U.S. DEPARTMENT OF COMMERCE
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
NATIONAL WEATHER SERVICE
NATIONAL METEOROLOGICAL CENTER

OFFICE NOTE 352

SPATIAL OBJECTIVE ANALYSIS

H. JEAN THIEBAUX

FEBRUARY 1989

SPATIAL OBJECTIVE ANALYSIS

H. J. Thiébaux National Meteorological Center

GLOSSARY

autoregressive: dependent on its own history, or determined by it up to random innovations.

diagnostics: measures whose values describe the state, energy, or distribution of a property of an observed system.

earth oriented system: one whose natural coordinates are designated with respect to Earth, in latitude, longitude, and depth or pressure.

global description: designation of the state of a system or the value of one of its variables at each point of its domain.

minimum variance estimation: the determination of a function or parameter values for a designated function of stochastic variables, which guarentee its having least statistical variance among all functions or sets of parameters.

spatial coherence: the continuity of values of variables such as temperature, salinity, or wind velocity, in a spatially continuous environment.

unbiased: the property of an estimate for a variable or parameter, of having the same statistical mean value.

SPATIAL OBJECTIVE ANALYSIS is a body of techniques of applied mathematics which have been developed to provide global descriptions of spatially coherent variables, with data from observing arrays which are sparce and generally inhomogeneous. Thus, in the specific, a spatial objective analysis, is a tensor estimate of the state of a system at each point of the domain of interest, which has been constructed from a set of irregularly spaced observations. The word "objective" identifies the construction as the outcome of applying a mathematical algorithm, in distinction to a "subjective" construction which admits the influence of individual scientific intuition.

I. Introduction

Historically, the techniques of spatial objective analysis have been developed as descriptive and diagnostic techniques of science, beginning with K. F. Gauss' pioneering work in Astronomy and extended by people who are now distinguished as founders of the discipline of mathematical statistics. Areas of scientific endeavor in which spatial objective analysis is currently used include geology, atmospheric and ocean sciences, and environmental studies. The relative isolation of research in these different fields has led to the assignment of credits to a spectrum of scientists, each of whom had the acumen to appreciate the importance of these techniques and applied them within their own fields with outstanding results. Prominent on this roster are G. Matheron, L. S. Gandin, and R. E. Kalman.

Scientific study of large scale, earth-oriented systems requires capabliities for constructing global descriptions, with algorithms built from assumptions about spatial inter-relationships of their component variables. These requirements must confront the de facto configuration of our habitat, which precludes spatially uniform data coverage. In fact, actual observations of global scale phenomena, including those of remote sensing technology, may provide information which is highly non-uniform in coverage. Nonetheless global descriptions of spatially dimensioned states of the system, which accurately reflect its properties and interrelationships of its component parts, must be constructed.

The techniques developed to meet these requirements may be classified by the characteristics of the information available to describe the system under study. There are two clear divisions: One presumes no history of information, of which an example would be an isolated, geological field study producing a single "snap shot" of observations at selected locations, for a system which is presumed relatively fixed in time. The other assumes an accumulated record of observations, on a time evolutionnary system for which the statistical relationships among states of the system at separate locations remain the same throughout the record.

II. The Fundamental Problem

A spatial objective analysis algorithm is a formula for providing estimates of the values of a spatially coherent variable, where each estimate is constructed from a set of proximal observations. If observations recorded true values perfectly then the objective might be to fit them perfectly, with a surface whose intermediate values would provide estimates of the variable everywhere else. We know the premise is incorrect. Also, we generally assume that the signal component of the field we seek to estimate differs from the true value at a given instant and point location.

Since the distinction between the signal component of the variable at a specific location and instant of time, and the value of an observation at that same location and instant, is basic to the theory from which all spatial objective analysis algorithms are derived, we consider this first. The signal component is the value one would like to know and which the estimate is constructed to approximate. It is not the same as the value of an observation, were it possible to obtain one; because an observation is a composite of the true value, errors due to the imprecision of sensors and recorders, and mistakes made in the transmission of information. Furthermore, the true value at a specific location and instant of time generally reflects microscale variability whose influence is regarded as obscuring the signal. Accordingly, the signal component has a theoretical definition as the average of the true values over a localized region and time interval. The construction of an

algorithm for estimating the signal takes into account these unwanted noisy constituents of the observations; so that it will act as a filter to separate signal from noise, as well as an interpolater.

