(2\sqrt{\theta_2})$ and $\theta_1 > 0$ is a scale parameter controlling the range of correlation. The smoothness of the field is controlled by $\theta_2 > 0$. $\mathcal{K}_{\theta_2}$ is the modified Bessel function of order θ_2 discussed in Abramowitz and Stegun (1964), §9. Based on numerical and graphical diagnostics this is a plausible specification for the spatial variation in this region. The use of a more complicated model would result is a bias–variance tradeoff with the greater model uncertainty. The class is motivated by the smooth nature of the spectral density, the wide range of behaviors covered and the interpretability of the parameters. The Exponential class corresponds to the sub-class with smoothness parameter $\theta_2 = 1/2$, that is $K_E(x) = \theta_1 \exp(-x/\theta_1)$. As $\theta_2 \to \infty$, $K_\theta(x) \to \exp(-x^2/\theta_1^2)$, often called the "Gaussian" covariance function. We shall refer to it as the Squared Exponential model. This model forms the upper limit of smoothness in the class. A draw back of this class is that it does not allow negative correlations. A related class of models that do allow negative correlations have been developed by Vecchia (1985), although the flexibility in the smoothness is restricted.

Based on meteorological arguments, we believe that the underlying meteorological field is continuous and may be differentiable many times . The magnitude of both random and systematic measurement error ("nugget effect") varies from year-to-year. It can be incorporated by adding a single additional parameter (θ_3) to the covariance function: $\mathrm{cov}\{Z(x), Z(y)\} = \alpha(\theta_3 \mathcal{I}(x = y) + K_\theta(x, y))$ where $\mathcal{I}(\cdot)$ is the indicator function. For example, for 1984 the maximum likelihood estimate had no nugget effect $(\widehat{\theta_3} = 0)$. Typically the estimated nugget effect was 25% - 50% of the point variance. In contrast, the mean winter temperatures reported in Handcock & Wallis (1994) the nugget effect appeared to be small relative to the year to year variation. As expected the estimate of the smoothness of the field increases when a nugget effect is included. For example, for 1988 the maximum likelihood estimate for the covariance structure based on the Matérn class is $(\widehat{\alpha}, \widehat{\theta}) = (1.33^\circ F^2,\ 1.02^\circ,\ 10.3,\ 77\%)$. The range of dependence (1.02°) spans approximately a ninth of the region under study. The point standard deviation of the mean spring temperature is $\sqrt{\widehat{\alpha} + \widehat{\alpha}\widehat{\theta}_3} = 1.5^\circ F$. The smoothness of the field is estimated to be ten mean-square derivatives. Under the Gaussian assumption this implies that the observed field has ten derivatives. The likelihood is very flat for $\theta_2 > 1$. The maximum likelihood Squared Exponential model ($\theta_2 = \infty$) with a nugget effect is $(\widehat{\alpha}, \widehat{\theta}) = (1.33^\circ F^2,\ 0.97^\circ,\ \infty,\ 76\%)$, with only a 0.01 decrease in log-likelihood from the previous model. This feature was typical of most years: the log-likelihood is very flat for large smoothness values indicating that there is little information in the spatial observations to discriminate between these smoothnesses *in the presence of a substantial micro-scale variation.*

We can add the temporal component of the model, generalizing the random field to $Z_t(x)$ where $t = 1948, \ldots$ represents the spring of observation. We consider the time-series of data from each station, independent of the spatial information. The spatial analysis was repeated independently for each year of data from 1948-88.

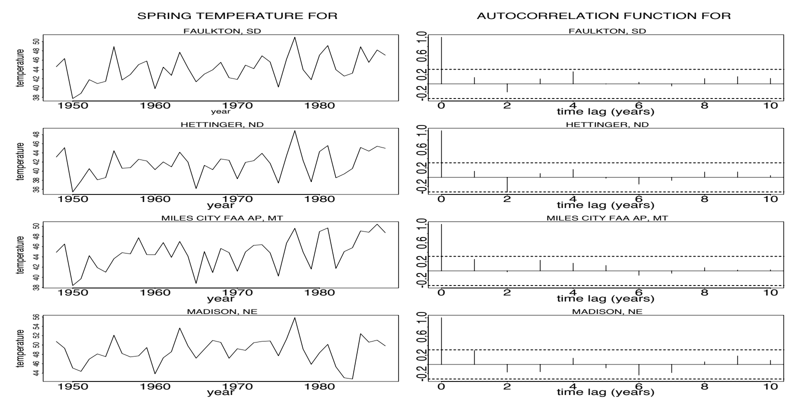

Figure 1. Time-series and empirical autocorrelation functions for four typical sites.

Figure 1 presents the time-series and empirical autocorrelation functions for four spatially separate stations. Individually the time-series are quite variable over time. Inspection of the series as a whole suggests a mild upward trend over the last half century time. The right hand side figures are the sample autocorrelation functions corresponding to the time-series. The dashed boundaries represent approximate 95% confidence limits. Note the similar patterns in the series over time and the lack of first lag autocorrelation. We investigated these fields for short-memory temporal structure. We found little evidence for AR(1) structure and indication that the series are close to uncorrelated over time. We also considered the presence of significant dependence between observations a long time span apart, postulating an autoregressive integrated moving average processes with non-integral degrees of differencing, d. (ARIMA (p, d, q)). There was little (likelihood) evidence for such structure.

4. MEASURING AREAL MEAN TEMPERATURE

Using the method described in Handcock & Wallis (1994) we can construct posterior distributions for the time–series of areal mean temperature over the region of interest. These are defined by: $\bar{Z}_t = \frac{1}{|R|}\int_R Z_t(x)dx,\ t = 1948, \ldots, 1988, \ldots$ where $|R|$ is the area of the region R. Thus at each point in time, $\bar{Z}_t$ represents the average temperature over the region and is a function of the field $Z(x)$. $\bar{Z}_t$ provides a natural measure for the detection of changing climatic patterns over the region. As the region is devoid of gross topographic features, it provides a convenient measure of overall temperature during the spring. It is important to note that $\bar{Z}_t$ is a characteristic of the temperature field itself, and not a characteristic of the stations in the network. The behavior of the areal mean temperature will provide an indication of the overall changes in climate over the region independent of the individual stations.

To further explore the temporal changes in $\bar{Z}_t$ we will consider the time-series of maximum *a posteriori* (MAP) values. The distributions have similar shape and the ratio of largest to smallest variance is 2.2. While this clearly represents a reduction in information relative to the full distribution, it facilitates examination. Figure 2 represents the MAP values for the last half century. Note the suggestion of a trend over time. Some interesting years have been indicated. There was little evidence, over this period, of short or long term dependence in the MAP values. However long term dependence in climatological series can occur over time scales of a centuries or more and such dependence would not be apparent from our half century of observation. The trend apparent in Figure 2 could simply be an artifact of such long term dependence.

As in Handcock & Wallis (1994) we can construct a model for the mean areal temperature over the last half century. If we compare the static areal mean temperature for 1948-67 to 1968-88 we can confirm that the posterior distributions have little overlap, with the information from the latter years indicating warmer temperatures. The uncertainty in the mean areal temperature for the latter period is larger, reflecting the increased variation in the annual values. The natural alternative model postulates a linear increase over time. However this trend is confounded with any long term dependence over a time scale much longer than the forty year observation period.

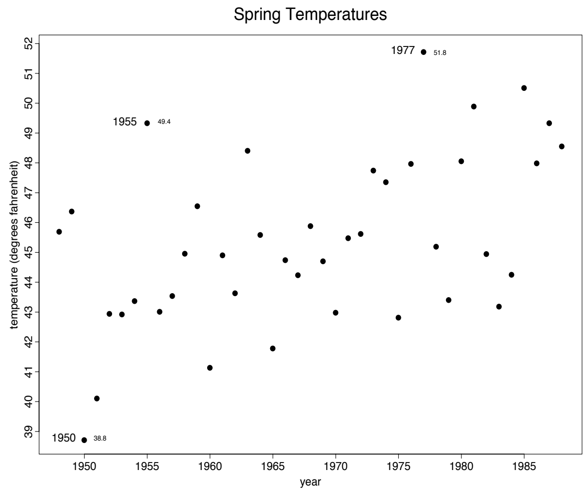

Figure 2. This is the time-series of (the MAP estimates for the) areal mean temperature. There is little short or long term dependence.

BIBLIOGRAPHY

Abramowitz, M. & Stegun, I.A. (1965). *Handbook of Mathematical Functions.* New York: Dover.