The coherency in space and in time which is assumed by the theory of objective analysis, is equivalent to the mathematical concept of continuity of functions. Thus, the goal of objective analysis is the creation of a spatially continuous estimator of a field variable, which may be evaluated at any location -- effectively interpolating only the signal components of the observations going into it. If it is reasonable to assume that noise is independent of signal in the observations, in the sense that the magnitude or strength of the signal does not affect the expected level of noise, this may be represented as estimating the signal component $\mu(\lambda,\phi,p,t)$ from a set of observed values which are "signal plus noise":

$$Z(\lambda,\phi,p,t) = \mu(\lambda,\phi,p,t) + \epsilon(\lambda,\phi,p,t) .$$

The indexing parameters denote earth-oriented coordinates: respectively, longitude, latitude, pressure or depth, and time. $\epsilon(\lambda,\phi,p,t)$ represents the total influence of small scale variability which is of no interest, together with instrument and transmission system errors, collectively referred to as "noise". Because of their tractability, analysis algorithms are generally based on the assumption that signal and noise are independent. Accordingly most of this article will deal with this formulation.

A more general expression for the value of an observation of a field variable is a Kalman Filter representation. To carry all indices in a brief notation, we collectively denote distinct arrays of space/time points by $\underset{\sim}{s}$ and $\underset{\sim}{s}+\underset{\sim}{\Delta}$. The formulation of the Kalman Filter assumes that the true value of the field variable at any location/time may be written as a composite of its deterministic evolution and a stochastic innovation:

$$X_{\underset{\sim}{s}+\underset{\sim}{\Delta}} = \Phi^{*}_{\underset{\sim}{s}} \circ X_{\underset{\sim}{s}} + W_{\underset{\sim}{s}} \quad ;$$

and that a recorded observation is a linear combination of the true value and noise from sensing, transmitting and recording devices:

$$Z_{\underset{\sim}{s}+\underset{\sim}{\Delta}} = X_{\underset{\sim}{s}+\underset{\sim}{\Delta}} + V_{\underset{\sim}{s}+\underset{\sim}{\Delta}} \, .$$

Here Φ^{*}_{o} denotes the algorithm for the deterministic component of the change in X from one location/time to another, W denotes the stochastic component in the evolution of the true field, and V denotes the collective impact of the noise elements. The Kalman Filter is composed of a prediction from the proximal/previous estimate, modified by the weighted discrepancy between what is observed and predicted:

$$\hat{X}_{\underset{\sim}{s}+\underset{\sim}{\Delta}} = \Phi_{\underset{\sim}{s}} \circ \hat{X}_{\underset{\sim}{s}} + K_{\underset{\sim}{s}+\underset{\sim}{\Delta}} \left(Z_{\underset{\sim}{s}+\underset{\sim}{\Delta}} - \Phi_{\underset{\sim}{s}} \circ \hat{X}_{\underset{\sim}{s}} \right).$$

A distinction is made here between the true "system operator" of the subject variable, Φ^{*}_{o}, and the "model system operator" used to generate predictions, Φ_{o}. Generally the true operator will not be known precisely; and assuming that they <u>are</u> the same will lead to erroneous deductions from a formulation that equates

them. $\mathbf{K}$ is customarily called the "gain matrix" of the filter; and we note that the indices used in the formulation indicate its time- and space-dependence, in addition to those of the system operators. The generality of this representation for a spatial objective analysis is mathematically impressive; however it is impractical to apply to the analysis of most physical systems, because the components are too imperfectly known.

Within the present framework, an optimal estimator is one for which the average squared departure of values of the estimates it produces, from the corresponding signal values, is smallest among all possible algorithms. Since the signal is a conceptual quantity which cannot be observed, the "average" is the statistical mean squared departure; and optimal objective analysis is minimum variance estimation in the classic statistical sense. Whether or not the minimum can be guarenteed depends on whether or not the multivariate statistical distribution of the field variable or of its increment from some known background field, can be specified. When the multivariate statistical distribution can be specified, the optimal estimator for any specified location is unique; and it is the conditional mean value of the variable at that location, with respect to its joint distribution with the values at all the locations at which it is observed. We refer to this as the ultimate objective analysis and reserve the definition optimal objective analysis for what is practical in terms of our knowledge of statistical properties of the field variable.

The most general way of representing the ultimate objective analysis value at a point indexed by 0, which is constructed with observations from m locations indexed 1,...,m, is in terms of the cumulative distribution function for z_0, conditioned on the observed values of its m covariates: $z_1, \ldots, z_m$, namely

$$F_0(z_0 \mid z_1, \ldots, z_m) = \text{Prob}(z_0 \leq z_0 \text{ given the values of } z_1, \ldots, z_m).$$

With this, the conditional mean is

$$E(z_0 \mid z_1, \ldots, z_m) = \int_{-\infty}^{+\infty} z_0 \, dF_0(z_0 \mid z_1, \ldots, z_m)$$

and this can be shown to provide the unique minimum of

$$E\{[z_0 - g(z_1, \ldots, z_m)]^2\}$$

for all possible functions g of the set of observations. However it clearly depends on knowing the family of multivariate distribution functions:

$$F(z_0, z_1, \ldots, z_m) = \text{Prob}(z_0 \leq z_0, z_1 \leq z_1, \ldots, z_m \leq z_m)$$

from which all conditional distribution functions and their conditional mean values can be derived. Many practical situations warrant the assumption that these distributions are reasonably approximated by multivariate normal distributions; and, in these cases, the objective analysis formulation assumes a particularly simple form.

In those circumstances in which the multivariate statistical distributions for the field variable cannot be specified, alternative approaches to objective analysis are employed.

Again, the choice of a technique will depend on what assumptions can reasonably be made about boundary conditions on the variable, for the region within which it is considered, and characteristics of its spatial variability. The non-statistical alternatives which will be cited in this article are referred to collectively as empirical interpolation techniques. There are several, each with advantages tuned to specific observation characteristics and requirements for the estimated field. All share the disadvantage that no analysis algorithm can out perform the statistical objective analysis, in circumstances in which the latter is a practical alternative.

III. Optimal Statistical Objective Analysis

Here the underlying assumption is that the family of multivariate statistical distributions for the field variable are known and/or that the situation warrants the assumption that they are reasonably approximated by multivariate normal distributions for which we have unbiased models for the mean vectors. In this case the optimal estimator for the "analysis increment", namely for the difference between the location specific value and the modelled mean value, is a weighted linear combination of observation increments. The latter are derived from the observations by differencing them with corresponding modelled means. Thus, denoting observed values by $z_1^o, \ldots, z_m^o$, and

the modelled mean values by $z_0^f, z_1^f, \ldots z_m^f$, for each point-of-estimate o and the m observing locations, respectively, the objective analysis algorithm generates point estimates

$$\hat{z}_o = z_o^f + \sum_{i=1}^{m} w_i (z_i^o - z_i^f) .$$

The coefficients which provide the weights bear a direct and simple relationship to the covariance parameters of the normal distribution function. With double subscripts identifying the locations of the variable to which each scalar covariance σ_{ij} pertains, the between-observing-location covariance array may be written as

$$\Sigma = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1m} \\ \vdots & & \vdots \\ \sigma_{m1} & \cdots & \sigma_{mm} \end{pmatrix}$$

and the array whose entries are the covariances between the values of the variable at the point-of-estimate and at the locations of observations as

$$\Sigma_0 = (\sigma_{01} \cdots \sigma_{0m}) .$$

With this notation the array of weights for the analysis increments is given by

$$W = \Sigma_0 \Sigma^{-1} .$$

As a theoretical formulation, (1) and (2) describe the "ultimate objective analysis".

In practice, the formulation requires specification of the covariance structure, which must itself be estimated, to create a

working, optimal statistical objective analysis (OSOA) algorithm. It is at this point that "optimal" takes on its practical connotation; and the dependence of the accuracy of the analysis on the representation of the covariance structure assumes central importance. As the word suggests covariance is a measure of the strength of the relationship between the variabilities of increments at different locations; and its tensor representation determines the weights assigned to observed increments in OSOA. Thus, if the assigned covariance values are not representative of the true statistical relationships of the increment field, the weights and therefore the analysis will be less than optimal.