Brown, P. J., Le, N. D. & Zidek, J. V. (1994), *Multivariate Spatial Interpolation and Exposure to Air Pollutants Canadian J. of Statistics, 22,* 489-516.

Guttorp, P. & Sampson, P. D. (1994). *Methods for Estimating Heterogeneous Spatial Covariance Functions with Environmental Applications* in *Handbook of Statistics, Vol. 12,* Eds. G. P. Patil and C. R. Rao, pp. 661-689. Elsevier.

Handcock, M.S. and Stein, M.L. (1993). *A Bayesian Analysis of Kriging.* Technometrics **35**, 403-410.

Handcock, M.S. and Wallis, J.R. (1994). *An Approach to Statistical Spatial–Temporal Modeling of Meteorological Fields* (with discussion) Journal of the American Statistical Association, 368-390.

Intergovernmental Panel on Climate Change, (1990) *Scientific Assessment of Climate Change,* World Meteorological Organization, United Nations Environment Programme, Geneva, Switzerland.

Jones, P.D., & Hulme, M. (1996). *Calculating Regional Climate Time Series for Temperature and Precipitation: Methods and Illustrations* International Journal of Climatology **16**, 361-377.

Karl, T.R. & Williams, C.N., Jr., Quinlan, F.T. & Boden, T.A. (1990). *United States Historical Climatology Network (HCN) serial temperature and precipitation data.* NDP-019/R1, p. 83 plus appendices. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, Oak Ridge, Tennessee.

Lettenmaier, D.P., Wood, E.F. & Wallis, J.R. (1994). *Hydro–Climatological Trends in the Continental United States, 1948-88* Journal of Climate **7**, 586-.

Luo, Z., Wahba, G. & Johnson, D. (1997). *Spatial–Temporal Analysis of Temperature Using Smoothing Spline ANOVA.* Technical Report, Department of Statistics, Penn State University.

Mardia, K. V. & Goodall, C. R. (1994) "Spatial–temporal Analysis of Multivariate Environmental Monitoring Data Models" in *Multivariate Environmental Statistics* Eds. G. P. Patil and C. R. Rao, pp. 347-386. Elsevier: New York.

Matérn, B. (1986). *Spatial Variation* Second Ed, Lecture Notes in Statistics, **36.** Berlin: Springer-Verlag.

Mitchell, J. F.B. (1989). *The "Greenhouse Effect" and Climate Change.* Reviews of Geophysics **27**, 115-139.

Niu, X. and Tiao, G. C. (1995). *Modeling Satellite Ozone Data,* Journal of the American Statistical Association, 969-983.

Ripley, B.D. (1981). *Spatial Statistics.* New York: Wiley.

Smith, R. L. (1996). *Estimating Nonstationary Spatial Correlations,* Technical Report, Department of Statistics, University of North Carolina.

Vecchia, A. V. (1985), *A General Class of Models for Stationary Two-Dimensional Random Processes,* Biometrika **72,** 281-91.

Wallis, J.R., Lettenmaier, D.P. & Wood, E.F. (1991). *A Daily Hydroclimatological Data Set for the Continental United States.* Water Resources Research **27**, 7, 1657-1663.

Summary

The traditional best linear unbiased prediction procedure ("Kriging") is used in this paper for inference, but within a Bayesian framework. See Brown, Le & Zidek (1994) for an alternative Bayesian formulation. Our approach is to exam how posterior predictive distributions of areal quantities change over time. The objective is to see if there have been changes in areal temperature that are discernible from the year-to-year variation. The approach takes into account the uncertainty about the covariance function expressed in the likelihood surface and ignored by point estimates of the covariance function.

These ideas are implemented for the spring temperature over the region in the northern United States based on the stations in the United States historical climatological network reported in Karl, Williams, Quinlan & Boden (1990). The results indicates that there is significant micro-scale variation over a spatially smooth field. There is substantial variation in the spring temperature from year-to-year that is spatially correlated (Figure 1). The areal mean temperature is correspondingly variable and appears to be increasing over the period considered (Figure 2). However, this finding has been confirmed using different statistical approaches (Lettenmaier, Wood and Wallis (1994)). It is also evident from this later study that, had other regions or periods been picked, the results would have been quite different.