Considerable research has been devoted to derivation and study of the properties of covariance functions for the representation of spatial covariance structure. Some of this work has started from first principles, in the following sense. The discrepancy between an observed field and an unbiased modelled mean field is described as a spatially coherent stochastic process, for which the covariance function may be derived analytically from the spatial analogue of an autoregressive representation for a time series. Selection of the order of the autoregressive representation and assignment of values to the parameters of the covariance function, requires comparative fitting of candidate functions to an array of covariances computed from data, together with study of implicit properties of these functions versus known or theoretical properties of the increment field. These derivations assume stationarity of the statistical properties of

the increment field and an extensive archive of observations from a fixed array.

IV. Kriging

In circumstances in which the increment field is known to be statistically non-stationary or for which an archive of observations is not available to provide guidance for selecting a representation for the between-location covariances of analysis increments, OSOA cannot be used; and an alternative technique must be selected.

One practical alternative which has been developed and used extensively in a number of natural and social science settings, is Kriging. The term is derived from the name Krige, the man to whom G. Matheron ascribes its inspirational use in the field of geology. This technique was developed to produce estimates of values of spatial processes over a geographic region with a single set of data available from a sampling array. It is a multi-stage procedure which first fits a trend surface to the data, then estimates the covariance structure of the residual field from the differences between the observations and the fitted surface, and finally corrects the point predictions of the trend surface by a constrained multi-variate interpolation of the observed residuals.

The trend surface is constructed by fitting a linear combination of basis functions, $h_\ell(\underset{\sim}{s})$, to the sample data, in the manner of a regression analysis. Thus the coefficients, α_ℓ , are determined by minimizing the sum of squared differences between the surface and the observed values:

$$\sum_{i=1}^{m} \left[z_i^{a} - \sum_{\ell} \alpha_\ell h_\ell(\underset{\sim}{s}_i) \right]^2$$

where $\underset{\sim}{s}_1, \ldots, \underset{\sim}{s}_m$, are the locations of the observations. This provides a trend surface estimate,

$$m(\underset{\sim}{s}) = \sum_{\ell} \hat{\alpha}_\ell h_\ell(\underset{\sim}{s})$$

which may be evaluated at any point of the region. In Kriging this is frequently referred to as the "drift" and is somewhat analogous to the role of the modelled mean field of OSOA. An important departure from OSOA is the accomodation made for the bias present in $m(\underset{\sim}{s})$. Since the minimization which parameterizes the fitted surface does not guarentee its unbiasedness, Kriging constrains the weights of the linear interpolation of residuals to sum to unity.

The residuals, namely, the discrepancies between the trend surface and the data to which it has been fitted, are regarded as multivariate normal variables with a stationary distribution. Thus the mean, which is the bias b in the fitted surface, and the variance σ^2 are assumed to be constant over the region. With the further assumption that the correlation between variables at separate locations is a function of scalar

displacement $\rho(\|\underset{\sim}{\Delta}\|)$ which approaches zero with increasing separation, covariances may be written as:

$$\sigma_{ij} = \sigma^2 \rho(\|\underset{\sim}{\Delta}\|), \qquad \|\underset{\sim}{\Delta}\| = \|\underset{\sim}{s}_i - \underset{\sim}{s}_j\|.$$

With the constraint of having only one observation at each location, the requirement for a representation of the covariances for interpolation of residuals to correct point predictions of the surface is satisfied by constructing a "variogram". For illustration we let $Y(\underset{\sim}{s})$ denote the residuals, so that

$$Y(\underset{\sim}{s}_i) = m(\underset{\sim}{s}_i) - z_i^0, \qquad i = 1, \ldots, m.$$

With this notation and the foregoing assumptions about the distribution of the residuals, it follows that the expected value of each squared difference of residuals can be written in terms of the variance and the correlation function as:

$$E\{[Y(\underset{\sim}{s}_i) - Y(\underset{\sim}{s}_j)]^2\} = E[\{[Y(\underset{\sim}{s}_i) - b] - [Y(\underset{\sim}{s}_j) - b]\}^2] = 2\sigma^2[1 - \rho(\|\underset{\sim}{\Delta}\|)].$$

Thus the expected value of $[Y(\underset{\sim}{s}_i) - Y(\underset{\sim}{s}_j)]^2/2$ is $\sigma^2[1 - \rho(\|\underset{\sim}{\Delta}\|)]$. Since $\rho(\|\underset{\sim}{\Delta}\|)$ is 1 at zero separation and approaches 0 at large separations, a plot of $[Y(\underset{\sim}{s}_i) - Y(\underset{\sim}{s}_j)]^2/2$ against the distances between observing locations $\|\underset{\sim}{\Delta}\| = \|\underset{\sim}{s}_i - \underset{\sim}{s}_j\|$ may be fitted with a function $\gamma(\|\underset{\sim}{\Delta}\|)$ whose asymptote provides an estimate of the variance $\hat{\sigma}^2$ and whose normalized values provide a representation for the correlation function:

$$\hat{\rho}(\|\underset{\sim}{\Delta}\|) = 1 - \gamma(\|\underset{\sim}{\Delta}\|) / \hat{\sigma}^2.$$

V. Empirical Interpolation Techniques

Spatial continuity of the variables for which analysed fields are constructed is basic to spatial objective analysis. However, having sufficient knowledge of the properties of the stochastic elements of the observed field for a statistical objective analysis, is not a requirement for the construction of a spatially continuous estimated field for the region from which observations are obtained. Under circumstances in which it is unreasonable to make explicit assumptions concerning either the definition of a mean field or the form of the statistical distribution of observed residuals from a defined mean field, other options are available.

Alternatives to statistical objective analysis are referred to here as empirical interpolation techniques. These include a) distance-weighted interpolation, b) spline fitting, and c) function surface fitting. The formal appearance of the first on this list is like that of a statistical objective analysis:

$$(\hat{z}_0 - z_0^f) = \sum_{i=1}^{m} a_i \left(z_i^o - z_i^f \right) .$$

However the coefficients which assign weights to observed values of the analysis increments are functions only of the distances between the point-of-estimate and the locations of the observations. They do not take into account the relative locations of the observations; and consequently are not able to adjust the analysis for the communality of information brought by observations from proximal locations. Nonetheless,

in circumstances in which observations are fairly uniformly spaced in the region of the analysis and when the covariance structure of the analysis increments is not known, this technique may provide reasonable estimates.

Spline fitting has become a scientific art, in recent years. It is a special case of "function surface fitting" in which several-dimensional polynomials are fitted to data within subregions. The polynomials have fewer parameters than the number of observations, so that they perform a smoothing role as well as data fitting; and they are matched at the boundaries of the subregions for continuity. Control for "excursions" of the piece-wise continuous surface beyond reasonable limits on the variability of the analyzed field is achieved by putting the spline "under tension". Spline surface fitting has considerable flexibility for tuning the surface to match theoretical properties or response to orography. However, it requires considerable insight and attention to mathematical and physical detail; and, like all constructions made by fitting functions to data, the resulting analysis algorithm cannot be used for estimation beyond the region of the observations with confidence in the result.

Function surface fitting has been considered, in outline, as the first step of Kriging in which a "trend surface" is obtained by choosing the parameters for a linear combination of basis functions, to be those which minimize the sum of its squared departures from the observations. Basis functions may be

selected to satisfy implicit spectral properties from physical theory or an analogue field; or they may be sets of orthogonal functions, selected for their suitability as statistical diagnostics of the analyzed field. Recent literature on this subject is extensive in its treatment of both the theory of function surface fitting and high speed computer techniques for data fitting.

BIBLIOGRAPHY

Chui, C.K. and L.L. Schumaker, eds. (1987) Topics in Multivariate Approximation. Academic Press (352 pp.)

Earnshaw, R.A., ed. (1985) Fundamental Algorithms for Computer Graphics. Springer-Verlag (1042 pp.)

Preisendorfer, R.W. and C.D. Mobley (1988) Principal Component Analysis in Meteorology and Oceanography. Elsevier Science (425 pp.)

Thiébaux, H.J. and M.A. Pedder (1987) Spatial Objective Analysis with Applications in Atmospheric Science. Academic Press (299 pp.)